Tampere University

LEI XU

Machine Learning Solutions for Classification and Regions of Interest Analysis on Imbalanced Datasets

Tampere University Dissertations 1003

Tampere University Dissertations 1003

LEI XU

Machine Learning Solutions for Classification and Regions of Interest Analysis on Imbalanced Datasets

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Information Technology and Communication Sciences
of Tampere University,
for public discussion in the TB109
of the Tietotalo Building, Korkeakoulunkatu 1, Tampere,
on Friday, 26 April 2024, at 12 o’clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Information Technology and Communication Sciences
Finland

Responsible supervisor and Custos	Professor Moncef Gabbouj Tampere University Finland	
Pre-examiners	Professor Mourad Oussalah University of Oulu Finland	Professor Naoufel Werghi Khalifa University United Arab Emirates
Opponent	Professor Guoying Zhao University of Oulu Finland	

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Cover design: Roihu Inc.

ISBN 978-952-03-3396-6 (print)
ISBN 978-952-03-3397-3 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-3397-3

HIILINEUTRAALI PAINOTUOTE
ClimateCalc CC-000025/FI
PunaMusta Printing

Carbon dioxide emissions from printing Tampere University dissertations have been compensated.

PunaMusta Oy – Yliopistopaino
Joensuu 2024

PREFACE/ACKNOWLEDGEMENTS

I would like to thank my esteemed supervisor - Prof. Moncef Gabbouj for his support, trust, and invaluable supervision during my M.Sc. and Ph.D. studies. He has provided me with the precious opportunity to enrich my life. My gratitude extends to the cooperators of all my works from the SAMI group: Prof. Jenni Raitoharju, Dr. Honglei Zhang, and Prof. Alexandros Iosifidis. Their immense knowledge and plentiful experience have encouraged me during my academic research.

I would also like to thank my friends, office mates, and colleagues for the cherished time spent together during my study. I am also thankful for the help from HR and IT departments.

Finally, my appreciation also goes out to my family and friends for their encouragement and support throughout my study. Special thanks to my husband for dedicating his time and effort for taking care of the baby, and thanks to my sweet baby for growing up healthily and happily.

Espoo, January 2024

Lei Xu

ABSTRACT

This dissertation investigates machine learning algorithms: Linear Discriminant Analysis (LDA) and Generative Adversarial Networks (GANs) to address real-world tasks suffering from imbalanced problems efficiently and robustly. The proposed LDA variants and GANs variants in this dissertation are used for classification and regions of interest analysis tasks with different severely imbalanced inputs.

This dissertation firstly contributes to addressing imbalance problems with LDA-related methods for binary-label and multi-label classification tasks. The traditional LDA has been widely used as a pre-processing step to enhance the efficiency and performance of sequential classifiers on datasets satisfying the Gaussian distribution. To extend the traditional LDA for uneven and diverse inputs, We initially introduce weight factors into the definition of scatter matrices based on a novel probabilistic saliency estimation method as saliency-based weighted LDA. Usually, the weight factors are exploited based on specific metrics with the label or feature correlation of input data. The redefinition can balance the sample contribution to mitigate the influence of outlier samples for binary-label classification tasks. The experimental performance of the proposed methods has been assessed over six publicly facial datasets, which demonstrates a robust improvement compared to the competing methods.

To address more complicated imbalanced problems widely existing in multi-label classification tasks, this dissertation introduces a saliency-based multilabel LDA framework. The proposed framework extends the probabilistic saliency estimation method for multi-label classification tasks based on multilabel LDA scatter matrices to alleviate the performance degradation caused by imbalanced problems. Six kinds of weight factors are obtained by the probabilistic saliency estimation method and the exploration of prior information with six metrics. The experimental results of the proposed methods over 17 imbalanced datasets show remarkable performance improvements compared to competing methods with seven quantitative evaluation metrics.

Alternative approaches to LDA include various deep learning models which have been explored for various regions of interest analysis tasks which may include imbalanced datasets in many applications. Especially, the GANs-related methods with appropriate deep neural networks can be used to address various regions of interest analysis tasks suffering from imbalance problems in computer vision areas efficiently and effectively, due to its excellent functionality of restoring balance in the problem. Another contribution of this dissertation is to explore GANs-based methods for two regions of interest analysis tasks with severely imbalanced inputs.

This dissertation investigates optimized Deep Convolutional Generative Adversarial Networks (DCGANs) to edit specific facial attributes, such as occlusions. The proposed method utilized a pre-trained DCGANs and an optimization loss function to detect the occluded facial regions and in-paint with corresponding facial attributes. The pre-trained DCGANs is trained with occlusion-free facial images to distinguish facial attributes and occlusions during the inference stage with the optimization function. The visual experimental results have shown that the proposed method can detect the required facial occlusions and then successfully in-paint them with the corresponding facial attributes.

Besides, a conditional Generative Adversarial Network (cGANs) based framework is introduced to detect anomalous regions (e.g., cracks) on pavement images efficiently and effectively in this dissertation. Such a task is also considered a binary semantic segmentation problem with imbalanced data due to uneven distribution between the number of the required anomalous region pixels and the background pixels. The proposed cGANs-based method consists of a UNet-based generator part for a multiscale feature representation, a discriminator part for real pairs and fake pairs judgment, and a novel auxiliary network for a refined feature representation. To increase performance while avoiding increasing network and computational complexity, the proposed framework is trained alternatively in two stages. The proposed methods have shown the effectiveness of GANs-based methods and their robustness in tackling binary semantics segmentation with severely imbalanced inputs through extensive experiments over six benchmark datasets.

CONTENTS

List of Figures

List of Tables

ABBREVIATIONS

AP	Average Precision
AUC	Area Under Curve
cGANs	Conditional Generative Adversarial Nets
DCGANs	Deep Convolutional Generative Adversarial Nets
diag	Diagonal Matrix
e.g.	For Example, from Latin *exempli gratia*
et al.	And Others, from Latin *et alii*
FN	False Negative
FP	False Positive
GA	Global Accuracy
HMM	Hidden Markov Model
IOU	Intersection-Over-Union
LDA	Linear Discriminant Analysis
LRR	Multioutput Linear Ridge Regressor
meanCIR	Mean Class Imbalance Ratio
meanIR	Mean Imbalance Ratio
ML-kNN	Multilabel k-nearest Neighbor Classifier
ODS	F-measure based on Optimal Dataset Scale
OIS	F-measure based on Optimal Image Scale
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic

SOTA	State-of-the-art
TN	True Negative
TP	True Positive
tr	Trace
UNet	U-shaped Encoder-decoder Network Architecture
VAE	Variational Autoencoder
wMLDA	A Weighted Multi-label LDA Framework

SYMBOLS

C	Number of classes
D	Discriminator
G	Generator
M	Mask matrix
N	Number of instances
N_c	Number of instances in class c
T	Threshold
$[\mathbf{P}]_{cj}$	All items in class c
$\hat{M}$	Learned mask matrix
$\mathbb{R}^D$	Original space
$\mathbb{R}^d$	Subspace
$\mathbf{1}$	Unit matrix
$\mathbf{A}$	Affinity matrix
$\mathbf{D}$	Diagonal matrix
$\mathbf{M}$	Weight factors matrix
$\mathbf{P}$	Probability matrix
$\mathbf{R}$	Correlation matrix
$\mathbf{S}_B$, $\mathbf{S}_b$	Between-class scatter matrix
$\mathbf{S}_T$, $\mathbf{S}_t$	Total scatter matrix
$\mathbf{S}_W$, $\mathbf{S}_w$	Within-class scatter matrix
$\mathbf{V}$	Prior information matrix

$\mathbf{W}$	Projection matrix
$\mathbf{X}$	Input data matrix
$\mathbf{Y}$	Label matrix
$\mathbf{Z}$	Discriminative feature matrix in an optimal subspace
$\hat{\mathbf{Y}}$	Predicted label matrix
$\mathbf{p}$	Probability vector
$\mathbf{p}^*_{pse}$	Optimized probability vector
$\mathbf{x}$	Input data vector
$\mathbf{y}$	Label vector
$\mathcal{G}_C$	Graph
$\mathcal{L}$	Loss function
$\boldsymbol{\mu}$	Total mean vector
$\boldsymbol{\mu}_c$	Mean vector of class c

ORIGINAL PUBLICATIONS

Publication I L. Xu, A. Iosifidis, and M. Gabbouj, "Weighted linear discriminant analysis based on class saliency information," in *2018 25th IEEE International Conference on Image Processing (ICIP 2018)*, Athens, Greece, 7-10 October 2018. DOI: 10.1109/ICIP.2018.8451614.

Publication II L. Xu, J. Raitoharju, A. Iosifidis, and M. Gabbouj, "Saliency-based multilabel linear discriminant analysis," *IEEE Transactions on Cybernetics*, vol. 52, Issue 10, October 2022, pp. 10 200–10 213, DOI: 10.1109/TCYB.2021.3069338.

Publication III L. Xu, H. Zhang, J. Raitoharju, and M. Gabbouj, "Unsupervised facial image de-occlusion with optimized deep generative models," in *2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Xi'an, China, 7-10 November 2018. DOI: 10.1109/IPTA.2018.8608127.

Publication IV L. Xu and M. Gabbouj, "Revisiting generative adversarial networks for binary semantic segmentation on imbalanced datasets," *arXiv preprint arXiv:2402.02245*, 2024.

1 INTRODUCTION

1.1 General Context

Pattern recognition, machine learning, and artificial intelligence-related algorithms have gained significant popularity from academic research areas to industrial applications during the past decades. When we develop pattern recognition or machine learning algorithms for problems in real applications, the heterogeneity of real-life datasets is an inevitable challenge. Hence researchers must investigate the features of datasets and then specifically develop efficient and robust algorithms that are validated and tested on real-world datasets without homogeneous distribution as assumptions. Failure to do so may result in algorithms' performance deterioration. This dissertation aims to explore specific pattern recognition and machine learning algorithms for classification and binary semantic segmentation tasks on imbalanced datasets that widely exist in real applications.

Imbalanced datasets are widely existing in various real applications. Usually, the number of instances from different classes may vary dramatically on imbalanced datasets. In certain scenarios, the minor classes (class) contain prominent information or contents as targets, such as object detection, fraud transaction detection, and medical diagnosis. Instead of deploying general methods under even distribution assumption directly, specific methods considering the characteristics of imbalance inputs are inevitable to reduce performance deterioration. Methods explored to address imbalance problems can be categorized as data-driven methods, algorithm-driven methods, and hybrid methods [1]. Data-driven methods [2] directly utilize the uneven distribution of datasets to restore the balance of the datasets through resampling and augmentation. Algorithm-driven methods [3], [P2] exploit specific algorithms to highlight the contribution of minority classes according to their prominence. Hybrid methods [4] utilize both data-driven and algorithm-driven methods.

Dimensionality reduction algorithms as traditional machine learning methods have been widely explored to solve classification tasks as a pre-processing step. Usually, it is used to extract the prominent information from a high-dimension original input to enhance the performance of classifiers ultimately. Linear discriminant analysis (LDA) is a traditional dimensionality reduction algorithm often used to efficiently solve binary-label classification tasks in datasets that follow a Gaussian distribution. This dissertation explores LDA-based techniques with specific weight factors to extend traditional LDA techniques for solving imbalance problems existing in both binary-label and multi-label classification tasks. LDA can be extended as an algorithm-driven method to tackle imbalance problems using novel weight factors to balance the contribution of prominent but minor instances. The weight factors usually can be exploited according to the features of class labels or instances [5]. After the projection of the original imbalanced data into an optimal subspace with LDA-related techniques, the subsequent classification performance can be enhanced.

Besides traditional machine learning methods, deep learning-related methods have achieved significant outcomes in tackling various imbalance problems in computer vision areas [1], [6], such as image segmentation and object detection. Regions of interest analysis is a typical kind of computer vision task whose performance often suffers, in part, due to imbalanced datasets wherein the regions of interest constitute only a small portion of pixels in an image. Generative adversarial networks (GANs) related techniques have been varied as effective methods for mitigating imbalance problems existing in the region of interest segmentation tasks [1], [7], [8]. In this dissertation, we introduce deep convolutional generative adversarial networks (DCGANs) with an optimized novel loss function to edit facial occlusions with the corresponding facial attributes. Moreover, inspired by the significant performance of conditional GANs-based networks on style transferring and medical image segmentation tasks, a framework based on the conditional GANs [9] is proposed to solve anomalous crack pixels detection in pavement datasets, which can be considered as a binary semantic segmentation task with imbalanced problems.

1.2 Objectives

The scope of the dissertation is limited to two areas: targeting imbalance problems existing in classification and regions of interest analysis. The overall schema of proposed methods and the corresponding publications in this dissertation is demonstrated in

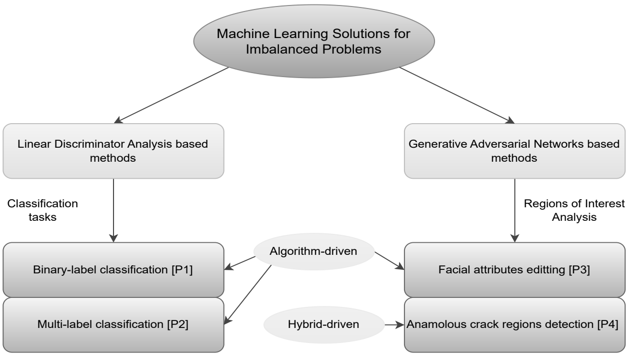

Machine Learning Solutions for Imbalanced Problems
Linear Discriminator Analysis based methods
Generative Adversarial Networks based methods
Classification tasks
Regions of Interest Analysis
Algorithm-driven
Binary-label classification [P1]
Facial attributes editting [P3]
Multi-label classification [P2]
Hybrid-driven
Anamolous crack regions detection [P4]

Figure 1.1 The overview of the dissertation with indication of proposed methods and the corresponding publications

Fig. 1.1.

The first research objective aims to enhance the performance of classifiers after using dimensionality reduction algorithms LDA to balance the distribution of imbalanced datasets. To achieve this goal, we explore a probabilistic estimation approach to investigate various weight factors for balancing the contribution of each instance in the original imbalanced datasets.

LDA is usually used in classification tasks as a pre-processing step to enhance the ultimate performance of classifiers on Gaussian distribution datasets. When tackling binary-label classification tasks on imbalanced datasets, weighting factors are introduced into the definition of scatter matrices as an extension for imbalanced datasets. Moreover, multi-label classification tasks can be efficiently solved with LDA-based techniques incorporating with prior information exploration. We therefore formulate the first research question as follows:

Research question 1: How to explore saliency information to properly highlight prominent but minor instances in an imbalanced dataset with the LDA for an optimal result of the subsequent classification tasks? To begin we wish to explore the saliency information of instancea inside each class with a probabilistic estimation model. The probabilistic estimation model was originally designed to segment salient objects in images. We extend the original proposal to describe the importance of each instance for its corresponding classes with a probability vector. Following that we investigate several types of prior information to obtain various probability vectors based on the label or feature relation. Finally, a probability matrix is obtained as the weight factors which is embedded into the definitions of scatter matrices. The ultimate objective is to alleviate the data imbalance problem for binary-label and multi-label classification tasks and possibly avoid the common over-counting problem generally existing in multi-label classification tasks with algorithm-driven methods.

The second research task is related to regions of interest analysis tasks with imbalanced image inputs. The objective of the second research task is to explore the GANs-based networks to solve such tasks with effectiveness and robustness on severely imbalanced datasets. The effectiveness of GANs-based techniques has been varied for solving various regions of interest analysis tasks suffering from imbalance problems through the massive, related works in computer vision areas recently. The diverse network topologies of the GANs-based method can be explored and incorporated with loss functions according to the requirements of a specific target. Our main second research question is then formulated as follows:

Research question 2: Can we develop GANs-based techniques working with different losses to carry out regions of interest analysis tasks with severely imbalanced inputs? The first application is to edit the specific facial attributes with occlusions using the DCGANs and a novel optimal loss function as an algorithm-driven method. The DCGANs trained with occlusion-free facial images can distinguish the pixels presenting occlusions and facial attributes under the support of the optimized loss function. The second application explores a conditional GANs-based framework to locate the pixels for anomalous cracks and generate a probability feature map indicating them in pavement images. A novel auxiliary network is introduced to train the model in two stages alternately and iteratively for a refined multiscale feature map. Moreover, the conditional GANs-based network works with different attention mechanisms to further obtain an enhanced and robust performance on diverse benchmark datasets. The second application is a hybrid-driven method with augmented data.

1.3 Dissertation Outline

The dissertation is structured as follows. Chapter 2 presents the backgrounds of the two main research topics investigated in the dissertation: linear discriminant analysis and generative models. Moreover, the evaluation metrics used in the dissertation are described in Chapter 2. Chapter 3 summarizes the main contributions of publications [P1, P2], which are related to linear discriminant analysis for classification tasks. The main results of publications [P3] and [P4] focusing on generative models for the regions of interest or saliency segmentation tasks are summarized in Chapter 4. Chapter 5 presents the conclusion of this dissertation and highlights possible future work.

Author's contribution

Publication I The proposed method is based on Linear Discriminant Analysis for enhancing the performance of classification tasks as a pre-processing step. The novelty of this work is the weight factor calculated based on a probabilistic estimation approach for the first time. The candidate developed the theory, performed the experiments, and wrote the paper. The co-authors have supervised, reviewed, and edited the publication.

Publication II In this work, a general framework based on weighted Linear Discriminant Analysis was proposed to boost the performance of classifiers on multi-label classification tasks. The framework is based on a probabilistic estimation approach by which the imbalance problem existing in multi-label datasets can be avoided. The candidate developed the theory, performed the experiments, and wrote the paper. The co-authors have supervised, reviewed, and edited the publication.

Publication III In this work, a method based on the Deep Convolutional Generative Adversarial Networks was proposed to detect and segment occluded facial attributes. The novelty of the proposed method is that it can segment the occluded facial attributes and generate the corresponding facial features simultaneously. The candidate developed the theory, performed the experiments, and wrote the paper. The co-authors have supervised, reviewed, and edited the publication.

Publication IV In this work, we revisited the cGANs-based algorithms to address anomalous crack detection in a pixel-to-pixel manner on severely imbalanced datasets. A cGANs-based autoencoder is adopted as the backbone incorporating attention mechanisms and entropy-based loss functions for a robust multiscale feature representation. The candidate developed the theory, performed the experiments, and wrote the paper. The co-authors have supervised, reviewed, and edited the publication.

2 RELATED WORKS AND BACKGROUND

This chapter presents an overview of the related works and background knowledge related to this dissertation. Firstly, the works related to LDA for classification tasks are described in detail in Section 2.1, to provide the theoretical foundation of our publications in the following chapter. Then, we provide an introduction to between-class imbalance problems in Section 2.2. The works related to regions of interest analysis using generative models are depicted in Section 2.3. The evaluation metrics used in the publications are described in Section 2.4.

2.1 Dimensionality Reduction for Classification Tasks

With the emergence of massive data in all walks of life, dimensionality reduction techniques have been widely explored to extract distinctive information from data with high dimensionality to decorrelate redundant raw data and enhance the performance of subsequent tasks on large-scale datasets [10]. Usually, dimensionality reduction-related algorithms can be categorized as supervised and unsupervised algorithms depending on whether label information is involved. Dimensionality reduction techniques have been widely used in multiple disciplines, such as image processing [11] and data compressing [12].

In this dissertation, we focus on binary-label and multi-label classification tasks solved with dimensionality reduction techniques. Binary-label classification tasks aim to determine whether one instance belongs to a given class label or not. The difference is that each instance can be associated with one or several class labels for a multi-label classification task. Another significant characteristic of multi-label datasets is the class-imbalanced problem [13], wherein the number of instances for each class varies dramatically. Therefore, the multi-label classification is more complicated. To tackle multi-label classification tasks efficiently and effectively, it is better to exploit the correlation and dependency of both data features and labels. Researchers have proposed various dimensionality reduction algorithms to tackle both binary-label and multi-label classification tasks on various datasets [3], [5], [10], [14].

Principle Component Analysis (PCA) is an unsupervised dimensionality reduction algorithm deriving the orthogonal direction of maximal variance using the feature covariance matrix for eigen-decomposition [10]. Canonical Correlation Analysis (CCA) is a supervised dimensionality reduction algorithm formulated as a least-squares problem, which uses both label and data feature information for maximally correlated projections [15]. CCA has been widely extended to improve classification performance [15], [16]. Multi-label informed latent semantic indexing (MLSI) algorithm [17] utilized the well-known unsupervised approach latent semantic indexing (LSI) to preserve the information of inputs and capture the correlations between the multiple outputs. Multi-label dimensionality reduction via dependence maximization (MDDM) [18] is proposed to tackle multi-label classification tasks with two kinds of projection strategies. The lower-dimensional space identified by MDDM is projected by maximizing the dependence between the original feature and class labels of each instance based on the Hilbert-Schmidt Independence Criterion (HSIC).

LDA is a typical supervised dimensionality reduction method using distinctive information on an optimal sub-space extracted from the high-dimensional input. The original LDA is under the assumption of the Gaussian distribution of input data, which cannot be used directly on imbalanced datasets. In the following subsection, we present the theoretical basis of how LDA can be used to tackle both binary-label and multi-label classification tasks with uneven distribution inputs. We present the mathematical theories and notations used in LDA techniques for binary-label classification tasks and multi-label classification tasks separately in the following subsections.

2.1.1 Linear Discriminant Analysis on Binary-label Datasets Classification

Standard linear discriminant analysis (LDA) is a well-known statistical algorithm for dimensionality reduction under the assumption of Gaussian distribution, which aims to identify an optimal sub-space using Fisher criterion optimization [19]. Given a binary-label dataset with N instances $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i, ..., \mathbf{x}_N\}$ and $\mathbf{x}_i \in \mathbb{R}^D$, where D is the original data dimensionality. The corresponding label vector is defined as $\mathbf{y} = \{y_1, y_2, ..., y_i, ..., y_N\}$ and $y_i \in \{1, ..., C\}$, where C is the number of classes. The standard LDA learns an optimal projection matrix $\mathbf{W} \in \mathbb{R}^{D \times d}$ mapping original input into a discriminant subspace $\mathbb{R}^d$.

To obtain the optimal projection matrix $\mathbf{W}$, the standard LDA firstly defines within-class scatter matrix $\mathbf{S}_W$, between-class scatter matrix $\mathbf{S}_B$, and total scatter matrix $\mathbf{S}_T$ as follows:

$$\mathbf{S}_W = \sum_{c=1}^{C} \sum_{\mathbf{x}_i, \alpha_i^c = 1} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T, \tag{2.1}$$

$$\mathbf{S}_B = \sum_{c=1}^{C} N_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T, \tag{2.2}$$

$$\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W. \tag{2.3}$$

Here, $\boldsymbol{\mu}_c$ denotes the mean vector of class c as

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{\mathbf{x}_i, \alpha_i^c = 1} \mathbf{x}_i, \tag{2.4}$$

where $N_c = \sum_{i=1}^{N} \alpha_i^c$ is the cardinality of class c, $\alpha_i^c = 1$ if $y_i = c$, otherwise $\alpha_i^c = 0$. The total mean vector $\boldsymbol{\mu}$ is computed as

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i, \tag{2.5}$$

Based on the above definitions, the optimal projection matrix $\mathbf{W}$ is learned by maximizing the Fisher's discriminant criterion [19] as

$$J(\mathbf{W}) = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{\operatorname{tr}(\mathbf{W}^T \mathbf{S}_B \mathbf{W})}{\operatorname{tr}(\mathbf{W}^T \mathbf{S}_W \mathbf{W})}, \tag{2.6}$$

where tr(.) denotes the trace of a matrix. This allows obtaining the projection matrix $\mathbf{W}$ by solving the generalized eigenvalue problem

$$\mathbf{S}_b \mathbf{w} = \mathbf{S}_w \lambda \mathbf{w}. \tag{2.7}$$

Ultimately the projection matrix $\mathbf{W}$ contains $d \leq C - 1$ eigenvectors as columns. The resulting subspace's maximal dimensionality equals the rank of $\mathbf{S}_b$ as $C - 1$.

Besides, the trace ratio problem can be solved using different iterative methods [20], [21].

The projection matrix $\mathbf{W}$ can be obtained with $\mathbf{S}_T$ instead of $\mathbf{S}_W$ using maximizing the Fisher's discriminant criterion, as

$$J(\mathbf{W}) = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{\operatorname{tr}(\mathbf{W}^T \mathbf{S}_B \mathbf{W})}{\operatorname{tr}(\mathbf{W}^T \mathbf{S}_T \mathbf{W})}. \tag{2.8}$$

Finally, the discriminative features on the optimal sub-space can be obtained as

$$\mathbf{Z} = \mathbf{W}^T \mathbf{X}, \tag{2.9}$$

where $\mathbf{Z} \in \mathbb{R}^{d \times N}$.

The underlying assumption of the standard LDA is that the dataset classes are equally distributed as a homoscedastic Gaussian model [22] with identical covariance matrices [23]. Otherwise, the performance of LDA is affected severely due to the imbalance of input datasets [24] originating from the excessive contribution of outlier classes and leading to an inferior projection matrix.

To balance the contribution of imbalance input, weight factors have been introduced into the definitions of scatter matrices in various related works [23]–[25] for binary-label classification tasks. Weight factors are used to balance the contribution of each class based on their real contribution by exploring appropriate prior information of the original input. For instance, in [26], the between-class scatter matrix is redefined for enhancing robustness in multi-class binary-label problems as

$$\mathbf{S}_b = \sum_{c=1}^{C-1} \sum_{j=c+1}^{C} L_{cj} p_c p_j (\boldsymbol{\mu}_c - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_c - \boldsymbol{\mu}_j)^T, \tag{2.10}$$

where p_c, p_j denote the prior probability of class c, class j, respectively. A distance function based on the Euclidean (or a Mahalanobis) space is used as the dissimilarity factor L_{cj} between class c and class j. Moreover, an outlier-class-resistant weighted LDA method is proposed in this work [23] based on Loog's work [26] for the influence reduction of outlier classes. The between-class scatter using Eq. (2.10). The within-class scatter is re-defined as follows:

$$\mathbf{S}_w = \sum_{c=1}^{C} \sum_{k=1}^{N_c} p_c r_c (\mathbf{x}_k - \boldsymbol{\mu}_c)(\mathbf{x}_k - \boldsymbol{\mu}_c)^T, \tag{2.11}$$

where $r_c = \sum_{i \neq c} \frac{1}{L_{ic}}$ is a relevance-weight between class c and class i, reducing attention to outlier classes.

Although various LDA variants have been proposed to conquer the drawbacks in traditional LDA, current weighted LDA variants for binary-label classification tasks merely have redefined the scatter matrices using weight factors based on the exploration of class similarity information as in [23], [25], [26]. The contribution of each instance to its class information has been neglected. To overcome the limitations, an algorithm-driven method was proposed in [P1] to further explore the prior information based on the correlation of each instance to its associated class. In [P1], the proposed method redefined class representation and scatter matrices to balance the prominence of each instance based on a novel saliency probabilistic estimation method.

2.1.2 Linear Discriminant Analysis on Multi-label Datasets Classification

As mentioned, tasks on multi-label datasets usually suffer from the occurrence of imbalanced problems [1]. Because the number of instances of major classes is larger than that of minor classes in an imbalanced dataset as shown in Fig. 2.1. Ignorance of this characteristic of multi-label datasets can lead to the deterioration of algorithm performance [1] while employing general algorithms directly. Various methods have been proposed specifically for solving multi-label classification tasks, such as variants of Support Vector Machine (SVM) [27] and several feature extraction methods [5], [28], [29]. As described in [30], multi-label classification methods are typically derived following either an algorithm adaptation (AA) approach or a problem transformation (PT) approach. Methods following the AA approach directly utilize the information of class labels and data instances to explore their correlation. Methods following the PT approach utilize single-label classification algorithms to tackle multi-label classification tasks using decomposition strategies.

LDA-related algorithms following the PT strategy have been proposed to tackle the multi-label classification tasks [10]. In a multi-label dataset, there exist two char-

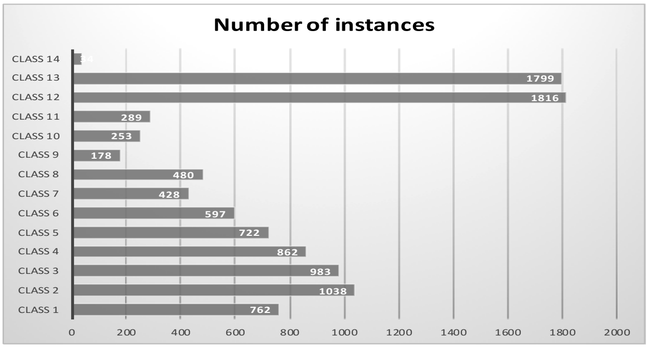


Figure 2.1 The details of the Yeast database [P2]

acteristics compared to a binary-label dataset as correlations or dependencies [31] of label information and imbalance problems. Hence, the scatter matrics definitions of the traditional LDA and its variants as Eqs. (2.1) - (2.2) or Eqs. (2.10) - (2.11) cannot be used directly for multi-label classification tasks [31].

To introduce LDA-related algorithms for solving multi-label classification tasks, it is necessary to redefine the scatter matrices with the characteristics exploration of datasets. Assuming a multi-label dataset with N data items as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i, ..., \mathbf{x}_N\}$ wherein $\mathbf{x}_i \in \mathbb{R}^D$ and $\mathbf{X} \in \mathbb{R}^{D\times N}$, where D is the original data dimensionality. $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_i, ..., \mathbf{y}_N\}$ is defined as the corresponding label matrix wherein $\mathbf{y}_i \in \{0, 1\}^C$ and $\mathbf{Y} \in \mathbb{R}^{C\times N}$, C is the number of classes. Each row in $\mathbf{Y}$ depicts as as $\mathbf{y}_{(j)}$, where $j \in 1, \ldots, C$. When a data item $\mathbf{x}_i$ is associated with class c, an element y_{ci} in the label matrix is 1, otherwise 0.

Multi-label linear discriminant analysis (MLDA) [28] is a typical LDA-related approach proposed to tackle classification tasks on multi-label datasets. The MLDA approach introduces weight factors based on label correlation information on multi-label datasets. Therefore, a matrix $\mathbf{M} \in \mathbb{R}^{C\times N}$ containing non-negative values with the same size as $\mathbf{Y}$ is introduced to present weight factors, which is defined as

$$\mathbf{M} = [\mathbf{m}_1, ..., \mathbf{m}_i, ..., \mathbf{m}_N] = [\mathbf{m}_{(1)}, ..., \mathbf{m}_{(j)}, ..., \mathbf{m}_{(C)}]^T, \qquad (2.12)$$

where each element m_{ci} presents weight factor for the i^{th} instance from class c, the weight vectors for i^{th} instance and the j^{th} class are depicted as $\mathbf{m}_i$ and $\mathbf{m}_{(j)}$ separately. Then a correlation matrix $\mathbf{R} \in \mathbb{R}^{C\times C}$ is defined based on label correlations of class pairs as

$$R_{kl} = \cos(\mathbf{y}_{(k)}, \mathbf{y}_{(l)}) = \frac{\mathbf{y}_{(k)}^T \mathbf{y}_{(l)}}{\|\mathbf{y}_{(k)}\| \|\mathbf{y}_{(l)}\|}, \tag{2.13}$$

where $\mathbf{y}_{(k)}$, $\mathbf{y}_{(l)}$ are label vectors for a class pair $k, l \in 1, ..., C$. The weight matrix M is obtained as $\mathbf{M} = \mathbf{RY}$. The weight factors are normalized with ℓ_1-norm to avoid the over-counting problem [28]:

$$\mathbf{m}'_i = \frac{\mathbf{m}_i}{\|\mathbf{y}_i\|_{\ell_1}}. \tag{2.14}$$

The label correlation matrix can reveal the relation between classes, as the more closely related the classes produce a higher label correlation value. However, the definition of the weight matrix M may cause an overcounting problem.

Direct MLDA [32] is an extension of MLDA used for performance enhancement in multi-label video classification tasks [28]. Due to the definition of between scatter matrix in the original MLDA, the subspace dimensionality is limited by the rank of between scatter matrix as $C - 1$. The re-definition of between scatter matrix in [32] can lead to a subspace with a higher dimensionality. There are many other MLDA-related extensions. For instance, multi-label discriminant analysis with locality consistency (MLDA-LC) [33] is another excellent work further enhancing the classification performance in multi-label data sets compared to MLDA and MLLS algorithms. MLDA-LC incorporates the graph Laplacian matrix into the MLDA method in the projection space to investigate the similarity among adjacent instances.

Although the current MLDA-related methods have been proven to tackle classification tasks on multi-label datasets [3], [28], [33] under various considerations, the imbalance problems still need more attention to enhance the robustness of LDA-related algorithms on multi-label classification tasks. Our method Saliency-Based Multilabel Linear Discriminant Analysis (SMLDA) in [P2] was proposed to explore either the correlation of data or the definition of scatter matrices as a novel variant for multi-label classification. The proposed SMLDA exploits the prior information of input data using the probabilistic saliency estimation method as weight factors to highlight the contribution of minority classes, which is a typical algorithm-driven method for imbalanced problems. In [P2], the weight factors are calculated within each class for each instance, which can alleviate the influence of outlier classes to avoid imbalance problems and conquer the overcounting problem. Additionally, six types of prior information are exploited to reveal the correlation of data features or labels. Furthermore, a simpler definition of scatter matrices in [5] is incorporated into the weight factors for the optimal projection matrix $\mathbf{W}$.

2.2 Overview of Between-Class Imbalance Problems

The datasets existing in real-world tasks can rarely satisfy the Gaussian distribution perfectly [1], [34], [35], which leads to imbalance problems with the dramatic degradation performance of generic machine learning algorithms. Therefore, it is necessary to investigate specific algorithms for datasets with uneven distribution when employing machine learning solutions for solving real-world applications. The imbalanced datasets can be categorized as between-class imbalanced datasets and within-class imbalance datasets [1]. The between-class imbalanced datasets contain a minor class with a smaller number of instances and a major class with a larger number of instances. The within-class imbalanced datasets are defined based on attributes imbalance of the class. This dissertation focuses on solving the between-class imbalance problems in [P3] and [P4].

The between-class imbalanced datasets widely exist in real-world applications, such as medical image segmentation [35]–[37], anomalous region detection [38], [39], and small objects detection [34]. Usually, the minor class in between-class imbalanced datasets contains significantly important information as the regions of interest to determine the effectiveness of the method. Intense efforts have been devoted to tackling the between-class imbalance problems. Data-driven techniques [40] enhancing the performance of deep learning models on imbalanced datasets, mainly constitute basic image manipulations, kernel filters, adversarial training, feature space augmentation, etc. The basic image manipulations utilize image transformation such as cropping, rotation, and flipping. In [P4], the basic image manipulations cropping and rotation are used to augment the raw input datasets. Moreover, algorithm-driven techniques are exploited to highlight the importance of the minor class by weights trade-off in objective functions, such as focal loss function [41], and Tversky loss function [42]. The proposed method is algorithm-driven in [P3] and a hybrid-driven method in [P4].

2.3 Generative Models for Regions of Interest Analysis

As mentioned earlier, regions of interest often represent a minor portion of data, hence the risk of encountering imbalanced problems is high in related applications. Usually, regions of interest can be defined differently according to the concrete tasks. For instance, retinal vessels on fundoscopic images are the interested regions when automatically detecting retinal diseases [36], abnormal wafer defect patterns need to be identified to prevent loss excursion in the semiconductor manufacturing [39], and various small objects exist on the road for autonomous driving [34]. The common characteristic of such tasks is that precise analysis of the interested regions is crucial to the success of subsequent tasks. The topic of anomalous regions of interest analysis discussed in this dissertation focuses on binary pixel-level tasks, as the pixels representing the required anomalous parts are positive instances and the pixels representing the background are negative instances [1].

Traditional methods for such applications explore different prior knowledge of input images, for instance, the statistical model as a log spectrum [43], gradient change between adjacent pixel values [44] and probabilistic estimation [45]. Naturally, with the vigorous development of deep learning, research works using deep learning frameworks to tackle regions of interest analysis have sprung up in the past decade [46]–[53]. In particular, UNet has gained popularity and effectiveness for regions of interest analysis. UNet was proposed by Ronneberger et. al initially in 2015 [46], and since then Unet related deep learning networks have been widely used to solve various regions of interest tasks [54]–[56]. UNet [46] is a symmetric architecture with a contracting path and an expanding path aiming to localize regions of interest precisely.

Deep generative models such as generative adversarial networks (GANs), diffusion models, and variational autoencoders (VAE), have been gaining popularity in the artificial intelligence or machine learning area recently [57]–[59]. Deep generative models, as statistical models, usually can be used to estimate the joint distribution of the target and original input data for new data generation [57]. With the prosperous development of deep learning in multiple disciplines, deep generative models have been successfully explored combining the UNet structure in regions of interest or anomaly detection applications [1], [48], [59]–[62]. Deep generative adversarial networks (GANs) have a significant advantage in restoring the imbalanced data and preventing performance degradation in an unsupervised or self-supervised manner when common computer vision algorithms are used [1].

In this section, we describe how deep GANs with the UNet structure can be used to tackle pixel-level regions of interest or anomaly detection tasks in the relevant literature.

2.3.1 Regions of Interest Detection Tasks on Facial images

Facial attributes editing is another imbalanced application in the computer vision area, due to the common existence of attribute level imbalance in facial datasets [1]. It is a challenging task due to the distributions of different attributes vary between local-wise and global-wise. The target of facial attribute editing is to manipulate specific facial attributes (regions of interest) for a new facial image with the desired attributes while persevering others.

Deep generative models [44], [63], [64] such as GANs, diffusion models and VAE are suitable solutions for tackling facial attribute editing, not only because of imbalance problems but also lack of ground truth images. Usually, facial attributes can be depicted in latent spaces with GANs or VAE as face attributes latent representation. The target of facial attribute editing can be achieved by proper manipulation of the face latent vectors. In [63], a GANs model with an encoder-decoder architecture was designed which manipulates facial attributes by modeling the relation between the attributes and the facial image latent representation. To preserve a great number of details, an attribute classification constraint was proposed to bridge the reconstruction learning and the adversarial learning in [63]. In [65], a robust LSTM-Autoencoders (RLA) containing two LSTM components was proposed to detect and restore partially occluded faces in the wild. The RLA model aims to generate a latent representation for occluded facial restoration. Moreover, GANs-based methods [52] have been proven through extensive research works as effective unsupervised solutions for anomaly detection or outlier detection tasks. Generally, GANs-related architectures can estimate the ideal distribution of normal instances in the feature latent space, therefore the anomaly samples can be distinguished through the latent space.

Although the current state-of-art works have achieved significant outcomings on facial attribute editing tasks, most of them either require constrained datasets for a fixed type of target or multiple and complicated models for multiple attribute manipulation [63]–[66]. To conquer these limitations, we proposed an unsupervised method for specific facial attribute manipulation based on DCGAN and an optimization method in our work [P3]. [P3] aims to detect and remove facial occlusion parts (scarf and sunglass) and in-paint the corresponding desired occluded attributes on general portrait datasets while preserving other attributes with a simple GANs-based model. Considering limited inputs and lack of ground truths, the network Deep Convolutional GANs (DCGANs) were used to train a normal facial feature latent space first, then during the inference phase, an optimization method embedded into the trained DCGANs was used to detect the anomalous parts (scarves or sunglasses) and in-paint the anomalous parts with corresponding facial attributes.

2.3.2 Anomalous Regions of Interest Detection Tasks

Anomaly detection has greatly benefited from the recent advances in machine learning and computer vision. Anomaly conditions can be defined diversely according to a specific task. In finance, typical anomalies can be defined as illegal activities beyond normal financial services such as inside trading, fraud, and money laundering [67]. In industrial manufacturing, anomalies may consist of defective samples or patterns with apparent discrepancies from the required samples or patterns [39], [68], whereas in medical image analysis, anomalies usually refer to the deviation from normality or normal state, such as lesions on brain images [69] and retinal vessels in fundus images [70]. A significant characteristic of such types of anomaly detection tasks is that anomaly conditions are often unusual phenomena compared to the more common conditions, which often leads to imbalanced problems. For example, the number of pixels representing the retinal vessels is less than the number of pixels for the other parts. Hence, a key challenge in such tasks is how to improve the accuracy of anomalous region detection in the presence of imbalance problems. Considering the large diversity of anomaly conditions in different areas, we only focus on anomaly detection works related to pavement crack detection in this section.

Traditional image processing and machine learning techniques have been intensely employed in this field. For instance, traditional sensors can be used to detect anomalous cracks and potholes on pavements. In [71], the authors designed a crack-detecting robot equipped with infrared, ultrasonic, and vibration sensors based on ARM processor to detect road cracks, which usually require a complex and expensive hardware system. In [72], the authors proposed a threshold-based algorithm to generate a set of candidate regions with high-intensity values, which consist of pothole regions, shadow regions, and other regions. Then a decision tree algorithm with four steps was employed to eliminate the false positive cases as shadow regions or other regions and retain the pothole regions. Edge information in an image can also be explored to distinguish abnormal road surface conditions. In [73], the authors adopted two stages of prepossessing and detection to detect various cracks. In the prepossessing stage, they used contrasted algorithms to obtain the edge information of cracks. Afterward, a decision tree algorithm was applied to detect and classify the road surface images with cracks automatically.

Besides the above traditional algorithms, various deep neural network (DNN) structures have been proposed for such anomalous region detection tasks, driven by the excellent performance of DNN on semantic segmentation. In [74], a feature pyramid and hierarchical boosting network (FPHBN) was proposed to detect cracks. The architecture of FPHBN used the feature pyramid network to integrate context information from top to bottom and layer by layer. To balance the contribution of easy samples and hard samples, hierarchical boosting was introduced following the feature pyramid network. In [39], the anomalous states are basic defect patterns and unseen defect patterns of wafer maps for semiconductor manufacturing. Then the performance of deep convolutional encoder-decoder neural networks based on Seg-Net, U-Net, and FCN were compared in the detection and segmentation of the defect wafer map patterns.

In particular, the between-class imbalance problems widely exist in medical image segmentation tasks. Attention UNet [75] is a variant of the original UNet architecture [46], which has embedded the outputs of attention gates into the expansive path to retain salient features of organs for medical image segmentation. In [76], a vision transformer-based architecture with a gated axial-attention model is proposed to explore the long-range dependencies of input images. Another vision transformer-based architecture in [77] exploits fine-grained context and coarse-grained context using local self-attention, global self-attention, and axial attention modules.

Moreover, generative models with deep neural network structures such as autoencoder (AE) or Generative Adversarial Networks (GANs) have been extensively developed to address pixel-level imbalances in segmentation, due to the potential to restore balance in imbalanced datasets [1]. In [78], auto-encoders and GANs were used to obtain pixel-level anomaly scores, to support doctors' diagnosing works from

X-ray images. Nguyen et al. [79] introduced an extension of conditional GANs for shadow detection tasks using a shadow detection generator with an additional sensitivity parameter for different shadow maps. Due to the imbalanced distribution of required shadow labels, weighted cross entropy was used in the loss function. Zhang et al. investigated the performance of transformer-based UNet architectures in pavement surface crack detection in [61]. As depicted in [61], the transformer-based UNet architectures can thoroughly explore the low-level information and global context information.

Although the current SOTA works have achieved significant detection results on various datasets solving between-class imbalance problems. These works still have limitations, such as a specific algorithm required to pre-process images, unstable performance on different datasets, and complexity increasing for better performance. Especially when the anomalous regions are related to the inhomogeneity of cracks under complex environments, the existing SOTA works cannot achieve robust and consistent results on various datasets. Inspired by the current state-of-art imbalance semantic segmentation works [1], [36], [61], [79], [80], we explored the cGANs-related architectures as a backbone to tackle the anomalous region detection tasks in [P4] for robust and outperforming results without increasing computational complexity. To achieve robust performance on six datasets, a novel and simple auxiliary network was proposed to assist the backbone for a refined probability feature map. Moreover, we investigated various attention mechanisms and loss functions to further explore the affecting factors of robustness on datasets with different characteristics (images captured under low-intensity, small and imperceptible cracks, or complex cracks).

2.4 Performance Metrics

This section presents the performance evaluation metrics used in the dissertation. We describe the general notations used in the evaluation metrics first. TP describes the number of correctly predicted samples or pixels, TN is the number of correctly predicted irrelevant samples or background pixels, FP is the number of wrongly predicted samples or pixels, and FN is the number of wrongly predicted irrelevant samples or background pixels.

In [P1], we used accuracy to evaluate the classification performance of the proposed methods. The accuracy calculated based on the confused matrix is commonly used to evaluate the performance of binary classification tasks [81], due to its less complexity and generalization. The definition of accuracy is shown as follows:

- ***Accuracy*** (↑) indicates the ratio of correct prediction [82] to all instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.15}$$

In [P2], we used seven evaluation metrics for multi-label classification tasks: ranking loss, one error, normalized coverage, Macro-AUC, Micro-AUC, Macro-F1, and Micro-F1. Moreover, we adopted the Friedman test and Wilcoxon signed-rank test [83] to verify the effectiveness of the proposed method SMLDA in [P2]. The evaluation metrics for multi-label classification can be divided as [84] *label-wise effective* and *instance-wise effective* based on the optimization methods. The metrics optimized by *label-wise effective* classifiers can distinguish the relevant classes from the irrelevant classes for every sample [84]. On the other hand, the metrics optimized by *instance-wise effective* classifiers can distinguish between relevant and irrelevant samples for each class.

To illustrate the evaluation metrics for multi-label classification tasks clearly, first, we describe the notations used in the evaluation metrics. $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_i, \dots, \mathbf{y}_M]$ presents M ground truth label matrix and its corresponding predicted label matrix is $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, \dots, \hat{\mathbf{y}}_i, \dots, \hat{\mathbf{y}}_M]$ for the test sample. Each test sample $\mathbf{x}_i$ has its corresponding ground truth label vector as $\mathbf{y}_i \in \mathbb{R}^C$. $\hat{\mathbf{p}}_i = f(\mathbf{x}_i)$ denotes the output of classifiers, wherein $\hat{p}_{i,c}$ indicates whether instance i is from class c with a probability. After setting thresholds on $\hat{\mathbf{p}}_i$, the final $\hat{\mathbf{y}}_i$ is generated. $\mathcal{L}_i = \{\mathrm{sort}_c(\hat{\mathbf{p}}_i)\}$ presents descending order of $\hat{\mathbf{p}}_i$. The relevant classes in $\mathbf{y}_i$ are $\mathcal{I}(\mathbf{y}_i)$ and the negative classes in $\mathbf{y}_i$ are $\neg\mathcal{I}(\mathbf{y}_i)$. (↓) denotes that the lower values with the metrics, the better performance, and the (↑) presents the opposite case.

- ***Ranking loss*** (↓) evaluates the fraction of reversely ordered relevant versus irrelevant pairs for each item i as in [84], [85]. Ranking loss is optimized based on label-wise effectiveness in [P4].

$$\text{ranking\_loss}_i = \frac{|\hat{p}_{i,\mathcal{I}(\mathbf{y}_i)} \leq \hat{p}_{i,\neg\mathcal{I}(\mathbf{y}_i)}|}{m * n}, \tag{2.16}$$

$$\text{ranking\_loss} = \frac{\sum_{i=1}^{M} \text{ranking\_loss}_i}{M}, \tag{2.17}$$

where $|\hat{p}_{i,\mathcal{I}(\mathbf{y}_i)} \leq \hat{p}_{i,\neg\mathcal{I}(\mathbf{y}_i)}|$ describes the number of reversely ranked pairs for each item i.

- *One error* (↓) indicates the fraction of the top ranked class for item i is not among the positive ground truth labels as in [84], [85] based on label-wise effectiveness.

$$\text{one\_error}_i = \begin{cases} 0, & \text{if } \mathcal{L}_i[1] \in \mathcal{I}(\mathbf{y}_i), \\ 1, & \text{otherwise}, \end{cases} \tag{2.18}$$

where $\mathcal{L}_i[1]$ denotes the first class in the sorted list $\mathcal{L}_i$.

$$\text{one\_error} = \frac{\sum_{i=1}^{M} \text{one\_error}_i}{M}. \tag{2.19}$$

- *Normalized coverage* (↓) describes the number of labels on average that should have been included in $\mathcal{L}_i$ to cover all the ground-truth labels of an instance i as in [84], [85] based on label-wise effectiveness.

$$\text{coverage} = \frac{\sum_{i=1}^{M} \max_j \{j | \mathcal{I}(\mathbf{y}_i) \in_j \mathcal{L}_i\} - 1}{M * (C - 1)}, \tag{2.20}$$

where $\{j | \mathcal{I}(\mathbf{y}_i) \in_j \mathcal{L}_i\}$ is the positions of relevant classes $\mathcal{I}(\mathbf{y}_i)$ in the ordered list $\mathcal{L}$.

- *Macro-AUC* (↑) is the average area under ROC curves (AUC) for different classes as in [84], [85] based on instance-wise effective classifiers. The ROC curve uses a true positive rate and false positive rate, which may be unreliable in the cases where rare classes are present [86].

- ***Micro-AUC*** (↑) is the area under ROC curves (AUC) averaged over the full predicted label matrix $\hat{\mathbf{Y}}$ defined in [84], [85] based on label-wise effectiveness. The Micro-AUC is calculated based on the aggregation of the predicted labels over all classes instead of for each class as in Macro-AUC.

- ***Macro-F1*** (↑) shows the average F1 value on each class c as in [84], [85] based on instance-wise effectiveness.

$$\text{macroF1} = \frac{2}{C}\sum_{c=1}^{C}\frac{\text{precision}_c * \text{recall}_c}{\text{precision}_c + \text{recall}_c}, \tag{2.21}$$

  where precision_c presents precision for class c defined as $TP_c/(TP_c + FP_c)$. recall_c presents recall for class c defined as $TP_c/(TP_c + FN_c)$.

- ***Micro-F1*** (↑) indicates the aggregation F1 score which is calculated as an average over the $\hat{\mathbf{Y}}$ with thresholds as defined in [84], [85].

$$\text{microF1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}, \tag{2.22}$$

  where $\text{precision} = TP/(TP+FP)$ and $\text{recall} = TP/(TP+FN)$ and TP, FP, and FN are the number of true positives, false positives, and false negatives predictions in the predicted label matrix $\hat{\mathbf{Y}}$. As shown in Eq. (2.22), the Micro-F1 is calculated with the aggregation of precision and recall from all classes, which is different from the Macro-F1 from each class.

- ***Friedman test*** depicts a rank-based non-parametric test to compare the performance of more classifiers over multiple datasets [83]. We verified whether the differences between our proposed SMLDAc methods and the competing methods are overall significant or not in [P2].

- ***Wilcoxon Signed-Rank test*** depicts a rank-based non-parametric test as an alternative to the paired t-test to rank the differences in performances of two classifiers for each dataset [83]. In [P2], we adopted the Wilcoxon Signed-Rank test to make a comparison between our proposed variants and all other competing methods.

According to [38], [49], [87], we adopted five evaluation metrics for semantic binary segmentation in [P4]: average precision (AP), F-measure based on Optimal

Dataset Scale (ODS), F-measure based on Optimal Image Scale (OIS), global accuracy (GA), mean Intersection-Over-Union (mean IOU).

The evaluation metrics used in [P4] are described as follows:

- *Average Precision* (↑) indicates the area under the precision-recall curve [88].

$$\mathrm{AP} = \sum_{n} (R_n - R_{n-1}) P_n, \tag{2.23}$$

  where P_n and R_n are the precision and recall at the n^{th} threshold.

- *ODS F1* (↑) indicates the optimal F1 score [38] with a fixed threshold for all images in a dataset.
- *OIS F1* (↑) indicates the F1 score is calculated as an aggregation based on each image [38] with the optimal threshold in a dataset.
- *Global Accuracy* (↑) indicates the ratio of correctly predicted pixels from cracks and backgrounds to the total number of pixels of all images.

$$\mathrm{GA} = \frac{\sum_n (TN_n + TP_n)}{\sum_n (TN_n + TP_n + FN_n + FP_n)}, \tag{2.24}$$

  where n is the number of test images.

- *Mean IOU* (↑) presents the mean ratio of true positive and the union of predicted crack pixels and ground truth crack pixels

$$\mathrm{Mean} - \mathrm{IOU} = \frac{1}{C} \sum_{i=1}^{C} \sum_{j=1}^{n} \frac{TP_{ij}}{TP_{ij} + FP_{ij} + FN_{ij}}, \tag{2.25}$$

  where C is the number of class, C equals to 2 in [P4].

3 LINEAR DISCRIMINANT ANALYSIS ALGORITHMS FOR CLASSIFICATION TASKS IN IMBALANCED DATASETS

This chapter summarizes the LDA-related dimensionality reduction methods proposed in [P1] and [P2]. Traditional LDA was originally proposed to address binary-label classification tasks on datasets under the Gaussian class distribution assumption. However, such an assumption does not hold in either imbalanced datasets with non-Gaussian distributions or multi-label datasets. To extend traditional LDA-related methods for non-Gaussian distribution datasets, we introduced a probabilistic class saliency estimation approach [45] as weight factors into LDA scatter matrices. Such an introduction can mitigate the influence of uneven distribution on the definition of scatter matrices for single-label dataset classification [P1]. Furthermore, we extended this probabilistic class saliency estimation approach with six kinds of prior information to multi-label classification tasks under a weighted linear discriminant analysis framework [5].

This chapter is structured as follows. In section 3.1, the probabilistic saliency estimation approach is generally described. The proposed method and experiments for single-label dataset classification tasks are presented in section 3.2. Section 3.3 presents the proposed methods and summarizes the experiments for solving multi-label imbalanced dataset classification tasks.

3.1 Probabilistic Saliency Estimation

The concept of saliency estimation in computer vision area utilizes the special perception of the human visual system in physiological science [89], [90]. The human visual system can distinguish the perceived scenes as prominent parts and non-prominent parts according to details e.g. colors and textures [91]. Based on the acknowledgment of saliency from the human visual system, the saliency information can be estimated using probabilistic models with the exploitation of prior probability distribution from images or videos [45], [92].

A novel probabilistic saliency estimation approach was proposed by Aytekin et al. [45] for image segmentation. In [45], a distinct region $\mathbf{x}_i$ (pixel, super-pixel, or patch of pixels) in an image can be presented by a probability mass function $P(\mathbf{x}_i)$. The higher values of $P(\mathbf{x}_i)$ for the region $\mathbf{x}_i$ indicates for prominent the region is. $P(\mathbf{x}_i)$ is obtained by optimizing the following objective function with two terms

$$\begin{aligned}
&\underset{P(x)}{\operatorname{argmin}}\left(\sum_i P(\mathbf{x}_i)^2 v_i + \sum_{i,j}\left(P(\mathbf{x}_i) - P(\mathbf{x}_j)\right)^2 a_{ij}\right) = \\
&\underset{P(x)}{\operatorname{argmin}}\left(\sum_i P(\mathbf{x}_i)^2 v_i + \sum_{i,j}\left(P(\mathbf{x}_i)^2 - P(\mathbf{x}_i)P(\mathbf{x}_j)\right)a_{ij}\right) \\
&\text{s.t.} \quad \sum_i P(\mathbf{x}_i) = 1,
\end{aligned} \tag{3.1}$$

where $v_i \geq 0$ in the first term depicts the prior information of region $\mathbf{x}_i$, which helps to suppress the influence of non-prominent regions with a higher value. a_{ij} describes the similarity between region $\mathbf{x}_i$ and $\mathbf{x}_j$ which forces the regions to have similar probabilities under a higher similarity value a_{ij}. We assume that the similarity values are symmetric, i.e., $a_{ij} = a_{ji}$. Given this objective function, the contributions of non-salient regions can be suppressed with lower probabilities and similar regions can be found out with similar probabilities.

The vanilla objective function of the probabilistic saliency estimation approach can be expressed in a matrix notation as

$$\begin{aligned}
&\mathbf{p}^* = \underset{\mathbf{p}}{\operatorname{argmin}} \ (\mathbf{p}^{\mathbf{T}}\mathbf{H}\mathbf{p}), \\
&\mathbf{H} = \mathbf{D} - \mathbf{A} + \mathbf{V}, \\
&\text{s.t.} \quad \mathbf{p}^T\mathbf{1} = 1,
\end{aligned} \tag{3.2}$$

where $\mathbf{p}$ is the probability vector to indicate whether region $\mathbf{x}_i$ to be salient or not, i.e., $p_i = P(\mathbf{x}_i)$. An affinity matrix $\mathbf{A}$ presents the similarity of each pair of regions $\mathbf{x}_i$ and $\mathbf{x}_j$ as $[\mathbf{A}]_{ij} = a_{ij}$. $\mathbf{D}$ is defined as a diagonal matrix with elements equal to $[\mathbf{D}_{ii}] = \sum_j a_{ij}$. The diagonal prior information matrix $\mathbf{V}$ has elements $[\mathbf{V}]_{ii} = v_i$, and $\mathbf{1}$ is a vector of ones. A final global optimum $\mathbf{p}^*$ is obtained by the Lagrangian multiplier method based on the following equation:

$$\mathcal{L}(\mathbf{p}, \gamma) = (\mathbf{p}^{\mathbf{T}}\mathbf{H}\mathbf{p}) - \gamma(\mathbf{p}^{\mathbf{T}}\mathbf{1} - 1), \tag{3.3}$$

where the partial derivative of this equation with respect $\mathbf{p}$ is zero.

The final optimized probability vector is depicted as

$$\mathbf{p}^*_{pse} = \frac{1}{\mathbf{1}^T\mathbf{H}^{-1}\mathbf{1}}\mathbf{H}^{-1}\mathbf{1}, \tag{3.4}$$

where $\mathbf{1}^T\mathbf{H}^{-1}\mathbf{1}$ satisfies the constraint $\mathbf{p}^T\mathbf{1} = 1$ and ensures that the result is a normalized probability vector. Moreover, the $\mathbf{p}^*$ always contains non-negative values as shown in [45].

3.2 Single-label Datasets Classification

This section presents the weighted LDA with class saliency information for single-label classification tasks which was proposed in [P1]. Inspired by [45], we first exploited the probabilistic saliency estimation model [45] to estimate the saliency of each instance to its associated class from six image datasets. Then an LDA variant method based on the probabilistic saliency estimation model was proposed to conquer the drawback of traditional LDA and its variants on imbalanced classes for single-label classification.

3.2.1 Proposed Method

LDA technique defines an optimal projection by means of Fisher criterion optimization from raw data to reduce the dimensions and extract discriminative features. Traditional LDA assumes Gaussian distribution for its input. Hence, when there is a large overlap of neighboring classes or there is a dominant outlier class [93], the definitions of scatter matrices could lead to sub-optimal results. To tackle such drawbacks of traditional LDA, weighting factors are introduced into the definitions of the within-class and between-class scatters in [23], [25], [26], [94]. In [P1], we calculated weighting factors using the proposed probabilistic method [45] to re-define the contribution of each sample to its associated class based on its *class saliency information*.

According to Section 3.1, we followed a similar path to define an affinity matrix $\mathbf{A}^c$, a priori saliency information matrix $\mathbf{V}_c$ and a diagonal matrix $\mathbf{D}_c$ for each class c. Firstly, the affinity matrix $\mathbf{A}^c \in \mathbb{R}^{N_c \times N_c}$ is defined using fully connected and k-NN graphs with RBF kernel function

$$[\mathbf{A}^c]_{ij} = \exp\left(-\frac{\|\mathbf{x}_i^c - \mathbf{x}_j^c\|^2}{2\sigma^2}\right), \tag{3.5}$$

where we used $\mathbf{x}_i^c$ and $\mathbf{x}_j^c$ to represent the i^{th} and j^{th} instances in class c, respectively. σ is the hyper-parameter and its value is set equal to the mean Euclidean distance between the training samples. Then a graph $\mathcal{G}_C = \{\mathbf{X}_c, \mathbf{A}^c\}$ is formed to represent the similarity of each instance in class c with the affinity matrix $\mathbf{A}^c \in \mathbb{R}^{N_c \times N_c}$, where $\mathbf{X}_c \in \mathbb{R}^{D \times N_c}$ is a matrix consisting of instances of class c. After obtaining $\mathbf{A}^c$, the diagonal matrix $\mathbf{D}^c$ is written as $[\mathbf{D}^c]_{ii} = \sum_j [\mathbf{A}^c]_{ij}$.

In [P1], the priori information called misclassification-based probability is used to calculate matrix $\mathbf{V}_c$ under the assumption that if a sample is closer to another class, it is less probable to be prominent to its associated class. The formula of the misclassification-based probability is defined as

$$[\mathbf{V}^c]_{ii} = \begin{cases} 0, & \text{if } d_{ic}^c < \min\limits_{k \neq c} d_{ic}^k, \\ \frac{d_{ic}^c}{\min\limits_{k \neq c} d_{ic}^k}, & \text{otherwise,} \end{cases} \tag{3.6}$$

where $\mathbf{x}_{ic}$ is the i^{th} instance of class c, and $\boldsymbol{\mu}_k$ is the mean vector of class k. d_{ic}^k presents the Euclidean distance between $\mathbf{x}_{ic}$ and $\boldsymbol{\mu}_k$ as $d_{ic}^k = \|\mathbf{x}_{ic} - \boldsymbol{\mu}_k\|_2^2$.

The probability of each sample $\mathbf{x}_i^c$ to its associated class c is given by: $\mathbf{p}^c = \mathbf{H^{c}}^{-1}\mathbf{1}$, where $\mathbf{H}^c = \mathbf{D}^c - \mathbf{A}^c + \mathbf{V}^c$. Once obtained $\mathbf{p}^c \in \mathbb{R}^{N_c}$, $c = 1, \ldots, C$, we define a new class representation as $\mathbf{m}^c = \mathbf{X}^c\mathbf{p}^c$.

Then we re-define the within-class scatter matrix in two different ways for a comparison. The first one is to incorporate $\mathbf{p}^c$ in $\mathbf{S}_w$ as:

$$\mathbf{S}_w^{(1)} = \sum_{c=1}^{C}\sum_{j=1}^{N_c} p_{c,j}(\mathbf{x}_i^c - \boldsymbol{\mu}_c)(\mathbf{x}_j^c - \boldsymbol{\mu}_c)^T, \tag{3.7}$$

where $\mathbf{x}_j^c$ is j-th sample in class c, $p_{c,j}$ is saliency score for the j-th sample in class c. The other within-class scatter matrix is defined as the relevance-weighted LDA in

[23] as:

$$\mathbf{S}_w^{(2)} = \sum_{c=1}^{C}\sum_{j=1}^{N_c} p_{c,j} r_c (\mathbf{x}_j^c - \boldsymbol{\mu}_c)(\mathbf{x}_j^c - \boldsymbol{\mu}_c)^T. \tag{3.8}$$

Here $r_c = \sum_{i\neq c}\frac{1}{L_{ic}}$ is called relevance-weight, where L_{ic} is the reverse of the Euclidean distance between pairwise mean vectors of class i and class c.

The four types of between-class scatter matrices in [P1] are depicted sequentially. The first type follows the definition of traditional LDA as

$$\mathbf{S}_b^{(1)} = \sum_{c=1}^{C} N_c(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T. \tag{3.9}$$

The second type used saliency scores $\mathbf{p}_c$ for new class representations, as follows:

$$\hat{\boldsymbol{\mu}}_c = \mathbf{X}^c\mathbf{p}^c, \tag{3.10}$$

$$\mathbf{S}_b^{(2)} = \sum_{c=1}^{C}(\hat{\boldsymbol{\mu}}_c - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}}_c - \boldsymbol{\mu})^T, \tag{3.11}$$

where matrix $\mathbf{X}_c$ contains all samples in class c, $\hat{\boldsymbol{\mu}}_c$ is the new class representation or weighted center of class c.

The third definition further exploits the relationships between pairs of new class representations for each class, as follows:

$$\mathbf{S}_b^{(3)} = \sum_{c_1=1}^{C}\sum_{c_2=1}^{C}(\hat{\boldsymbol{\mu}}_{c_1} - \hat{\boldsymbol{\mu}}_{c_2})(\hat{\boldsymbol{\mu}}_{c_1} - \hat{\boldsymbol{\mu}}_{c_2})^T. \tag{3.12}$$

The last definition maximizes the discrimination between every sample in one class with the other class representations, while considering each sample's saliency scores, as follows:

$$\mathbf{S}_b^{(4)} = \sum_{c_1=1}^{C}\sum_{\substack{c_2=1,\\ c_2\neq c_1}}^{C}\sum_{j=1}^{N_{c_1}} p_{c_1,j}(\mathbf{x}_i^{c_1} - \hat{\boldsymbol{\mu}}_{c_2})(\mathbf{x}_j^{c_1} - \hat{\boldsymbol{\mu}}_{c_2})^T, \tag{3.13}$$

where N_{c_1} is the cardinality of class c_1.

Table 3.1 Datasets used for the experiment [P1]

Database	Contents	Numbers #	Subjects #	Classes #
BU [95]	facial expression images	2500	100	6
KANADE [96]	facial expressions	500	100	6
JAFFE [97]	facial expressions	213	10	7
ORL [98]	facial expressions and details	400	40	6
YALE [99]	facial expressions and details	165	15	11
AR [100]	facial expressions and details	4000	126	13

The final optimization criteria can be formed as follows:

$$J(\mathbf{W}) = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{tr(\mathbf{W}^T \mathbf{S}_b^{(i)} \mathbf{W})}{tr(\mathbf{W}^T \mathbf{S}_t^{(ij)} \mathbf{W})}, \tag{3.14}$$

where $\mathbf{S}_t^{(ij)} = \mathbf{S}_w^{(j)} + \mathbf{S}_b^{(i)}$, $i \in \{1,2,3,4\}$ and $j \in \{1,2\}$. Then the projection matrix $\mathbf{W}$ was obtained by eigenvalue decomposition as in Eq. (2.7). The nearest centroid classifier is applied for classification after dimensionality reduction on test samples using the matrix $\mathbf{W}$. We added a small constant $\epsilon = 0.01$ to $\mathbf{S}_w$ in the diagonal direction, as to avoid singularity problem as

$$\mathbf{S}_b \mathbf{w} = (\mathbf{S}_w + \epsilon \mathbf{I}) \lambda \mathbf{w}, \tag{3.15}$$

3.2.2 Experimental Results

Six public facial image datasets are used in [P1] as shown in Table 3.1. Each image from these datasets is resized to 40×30 pixels (gray-scale images) and vectorized to obtain facial vectors $\mathbf{x}_i \in \mathbb{R}^{1200}$. We normalized each dataset to have zero mean and unit standard derivation and split each dataset into 5 folds for cross-validation.

We evaluated the experimental results using an accuracy metric. The results of traditional LDA, Tang et al. [23], Jarchi and Boostani's work [25] are considered as the competing methods. The comparison results are presented in Table 3.2 and Table 3.3. $SwLDA_{ij}$ was obtained by using the matrices $\mathbf{S}_t^{(ij)}$ and $\mathbf{S}_b^{(i)}$, $i \in \{1,2,3,4\}$, $j \in \{1,2\}$.

As shown in Table 3.2, $SwLDA_{42}$ is the most effective method over datasets

BU and KANADE with fully connected graphs. $SwLDA_{41}$ with fully connected graphs achieves the best performance over dataset JAFFE. Our methods can achieve the best performance over ORL, YALE, and AR compared to competing methods shown in Table 3.3.

Table 3.2 Classification accuracy of proposed $SwLDA$ [P1]

Dataset	BU		KANADE		JAFFE		ORL		YALE		AR	
K	1	$\min(5, 0.1 * N_c)$	1	$\min(5, 0.1 * N_c)$	1	$\min(5, 0.1 * N_c)$	1	$\min(5, 0.1 * N_c)$	1	$\min(5, 0.1 * N_c)$	1	$\min(5, 0.1 * N_c)$
$SwLDA_{11}$	0.5714	0.5714	0.6816	0.6939	0.5619	0.5762	0.9700	0.9700	**0.9597**	0.9564	**0.9696**	**0.9696**
$SwLDA_{21}$	0.5714	0.5686	0.6816	0.6816	0.5619	0.5762	0.9700	0.9700	**0.9597**	0.9568	**0.9696**	**0.9696**
$SwLDA_{31}$	0.5886	0.5829	0.6776	0.6776	0.5524	0.5762	**0.9850**	**0.9850**	**0.9597**	0.9556	**0.9696**	0.9692
$SwLDA_{41}$	0.6500	0.6529	0.7020	0.6980	**0.5905**	0.5857	**0.9850**	**0.9850**	**0.9597**	0.9568	**0.9696**	**0.9696**
$SwLDA_{12}$	0.5800	0.5814	0.6816	0.6816	0.5667	0.5667	**0.9850**	**0.9850**	0.9589	0.9564	0.9692	0.9688
$SwLDA_{22}$	0.5800	0.5814	0.6816	0.6816	0.5667	0.5571	**0.9850**	**0.9850**	0.9589	0.9572	0.9684	0.9684
$SwLDA_{32}$	0.6243	0.6200	0.6776	0.6776	0.5286	0.5238	0.9600	0.9600	0.9589	0.9572	0.9684	0.9684
$SwLDA_{42}$	**0.6786**	0.6743	**0.7224**	0.7184	0.5476	0.5524	0.9450	0.9450	0.9593	0.9572	**0.9696**	0.9692

3.3 Multi-label Datasets Classification

Each data item can belong to either one or several classes in multi-label datasets. For example, a scenic image can contain several classes, such as a beach, people, sunlight, a boat, and so on. Another characteristic of multi-label datasets is imbalance [13]. Hence, it is imperative to specifically analyze multi-label datasets when applying machine learning algorithms to avoid performance deterioration [1].

Dimensionality reduction is a typical technique for tackling multi-label classification problems [10] as a pre-processing step. After the introduction of weighting factors, traditional LDA can be extended to tackle multi-label problems as in [28]. In work [P2], we have introduced weighting factors derived from probabilistic saliency estimation [45] into a multi-label LDA framework. The proposed method is called Saliency-based Multi-label Linear Discriminant Analysis (SMLDA).

Table 3.3 Classification accuracy comparison with competing methods [P1]

Dataset	BU	KANADE	JAFFE	ORL	YALE	AR
LDA	0.5729	0.6898	0.5571	0.9725	0.9593	0.9688
[23]	0.5743	0.6857	0.5714	0.9800	0.9564	0.9681
[25]	0.5957	0.6898	0.5381	0.9800	**0.9597**	0.9692
$SwLDA_{41}$	0.6500	0.7020	**0.5905**	**0.9850**	**0.9597**	**0.9696**
$SwLDA_{42}$	**0.6786**	**0.7224**	0.5476	0.9450	0.9593	**0.9696**

3.3.1 Proposed Method

Our proposed method aims to obtain a final lower-dimension subspace $\mathbb{R}^d$ retaining the most distinguishable features from the original higher-dimension space $\mathbb{R}^D$ with a data projection matrix $\mathbf{W} \in \mathbb{R}^{D\times d}$. Moreover, our SMLDA approach exploited various prior weighting factors to balance the contribution of each sample to its associated classes to mitigate problems related to imbalanced classes since we calculated weighting factors within each class. Two main steps are carried out in the proposed method for the target. Firstly, a probability matrix indicating the importance of each sample for its associated classes was calculated based on the probabilistic saliency estimation approach [5], [45]. Then, we embedded the probability matrix as weighting factors in a multi-label LDA framework from work [5] for scatter matrices calculation.

For the first step, a probability matrix $\mathbf{P} \in \mathbb{R}^{C\times N}$ is defined as

$$\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_i, ..., \mathbf{p}_N] = [\mathbf{p}_{(1)}, ..., \mathbf{p}_{(j)}, ..., \mathbf{p}_{(C)}]^T, \tag{3.16}$$

where $\mathbf{p}_i \in \mathbb{R}^C$ is a vector with probabilities indicating whether instance i to be salient for class c or not. $\mathbf{p}_{(j)} \in \mathbb{R}^N$ is the probability vector for the j^{th} class, which are normalized to sum up to one, i.e. $\sum_{i=1}^{N} p_{ci} = 1 \quad \forall c \in 1, \ldots C$. The probability matrix $\mathbf{P}$ is calculated by the probabilistic multi-label class-saliency estimation approach as mentioned in Section 3.1 based on each class.

Assume that $\mathbf{X} \in \mathbb{R}^{D\times N}$ and $\mathbf{Y} \in \mathbb{R}^{C\times N}$ are the arranged matrices separately. $\mathbf{x}_i \in \mathbb{R}^D, i \in 1, ..., N$ is an input data sample i and $\mathbf{y}_i \in \{0, 1\}^C$ is its corresponding binary label vector, where D is the number of original high dimensionality and C is the number of classes. $y_{ci} = 1$, when the sample $\mathbf{x}_i$ is associated with class c. The vector $\mathbf{y}_{(j)}$ presents the rows of $\mathbf{Y}$ contain 1s for all data samples that are associated with the particular class, where $j \in 1, \ldots, C$.

In [P2], we investigated the salient information of each sample inside each class. Hence, if a data sample i is salient to a class c, the element $p_{ci} > 0$ and $y_{ci} \neq 0$, otherwise, $p_{ci} = 0$ if $y_{ci} = 0$. For each class c, $\mathbf{X}^c \in \mathbb{R}^{D\times N^c}$ presents the data matrix and a probability vector $\mathbf{p}^c \in \mathbb{R}^{N^c}$ presents N^c data samples from class c. The final optimization function for the multi-label class saliency estimation is defined as

$$\begin{aligned}
&\underset{\mathbf{p}^c}{\operatorname{argmin}}\left(\sum_i^{N^c} (p_i^c)^2 v_i^c + \frac{1}{2}\sum_i^{N^c}\sum_j^{N^c}\left(p_i^c - p_j^c\right)^2 a_{ij}^c \right) = \\
&\underset{\mathbf{p}^c}{\operatorname{argmin}}\left(\sum_i^{N^c} (p_i^c)^2 v_i^c + \frac{1}{2}\sum_i^{N^c}\sum_j^{N^c}\left((p_i^c)^2 a_{ij}^c + (p_j^c)^2 a_{ij}^c\right) - \right. \\
&\left. \sum_i^{N^c}\sum_j^{N^c}\left(p_i^c p_j^c\right) a_{ij}^c \right) \\
&\text{s.t.} \quad \sum_i^{N^c} p_i^c = 1,
\end{aligned} \tag{3.17}$$

where p_i^c is the i^{th} element in $\mathbf{p}^c$ and $v_i^c \geq 0$ is the corresponding prior information to suppress the probabilities of non-salient instances from class c. a_{ij}^c is the similarity value forcing the instances $\mathbf{x}_i^c$ and $\mathbf{x}_j^c$ have similar probabilities if they are similar. Unlike the original probabilistic saliency estimation in Eq. (3.1), our definition of similarity values can be asymmetric.

Eq. (3.17) can be re-written in matrix notation as

$$\begin{aligned}
&\mathbf{p}^{\mathbf{c}^*} = \underset{\mathbf{p}^c}{\operatorname{argmin}} \ (\mathbf{p}^{c^T}\mathbf{H}^c\mathbf{p}^c), \\
&\mathbf{H}^c = \frac{1}{2}\mathbf{D_1}^c + \frac{1}{2}\mathbf{D_2}^c - \mathbf{A}^c + \mathbf{V}^c, \\
&\text{s.t.} \quad \mathbf{p}^{cT}\mathbf{1} = 1,
\end{aligned} \tag{3.18}$$

where $\mathbf{V}^c \in \mathbb{R}^{N^c \times N^c}$ is the prior information matrix having elements $[\mathbf{V}^c]_{ii} = v_i^c$ along diagonal. The affinity matrix of class c is $\mathbf{A}^c$ as $[\mathbf{A}^c]_{ij} = a_{ij}^c$ expressing the similarity of i^{th} and j^{th} samples in class c. $\mathbf{D_1}^c$ and $\mathbf{D_2}^c$ are the diagonal matrices which can be computed as $[\mathbf{D_1}^c]_{ii} = \sum_j [\mathbf{A}_c]_{ij}$ over rows and $[\mathbf{D_2}^c]_{ii} = \sum_j [\mathbf{A}_c]_{ji}$ over columns separately.

In [P2], we computed the affinity matrix $\mathbf{A}^c \in \mathbb{R}^{N_c \times N_c}$ for each class c with the RBF kernel function as Eq. (3.5). The affinity matrix could also be formed either by a fully connected one or sparse variants. For instance, an affinity matrix from [101] or a k-NN graph, where the sensitive parameter σ is avoided.

We explored six different kinds of prior information to set the values of $\mathbf{V}^c$, as follows.

- *Correlation-based prior information (SMLDAc)* was introduced in the original MLDA algorithm [28] to explore label information for weigh factors. A label correlation matrix $\mathbf{R}$ was calculated as Eq. (2.13) first and then the normalized weight vector $\mathbf{m'}_j \in \mathbb{R}^C$ is calculated for all data items from class c as Eq. (2.14).

  Finally, the prior information matrix values are depicted as

  $$[\mathbf{V}^c]_{ii} = 1 - m'_{cj}, \tag{3.19}$$

  The label correlation for classes k and l is high if the classes are closely related.

- *Binary-based prior information (SMLDAb)* utilizes the label information as the label matrix in [102]. With this prior information, only instances belonging to class c are considered in $\mathbf{V}^c$ with normalization for each class.

- *Entropy-based prior information (SMLDAe)* utilized the entropy-based prior information in [5], [103] to assume that data samples associated with more classes are less salient for any class. The expression is as follows:

  $$[\mathbf{V}^c]_{ii} = 1 - \frac{1}{||\mathbf{y}_i^c||_{\ell_1}}, \tag{3.20}$$

  where $\mathbf{y}_i^c$ is the class c label vector and $||\mathbf{y}_i^c||_{\ell_1}$ is the total number of classes the i^{th} item is associated with.

- *Fuzzy-based prior information (SMLDAf)* uses a fuzzy C-means clustering algorithm (SFCM) exploiting both label and features information as in [5], [104] to learn the membership degree of each item in each class. We use the membership directly as our prior information as

  $$[\mathbf{V}^c]_{ii} = 1 - g_j^c, \tag{3.21}$$

  where g_j^c is the membership degree of item j and item j is the i^{th} item associated with class c.

- *Dependence-based prior information (SMLDAd)* is from the Hilbert-Schmidt independence criterion (HSIC) [105], by which the statistical dependence between features and labels is explored. A multi-label task can be divided into several single-label tasks as described in [5]. Only the most prominent class for each item is labeled with 1 after the final iteration. Hence the prior information by HSIC is described as

$$[\mathbf{V}^c]_{ii} = 1 - h_j^c, \tag{3.22}$$

where h_j^c is 1 if j is estimated as the most prominent one to class c and zero otherwise and item j is the i^{th} item associated with class c.

- *Misclassification-based prior information (SMLDAm)* is similar to the prior information used in [P1] for single-label data as Eq. (3.6). Using this prior information type, the samples around boundaries can be emphasized to avoid the ambiguous status of outliers. Finally, we used the full data $\mathbf{X}$ to compute the prior information matrix for the whole input dataset.

The probability vector $\mathbf{p^{c*}}$ can be solved based on each class c as

$$\mathbf{p}^{c*} = \frac{1}{\mathbf{1}^T \mathbf{H}^{c^{-1}} \mathbf{1}} \mathbf{H}^{\mathrm{c}^{-1}} \mathbf{1}. \tag{3.23}$$

The full probability matrix is $\mathbf{P} \in \mathbb{R}^{C \times N}$ shown in Eq. (3.16) with the collection of each probability vector $\mathbf{p}^{c*}$. Hence, $[\mathbf{P}]_{cj} = p_i^c$ for all items in class c, where the i^{th} item in class c is the j^{th} item in the whole dataset.

For the second step, the probability matrix is directly used as weights embedding into a multi-label LDA [5]. The definitions of scatter matrices $\mathbf{S}_w$ and $\mathbf{S}_b$ are depicted as

$$\mathbf{S}_w = \mathbf{X}\Big(\mathrm{diag}(\widehat{\mathbf{p}}) - \mathbf{P}^\top \mathbf{P}\Big)\mathbf{X}^\top, \tag{3.24}$$

$$\mathbf{S}_b = \mathbf{X}\Big(\mathbf{P}^\top\, \mathbf{P} - \frac{1}{n}\widehat{\mathbf{p}}^\top \widehat{\mathbf{p}}\Big)\mathbf{X}^\top, \tag{3.25}$$

where $\widehat{\mathbf{p}} = \sum_{c=1}^{C} \mathbf{p}_{(c)}$ and $n = \sum_{c=1}^{C} \sum_{i=1}^{N} p_{ci}$. The summation of each probability vector for each class is always 1.

Using Eq. (3.15), we calculated the optimal projection matrix $\mathbf{W}$ by solving the regularized version of the generalized eigenproblem. d largest eigenvalues containing 0.999 of the information were considered to keep the corresponding eigenvectors for the projection matrix $\mathbf{W}$. The optimal subspace features can be obtained as

$$\mathbf{Z} = \mathbf{W}^T \mathbf{X}. \tag{3.26}$$

Table 3.4 Characteristics of datasets used for experiments [P2]

Database	Contents	Train #	Test #	Classes	Features	Cardinality	Min #	Max #	meanIR	meanCIR
Bibtex [106]	Text	4880	2515	159	1836	2.4	28	691	12.8	89.3
Birds [107]	Audio	179	172	19	260	1.9	4	64	6.1	16.1
Cal500 [108]	Music	300	202	174/173	68	26.1	2	263	21.1	23.1
CHD_49 [109]	Medicine	371	181	6	49	2.6	12	281	5.3	6.6
Corel16k(001) [5], [110]	Scene	5188	1744	153	500	3.1	21	1124	23.8	108.8
Emotions [111]	Music	398	195	6	72	1.9	96	181	1.5	2.4
Enron [112]	Text	988	660	57/53	1001	27	0	535	74.8	137.1
Eukaryote [3]	Biology	4658	3108	22	440	1.1	6	1387	45.1	150.5
Human [113]	Biology	1862	1244	14	440	1.2	14	623	15.4	45.2
Image [114]	Scene	1200	800	5	294	1.2	249	345	1.2	3.1
Medical [115]	Text	645	333	45/34	1449	1.2	0	170	60.9	230.2
PlantPseAAC [113]	Biology	588	390	12	440	1.1	12	172	6.7	21.8
Scene [116]	Scene	1211	1196	6	294	1.1	165	277	1.3	4.8
Stackex_coffee [13]	Text	151	74	123/63	1763	2.0	0	32	22.6	105.6
TMC2007-500 [117]	Text	21519	7077	22	500	2.2	304	12876	17.1	27.6
Yeast [118]	Biology	1500	917	14	103	4.2	21	1128	7.3	9.0
Yelp [119]	Text	6724	3281	5	671	1.8	580	4263	2.8	3.7

3.3.2 Experimental Results

3.3.2.1 Databases and data preprocessing

We performed our experiments on 17 publicly available multi-label databases [1,2]. Table 3.4 shows the details of datasets used in [P2]. We pre-process the datasets with the same techniques as in [5]. Moreover, some instances without labels or with NaN values were deleted.

3.3.2.2 Experiments and Results

Two multi-label classifiers were applied to the projected test data for the final results: multi-label k-nearest neighbor classifier (ML-kNN) [114] and multi-output linear ridge regressor (LRR) [120], [121]. The hyper-parameter k of ML-kNN was set to 15 as in [5]. A threshold (≥ 0.5) is required to generate the prediction labels based on the predicted probabilities on a test dataset. The LRR classifier has two hyperparameters: a threshold (≥ 0) for the predicted label and the regularization constant $\mu = 0.1$. The experiments of our proposed method and all other comparisons are carried out with the Matlab codes provided by [5][1]. We adopted seven

[1] http://ceai.njnu.edu.cn/Lab/LABIC/LABIC_Software.html

[2] http://www.uco.es/kdis/mllresources/#KatakisEtAl2008

evaluation metrics for performance comparison. Besides, we applied the Friedman test and Wilcoxon Signed-Rank test to evaluate the statistical significance of observed differences.

LDA-based dimensionality reduction techniques: DMLDA [32], wMLDAc [5], wMLDAb [5], wMLDAe [5], wMLDAf [5], and wMLDAd [5] were used for a comparison with the proposed SMLDA in [P2]. All the other LDA-based methods were solved using the regularized generalized eigenproblem as (3.15). Considering the limitation of contents in this dissertation, only the results using the ranking loss evaluation metric are shown in Tables 3.5 and 3.6, and the other results can be found in the work [P2]. Moreover, five non-LDA-based techniques: PCA, CCA, MLSI, MDDMp, and MVMD, along with DMLDA, wMLDAc, wMLDAd and SMLDAc are compared in Tables 3.7 and 3.8 and Tables I-XII from the supplementary material of work [P2]. Furthermore, we demonstrate the evaluation results on seven of the most imbalanced datasets with macro-F1 to verify our method can mitigate the imbalance problem. We collect from Tables XVII and XVIII the supplementary material of [P2] for the classes having meanIR over 15 and provide them in Tables 3.9 and 3.10.

Table 3.5 Comparison of different variants of the proposed method results with ML-kNN using ranking loss (↓) [P2]

	wMLDAc	SMLDAc	wMLDAb	SMLDAb	wMLDAe	SMLDAe	wMLDAf	SMLDAf	wMLDAd	SMLDAd	SMLDAm
Bibtex	0.164	**0.149**	**0.151**	0.152	0.153	**0.149**	0.151	**0.150**	0.147	**0.146**	0.147
Birds	0.217	**0.200**	0.206	**0.197**	**0.193**	0.197	0.201	**0.196**	0.232	**0.204**	0.204
CHD_49	0.212	**0.195**	**0.200**	0.206	0.198	**0.194**	**0.195**	0.200	0.226	**0.206**	0.205
Cal500	0.190	**0.187**	**0.187**	0.187	0.188	**0.187**	0.187	**0.187**	**0.186**	0.188	0.186
Corel16k(001)	0.190	**0.187**	**0.186**	0.186	**0.187**	0.187	**0.187**	0.187	0.186	**0.184**	0.182
Emotions	**0.173**	0.190	**0.162**	0.187	**0.164**	0.177	0.182	**0.177**	0.205	**0.184**	0.182
Enron	0.218	**0.142**	0.188	**0.145**	0.177	**0.142**	0.178	**0.142**	0.161	**0.139**	0.142
Eukaryote	0.122	**0.121**	0.122	**0.121**	**0.121**	0.121	**0.120**	0.121	**0.119**	0.121	0.120
Human	0.173	**0.160**	0.172	**0.162**	0.171	**0.162**	0.172	**0.159**	0.171	**0.162**	0.157
Image	0.193	**0.173**	0.199	**0.160**	0.195	**0.167**	0.199	**0.166**	0.203	**0.162**	0.172
Medical	0.071	**0.060**	0.066	**0.059**	0.065	**0.060**	0.064	**0.059**	0.071	**0.058**	0.057
PlantPseAAC	0.280	**0.228**	0.260	**0.230**	0.284	**0.225**	0.291	**0.229**	0.271	**0.234**	0.224
Scene	0.135	**0.088**	0.137	**0.087**	0.135	**0.089**	0.135	**0.088**	0.132	**0.089**	0.092
Stackex_coffee	**0.241**	0.273	**0.268**	0.272	**0.269**	0.272	**0.271**	0.272	0.284	**0.270**	0.275
TMC2007	0.027	**0.026**	0.026	**0.026**	0.026	**0.026**	0.026	**0.026**	0.028	**0.026**	0.027
Yeast	0.183	**0.178**	0.185	**0.177**	0.183	**0.178**	0.185	**0.178**	0.185	**0.177**	0.178
Yelp	0.126	**0.124**	0.130	**0.123**	0.126	**0.125**	**0.125**	0.126	0.139	**0.131**	0.139
Average	0.171	**0.158**	0.167	**0.158**	0.167	**0.156**	0.169	**0.157**	0.173	**0.158**	0.158
	Statistical analysis: Friedman: p = **4.6e-05**, Wilcoxon Signed-Ranks test wrt. SMLDAc:										
	25.0	X	**25.0**	-66.0	36.0	-53.0	**30.0**	-57.0	**10.0**	-73.0	-67.0

Table 3.6 Comparison of different variants of the proposed method results with LRR using ranking loss (↓) [P2]

	wMLDAc	SMLDAc	wMLDAb	SMLDAb	wMLDAe	SMLDAe	wMLDAf	SMLDAf	wMLDAd	SMLDAd	SMLDAm
Bibtex	0.120	**0.115**	0.120	**0.115**	0.120	**0.115**	0.119	**0.115**	0.124	**0.114**	0.112
Birds	0.332	**0.268**	0.321	**0.265**	0.321	**0.270**	0.318	**0.270**	0.321	**0.263**	0.258
CHD_49	0.209	**0.207**	0.208	**0.203**	0.208	**0.204**	0.208	**0.204**	0.213	**0.196**	0.189
Cal500	**0.242**	0.267	**0.250**	0.267	**0.266**	0.266	**0.265**	0.267	**0.194**	0.267	0.266
Corel16k(001)	**0.201**	0.202	0.206	**0.202**	0.204	**0.202**	0.204	**0.202**	**0.190**	0.199	0.197
Emotions	0.172	**0.163**	0.166	**0.161**	0.170	**0.168**	**0.168**	0.168	0.171	**0.162**	0.159
Enron	0.360	**0.198**	0.344	**0.195**	0.327	**0.196**	0.326	**0.196**	0.306	**0.195**	0.189
Eukaryote	0.129	**0.125**	0.130	**0.126**	0.129	**0.125**	0.129	**0.125**	0.128	**0.125**	0.123
Human	0.199	**0.179**	0.203	**0.181**	0.200	**0.178**	0.199	**0.178**	0.194	**0.174**	0.171
Image	0.208	**0.174**	0.203	**0.174**	0.205	**0.172**	0.203	**0.172**	0.203	**0.175**	0.188
Medical	0.044	**0.027**	0.036	**0.027**	0.035	**0.027**	0.033	**0.027**	0.044	**0.026**	0.028
PlantPseAAC	0.356	**0.339**	0.352	**0.336**	0.352	**0.339**	0.351	**0.339**	0.352	**0.338**	0.329
Scene	0.137	**0.091**	0.136	**0.092**	0.137	**0.092**	0.136	**0.091**	0.133	**0.092**	0.092
Stackex_coffee	0.199	**0.157**	**0.158**	0.160	**0.159**	0.160	0.188	**0.162**	0.188	**0.156**	0.157
TMC2007	0.040	0.040	**0.038**	0.040	**0.039**	0.040	**0.039**	0.040	0.047	**0.040**	0.042
Yeast	0.184	**0.178**	0.182	**0.178**	0.184	**0.178**	0.185	**0.178**	0.188	**0.178**	0.177
Yelp	0.137	**0.136**	0.137	**0.136**	0.137	**0.135**	0.137	**0.135**	0.148	**0.142**	0.147
Average	0.192	**0.169**	0.188	**0.168**	0.188	**0.169**	0.189	**0.169**	0.185	**0.167**	0.166
	Statistical analysis: Friedman: p = **1.8e-10**, Wilcoxon Signed-Ranks test wrt. SMLDAc:										
	14.0	X	**16.0**	-53.0	**5.0**	-69.5	**3.0**	-56.0	**23.0**	**-31.0**	-44.0

Table 3.7 Comparative results with ML-kNN using ranking loss (↓) [P2]

	Competing methods								Proposed
	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAd	SMLDAc
Bibtex	0.204	0.197	0.199	**0.116**	0.186	0.271	0.164	0.147	0.149
Birds	0.323	0.203	0.323	0.322	0.322	0.248	0.217	0.232	**0.200**
CHD_49	0.224	0.212	0.214	0.209	0.225	0.224	0.212	0.226	**0.195**
Cal500	0.183	0.187	0.184	**0.182**	0.183	0.187	0.190	0.186	0.187
Corel16k(001)	0.198	0.188	0.196	**0.185**	0.198	0.197	0.190	0.186	0.187
Emotions	0.299	0.178	0.299	0.301	0.295	0.245	**0.173**	0.205	0.190
Enron	0.133	0.170	0.135	**0.124**	0.136	0.191	0.218	0.161	0.142
Eukaryote	0.113	0.126	0.113	**0.106**	0.111	0.141	0.122	0.119	0.121
Human	0.159	0.178	0.159	**0.149**	0.158	0.191	0.173	0.171	0.160
Image	0.167	0.201	0.170	0.186	**0.166**	0.284	0.193	0.203	0.173
Medical	0.057	0.076	**0.039**	0.051	0.058	0.072	0.071	0.071	0.060
PlantPseAAC	0.197	0.277	0.198	**0.180**	0.198	0.258	0.280	0.271	0.228
Scene	0.084	0.141	0.083	0.102	**0.077**	0.234	0.135	0.132	0.088
Stackex_coffee	0.279	0.304	0.259	0.257	0.276	0.298	**0.241**	0.284	0.273
TMC2007	0.035	0.026	0.035	0.030	0.030	0.038	0.027	0.028	**0.026**
Yeast	0.174	0.184	**0.173**	0.179	0.174	0.188	0.183	0.185	0.178
Yelp	0.178	**0.117**	0.171	0.148	0.176	0.139	0.126	0.139	0.124
Average	0.177	0.174	0.174	0.166	0.175	0.200	0.171	0.173	**0.158**
	Statistical analysis: Friedman: p = **1.8e-04**, Wilcoxon Signed-Ranks test wrt. SMLDAc:								
	52.0	**16.0**	63.0	-74.0	61.0	**0.0**	**25.0**	**10.0**	X

Table 3.8 Comparative results with LRR using ranking loss (↓) [P2]

	Competing methods								Proposed
	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAd	SMLDAc
Bibtex	0.117	0.120	0.117	**0.079**	0.091	0.120	0.120	0.124	0.115
Birds	0.236	0.301	0.288	**0.171**	0.199	0.318	0.332	0.321	0.268
CHD_49	0.208	0.210	0.208	**0.196**	0.205	0.210	0.209	0.213	0.207
Cal500	0.258	0.269	0.265	0.248	0.250	0.245	0.242	**0.194**	0.267
Corel16k(001)	0.208	0.208	0.208	0.195	0.208	0.206	0.201	**0.190**	0.202
Emotions	0.163	0.167	0.163	0.174	0.177	0.166	0.172	0.171	**0.163**
Enron	0.250	0.324	0.332	**0.121**	0.138	0.400	0.360	0.306	0.198
Eukaryote	0.134	0.130	0.134	**0.111**	0.120	0.131	0.129	0.128	0.125
Human	0.211	0.209	0.210	**0.157**	0.185	0.210	0.199	0.194	0.179
Image	0.206	0.207	0.217	0.198	0.177	0.223	0.208	0.203	**0.174**
Medical	0.031	0.039	0.057	0.025	**0.024**	0.063	0.044	0.044	0.027
PlantPseAAC	0.340	0.351	0.343	**0.194**	0.315	0.362	0.356	0.352	0.339
Scene	0.136	0.136	0.138	0.097	**0.088**	0.141	0.137	0.133	0.091
Stackex_coffee	0.170	0.168	0.169	**0.157**	0.171	0.163	0.199	0.188	0.157
TMC2007	0.038	0.037	0.038	0.049	0.048	**0.037**	0.040	0.047	0.040
Yeast	0.184	0.183	0.184	0.180	0.179	0.182	0.184	0.188	**0.178**
Yelp	0.130	0.129	0.130	0.165	0.135	**0.129**	0.137	0.148	0.136
Average	0.178	0.188	0.188	**0.148**	0.159	0.195	0.192	0.185	0.169
	Statistical analysis: Friedman: p = **6.6e-05**, Wilcoxon Signed-Ranks test wrt. SMLDAc:								
	37.0	**11.0**	**16.0**	-44.0	-55.0	**20.0**	**14.0**	**23.0**	X

Table 3.9 Comparative results with ML-kNN using macro-F1 (↑) [P2]

	Competing methods								Proposed
	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAd	SMLDAc
Cal500	0.056	0.050	0.055	**0.062**	0.055	0.051	0.051	0.056	0.051
Corel16k(001)	0.013	0.030	0.017	0.036	0.018	0.018	**0.037**	0.034	0.036
Enron	0.046	0.073	0.042	**0.095**	0.054	0.012	0.039	0.036	0.062
Eukaryote	0.053	0.074	0.053	0.060	0.065	0.002	0.090	**0.092**	0.072
Human	0.043	0.146	0.041	0.095	0.071	0.001	0.145	0.133	**0.159**
Medical	0.219	0.294	**0.307**	0.280	0.226	0.186	0.259	0.263	0.302
Stackex_coffee	0.000	0.023	0.017	0.013	0.000	0.010	0.036	0.040	**0.048**
Average	0.061	0.099	0.076	0.092	0.070	0.040	0.094	0.093	**0.104**
	Statistical analysis: Friedman: p = **2.7e-03**, Wilcoxon Signed-Ranks test wrt. SMLDAc:								
	1.0	7.0	3.0	7.0	**1.0**	**0.0**	8.0	6.0	X

3.4 Summary and Discussion

In this chapter, we aim to explore saliency information to reallocate the prominence of minor instances in imbalanced datasets with the LDA-related methods for performance enhancement of the subsequent classifiers with two main contributions

Table 3.10 Comparative results with LRR using macro-F1 (↑) [P2]

	Competing methods								Proposed
	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAd	SMLDAc
Cal500	0.122	0.125	0.126	0.120	0.120	0.104	0.103	0.070	**0.127**
Corel16k(001)	0.043	0.044	0.043	0.035	0.041	**0.044**	0.042	0.037	0.044
Enron	**0.123**	0.117	0.121	0.086	0.101	0.097	0.101	0.095	0.113
Eukaryote	0.113	**0.119**	0.113	0.097	0.111	0.117	0.119	0.117	0.111
Human	0.143	0.149	0.145	0.136	0.151	0.148	0.147	0.150	**0.156**
Medical	0.531	**0.551**	0.489	0.444	0.487	0.488	0.440	0.443	0.536
Stackex_coffee	0.171	0.179	0.186	0.124	0.165	0.190	0.144	0.159	**0.196**
Average	0.178	**0.184**	0.175	0.149	0.168	0.170	0.156	0.153	0.183
	Statistical analysis: Friedman: p = **1.2e-03**, Wilcoxon Signed-Ranks test wrt. SMLDAc:								
	7.0	-14.0	7.0	0.0	**1.0**	4.0	**2.0**	**1.0**	X

Table 3.11 Summary of Wilcoxon Signed-Ranks test results:
the number of times when SMLDAc was better in a statistically significant way [P2]

	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAb	wMLDAe	wMLDAf	wMLDAd
ML-kNN	4/7	5/7	2/7	1/7	2/7	7/7	4/7	4/7	1/7	4/7	7/7
LRR	1/7	4/7	5/7	1/7	0/7	4/7	7/7	5/7	7/7	5/7	7/7
Total	5/14	9/14	7/14	2/14	2/14	11/14	11/14	9/14	8/14	9/14	14/14

published in [P1] and [P2]. In the first contribution [P1], we initially addressed the research question using the saliency-based weighted linear discriminant analysis under a probabilistic estimation approach. The proposed method explores the prior saliency information of each sample as weight factors. Our proposed method can further describe the contribution of each instance and redefine the class saliency information to balance the input dataset using the weight factors, which benefits the subsequent binary-label classification. In [P1], we adapted misclassification-based prior saliency information to highlight the outliers around boundaries.

According to the experimental results, our proposed method can enhance the classification results compared to the competing methods. For instance, the proposed method has achieved the best classification accuracy of 0.6786 and 0.7224 on the BU dataset and the KANADE dataset, which are 13.92% and 4.73% higher than the competing SOTA method [25] separately. However, the input datasets are limited to facial expressions-related topics, and the work [P1] lacks generality for imbalance problems.

To further explore the research question, we have extended the method in [P1] for multi-label classification tasks in [P2] with diverse datasets. The objective of work [P2] is to build up a general framework based on weighted linear discriminant analysis (WLDA) algorithms and the probabilistic estimation approach in [P1] to support classifiers for multi-label classification tasks. In [P2], the weight factors are calculated using six kinds of prior information to highlight the importance of prominent instances as six variants. Because our proposed framework has explored the prominence of each sample within each class, the method in the work [P2] can mitigate the influence of imbalanced problems existing widely in multi-label datasets. Furthermore, due to the generality of our framework, it can work on diverse datasets, such as audio, video, and text, with the exploration of various prior information.

We have validated our proposed framework on 17 datasets with seven evaluation metrics and two statistical tests. The experimental results have shown that our proposed framework based on the dimensionality reduction technique can not only enhance the performance of sequential multi-label classifiers but also outperform other LDA-related or no-LDA-related algorithms after comparisons. For instance, our proposed method achieves better average performance on 17 datasets with ML-kNN classifier using ranking loss metric, where the improvement rate is 7.6% compared to the competing SOTA method wMLDA [5]. Besides, the proposed method has achieved a significant improvement on the seven most imbalanced datasets with ML-kNN classifier using the Macro-F1 metric as the enhancement rate is 10.64% compared to the SOTA method wMLDA [5]. Moreover, the Friedman test and Wilcoxon signed-ranks test verify the excellent performance of our proposed method in most cases. Although the effectiveness of the proposed method for the research question has been verified, the inherent limitation still exists, which is caused by the computational complexity of the kernel matrix.

4 GENERATIVE MODELS FOR REGIONS OF INTEREST ANALYSIS

Generative models have gained more popularity currently to enrich the machine learning and artificial intelligence community. Especially, the cGANs architecture has been widely used for various topics with imbalanced inputs, such as anomaly detection, semantic binary segmentation, and facial attribute editing [1], [50], [52], due to the ability of distribution estimation. This chapter studies GANs-based methods for two topics: regions of interest and anomaly detection. The first one [P3] particularly focuses on locating and in-painting special facial attributes as the interest regions with a deep convolutional cGANs architecture. The second one [P4] presents a cGANs-related architecture to address a semantic binary segmentation task with extremely imbalanced inputs.

This chapter is structured as follows. In section 4.1, the topology of GANs is described. The proposed method and experiments targeting facial attribute detection and in-painting are depicted in section 4.2. Section 4.3 presents the proposed methods and experiments for solving anomaly detection on road surface images. Section 4.4 summarizes our contributions.

4.1 General Description of GANs

GANs was originally proposed by Goodfellow et al. [122] with the employment of the zero-sum game strategy to achieve a Nash equilibrium result. The original topology of GANs consists of a generator (G) and a discriminator (D). The real distribution p_{data} of real target data $\mathbf{x}$ can be estimated with a random data vector $\mathbf{z} \sim p_z(\mathbf{z})$. Then the generated fake data $\hat{\mathbf{x}}$ is used to deceive the judgment of the discriminator. The discriminator aims to distinguish two kinds of inputs fake data from the generator and real data.

The general objective function of GANs is formulated as

$$\min_{G} \max_{D} V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})}[log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})}[log(1 - D(G(\mathbf{z})))]. \quad (4.1)$$

Here $D(.)$ indicates the probability output of the discriminator whether the input fits the distribution of real data $p_{data}(\mathbf{x})$ or the generated fake data $p_g(\mathbf{z})$ from $G(.)$.

In [9], Mirza et al. proposed a conditional version of GANs with data $\mathbf{y}$ to condition both the generator and discriminator. The data used to condition the model could be based on class labels, some part of data, or data from other modalities. The objective function of cGANs is written as in [123]

$$\min_{G} \max_{D} V(D, G) = \mathbb{E}_{\mathbf{x},\mathbf{y}}[log(D(\mathbf{x}, \mathbf{y}))] + \mathbb{E}_{\mathbf{x},\mathbf{z}}[log(1 - D(x, G(\mathbf{x}, \mathbf{z})))]\Big]. \quad (4.2)$$

The architecture of cGANs can be based on an encoder-decoder generator or UNet-based generator [123]. Some variants of cGANs have been successful on various computer vision tasks such as [124]–[126].

Radford et al. proposed a novel topology of GANs based on deep convolutional networks in [127] called DCGANs. The generator of DCGANs consists of transposed convolutional layers and the discriminator has strived convolutional layers as shown in the following figure The architecture of DCGANs uses convolutional neural networks (CNN) for stable training across a range of datasets [127]. This modification has brought three main advantages compared to the original GANs architecture. The first advantage is the generator of DCGANs can learn its spa-

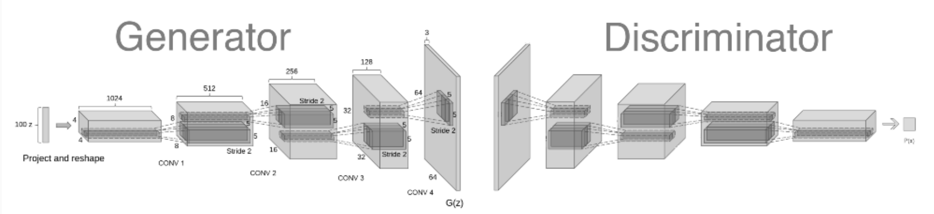


Figure 4.1 The topology of DCGANs [127]

tial up-sampling with fractional-strided convolutions. The second advantage is the deeper architectures after removing fully connected hidden layers. The third advantage is the stability of the generator from batch normalization with zero mean and unit variance.

4.2 Regions of Interest Detection and Editing on Facial Images

This section presents the second contribution [P3] which utilized GANs to detect and in-paint regions of interest on facial images unsupervised with DCGANs. The regions of interest in this work are defined as occluded facial parts covered by sunglasses and scarves. Generally, in-painting tasks require manual blocks with zero-value pixels as the occluded parts for the subsequent in-painting or constrained datasets [63]. In [P3], we proposed an unsupervised learning method to infer a mask with 0 values covering the regions of occlusions from an occluded facial image in the wild using a loss function for the subsequent in-painting.

4.2.1 Proposed Method

Our proposed method in [P3] introduces DCGANs for facial attributes detection and in-painting. Here, we aim to locate and remove facial occlusions with corresponding masks and then semantically in-paint the occluded pixels with appropriate contents. To achieve our goal, we adopted three steps in [P3]: 1) training DCGANs using occlusion-free facial images, 2) learning binary occlusion masks and corresponding facial images for completing, and 3) obtaining de-occluded facial images by merging. According to the survey in [1], facial attribute editing-related tasks can be considered as intra-class imbalance problems, due to lack of detailed data with annotated labels. Considering the advantage of balancing input with a GANs-related model, we used DCGANs to tackle the specific facial attribute editing tasks in [P3]. The workflow is demonstrated in Fig. 4.2.

- *Architecture:* We adopted the standard DCGANs architecture [127] for training, as shown in Fig. 4.2. The DCGANs contains a generator G and a discriminator D with a reverse network to G. We used occlusion-free facial images to train the DCGANs to learn a stable model representing the diverse features of normal facial images. Then the learned generator works with an optimization

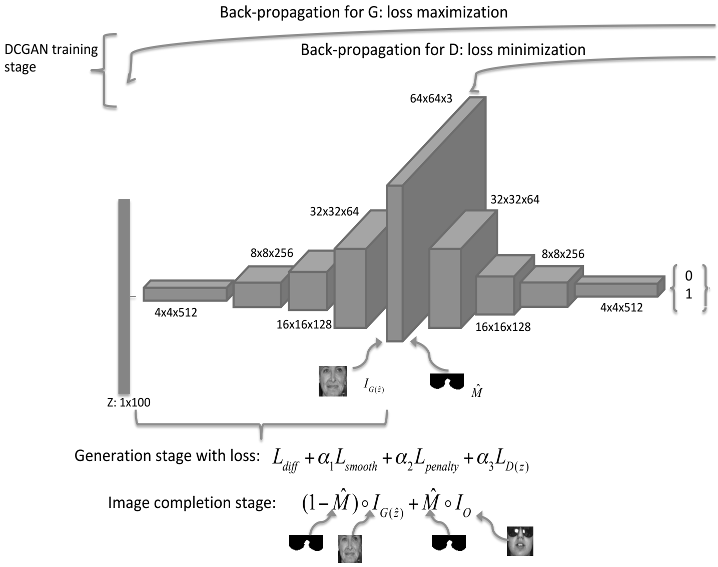

Back-propagation for G: loss maximization
DCGAN training stage
Back-propagation for D: loss minimization
64x64x3
32x32x64
32x32x64
8x8x256
8x8x256
4x4x512
16x16x128
16x16x128
4x4x512
0
1
$I_{G(\hat{z})}$
$\hat{M}$
Z: 1x100
Generation stage with loss: $L_{diff} + \alpha_1 L_{smooth} + \alpha_2 L_{penalty} + \alpha_3 L_{D(z)}$
Image completion stage: $(1-\hat{M}) \circ I_{G(\hat{z})} + \hat{M} \circ I_O$

Figure 4.2 Architecture of the proposed method [P3]

method to distinguish occlusion pixels and facial attribute pixels for a binary occlusion mask.

- *Training:* The mask M was initially set with a constant and the input vector $\mathbf{z}$ with 100 dimensions was randomly sampled based on the uniform distribution $\mathcal{U}[-1, 1]$ at the beginning. The optimization process aims to find an optimal $\hat{\mathbf{z}}$ and a binary mask M with the proposed loss function, by which the mask and in-painted occluded parts were generated at the same time. The final binary mask M contains zero-valued pixels denoting occluded areas and one-valued pixels for the occlusion-free areas. The optimal $\hat{\mathbf{z}}$ can generate an image with the corresponding facial attributes to the occluded inputs.

Two optimization algorithms: Adam gradient descent [128] and stochastic gradient descent (SGD) [129] were used to learn $\hat{\mathbf{z}}$ and M separately. The $\mathbf{z}$ and M were updated alternatively during each iteration. Moreover, two kinds of morphological filters were used to remove noises: a closing filter and an erosion filter for learning the normalized M in each iteration. The final binary M was obtained using a threshold T, as pixel values larger than T to 1, otherwise to 0.

- *Loss Function*

  We adopted four terms in our loss function as contextual loss function $\mathcal{L}_{diff}$, prior loss function $\mathcal{L}_{D(\mathbf{z})}$, smoothness loss function $\mathcal{L}_{smooth}$, and occlusion loss function $\mathcal{L}_{penalty}$.

  - **Contextual Loss Function** $\mathcal{L}_{diff}$ indicates the difference of the areas without occlusion in an occluded image I_O and its corresponding generated one $I_{G_{\mathbf{z}}}$, so that the generated image $I_{G(\hat{\mathbf{z}})}$ can be closer to the occluded image I_O. It is defined as

    $$\mathcal{L}_{diff} = \|W \odot (I_{G(\mathbf{z})} - I_O)\|_1, \tag{4.3}$$

    Here, $\odot$ means element-wise multiplication. W is the weighting term based on M to highlight the importance of surrounding pixels of the missing parts, as in [130]

    $$W_i = \begin{cases} \sum\limits_{j \in N(i)} \frac{(1-M_j)}{|N(i)|}, & if \;\; M_i \neq 0 \\ 0, & if \;\; M_i = 0 \end{cases}, \tag{4.4}$$

    where W_i denotes the importance weight of pixel i, $N(i)$ represents a window around pixel i, and $|N(i)|$ is the cardinality of the window.

  - **Prior Loss Function** $\mathcal{L}_{D(\mathbf{z})}$ is the loss of the trained discriminator D, acting as a penalty, defined as follows:

    $$\mathcal{L}_{D(\mathbf{z})} = log(1 - D(G(\mathbf{z}))), \tag{4.5}$$

    which leads the generated image to be as realistic as possible, satisfying the human visual experience.

  - **Smoothness Loss Function** $\mathcal{L}_{smooth}$ is designed to learn a smooth mask with the same size as the occluded image. It forces the occluded part in the mask toward a uniform value.

    $$\mathcal{L}_{smooth} = \sum_{i}^{N_1} \sum_{j}^{N_2} \sum_{k}^{-1,1} \|x_{i,j} - x_{i+k,j+k}\|_2, \tag{4.6}$$

    where $x_{i,j}$ refers to each pixel value of mask M. N_1 and N_2 are width and height, respectively. This smoothness loss function measures the similarity of each pixel with its four neighbors horizontally and vertically.

- **Occlusion Loss Function** $\mathcal{L}_{penalty}$ uses ℓ_1-norm as a penalty for large occlusion areas in the mask as:

$$\mathcal{L}_{penalty} = \sum_{i}^{N_1} \sum_{j}^{N_2} \|x_{i,j}\|_1. \tag{4.7}$$

This term is needed to avoid assigning all the pixels as occlusion. Otherwise, setting all M values to zero would be an easy way to minimize $\mathcal{L}_{diff}$.

The entire loss function is formed as:

$$\mathcal{L} = \mathcal{L}_{diff} \ + \ \alpha_1 \mathcal{L}_{smooth} \ + \ \alpha_2 \mathcal{L}_{penalty} \ + \ \alpha_3 \mathcal{L}_{D(\mathbf{z})} \tag{4.8}$$

The $\hat{\mathbf{z}}$ can be obtained by minimizing Eq. 4.8 as

$$\hat{\mathbf{z}} = arg \ \underset{\mathbf{z}}{min} \mathcal{L}. \tag{4.9}$$

The $\hat{\mathrm{M}}$ can be obtained by minimizing $\mathcal{L}$:

$$\hat{\mathrm{M}} = arg \ \underset{\mathbf{M}}{min} \mathcal{L}. \tag{4.10}$$

- *Image Completion* In the completion stage, we adopted the general strategy used in in-painting works [130], [131] as $I_{G(\hat{\mathbf{z}})}$ and I_O are merged using $\hat{M}$

$$I_{rec} = (1 - M) \ \odot I_{G(\hat{\mathbf{z}})} \ + M \odot I_O. \tag{4.11}$$

4.2.2 Experimental Results

4.2.2.1 Databases and Data Preprocessing

In work [P3], we used two public datasets CelebFaces Attributes Datasets (CelebA) [132] and AR face database [100] in our experiments. The CelebA dataset contains

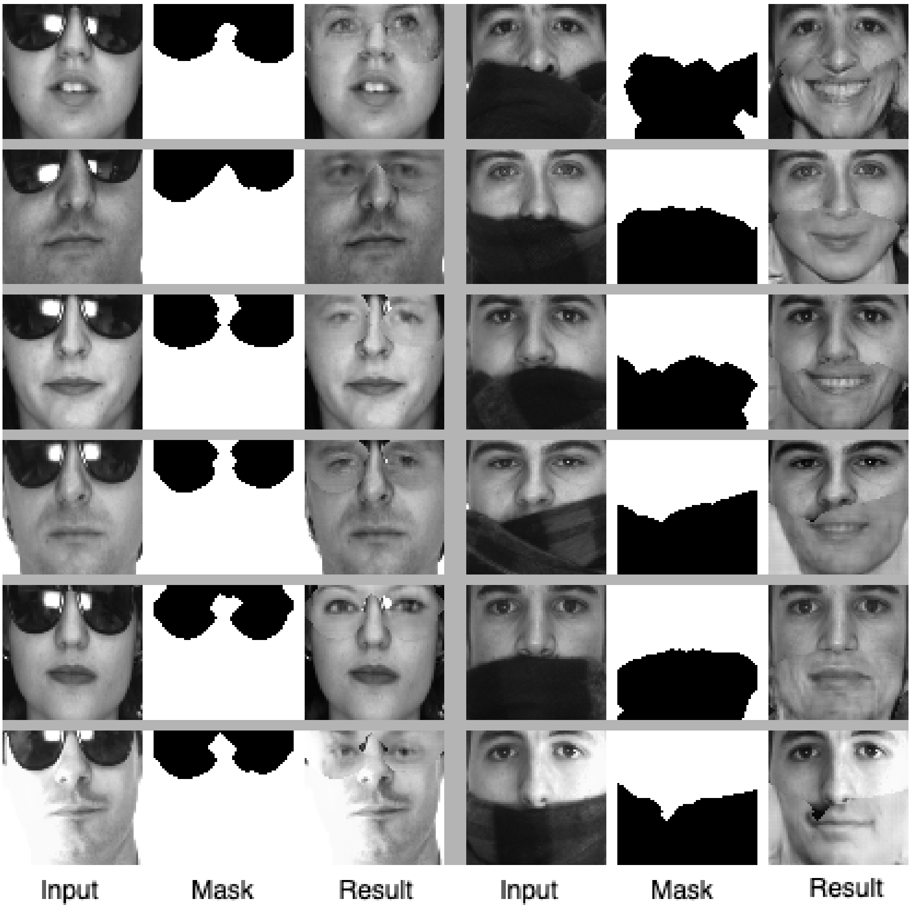


Figure 4.3 Visual results from AR dataset [P3]

202, 599 number of face images, and each image is with 40 attribute annotations. The AR face dataset consists of 4,000 face images and each face image has 13 attributes.

Each input image in [P3] was aligned and cropped using OpenFace [133] for the size of 64×64 pixels. We used the dataset CelebA [132] to train DCGANs model for the first step. Any image with occlusions (sunglasses and scarves) in CelebA has been removed before training. AR Face Dataset [100] was used as the test dataset for de-occlusion and in-painting with the trained model. Furthermore, we randomly selected several facial images with occlusions from CelebA to validate the algorithm for de-occlusion and in-painting with the trained model.

4.2.2.2 Experimental Setup

In [P3], we set the hyperparameters of the loss functions as $\alpha_1 = 1$, $\alpha_2 = 5$, and $\alpha_3 = 0.1$. Because the importance of occlusion size loss function $\mathcal{L}_{penalty}$ needs to be highlighted during training in case the values of the learned mask could be very close to being 0, we set $\alpha_2 = 5$. We have set different values for $\alpha_1 \in [1, 5]$ and $\alpha_3 \in [0, 1]$ to obtain the optimal hyper-parameters. When $\alpha_1 = 1$ and $\alpha_3 = 0.1$, the convergence process was steady. Threshold T is set to be 0.7 finally for the best performance after setting $T \in [0.5, 0.8]$ for several times experiments. The hyperparameters for both morphological filters [134] are set to be 1 pixel [135]. The hyperparameters of training the model are set as follows: 25 epochs to train the DCGANs and 1000 iterations in the generation (testing) stage for each occluded image.

4.2.2.3 Experimental Result

The proposed method in [P3] is a typical unsupervised solution without the corresponding ground truth. Hence, the experimental results are evaluated based on visual inspection instead of quantitative metrics. The experimental results are demonstrated using three figures Fig. 4.3, Fig. 4.4, and Fig. 4.5. Fig. 4.3 presents the results from the AR dataset. The process of mask generation is demonstrated in Fig. 4.4. One failure case is shown in Fig. 4.5.

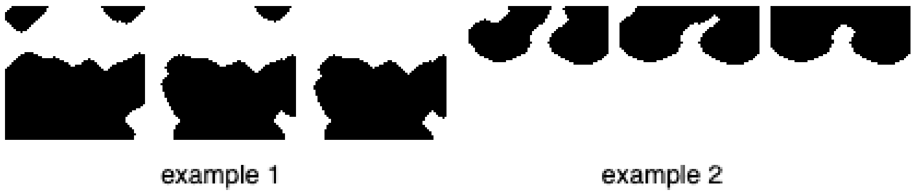


Figure 4.4 Examples of mask optimization process [P3]

4.3 Anomalous Regions of Interest Detection on Pavement Images

This section presents the fourth contribution in [P4] which utilized a cGANs-related variant to detect and segment anomalous cracks on pavement images. A key challenge of such tasks is to improve the deteriorating performance caused by the existence of

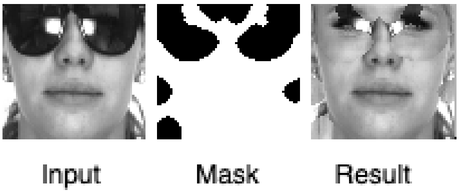


Figure 4.5 Failure examples [P3]

pixel level imbalances [1] using algorithms. GANs-related algorithms have shown good potential in solving various computer vision tasks in imbalanced datasets [1], such as shadow detection with cGANs [79], small object detection with perceptual GANs [136]. We proposed a GANs-related variant incorporating with attention mechanism and entropy-based loss functions to tackle anomalous crack detection on road surface images in [P4].

4.3.1 Proposed Method

In [P4], our proposed cGANs-related method is based on an encoder-decoder generator as the backbone, which has achieved a certain success on style transferring tasks [123]. The architecture of the generator part is shown in Fig. 4.7 [P4] and the discriminator is in Fig. 4.8 [P4]. We present the details of the work [P4] as follows

- Network Architectures

  The overall architecture of the network consists of two pipelines in the training phase as shown in Fig. 4.6. The first pipeline contains a generator and a discriminator working together based on the zero-sum game mechanism [P4]. The input of this generator is a raw pavement image $\mathbf{X}$ and its output is a generated feature probability map $\hat{\mathbf{Y}}$ with values from 0 to 1. The generator aims to deceive the judgment of the discriminator with its produced image $\hat{\mathbf{Y}}$. The inputs of the discriminator are a real image pair as $\{\mathbf{X}, \mathbf{Y}\}$ and a fake image pair $\{\mathbf{X}, \hat{\mathbf{Y}}\}$. The probability map indicates whether each pixel belongs to cracks or not. We adopted two kinds of discriminator architectures in [P4]: pixel-level discriminator in Fig. 4.8(a) or image-level discriminator Fig. 4.8(b) to investigate which one has more effect on the severely imbalanced dataset as shown in Fig. 4.8. The image-level discriminator aims to distinguish the image

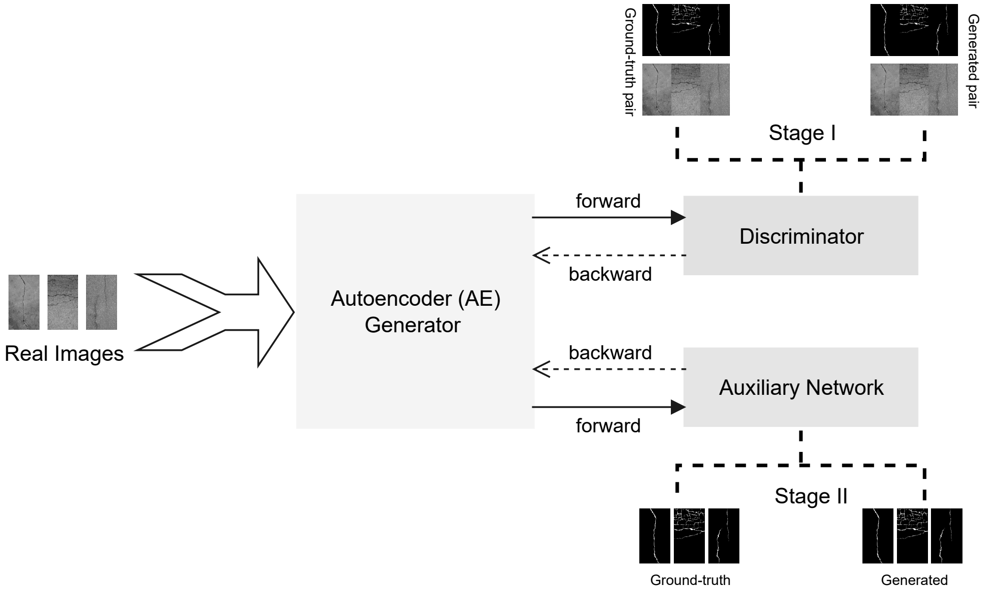


Figure 4.6 The general scheme of our proposed method [P4]

pairs with a probability. The pixel-level discriminator is to distinguish the image pairs with a probability matrix to indicate whether each pixel belongs to a real or fake.

This generator contains 22 convolutional blocks, where each contains a 2D convolutional layer, a batch normalization layer, and RELU activation sequentially. The last convolutional layer for output uses sigmoid activation. The four AGs are used to take four skipping layers in the contracting path as the gating signals. Each convolutional layer uses a 3×3 filter and 1×1 stride. The architecture of AGs is referred from [137]–[139]. The pixel-level discriminator consists of four convolutional layers, except the last layer with a sigmoid activation, others followed with RELU activation. The image-level discriminator contains five convolutional blocks, and each block has two convolutional layers followed by batch normalization, RELU activation, and max-pooling. The auxiliary network consists of four convolutional layers, except the last layer with a sigmoid activation, and others followed with a leaky RELU activation.

A significant characteristic of the pavement dataset is the imbalanced propor-

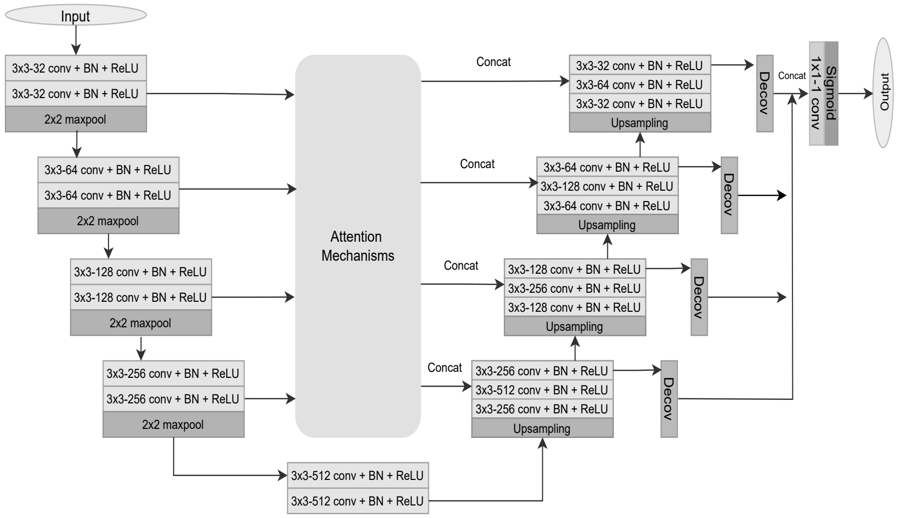


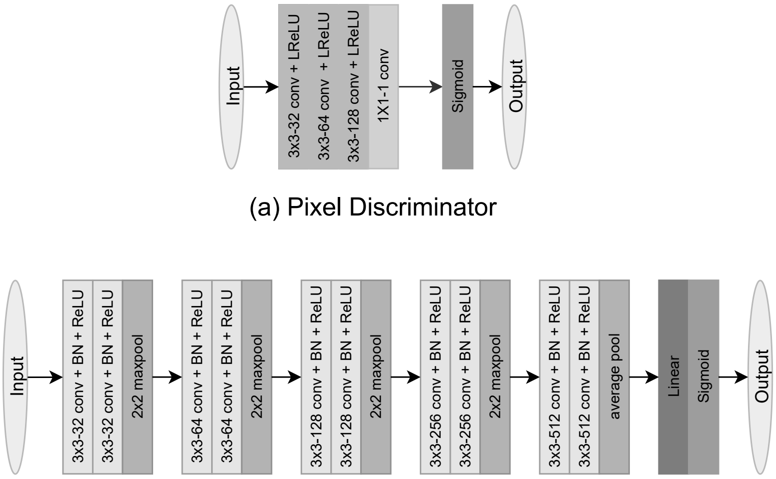
Figure 4.7 The architecture of attention UNet generator [P4].

tion of the number of pixels between cracks and backgrounds. Moreover, the texture and intensity of the regions of interest vary dramatically according to diverse capturing conditions. To address the pixel-level imbalance tasks robustly, we introduced an auxiliary network to refine the regions of interest in a pixel-by-pixel manner as the second branch to audit the generator further. The auxiliary network is a pixel-by-pixel convolutional network with three convolutional layers as shown in Figure 4.8(a). The input of the auxiliary network is the output $\hat{\mathbf{Y}}$ from the generator and its corresponding ground truth $\mathbf{Y}$.

- Loss Functions

  The conditional GAN loss function $\mathcal{L}_{cGAN}$ [123] is used for the first branch. The input of the discriminator is image pairs $\{\mathbf{X},\mathbf{Y}\}$ or $\{\mathbf{X},\hat{\mathbf{Y}}\}$, the discriminator determines whether the input pair is from the real pair $\{\mathbf{X},\mathbf{Y}\}$ or the fake pair $\{\mathbf{X},\hat{\mathbf{Y}}\}$. Hence the conditional GANs loss function is formulated as [36]

$$\mathcal{L}_{cGANs}(G,D) = \mathbb{E}_{\mathbf{X}\sim p_{data}(\mathbf{X},\mathbf{Y})}[log(D(\mathbf{X},\mathbf{Y}))]+ \mathbb{E}_{\mathbf{X}\sim p_{data}(\mathbf{X})}[log(1-D(\mathbf{X},G(\mathbf{X})))]. \tag{4.12}$$



(a) Pixel Discriminator

(b) Image Discriminator

Figure 4.8 The architectures of two discriminators [P4].

Here $\mathbf{X}$ is a raw pavement image containing cracks. $\mathbf{Y}$ depicts ground truth cracks image with 0 (normal road surface) or 1 (cracks) pixel values, and $\hat{\mathbf{Y}}$ is for the corresponding probability feature map from the generator G.

Then we utilized the definition of entropy to regularize the results from the generator in the second pipeline. Here the target of the auxiliary network is to minimize the distribution difference between the ground truth $\mathbf{Y}$ and the probability feature map $\hat{\mathbf{Y}}$ by cross-entropy and KL-divergence [140].

The auxiliary network works as a mapping function $\phi(.)$. A perceptual loss function derived from KL-divergence is defined as

$$\mathcal{L}_{KL} = \sum \phi(\mathbf{Y}) log \frac{\phi(\mathbf{Y})}{\phi(\hat{\mathbf{Y}})}. \tag{4.13}$$

The perceptual loss function is used to train the auxiliary network while the generator training is fixed. It aims to learn an optimal model to minimize the discrepancy between $\phi(\mathbf{Y})$ and $\phi(\hat{\mathbf{Y}})$. Besides, the auxiliary network is used to audit the generator with a reconstruction loss function. The reconstruction loss function can penalize the output discrepancy of the auxiliary network which is defined according to the sigmoid cross-entropy as

$$\mathcal{L}_{CE} = -log(1 - \mathbf{sigmoid}(|\phi(\mathbf{Y}) - \phi(\hat{\mathbf{Y}})|)). \tag{4.14}$$

Here $|\phi(\mathbf{Y}) - \phi(\hat{\mathbf{Y}})|$ presents the difference between the auxiliary network outputs. Furthermore, side network loss function [141] is a famous strategy for line detection initially, which is effective for crack detection [74], [88], [138]. In [P4], the side network loss function has affected the effectiveness of crack detection, especially for severely imbalanced and complex pavement datasets. The side network loss function contains losses from four side layers and the final fused layer, which are defined based on binary cross entropy as

$$\mathcal{L}_{bce} = -\frac{1}{N}\sum_{n=1}^{N} y_n log(p_n) + (1 - y_n) log(1 - p_n), \tag{4.15}$$

where p_n is the predicted pixel n in an output probability map and y_n is the corresponding ground truth pixel n. Then the loss function from the side network is denoted as

$$\mathcal{L}_{Side} = \sum_{i=1}^{4}(\mathcal{L}_{bce})^{i}_{side} + (\mathcal{L}_{bce})_{fuse}. \tag{4.16}$$

The last term of the loss function is Tversky loss function $\mathcal{L}_{TL}$ defined based on the Tversky index (TI) [42] The Tversky loss function is then defined

$$\mathcal{L}_{TL} = 1 - TI. \tag{4.17}$$

We wish the side networks, and the auxiliary networks could have an equal influence on the network parameters updating. Hence, the entire loss function consists of the above four terms with a hyperparameter γ to control the contribution of cGANs loss function as

$$\mathcal{L} = \gamma\mathcal{L}_{cGANs} + \mathcal{L}_{KL} + \mathcal{L}_{CE} + \mathcal{L}_{Side} + \mathcal{L}_{TL}. \tag{4.18}$$

4.3.2 Experimental Results

4.3.2.1 Databases and Data Preprocesing

We trained and tested the proposed framework on six datasets: CRACK500[74], CFD[38], CrackTree260[38], CrackLS315[38], CRKWH100 [38], and DeepCrack-DB[142]. To increase the number of images, we cropped the original images with overlap and retained the cropped images containing more than 1000 crack pixels on each dataset. Then the data augmentation was carried out by rotation with 90°, 180°, and 270°. Finally, we randomly selected 10% for testing and 90% for training and validation on each dataset.

4.3.2.2 Experimental Setup

The hyperparameters for the training process are set as follows: the learning rate for the Adam optimizer is 0.0001, the momentum term is 0.2 [36], and the number of iterations is 50000. Besides, we used validation to select the best model during training according to the average value of dice, accuracy, sensitivity, and specificity calculated between the predicted feature maps binarizing with the Otsu filter [143] and the corresponding ground truth every 2000 iterations.

The work [P4] has several hyperparameters for the LSA model and the loss function, which are set as follows. The window size of an LSA module is set to 8 emphasizing the correlation of a neighbor area. Moreover, the $\alpha = 0.3$ and $\beta = 0.7$ of the Tversky loss function emphasize recall (false negative pixels) more than precision. Furthermore, we set $\gamma = 0.25$ for the total loss function 4.18 to highlight the influence from the side networks and the auxiliary network more.

As mentioned above, the training process has two stages. The first stage follows a common training process of the cGANs [9]. The second stage is based on the auxiliary network to update the parameters of the generator and the auxiliary network simultaneously.

4.3.2.3 Experimental Result Analysis

We compared the performance of seven competing methods: UNet [46], HED [141], FPHB [74], V-GAN with pixel-level discriminator [36], V-GAN with image-

Table 4.1 Complexity Comparison with a 512x512 Input [P4]

Model	FLOPs	Params	Time/image (s)
UNet [46]	60.91G	5.76M	0.0090
HED [141]	0.27G	2.8K	0.0114
FPHB [74]	273.91G	44.70M	0.0457
V-GAN (pixel) [36]	160.98G	8.04M	0.0178
V-GAN (image) [36]	115.02G	12.67M	0.0180
DeepCrack [142]	160.65G	14.72M	0.0468
Crackformer II [138]	176.60G	4.96M	0.1060
cGAN_LSA (pixel)	207.15G	25.26M	0.0630
cGAN_CBAM (pixel)	179.95G	8.85M	0.0194
cGAN_CBAM_Ig (pixel)	179.95G	8.85M	0.0203
cGAN_LSA (image)	163.08G	29.88M	0.0626
cGAN_CBAM (image)	131.46G	13.47M	0.0188
cGAN_CBAM_Ig (image)	131.46G	13.47M	0.0188

Table 4.2 Result on DeepCrack-DB [P4]

Model	ODS	OIS	AP	Global Accuracy	Mean IOU
UNet [46]	0.7645	0.7706	0.8011	0.9851	0.8017
HED [141]	0.7907	0.7770	0.8429	0.9860	0.8194
FPHB [74]	0.8089	0.7595	0.8882	0.9867	0.8326
V-GAN (pixel) [36]	0.7007	0.7301	0.6063	0.9801	0.7592
V-GAN (image) [36]	0.7053	0.7063	0.5237	0.9818	0.7630
DeepCrack [142]	0.8212	0.8110	0.8928	0.9873	0.8416
Crackformer II [138]	0.8751	0.8537	**0.9195**	0.9911	0.8844
cGAN_LSA (pixel)	0.8662	0.8481	0.8360	0.9905	0.8771
cGAN_CBAM (pixel)	**0.8926**	0.8759	0.8662	**0.9924**	**0.8991**
cGAN_CBAM_Ig (pixel)	0.8921	**0.8765**	0.8833	**0.9924**	0.8986
cGAN_LSA (image)	0.8505	0.8251	0.8000	0.9894	0.8644
cGAN_CBAM (image)	0.8760	0.8556	0.8316	0.9912	0.8851
cGAN_CBAM_Ig (image)	0.8014	0.7917	0.7457	0.9864	0.8273

Table 4.3 Result on CrackLS315 [P4]

Model	ODS	OIS	AP	Global Accuracy	Mean IOU
UNet [46]	0.2452	0.2324	0.0908	0.9939	0.5650
HED [141]	0.0009	0.0008	0.0025	0.9977	0.4991
FPHB [74]	0.0119	0.0089	0.0025	0.9977	0.4989
V-GAN (pixel) [36]	0.1849	0.1449	0.0404	0.9977	0.5495
V-GAN (image) [36]	0.2411	0.2232	0.0982	0.9973	0.5669
DeepCrack [142]	0.3668	0.3534	0.2707	0.9977	0.6104
Crackformer II [138]	0.3156	0.2853	0.1971	0.9959	0.5916
cGAN_LSA (pixel)	0.4553	0.4225	0.2300	0.9976	0.6461
cGAN_CBAM (pixel)	**0.5418**	**0.5006**	**0.3520**	**0.9980**	**0.6846**
cGAN_CBAM_Ig (pixel)	0.5322	0.4889	0.3196	**0.9980**	0.6802
cGAN_LSA (image)	0.4024	0.3815	0.1934	0.9974	0.6245
cGAN_CBAM (image)	0.4366	0.4085	0.2240	0.9976	0.6382
cGAN_CBAM_Ig (image)	0.5044	0.4658	0.2944	0.9979	0.6675

Table 4.4 Alation study of loss function on CrackLS315 [P4]

Model	ODS	OIS	AP	Global Accuracy	Mean IOU
cGAN_CBAM_Ig (pixel)	0.5322	0.4889	0.3196	0.9980	0.6802
cGAN_CBAM_Ig (w/o Side loss)	0.5251	0.4823	0.3172	0.9980	0.6769
cGAN_CBAM_Ig (w/o Side loss + Tversky loss)	0.5116	0.4734	0.2941	0.9977	0.6703

Table 4.5 Alation study of loss function on DeepCrack-DB [P4]

Model	ODS	OIS	AP	Global Accuracy	Mean IOU
cGAN_CBAM_Ig (pixel)	0.8921	0.8765	0.8833	0.9924	0.8986
cGAN_CBAM_Ig (w/o Side loss)	0.8420	0.8066	0.7110	0.9881	0.8575
cGAN_CBAM_Ig (w/o Side loss + Tversky loss)	0.8161	0.7772	0.7361	0.9865	0.8377

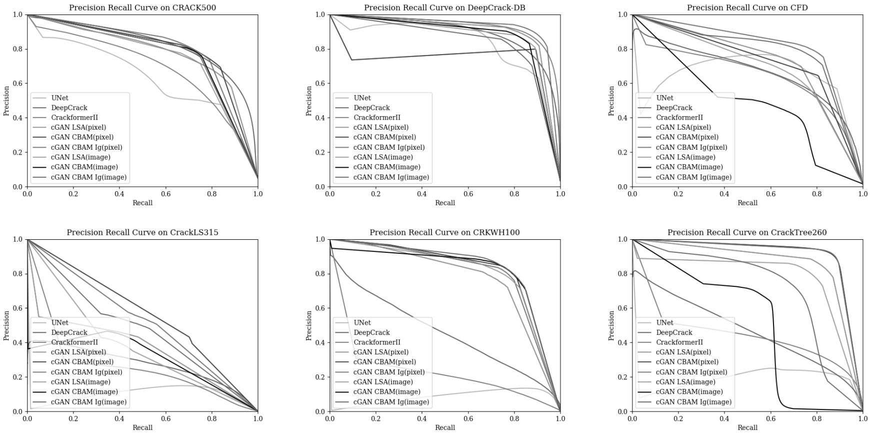


Figure 4.9 Precision and Recall Curves [P4]

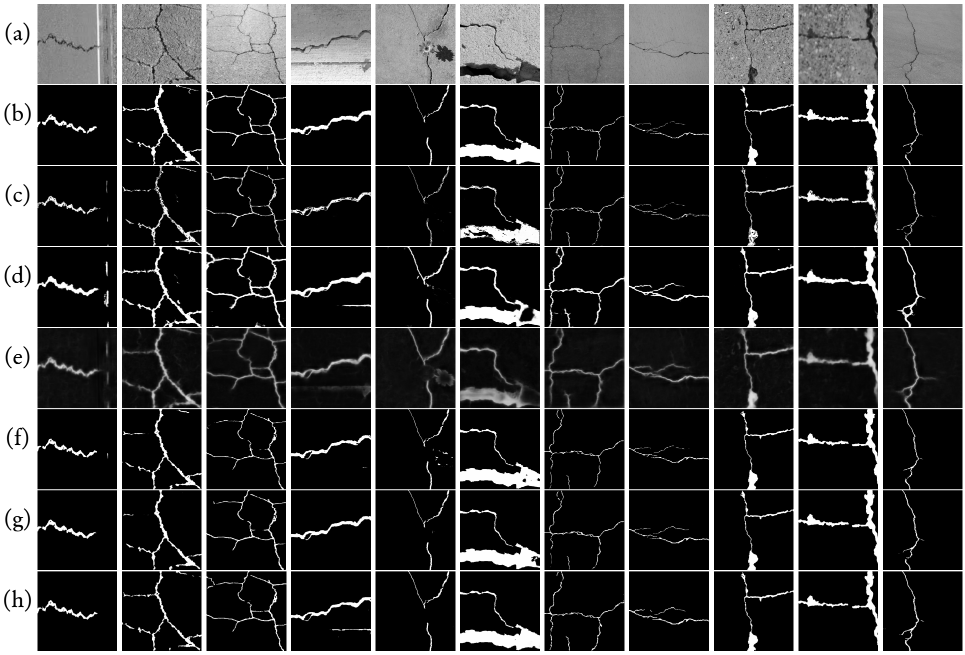


Figure 4.10 **Visual results on DeepCrack-DB** (a) Input image. (b) Ground truth. (c) UNet. (d) CrackformerII. (e) DeepCrack. (f) cGAN_CBAM (pixel). (g) cGAN_CBAM_Ig (pixel). (h)cGAN_LSA (pixel). [P4]

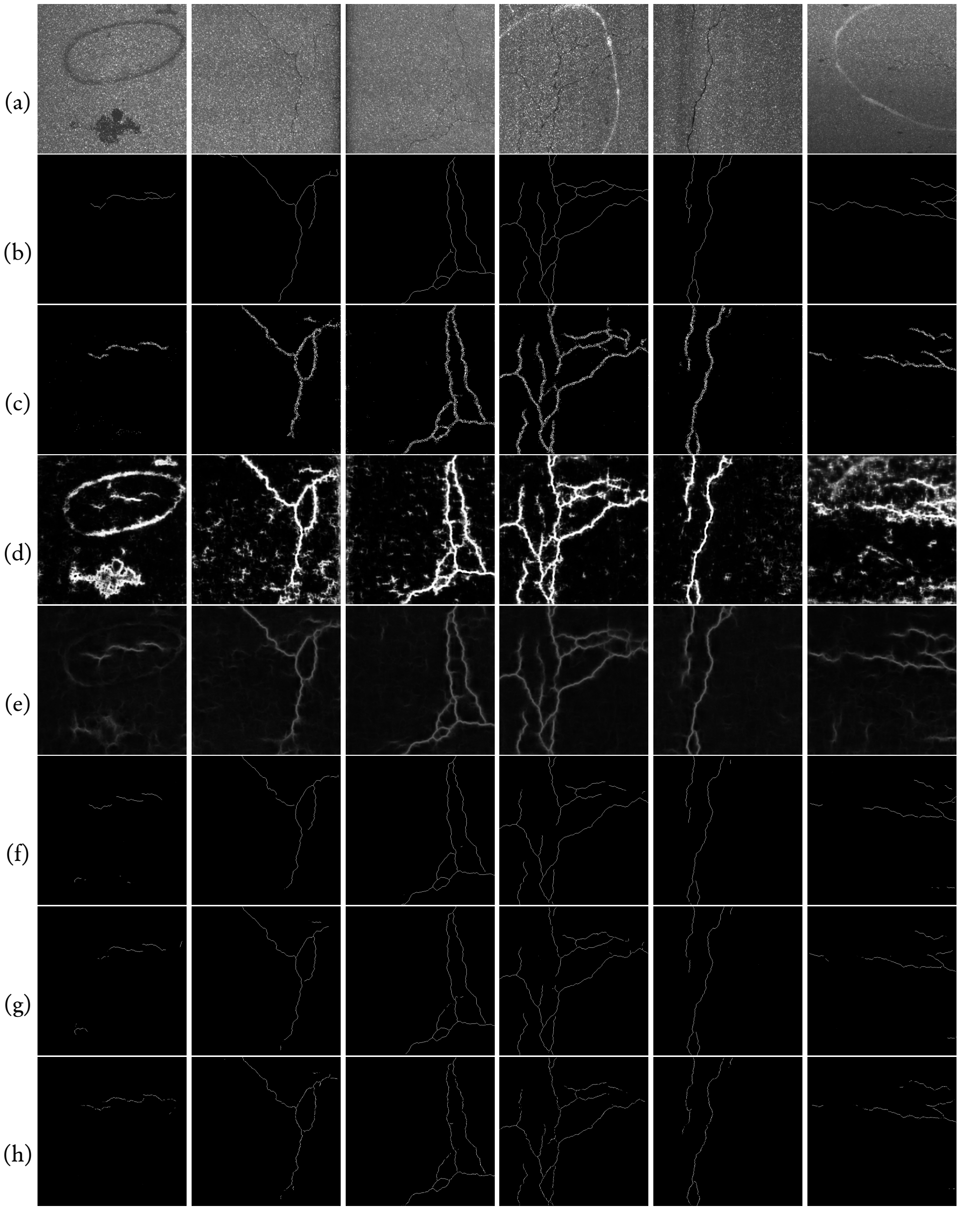


Figure 4.11 **Visual results on CrackLS315** (a) Input image. (b) Ground truth. (c) UNet. (d) CrackformerII. (e) DeepCrack. (f) cGAN_CBAM (pixel). (g) cGAN_CBAM_lg (pixel). (h)cGAN_LSA (pixel). [P4]

level discriminator [36], DeepCrack [142], and CrackformerII [138] to our proposed framework with five quantitative evaluation metrics, precision-and-recall curves, and visual results. The quantitative evaluation metrics are shown in Section 2.4.

- *Complexity Analysis:* We evaluate the model complexity, computational complexity, and inference efficiency between our proposed framework and competing methods in Table. 4.1. As shown, the computational complexity of our proposed framework with an image-level discriminator is lower than with a pixel-level discriminator, whereas the proposed framework has fewer parameters with an image-level discriminator. In addition, the computational complexity of the LSA is higher than other attention mechanisms.

- *Ablation Study* We conducted an alation study of attention mechanisms and discriminator structures for our proposed framework. We only show results on DeepCrack-DB and CrackLS315 in this dissertation as shown in Table. 4.2 and Table. 4.3. The results from others are demonstrated in [P4]. From these quantitative results, we can conclude that pixel-level discriminator works better than image-level discriminator in most cases. Moreover, the common CBAM attention mechanism achieves the best performance in most cases and the CBAM ignoring attention mechanism reaches up to similar results compared to the common CBAM. The alation study was carried out on DeepCrack-DB and CrackLS315 to verify the effectiveness of the side loss function and Tversky loss function to our proposed framework in Table. 4.4 and Table. 4.5.

- *Comparison with Competing Methods* From the precision-and-recall curves in Fig. 4.9, our proposed framework has achieved the best performance on the other four benchmark datasets: CFD, CrackLS315, CRKWH100, and Crack-Tree260 with the pixel-level discriminator, compared to the competing methods. As shown in Table. 4.2, when the backbone with the pixel-level discriminator and the common CBAM, the cGAN_CBAM (pixel) variant has achieved the best performance with ODS, Global Accuracy, and Mean IOU. Furthermore, it can be observed from Table.4.3 that our framework achieves the best performance compared to the competing methods. The other results are demonstrated in [P4].

4.4 Summary and Discussion

In this chapter, we aim to explore GANs-based techniques for two kinds of regions of interest analysis tasks with imbalanced inputs in [P3] and [P4]. In [P3], we addressed the research question of developing a GANs-based architecture with a novel optimal loss function to edit specific facial attributes with occlusions. The proposed method is based on the DCGANs trained with facial images without occlusions. During the inference stage, the trained DCGANs working with the loss function can locate and in-paint the occluded facial attribute regions in an unsupervised manner.

The experimental results under subjective visual inspection have shown that our algorithm can successfully detect and complete the occluded facial attributes on specific facial images containing occlusions as imbalanced inputs in the absence of ground truth. The final results are evaluated with subjective visual inspection to assess the performance as shown in Fig. 4.3. Besides, the generation process of two occluded masks is demonstrated in Fig. 4.4. Although this method has achieved the expected results with proper inputs, it contains some inherent limitations. First, the visual effects of the results are influenced by illumination situations and texture contrast conditions. In particular, it is difficult to evaluate the results with quantitative metrics and make a comparison with other methods due to lacking ground truth.

The fourth contribution [P4] presents another solution to research question 2 by exploring a GANs-based architecture with entropy-based loss functions to segment the anomalous crack pixels on pavement images. The proposed method adopts a cGANs-based architecture as a backbone with severely imbalanced image inputs. The cGANs-based architecture consists of two training stages for a refined multiscale feature probability map to indicate crack pixels. Besides, we further investigated the effectiveness of attention mechanisms for the imbalanced problems in the proposed framework.

We have carried out extensive experiments on six benchmark datasets with five quantitative evaluation metrics, precision-and-recall curves, and visual results. The experimental results have shown that our proposed method can achieve a stable performance on diverse benchmark datasets compared to competing methods. For instance, the proposed method cGAN_CBAM (pixel) can enhance the OSD, OIS, and Mean IOU on the DeepCrack-DB by 2.0%, 2.6%, and 1.7% respectively, compared to the SOTA competing method Crackformer II [138]. Our proposed method can improve the performance in severely imbalanced datasets according to the experimental result on CrackLS315. The proposed method cGAN_CBAM (pixel) has achieved the best ODS: 0.5418, OIS: 0.5006, and Mean IOU: 0.6846 compared to the SOTA method DeepCrack [142]. Besides, our proposed methods can obtain probability feature maps closer to the ground truth with hard inputs such as CrackLS315 according to the visual results. The alation study has shown the effectiveness of side networks and Tversky loss function with a variant of our proposed method cGAN_CBAM_Ig (pixel). The side network and Tversky loss function can increase the ODS, OIS, and Mean IOU by 9.3%, 12.8%, and 7.3% separately on DeepCrack-DB. Although extensive experiments have verified the effectiveness and robustness of the proposed method, it still presents inherent limitations. In particular, the network complexity is high since it uses an attention mechanism with skip-connection.

5 CONCLUSIONS

This dissertation presents machine-learning solutions for the classification and regions of interest analysis on imbalanced datasets. The proposed solutions aim to address two research questions: explore saliency information with LDA for an optimal result of the subsequent classification tasks and develop GANs-based techniques working with different losses to carry out regions of interest analysis tasks with severely imbalanced inputs.

Considering the shortcomings of existing LDA-related variants used for classification tasks, we first proposed saliency-based weighted linear discriminant analysis (LDA) variants to solve binary-label classification tasks in [P1]. Our proposed method used the probabilistic saliency estimation to explore the prior information hidden behind the raw input data to reveal the real prominence of each sample to its class so that we can incorporate this prior information into the scatter matrices to balance the contribution of each sample for more authentic information on the optimal sub-space. Our proposed method has achieved a promising performance on facial image datasets BU and KANADE with 13.92% and 4.73% improvement separately compared to the SOTA method. We further proposed a general framework based on the probabilistic saliency estimation approach with a multi-label linear discriminant analysis (MLDA) framework for multi-label classification tasks in [P3]. Our proposed framework not only provides a general way to utilize different kinds of prior information of input data but also mitigates the imbalanced problem widely existing in multi-label datasets because the weight factor is analyzed within each class. We implemented our proposed framework on 17 diverse datasets covering audio, video, and text with a varied number of instances for each dataset. Compared to the SOTA method wMLDA, our proposed method has achieved the best average over all datasets working with different prior information with ML-kNN and LRR classifiers using the ranking loss metric. Moreover, our proposed method outperforms all other non-LDA-based methods with ML-kNN classifier and most other non-LDA-based methods with LRR using the ranking loss metric. Furthermore, our proposed method has achieved the best performance in the seven most imbalanced datasets compared to other non-LDA-based methods with ML-kNN using macro-F1. This contribution further verifies the effectiveness of the probabilistic saliency estimation approach with the LDA technique addressing multi-label classification tasks.

Recall the first research question, how to explore saliency information to properly highlight prominent but minor instances in an imbalanced dataset with the LDA for an optimal result of the subsequent classification tasks, we can conclude that the first contribution [P1] in this dissertation initially explores saliency information based on the misclassification prior information with LDA for an optimal result of the subsequent binary classification tasks. Furthermore, the second contribution [P2] explores saliency information based on six kinds of prior information with LDA for optimal results of the subsequent multilabel classification tasks. These two contributions and the respective results summarized above fully address the first research question.

Although the experimental results from contributions [P1] and [P2] have shown significant improvements in the subsequent classification tasks, the limitation of the proposed methods stems from the complex kernel matrix computation with prior information. For future work, other similarity measurements can be investigated, such as the use of sparse matrices to explore the saliency information in the input. Furthermore, kernel LDA can also be explored as a framework for nonlinear subspace for classification tasks in future work.

The second research question asks if we can develop GANs-based techniques working with different losses to carry regions of interest analysis tasks with severely imbalanced inputs. We addressed this question in publications [P3] and [P4]. In [P3], we have investigated the potential of GAN-based methods for editing specific facial attributes in an unsupervised learning manner. Our proposed method is based on the DCGANs architecture with a novel optimal loss function. The loss function is used to detect the occluded facial parts and then the trained DCGANs generate the corresponding occluded facial features to complete a facial image without any occlusion. Our proposed method does not require large-scale datasets with annotations for recovering the ground truth images but only for facial image completion, hence it is impossible to be evaluated with metrics. In work [P4], we further explored the potential of GAN-based methods for solving a binary semantic segmentation task with pavement images with anomalous crack pixels, which is a typical imbalanced problem. We have proposed a cGANs-based method with a novel auxiliary network working in two stages iteratively for a refined probability feature map for crack detection. Moreover, we further investigated the effectiveness of attention mechanisms and losses for robust results on diverse datasets. We have implemented the method on six benchmark datasets and evaluated the experimental results with five quantitative metrics, precision-and-recall curves, and visual results.

These contributions fully address the second research question as follows. From the visual results in [P3], we can conclude that the proposed method successfully provides specific facial attributes editing tasks in an unsupervised manner based on the DCGANs architecture and an optimal loss function. While a cGANS-based framework solution proposed in [P4] can segment anomalous crack pixels effectively based on attention mechanisms and entropy-based losses.

The extensive experimental results have verified the proposals in the contributions [P3] and [P4]. However, there are still several inherent limitations existing in [P3] and [P4]. In particular, the third contribution does not contain any quantitative evaluation due to a lack of ground truth. Besides, the proposed method in [P3] works well on specific images captured under a proper condition which lacks generality. Moreover, the computation complexity is still high, especially with the LSA attention module. Therefore, it is worthy to further extend the contribution [P3] with a semi-supervised learning manner and evaluate the final results using quantitative metrics. Additional potential aspects for further study could include investigating the compact or lightweight networks instead of the current ones in the contribution [P4]. Furthermore, further studies can target carrying out the proposed framework in [P4] as a solution for binary semantics segmentation on medical images.

REFERENCES

[1] V. Sampath, I. Maurtua, J. José, A. Martín, and A. Gutierrez, *A survey on generative adversarial networks for imbalance problems in computer vision tasks*. Springer International Publishing, 2021, ISBN: 4053702100. DOI: 10.1186/s40537-021-00414-0.

[2] N. Chawla, K. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002. DOI: 10.1613/jair.953.

[3] J. Xu, J. Liu, J. Yin, and C. Sun, "A multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence simultaneously," *Knowledge-Based Systems*, vol. 98, pp. 172–184, 2016, ISSN: 09507051. DOI: 10.1016/j.knosys.2016.01.032.

[4] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. M. Schmidt-Erfurth, "F-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical Image Analysis*, vol. 54, pp. 30–44, 2019. DOI: 10.1016/j.media.2019.01.010.

[5] J. Xu, "A weighted linear discriminant analysis framework for multi-label feature extraction," *Neurocomputing*, vol. 275, pp. 107–120, 2018, ISSN: 18728286. DOI: 10.1016/j.neucom.2017.05.008.

[6] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018, ISSN: 0893-6080. DOI: 10.1016/j.neunet.2018.07.011.

[7] Y. Cai, L. Dai, H. Wang, L. Chen, and Y. Li, "A Novel Saliency Detection Algorithm Based on Adversarial Learning Model," *IEEE Transactions on Image Processing*, vol. 29, pp. 4489–4504, 2020, ISSN: 19410042. DOI: 10.1109/TIP.2020.2972692.

[8] M. Rezaei *et al.*, "A conditional adversarial network for semantic segmentation of brain tumor," in *Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2017. Lecture Notes in Computer Science()*, vol. 10670, Springer, Cham, 2018. DOI: https://doi.org/10.1007/978-3-319-75238-9_21.

[9] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," *arXiv preprint arXiv:1411.1784v1*, pp. 1–7, 2014.

[10] W. Siblini, P. Kuntz, and F. Meyer, "A Review on Dimensionality Reduction for Multi-label Classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 8, 2019, ISSN: 15582191. DOI: 10.1109/TKDE.2019.2940014.

[11] S. V.B. and J. M. David, "Significance of Dimensionality Reduction in Image Processing," *Signal & Image Processing : An International Journal*, vol. 6, no. 3, pp. 27–42, 2015. DOI: 10.5121/sipij.2015.6303.

[12] R. Houari, A. Bounceur, M. T. Kechadi, A. K. Tari, and R. Euler, "Dimensionality reduction in data mining: A Copula approach," *Expert Systems with Applications*, vol. 64, pp. 247–260, 2016, ISSN: 09574174. DOI: 10.1016/j.eswa.2016.07.041.

[13] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Addressing imbalance in multilabel classification: Measures and random resampling algorithms," *Neurocomputing*, vol. 163, pp. 3–16, 2015, ISSN: 18728286. DOI: 10.1016/j.neucom.2014.08.091.

[14] X. Y. Jing, D. Zhang, and Y. Y. Tang, "An improved LDA approach," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 5, pp. 1942–1951, 2004, ISSN: 10834419. DOI: 10.1109/TSMCB.2004.831770.

[15] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 194–200, 2011, ISSN: 01628828. DOI: 10.1109/TPAMI.2010.160.

[16] C. Shen, M. Sun, M. Tang, and C. E. Priebe, "Generalized canonical correlation analysis for classification," *Journal of Multivariate Analysis*, vol. 130, pp. 310–322, 2014, ISSN: 10957243. DOI: 10.1016/j.jmva.2014.05.011.

[17] K. Yu, S. Yu, and V. Tresp, "Multi-label informed latent semantic indexing," *SIGIR 2005 - Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 258–265, 2005. DOI: 10.1145/1076034.1076080.

[18] Y. Zhang and Z. H. Zhou, "Multilabel dimensionality reduction via dependence maximization," *ACM Transactions on Knowledge Discovery from Data*, vol. 4, no. 3, pp. 1–21, 2010, ISSN: 15564681. DOI: 10.1145/1839490.1839495.

[19] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936, ISSN: 20501420. DOI: 10.1111/j.1469-1809.1936.tb02137.x.

[20] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007, ISSN: 10636919. DOI: 10.1109/CVPR.2007.382983.

[21] Z. Li, F. Nie, X. Chang, and Y. Yang, "Beyond trace ratio: Weighted harmonic mean of trace ratios for multiclass discriminant analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 10, pp. 2100–2110, 2017, ISSN: 10414347. DOI: 10.1109/TKDE.2017.2728531.

[22] S. Petridis and S. J. Perantonis, "On the relation between discriminant analysis and mutual information for supervised linear feature extraction," *Pattern Recognition*, vol. 37, no. 5, pp. 857–874, 2004, ISSN: 00313203. DOI: 10.1016/j.patcog.2003.12.002.

[23] E. K. Tang, P. N. Suganthan, X. Yao, and A. K. Qin, "Linear Dimensionality Reduction Using Relevance Weighted LDA," *Pattern Recognition*, vol. 38, no. 4, pp. 485–493, 2005, ISSN: 00313203. DOI: 10.1016/j.patcog.2004.09.005.

[24] E. Tang, P. Suganthan, and X. Yao, “Generalized LDA using relevance weighting and evolution strategy,” *Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No.04TH8753)*, no. 1, pp. 2230–2234, 2005. DOI: 10.1109/cec.2004.1331174.

[25] D. Jarchi and R. Boostani, “A New Weighted LDA Method in Comparison to Some Versions of LDA,” *Proceedings of Word Academy of Science, Engineering and Technology*, vol. 18, no. 12, pp. 233–238, 2006.

[26] M. Loog, R. Duin, and R. Haeb-Umbach, “Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 762–766, 2001, ISSN: 1521-4141. DOI: 10.1002/eji.1830070516.

[27] S. Godbole and S. Sarawagi, “Discriminative methods for multi-labeled classification,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3056, pp. 22–30, 2004, ISSN: 16113349.

[28] H. Wang, C. Ding, and H. Huang, “Multi-label linear discriminant analysis,” *Lecture Notes in Computer Science*, vol. 6316 LNCS, no. PART 6, pp. 126–139, 2010, ISSN: 03029743. DOI: 10.1007/978-3-642-15567-3_10.

[29] G. P. Zhang, “data mining and knowledge discovery handbook,” in *Soft Computing for Knowledge Discovery and Data Mining*, 2008, pp. 667–685, ISBN: 9780387699349. DOI: 10.1007/978-0-387-69935-6_2.

[30] M. L. Zhang and Z. H. Zhou, “A review on multi-label learning algorithms,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014, ISSN: 10414347. DOI: 10.1109/TKDE.2013.39.

[31] Q. Wu, M. Tan, H. Song, J. Chen, and M. K. Ng, “ML-FOREST: A multi-label tree ensemble method for multi-label classification,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2665–2680, 2016, ISSN: 10414347. DOI: 10.1109/TKDE.2016.2581161.

[32] M. Oikonomou and A. Tefas, “Direct Multi-label Linear Discriminant Analysis,” *Communications in Computer and Information Science*, vol. 383 CCIS, no. PART 1, pp. 414–423, 2013, ISSN: 18650929. DOI: 10.1007/978-3-642-41013-0_43.

[33] Y. Yuan, K. Zhao, and H. Lu, “Multi-label Linear Discriminant Analysis with Locality Consistency,” in *International Conference on Neural Information Processing*, 2014, pp. 386–394.

[34] W. Wei, Y. Cheng, J. He, and X. Zhu, “A review of small object detection based on deep learning,” *Neural Computing and Applications*, pp. 1–21, 2024.

[35] W. Hugo, L. Pinaya, P.-d. Tudosiu, R. Gray, and G. Rees, “Unsupervised Brain Anomaly Detection and Segmentation with Transformers,” *Proceedings of Machine Learning Research – Under Review*, pp. 1–22, 2021.

[36] J. Son, S. J. Park, and K.-H. Jung, “Retinal Vessel Segmentation in Fundoscopic Images with Generative Adversarial Networks,” *arXiv:1706.09318*, 2017.

[37] M. Rezaei, H. Yang, K. Harmuth, and C. Meinel, “Conditional generative adversarial refinement networks for unbalanced medical image semantic segmentation,” *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*, pp. 1836–1845, 2019. DOI: 10.1109/WACV.2019.00200.

[38] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang, and S. Wang, “DeepCrack: Learning hierarchical convolutional features for crack detection,” *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1498–1512, 2019, ISSN: 10577149.

[39] T. Nakazawa and D. V. Kulkarni, “Anomaly detection and segmentation for wafer defect patterns using deep Convolutional Encoder-Decoder Neural Network Architectures in Semiconductor Manufacturing,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 32, no. 2, pp. 250–256, 2019, ISSN: 15582345. DOI: 10.1109/TSM.2019.2897690.

[40] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[41] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020, ISSN: 19393539. DOI: 10.1109/TPAMI.2018.2858826.

[42] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10541 LNCS, 2017, pp. 379–387, ISBN: 9783319673882. DOI: 10.1007/978-3-319-67389-9_44.

[43] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8. DOI: 10.1109/CVPR.2007.383267.

[44] R. Sun, T. Lei, Q. Chen, Z. Wang, X. Du, W. Zhao, and A. K. Nandi, "Survey of Image Edge Detection," *Frontiers in Signal Processing*, vol. 2, no. March, pp. 1–13, 2022. DOI: 10.3389/frsip.2022.826967.

[45] C. Aytekin, A. Iosifidis, and M. Gabbouj, "Probabilistic saliency estimation," *Pattern Recognition*, vol. 74, pp. 359–372, 2018, ISSN: 00313203. DOI: 10.1016/j.patcog.2017.09.023.

[46] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 9351, 2015, pp. 234–241, ISBN: 9783319245737. DOI: 10.1007/978-3-319-24574-4_28.

[47] S. Mohammadi, M. Noori, A. Bahri, S. Ghofrani Majelan, and M. Havaei, "CAGNet: Content-Aware Guidance for Salient Object Detection," *Pattern Recognition*, vol. 103, pp. 1–25, 2020, ISSN: 00313203. DOI: 10.1016/j.patcog.2020.107303.

[48] X. Zhao, T. Yang, B. Li, and X. Zhang, "SwinGAN: A dual-domain Swin Transformer-based generative adversarial network for MRI reconstruction," *Computers in Biology and Medicine*, vol. 153, no. August 2022, p. 106 513, 2023.

[49] R. Zhao, B. Qian, X. Zhang, Y. Li, R. Wei, Y. Liu, and Y. Pan, "Rethinking dice loss for medical image segmentation," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, vol. 2020-Novem, 2020, pp. 851–860, ISBN: 9781728183169. DOI: 10.1109/ICDM50108.2020.00094.

[50] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, "Imbalance Problems in Object Detection: A Review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1, 2020, ISSN: 0162-8828. DOI: 10.1109/tpami.2020.2981890.

[51] W. D. Jin, J. Xu, M. M.-m. Cheng, Y. Zhang, and W. Guo, "ICNet: Intra-saliency correlation network for co-saliency detection," in *Advances in Neural Information Processing Systems*, vol. 2020-Decem, 2020, pp. 1–11.

[52] X. Xia, X. Pan, N. Li, X. He, L. Ma, X. Zhang, and N. Ding, "GAN-based anomaly detection: A review," *Neurocomputing*, vol. 493, pp. 497–535, 2022, ISSN: 18728286.

[53] Z. Lin, H. Wang, and S. Li, "Pavement anomaly detection based on transformer and self-supervised learning," *Automation in Construction*, vol. 143, no. April, p. 104 544, 2022, ISSN: 09265805. DOI: 10.1016/j.autcon.2022.104544.

[54] M. Futrega, A. Milesi, M. Marcinkiewicz, and P. Ribalta, "Optimized U-Net for Brain Tumor Segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12963 LNCS, 2022, pp. 15–29, ISBN: 9783031090011.

[55] V. Tankovich, C. Häne, Y. Zhang, A. Kowdle, S. Fanello, and S. Bouaziz, "HitNet: Hierarchical Iterative Tile Refinement Network for Real-time Stereo Matching," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 14 357–14 367, Jul. 2021, ISSN: 10636919. DOI: 10.1109/CVPR46437.2021.01413.

[56] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020, ISSN: 18792782. DOI: 10.1016/j.neunet.2019.08.025.

[57] G. Harshvardhan, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, “A comprehensive survey and analysis of generative models in machine learning,” *Computer Science Review*, vol. 38, p. 100 285, 2020. DOI: 10.1016/j.cosrev.2020.100285.

[58] F. Luleci and F. N. Catbas, “A brief introductory review to deep generative models for civil structural health monitoring,” *AI in Civil Engineering*, vol. 2, no. 9, 2023.

[59] B. Li, Z. Sun, and Y. Guo, “SuperVAE: Superpixelwise variational autoencoder for salient object detection,” in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, vol. 1, 2019, pp. 8569–8576, ISBN: 9781577358091. DOI: 10.1609/aaai.v33i01.33018569.

[60] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, “Diffusion Models for Medical Anomaly Detection,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13438 LNCS, 2022, pp. 35–45, ISBN: 9783031164514. DOI: 10.1007/978-3-031-16452-1_4.

[61] Y. Zhang and L. Zhang, “Detection of Pavement Cracks by Deep Learning Models of Transformer and UNet,” *arXiv preprint arXiv:2304.12596*, 2023.

[62] Z. Pan, S. L. Lau, X. Yang, N. Guo, and X. Wang, “Automatic pavement crack segmentation using a generative adversarial network (GAN)-based convolutional neural network,” *Results in Engineering*, vol. 19, no. September 2022, p. 101 267, 2023.

[63] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, “AttGAN: Facial Attribute Editing by only Changing What You Want,” *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464–5478, 2019, ISSN: 19410042. DOI: 10.1109/TIP.2019.2916751.

[64] Z. Huang, K. C. Chan, Y. Jiang, and Z. Liu, “Collaborative diffusion for multi-modal face generation and editing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[65] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, “Robust LSTM-Autoencoders for Face De-Occlusion in the Wild,” *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 778–790, 2018, ISSN: 10577149. DOI: 10.1109/TIP.2017.2771408.

[66] Q. Sun, J. Guo, and Y. Liu, “PattGAN: Pluralistic Facial Attribute Editing,” *IEEE Access*, vol. 10, no. June, pp. 68 534–68 544, 2022.

[67] A. Anandakrishnan, S. Kumar, A. Statnikov, and D. Xu, “Anomaly Detection in Finance: Editors’ Introduction Capital One Capital One,” *Proceedings of Machine Learning Research*, vol. 71, pp. 1–7, 2017.

[68] B. Staar, M. Lütjen, and M. Freitag, “Anomaly detection with convolutional neural networks for industrial surface inspection,” *Procedia CIRP*, vol. 79, pp. 484–489, 2019, ISSN: 22128271. DOI: 10.1016/j.procir.2019.02.123.

[69] X. Chen and E. Konukoglu, “Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders,” *Medical Imaging with Deep Learning*, pp. 1–9, 2018, ISSN: 23318422. DOI: 10.3929/ethz-b-000321650.

[70] K. Zhou, Y. Xiao, J. Yang, J. Cheng, W. Liu, W. Luo, Z. Gu, J. Liu, and S. Gao, “Encoding Structure-Texture Relation with P-Net for Anomaly Detection in Retinal Images,” in *European Conference on Computer Vision*, Springer International Publishing, 2020, pp. 360–377, ISBN: 9783030585648. DOI: 10.1007/978-3-030-58565-5_22.

[71] P. S. S. Nithya, P. Sathya, and S. Pradeepa, “A nondestructive sensing robot for crack detection and deck maintenance,” *International Journal For Research in Applied Science and Engineering Technology*, vol. 3, no. Iv, pp. 663–673, 2015.

[72] I. Schiopu, J. P. Saarinen, L. Kettunen, and I. Tabus, “Pothole detection and tracking in-car video sequence,” *2016 39th International Conference on Telecommunications and Signal Processing, TSP 2016*, pp. 701–706, 2016. DOI: 10.1109/TSP.2016.7760975.

[73] A. Cubero-Fernandez, F. J. Rodriguez-Lozano, R. Villatoro, J. Olivares, and J. M. Palomares, "Efficient pavement crack detection and classification," *Eurasip Journal on Image and Video Processing*, vol. 2017, no. 1, 2017, ISSN: 16875281. DOI: 10.1186/s13640-017-0187-0.

[74] F. Yang, L. Zhang, S. Yu, D. Prokhorov, X. Mei, and H. Ling, "Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2019, ISSN: 1524-9050. DOI: 10.1109/tits.2019.2910595.

[75] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning Where to Look for the Pancreas," *arXiv:1804.03999*, 2018.

[76] H. Tong, Z. Fang, Z. Wei, Q. Cai, and Y. Gao, "Sat-net: A side attention network for retinal image segmentation," *Applied Intelligence*, vol. 51, pp. 5146–5156, 2021.

[77] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X. H. Han, Y. W. Chen, and R. Tong, "Mixed Transformer U-Net for Medical Image Segmentation," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2022-May, pp. 2390–2394, 2022.

[78] D. Davletshina, V. Melnychuk, V. Tran, H. Singla, M. Berrendorf, E. Faerman, M. Fromm, and M. Schubert, "Unsupervised anomaly detection for X-ray images," *arXiv preprint arXiv:2001.10883.*, 2020.

[79] V. Nguyen, T. F. Y. Vicente, M. Zhao, M. Hoai, D. Samaras, and S. Brook, "Shadow Detection with Conditional Generative Adversarial Networks," *IEEE International Conference on Computer Vision Shadow*, 2017. DOI: 10.1109/ICCV.2017.483.

[80] Y. Yan, S. Zhu, S. Ma, Y. Guo, and Z. Yu, "CycelADC-Net: A crack segmentation method based on multi-scale feature fusion," *Measurement: Journal of the International Measurement Confederation*, vol. 204, no. October, p. 112 107, 2022, ISSN: 02632241.

[81] H. M. and S. M.N, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, pp. 01–11, 2015. DOI: 10.5121/ijdkp.2015.5201.

[82] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, ISSN: 01678655. DOI: 10.1016/j.patrec.2005.10.010.

[83] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006, ISSN: 15337928.

[84] X.-Z. Wu and Z.-H. Zhou, "A Unified View of Multi-Label Performance Measures," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, 2017, pp. 3780–3788.

[85] E. Gibaja and S. Ventura, "Multi-label learning: A review of the state of the art and ongoing research," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 6, pp. 411–444, 2014, ISSN: 19424795. DOI: 10.1002/widm.1139.

[86] J. Davis and M. Goadrich, "The Relationship between Precision-Recall and ROC Curves," in *Proceedings of the 23rd international conference on Machine learning*, New York, NY, USA, 2006, pp. 233–240. DOI: 10.1145/1143844.1143874.

[87] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, 2015. DOI: 10.1371/journal.pone.0118432.

[88] H. Liu, X. Miao, C. Mertz, C. Xu, and H. Kong, "CrackFormer: Transformer Network for Fine-Grained Crack Detection," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3763–3772, 2021, ISSN: 15505499. DOI: 10.1109/ICCV48922.2021.00376.

[89] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[90] A. M. Treisman and G. Gelade, "A Feature-Integration Theory of Attention," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[91] M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu, “Global Contrast based Salient Region Detection,” *2011 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 37, no. 3, pp. 409–416, 2011, ISSN: 1063-6919. DOI: 10.1109/CVPR.2011.5995344.

[92] S. Jetley, N. Murray, and E. Vig, “End-to-end saliency mapping via probability distribution prediction,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 5753–5761, 2016, ISSN: 10636919. DOI: 10.1109/CVPR.2016.620.

[93] B. Yu, L. Jin, and P. Chen, “A new LDA-based method for face recognition,” *Proceedings - International Conference on Pattern Recognition*, vol. 16, no. 1, pp. 168–171, 2002, ISSN: 10514651. DOI: 10.1109/icpr.2002.1044639.

[94] Z. Li, D. Lin, and X. Tang, “Nonparametric discriminant analysis for face recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 755–761, 2009, ISSN: 01628828. DOI: 10.1109/TPAMI.2008.174.

[95] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, “A 3D facial expression database for facial behavior research,” *FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, vol. 2006, pp. 211–216, 2006. DOI: 10.1109/FGR.2006.6.

[96] T. Kanade and J. Cohn, “Comprehensive Database for Facial Expression Analysis,” *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 46–53, 2000, ISSN: 0-7695-0580-5. DOI: 10.1109/AFGR.2000.840611.

[97] M. J. Lyons, M. Kamachi, and J. Gyoba, “Coding facial expressions with Gabor wavelets (IVC special issue),” *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, ISSN: 23318422. DOI: 10.5281/zenodo.4029679.

[98] F. S. Samaria and A. C. Harter, “Parameterisation of a stochastic model for human face identification,” *IEEE Workshop on Applications of Computer Vision - Proceedings*, pp. 138–142, 1994. DOI: 10.1109/acv.1994.341300.

[99] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001, ISSN: 01628828. DOI: 10.1109/34.927464.

[100] A. Martinez and R. Benavente, "The AR Face Database," *CVC Technical Report #24*, 1998.

[101] F. Nie, X. Wang, M. I. Jordan, and H. Huang, "The constrained laplacian rank algorithm for graph-based clustering," *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, no. 1, pp. 1969–1976, 2016.

[102] C. H. Park and M. Lee, "On applying linear discriminant analysis for multi-labeled problems," *Pattern Recognition Letters*, vol. 29, no. 7, pp. 878–887, 2008, ISSN: 01678655. DOI: 10.1016/j.patrec.2008.01.003.

[103] W. Chen, J. Yan, B. Zhang, Z. Chen, and Q. Yang, "Document transformation for multi-label feature selection in text categorization," *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 451–456, 2007, ISSN: 15504786. DOI: 10.1109/ICDM.2007.18.

[104] X. Lin and X.-w. Chen, "Mr. KNN - Soft relevance for multi-label classification.," in *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, 2010, pp. 349–358, ISBN: 9781450300995.

[105] A. Gretton, O. Bousquet, A. Smola, and B. Sclkopf, "Measuring statistical dependence with Hilbert-Schmidt norms," *Proceedings of the Sixteenth International Conference on Algorithmic Learning Theory (ALT 2005)*, pp. 63–77, 2005, ISSN: 03029743. DOI: 10.1007/11564089_7.

[106] I. Katakis, G. Tsoumakas, and V. Ioannis, "Multilabel Text Classification for Automated Tag Suggestion," in *Proceedings of the ECML/PKDD 2008 Discovery Challenge*, 2008. DOI: 10.1109/cca.2001.973936.

[107] F. Briggs *et al.*, "The 9th annual MLSP competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment," Southampton, UK, 2013, pp. 1–8. DOI: 10.1109/MLSP.2013.6661934.

[108] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008, ISSN: 15587916. DOI: 10.1109/TASL.2007.913750.

[109] H. Shao, G. Li, G. Liu, and Y. Wang, "Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine," *Science China Information Sciences*, vol. 56, no. 5, pp. 1–13, 2013, ISSN: 1098-6596.

[110] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan, "Matching Words and Pictures," *Journal of Machine Learning Research*, vol. 3, no. 6, pp. 1107–1135, 2003, ISSN: 15324435. DOI: 10.1162/153244303322533214.

[111] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Effective and efficient multilabel classification in domains with large number of labels," in *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)*, 2008, pp. 30–44.

[112] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier Chains for Multi-label Classification," *Joint European Conference on Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2009. Lecture Notes in Computer Science*, vol. 5782, pp. 254–269, 2009. DOI: 10.1007/978-3-642-04174-7_17.

[113] J. Xu, "Fast multi-label core vector machine," *Pattern Recognition*, vol. 46, no. 3, pp. 885–898, 2013, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2012.09.003.

[114] M. L. Zhang and Z. H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, pp. 2038–2048, 2007, ISSN: 00313203. DOI: 10.1016/j.patcog.2006.12.019.

[115] J. P. Pestian, C. Brew, P. Matykiewicz, D. J. Hovermale, N. Johnson, K. B. Cohen, and W. Duch, "A shared task involving multi-label classification of clinical free text," *ACL 2007 - Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, no. June, pp. 97–104, 2007.

[116] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, “Learning multi-label scene classification,” *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004, ISSN: 00313203. DOI: 10.1016/j.patcog.2004.03.009.

[117] A. N. Srivastava and B. Zane-Ulman, “Discovering recurring anomalies in text reports regarding complex space systems,” in *IEEE Aerospace Conference*, 2005, pp. 3853–3862, ISBN: 0780388704. DOI: 10.1109/AERO.2005.1559692.

[118] K. Nakai and M. Kanehisa, “A Knowledge Base for Predicting Protein Localization Sites in Eukaryotic Cells,” *Genomics*, vol. 14, pp. 897–911, 1992.

[119] H. Sajnani, V. Saini, K. Kumar, E. Gabrielova, P. Choudary, and C. Lopes, “Classifying Yelp reviews into relevant categories,” *Mondego Group, Univ. California Press, Berkeley, CA USA, Tech. Rep*, 2013.

[120] H. Borchani, G. Varando, C. Bielza, and B. Monte, “A survey on multi-output regression,” in *WIREs Data Min. Knowl. Discov.* 5, vol. 5, 2015, pp. 216–233.

[121] C. Tan, S. Chen, G. Ji, and X. Geng, “Multilabel Distribution Learning Based on Multioutput Regression and Manifold Learning,” *IEEE Transactions on Cybernetics*, pp. 1–15, 2020, ISSN: 2168-2267. DOI: 10.1109/tcyb.2020.3026576.

[122] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” in *NIPS*, 2013, p. iii, ISBN: 978-0-12-088571-8. DOI: 10.1016/B978-0-12-088571-8.01001-9.

[123] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. DOI: 10.1109/CVPR.2017.632.

[124] M. A. Souibgui and Y. Kessentini, “DE-GAN: A Conditional Generative Adversarial Network for Document Enhancement,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, ISSN: 19393539. DOI: 10.1109/TPAMI.2020.3022406.

[125] M. Hu, D. Zhou, and Y. He, “Variational conditional GAN for fine-grained controllable image generation,” in *Machine Learning Research*, vol. 101, 2019, pp. 109–124.

[126] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs,” *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[127] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” in *arXiv preprint arXiv:1511.06434*, 2015, pp. 1–16.

[128] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” *Proceedings of COMPSTAT 2010 - 19th International Conference on Computational Statistics, Keynote, Invited and Contributed Papers*, pp. 177–186, 2010. DOI: 10.1007/978-3-7908-2604-3_16.

[129] X. Shu, H. Xu, and L. Tao, “A least squares formulation of multi-label linear discriminant analysis,” *Neurocomputing*, vol. 156, pp. 221–230, 2015, ISSN: 18728286. DOI: 10.1016/j.neucom.2014.12.057.

[130] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Globally and locally consistent image completion,” *ACM Transactions on Graphics*, vol. 36, no. 4, 2017, ISSN: 15577368. DOI: 10.1145/3072959.3073659.

[131] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, “Semantic image inpainting with deep generative models,” *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 6882–6890, 2017. DOI: 10.1109/CVPR.2017.728.

[132] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep Learning Face Attributes in the Wild,” *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[133] A. Brandon, B. Ludwiczuk, and S. Mahadev, “OpenFace: A general-purpose face recognition library with mobile applications,” *CMU-CS-16-118, CMU School of Computer Science*, 2016.

[134] J, Serra and L. Vincent, "An overview of morphological filtering," *Circuits Systems and Signal Process 11*, pp. 47–108, 1992. DOI: https://doi.org/10.1007/BF01189221. [Online]. Available: https://scikit-image.org/docs/stable/auto_examples/applications/plot_morphology.html (visited on 01/31/2024).

[135] N. Efford, *Digital Image Processing: A Practical Introduction Using Java (with CD-ROM).* Addison-Wesley Longman Publishing Co., Inc., 2000.

[136] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 1951–1959, 2017. DOI: 10.1109/CVPR.2017.211.

[137] S. Woo, J. Park, J.-y. Lee, and I. S. Kweon, "CBAM : Convolutional Block Attention Module," in *European Conference on Computer Vision*, 2018.

[138] H. Liu, J. Yang, X. Miao, C. Mertz, and H. Kong, "CrackFormer Network for Pavement Crack Segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 9, pp. 9240–9252, 2023, ISSN: 15580016. DOI: 10.1109/TITS.2023.3266776.

[139] F. Laakom, K. Chumachenko, J. Raitoharju, A. Iosifidis, and M. Gabbouj, "Learning to ignore: rethinking attention in CNNs," in *The British Machine Vision Conference (BMVC)*, 2021, pp. 1–13.

[140] T. M. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc, 1991, vol. 1, pp. 12–49, ISBN: 0471062596.

[141] S. Xie and Z. Tu, "Holistically-nested edge detection," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 1395–1403, 2015, ISSN: 15505499. DOI: 10.1109/ICCV.2015.164.

[142] Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li, "DeepCrack: A deep hierarchical feature learning architecture for crack segmentation," *Neurocomputing*, vol. 338, pp. 139–153, 2019. DOI: 10.1016/j.neucom.2019.01.036.

[143] N. Otsu, P. L. Smith, D. B. Reid, C. Environment, L. Palo, P. Alto, and P. L. Smith, "Otsu_1979_otsu_method," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. C, no. 1, pp. 62–66, 1979, ISSN: 0018-9472.

PUBLICATIONS

PUBLICATION
I

Weighted linear discriminant analysis based on class saliency information

L. Xu, A. Iosifidis, and M. Gabbouj

In *2018 25th IEEE International Conference on Image Processing (ICIP 2018)*, Athens, Greece, 7-10 October 2018

DOI: 10.1109/ICIP.2018.8451614

WEIGHTED LINEAR DISCRIMINANT ANALYSIS BASED ON CLASS SALIENCY INFORMATION

Lei Xu[1], Alexandros Iosifidis[2] and Moncef Gabbouj[1]

[1]Laboratory of Signal Processing, Tampere University of Technology, Tampere, Finland
[2] Department of Engineering, Electrical & Computer Engineering, Aarhus University, Aarhus, Denmark

ABSTRACT

In this paper, we propose a new variant of Linear Discriminant Analysis to overcome underlying drawbacks of traditional LDA and other LDA variants targeting problems involving imbalanced classes. Traditional LDA sets assumptions related to Gaussian class distribution and neglects influence of outlier classes, that might hurt in performance. We exploit intuitions coming from a probabilistic interpretation of visual saliency estimation in order to define saliency of a class in multi-class setting. Such information is then used to redefine the between-class and within-class scatters in a more robust manner. Compared to traditional LDA and other weight-based LDA variants, the proposed method has shown certain improvements on facial image classification problems in publicly available datasets.

Index Terms**—** Visual saliency estimation, Fisher's discriminant criterion, Linear Discriminant Analysis(LDA)

1. INTRODUCTION

Linear Discriminant Analysis (LDA), as a traditional statistical machine learning technique, has been employed for several classification tasks, such as human action recognition [1], face recognition [2], [3], and person identification [4], due to its effectiveness in reducing dimensions and extracting discriminative features. In a classification task, LDA is used to define an optimal projection by means of Fisher criterion optimization. Despite the widespread application of traditional LDA, its performance is affected by several issues related to its underlying assumptions. Traditional LDA represents each class with the corresponding class mean and discriminates between classes based on the scatters of these class representations with respect to the total data mean. Such a class discrimination definition may cause large overlaps of neighboring classes [5], and receive a sub-optimal result, since an outlier class being far from the others dominates the solution [5]. Furthermore, in traditional LDA all classes equally contribute to the within-class scatter definition [6] based on the assumption of the same Gaussian distribution for all classes. This assumption overemphasizes well-separated outlier classes, which should have lower contribution in the overall within-class scatter definition. A method that automatically determines optimized class representations for LDA-based projections was proposed in [7], [8]; however, it also suffers from the class imbalance problems discussed above. In order to overcome aforementioned drawbacks of traditional LDA, extensions imposing weighting strategies for the definition of the within-class and between-class scatters have been proposed in [9], [10], [11], [12], [13]. In these methods, the weighting factors incorporated to the scatter matrices definitions are based on class statistics, e.g. class cardinality, and class representation is still assumed to be the class mean.

A novel extension of LDA that exploits intuitions from saliency [14] is proposed in this paper. A probabilistic criterion is formulated in order to express the samples around boundary within its original class following a probabilistic saliency estimation framework [15]. Such a definition is naturally expressed by graph notation, in which several types of graphs can be exploited. Both fully connected and k-NN graphs are considered. After defining the probability of each sample belonging to its corresponding class, this information is used to define new class representations, as well as new within-class and between-class scatters. Compared to traditional LDA and its weighted variants, the proposed Saliency-based weighted LDA ($SwLDA$) has shown enhanced performance on facial image classification problems.

The remainder of this paper is structured as follows. In Section 2, we briefly present related works. In Section 3, we rigorously derive the proposed $SwLDA$ method on the basis of various weighted LDA methods and saliency estimation. Experimental results on publicly available facial image datasets are provided in Section 4, and Section 5 concludes this work.

2. RELATED WORK

In this section, first we briefly describe original LDA and two of its weighted variants, which have been proposed in order to overcome shortcomings of LDA related to class imbalance problems. Later, visual saliency estimation based on the recently proposed probabilistic interpretation [15] is presented.

In the following, we assume that each training sample is represented by a vector $\mathbf{x}_i \in \mathbb{R}^D$ and is followed by a class label $y_i \in \{1, \ldots, C\}$. A set of training vectors $\mathbf{x}_i$, $i = 1, \ldots, N$ are used in order to define a linear projection from the input space $\mathbb{R}^D$ to a discriminant subspace $\mathbb{R}^d$ such that the representation of the i-th sample is given by $\boldsymbol{z}_i = \mathbf{W}^T \mathbf{x}_i$, where $\mathbf{W} \in \mathbb{R}^{D \times d}$ is the projection matrix to be learned by optimizing class discrimination criteria.

2.1. Linear Discriminant Analysis

LDA defines the optimal data projection matrix $\mathbf{W}$ by maximizing the following criterion

$$J(\mathbf{W}) = \underset{\mathbf{W}}{max} \ \frac{tr(\mathbf{W}^T \mathbf{S}_B \mathbf{W})}{tr(\mathbf{W}^T \mathbf{S}_W \mathbf{W})}, \tag{1}$$

where $\mathbf{S}_W$, $\mathbf{S}_B$ are within-class and between-class scatter matrices respectively, and defined as follows:

$$\mathbf{S}_W = \sum_{c=1}^{C} \sum_{\mathbf{x}_i, \alpha_i^c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T, \tag{2}$$

$$\mathbf{S}_B = \sum_{c=1}^{C} N_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T. \tag{3}$$

In the above, α_i^c is an index denoting whether sample i belongs to class c, i.e. $\alpha_i^c = 1$ if $y_i = c$ and $\alpha_i^c = 0$ otherwise. N_c denotes the cardinality of class c, i.e. $N_c = \sum_{i=1}^{N} \alpha_i^c$ and $\boldsymbol{\mu}_c$ denotes the mean vector of class c, i.e. $\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{\mathbf{x}_i, \alpha_i^c = 1} \mathbf{x}_i$. $\boldsymbol{\mu}$ is the total mean vector $\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$.

The optimal projection matrix $\mathbf{W}$ is obtained through applying eigenvalue decomposition of the matrix $\mathbf{S} = \mathbf{S}_W^{-1} \mathbf{S}_B$ and keeping the eigenvectors corresponding to the largest (up to $C - 1$ in total) eigenvalues.

2.2. Weighted LDA Variants

Weighted versions of LDA aim at scaling the contribution of each class based on their influences on projection, by defining appropriate weights. In [12], between-class scatter matrix is redefined for enhancing robustness in multi-class problems, as follows:

$$\mathbf{S}_b = \sum_{c=1}^{C-1} \sum_{j=c+1}^{C} L_{cj} p_c p_j (\boldsymbol{\mu}_c - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_c - \boldsymbol{\mu}_j)^T, \tag{4}$$

where p_c, p_j denote the prior probability of class c, class j, respectively. L_{cj} expresses the dissimilarity between class c and class j, using a distance function in the Euclidean (or a Mahalanobis) space. In order to reduce the influence of outlier classes, an outlier-class-resistant weighted LDA method is proposed in this work [10] based on Loog's work [12]. They express the between-class scatter using (4) and a new within-class scatter definition is proposed as follows:

$$\mathbf{S}_w = \sum_{c=1}^{C} \sum_{k=1}^{N_c} p_c r_c (\mathbf{x}_k - \boldsymbol{\mu}_c)(\mathbf{x}_k - \boldsymbol{\mu}_c)^T, \tag{5}$$

where $r_c = \sum_{i \neq c} \frac{1}{L_{ic}}$ is a relevance-weight between class c and class i, reducing attention to outlier classes.

Another version of weighted LDA aiming at alleviating the influence of outlier class is proposed in [11]. They define the between-class scatter and within-class scatter as follows:

$$\mathbf{S}_b = \sum_{c=1}^{C-1} \sum_{j=c+1}^{C} n_c n_j w_1(\Delta_{cj})(\boldsymbol{\mu}_c - \boldsymbol{\mu}_j)(\boldsymbol{\mu}_c - \boldsymbol{\mu}_j)^T, \tag{6}$$

$$\mathbf{S}_w = \sum_{c=1}^{C} \sum_{k=1}^{N_c} p_c w_2(\Delta_{c:})(\mathbf{x}_k - \boldsymbol{\mu}_c)(\mathbf{x}_k - \boldsymbol{\mu}_c)^T, \tag{7}$$

where n_c, n_j are the number of samples for class c and class j, in addition, $w_1(\Delta_{cj})$ and $w_2(\Delta_{c:})$ are defined as $\frac{1}{\Delta_{cj}}$ and $\frac{1}{\sum_{j \neq c} \Delta_{cj}}$, respectively. Δ_{cj} is the Fisher's discriminant criterion in the discriminant space determined through applying LDA using the between-class scatter matrix $\mathbf{S}_B$ and the total scatter matrix $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W$, i.e.:

$$\mathbf{w}^* = \underset{\mathbf{w}}{argmax} \{ \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_T \mathbf{w}} \} = \mathbf{S}_T^{-1} (\boldsymbol{\mu}_c - \boldsymbol{\mu}_j), \tag{8}$$

$$\Delta_{cj} = \frac{\mathbf{w}^{*T} \mathbf{S}_B \mathbf{w}^*}{\mathbf{w}^{*T} \mathbf{S}_T \mathbf{w}*}. \tag{9}$$

Using the above definition of Δ_{cj}, in the case where a class is well separated from all others, a smaller value of $w(\Delta_{cj})$ will be used, reducing the influence of that class on the result. Once the new $\mathbf{S}_w$ and $\mathbf{S}_t$ ($\mathbf{S}_t$ = $\mathbf{S}_w$ + $\mathbf{S}_b$) are obtained, the final projection matrix $\mathbf{W}$ can be determined by optimizing the following Fisher's discriminant criterion:

$$J(\mathbf{W}) = \underset{\mathbf{W}}{argmax} \frac{tr(\mathbf{W}^T \mathbf{S}_b \mathbf{W})}{tr(\mathbf{W}^T \mathbf{S}_t \mathbf{W})}. \tag{10}$$

2.3. Visual Saliency Estimation

Visual saliency estimation has gained attention during the last decade, since it can be applied as a pre-processing step for higher level Computer Vision tasks. Recently, Aytekin et al.[15] formulated the salient object segmentation problem based on probabilistic interpretation. Specifically, they defined a probability mass function $P(x)$ encoding the probability that an image region (in the sense of pixel, super-pixel or patch) to depict a salient region. Estimation of $P(x)$ is formulated as an optimization problem enforcing similar regions to have similar probabilities, while any prior information regarding saliency (defined based on the location of each region in the image lattice) can be exploited. This joint optimization is expressed as:

$$\underset{r(x)}{argmin} \left(\sum_i (P(x=x_i))^2 v_i + \frac{1}{2} \left(\sum_{i,j} \left(\left(P(x=x_i) \right)^2 - P(x=x_i) P(x=x_j) \right) w_{i,j} \right) \right) \tag{11}$$

$$s.t. \quad \sum_i P(x = x_i) = 1,$$

where $v_i \geq 0$ denotes prior information for region i by non-negative values and w_{ij} expresses the similarity of regions i and j. The optimization problem in (11) can be expressed using a matrix notation as follows:

$$\mathbf{p}^* = \underset{p}{argmin} (\mathbf{p^T H p}), \tag{12}$$

$$\mathbf{H} = \mathbf{D} - \mathbf{W} + \mathbf{V}, \tag{13}$$

$$s.t. \quad \mathbf{p^T 1} = 1,$$

where $\mathbf{p}$ is a vector having elements $p_i = P(x = x_i)$ corresponding to the probability of each region to be salient. $\mathbf{W}$ is the affinity matrix of a graph having as vertices for the region representations and $\mathbf{D}$ is the corresponding diagonal matrix having elements equal to $D_{ii} = \sum_j \mathbf{W}_{ij}$. $\mathbf{V}$ is a diagonal matrix having elements $[\mathbf{V}]_{ii} = v_i$. In visual saliency, the element V_{ii} expresses the a priori knowledge that an image location belongs to background, that is introduced by the user.

As has been shown in [15], the optimization problem in (12) has a global optimum given by: $\mathbf{p}^*_{pse} = \mathbf{H}^{-1}\mathbf{1}$. Interestingly, the above solution is equivalent to an one-class classification model, making a connection between salient object segmentation and one-class classification problems. In the following, we will use this connection in order to derive a new definition for class-representation and scatter matrices calculation in LDA.

3. SALIENCY-BASED WEIGHTED LINEAR DISCRIMINANT ANALYSIS

This section describes in detail the proposed weighted versions of LDA. We define the contribution of each sample to the corresponding class, and then new class representations and scatter matrices are proposed accordingly. We start by describing the proposed sample weights.

3.1. Sample Weights and Class Representation

Weighted LDA variants represent each class with the corresponding mean vector and define weights based on pair-wise class distances to address the outlier class problem. Such mutation yields a certain improvement over traditional LDA. Nevertheless, it neglects the influences of outlier samples within each class [13], which may affect the classification result greatly. This is due to the fact that all class samples equally contribute to the definition of the class representation and scatter matrix calculation.

In our work, we determine the contribution of each sample based on its *class saliency information*. We define the class saliency information of a sample $\mathbf{x}_i$ based on its probability to belong to its true class y_i. In order to do so, we calculate the probability mass function $P_c(x)$ of each class c independently following the probabilistic saliency estimation (PSE) in [15]. That is, for each class c, we form the corresponding graph $\mathcal{G}_C = \{\mathbf{X}_c, \mathbf{W}_c\}$, where $\mathbf{X}_c \in \mathbb{R}^{D \times N_c}$ is a matrix formed by the samples belonging to class c and $\mathbf{W}_c \in \mathbb{R}^{N_c \times N_c}$ is the graph weight matrix expressing the similarity between the class samples. Any type of graph can be used to this end. In our experiments we have used fully connected and the k-NN graphs, using the heat kernel function:

$$W_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{2\sigma^2}\right), \tag{14}$$

where the value of σ is set equal to the mean Euclidean distance between the class samples, which is the natural scaling factor for each class.

We define a priori saliency information as misclassification-based probability for the class data to be set in the diagonal elements of the matrix $\mathbf{V}_c$. Misclassification-based probability assumes that a sample is less probable to have high saliency information if it is closer to another class, when compared to its true class. In this case, the elements of $\mathbf{V}_c$ are set equal to:

$$V_{c,ii} = \begin{cases} 0, & if \ d^c_{c,i} < \underset{k \neq c}{min} \ d^k_{c,i}, \\ \frac{d^c_{c,i}}{\underset{k \neq c}{min} \ d^k_{c,i}}, & otherwise, \end{cases} \tag{15}$$

where $d^k_{c,i} = \|\mathbf{x}_{c,i} - \boldsymbol{\mu}_k\|^2_2$. In this case, a sample which is close to another class is assigned to low saliency information, even if it may be close to the center of its class.

After having defined the matrices $\mathbf{W}_c$ and $\mathbf{V}_c$, the probability of each sample $\mathbf{x}_{c,i}$ to belong to class c is given by: $\mathbf{p}_c = \mathbf{H}_c^{-1}\mathbf{1}$, where $\mathbf{H}_c = \mathbf{D}_c - \mathbf{W}_c + \mathbf{V}_c$ and $\mathbf{D}_{c,ii} = \sum_j \mathbf{W}_{c,ij}$. Having obtained $\mathbf{p}_c \in \mathbb{R}^{N_c}$, $c = 1, \dots, C$, we define a new class representation as $\mathbf{m}_c = \mathbf{X}_c\mathbf{p}_c$.

3.2. Scatter Matrices Definition

By exploiting class-specific saliency information described above, we can define within-class scatter matrix in two different ways. The first one is to incorporate $\mathbf{p}_c$ in $\mathbf{S}_w$ as:

$$\mathbf{S}_w^{(1)} = \textstyle\sum_{c=1}^C \sum_{j=1}^{N_c} p_{c,j}(\mathbf{x}_{c,j} - \boldsymbol{\mu}_c)(\mathbf{x}_{c,j} - \boldsymbol{\mu}_c)^T, \tag{16}$$

where $\mathbf{x}_{c,j}$ denotes j-th sample in class c, $p_{c,j}$ is saliency score for j-th sample in class c. The other one is inspired by relevance weighted LDA mentioned in section 2, as:

$$\mathbf{S}_w^{(2)} = \textstyle\sum_{c=1}^C \sum_{j=1}^{N_c} p_{c,j} r_c(\mathbf{x}_{c,j} - \boldsymbol{\mu}_c)(\mathbf{x}_{c,j} - \boldsymbol{\mu}_c)^T. \tag{17}$$

Here $r_c = \sum_{i \neq c} \frac{1}{L_{ic}}$ is a relevance-weight, where L_{ic} is defined based on the Euclidean distance between pairwise mean vectors of class i and class c, as (18):

$$L_{ic} = \sqrt{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_c)^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_c)}. \tag{18}$$

Definitions of between-class scatter matrix in aforementioned LDA methods simply maximize either the variations between each class mean vector and the total mean vector, or the variations between class pairs. Here, we propose four types of between-class scatter matrices, which are not only based on the aforementioned definition of $\mathbf{S}_b$, but also capture the structure inside each class. The first definition is the same as (3):

$$\mathbf{S}_b^{(1)} = \textstyle\sum_{c=1}^C N_c(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T. \tag{19}$$

The second one uses saliency scores $\mathbf{p}_c$, when generating new class representations, as follows:

$$\hat{\boldsymbol{\mu}}_c = \mathbf{X}_c\mathbf{p}_c, \tag{20}$$

$$\mathbf{S}_b^{(2)} = \textstyle\sum_{c=1}^C (\hat{\boldsymbol{\mu}}_c - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}}_c - \boldsymbol{\mu})^T, \tag{21}$$

Table 1. Classification accuracy of proposed $SwLDA$

Dataset	BU		KANADE		JAFFE		ORL		YALE		AR	
K	1	$\min(5, 0.1 * N_c)$	1	$\min(5, 0.1 * N_c)$	1	$\min(5, 0.1 * N_c)$	1	$\min(5, 0.1 * N_c)$	1	$\min(5, 0.1 * N_c)$	1	$\min(5, 0.1 * N_c)$
$SwLDA_{11}$	0.5714	0.5714	0.6816	0.6939	0.5619	0.5762	0.9700	0.9700	**0.9597**	0.9564	**0.9696**	**0.9696**
$SwLDA_{21}$	0.5714	0.5686	0.6816	0.6816	0.5619	0.5762	0.9700	0.9700	**0.9597**	0.9568	**0.9696**	**0.9696**
$SwLDA_{31}$	0.5886	0.5829	0.6776	0.6776	0.5524	0.5762	**0.9850**	**0.9850**	**0.9597**	0.9556	**0.9696**	0.9692
$SwLDA_{41}$	0.6500	0.6529	0.7020	0.6980	**0.5905**	0.5857	**0.9850**	**0.9850**	**0.9597**	0.9568	**0.9696**	**0.9696**
$SwLDA_{12}$	0.5800	0.5814	0.6816	0.6816	0.5667	0.5667	**0.9850**	**0.9850**	0.9589	0.9564	0.9692	0.9688
$SwLDA_{22}$	0.5800	0.5814	0.6816	0.6816	0.5667	0.5571	**0.9850**	**0.9850**	0.9589	0.9572	0.9684	0.9684
$SwLDA_{32}$	0.6243	0.6200	0.6776	0.6776	0.5286	0.5238	0.9600	0.9600	0.9589	0.9572	0.9684	0.9684
$SwLDA_{42}$	**0.6786**	0.6743	**0.7224**	0.7184	0.5476	0.5524	0.9450	0.9450	0.9593	0.9572	**0.9696**	0.9692

Table 2. Classification accuracy comparison

Dataset	BU	KANADE	JAFFE	ORL	YALE	AR
LDA	0.5729	0.6898	0.5571	0.9725	0.9593	0.9688
[10]	0.5743	0.6857	0.5714	0.9800	0.9564	0.9681
[11]	0.5957	0.6898	0.5381	0.9800	**0.9597**	0.9692
$SwLDA_{41}$	0.6500	0.7020	**0.5905**	**0.9850**	**0.9597**	**0.9696**
$SwLDA_{42}$	**0.6786**	**0.7224**	0.5476	0.9450	0.9593	**0.9696**

where $\mathbf{X}_c$ contains all samples in class c, $\hat{\boldsymbol{\mu}}_c$ is the new class representation or weighted center of class c. The third definition extends (21) to exploit the relationships between pairs of new class representation for each class, as follows:

$$\mathbf{S}_b^{(3)} = \sum_{c1=1}^{C} \sum_{c2=1}^{C} (\hat{\boldsymbol{\mu}}_{c_1} - \hat{\boldsymbol{\mu}}_{c_2})(\hat{\boldsymbol{\mu}}_{c_1} - \hat{\boldsymbol{\mu}}_{c_2})^T. \tag{22}$$

The last definition, $\mathbf{S}_b^{(4)}$, intends to maximize discrimination between every sample in one class with other new class representations, meanwhile takes into account of each sample's saliency scores, as follows:

$$\mathbf{S}_b^{(4)} = \sum_{c_1=1}^{C} \sum_{\substack{c_2=1,\\ c_2 \neq c_1}}^{C} \sum_{j=1}^{N_{c1}} p_{c_1,j} (\mathbf{x}_{c_1,j} - \hat{\boldsymbol{\mu}}_{c_2})(\mathbf{x}_{c_1,j} - \hat{\boldsymbol{\mu}}_{c_2})^T, \tag{23}$$

where N_{c_1} is the cardinality of class c_1.

3.3. Discriminant Criterion

Using the above described scatter matrices, several optimization criteria can be formed as follows:

$$J(\mathbf{W}) = \underset{\mathbf{W}}{argmax} \frac{tr(\mathbf{W}^T \mathbf{S}_b^{(i)} \mathbf{W})}{tr(\mathbf{W}^T \mathbf{S}_t^{(ij)} \mathbf{W})}, \tag{24}$$

where $\mathbf{S}_t^{(ij)} = \mathbf{S}_w^{(j)} + \mathbf{S}_b^{(i)}$, $i \in \{1, 2, 3, 4\}$ and $j \in \{1, 2\}$. After obtaining projection matrix $\mathbf{W}$ by eigenvalue decomposition, we map corresponding class representations and test samples by the optimal $\mathbf{W}$, and then nearest centroid classifier is applied for classification. It should be noted that when $\mathbf{H}$ or $\mathbf{S}_t$ are singular, a regularized version is used.

4. EXPERIMENT RESULTS

In our experiments, we evaluate the performance of proposed $SwLDA$, traditional LDA and two weighted LDA approaches mentioned in section 2 on six public facial image datasets: BU, KANADE, JAFFE, ORL, YALE and AR. We evaluate the performance of the proposed $SwLDA$ approaches, as illustrated in Table 1. The results of $SwLDA_{ij}$ illustrate classification accuracy obtained by using the matrices $\mathbf{S}_t^{(ij)}$ and $\mathbf{S}_b^{(i)}, i \in \{1, 2, 3, 4\}, j \in \{1, 2\}$. The result of traditional LDA is considered as baseline. The results comparison of baseline, Tang et al. [10], Jarchi and Boostani work [11] and our work are presented in Table 2. The best result in each dataset is presented in bold font. We implement standardization on all datasets before training and split each dataset into 5 folds for cross-validation. When obtaining $\mathbf{W}_c$, we select k-NN graphs with $k \in \min(5, 0.1 * N_c)$ or fully connected graphs to evaluate its impact on the results. As shown, the best performances over datasets BU and KANADE are both achieved by using $SwLDA_{42}$ with fully connected graphs. $SwLDA_{41}$ is the most effective over dataset JAFFE. The maximal improvement is 10.57% on dataset BU using $SwLDA_{42}$ with fully connected graphs, compared to the result of traditional LDA. That over Tang's work [10] is 0.14% and over Jarchi's work [11] is 2.28%. $SwLDA_{12}$ and $SwLDA_{22}$ work better than $SwLDA_{32}$ and $SwLDA_{42}$ apparently on datasets JAFFE and ORL. Fully connected graphs works better than k-NN graphs does over YALE dataset for all cases. Graph connection does not affect the classification accuracy using $SwLDA_{11}$, $SwLDA_{21}$, $SwLDA_{41}$, $SwLDA_{22}$ and $SwLDA_{32}$ over dataset AR.

5. CONCLUSION

In this paper, we propose weighted LDA variants based on a probabilistic definition of visual saliency estimation. We follow a class-specific saliency estimation process in order to determine the contribution of each sample in the optimization problems solved for discriminant subspace learning. Then, we employ our new approaches to six public datasets for evaluation and comparison with related LDA methods. Our new definitions target to reveal connections between each sample in every class, and further solve shortcomings in weighted LDA variants. Experimental results sufficiently demonstrate that the highest classification accuracy is always with one of our proposed approaches over these six facial image datasets.

6. REFERENCES

[1] A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, “Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis,” *Computer Vision and Image Understanding*, vol. 116, pp. 347–360, March 2012.

[2] Peter N. Belhumeur, João P. Hespanha, and David J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” *Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711–720, July 1997.

[3] Kamran Etemad and Rama Chellappa, “Discriminant analysis for recognition of human face images,” *Journal of the Optical Society of American A*, vol. 14, pp. 1724–1733, August 1997.

[4] A. Iosifidis, A. Tefas, and I. Pitas, “Activity-based person identification using fuzzy representation and discriminant learning,” *IEEE Transactions on Information Forensics and Security*, vol. 7, pp. 530–542, April 2012.

[5] B. Yu, L. Jin, and P. Chen, “A new LDA-based method for face recognition,” in *Proceedings 16th International Conference on Pattern Recognition*. IEEE, 2002, vol. 1, pp. 168–171.

[6] E. K. Tang, P. N. Suganthan, and X. Yao, “Generalized LDA using relevance weighting and evolution strategy,” in *Proceedings Congress on Evolutionary Computation*. IEEE, 2004, vol. 2, pp. 2230–2234.

[7] A. Iosifidis, A. Tefas, and I. Pitas, “On the optimal class representation in linear discriminant analysis,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, pp. 1491–1497, September 2013.

[8] A. Iosifidis, A. Tefas, and I. Pitas, “Kernel reference discriminant analysis,” *Pattern Recognition Letters*, vol. 49, pp. 85–91, November 2014.

[9] H. Ahmed, J. Mohamed, and Z. Noureddine, “Face recognition systems using relevance weighted two dimensional linear discriminant analysis algorithm,” *Signal and Information Processing*, vol. 3, pp. 130–135, November 2012.

[10] E. K. Tang, P. N. Suganthan, X. Yao, and A. K. Qin, “Linear dimensionality reduction using relevance weighted LDA,” *Pattern Recognition*, vol. 38, pp. 485–493, April 2005.

[11] D. Jarchi and R. Boostani, “A new weighted LDA method in comparison to some versions of lda,” *Proceedings of Word Academy of Science, Engineering and Technology*, vol. 12, pp. 233–238, 2006.

[12] M. Loog, R. Duin, and R. Haeb-Umbach, “Multiclass linear dimension reduction by weighted pairwise fisher criteria,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 762 – 766, July 2001.

[13] Z. Li, D. Lin, and X. Tang, “Nonparametric discriminant analysis for face recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 755–761, February 2009.

[14] C. Aytekin, S. Kiranyaz, M. Gabbouj, and A. Iosifidis, “Recent advances in salient object detection,” *FuturaBigData*, vol. 35, pp. 80–92, 2016.

[15] C. Aytekin, A. Iosifidis, and M. Gabbouj, “Probabilistic saliency estimation,” *Pattern Recognition*, vol. 74, pp. 359–372, September 2017.

PUBLICATION
II

Saliency-based multilabel linear discriminant analysis

L. Xu, J. Raitoharju, A. Iosifidis, and M. Gabbouj

IEEE Transactions on Cybernetics, vol. 52, Issue 10, October 2022, pp. 10 200–10 213
DOI: 10.1109/TCYB.2021.3069338

Saliency-Based Multilabel Linear Discriminant Analysis

Lei Xu, *Student Member, IEEE*, Jenni Raitoharju, *Member, IEEE*,
Alexandros Iosifidis, *Senior Member, IEEE*, and Moncef Gabbouj, *Fellow, IEEE*

***Abstract*—Linear discriminant analysis (LDA) is a classical statistical machine-learning method, which aims to find a linear data transformation increasing class discrimination in an optimal discriminant subspace. Traditional LDA sets assumptions related to the Gaussian class distributions and single-label data annotations. In this article, we propose a new variant of LDA to be used in multilabel classification tasks for dimensionality reduction on original data to enhance the subsequent performance of any multilabel classifier. A probabilistic class saliency estimation approach is introduced for computing saliency-based weights for all instances. We use the weights to redefine the between-class and within-class scatter matrices needed for calculating the projection matrix. We formulate six different variants of the proposed saliency-based multilabel LDA (SMLDA) based on different prior information on the importance of each instance for their class(es) extracted from labels and features. Our experiments show that the proposed SMLDA leads to performance improvements in various multilabel classification problems compared to several competing dimensionality reduction methods.**

***Index Terms*—Class saliency, dimensionality reduction, linear discriminant analysis (LDA), multilabel classification.**

I. Introduction

MULTILABEL classification tasks have become more and more common in the machine-learning field recently, for example, in text information categorization [1], image and video annotation [2], sequential data prediction [3], or music information retrieval [4]. Compared to single-label problems, the characteristics of multilabel problems are more complicated and unpredictable. In a single label problem, each instance merely belongs to a single class. In a multilabel dataset, data items can be associated with either one or several classes. For example, an image can represent both a beach and a sunset and, thus, be associated with both of these classes. Moreover, different classes typically contain a varying number of data items, leading to class-imbalanced problems [5]. Hence, in order to solve a multilabel classification problem efficiently and effectively, we need not only to consider the correlation of class labels and features of each data item but also to take into account the different cardinalities of the classes. The problem of multilabel learning (MLL) has been widely studied and various multilabel classifiers have been suggested [6]–[8].

In this article, we focus on dimensionality reduction for multilabel classification. Dimensionality reduction techniques in general aim at transforming the data to a lower dimensional form that is easier to process by the learning techniques without losing relevant information. The dimensionality reduction techniques for multilabel classification aim at optimizing the data transformation for subsequent multilabel classification. At least 50 such methods have been proposed [9].

A well-known supervised dimensionality reduction technique linear discriminant analysis (LDA) and its variants have been widely used to extract discriminant data representations for solving various problems, for example, in human action recognition [10] or biological data classification [11]. However, they are not optimal for multilabel problems due to the characteristics of multilabel data. This is due to two factors: 1) the contribution of each data item in the calculation of the scatter matrices involved in the optimization problem of single-label LDA and its variants cannot be appropriately determined and 2) the cardinality of the various classes forming the multilabel problem can be quite imbalanced. In multilabel LDA (MLDA) [12] and its variants, these problems have been tackled by introducing different weights to take into account the label and/or feature correlation of different items.

In this article, we propose a novel dimensionality reduction method for multilabel classification based on a probabilistic approach that is able to estimate the contribution of each data item to the classes it is associated with by taking into account prior information encoded using various types of metrics. The proposed calculation of the contribution of each data item to the classes it belongs to can not only weigh its importance but can also avoid problems related to imbalanced classes. To this end, we exploit the concept of class saliency introduced in [13]. Hence, the proposed method is called saliency-based MLDA (SMLDA). Our proposed SMLDA approach exploits

Manuscript received 8 August 2020; revised 8 January 2021; accepted 19 March 2021. Date of publication 20 April 2021; date of current version 16 September 2022. This work was supported by the NSF-Business Finland Center for Visual and Decision Informatics (CVDI) Project AMALIA and EUREKA ITEA 3 Project Mad@Work. The work of Jenni Raitoharju was supported by the Academy of Finland Project under Grant 324475. This article was recommended by Associate Editor B. Ribeiro. *(Corresponding author: Lei Xu.)*

Lei Xu and Moncef Gabbouj are with the Department of Computing Sciences, Tampere University, 33100 Tampere, Finland (e-mail: lei.xu@tuni.fi; moncef.gabbouj@tuni.fi).

Jenni Raitoharju is with the Programme for Environmental Information, Finnish Environment Institute, 40500 Jyväskylä, Finland (e-mail: jenni.raitoharju@syke.fi).

Alexandros Iosifidis is with the Department of Electrical and Computer Engineering, Aarhus University, 8000 Aarhus, Denmark (e-mail: ai@ece.au.dk).

This article has supplementary downloadable material available at https://doi.org/10.1109/TCYB.2021.3069338, provided by the authors.

Digital Object Identifier 10.1109/TCYB.2021.3069338

both label and feature information with various prior weighting factors. The proposed method yields features optimized for multilabel classification that can be subsequently classified using any multilabel classifier.

We have made the following contributions on dimensionality reduction for multilabel classification tasks with our novel SMLDA approach.

1) We propose a general framework for using the probabilistic saliency estimation to weigh the importance of each data item for the classes it is associated with for the first time in MLL.
2) We formulate a novel SMLDA method that uses the saliency-based weights in the scatter matrices and can alleviate the problems related to imbalanced datasets.
3) We integrate different label and feature information previously used as weights in dimensionality reduction to SMLDA by using them as prior information for probabilistic saliency estimation and show experimentally that our approach leads to a better performance.
4) We compare our proposed approach to 11 competing dimensionality reduction methods on 17 diverse multilabel datasets using seven evaluation metrics and applying two different multilabel classifiers on the produced features, and the results show considerable improvements in multilabel classification tasks using our approach.

The remainder of this article is structured as follows. In Section II, we briefly review the related works. We include a precise explanation of the LDA and weighted MLDA with adequate mathematical notations to support the derivations of the proposed method. In Section III, we describe our proposed methods in detail. Section IV presents the experimental setup and results. In Section V, we conclude this article and discuss the potential future studies.

II. Related Works

In this section, we first briefly present several dimensionality reduction techniques previously used in multilabel classification tasks in Section II-A. In Section II-B, we give a detailed description of the standard LDA, weighted LDA, and MLDA, since they form the theoretical foundation for the proposed work. Subsequently, we introduce the general concepts of saliency estimation and the probabilistic saliency estimation approach needed to develop the proposed method.

A. Dimensionality Reduction Methods for Multilabel Classification

Dimensionality reduction techniques are commonly used as a preprocessing step for multilabel classification to map the raw high-dimensional data into an optimal lower-dimensional subspace preserving the distinguishing features [9]. The techniques can be categorized as unsupervised or supervised approaches depending on whether class label information is used or not [14]. Furthermore, the techniques can be divided into methods that are independent of the classifiers or dependent of the classifiers [9]. In this article, we consider only dimensionality reduction techniques that are all independent of the classifiers.

Principal component analysis (PCA) [15] is the most well-known unsupervised dimensionality reduction algorithm that minimizes the information lost by preserving as much of the data's variations as possible. Canonical correlation analysis (CCA) [16] is a widely known supervised dimensionality reduction algorithm, projecting the raw data into a subspace exploiting the correlations between the features and labels.

Dimensionality reduction techniques specifically designed for multilabel data include the multilabel-informed latent semantic indexing (MLSI) algorithm [17] that preserves the discriminate feature information by considering the correlations between the multiple labels and multilabel dimensionality reduction via the dependence maximization (MDDM) algorithm [18] that maximizes the dependence between the original features and class labels using the Hilbert–Schmidt independence criterion (HSIC) for measuring dependence. MDDM has two variants with different constraints: 1) MDDMp with an uncorrelated projection constraint and 2) MDDMf with an uncorrelated feature constraint. MDDMp variant was observed to perform better in [18]. Xu *et al.* [19] proposed a multilabel feature extraction method that integrates least-squares formulations of PCA and MDDM linearly, which both maximizes feature variance and maximizes feature-label dependence (MVMD) at the same time.

B. Linear Discrimination Analysis-Based Algorithms for Multilabel Classification

Standard LDA and its variants have been applied to tackle various multilabel classification problems [9], [12], [20]–[23]. These methods operate on N data items $\mathbf{x}_i \in \mathbb{R}^D$ and their corresponding binary label vectors $\mathbf{y}_i \in \{0, 1\}^C$, where D is the original data dimensionality and C is the number of classes. These are arranged into matrices $\mathbf{X} \in \mathbb{R}^{D\times N}$ and $\mathbf{Y} \in \mathbb{R}^{C\times N}$. An element y_{ci} of $\mathbf{Y}$ is 1 only if the corresponding data item $\mathbf{x}_i$ is associated with class c. Thus, in single-label classification tasks, there is a single 1 on each column, but in multilabel classification, the number of 1s is not constrained. The rows of $\mathbf{Y}$ contain 1s for all data items that are associated with the particular class and we denote them as $\mathbf{y}_{(j)}$, where $j \in 1, \ldots, C$. The objective of LDA-based methods is to find a data projection matrix $\mathbf{W} \in \mathbb{R}^{D\times d}$ that maps the data from the original feature space $\mathbb{R}^D$ to a subspace $\mathbb{R}^d$, where $D > d$, in a manner that maximizes the class discrimination.

1) Linear Discrimination Analysis: LDA is an effective technique to reduce the dimensionality of original data as a prepossessing step for single-label classification problems. LDA operates on within-class, between-class, and total scatter matrices $\mathbf{S}_w$, $\mathbf{S}_b$, and $\mathbf{S}_t$ defined as follows:

$$\mathbf{S}_w = \sum_{c=1}^{C}\sum_{i=1}^{N} y_{ci}\big(\mathbf{x}_i - \boldsymbol{\mu}_c\big)\big(\mathbf{x}_i - \boldsymbol{\mu}_c\big)^T \tag{1}$$

$$\mathbf{S}_b = \sum_{c=1}^{C}\left(\sum_{i=1}^{N} y_{ci}\right)\big(\boldsymbol{\mu}_c - \boldsymbol{\mu}\big)\big(\boldsymbol{\mu}_c - \boldsymbol{\mu}\big)^T \tag{2}$$

$$\mathbf{S}_t = \sum_{c=1}^{C}\sum_{i=1}^{N} y_{ci}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T. \tag{3}$$

Here, $\boldsymbol{\mu}_c$ denotes the mean vector of class c as

$$\boldsymbol{\mu}_c = \frac{1}{N_c}\sum_{i=1}^{N} y_{ci}\mathbf{x}_i \tag{4}$$

where $N_c = \sum_{i=1}^{N} y_{ci}$ is the cardinality of class c. The total mean vector $\boldsymbol{\mu}$ is computed as

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i. \tag{5}$$

The optimal projection matrix $\mathbf{W}$ is learned by maximizing Fisher's discriminant criterion [24] that minimizes the within-class scatter while maximizing the between-class scatter

$$J(\mathbf{W}) = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{\operatorname{tr}\left(\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right)}{\operatorname{tr}\left(\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right)} \tag{6}$$

where tr(.) denotes the trace of a matrix. Typically, the solution to this trace ratio optimization is approximated by solving the corresponding ratio trace optimization. This allows obtaining the projection matrix $\mathbf{W}$ by solving the generalized eigenvalue problem

$$\mathbf{S}_b\mathbf{w} = \mathbf{S}_w\lambda\mathbf{w} \tag{7}$$

and taking the eigenvectors corresponding to the $d \leq C-1$ largest eigenvalues as columns of the projection matrix $\mathbf{W}$. The rank of $\mathbf{S}_b$ is equal to $C-1$, which is the maximal dimensionality of the resulting subspace. Also, different iterative methods for solving directly the trace ratio problem have been proposed [25], [26].

Since $\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b$, an alternative approach is to use $\mathbf{S}_t$ instead of $\mathbf{S}_w$ and maximize Fisher's discriminant criterion as

$$J(\mathbf{W}) = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{\operatorname{tr}\left(\mathbf{W}^T\mathbf{S}_b\mathbf{W}\right)}{\operatorname{tr}\left(\mathbf{W}^T\mathbf{S}_t\mathbf{W}\right)}. \tag{8}$$

Finally, the optimized features can be obtained as

$$\mathbf{Z} = \mathbf{W}^T\mathbf{X}. \tag{9}$$

The datasets used in most traditional LDA classification tasks are assumed to have equal class distribution as a homoscedastic Gaussian model [27], in which the covariance matrices of each class should be identical [28]. The performance is affected severely due to the imbalance of input datasets [29].

2) Weighted Linear Discrimination Analysis: In order to enhance the robustness of traditional LDA on different kinds of datasets, various weight factors based on class statistics [26], [28], [30], for example, class cardinality, a prior probability, have been introduced into the definitions of the scatter matrices to balance the contribution of each class. Weighted LDA approaches have diminished the influence of outlier classes on the scatter matrices of imbalanced datasets to some extent; however, they still neglect the varying importance of individual samples in the class description. Saliency-based weighted LDA (SwLDA) [13] was proposed to explore the contribution of each instance based on probabilistic saliency estimation [31]. Our work uses a similar idea for multilabel classification.

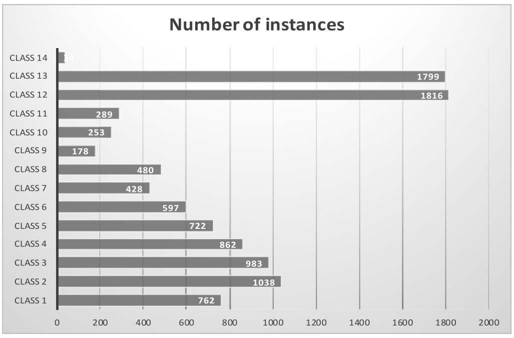


Fig. 1. Number of instances for each class in the Yeast database.

3) Multilabel Linear Discrimination Analysis: Although weighted LDA algorithms enhance the performance in single-label classification tasks [32] compared to traditional LDA, such variants are still not directly applicable for multilabel classification tasks [12]. In a multilabel dataset, label information contains certain correlations or dependencies [33], for example, an image instance labeled as "car" highly correlates to label "road" [12]. Besides, it is quite common that the number of samples in each class in a multiclass dataset is imbalanced. For example, the largest class size is 1128 and the smallest is 21 in the widely used Yeast database [34], as shown in Fig. 1. Due to the specific characteristics of multilabel databases, it is imperative to take into account the correlation of class labels and/or discriminative feature information of each instance to tackle the suboptimal classification result on imbalanced datasets.

If traditional LDA and its variants are applied to multilabel classification tasks by simply using (1) and (2) with the multilabel label matrix $\mathbf{Y}$, an overcounting problem is encountered, that is, the contribution of one instance can be repeatedly counted in computing the scatter matrices. Hence, an MLDA [12] and its variants use weight factors to express redundancy or/and correlation information so that the scatter matrices can be calculated without redundancy on multilabel databases. These weight factors can be organized to a non-negative weight matrix $\mathbf{M} \in \mathbb{R}^{C\times N}$ with the same size as the label matrix $\mathbf{Y}$

$$\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_i, \ldots, \mathbf{m}_N] = \left[\mathbf{m}_{(1)}, \ldots, \mathbf{m}_{(j)}, \ldots, \mathbf{m}_{(C)}\right]^T \tag{10}$$

where $\mathbf{m}_i$ represents a weight vector for the ith instance, $\mathbf{m}_{(j)}$ is a weight vector for the jth class, and m_{ci} is the weight factor of the ith instance for class c.

We denote by n_i, $n_{(c)}$, and n the summations of the weights for the ith instance, weights for class c, and all weights, respectively

$$n_i = \sum_{c=1}^{C} m_{ci} \tag{11}$$

$$n_{(c)} = \sum_{i=1}^{N} m_{ci} \tag{12}$$

$$n = \sum_{c=1}^{C}\sum_{i=1}^{N} m_{ci} = \sum_{c=1}^{C} n_{(c)}. \tag{13}$$

We also define row vectors $\hat{\mathbf{n}}$ and $\hat{\mathbf{m}}$ and matrix $\hat{\mathbf{M}}$ for simplifying notations as

$$\hat{\mathbf{n}} = \left[\frac{1}{n_{(1)}}, \ldots, \frac{1}{n_{(C)}}\right]. \tag{14}$$

$$\hat{\mathbf{m}} = [n_1, \ldots, n_i, \ldots, n_N] = \sum_{c=1}^{C} \mathbf{m}_{(c)} \tag{15}$$

$$\hat{\mathbf{M}} = \mathbf{M}\mathrm{diag}\Big(\hat{\mathbf{n}}^{\frac{1}{2}}\Big) \tag{16}$$

where $\hat{\mathbf{M}}$ has row vectors $([\mathbf{m}_{(c)}]/[\sqrt{n_{(c)}}])$ for $c = 1, \ldots, C$.

The scatter matrices for MLDA can now be given as

$$\begin{aligned}\mathbf{S}_w &= \sum_{c=1}^{C}\sum_{i=1}^{N} m_{ci}\big(\mathbf{x}_i - \boldsymbol{\mu}_c\big)\big(\mathbf{x}_i - \boldsymbol{\mu}_c\big)^T \\ &= \mathbf{X}\Big(\mathrm{diag}\big(\hat{\mathbf{m}}\big) - \hat{\mathbf{M}}^{\mathsf{T}}\hat{\mathbf{M}}\Big)\mathbf{X}^{\mathsf{T}} \end{aligned} \tag{17}$$

$$\begin{aligned}\mathbf{S}_b &= \sum_{c=1}^{C}\left(\sum_{i=1}^{N} m_{ci}\right)\big(\boldsymbol{\mu} - \boldsymbol{\mu}_c\big)\big(\boldsymbol{\mu} - \boldsymbol{\mu}_c\big)^T \\ &= \mathbf{X}\left(\hat{\mathbf{M}}^{\mathsf{T}}\hat{\mathbf{M}} - \frac{1}{n}\hat{\mathbf{m}}^{\mathsf{T}}\hat{\mathbf{m}}\right)\mathbf{X}^{\mathsf{T}} \end{aligned} \tag{18}$$

$$\begin{aligned}\mathbf{S}_t &= \mathbf{S}_w + \mathbf{S}_b \\ &= \mathbf{X}\left(\mathrm{diag}\big(\hat{\mathbf{m}}\big) - \frac{1}{n}\hat{\mathbf{m}}^{\mathsf{T}}\hat{\mathbf{m}}\right)\mathbf{X}^{\mathsf{T}} \end{aligned} \tag{19}$$

where $\boldsymbol{\mu}$ is the total mean vector of all training instances and $\boldsymbol{\mu}_c$ is the mean vector of class c

$$\boldsymbol{\mu} = \frac{\sum_{c=1}^{C}\sum_{i=1}^{N} m_{ci}\mathbf{x}_i}{\sum_{c=1}^{C}\sum_{i=1}^{N} m_{ci}}, \quad \boldsymbol{\mu}_c = \frac{\sum_{i=1}^{N} m_{ci}\mathbf{x}_i}{\sum_{i=1}^{N} m_{ci}}. \tag{20}$$

A detailed derivation of the matrix forms in (17)–(19) can be found in [14]. The optimal projection matrix $\mathbf{W}$ can still be obtained by solving the generalized eigenproblem in (7) as discussed in Section II.

In the original MLDA [12], the weight factors are solved using label correlations for different classes. First, a correlation matrix $\mathbf{R} \in \mathbb{R}^{C\times C}$ is computed using the class labels of each pair of classes

$$R_{kl} = \cos\big(\mathbf{y}_{(k)}, \mathbf{y}_{(l)}\big) = \frac{\mathbf{y}_{(k)}^T\mathbf{y}_{(l)}}{\|\mathbf{y}_{(k)}\|\,\|\mathbf{y}_{(l)}\|} \tag{21}$$

where $\mathbf{y}_{(k)}$, $\mathbf{y}_{(l)}$ are label vectors for classes $k, l \in 1, \ldots, C$. The label correlation for classes k and l is high if the classes are closely related. The correlation matrix $\mathbf{R}$ can be used to compute the weight matrix $\mathbf{M}$ as $\mathbf{M} = \mathbf{RY}$. However, also this approach may lead to the overcounting problem. To tackle the overcounting problem [12], the weight factors are normalized with the ℓ_1-norm

$$\mathbf{m}'_i = \frac{\mathbf{m}_i}{\|\mathbf{y}_i\|_{\ell_1}}. \tag{22}$$

Other metrics for evaluating the relationships among instances from the labels and/or features were used for determining the weights in [14] under the name weighted multilabel LDA (wMLDA). In addition to the label correlation-based weight factors used in MLDA [12], Xu [14] considered entropy-based [35], binary-based [20], fuzzy-based [36], and dependence-based weight factors [14]. Similar metrics can be used as prior information within our probabilistic saliency estimation framework. Therefore, the detailed explanations of these metrics are left to Section III-A1.

In [21], MLDA was extended to Direct MLDA by changing the definition of $\mathbf{S}_b$ in a way that allows obtaining a higher dimensional subspace than the original MLDA, where the subspace dimensionality is limited by the rank of $\mathbf{S}_b$ to $C-1$. This extension work further enhanced the results in multilabel video classification tasks. Another extension, multilabel discriminant analysis with locality consistency (MLDA-LC) [22] not only preserves the global class label information as MLDA does but also incorporates a graph regularized term to utilize the local geometric information. MLDA-LC reveals the similarity among nearby instances with transformation in the projection space using incorporation of the graph Laplacian matrix into the MLDA approach, which further enhances the classification performance in multilabel datasets compared to MLDA and MLLS algorithms.

C. Saliency Estimation

Saliency estimation, as a standard computer vision task, is inspired by neurobiological studies [37] and cognition psychology [38]. Generally, saliency estimation is a preprocessing step for various high-level computer vision tasks, such as object detection [31], [39] and omni directional images [40]. Saliency in physiological science is defined as a special kind of perception of the human visual system, by which humans can perceive particular parts in a scene in details due to colors, textures, or other prominent information contained in these parts [41]. These particular parts can be distinguished as a foreground from nonsalient background parts.

Computational saliency estimation approaches can be categorized as local approaches and global approaches based on the way they process saliency information [41]. Local saliency estimation approaches explore the prominent information around the neighborhood of specific pixels/regions whilst global approaches exploit the rarity of a pixel/patch/region in the entire scene. Since the emergence of the computational saliency estimation field [42], various probabilistic approaches have been explored in this topic.

Aytekin *et al.* [31] proposed a probabilistic saliency estimation approach for segmenting salient objects in an image, where a probability mass function $P(\mathbf{x})$ depicts whether a region $\mathbf{x}_i$ (pixel, super-pixel, or patch) in an image is considered as a distinct region. The higher the values of $P(\mathbf{x}_i)$ for a region, the more prominent the region is. $P(\mathbf{x})$ is solved by simultaneously optimizing two terms to allocate not only lower probabilities to nonsalient regions but also similar probabilities to similar regions

$$\begin{aligned} &\underset{P(x)}{\mathrm{argmin}}\left(\sum_i P(\mathbf{x}_i)^2 v_i + \sum_{i,j}\big(P(\mathbf{x}_i) - P(\mathbf{x}_j)\big)^2 a_{ij}\right) \\ &\quad = \underset{P(x)}{\mathrm{argmin}}\left(\sum_i P(\mathbf{x}_i)^2 v_i + \sum_{i,j}\Big(P(\mathbf{x}_i)^2 - P(\mathbf{x}_i)P(\mathbf{x}_j)\Big)a_{ij}\right) \\ &\text{s.t.}\ \sum_i P(\mathbf{x}_i) = 1 \end{aligned} \tag{23}$$

where the first term suppresses the probability of a nonprominent region $\mathbf{x}_i$ using its prior information $v_i \geq 0$. In the second term, a high similarity of regions $\mathbf{x}_i$ and $\mathbf{x}_j$, given as a high similarity value a_{ij}, forces the regions to have similar probabilities. To go from the first form to the second form, the similarity values are assumed symmetric, that is, $a_{ij} = a_{ji}$.

The optimization task in (23) can be expressed using matrix notations as

$$\begin{aligned} \mathbf{p}^* = \underset{\mathbf{p}}{\operatorname{argmin}} \ & \left(\mathbf{p^T H p}\right) \\ & \mathbf{H} = \mathbf{D} - \mathbf{A} + \mathbf{V} \\ \text{s.t.} \ & \mathbf{p}^T \mathbf{1} = 1 \end{aligned} \tag{24}$$

where $\mathbf{p}$ is a probability vector that contains the probabilities of each element or region $\mathbf{x}_i$ to be salient, that is, $p_i = P(\mathbf{x}_i)$, $\mathbf{A}$ is an affinity matrix, which denotes the similarity of each pair of regions $\mathbf{x}_i$ and $\mathbf{x}_j$ as $[\mathbf{A}]_{ij} = a_{ij}$. $\mathbf{D}$ is a diagonal matrix having elements equal to $[\mathbf{D}_{ii}] = \sum_j a_{ij}$, $\mathbf{V}$ is a diagonal prior information matrix having elements $[\mathbf{V}]_{ii} = v_i$, and $\mathbf{1}$ is a vector of ones. Then, the Lagrangian multiplier method is employed

$$\mathcal{L}(\mathbf{p}, \gamma) = \left(\mathbf{p^T H p}\right) - \gamma\left(\mathbf{p^T 1} - 1\right). \tag{25}$$

A global optimum $\mathbf{p}^*$ is obtained by setting the partial derivative of (25) with the respect $\mathbf{p}$ to 0. The final optimized probability vector is

$$\mathbf{p}^*_{pse} = \frac{1}{\mathbf{1}^T \mathbf{H}^{-1} \mathbf{1}} \mathbf{H}^{-1} \mathbf{1} \tag{26}$$

where the normalization constant $\mathbf{1}^T \mathbf{H}^{-1} \mathbf{1}$ follows from the constraint $\mathbf{p}^T \mathbf{1} = 1$ and ensures that the resulting values are actual probabilities. Due to the properties of matrix $\mathbf{H}^{-1}$, the elements of $\mathbf{p}^*$ are always non-negative as shown in [31]. A more detailed derivation of (26) can be also found in [31].

III. Proposed Method

We propose a novel SwLDA method for multilabel classification tasks. The proposed method has two main steps. For the first step, we propose a probabilistic saliency estimation approach to evaluate the importance of each sample for each class in a multilabel dataset. This is a general framework for multilabel class-saliency and, as future work, can be easily integrated also with other dimensionality reduction techniques or directly with multilabel classifiers that weigh the samples based on their importance. In the second step, we use the class-saliency analysis as weights in an MLDA technique.

In our prior work [13], we used the idea of probabilistic class-saliency estimation for single-label datasets to tackle the suboptimal results of LDA-based algorithms caused by imbalanced datasets or/and outliers. In this article, we formulate multilabel extensions of both the probabilistic class-saliency estimation and the subsequent LDA-based dimensionality reduction technique. Furthermore, we show how to use as prior information in the probabilistic multilabel class-saliency estimation framework different types of information extracted from the data and/or labels that have been previously used directly as sample weights in MLL and we propose a new misclassification-based multilabel information extraction approach, which is based on the prior information type used for single-label data in [13].

A. Probabilistic Multilabel Class-Saliency Estimation

The goal of probabilistic multilabel class-saliency estimation is to define the probability of each data item to be salient for each class. In other words, we want to find a probability matrix $\mathbf{P} \in \mathbb{R}^{C \times N}$

$$\mathbf{P} = \left[\mathbf{p}_1, \ldots, \mathbf{p}_i, \ldots, \mathbf{p}_N\right] = \left[\mathbf{p}_{(1)}, \ldots, \mathbf{p}_{(j)}, \ldots, \mathbf{p}_{(C)}\right]^T \tag{27}$$

where $\mathbf{p}_i \in \mathbb{R}^C$ is a vector containing the probabilities for instance i to be salient for class c and $\mathbf{p}_{(j)} \in \mathbb{R}^N$ is the probability vector for the jth class. The probabilities for each class are normalized to sum up to one, that is, $\sum_{i=1}^N p_{ci} = 1 \ \forall c \in 1, \ldots C$.

First, we make an assumption that only data items associated with a class can be salient, that is, $p_{ci} = 0$ if $y_{ci} = 0$. As we need to solve the probabilities p_{ci} only for data items associated with class c, we form a reduced data matrix $\mathbf{X}^c \in \mathbb{R}^{D \times N^c}$ and reduced probability vector $\mathbf{p}^c \in \mathbb{R}^{N^c}$ corresponding to N^c data items associated with class c. Now, we can write the optimization problem of probabilistic multilabel class-saliency estimation as

$$\begin{aligned} & \underset{\mathbf{p}^c}{\operatorname{argmin}} \left(\sum_i^{N^c} (p_i^c)^2 v_i^c + \frac{1}{2} \sum_i^{N^c} \sum_j^{N^c} \left(p_i^c - p_j^c\right)^2 a_{ij}^c \right) \\ & \quad = \underset{\mathbf{p}^c}{\operatorname{argmin}} \left(\sum_i^{N^c} (p_i^c)^2 v_i^c + \frac{1}{2} \sum_i^{N^c} \sum_j^{N^c} \left((p_i^c)^2 a_{ij}^c + (p_j^c)^2 a_{ij}^c \right) \right. \\ & \qquad \left. - \sum_i^{N^c} \sum_j^{N^c} \left(p_i^c p_j^c\right) a_{ij}^c \right) \\ & \text{s.t.} \ \sum_i^{N^c} p_i^c = 1 \end{aligned} \tag{28}$$

where p_i^c is the ith element in $\mathbf{p}^c$ and $v_i^c \geq 0$ is the corresponding prior information to suppress the probabilities of nonsalient instances from class c. The similarity value a_{ij}^c forces the instances $\mathbf{x}_i^c$ and $\mathbf{x}_j^c$ have similar probabilities, if they are similar. Unlike the original probabilistic saliency estimation in (23), we do not require the similarity values to be symmetric.

Equation (28) can be expressed in matrix notation as

$$\begin{aligned} \mathbf{p^{c*}} = \underset{\mathbf{p}^c}{\operatorname{argmin}} \ & \left(\mathbf{p}^{c^T} \mathbf{H}^c \mathbf{p}^c\right) \\ & \mathbf{H}^c = \frac{1}{2}\mathbf{D_1}^c + \frac{1}{2}\mathbf{D_2}^c - \mathbf{A}^c + \mathbf{V}^c \\ \text{s.t.} \ & \mathbf{p}^{c^T} \mathbf{1} = 1 \end{aligned} \tag{29}$$

where $\mathbf{A}^c$ is an affinity matrix of the items associated with class c with $[\mathbf{A}^c]_{ij} = a_{ij}^c$ expressing the similarity of the ith and jth class items, the diagonal matrix $\mathbf{D_1}^c$ can be then computed as $[\mathbf{D_1}^c]_{ii} = \sum_j [\mathbf{A}_c]_{ij}$ and $\mathbf{D_2}^c$ can be then computed as $[\mathbf{D_2}^c]_{ii} = \sum_j [\mathbf{A}_c]_{ji}$, that is, $\mathbf{D_1}^c$ has summations over rows while $\mathbf{D_2}^c$ has summations over columns. $\mathbf{V}^c$ is a diagonal prior information matrix having elements $[\mathbf{V}^c]_{ii} = v_i^c$.

In this work, the compute the affinity matrix $\mathbf{A}^c \in \mathbb{R}^{N_c \times N_c}$ with the RBF kernel function as

$$\left[\mathbf{A}^c\right]_{ij} = \exp\left(-\frac{\left\|\mathbf{x}_i^c - \mathbf{x}_j^c\right\|^2}{2\sigma^2}\right) \tag{30}$$

where $\mathbf{x}_i^c$ and $\mathbf{x}_j^c$ are the ith and jth instance in class c and σ is a hyper parameter. While (30) is sensitive to the parameter σ, we follow a common approach of setting its values to the mean distance value between the training samples. The affinity matrix could be also replaced by sparse variants, for example, by forming a kNN graph and keeping only the values for k nearest neighbors or by using an affinity matrix proposed in [43], where the sensitive parameter σ is avoided.

$\mathbf{V}^c \in \mathbb{R}^{N^c \times N^c}$ is a diagonal matrix, which carries the prior information on whether each instance in class c is salient for the class. The values of $\mathbf{V}^c$ are higher for samples, which are expected to *not* be salient, that is, the lower a value $[\mathbf{V}^c]_{ii}$, the more prominent the corresponding ith instance is expected to be. Values for $[\mathbf{V}^c]_{ii} = v_i^c \; \forall i \in 1, \ldots N_c$ can be estimated from different prior information. For example, a data item that belongs to all the classes it is unlikely to be salient for any particular class or if an item is very different from other samples in a class it is unlikely to be salient for that class. It should be noted that while we set prior information values v_i^c only for items associated with class c, we can exploit the information extracted from other data items while setting the values of $\mathbf{V}^c$. For example, items having a high similarity with many items not associated with the class could be considered less likely to be prominent. We introduce six different approaches to set the values of $\mathbf{V}^c$ in Section III-A1.

After computing the matrices $\mathbf{A}^c$ and $\mathbf{V}^c$, the probability vector $\mathbf{p^{c*}}$ can be solved as

$$\mathbf{p}^{c*} = \frac{1}{\mathbf{1}^T\mathbf{H}^{c^{-1}}\mathbf{1}}\mathbf{H^{c^{-1}}1}. \tag{31}$$

In order to avoid singularity during this process, a regularized version of $\mathbf{H}^c$ with a small value ϵ added to the diagonal elements if $\mathbf{H}^c$ is rank-deficient.

As the probability vector $\mathbf{p}^{c*}$ obtained by solving (31) has only N^c elements, but we want to form a probability matrix $\mathbf{P} \in \mathbb{R}^{C \times N}$ shown in (27), we need to put the values $\mathbf{p}^{c*}$ to the correct places in $\mathbf{P}$. If the ith item in class c is the jth item in the entire dataset, this can be done by setting $[\mathbf{P}]_{cj} = p_i^c$ for all items in class c. To obtain full matrix $\mathbf{P}$, the above-described process is repeated for each class $c \in \{1, \ldots, C\}$. The sum of the values for each row in $\mathbf{P}$ is one, which is expected to alleviate the overcounting problem.

1) Prior Information Types: Probabilistic saliency estimation [31] was originally proposed for segmenting salient parts from images. In this setup, the prior information used was that the pixel at the image borders is typically nonsalient. The prior information value v_i was set to 1 for any border pixels and to 0 for all the others. In multilabel class-saliency estimation, we similarly want to use v_i^c to integrate our prior knowledge on which data items are likely to be salient for class c. To this end, we propose a novel information type for MLL context: misclassification-based prior information. Furthermore, we introduce five prior information types based on weight factors proposed for MLDA and wMLDA. Our experimental results show that using these information types as prior information for our proposed saliency estimation framework instead of using them directly as weight factors consistently leads to better results.

Correlation-based prior information (SMLDAc) was used as weight factors in the original MLDA algorithm [12]. As in [12], we first compute the label correlation matrix $\mathbf{R}$ defined in (21). We then compute the normalized weight vector $\mathbf{m'}_j \in \mathbb{R}^C$ using (22) for all data items and set our prior information matrix values as

$$\left[\mathbf{V}^c\right]_{ii} = 1 - m'_{cj} \tag{32}$$

where item j of the full dataset is the ith item associated with class c. Label correlation information is widely exploited to tackle the redundancy of label information in multilabel tasks [12], [44], but it can lead to a suboptimal result due to nonzero values in the correlation weight factor matrix for irrelevant labels [14]. As we pick only the values for data items associated with class c, the problem of unwanted nonzero values can be avoided.

Binary-based prior information (SMLDAb) utilizes the label information as in [20]. In our formulation, this approach reduces to having an equal value in $\mathbf{V}^c$ for all instances as only instances belonging to class c are considered in $\mathbf{V}^c$. For wMLDA, such direct use of class labels leads to an overcounting problem in the scatter matrices. In our formulation, this problem is avoided because $\mathbf{V}^c$ merely represents the prior information for class saliency estimation and the final weight matrix $\mathbf{P}$ is normalized for each class.

Entropy-based prior information (SMLDAe) assumes that data items, which are associated with more classes are less salient for any class as in [14] and [35]. We use this assumption as our prior information as

$$\left[\mathbf{V}^c\right]_{ii} = 1 - \frac{1}{\left\|\mathbf{y}_i^c\right\|_{\ell_1}} \tag{33}$$

where $\mathbf{y}_i^c$ is the label vector of the ith sample associated with class c and, thus, $\|\mathbf{y}_i^c\|_{\ell_1}$ is the total number of classes the item is associated with.

Fuzzy-based prior information (SMLDAf) uses a supervised version of fuzzy C-means clustering algorithm (SFCM) as in [14] and [36] to learn the membership degree of each item in each class. We use the membership directly as our prior information as

$$\left[\mathbf{V}^c\right]_{ii} = 1 - g_j^c \tag{34}$$

where g_j^c is the membership degree of item j in class j and item j is the ith item associated with class c.

Dependence-based prior information (SMLDAd) uses HSIC [45], which is used to describe statistical dependence between features and labels based on the estimation of the Hilbert–Schmidt norms. To maximize HSIC, we follow an iterative algorithm described in [14].

This approach transforms a multilabel task to several single-label tasks. It allocates 1 to only one prominent class for each item after the final iteration. In our probabilistic formulation, we set

$$\left[\mathbf{V}^c\right]_{ii} = 1 - h_j^c \tag{35}$$

where h_j^c is 1 if item j has been assigned to class c and 0 otherwise and item j is the ith item associated with class c.

Misclassification-based prior information (SMLDAm) is similar to the prior information used in [13] for single-label data to alleviate the suboptimal result in LDA arising from outlier items on imbalanced datasets

$$[\mathbf{V}^c]_{ii} = \begin{cases} 0, & \text{if } d_{ic}^c < \min\limits_{k \neq c} d_{ic}^k \\ \dfrac{d_{ic}^c}{\min\limits_{k \neq c} d_{ic}^k}, & \text{otherwise} \end{cases} \tag{36}$$

where $d_{ic}^k = \|\mathbf{x}_{ic} - \boldsymbol{\mu}_k\|_2^2$, $\mathbf{x}_{ic}$ is the ith instance of class c, and $\boldsymbol{\mu}_k$ is the mean vector of class k. Using this prior information type, a sample that is closer to another class is considered less salient for class c even if it is relatively close to the center of class c. Note that when computing this prior information matrix, we consider the full data $\mathbf{X}$ and not only the data items in $\mathbf{X^c}$ for which we are defining the prior information values.

B. Saliency-Based Multilabel Linear Discriminant Analysis

After forming the probability matrix $\mathbf{P}$ using the proposed probabilistic multilabel class-saliency estimation, we use the probabilities directly as weights for our MLDA. We compute the scatter matrices $\mathbf{S}_w$ and $\mathbf{S}_b$ as

$$\mathbf{S}_w = \mathbf{X}\big(\text{diag}(\hat{\mathbf{p}}) - \mathbf{P}^\intercal\mathbf{P}\big)\mathbf{X}^\intercal \tag{37}$$

$$\mathbf{S}_b = \mathbf{X}\left(\mathbf{P}^\intercal\,\mathbf{P} - \frac{1}{n}\hat{\mathbf{p}}^\intercal\hat{\mathbf{p}}\right)\mathbf{X}^\intercal \tag{38}$$

where $\hat{\mathbf{p}} = \sum_{c=1}^{C} \mathbf{p}_{(c)}$ and $n = \sum_{c=1}^{C}\sum_{i=1}^{N} p_{ci}$. Note that the probability values for each class are always normalized to sum to one. By setting $m_{ci} = p_{ci}$, we get $n_{(c)} = 1$ from (12) for all classes and, thus, $\hat{\mathbf{n}}$ in (14) is a vector of ones and $\text{diag}(\hat{\mathbf{n}}^{[1/2]})$ in (16) is an identity matrix. This gives us simpler formulas for $\mathbf{S}_w$ and $\mathbf{S}_b$ than the ones used in MLDA.

The optimal projection matrix $\mathbf{W}$ can be obtained by solving the regularized version of the generalized eigenproblem in (7)

$$\mathbf{S}_b\mathbf{w} = (\mathbf{S}_w + \epsilon\mathbf{I})\lambda\mathbf{w} \tag{39}$$

where ϵ is a small constant added to the diagonal values of $\mathbf{S}_w$ to avoid problems caused by singularity. We select the eigenvectors corresponding to d largest eigenvalues containing 0.999 of the information to form the projection matrix $\mathbf{W}$ and, finally, the features optimized for multilabel classification can be obtained as

$$\mathbf{Z} = \mathbf{W}^T\mathbf{X}. \tag{40}$$

The pseudocode for the overall SMLDA algorithm is provided in Algorithm 1. In the pseudocode, we give the correlation-based prior information type as our default type, but other prior information types can be used by simply replacing (32) on the pseudocode line 4 with a formula of another prior information type.

Algorithm 1: The Pseudocode of SMLDA

```
   /* Training procedure for obtaining
      optimal projection matrix W                    */
   Input: $\mathbf{X}_{train} \in \mathbb{R}^{D\times N}$,  $\mathbf{Y}_{train} \in \mathbb{R}^{C\times N}$
   Output: Projection matrix $\mathbf{W} \in \mathbb{R}^{D\times d}$
1  Create the probability matrix $\mathbf{P} \in \mathbb{R}^{C\times N}$ and fill it with
   zeros;
2  for each class $c \in \{1, \ldots, C\}$ do
3      Calculate the affinity matrix $\mathbf{A}^c \in \mathbb{R}^{N^c\times N^c}$ using (30);
4      Calculate the prior information matrix $\mathbf{V}^c \in \mathbb{R}^{N^c\times N^c}$
       using (32);
5      Calculate diagonal matrices $\mathbf{D_1}^c, \mathbf{D_2}^c \in \mathbb{R}^{N^c\times N^c}$ as
       $[\mathbf{D_1}^c]_{ii} = \sum_j [\mathbf{A}_c]_{ij}$ and $[\mathbf{D_2}^c]_{ii} = \sum_j [\mathbf{A}_c]_{ji}$;
6      Calculate $\mathbf{H}^c = \frac{1}{2}\mathbf{D_1}^c + \frac{1}{2}\mathbf{D_2}^c - \mathbf{A}^c + \mathbf{V}^c$;
7      Using Eq. (31), solve the probability matrix $\mathbf{p}^{c*}$;
8      Put the values of $\mathbf{p}^{c*}$ to correct places in $\mathbf{P}$;
9  end
10 Calculate the scatter matrices $\mathbf{S}_w$ and $\mathbf{S}_b$ using Eqs. (37)
   and (38);
11 Solve the projection matrix $\mathbf{W}$ using Eq. (39);
```

C. Computational Complexity Analysis

The computational complexity of the proposed SMLDA algorithm is formed as follows: for a class with N^c associated data items, the computational complexity of computing the kernel matrix is $([N^{c2} - N^c]/2)$, that is, the complexity of computing the affinity matrix is $\mathcal{O}(N^{c2})$. The complexity of computing the prior information matrix using (32) is $\mathcal{O}(C^2N)$ as it requires computing the correlation between each pair of classes using (21) and multiplying $C \times C$ and $C \times N$ matrices in (22). The computational complexity of solving (31) is $\mathcal{O}(N^{c3})$ due to the required matrix inversion. The overall complexity of applying the probabilistic multilabel class-saliency estimation for all the classes becomes $\mathcal{O}(\max_c N^{c3})$. The complexity of the LDA operation for D-dimensional data items is $\mathcal{O}(D^3)$. The overall complexity of SMLDA is $\mathcal{O}(\max_c N^{c3} + D^3)$. Thus, if $\max_c N^c < D$, the proposed method does not significantly affect the complexity compared to the standard LDA operation, but for $\max N^c > D$ the complexity is higher.

IV. Experiments

A. Databases and Data Preprocessing

We performed our experiments on 17 publicly available multilabel databases[1,2]. The datasets and their characteristics are given in Table I, where "Cardinality" means the mean numbers of class labels per instance for the training set and "Min #/Max #" shows the smallest/largest class size in the training set. The mean imbalance ratio ("MeanIR") measures the dataset imbalance following [59], where the imbalance for a class is computed by dividing the largest class size by the

[1] http://ceai.njnu.edu.cn/Lab/LABIC/LABIC_Software.html

[2] http://www.uco.es/kdis/mllresources/#KatakisEtAl2008

TABLE I
CHARACTERISTICS OF DATASETS USED FOR EXPERIMENTS

Database	Contents	Train #	Test #	Classes	Features	Cardinality	Min #	Max #	meanIR	meanCIR
Bibtex [46]	Text	4880	2515	159	1836	2.4	28	691	12.8	89.3
Birds [47]	Audio	179	172	19	260	1.9	4	64	6.1	16.1
Cal500 [48]	Music	300	202	174/173	68	26.1	2	263	21.1	23.1
CHD_49 [49]	Medicine	371	181	6	49	2.6	12	281	5.3	6.6
Corel16k(001) [50], [14]	Scene	5188	1744	153	500	3.1	21	1124	23.8	108.8
Emotions [51]	Music	398	195	6	72	1.9	96	181	1.5	2.4
Enron [52]	Text	988	660	57/53	1001	27	0	535	74.8	137.1
Eukaryote [19]	Biology	4658	3108	22	440	1.1	6	1387	45.1	150.5
Human [53]	Biology	1862	1244	14	440	1.2	14	623	15.4	45.2
Image [54]	Scene	1200	800	5	294	1.2	249	345	1.2	3.1
Medical [55]	Text	645	333	45/34	1449	1.2	0	170	60.9	230.2
PlantPseAAC [53]	Biology	588	390	12	440	1.1	12	172	6.7	21.8
Scene [56]	Scene	1211	1196	6	294	1.1	165	277	1.3	4.8
Stackex_coffee [5]	Text	151	74	123/63	1763	2.0	0	32	22.6	105.6
TMC2007-500 [57]	Text	21519	7077	22	500	2.2	304	12876	17.1	27.6
Yeast [34]	Biology	1500	917	14	103	4.2	21	1128	7.3	9.0
Yelp [58]	Text	6724	3281	5	671	1.8	580	4263	2.8	3.7

size of the class (i.e., this value is 1 for the largest class and larger for other classes). MeanIR is the mean over all the classes. Mean class imbalance ratio ("MeanCIR") denotes mean imbalance as in [60], where the imbalance of a class is computed by dividing the number of negative samples by the number of positive samples if the number of negative samples is larger and vice versa if the number of positive samples is larger. MeanCIR is the mean over all the classes. Thus, meanIR measures the imbalance between classes, which is our main interest. MeanCIR, on the other hand, focuses on the imbalance between positive and negative samples and maybe high even if all the classes have equal size.

We centralized the datasets and, for non-LDA-based techniques, we centralized also the label matrix used for training. We deleted some instances without labels or with NaN values. Some of the datasets have empty classes with no samples in either train or test set. For such datasets, we used all the samples and the full label matrix for training, but for computing the evaluation metrics we considered only classes with at least one test sample. If the number of test classes for a dataset is lower than the overall class number, we show also the number of test classes in the "Classes" column of Table I.

B. Evaluation Metrics

We adopt seven different evaluation metrics [61] to evaluate the performance of our proposed algorithm. Here, we denote the ground-truth label matrix for the M test samples as $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_M]$, where the ith column $\mathbf{y}_i \in \mathbb{R}^C$ represents the label vector of test sample $\mathbf{x}_i$. The multilabel classifiers give as their outputs for each input vector $\mathbf{x}_i$, a vector $\hat{\mathbf{p}}_i = f(\mathbf{x}_i)$, where $\hat{p}_{i,c}$ denotes the membership of instance i in class c. This is then converted to a binary predicted label vector $\hat{\mathbf{y}}_i$ by thresholding. $\mathcal{L}_i = \{\text{sort}_c(\hat{\mathbf{p}}_i)\}$ denotes an ordered list of classes ranked in the order of descending probability in $\hat{\mathbf{p}}_i$. $\mathcal{I}(\mathbf{y}_i)$ is used to denote the indices of relevant classes in $\mathbf{y}_i$ and $\neg\mathcal{I}(\mathbf{y}_i)$ denotes the indices of negative classes in $\mathbf{y}_i$. We use (↓) to denote metrics, where lower values indicate better results and (↑) in the opposite case.

1) *Ranking loss (↓)* evaluates for each item i relevant versus irrelevant class pair and gives the fraction of pairs, where the irrelevant class if ranked above the relevant one. Here, we use m to denote the number of relevant classes in $\mathbf{y}_i$ and $n = C - m$

$$\text{ranking\_loss}_i = \frac{|\hat{p}_{i,\mathcal{I}(\mathbf{y}_i)} \leq \hat{p}_{i,\neg\mathcal{I}(\mathbf{y}_i)}|}{m * n} \tag{41}$$

$$\text{ranking\_loss} = \frac{\sum_{i=1}^{M} \text{ranking\_loss}_i}{M} \tag{42}$$

where $|\hat{p}_{i,\mathcal{I}(\mathbf{y}_i)} \leq \hat{p}_{i,\neg\mathcal{I}(\mathbf{y}_i)}|$ is used to denote the count of wrong rankings for item i.

2) *One error (↓)* shows how often the top-ranked class for an item is not among the positive ground-truth labels

$$\text{one\_error}_i = \begin{cases} 0, & \text{if } \mathcal{L}_i[1] \in \mathcal{I}(\mathbf{y}_i) \\ 1, & \text{otherwise} \end{cases} \tag{43}$$

where $\mathcal{L}_i[1]$ denotes the first class in the sorted list $\mathcal{L}_i$

$$\text{one\_error} = \frac{\sum_{i=1}^{M} \text{one\_error}_i}{M}. \tag{44}$$

3) *Normalized coverage (↓)* demonstrates how far on average in the predicted label ranking $\mathcal{L}_i$ one needs to go to cover all the ground-truth labels of an instance

$$\text{coverage} = \frac{\sum_{i=1}^{M} \max_j \{j | \mathcal{I}(\mathbf{y}_i) \in_j \mathcal{L}_i\} - 1}{M * (C-1)} \tag{45}$$

where $\{j | \mathcal{I}(\mathbf{y}_i) \in_j \mathcal{L}_i\}$ gives the positions of relevant classes $\mathcal{I}(\mathbf{y}_i)$ in the ordered list $\mathcal{L}$.

4) *Macro-AUC (↑)* is the average area under ROC curves (AUC) for different classes [61]. The ROC curve uses the true-positive rate and false-positive rate, which may be unreliable in the cases, where very rare classes are present (high meanCIR) [62].

5) *Micro-AUC (↑)* is the area under ROC curves (AUC) averaged over the full predicted label matrix $\hat{\mathbf{Y}}$ [61].

6) *Macro-F1 (↑)* shows the average $F1$ value on each class

$$\text{macro}F1 = \frac{2}{C} \sum_{c=1}^{C} \frac{\text{precision}_c * \text{recall}_c}{\text{precision}_c + \text{recall}_c} \tag{46}$$

where $\text{precision}_c = \text{TP}_c/(\text{TP}_c + \text{FP}_c)$ and $\text{recall}_c = \text{TP}_c/(\text{TP}_c + \text{FN}_c)$ are precision and recall for class c,

TABLE II
SUMMARY OF THE EVALUATION METRIC PROPERTIES

	Optimized by		Uses
Metric	label-wise eff.	instance-wise eff.	threshold
Ranking loss	✓		
One error	✓		
Normalized coverage	✓		
Macro-AUC		✓	
Micro-AUC	✓		
Macro-FI		✓	✓
Micro-FI			✓

and TP_c, FP_c, and FN_c are the number of true positives, false positives, and false negatives for class c.

7) *Micro-F1 (↑)* indicates the overall $F1$ score averaged over the full predicted label matrix $\hat{\mathbf{Y}}$

$$\mathrm{micro}F1 = 2 * \frac{\mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{47}$$

where precision $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ and recall $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ and TP, FP, and FN are the number of true positives, false positives, and false negatives predictions in the predicted label matrix $\hat{\mathbf{Y}}$.

Some characteristics of the used metrics are summarized in Table II following the analysis provided in [61]. Most of the metrics are based on the predicted membership vectors $\hat{\mathbf{p}}_i$, while the last two use the predicted class labels that can be obtained from the predicted memberships by setting a threshold. It is possible to get different predicted labels from the same $\hat{\mathbf{p}}_i$ with different thresholds, but this does not depend on the input features, that is, the quality of the dimensionality reduction techniques. Therefore, the metrics based on the predicted memberships are well suited for evaluating the differences of the dimensionality reduction techniques.

Most of the metrics can be optimized by *labelwise effective* classifiers, which roughly means that the classifier can give higher membership values for the relevant classes than for the irrelevant classes for every sample. *Instancewise effective* classifiers, on the other hand, can distinguish between relevant and irrelevant samples for each class. Some classifiers, such as Micro-FI, are optimized only by *double effective* classifiers that are both labelwise and instancewise effective. The metrics that optimize instancewise effectiveness give more weight to the samples in smaller classes and, thus, are suitable for evaluating the performance in imbalanced (high meanIR) datasets, when it is not desired to obtain an overall high performance by predicting the majority classes correctly and failing in the rare classes. Due to the aforementioned unreliability of ROC curves in the presence of a very small class (high meanCIR), we use macro-F1 as the main metric for imbalance-aware evaluation.

C. Experimental Setup

We carried out all the experiments using two multilabel classifiers applied to the projected data: 1) multilabel k-nearest neighbor classifier (ML-kNN) [54] and 2) multioutput linear ridge regressor (LRR) [2], [63]. ML-kNN utilizes the k-nearest neighbor algorithm and maximum a posterior (MAP) principle to tackle the multilabel categorization task. ML-kNN first estimates prior and posterior probabilities of each instance i for each class c from a training dataset based on frequency counting [54]. Then, the predicted probabilities on a test dataset are calculated using the Bayesian rule. In our work, the predicted labels were obtained by setting a threshold (≥ 0.5) for the predicted probabilities.The hyperparameter k of ML-kNN was set to 15 as in [14]. As multilabel classification is a specific case of multitarget regression [64], the multioutput LRR can be trained to solve the multilabel classification tasks. In our work, we used the LRR classifier with a hyperparameter $\mu = 0.1$. The predicted labels were obtained by setting a threshold (≥ 0) for the predicted values from the LRR classifiers.

For comparisons, we used the following LDA-based dimensionality reduction techniques: DMLDA [21], wMLDAc, wMLDAb, wMLDAe, wMLDAf, and wMLDAd [14], where the subscripts denote the types of prior information used as weight factors following Section III-A1. Note that wMLDAc is equivalent to the original MLDA [12]. For all the LDA-based methods, we solved the regularized generalized eigenproblem (39) with $\epsilon = 0.1$. After solving the eigenproblem, we kept the eigenvectors corresponding to the top 0.999 informative eigenvalues to form the projection matrix $\mathbf{W}$. Besides the LDA-based methods, we conducted experiments with five other dimensionality reduction techniques: PCA, CCA [16], MLSI [17], MDDM_d [18], and MVMD [19]. We used the MATLAB codes provided for [14][1] in the comparative experiments and exploit the relevant parts also in the implementation of our proposed method.

D. Classification Results and Analysis

1) Comparisons of Different Variants of SMLDA and wMLDA: We first compare the different variants of our proposed SMLDA approach. Furthermore, we compare our methods against the variants of wMLDA that use the same prior information types directly as weights. We show the results using the ranking loss evaluation metric in Tables III and IV of the main paper and the results using the six other evaluation metrics in Tables I–XII of the supplementary material. In each table, we place next to each other the variants of SMLDA and wMLDA with the same prior information type and highlight the better approach for each dataset. The prior information for SMLDAm was proposed by us and has not been previously used with wMLDA. Therefore, we do not show such a comparison for it.

We first observe that our proposed SMLDA variants clearly outperform the corresponding wMLDA variants. In all test cases by both classifiers and any evaluation metric, the average performance of the proposed approach is better. This clearly confirms the value of using the probabilistic saliency estimation instead of just using the same prior information type directly as a weight as in wMLDA.

Next, we observe that there are no major differences among the variants of SMLDA. Therefore, we do not recommend using SMLDAd or SMLDAf because the fuzzy and dependence-based prior information types are computationally much more expensive than the other prior information types. Among the remaining variants, we select SMLDAc as our default variant.

TABLE III
COMPARISON OF DIFFERENT VARIANTS OF THE PROPOSED METHOD RESULTS WITH ML-kNN USING RANKING LOSS (↓)

	wMLDAc	SMLDAc	wMLDAb	SMLDAb	wMLDAe	SMLDAe	wMLDAf	SMLDAf	wMLDAd	SMLDAd	SMLDAm
Bibtex	0.164	**0.149**	**0.151**	0.152	0.153	**0.149**	0.151	**0.150**	0.147	**0.146**	0.147
Birds	0.217	**0.200**	0.206	**0.197**	**0.193**	0.197	0.201	**0.196**	0.232	**0.204**	0.204
CHD_49	0.212	**0.195**	**0.200**	0.206	0.198	**0.194**	**0.195**	0.200	0.226	**0.206**	0.205
Cal500	0.190	**0.187**	**0.187**	0.187	0.188	**0.187**	0.187	**0.187**	**0.186**	0.188	0.186
Corel16k(001)	0.190	**0.187**	**0.186**	0.186	**0.187**	0.187	**0.187**	0.187	0.186	**0.184**	0.182
Emotions	**0.173**	0.190	**0.162**	0.187	**0.164**	0.177	0.182	**0.177**	0.205	**0.184**	0.182
Enron	0.218	**0.142**	0.188	**0.145**	0.177	**0.142**	0.178	**0.142**	0.161	**0.139**	0.142
Eukaryote	0.122	**0.121**	0.122	**0.121**	**0.121**	0.121	**0.120**	0.121	**0.119**	0.121	0.120
Human	0.173	**0.160**	0.172	**0.162**	0.171	**0.162**	0.172	**0.159**	0.171	**0.162**	0.157
Image	0.193	**0.173**	0.199	**0.160**	0.195	**0.167**	0.199	**0.166**	0.203	**0.162**	0.172
Medical	0.071	**0.060**	0.066	**0.059**	0.065	**0.060**	0.064	**0.059**	0.071	**0.058**	0.057
PlantPseAAC	0.280	**0.228**	0.260	**0.230**	0.284	**0.225**	0.291	**0.229**	0.271	**0.234**	0.224
Scene	0.135	**0.088**	0.137	**0.087**	0.135	**0.089**	0.135	**0.088**	0.132	**0.089**	0.092
Stackex_coffee	**0.241**	0.273	**0.268**	0.272	**0.269**	0.272	**0.271**	0.272	0.284	**0.270**	0.275
TMC2007	0.027	**0.026**	0.026	**0.026**	0.026	**0.026**	0.026	**0.026**	0.028	**0.026**	0.027
Yeast	0.183	**0.178**	0.185	**0.177**	0.183	**0.178**	0.185	**0.178**	0.185	**0.177**	0.178
Yelp	0.126	**0.124**	0.130	**0.123**	0.126	**0.125**	**0.125**	0.126	0.139	**0.131**	0.139
Average	0.171	**0.158**	0.167	**0.158**	0.167	**0.156**	0.169	**0.157**	0.173	**0.158**	0.158
	Statistical analysis: Friedman: $p =$ **4.6e-05**, Wilcoxon Signed-Ranks test wrt. SMLDAc:										
	25.0	X	**25.0**	-66.0	36.0	-53.0	**30.0**	-57.0	**10.0**	-73.0	-67.0

TABLE IV
COMPARISON OF DIFFERENT VARIANTS OF THE PROPOSED METHOD RESULTS WITH LRR USING RANKING LOSS (↓)

	wMLDAc	SMLDAc	wMLDAb	SMLDAb	wMLDAe	SMLDAe	wMLDAf	SMLDAf	wMLDAd	SMLDAd	SMLDAm
Bibtex	0.120	**0.115**	0.120	**0.115**	0.120	**0.115**	0.119	**0.115**	0.124	**0.114**	0.112
Birds	0.332	**0.268**	0.321	**0.265**	0.321	**0.270**	0.318	**0.270**	0.321	**0.263**	0.258
CHD_49	0.209	**0.207**	0.208	**0.203**	0.208	**0.204**	0.208	**0.204**	0.213	**0.196**	0.189
Cal500	**0.242**	0.267	**0.250**	0.267	**0.266**	0.266	**0.265**	0.267	**0.194**	0.267	0.266
Corel16k(001)	**0.201**	0.202	0.206	**0.202**	0.204	**0.202**	0.204	**0.202**	**0.190**	0.199	0.197
Emotions	0.172	**0.163**	0.166	**0.161**	0.170	**0.168**	**0.168**	0.168	0.171	**0.162**	0.159
Enron	0.360	**0.198**	0.344	**0.195**	0.327	**0.196**	0.326	**0.196**	0.306	**0.195**	0.189
Eukaryote	0.129	**0.125**	0.130	**0.126**	0.129	**0.125**	0.129	**0.125**	0.128	**0.125**	0.123
Human	0.199	**0.179**	0.203	**0.181**	0.200	**0.178**	0.199	**0.178**	0.194	**0.174**	0.171
Image	0.208	**0.174**	0.203	**0.174**	0.205	**0.172**	0.203	**0.172**	0.203	**0.175**	0.188
Medical	0.044	**0.027**	0.036	**0.027**	0.035	**0.027**	0.033	**0.027**	0.044	**0.026**	0.028
PlantPseAAC	0.356	**0.339**	0.352	**0.336**	0.352	**0.339**	0.351	**0.339**	0.352	**0.338**	0.329
Scene	0.137	**0.091**	0.136	**0.092**	0.137	**0.092**	0.136	**0.091**	0.133	**0.092**	0.092
Stackex_coffee	0.199	**0.157**	**0.158**	0.160	**0.159**	0.160	0.188	**0.162**	0.188	**0.156**	0.157
TMC2007	0.040	**0.040**	**0.038**	0.040	**0.039**	0.040	**0.039**	0.040	0.047	**0.040**	0.042
Yeast	0.184	**0.178**	0.182	**0.178**	0.184	**0.178**	0.185	**0.178**	0.188	**0.178**	0.177
Yelp	0.137	**0.136**	0.137	**0.136**	0.137	**0.135**	0.137	**0.135**	0.148	**0.142**	0.147
Average	0.192	**0.169**	0.188	**0.168**	0.188	**0.169**	0.189	**0.169**	0.185	**0.167**	0.166
	Statistical analysis: Friedman: $p =$ **1.8e-10**, Wilcoxon Signed-Ranks test wrt. SMLDAc:										
	14.0	X	**16.0**	-53.0	**5.0**	-69.5	**3.0**	-56.0	**23.0**	**-31.0**	-44.0

2) Comparisons Against Competing Dimensionality Reduction Techniques: We then compare wMLDAc, which we recommend using as our default variant, against other competing dimensionality reduction techniques. Here, we consider five non-LDA-based techniques: 1) PCA; 2) CCA; 3) MLSI; 4) MDDMp; and 5) MVMD, along with DMLDA, MLDA, which is equivalent to wMLDAc and uses the same prior information as our proposed variant wMLDAc, and wMLDAd, which was the proposed wLMDA variant in [14]. We provide the results in Tables V and VI of the main paper and Tables XIII–XXIV of the supplementary material.

The results show that our proposed method has the best average performance with ML-kNN evaluated by all the performance metrics and with LRR evaluated by macro-F1. MDDMp is the best performing competing method. However, in all cases, our proposed approach achieves a similar performance, while our method is clearly better when evaluated with macro-F1. Our proposed method also clearly outperforms other LDA-based techniques.

We then focus on the most imbalanced datasets evaluated by our main metric for imbalanced classification, macro-F1. We collect from Tables XVII and XVIII of the supplementary material the results for the classes having meanIR over 15 and provide them in Tables VII and VIII. Our proposed method has the best average performance with ML-kNN and the second best with LRR, which shows that the proposed method indeed can help to deal with class imbalance.

3) Statistical Analysis of the Results: To evaluate whether the observed differences are statistically significant, we followed the recommendations of [65]. We first applied to each table the Friedman test, which is a rank-based nonparametric test showing whether the differences are overall significant. At the bottom of each table, we report the Friedman p value. We have highlighted the value if it shows that the null hypothesis can be rejected at the 0.05 significance level. Next, we perform the Wilcoxon sign-ranks test to evaluate the pairwise differences between the methods. This test ranks the differences between two classifiers ignoring the signs and uses the ranks to determine value T as described, for example, in [65]. Finally, the T value is compared to a critical value that depends on the number of datasets. In our experiments, we used 17 datasets, which means that the null hypothesis

TABLE V
COMPARATIVE RESULTS WITH ML-kNN USING RANKING LOSS (↓)

	Competing methods								Proposed
	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAd	SMLDAc
Bibtex	0.204	0.197	0.199	**0.116**	0.186	0.271	0.164	0.147	0.149
Birds	0.323	0.203	0.323	0.322	0.322	0.248	0.217	0.232	**0.200**
CHD_49	0.224	0.212	0.214	0.209	0.225	0.224	0.212	0.226	**0.195**
Cal500	0.183	0.187	0.184	**0.182**	0.183	0.187	0.190	0.186	0.187
Corel16k(001)	0.198	0.188	0.196	**0.185**	0.198	0.197	0.190	0.186	0.187
Emotions	0.299	0.178	0.299	0.301	0.295	0.245	**0.173**	0.205	0.190
Enron	0.133	0.170	0.135	**0.124**	0.136	0.191	0.218	0.161	0.142
Eukaryote	0.113	0.126	0.113	**0.106**	0.111	0.141	0.122	0.119	0.121
Human	0.159	0.178	0.159	**0.149**	0.158	0.191	0.173	0.171	0.160
Image	0.167	0.201	0.170	0.186	**0.166**	0.284	0.193	0.203	0.173
Medical	0.057	0.076	**0.039**	0.051	0.058	0.072	0.071	0.071	0.060
PlantPseAAC	0.197	0.277	0.198	**0.180**	0.198	0.258	0.280	0.271	0.228
Scene	0.084	0.141	0.083	0.102	**0.077**	0.234	0.135	0.132	0.088
Stackex_coffee	0.279	0.304	0.259	0.257	0.276	0.298	**0.241**	0.284	0.273
TMC2007	0.035	0.026	0.035	0.030	0.030	0.038	0.027	0.028	**0.026**
Yeast	0.174	0.184	**0.173**	0.179	0.174	0.188	0.183	0.185	0.178
Yelp	0.178	**0.117**	0.171	0.148	0.176	0.139	0.126	0.139	0.124
Average	0.177	0.174	0.174	0.166	0.175	0.200	0.171	0.173	**0.158**
	Statistical analysis: Friedman: $p =$ **1.8e-04**, Wilcoxon Signed-Ranks test wrt. SMLDAc:								
	52.0	**16.0**	63.0	-74.0	61.0	**0.0**	**25.0**	**10.0**	X

TABLE VI
COMPARATIVE RESULTS WITH LRR USING RANKING LOSS (↓)

	Competing methods								Proposed
	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAd	SMLDAc
Bibtex	0.117	0.120	0.117	**0.079**	0.091	0.120	0.120	0.124	0.115
Birds	0.236	0.301	0.288	**0.171**	0.199	0.318	0.332	0.321	0.268
CHD_49	0.208	0.210	0.208	**0.196**	0.205	0.210	0.209	0.213	0.207
Cal500	0.258	0.269	0.265	0.248	0.250	0.245	0.242	**0.194**	0.267
Corel16k(001)	0.208	0.208	0.208	0.195	0.208	0.206	0.201	**0.190**	0.202
Emotions	0.163	0.167	0.163	0.174	0.177	0.166	0.172	0.171	**0.163**
Enron	0.250	0.324	0.332	**0.121**	0.138	0.400	0.360	0.306	0.198
Eukaryote	0.134	0.130	0.134	**0.111**	0.120	0.131	0.129	0.128	0.125
Human	0.211	0.209	0.210	**0.157**	0.185	0.210	0.199	0.194	0.179
Image	0.206	0.207	0.217	0.198	0.177	0.223	0.208	0.203	**0.174**
Medical	0.031	0.039	0.057	0.025	**0.024**	0.063	0.044	0.044	0.027
PlantPseAAC	0.340	0.351	0.343	**0.194**	0.315	0.362	0.356	0.352	0.339
Scene	0.136	0.136	0.138	0.097	**0.088**	0.141	0.137	0.133	0.091
Stackex_coffee	0.170	0.168	0.169	**0.157**	0.171	0.163	0.199	0.188	0.157
TMC2007	0.038	0.037	0.038	0.049	0.048	**0.037**	0.040	0.047	0.040
Yeast	0.184	0.183	0.184	0.180	0.179	0.182	0.184	0.188	**0.178**
Yelp	0.130	0.129	0.130	0.165	0.135	**0.129**	0.137	0.148	0.136
Average	0.178	0.188	0.188	**0.148**	0.159	0.195	0.192	0.185	0.169
	Statistical analysis: Friedman: $p =$ **6.6e-05**, Wilcoxon Signed-Ranks test wrt. SMLDAc:								
	37.0	**11.0**	**16.0**	-44.0	-55.0	**20.0**	**14.0**	**23.0**	X

TABLE VII
COMPARATIVE RESULTS WITH ML-kNN USING MACRO-F1 (↑)

	Competing methods								Proposed
	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAd	SMLDAc
Cal500	0.056	0.050	0.055	**0.062**	0.055	0.051	0.051	0.056	0.051
Corel16k(001)	0.013	0.030	0.017	0.036	0.018	0.018	**0.037**	0.034	0.036
Enron	0.046	0.073	0.042	**0.095**	0.054	0.012	0.039	0.036	0.062
Eukaryote	0.053	0.074	0.053	0.060	0.065	0.002	0.090	**0.092**	0.072
Human	0.043	0.146	0.041	0.095	0.071	0.001	0.145	0.133	**0.159**
Medical	0.219	0.294	**0.307**	0.280	0.226	0.186	0.259	0.263	0.302
Stackex_coffee	0.000	0.023	0.017	0.013	0.000	0.010	0.036	0.040	**0.048**
Average	0.061	0.099	0.076	0.092	0.070	0.040	0.094	0.093	**0.104**
	Statistical analysis: Friedman: $p =$ **2.7e-03**, Wilcoxon Signed-Ranks test wrt. SMLDAc:								
	1.0	7.0	3.0	7.0	**1.0**	**0.0**	8.0	6.0	X

can be rejected at 0.01 significance level if $T_1 \leq 23$ and at 0.05 significance level if $T_2 \leq 34$. For seven datasets, as in Tables VII and VIII, $T_2 \leq 2$. We applied the Wilcoxon sign-ranks test between our proposed SMLDAc method and every other method dimensionality reduction technique. We give these values at the bottom of every table and bold the values if they show that the difference between the methods is statistically significant at a 0.05 significance level. Negative

TABLE VIII
COMPARATIVE RESULTS WITH LRR USING MACRO-F1 (↑)

	Competing methods								Proposed
	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAd	SMLDAc
Cal500	0.122	0.125	0.126	0.120	0.120	0.104	0.103	0.070	**0.127**
Corel16k(001)	0.043	0.044	0.043	0.035	0.041	**0.044**	0.042	0.037	0.044
Enron	**0.123**	0.117	0.121	0.086	0.101	0.097	0.101	0.095	0.113
Eukaryote	0.113	**0.119**	0.113	0.097	0.111	0.117	0.119	0.117	0.111
Human	0.143	0.149	0.145	0.136	0.151	0.148	0.147	0.150	**0.156**
Medical	0.531	**0.551**	0.489	0.444	0.487	0.488	0.440	0.443	0.536
Stackex_coffee	0.171	0.179	0.186	0.124	0.165	0.190	0.144	0.159	**0.196**
Average	0.178	**0.184**	0.175	0.149	0.168	0.170	0.156	0.153	0.183
	Statistical analysis: Friedman: $p =$ **1.2e-03**, Wilcoxon Signed-Ranks test wrt. SMLDAc:								
	7.0	-14.0	7.0	**0.0**	**1.0**	4.0	**2.0**	**1.0**	X

TABLE IX
SUMMARY OF THE WILCOXON SIGNED-RANKS TEST RESULTS: THE NUMBER OF TIMES WHEN SMLDAc WAS BETTER IN A STATISTICALLY SIGNIFICANT WAY

	PCA	CCA	MLSI	MDDMp	MVMD	DMLDA	wMLDAc	wMLDAb	wMLDAe	wMLDAf	wMLDAd
ML-kNN	4/7	5/7	2/7	1/7	2/7	7/7	4/7	4/7	1/7	4/7	7/7
LRR	1/7	4/7	5/7	1/7	0/7	4/7	7/7	5/7	7/7	5/7	7/7
Total	5/14	9/14	7/14	2/14	2/14	11/14	11/14	9/14	8/14	9/14	14/14

values indicate that the other method was performing better than SMLDAc.

The results of the Friedman test show that the overall differences are statistically significant in most cases. The only exceptions among 28 result tables are Tables I, II, XI, and XIV in the supplementary material. Tables I and II in the supplementary material, compare the variants of the proposed method using ML-kNN and LRR with one error evaluation metric. Table XI in the supplementary material, compares the variants of the proposed method using ML-kNN with the Micro-F1 evaluation metric. Table XIV in the supplementary material, compares SMLDAc against competing methods using LRR and one error evaluation metric.

The results of the Wilcoxon signed-ranks test confirm the good performance of our proposed SMLDAc. There is no such case, where a competing method would outperform SMLDAc in a statistically significant manner (only another variant of our proposed method, SMLDAd, can do this in two cases). On the other hand, SMLDAc can outperform every competing method in a statistically significant manner at least twice. The results of the conducted Wilcoxon signed-ranks test are summarized in Table IX showing the number of times when a statistically significant difference was detected between SMLDAc and all competing methods.

V. CONCLUSION

In this article, we proposed a novel probabilistic framework for the LDA-related dimensionality reduction algorithm aiming to improve the performance of multilabel classifiers on various multilabel datasets. The probabilistic approach uses an affinity matrix to ensure similar results for similar instances and a prior information matrix to integrate prior information on the prominence of each instance for each class. Our solution can alleviate the data imbalance problem, which is commonly encountered in multilabel datasets, as the weight factor vectors are calculated separately for each class. Our method can also alleviate the common overcounting problem. We proposed variants of our methods using different prior information matrices based on both labels and features.

We used seven metrics to evaluate the performance of our method with competing methods on 17 multilabel datasets. The experimental results showed that our method enhanced the classification performance compared to the competing algorithms and handles imbalanced classification well. Our algorithm is still based on the linear subspace learning technique. In the future, we will make a nonlinear extension using the kernel trick.

REFERENCES

[1] L. Li, H. Wang, X. Sun, B. Chang, S. Zhao, and L. Sha, "Multi-label text categorization with joint learning predictions-as-features method," in *Proc. Conf. Empirical Methods Nat. Language Process.*, 2015, pp. 835–839.

[2] C. Tan, S. Chen, G. Ji, and X. Geng, "Multilabel distribution learning based on multioutput regression and manifold learning," *IEEE Trans. Cybern.*, early access, Oct. 23, 2020, doi: 10.1109/TCYB.2020.3026576.

[3] J. Read, L. Martino, and J. Hollmén, "Multi-label methods for prediction with sequential data," *Pattern Recognit.*, vol. 63, pp. 45–55, Sep. 2016.

[4] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multi-label classification of music by emotion," *EURASIP J. Audio Speech Music Process.*, vol. 2011, no. 1, pp. 1–9, 2011. [Online]. Available: https://asmp-eurasipjournals.springeropen.com/articles/10.1186/1687-4722-2011-426793

[5] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Addressing imbalance in multilabel classification: Measures and random resampling algorithms," *Neurocomputing*, vol. 163, pp. 3–16, Sep. 2015.

[6] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognit.*, vol. 45, no. 9, pp. 3084–3104, 2012.

[7] F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus, *Multilabel Classification*. Cham, Switzerland: Springer, 2016. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-41111-8_1#citeas

[8] E. Gibaja and S. Ventura, "Multi-label learning: A review of the state of the art and ongoing research," *Interdiscipl. Rev. Data Min. Knowl. Disc.*, vol. 4, no. 6, pp. 411–444, 2014.

[9] W. Siblini, P. Kuntz, and F. Meyer, "A review on dimensionality reduction for multi-label classification," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 8, pp. 839–857, Mar. 2021.

[10] A. Iosifidis, A. Tefas, and I. Pitas, "Regularized extreme learning machine for multi-view semi-supervised action recognition," *Neurocomputing*, vol. 145, pp. 250–262, Dec. 2014.

[11] H. Wang, L. Yan, H. Huang, and C. Ding, "From protein sequence to protein function via multi-label linear discriminant analysis," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 14, no. 3, pp. 503–513, May/Jun. 2017.

[12] H. Wang, C. Ding, and H. Huang, *Multi-Label Linear Discriminant Analysis* (LNCS 6316). Heidelberg, Germany: Springer, 2010, pp. 126–139. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-15567-3_10#citeas

[13] L. Xu, A. Iosifidis, and M. Gabbouj, "Weighted linear discriminant analysis based on class saliency information," in *Proc. Int. Conf. Image Process. (ICIP)*, 2018, pp. 2306–2310.

[14] J. Xu, "A weighted linear discriminant analysis framework for multi-label feature extraction," *Neurocomputing*, vol. 275, pp. 107–120, Jan. 2018.

[15] I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments," *Philosop. Trans. Roy. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, 2016.

[16] L. Sun, S. Ji, and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 194–200, Jan. 2011.

[17] K. Yu, S. Yu, and V. Tresp, "Multi-label informed latent semantic indexing," *Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR)*, 2005, pp. 258–265.

[18] Y. Zhang and Z. H. Zhou, "Multilabel dimensionality reduction via dependence maximization," *ACM Trans. Knowl. Disc. Data*, vol. 4, no. 3, pp. 1–21, 2010.

[19] J. Xu, J. Liu, J. Yin, and C. Sun, "A multi-label feature extraction algorithm via maximizing feature variance and feature-label dependence simultaneously," *Knowl. Based Syst.*, vol. 98, pp. 172–184, Apr. 2016.

[20] C. H. Park and M. Lee, "On applying linear discriminant analysis for multi-labeled problems," *Pattern Recognit. Lett.*, vol. 29, no. 7, pp. 878–887, 2008.

[21] M. Oikonomou and A. Tefas, "Direct multi-label linear discriminant analysis," *Commun. Comput. Inf. Sci.*, vol. 383, no. 1, pp. 414–423, 2013.

[22] Y. Yuan, K. Zhao, and H. Lu, "Multi-label linear discriminant analysis with locality consistency," in *Proc. Int. Conf. Neural Inf. Process.*, 2014, pp. 386–394.

[23] F. Nie, S. Xiang, Y. Jia, and C. Zhang, "Semi-supervised orthogonal discriminant analysis via label propagation," *Pattern Recognit.*, vol. 42, no. 11, pp. 2615–2627, 2009.

[24] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[25] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–9.

[26] Z. Li, F. Nie, X. Chang, and Y. Yang, "Beyond trace ratio: Weighted harmonic mean of trace ratios for multiclass discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2100–2110, Oct. 2017.

[27] S. Petridis and S. J. Perantonis, "On the relation between discriminant analysis and mutual information for supervised linear feature extraction," *Pattern Recognit.*, vol. 37, no. 5, pp. 857–874, 2004.

[28] E. K. Tang, P. N. Suganthan, X. Yao, and A. K. Qin, "Linear dimensionality reduction using relevance weighted LDA," *Pattern Recognit.*, vol. 38, no. 4, pp. 485–493, 2005.

[29] E. Tang, P. Suganthan, and X. Yao, "Generalized LDA using relevance weighting and evolution strategy," in *Proc. Congr. Evol. Comput.*, vol. 1, 2005, pp. 2230–2234.

[30] Z. Li, D. Lin, and X. Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 755–761, Apr. 2009.

[31] C. Aytekin, A. Iosifidis, and M. Gabbouj, "Probabilistic saliency estimation," *Pattern Recognit.*, vol. 74, pp. 359–372, Feb. 2018.

[32] H. Ahmed, J. Mohamed, and Z. Noureddine, "Face recognition systems using relevance weighted two dimensional linear discriminant analysis algorithm," *J. Signal Inf. Process.*, vol. 3, no. 1, pp. 130–135, 2012.

[33] Q. Wu, M. Tan, H. Song, J. Chen, and M. K. Ng, "ML-FOREST: A multi-label tree ensemble method for multi-label classification," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 10, pp. 2665–2680, Oct. 2016.

[34] K. Nakai and M. Kanehisa, "A knowledge base for predicting protein localization sites in eukaryotic cells," *Genomics*, vol. 14, no. 4, pp. 897–911, 1992.

[35] W. Chen, J. Yan, B. Zhang, Z. Chen, and Q. Yang, "Document transformation for multi-label feature selection in text categorization," in *Proc. IEEE Int. Conf. Data Min. (ICDM)*, 2007, pp. 451–456.

[36] X. Lin and X.-W. Chen, "Mr.KNN—Soft relevance for multi-label classification." in *Proc. 19th ACM Conf. Inf. Knowl. Manage.*, 2010, pp. 349–358.

[37] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[38] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[39] C. Li *et al.*, "ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 88–100, Jan. 2021.

[40] F. Battisti, S. Baldoni, M. Brizzi, and M. Carli, "A feature-based approach for saliency estimation of omni-directional images," *Signal Process. Image Commun.*, vol. 69, pp. 53–59, Mar. 2018.

[41] M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 37, 2011, pp. 409–416.

[42] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.

[43] F. Nie, X. Wang, M. I. Jordan, and H. Huang, "The constrained Laplacian rank algorithm for graph-based clustering," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 1969–1976.

[44] Y. Zhu, J. T. Kwok, and Z.-H. Zhou, "Multi-label learning with global and local label correlation," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1081–1094, Jun. 2019.

[45] A. Gretton, O. Bousquet, A. Smola, and B. Scḧlkopf, "Measuring statistical dependence with Hilbert–Schmidt norms," *Proc. 16th Int. Conf. Algorithmic Learn. Theory (ALT)*, 2005, pp. 63–77.

[46] I. Katakis, G. Tsoumakas, and V. Ioannis, "Multilabel text classification for automated tag suggestion," in *Proc. ECML/PKDD Disc. Challenge*, 2008, pp. 1107–1135.

[47] F. Briggs *et al.*, "The 9th annual MLSP competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, 2013, pp. 1–8.

[48] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 2, pp. 467–476, Feb. 2008.

[49] H. Shao, G. Li, G. Liu, and Y. Wang, "Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine," *Sci. China Inf. Sci.*, vol. 56, no. 5, pp. 1–13, 2013.

[50] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, no. 6, pp. 1107–1135, 2003.

[51] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Effective and efficient multilabel classification in domains with large number of labels," in *Proc. ECML/PKDD Workshop Min. Multidimensional Data (MMD)*, 2008, pp. 30–44. [Online]. Available: http://lpis.csd.auth.gr/publications/tsoumakas-mmd08.pdf

[52] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Disc. Databases (ECML PKDD)*, vol. 5782, 2009, pp. 254–269. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-04174-7_17#citeas

[53] J. Xu, "Fast multi-label core vector machine," *Pattern Recognit.*, vol. 46, no. 3, pp. 885–898, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320312003950

[54] M. L. Zhang and Z. H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognit.*, vol. 40, no. 7, pp. 2038–2048, 2007.

[55] J. P. Pestian *et al.*, "A shared task involving multi-label classification of clinical free text," in *Proc. ACL Workshop BioNLP Biol. Transl. Clin. Lang. Process.*, Jun. 2007, pp. 97–104.

[56] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognit.*, vol. 37, no. 9, pp. 1757–1771, 2004.

[57] A. N. Srivastava and B. Zane-Ulman, "Discovering recurring anomalies in text reports regarding complex space systems," in *Proc. IEEE Aerosp. Conf.*, 2005, pp. 3853–3862.

[58] H. Sajnani, V. Saini, K. Kumar, E. Gabrielova, P. Choudary, and C. Lopes. (2013). *Classifying Yelp Reviews Into Relevant Categories.* [Online]. Available: http://www.ics.uci.edu/ vpsaini/.

[59] F. Charte, A. Rivera, M. J. del Jesus, and F. Herrera, "A first approach to deal with imbalance in multi-label datasets," in *Hybrid Artificial Intelligent Systems*, J.-S. Pan, M. M. Polycarpou, M. Woźniak, A. C. P. L. F. de Carvalho, H. Quintián, and E. Corchado, Eds.

Heidelberg, Germany: Springer, 2013, pp. 150–160. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-40846-5_16#citeas

[60] M.-L. Zhang, Y.-K. Li, and X.-Y. Liu, "Towards class-imbalance aware multi-label learning," in *Proc. 24th Int. Conf. Artif. Intell. (IJCAI)*, 2015, pp. 4041–4047.

[61] X.-Z. Wu and Z.-H. Zhou, "A unified view of multi-label performance measures," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3780–3788.

[62] J. Davis and M. Goadrich, "The relationship between precision-recall and RoC curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, New York, NY, USA, 2006, pp. 233–240. [Online]. Available: https://doi.org/10.1145/1143844.1143874

[63] H. Borchani, G. Varando, C. Bielza, and B. Monte, "A survey on multi-output regression," *Interdiscipl. Rev. Data Min. Knowl. Discov.*, vol. 5, no. 5, pp. 216–233, 2015.

[64] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas, "Multi-target regression via input space expansion: Treating targets as inputs," *Mach. Learn.*, vol. 104, no. 1, pp. 55–98, 2016.

[65] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, no. 1, pp. 1–30, 2006.

Lei Xu (Student Member, IEEE) received the B.S.E.E. degree from East China Normal University, Shanghai, China, in 2006, and the M.S.E.E. degree from the Tampere University of Technology, Tampere, Finland, in 2017. She is currently pursuing the Ph.D. degree with Tampere University, Tampere.

From 2006 to 2013, she was an Engineer in Shanghai, where she was involved with on-train communication systems design. Her current research interests include artificial intelligence, data science, and machine learning.

Jenni Raitoharju (Member, IEEE) received the Ph.D. degree from the Tampere University of Technology, Tampere, Finland, in 2017.

She works as a Senior Research Scientist with the Finnish Environment Institute, Jyväskylä, Finland. She has coauthored 25 international journal papers and 34 papers in international conferences. She currently leads two research projects funded by the Academy of Finland, focusing on automatic taxa identification. Her research interests include machine learning and pattern recognition methods along with applications in biomonitoring and autonomous systems.

Dr. Raitoharju is the Chair of Young Academy Finland from 2019 to 2021.

Alexandros Iosifidis (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Democritus University of Thrace, Komotini, Greece, in 2008 and 2010, respectively, and the Ph.D. degree from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2014.

He is currently an Associate Professor with Aarhus University, Aarhus, Denmark. He has coauthored 84 articles in international journals and 100 papers in international conferences and workshops. His research interests include topics of neural networks and statistical machine learning finding applications in computer vision, financial engineering, and graph analysis problems.

Prof. Iosifidis's work has received the H.C. Oersted Young Researcher Prize 2018 and the EURASIP Early Career Award 2021. He served as an Officer of the Finnish IEEE SP/CAS Chapter from 2016 to 2018. He is currently a member of the EURASIP Technical Area Committee on Visual Information Processing. He serves as the Associate Editor-in-Chief for *Neurocomputing*, as an Area Editor for *Signal Processing: Image Communications*, and an Associate Editor for *BMC Bioinformatics*.

Moncef Gabbouj (Fellow, IEEE) received the B.S. degree in electrical engineering from Oklahoma State University, Stillwater, OK, USA, in 1985, and the M.S. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1986 and 1989, respectively.

He is a Professor of Signal Processing with the Department of Computing Sciences, Tampere University, Tampere, Finland. He was an Academy of Finland Professor from 2011 to 2015. He is the Finland Site Director of the NSF IUCRC funded Center for Visual and Decision Informatics and leads the Artificial Intelligence Research Task Force of the Ministry of Economic Affairs and Employment funded Research Alliance on Autonomous Systems. His research interests include big data analytics, multimedia content-based analysis, indexing and retrieval, artificial intelligence, machine learning, pattern recognition, nonlinear signal and image processing and analysis, voice conversion, and video processing and coding.

Prof. Gabbouj is the past Chairman of the IEEE CAS TC on DSP and a committee member of the IEEE Fourier Award for Signal Processing. He has served as an associate editor and a guest editor for many IEEE and international journals and a Distinguished Lecturer for the IEEE CASS. He is a member of the Academia Europaea and the Finnish Academy of Science and Letters.

PUBLICATION III

Unsupervised facial image de-occlusion with optimized deep generative models

L. Xu, H. Zhang, J. Raitoharju, and M. Gabbouj

In *2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Xi'an, China, 7-10 November 2018
DOI: 10.1109/IPTA.2018.8608127

Unsupervised Facial Image De-occlusion with Optimized Deep Generative Models

Lei Xu, Honglei Zhang, Jenni Raitoharju and Moncef Gabbouj
Laboratory of Signal Processing, Tampere University of Technology, Tampere, Finland
e-mail: {lei.xu, honglei.zhang, jenni.raitoharju, moncef.gabbouj}@tut.fi

Abstract—**In recent years, Generative Adversarial Networks (GANs) or various types of Auto-Encoders (AEs) have gained attention on facial image de-occlusion and/or in-painting tasks. In this paper, we propose a novel unsupervised technique to remove occlusion from facial images and complete the occluded parts simultaneously with optimized Deep Convolutional Generative Adversarial Networks (DCGANs) in an iterative way. Generally, GANs, as generative models, can estimate the distribution of images using a generator and a discriminator. DCGANs, as its variant, are proposed to conquer its instability during training. Existing facial image in-painting methods manually define a block of pixels as the missing part and the potential content of this block is semantically generated using generative models, such as GANs or AEs. In our method, a mask is inferred from an occluded facial image using a novel loss function, and then this mask is utilized to in-paint the occlusions automatically by pre-trained DCGANs. We evaluate the performance of our method on facial images with various occlusions, such as sunglasses and scarves. The experiments demonstrate that our method can effectively detect certain kinds of occlusions and complete the occluded parts in an unsupervised manner.**

Keywords—**Generative Adversarial Networks, Deep Convolutional Generative Adversarial Networks, Facial Image De-occlusion, Facial Image Completion**

I. Introduction

Facial image de-occlusion refers to removing unnecessary occluded parts (e.g. scarf, glasses, cover) from a facial image, and then reconstructing its corresponding contents to a complete and realistic facial image conditioning on predefined image sets [1], [2], [3] and [6]. Such tasks are quite challenging, due to the diversity of occlusions and complexity of reconstructing the details of facial images. Over decades, various approaches have been proposed to deal with the facial image de-occlusion tasks, such as Principle Component Analysis (PCA) or its variants, sparse coding, and Auto-Encoders (AEs).

Image in-painting proposed by Bertalmio et al. firstly [13], also known as image completion, familiarly means to synthesize the missing pixels or remove unwanted pixels through learning context features from their surrounding regions, the whole image content, or external databases. Recently, Generative Adversarial Networks (GANs) [20] have received considerable results on various machine learning tasks. GANs or its variants are extensively applied to tackle facial image in-painting tasks, as in [10], [19]. Compared to facial image de-occlusion, one challenging problem of facial images in-painting is to locate missing regions beforehand. Compared to facial image de-occlusion, facial image in-painting does not implicitly locate the occlusion.

Although above methods have achieved significant results in tackling facial image de-occlusion or in-painting tasks, there still exist various limitations. For example, the linearity of PCA and sparse coding restricts the further improvement of the results; moreover, the result quality of both methods heavily depends on the consistency of training and testing dataset subjects. Generally, AE techniques with deep neural networks produce blurry results, unless they are combined with other strategies to preserve the details, as in [7]. Another point is that these mentioned methods require occlusion free image datasets as ground truth. Especially for AE related approaches, a large number of occlusion free images and corresponding occluded images are crucial in training a decent model.

Inspired by the above works, we propose an unsupervised approach on the basis of Deep Convolutional Generative Adversarial Networks (DCGANs), which does not require a large dataset with occlusion free images and their corresponding occluded images. Our approach not only aims to detect and segment the occlusion part automatically, but also in-paint the occluded part with the pre-trained DCGANs. The approach finds an optimized result in an iterative way.

This paper is organized as follows: in Section 2, we present related works, which inspired us to propose this novel idea. In the next section, we describe the theory related to this work and how it is exploited in our work. In Section 4, the neural networks structure, loss function, and algorithm are provided to meticulously explain the work-flow of the proposed work. Experimental results and the corresponding analysis are depicted in Section 5. Finally, in Section 6, we make conclusions about this work and discuss how to further improve it.

II. Related work

PCA or PCA variants can restore occluded images through manipulating eigenspaces of the training images as in [1] and [2]. A Robust-PCA framework [3] is used to detect occlusion masks of input images taking advantage of a non-occluded facial image set, and then inpaint the occluded parts based on prior information. Sparse coding techniques have been widely used in the field of image restoration [4], [5] to restore occluded images using sparse coefficients of a learned low-rank dictionary.

As depicted in [6], [7], and [8], AE and its variants are explicitly powerful enough to remove noise and reconstruct relatively clean images in various schemes. Besides, more efforts have been devoted to establish AE-related models for facial image de-occlusion tasks, as shown in [7], [8], and [11]. Zhao et al. [7] proposed a long short-term memory AE with two decoding channels to detect occlusions and to reconstruct faces simultaneously. Their method obtains a decent result without constraints (e.g. consistent occlusions in train and test sets) on the training and testing datasets. AEs usually produce blurry images, which can be explained by the ℓ_2-norm loss function used in AEs to calculate the similarity between the generated images and corresponding ground-truth images [12]. Due to this fact, Zhao et al. [11] introduced a supervised Convolutional Neural Network (CNN) and an adversarial CNN to lessen the blurring of the results. Zhang et al. [8] introduced a multi-AE structure to detect occlusion and restore partitioned parts of face on the basis of 68 facial landmarks. All the methods mentioned require facial images without any kind of occlusion and the corresponding occluded images of the same persons as the ground-truth dataset. It is difficult to collect such a dataset with a large number of images and various occlusions. So to train such methods, it is common to use artificially generated occluded images with different occlusions (e.g. scarves, glasses, sunglasses) from commonly used facial databases.

Bertalmio et al. [13] proposed an in-painting algorithm on the basis of professional restorators. Sun et al. [14] introduced an image in-painting algorithm with a global optimization method, which pays more emphasis on highlighting the structural integrity of the salient object than on the surrounding pixel values of the regions desired to be completed. After GANs [20] have been proposed in 2014, there has been an increasing number of works involving in GANs and its variants for facial image in-painting tasks. Unlike AEs and its variants, GANs and its variants can retain sharp details of the facial images.

GANs/AE-related models have significant effects in general image in-painting tasks, especially for completing large missing regions. Demir et al. [15] proposed a novel GANs structure consisting of a global GAN (G-GAN) and a patch GAN (P-GAN) to improve the quality of the completed images with artificial masks denoting the occlusions. A conditional AE is used by Pathak et al [16]. In this work, the latent space between the encoder and decoder is channel-wise fully connected, which can describe image more precisely. Inspired by [17] and [19], Lahiri et al. [12] constrained DCGANs with facial expression features as conditional information to improve the consistency and correctness of in-painted images. Yeh et al. [17] proposed a DCGANs structure to in-paint the missing part of a facial image semantically. In this work, the missing part of each image is given beforehand. Our work follows a similar approach for image in-painting, but, instead of using a pre-defined mask denoting missing pixels, we detect the occlusion as a mask automatically in an unsupervised way.

III. Preliminaries

Our proposed approach involves both facial image de-occlusion and facial image in-painting techniques. It targets to locate and remove the occlusions in facial images automatically, and then semantically complete the detected regions with appropriate contents. Hence, pre-trained DCGANs are iteratively used to locate occlusions and to generate the missing facial parts. In this section we introduce the key elements of the framework proposed in [17], which are also exploited in the proposed method.

A. Deep Convolutional Generative Adversarial Networks

Generative models aim to discover the statistical laws within the observed data and then to generate new data similar to the observed data on the basis of the obtained probability distribution model.

The GANs [20] are designed to include a generative model G and a discriminative model D. The main task of the discriminator D is to evaluate whether the data come from the real data distribution p_{data} or from a data distribution p_G generated by the generator G. During the training process, it aims to maximize the accuracy of discriminating the real data and the generated data, assigning 1 for the real data and 0 for the generated data. On the contrary, the generator G generates fraud data using a random vector $z \sim \mathcal{U}[-1, 1]$ as input. Its objective is to generate data that appears so authentic that the discriminator D is not able to discriminate it from the real data. These models acting against each other with the opposite goals bring GANs its name. In the training process of GANs, D and G are optimized alternately. When G is optimized, D is fixed and vice versa.

The objective function of the GANs is a zero-sum game between the generator G and the discriminator D, which can be considered also as a minimax two-player game. GANs is trained by optimizing the following loss function [20]:

$$\underset{G}{min}\, \underset{D}{max}\, V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})}[log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})}[log(1 - D(G(\mathbf{z})))]. \quad (1)$$

Here $G(\mathbf{z})$ means the output from G, i.e., the generated fraud data, with random data as input, $D(\mathbf{x})$ is the output of D, which denotes the probability of the input $\mathbf{x}$ being real data. Due to instability of the training stage and randomness of the generated images, the original GANs topology is not suitable for image in-painting tasks [17]. Radford et al. improved the network structure of GANs in [19] to enhance the stability of the original GANs during training and the quality of the results. Compared to the original GANs, the improvements in DCGANs are mainly:

- Using a transposed convolutional layer in G instead of the up-sampling layer.
- Removing all the pooling layers. In the D network, the pooling is replaced by strided convolution.
- Using batch normalization in both G and D.

- Removing the fully-connected layer to turn the network into a fully convolutional network.
- Using mostly ReLU as the activation function in G and tanh as the activation function of the last layer.
- Using LeakyReLU as the activation function in D.

B. Semantic Image In-painting with DCGANs

Inspired by the strategy of back-propagation to the input data from [21], [22], and [23], Yeh et al. [17] proposed a specific loss function for the DCGANs topology for facial image completion. Yeh et al. assumed that an efficient G can generate images analogous to the occluded input, even if they are not from p_{data}. They try to find an encoding $\hat{z}$ that can produce an image closest to the occluded image, while $\hat{z}$ is constrained to the encoding manifold z learned by G [17]. Hence, in their work, G and D are trained using occlusion-free images before utilizing them to discover $\hat{z}$ with a novel loss function for in-paining.

1) Loss Function: In [17], the emphasis is on how loss terms work during the completion stage. The loss functions require knowledge about the missing part location. This is defined by a mask M, where the missing areas (occlusions) are denoted by the pixel value zero and other (non-occluded) areas by the pixel value one. In the following, we use I_O to represent the occlude image, I_{rec} to be the result image, $I_{G(\mathbf{z})}$ to denote an image generated by G from any input $\mathbf{z}$, $I_{G(\hat{\mathbf{z}})}$ to denote the final image generated by G using the encoding manifold $\hat{\mathbf{z}}$.

The loss function consists of two terms: contextual loss and prior loss.

- **Contextual Loss** $\mathcal{L}_{diff}$ indicates the difference of a generated image and an occluded images in the occlusion-free areas. In order to highlight the importance of pixels surrounding the missing part, a weighting term W is introduced on the basis of M in [17], given as:
$$W_i = \begin{cases} \sum\limits_{j \in N(i)} \frac{(1-M_j)}{|N(i)|}, if \ \ M_i \neq 0 \\ 0, \qquad\qquad\quad if \ \ M_i = 0 \end{cases}, \tag{2}$$
where W_i denotes the importance weight of pixel i, $N(i)$ represents a window around pixel i, and $|N(i)|$ is the cardinality of the window. Then, the contextual loss is defined as:
$$\mathcal{L}_{diff} = \|W \odot (I_{G(\mathbf{z})} - I_O)\|_1, \tag{3}$$
which aims to force the difference of occlusion-free areas in $I_{G_\mathbf{z}}$ and I_O to be zero, so that the occlusion-free final image $I_{G(\hat{\mathbf{z}})}$ can mimic the occluded image I_O. $\odot$ means element-wise multiplication.
- **Prior Loss** $\mathcal{L}_{D(\mathbf{z})}$ is the loss of the trained discriminator D, acting as penalty, defined as follows:
$$\mathcal{L}_{D(\mathbf{z})} = log(1 - D(G(\mathbf{z}))), \tag{4}$$
which leads the generated image to be as realistic as possible, until satisfying human visual experience.

Using the above two loss terms, the entire loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{diff} \ + \ \alpha\mathcal{L}_{D(\mathbf{z})}. \tag{5}$$

Here, a scaling factor α is used to balance the contribution of the two loss terms. The target of optimizing this loss function is to discover $\hat{\mathbf{z}}$, which can minimize $\mathcal{L}$:

$$\hat{\mathbf{z}} = arg \min_{\mathbf{z}} \mathcal{L}. \tag{6}$$

2) Facial Image Completion: Given the pre-trained DCGANs, an initial random input vector $\mathbf{z}$ is fed to G and the loss is iteratively back-propagated to $\mathbf{z}$ using Adam gradient descent algorithm [25], until obtaining an optimal $\hat{\mathbf{z}}$. At this point, the final in-painted image is obtained by combining $I_{G(\hat{\mathbf{z}})}$ and I_O as follows:

$$I_{rec} = (1 - M) \ \odot I_{G(\hat{\mathbf{z}})} \ + M \odot I_O. \tag{7}$$

Here, I_{rec} refers to a stacked image, which contains the occlusion-free pixels from the occluded input image and generated pixels segmented by M from $I_{G(\hat{\mathbf{z}})}$. To reduce the graininess at the edges of stacking, in [17], I_{rec} is finally processed by Poisson blending [24] to preserve image details, as in [16].

IV. Proposed Approach

Our approach has three main stages: training of DCGANs using occlusion-free images, generating the occlusion mask and the image used in completion, and merging the input and the generated image by exploiting the generated mask. Our main contribution lies in the generation stage. Instead of taking the (artificial) missing part mask as an input as in [17], our approach automatically generates the occlusion mask from an occluded image. This is achieved by introducing novel loss terms. No training samples with occlusion are needed. The overall approach is illustrated in Fig. 1 and the details of each stage are discussed in the following sub-sections.

A. DCGANs Architecture and Training Process

The DCGANs architecture and training process are adopted from [19] (as in [17]) with the exception of the layer depths. The structures of D and G follow the reverse order strategy. In total, there are eight convolutional layers. The filter sizes and layers depths are shown in Fig. 1. The loss function used in the training is the basic GANs loss given in Eq. (1).

B. Generation of Mask and Occlusion-free Image

When the DCGANs model have been trained, the generator G is used to generate an occlusion-free image that can be used in the in-painting process. As in [17], the main idea is to find an optimal input $\hat{\mathbf{z}}$ that produces an image similar to the occluded image, while keeping $\hat{\mathbf{z}}$ constrained to the encoding manifold learned by G, so that the discriminator D will find the generated image to have a natural appearance.

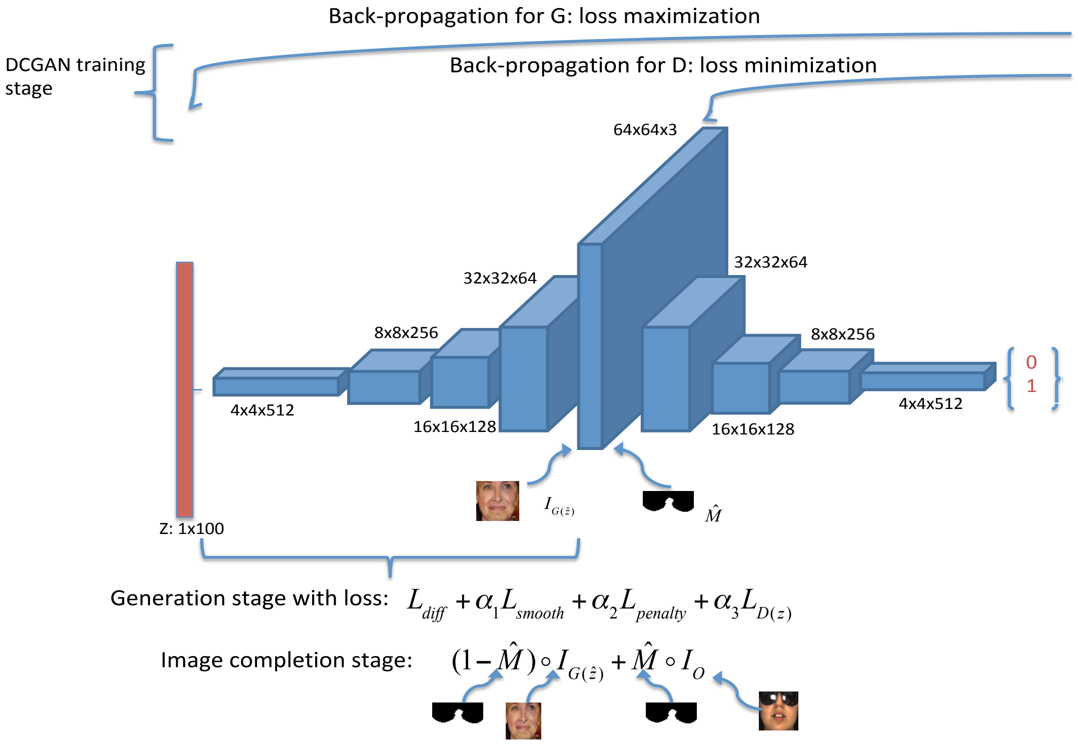

Back-propagation for G: loss maximization
DCGAN training stage
Back-propagation for D: loss minimization
64x64x3
32x32x64
32x32x64
8x8x256
8x8x256
4x4x512
16x16x128
16x16x128
4x4x512
0
1
$I_{G(\hat{z})}$
$\hat{M}$
Z: 1x100
Generation stage with loss: $L_{diff} + \alpha_1 L_{smooth} + \alpha_2 L_{penalty} + \alpha_3 L_{D(z)}$
Image completion stage: $(1-\hat{M}) \circ I_{G(\hat{z})} + \hat{M} \circ I_O$

Fig. 1. Work-flow of the proposed approach

In our work, we propose a novel loss function so that we can generate and optimize the mask at the same time. Our mask is a binary image M, where pixels with value zero denote occluded areas and pixels with value one denote the occlusion-free areas. With the occluded image as a reference, we iteratively update $\mathbf{z}$ and M through minimizing the loss function to discover $\hat{\mathbf{z}}$ generating an image similar to the occlusion-free areas of the occluded image and mask $\hat{M}$ matching with the occlusion in the occluded image.

We start the optimization process with a constant mask M and an input vector $\mathbf{z}$ with 100 dimensions randomly sampled from a uniform distribution in the range of $[-1, 1]$. We adopt Adam gradient descent algorithm [25] to discover $\hat{\mathbf{z}}$ and stochastic gradient descent (SGD) [26] to update the binary mask M. We iteratively update $\mathbf{z}$ and M, while the other is kept fixed. For each iteration, the updated $\mathbf{z}$ is limited to $[-1, 1]$ to ensure stability [17]. For M, we apply Morphological Filtering to eliminate edge noise caused by occlusion edge or similarity of pixels on the occluded area and far from it. We use two kinds of morphological filters: a closing filter to remove "pepper" noise and an erosion filter to eliminate occlusion edge noise. After filtering, we use a threshold value T to set pixel values larger than T to 1, otherwise to 0. The loss function optimized to update z and M is presented in Section IV-B-1.

The overall procedure in the generation phase is shown in Algorithm 1.

```
Inputs: trained DCGANs, $I_O$, initial $M$, $\mathbf{z}$, and $T$.
Outputs: $\hat{M}$ and $I_{G(\hat{\mathbf{z}})}$
for the number of iteration do:
    Feed the DCGANs with $\mathbf{z}$ to generate $I_{G(\mathbf{z})}$ and $\mathcal{L}_D$;
    Calculate the entire loss $\mathcal{L}$ with $I_O$, $I_{G(\mathbf{z})}$, and $M$;
    Update $\mathbf{z}$ using Adam optimizer on Eq. (10);
    Restrict values of $\mathbf{z}$ to [-1,1];
    Update $M$ using SGD on Eq. (10);
    Normalize $M$ to $[0, 1]$;
    Apply morphological filtering to $M$;
    Use threshold $T$ to obtain a binary $M$;
Set $\hat{\mathbf{z}} = \mathbf{z}$ and $\hat{M} = M$;
Feed $\hat{\mathbf{z}}$ to G to generate $I_{G(\hat{\mathbf{z}})}$;
```

Algorithm 1: Generation of an occlusion-free image $I_G(\hat{z})$ and a $\hat{M}$

1) Loss Function: Besides the contextual loss and prior loss mentioned in Section III, we utilize a smoothness term and an occlusion size penalty to generate a binary mask instead of using a pre-defined mask in [17]. In addition, we set $W = M$ directly.

- **Smoothness loss** $\mathcal{L}_{smooth}$ is designed to learn a smooth mask with the same size as the occluded image. It forces the occluded part in the mask toward a uniform value.

$$\mathcal{L}_{smooth} = \sum_{i}^{N_1} \sum_{j}^{N_2} \sum_{k}^{-1,1} \|x_{i,j} - x_{i+k,j+k}\|_2, \quad (8)$$

where $x_{i,j}$ refers to each pixel value of mask M. N_1 and

N_2 mean the number of pixels in rows/columns. This term measures the similarity of each pixel with its four neighbors.

- **Occlusion size loss** $\mathcal{L}_{penalty}$ penalizes large occlusion areas in the mask using ℓ_1-norm

$$\mathcal{L}_{penalty} = \sum_{i}^{N_1} \sum_{j}^{N_2} \|x_{i,j}\|_1, \tag{9}$$

This term is needed to avoid assigning all the pixels as occlusion. Otherwise, setting all M values to zero would be an easy way to minimize $\mathcal{L}_{diff}$.

The entire loss function is formed as:

$$\mathcal{L} = \mathcal{L}_{diff} \ + \alpha_1 \mathcal{L}_{smooth} \ + \alpha_2 \mathcal{L}_{penalty} \ + \alpha_3 \mathcal{L}_{D(\mathbf{z})} \tag{10}$$

The optimal input $\hat{\mathbf{z}}$ is solved as shown in Eq. (6) and, similarly, $\hat{M}$ as follows:

$$\hat{M} = arg \min_{M} \mathcal{L}. \tag{11}$$

C. Image Completion

After finding the optimized $\hat{\mathbf{z}}$ and $\hat{M}$, we generate the occlusion-free image $I_{G(\hat{\mathbf{z}})}$, which should be similar to the occluded image I_O. In the final stage, $I_{G(\hat{\mathbf{z}})}$ and I_O are merged using $\hat{M}$ according to Eq. (7). In this work, we do not use the Poisson blending applied in [17] to clearly observe the ability of the proposed method for facial image completion.

V. Experiments

A. Datasets

In the DCGANs training stage, we use the dataset Celeb-Faces Attributes Datasets (CelebA) [27]. Before training, each image is aligned using OpenFace [28] to ensure the size of 64 × 64 pixels. When training DCGANs, we removed any images with sunglasses in CelebA, because we consider sunglasses as occlusion and do not want our G to generate images with sunglasses.

We apply the in-painting algorithm on AR Face Database [29], because it contains an adequate number of occluded images. Furthermore, we randomly select several frontal facial images with sunglasses/covers from CelebA or e-commerce web-page as eBay. All occluded images for in-painting are aligned using OpenFace [28] and resized to 64 × 64 pixels, also.

B. Parameters

In the generation stage, we set the loss terms' scaling to $\alpha_1 = 1$, $\alpha_2 = 5$, and $\alpha_3 = 0.1$. Threshold T is set to be 0.7. The parameters of creating a circular structure for both morphological filters are set to be 1 pixel [30]. We use 25 epochs to train the DCGANs and 1000 iterations in the generation stage for each occluded image.

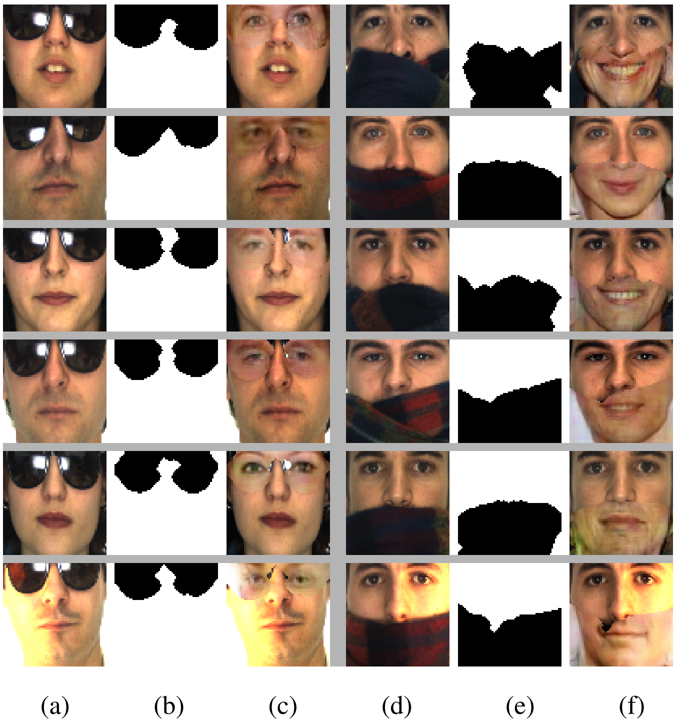


Fig. 2. In-painting results from AR dataset. Columns (a), (d) represent occluded images, columns (b), (e) show learned mask, and columns (c), (f) are result images.

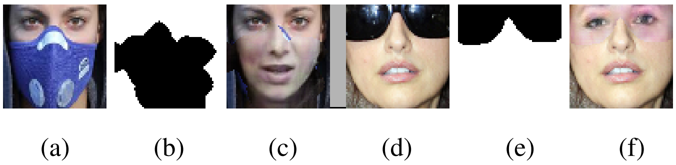


Fig. 3. In-paining results from web images

C. Experimental Results

We demonstrate the results of this work using four figures: results from facial images are shown in Fig. 2 and 3. Fig. 4 illustrates the process of mask generation from different iterations sequentially. Fig. 4(a) - (c) and Fig. 4(d) - (f) demonstrate the mask generation of two input images separately. Fig. 5 shows a few cases, where the proposed method was not able to predict the real occluded area.

Our method is an unsupervised approach, so it attempts to reconstruct realistic faces semantically, instead of restoring the ground-truth images from its occluded version. In fact, there is no ground-truth, but the quality of the results depends on the observer's subjective opinion on the image credibility. Therefore, we do not provide any numerical results here.

As shown in Fig. 2, 3, and 5, the results are influenced by illumination. Strong illumination either leads to a blurry result or a mask not corresponding with the real occlusion. The noise around the fringe of a mask also influences the final result. As shown in Fig. 3(c), the final image obviously contains a purple hue on the lower half due to the remaining occlusion pixels on the upper half. Moreover, the light spots on the sunglasses also seriously affect the correctness of mask as in Fig. 2(b) and Fig. 5(b).

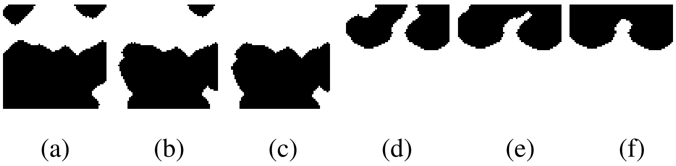


Fig. 4. Examples of mask optimization process

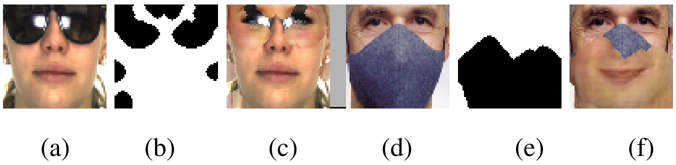


Fig. 5. Failure examples

VI. Conclusions and Discussion

In this work, we proposed a novel approach to learn a mask depicting occlusion and reconstruct an occlusion-free image semantically using trained DCGANs in an unsupervised manner. Our proposed method relates to both facial image de-occlusion and image in-painting. We apply this work on diverse occluded images and obtain proper results. In future, we wish to further improve the loss function to obtain more precise occlusion mask, in addition to enhancing the quality of the final image. The extension could detect more dynamic occlusions in wild, reduce the impact of illumination, and produce high-resolution results. Furthermore, we will consider how to evaluate the result quality in a more systematic manner. Possibly a separate DCGANs structure can be used to evaluate the credibility of the resulting images.

Acknowledgment

This work was supported by the NSF-TEKES Center for Visual and Decision Informatics project (CVDI) sponsored by Nokia.

References

[1] T. Hosoi, S. Nagashima, K. Kobayashi, K. Ito, and T. Aoki. Restoring Occluded Regions using FW-PCA for Face Recognition. In *IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops*, proceedings, Providence-USA, June 2012.

[2] Y. Saito, Y. Kenmochi, and K. Kotani. Estimation of Eyeglassless Facial Images using Principal Component Analysis. In *IEEE Conf. Image Processing*, proceedings, Kobe-Japan, October 1999.

[3] R. Min and J. Dugelay. Inpainting of Sparse Occlusion in Face Recognition. In *IEEE Conf. Image Processing*, proceedings, Orlando-USA, October 2012.

[4] M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust Sparse Coding for Face Recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, proceedings, Colorado Springs-USA, June 2011.

[5] L. Ma, C. Wang, B. Xiao, and W. Zhou. Sparse Representation for Face Recognition based on Discriminative Low-rand Dictionary Learning. In *IEEE Conf. on Computer Vision and Pattern Recognition*, proceedings, Washington, D.C.-USA, June 2012.

[6] Y. Zhang, R. Liu, S. Zhang, and M. Zhu. Occlusion-Robust Face Recognition Using Iterative Stacked Denoising Autoencoder. In *Int. Conf. on Neural Information Processing*, proceedings, Daegu-Korea, November 2013.

[7] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan. Robust LSTM-Autoencoders for Face De-Occlusion in the Wild. *IEEE Trans. on Image Processing*, 27(2): 778–790, 2018.

[8] J. Zhang, M. Kan, S. Shan, and X. Chen. Occlusion-free Face Alignment: Deep Regression Networks Coupled with De-corrupt AutoEncoders. In *IEEE Conf. on Computer Vision and Pattern Recognition*, proceedings, Las Vegas-USA, June 2012.

[9] W. Leow, G. Li, J. Lai, T. Sim, and V. Sharma. Hide and Seek: Uncovering Facial Occlusion with Variable-Threshold Robust PCA. In *IEEE Winter Conf. on Applications of Computer Vision*, Lake Placid-USA, March 2016.

[10] Y. Li, S. Liu, J. Yang, and M. Yang. Generative Face Completion. In *IEEE Conf. on Computer Vision and Pattern Recognition*, proceedings, Hawaii-USA, July 2017.

[11] L. Cheng, J. Wang, Y. Gong, and Q. Hou. Robust Deep Auto-encoder for Occluded Face Recognition. In *ACM Int. Conf. on multimedia*, proceedings, Brisbane-Australia, October 2015.

[12] A. Lahiri, A. Jain, P. Biswas, and P. Mitra. Improving Consistency and Correctness of Sequence Inpainting using Semantically Guided Generative Adversarial Network. *arXiv preprint arXiv:1711.06106*, 2017.

[13] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image Inpainting. In *the 27th annual conf. on Computer graphics and interactive techniques*, proceedings, 2000.

[14] J. Sun, L. Yuan, J. Jia, and H. Shunm. Image Completion with Structure Propagation. ACM Trans. on Graphics, 24(3): 861–868, 2005.

[15] U. Demir and G. Unal. Patch-Based Image Inpainting with Generative Adversarial Networks. *arXiv preprint arXiv:1803.07422v1*, 2018.

[16] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. Efros. Context Encoders: Feature Learning by Inpainting. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, proceedings, 2016.

[17] R. Yeh, C. Chen, T. Lim, A. Schwing, M. Hasegawa-Johnson, and M. Do. Semantic Image Inpainting with Deep Generative Models. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, proceedings, Hawaii-USA, July 2017.

[18] S. Iizuka, E. Simo-serra, and H. Ishikawa. Globally and Locally Consistent Image Completion. *ACM Trans. on Graphics*, 36(4):107:1–107:14, 2017.

[19] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *Int. Conf. on Learning Representations (ICLR)*, proceedings, 2016.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. In *Int. Conf. on Neural Information Processing Systems (NIPS)*, proceedings, 2014.

[21] L. A. Gatys, A. S. Ecker, and M. Bethge. Image Style Transfer Using Convolutional Neural Networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, proceedings, June 2016.

[22] L. Gatys, A. S. Ecker, and M. Bethge. Texture Synthesis using Convolutional Neural Networks. In *Int. Conf. on Neural Information Processing Systems (NIPS)*, proceedings, 2015

[23] C. Li and M. Wand. Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, proceedings, June 2016.

[24] P. Pérez, M. Gangnet, and A. Blake Poisson Image Editing. *ACM Trans. on Graphics*, 22(3): 313–318, 2003.

[25] D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *Int. Conf. on Learning Representations (ICLR)*, proceedings, 2015.

[26] L. Bottou. Large-scale Machine Learning with Stochastic Gradient Descent. In *COMPSTAT*, proceedings, 2010.

[27] S. Yang, P. Luo, C. C. Loy, and X. Tang. From Facial Parts Responses to Face Detection: A Deep Learning Approach. In *IEEE Int. Conf. on Computer Vision (ICCV)*, proceedings, 2015.

[28] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. OpenFace: A General-purpose Face Recognition. *CMU-CS-16-118, CMU School of Computer Science, Tech. Rep.*, 2016.

[29] A. Martinez and R. Benavente. The AR Face Database. *CVC Technical Report #24*, June 1998.

[30] N. Efford. Digital Image Processing: A Practical Introduction Using JavaTM. *Pearson Education*, Chapter 11, 2000.

PUBLICATION
IV

Revisiting generative adversarial networks for binary semantic segmentation on imbalanced datasets

L. Xu and M. Gabbouj

arXiv preprint arXiv:2402.02245