Tampereen yliopisto

Lauri Lehtonen

IMPLICIT NEURAL REPRESENTATIONS FOR NON-BLIND DEPTH-AWARE IMAGE DEBLURRING

Bachelor's thesis
Faculty of Information Technology and Communication Sciences
Examiner: Felipe Torres
May 2024

ABSTRACT

Lauri Lehtonen: Implicit Neural Representations For Non-blind Depth-aware Image Deblurring
Bachelor's Thesis
Tampere University
Bachelor's Programme in Computing and Electrical Engineering
May 2024

The purpose of this thesis is to evaluate different implicit neural representation architectures in a self-supervised framework for image deblurring, where depth information is available. Image deblurring is the task of recovering sharp details from a blurry image. Classical methods for image deblurring often rely on computationally-expensive iterative algorithms that perform poorly on natural images, while fully-supervised deep learning approaches require large-scale datasets for training. Implicit neural representations have emerged as a powerful tool to represent multidimensional signals through neural networks. By coupling an implicit neural representation with a differentiable blur formation model, deep-learning optimization methods can be used to learn a neural representation that produces a sharp image, using only the blurry input for supervision. In particular, this thesis explores two computationally efficient hybrid implicit neural representations: 1) Instant Neural Graphic Primitives, and 2) Dictionary fields.

The thesis is divided into 3 parts, First going through the background theory behind image blur, image deblurring and implicit neural representations. Secondly an introduction of the hybrid implicit neural representations used for image deblurring in this thesis. Lastly the results obtained from the experiments will be analyzed and discussed.

The research done shows impressive results and promising possibilities for further optimization. The results gained from the state-of-the-art methods used have shown to be computationally more efficient and have been able to produce superior quality to the baseline models used as comparisons, outperforming them both in training time and result quality.

Keywords: Implicit Neural Representations, Image Deblurring, Machine Learning

TIIVISTELMÄ

Lauri Lehtonen: implisiittiset neuroesitykset ei-sokealle syvyystietoiselle kuvan sumennuksen poistamiselle
Kandidaatintyö
Tampereen yliopisto
Tieto- ja sähkötekniikan kandidaattiohjelma
Toukokuu 2024

Tämän työn tarkoituksena on arvioida erilaisia implisiittisiä neuraaliesitysarkkitehtuureja itseohjautuvassa kehyksessä kuvan sumeuden poistoa varten, kun syvyystieto on saatavilla. Kuvan sumeuden poisto on tehtävä, jossa tarkat yksityiskohdat palautetaan epätarkasta kuvasta. Klassiset kuvien sumeudenpoistomenetelmät perustuvat usein laskennallisesti kalliisiin iteratiivisiin algoritmeihin, jotka toimivat huonosti luonnollisissa kuvissa, kun taas täysin valvotut syväoppimismenetelmät vaativat koulutukseen laajoja data-aineistoja. Implisiittiset neuraaliesitykset ovat nousseet tehokkaaksi välineeksi moniulotteisten signaalien esittämiseen neuroverkkojen avulla. Yhdistämällä implisiittiset neuraaliesitykset differentoituvaan sumeuden muodostusmalliin voidaan syväoppimisen optimointimenetelmiä käyttää oppimaan neuraaliesitys, joka tuottaa terävän kuvan, käyttäen vain epätarkkaa syötettä koulutuksen valvontaan. Tässä tutkielmassa tutkitaan erityisesti kahta laskennallisesti tehokasta hybridi-implisiittistä neuraalista edustusta: 1) Instant Neural Graphic Primitives ja 2) Dictionary Fields.

työ on jaettu kolmeen osaan: ensin käydään läpi taustateoriaa kuvan sumeuden, kuvan sumeuden poiston ja implisiittisten neuraaliesitysten takana. Toiseksi esitellään hybridi-implisiittiset neuraaliesitykset, joita käytetään tässä työssä kuvan sumeuden poistoon. Lopuksi analysoidaan ja käsitellään kokeista saatuja tuloksia.

Tehty tutkimus osoittaa vaikuttavia tuloksia ja lupaavia mahdollisuuksia optimoinnin jatkamiseen. Käytetyillä uusimmilla menetelmillä saadut tulokset ovat osoittautuneet laskennallisesti tehokkaammiksi, ja ne ovat pystyneet tuottamaan parempaa laatua kuin vertailukohtana käytetyt perusmallit, jotka ovat olleet niitä parempia sekä koulutusajassa että tulosten laadussa.

Avainsanat: Implisiittiset Neuraaliesitykset, Kuvan Sumeuden Poisto, Koneoppiminen

TABLE OF CONTENTS

1. INTRODUCTION 1
2. BACKGROUND 3
    2.1 Image blur 3
    2.2 Blur formation models caused by camera shake 3
        2.2.1 Uniform blur kernel 3
        2.2.2 Pixel-wise blur (PWB) model 4
        2.2.3 Parallax image compositing blur (ICB) model 4
    2.3 Implicit Neural Representations 5
        2.3.1 sinusoidal representation networks (SIREN) 5
        2.3.2 Fourier features 6
    2.4 Image deblurring 6
        2.4.1 Classical image deblurring 6
        2.4.2 Supervised deep-learning deblurring 7
        2.4.3 Self-supervised deblurring through implicit neural representations 7
3. METHODS 9
    3.1 Instant Neural Graphic Primitives (InstantNGP) 9
    3.2 Dictionary fields (DiF) 10
    3.3 CNN for model reguralization 12
4. EXPERIMENTS 14
    4.1 Evaluation datasets 14
    4.2 InstantNGP 15
    4.3 DiF 15
    4.4 Comparisons of INRs 16
    4.5 Analysis of the CNN prior regularization 18
5. CONCLUSION 20
REFERENCES 21

ABBREVIATIONS AND NOTATIONS

3D	Three Dimensional
CMB	Camera Motion Blur
CPU	Central Processing Unit
CNN	Convolutional Neural Network
CUDA	Compute Unified Device Architecture
DiF	Dictionary Field
GB	Gigabyte
GPU	Graphics Processing Unit
ICB	Image Composition Blur
INR	Implicit Neural Representation
LPIPS	Learned Perceptual Image Patch Similarity
MAE	Mean Absolute Error
MB	Megabyte
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Square Error
NeRF	Neural Radiance Field
NGP	Neural Graphics Primitives
NN	Neural Network
PC	Personal Computer
PSNR	Peak Signal-To-Noise Ratio
RAM	Random Access Memory
ReLU	Rectified Linear Unit
SIREN	Sinusoidal Representation Networks
SSIM	Structural Similarity Index Measure
TV	Total Variation

α	Learning rate
s	second
b	Blurry image
i	Sharp image
x	image coordinates
K	Blur Kernel
$*$	Convolution Operator

1. INTRODUCTION

Image deblurring is the process of recovering a sharp image from an image that has been distorted due to either camera motion or the movement of the objects within the scene while the device captured the image. Its primary goal is to enhance the sharpness and intricate details of an image, making it suitable for aesthetic purposes in consumer-grade cameras. Additionally, it serves as preprocessing step in advanced computer vision tasks, e.g., object detection [1] and feature matching [2], amplifying its utility beyond mere visualization. Therefore, image deblurring is itself an important research subject that remains as an open problem, despite considerable efforts over time.

Classical methods to image deblurring are formulated as an optimization problem which involves: 1) a model that describes the blur formation, and 2) an image prior that regularizes the solution space, essential due to the ill-pose nature of the deblurring task. When the parameters of the blur model are known in advance, the problem falls under non-blind image deblurring, otherwise the problem is referred as blind deblurring since it requires a joint optimization of the blur formation parameters and the latent sharp image. Classical approaches are often characterized by their time-consuming and computationally intensive nature since they require the execution of the entire optimization process for each individual image. Furthermore, traditional image priors may inadequately capture the true essence of natural images, thereby resulting in artifacts and distortions.

Nowadays deep learning has become the primary method for image deblurring, offering results with higher image quality [3]. At its core, a neural network architecture learns an end-to-end mapping from blurry input images to sharp ones through supervised training using gradient-descent methods over large-scale datasets. Nevertheless, fully- supervised deep learning approaches still raise some issues such as the need for datasets containing a high volume of images and substantial computational cost and training time.

More recently, Torres and Kämäräinen [4] introduced a self-supervised method to produce a sharp image from a single blurry image. In principle, they use the same formulation of a classical deblurring approach, including a blur formation model (in a non-blind stetting) and a data prior within an optimization problem. Notwithstanding, instead of deriving an iterative algorithm to progressively remove the blur as in classical deblurring, they parametrized the latent sharp image through a primitive implicit neural representation [5]. By utilizing deep-learning optimization, they train this neural representation, such that it produces a sharp image whose blur formation output is consistent with the blurry input. Their results demonstrate competitive results compared to supervised deep-learning approaches, but the usage of primitive implicit representations might still pose limitations in terms of computational time.

The purpose of this thesis is to integrate and evaluate two state-of-the-art hybrid implicit representations in the self-supervised approach of Torres and Kämäräinen [4]. 1) Instant Neural Graphic Primitives (InstantNGP) integrates a compact neural network with a multiresolution hash table of feature vectors. 2) Dictionary Fields (DiF) is a two-factor representation comprised of basis (with periodic transformation) and coefficient fields to effectively capture both global and local signal properties.

Moreover, the usage of the deep image prior [6] as a regularization technique is further investigated in this thesis. In work of Torres and Kämäräinen [4], the TV regularization is used to encourage fewer sharp transitions in the latent image. However, such regularization yields to ringing artifacts that harm the final result. By using the deep image prior technique, i.e., by stacking a CNN on top the implicit representation, the CNN provides structure to the data such that it produces more visually appealing result.

In summary, the contributions of the thesis are:
1) An evaluation of InstantNGP and DiF as powerful implicit representations that are integrated in the deblurring approach of Torres and Kämäräinen. Indeed, such representations yield to a substantial decrease in training time, while achieving on-par performance in terms of accuracy.
2) The integration of the deep image prior in the deblurring approach, resulting in a more visually pleasing outcome.

2. BACKGROUND

2.1 Image blur

Image blur refers to the loss of sharpness and detail in an image, which occurs when the lines between objects in an image become less discernible. Several factors contribute to the formation of motion blur, for instance motion blur can occur if the camera or subject moves during the image taking process, if the camera lens focuses on the wrong point there can out-of-focus blur and if the light rays converge incorrectly at the camera lens there may be aberrations in the image, leading to blur.

Image blur is often undesired in consumer-grade applications or high-level computer vision tasks. However, in some instances it may be desirable to make the image more visually appealing, when the background of an image is blurred the subject will pop out, or motion blur can be introduced to an image to create an illusion of motion in the image. In addition, a long exposure time during shooting introduces motion blur, crafting the illusion of movement within the image [7].

In this thesis, blur resulting from camera movement during sensor exposure is addressed, while the scene remains static and exhibiting pronounced 3D structure. Consequently, information about the scene depth is crucial for accurately reproducing real blur.

2.2 Blur formation models caused by camera shake

Various mathematical models for the generation of blur have been explored in the literature. These models are primarily used in classical optimization methods for image deblurring but can also be applied to create blur filters in editing software for artistic purposes or in data augmentation pipelines for advanced computer vision tasks [8]. This section details some models capable of generating realistic image blur.

2.2.1 Uniform blur kernel

Non-blind deconvolution is used when the blur kernel or point spread function is known and this process can be expressed with:

$$\boldsymbol{b} = \boldsymbol{i} * \boldsymbol{K} \tag{1}$$

where b is the blurred image, I is the sharp image and K is the point spread function or blur kernel. generally when working with non-blind deconvolution the problem is to address reducing ringing artifacts or other remaining artifacting, or reducing the computational cost of the deblurring algorithm.

Blind deconvolution is when neither the blur kernel or un-blurred image is known, and can be expressed as:

$$\boldsymbol{b} = \boldsymbol{i} * \boldsymbol{K} + \boldsymbol{n} \tag{2}$$

where n is additive noise and b, I and f represent the same as in non-blind image deconvolution. Clasically the strategy has been to estimate the blur point spread function and the un-blurred image seperately, resulting in alternating optimization [9].

2.2.2 Pixel-wise blur (PWB) model

Image blur generally has a spatially-varying nature, to take this into account the PWB models image blur as:

$$b(x) = i(x) * k(x,u) + \eta \tag{3}$$

where the blurred image b is modeled via convolutions with pixel-wise kernels $k(x,u)$, with $x = (i,j)$ are the pixel coordinates and $u = (i,j)$ are the kernel coordinates and η is the error of the model [4]. Although the PWB model captures more realistic blur, it is memory-intensive because it requires a blur kernel for each pixel. In fact, the memory needed to fit the motion kernels increases quadratically with the image size.

2.2.3 Parallax image compositing blur (ICB) model

The parallax ICB model, introduced by Torres and Kämäräinen [4], is an image blur model that incorporates depth. The model is defined as:

$$b = \sum_{l=0}^{L-1} (i * k_i) \cdot A_l + \eta \tag{4}$$

Where $\{A_l\}_{l=0}^{L-1}$ and $\{k_l\}_{l=0}^{L-1}$ are alpha-matting terms and blur kernels respectively, and "·" is pixel-wise multiplication, η is the error of the model, i is the full depth map of the latent image, b is the resulting blurred image.

Compared to PWB the parallaxICB model is better able to trace generated blur over the depth discontinunuities and merges blur from different depth layers more realistically leading to more natural blur while obtaining significant improvements in term of memory consumption [4].

2.3 Implicit Neural Representations

An Implicit Neural Representation (INR) aims to model multidimentional signals as a continuous function instead of discrete values. INRs are beneficial across various machine learning and signal processing subfields because, unlike discrete signals, which are tied to a specific spatial resolution and contain limited information, the memory required for these representations depends only on the complexity of the signal.

Vanilla INRs for signal fitting comprises a Neural Network (NN) Φ_θ that is used to estimate the functional F for from discrete samples I. This NN learns a class of functions Φ_θ which satisfies the functional:

$$F(x, \Phi) = \Phi(x) - I(x) = 0, for\ all\ x \in \Omega \tag{5}$$

Once this function is learned, the NN can produce the signal values by evaluating the NN at each signal coordinate. Effectively, this enables the NN to handle signal resolution agnostically, scaling the storage requirements based on the complexity of the signal rather than the spatial resolution, and benefits from the use of well-established gradient-based optimization [5][10][11].

2.3.1 sinusoidal representation networks (SIREN)

SIREN refers to a NN architecture for INR with sinusoidal activation functions of the form:

$$\Phi_\theta(x) = W_n(\phi_{n-1} \circ \phi_{n-2} \circ \ldots \circ \phi_0)(x) + b_n, \quad x_i \mapsto \phi_i(x_i) = \sin(W_i x_i + b_i) \tag{6}$$

Where $\phi_i : \mathbb{R}^{M_i} \mapsto \mathbb{R}^{N_i}$ is the n[th] layer of the network, $\{W_i\}_{i=0}^{n}$ are the weight matrices and $\{b\}_{i=0}^{n}$ are the scalar biases that correspond to the NN parameters θ, i.e., $\theta = \{W_0, \dots, W_n, b_0, \dots, b_{n\backslash}\}$. SIRENs excel at representing complex natural signals and they can be used to represent images, audio, video, NeRFs and solving many of the common problems in machine learning such as the Poisson equation, wave equation and Helmholtz equation. An interesting note on SIRENs is that their derivatives are also SIRENs as the derivative of the sine function is a cosine which is a phase shifted sine, this property of the SIREN networks allows for supervising of SIRENs derivatives. Furthermore, SIREN converges faster than baseline NN models, allowing for faster operations [5][12].

2.3.2 Fourier features

Fourier feature Networks correspond to another family of NN for INR. Specifically, they consists of a regular MLP combined with a Fourier-alike transformation that is applied to the input coordinates x:

$$\Phi_\theta(x) = MLP_\theta\big(\gamma(x)\big), \qquad \gamma(x) = [\cos(2\pi Bx), \sin(2\pi Bx)] \tag{7}$$

where B is sampled from a Gaussian matrix $\mathcal{N}(0, \sigma^2)$, and θ denotes the parameters of the NN. Interestingly, the Fourier transformation allows the MLP to learn the high-frequency details of the signal, whose frequency range is controlled by σ. Additionally, this transformation enables faster convergence [10].

2.4 Image deblurring

Image deblurrring is the process of removing blur from an image, making it sharper and clearer. This section discusses various approaches to address image deblurring

2.4.1 Classical image deblurring

Classical image deblurring methods are formulated as an optimization problem, comprising a data fitting term, and a regularization term. The regularization term incorporates a blur formation model (described in Section 2.2) to ensure that the optimized sharp image is consistent with the blurry measurement. The regularization term constrains the solution space to promote solutions that have a natural appearance.

Classical methods are further divided into blind and non-blind deblurring. When the parameters of the blur model, such as the kernel k (Eq. 2) in the uniform blur kernel model, are assumed to be known, the problem is termed non-blind image deblurring. Conversely, if the blur model parameters are unknown in advance, which is typically the case, the problem is referred as non-blind image deblurring [9]. This requires to estimate both, the latent sharp image and the blur model parameters, resulting in a complex and computationally intensive iterative process to resolve the unknowns.

Assuming non-blind settings and the uniform blur kernel model (Section 2.2.1), the optimization problem can be formulated as follows:

$$min_I\{i * f - b\}^2 + \lambda\Psi(I) \quad (8)$$

where i is the latent sharp image, f is the point spread function, b is the blurred image and $\lambda\Psi(I)$ is a weighted regularization term.

2.4.2 Supervised deep-learning deblurring

Deep learning methods have been used for image deblurring with great efficiency. A majority of deep deblurring architectures follow an encoder-decoder structure. In these models, the encoder extracts features from the blurred image, and the decoder reconstructs a sharp image from these features. Some models, like DeblurGANV2 [13], employ an adversarial loss to produce images that appear more pleasant perceptually. Generally deep deblurring methods primarily use MSE or MAE losses between the predicted sharp image and the ground truth [14][15].

Training these models is done using supervised learning, where gradient descent is used to adjust the weights of the network. minimizing loss and making the produced image more similar to the ground truth image. A very large dataset of paired blurry and sharp images is needed to produce decent result with these methods making them require large amounts of storage and computational power [16].

2.4.3 Self-supervised deblurring through implicit neural representations

This approach represents a combination of classical image deblurring and supervised deep-learning deblurring. Overall, this approach adopts formulation of classical image

deblurring that allows for the sharp estimation without any groundtruth supervision. However, rather than directly optimizing for the latent sharp image, the latent is parametrized through an INR, described in section 2.3. Such INR is comprised of a NN which can be optimized through gradient-descent methods. Therefore, this approach borrows the computational parametrization and optimization techniques of deep-learning to solve for the latent sharp image. Specifically, the weights θ of the INR Φ_θ are optimized so that Φ_θ learns to produce the colors of the sharp image at pixel coordinate p. These weights θ are optimized by minimizing the loss function [4]:

$$\mathcal{L} = \sum_{p} \left|\left| f\left(\Phi_\theta(\mathrm{x})\right) - b(\mathrm{x}) \right|\right|_2^2 + \lambda \left|\left| |\nabla_P \Phi_\theta(x)| \right|\right|_1^1 \quad (9)$$

Where f is the blur function, Φ_θ is a coordinate based MLP that optimizes its parameters θ to fit a sharp image, λ is a hyperparameter that controls how smooth the gradient is and the second term is a TV-like regularization term that promotes the preservation of sharp edges in the deblurring process.

3. METHODS

In this work, non-blind camera-motion image deblurring is addressed using the self-supervised approach outlined in Chapter 2.4.3. The parallax ICB model, discussed in Chapter 2.2.3, is used for blur formation, assuming known depth and camera trajectory. This approach, initially explored by Torres and Kamarainen [4], employs SIREN to parameterize the latent sharp image. While SIREN offers greater accuracy and efficiency than a plain MLP for image fitting, it remains computationally expensive, particularly for high-resolution images.

This chapter introduces two hybrid Implicit Neural Representations (INRs) that are integrated into this thesis to improve the results obtained with SIRENs. These INRs are Instant Neural Graphic Primitives (InstantNGP) and Dictionary fields (DiF), detailed in Sections 3.1 and 3.2, respectively. Additionally, an alternative regularization method using a CNN prior is presented in Section 3.3.

3.1 Instant Neural Graphic Primitives (InstantNGP)

InstantNGP is a hybrid INR that has been proposed to fit multidimensional signals such as images, Signed Distance Functions (SDFs) and NeRFs [11]. At its core, InstantNGP uses a multiresolution hash encoding method. This method encodes the input coordinates with a hashing function which leads to significant improvements in training time and computational efficiency. [17].

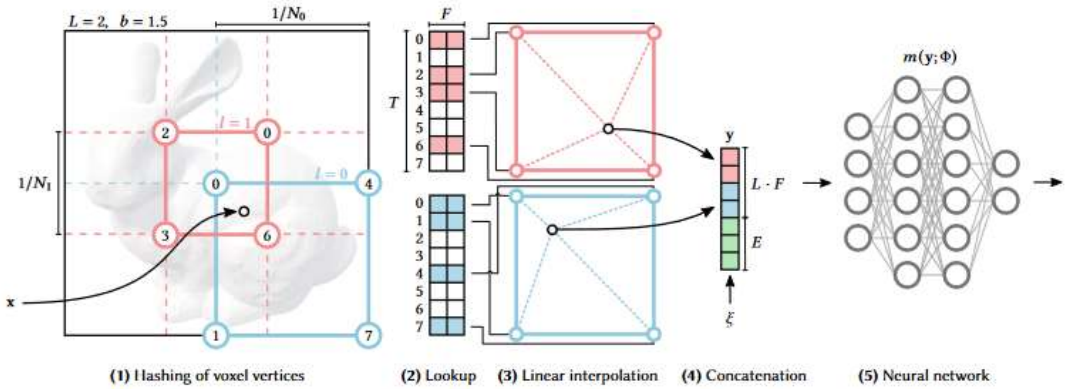

Fig 3.1. Process for multiresolutional hash encoding, obtained from source [17]

The method Operates in several steps:

1) It finds the surrounding grids at L resolution levels and assign indices to their corners by hashing their coordinates. the resolution for the grids is defined as the intermediate resolutions between the coarsest(N_{min}) and finest(N_{max}) levels obtained by geometric progression with the functions:

$$N_1 \coloneqq \lfloor N_{min} * b^l \rfloor,$$

(10)

$$b \coloneqq \exp\left(\frac{lnN_{max} - lnN_{min}}{L-1}\right) \rightarrow \left(\frac{N_{\max}}{N_{\min}}\right)^{\frac{1}{L-1}}$$

(11)

2) For each indice a corresponding F-dimensional, in the case of images F is 2, feature vectors from hash tables. the spatial hash function used for hashing coordinates to feature vectors is defined as:

$$h(x) = \left(\bigoplus_{i=0}^{d} x_i \pi_i \right) \ mod\ T$$

(12)

Where $\oplus$ denotes the bit-wise XOR operator and π_i are unique large prime numbers, is used to map each 2 dimensional vertex to on of the feature vectors of the respective level, giving the method computational complexity of O(1) for the lookup, as it functions similarly to a hash table, altough having no collision handling.

3) for each level, vectors can be linearly interpolated to get the final layer representing vector.

4) Vectors are concatenated for each layer and placed as input for the MLP, which produces the signal estimate at the point.

3.2 Dictionary fields (DiF)

DiF models a signal using a two-factor representation. DiF consists of 1) a coefficient field to express localized spatially-varying features, and 2) a basis field with periodic transformation to model commonalities of shared patterns over the whole signal domain. On top of the two-factor representatoin, a learned projection function, in the form of an MLP, is then used to obtain the target signal via regression of the factor product.

The modeling of two factors allow for efficient representation of both global and local properties of a signal, leading to superior quality as well as fast and compact reconstruction of the signal. DiF are expressed as:

$$\Phi_\theta(x) = \mathrm{MLP}_\theta(C(x) \circ B(t(x)) \quad (13)$$

where Φ_θ is the output signal, MLP_θ is the learned projection function modeled by a shallow MLP with parameters θ. The projection function maps the K-dimensional Hadamant product C(x) ∘ B(γ(x)) to the Q-dimensional target signal. C(x) and B(γ(x)) are the field representations for coefficients and basises and t(x) is a coordinate transform function. $t : \mathbb{R}^D \to \mathbb{R}$

Note that t(x) is a deterministic function whereas $\mathcal{P}$, C(x) and B(t(x)) are parametric mappings.

The field representations C(x) and B(t(x) can be represented by multiple different representations, such as MLPs. This reduces the memory requirements of the representation and induce a smoothness bias, but are slower to train. 3D voxel grids are faster to optimize but require more memory. Other representations available are Polynomial, 1D vectors and 2D maps shown in Fig. 3.3.

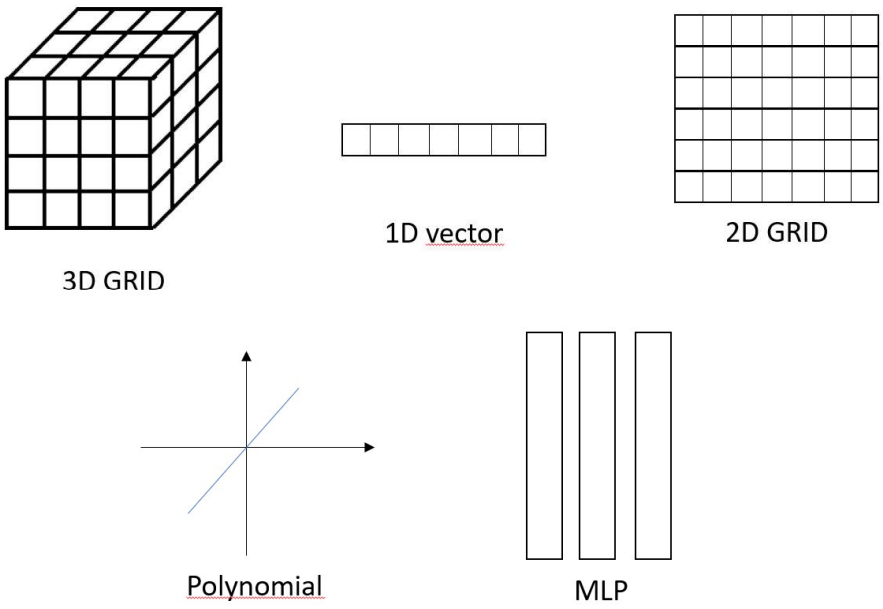

Fig 3.2. Basis and Coefficient representations [12]

The coordinate transform function γ is used to transform the input basis field. γ can be either periodic, as shown in figure 3.4, or non-periodic transform function such as hashing or orthogonal transforms. Periodic transformations allow applying the same basis at multiple spatial resolutions, which is important as signals often contain high and low frequency components.

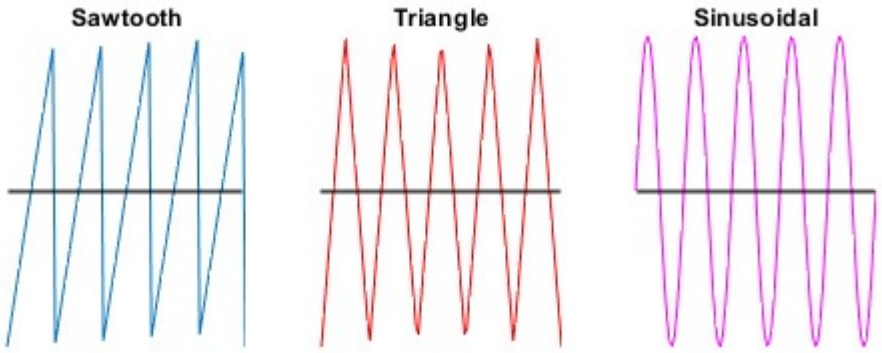

Fig 3.3. Various periodic coordinate transformation that can be used as t(x) [12]

The DiF model leads to state-of-the-art reconstruction results in most multidimensional signal, while being faster and more compact than most of the other state-of-the-art methods [10].

3.3 CNN for model reguralization

The self-supervised approach for non-blind image deblurring explored in this thesis utilizes a regularization term to constrain the solution space. The original work of Torres and Kamarainen employs a TV-like regularization to promote sharp edges and uniform regions, as outlined in eq. (6). Nevertheless, such regularization technique may not truly capture the statistics of natural images and can lead to artifacts.

In contrast, this thesis also evaluates the CNN prior proposed by Ulyanov et al. [6], which leverages randomly initialized CNNs to capture the low-level statistics of natural images. These statistics are sufficient to model conditional image distributions that arise in image restoration problems. In our specific non-blind deblurring scenario, the additional regularization is dropped from eq. (9), and a randomly initialized CNN is fed with the INR that parametrizes the sharp image. Thus the loss function in eq. (9) becomes:

$$\mathcal{L} = \sum_{p} \left|\left| f\left(\Phi_{\theta}(\mathrm{x})\right) - b(\mathrm{x}) \right|\right|_{2}^{2} + \theta^{*}$$

(14)

Where the minimizer θ^{*} replaces the regularizer and is obtained using an optimizer starting from a random initialization of the parameters.

4. EXPERIMENTS

This chapter presents the results, both quantitative and qualitative, of incorporationg InstantNGP and DiF into the considered self-supervied image deblurring method on synthetic and rea limages. The quantitative evaluation includes measurements of average PSNR, SSIM, LPIPS, training time, and model size. As baseline comparison, experiments with the provided implementations of SIREN and Fourier feature networks were conducted. Moreover, experiments were conducted to evaluate the impact of CNN prior regularization, compared to the TV-like regularization approach. By default, the parallax ICB model, introduced in Chapter 2.2.3, is considered as the blur formation model in all the experiments.

The experiments were conducted using a pc with an Nvidia RTX 2060 super GPU, AMD Ryzen 5 3600 CPU and 16 GB of DDR4 ram running Ubuntu 22.04. The implementation was executed in Python, utilizing the PyTorch library and Nvidia CUDA extensions.

The source code associated with this thesis is also made publicly available in the github repository: https://github.com/Lauri-Lehtonen/INR-deblurring, which is built upon the code repository released by Torres and Kämäräinen [4]

4.1 Evaluation datasets

The evaluation data comprises two datasets introduced by Torres and Kämäräinen [4]: Virtual Camera Motion Blur (VirtualCMB) and Real Camera Motion Blur (RealCMB), containing synthetic and real images, respectively. Each dataset includes data tuples consisting of a blurry image, a corresponding sharp image, a depth map, and the camera trajectory followed during the exposure. VirtualCMB is rendered using the UNITY graphics engine, while RealCMB is captured by an iOS app [18] deployed on iPhone devices equipped with LiDAR depth sensors. In total, the VirtualCMB and RealCMB contains 983 and 58 data tuples, respectively. However, due to hardware limitations, only the first 10 data tuples from VirtualCMB are used in our experiments, as this dataset includes HD resolution images, which demand significant computational resources.

4.2 InstantNGP

InstantNGP incorporates the PyTorch extension of tiny-cuda-nn [19] to initialize a fast NN with the multiresolution hash encoding. The NN consists of a single MLP with 2 hidden layers, each containing 128 neurons, and ReLU activation functions. For optimization, parameters are set as follows: α=1×10−3, β1=0.90, β2=0.99, and ϵ=1×10−15.

In respect to the hash encoding, its parameters are set to: 14 levels of feature vectors, 2 features vectors per level, a log2 hashmap size of 20, and a base resolution of 12,. Although these settings differ from the default values used in the InstantNGP implementation [8], the employed values are set to have a good compromise between training time and image quality. Table 4.1 compares the evaluation metrics for a single image in the RealCMB. Notably, the used settings result in a substantial decrease in training time and model size, with negligible effects on image quality.

Settings	PSNR	SSIM	LPIPS ($x10^{-4}$)	Time (s)	Size (MB)
Ours	33.154	0.964632	2.13	6.943	30.67
INGP	33.189	0.958875	2.25	17.299	389.28

Table 4.1. Comparison of performance in single image deblurring between our used parameters and the Instant-NGP default parameters

4.3 DiF

Dictionary fields were implemented using the Python code provided by the authors of the factor fields paper [12]. An MLP is used for coefficiet and basis as the field representations for their memory efficiency and their smoothness bias. the coordinate transform function is set as a sawtooth function as the periodic function allows us to apply the same basis at multiple resolutions. The model has 61445328 total parameters with 1024 hidden dimensions and $\alpha = 0.002$. These settings were found to have the best balance between measured performance and computational time.

4.4 Comparisons of INRs

Deblurring was ran for all images in RealCMB dataset and 10 random images in VirtualCMB dataset.

	InstantNGP	DiF	SIREN	Fourier feature network
PSNR	30.99	28.48	31.42	20.62
SSIM	0.9442	0.9073	0.9502	0.7049
LPIPS ($x10^{-4}$)	4.014	0.2137	4.615	0.7041
Training time (s)	8.86	1274	14.07	27.13
Model size (MB)	30.67	8.689	0.5698	3.079

Table 4.2. Average performance in RealCMB

Table 4.2. presents the average evaluation metrics for each INR in the RealCMB. It can be seen that InstantNGP reduces significantly the training time while achieving comparable performance with SIREN, in terms of PSNR and SSIM. Conversely, DIF outperforms the Fourier feature network but does not reach the level of SIREN and InstantNGP. Besides, DiF substantially requires longer training times.

	InstantNGP	DiF	SIREN [4]	Fourier feature Networt
PSNR	28.73	30.58	27.14	24.59
SSIM	0.8420	0.9420	0.8001	0.8591
LPIPS (x10^{-4})	10.21	8.396	36.56	13.28
Training time (s)	85.19	4904	459.9	996.2
Model size (MB)	30.67	8.689	0.5698	3.079

Table 4.3. Average performance in VirtualCMB

Table 4.3. presents the average performance in VirtualCMB. In this case multiresolutional hash encoding outperforms SIREN in all observed metrics as well as being magnitudes faster at performing deblurring. Dictionary fields outperforms all other methods by a wide margin, but requires training times magnitudes of order larger than the others.

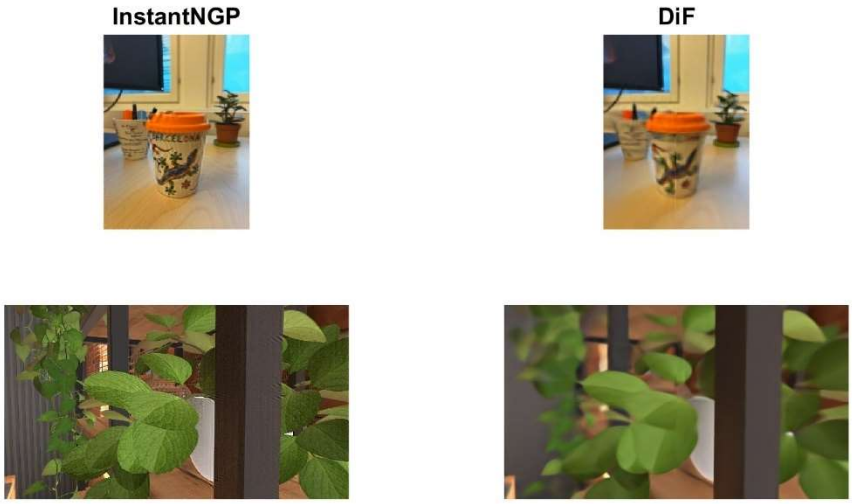

Fig 4.1. results of deblurring images with InstanNGP and DiF in both datasets.

Fig 4.1 provides a visual sample comparison in the RealCMB and VirtualCMB datasets for the INRs implemented, with the images in the first row being from RealCMB and second from VirtualCMB. In the first column te results with InstantNGP can be seen and DiF in the second column. It can be seen that the results from DiF have a tendency to smooth the images which lead to higher measured quality in larger images but loses information in smaller details which leads to InstantNGP having superior visual quality in the RealCMB dataset.

Overall the analyzed methods, those being InstantNGP, that provides noticeable improvements in training time and marginal improvements in performance for virtual images over SIREN, and DiF that shows the best quality metrics under the VirtualCMB dataset, but requires significant training time and may need a future C++ or CUDA implementation to improve training time.

4.5 Analysis of the CNN prior regularization

Altough overall the measured performance of Multiresolutional hash encoding fell with the use of the CNN Prior regularization, the CNN also helps address ringing artifacts in the resulting images, that were observed with other regularization methods leading to visually more pleasing results.

To validate the untrained CNN it is tested against both net gradient and filter regularization using the SIREN model.

	Untrained CNN	Net gradient	filter
PSNR	32.64	32.89	32.78
SSIM	0.9564	0.9555	0.96553
LPIPS ($x10^{-4}$)	2.40	2.52	2.38
Elapsed time	11.32	16.08	7.32

Table 4.4. Untrained CNN for regularization tested against other regularization method

From the experiment we can see that it performs only slightly worse than the other methods while allowing us to ignore changes input resolutions between different INR architechtures making experimentation faster.

5. CONCLUSION

This thesis integrates and evaluates InstantNGP and DiF as alternative INR in a self-supervised approach for non-blind image deblurring. Experimental results on real and synthetic images demonstrate significant improvements in both training time and performance for InstantNGP and even more improvements in large virtual images for DiF at the cost of high training time. In addition, it explores the use of a CNN prior as an alternative regularization technique. Our results exhibit a minor decrease in measured performance, which is offset by the adaptibility it provides to the deblurring process.

REFERENCES

1. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: Deblurgan: Blind motion deblurring using conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8183–8192 (2018)

2. Mustaniemi, J., Kannala, J., Särkkä, S., Matas, J., Heikkilä, J.: Gyroscope-aided motion deblurring with deep networks. In: IEEE Winter Conference on Applications of Computer Vision (WACV)

3. Makarkin M, Bratashov D. State-of-the-Art Approaches for Image Deconvolution Problems, including Modern Deep Learning Architectures. Micromachines (Basel). 2021;12(12):1558–.

4. Torres GF, Kämäräinen J. Depth-Aware Image Compositing Model for Parallax Camera Motion Blur. In: Image Analysis. Cham: Springer Nature Switzerland; p. 279–96.

5. Sitzmann, V., Martel, J. N. P., Bergman, A. W., Lindell, D. B., & Wetzstein, G. (2020). Implicit Neural Representations with Periodic Activation Functions. ArXiv.Org. https://doi.org/10.48550/arxiv.2006.09661

6. Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2020). Deep Image Prior. International Journal of Computer Vision, 128(7), 1867–1888. https://doi.org/10.1007/s11263-020-01303-4

7. Zhang S, Shen X, Lin Z, Mech R, Costeira JP, Moura JMF. Learning to Understand Image Blur. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2018. p. 6586–95.

8. Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Wang, J., Yu, B., ... & Liu, Y. (2020). Watch out! motion is blurring the vision of your deep neural networks. Advances in Neural Information Processing Systems, 33, 975-985.

9. Rajagopalan AN, Chellappa R, editors. Motion deblurring : algorithms and systems. Cambridge: Cambridge University Press; 2014.

10. Tancik, M., Srinivasan, P. P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J. T., & Ng, R. (2020). Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. ArXiv.Org. https://doi.org/10.48550/arxiv.2006.10739

11. Vincent Sitzmann, Chiyu Max Jiang. Awesome Implicit Neural Representations. Accessed 11.11.2023. https://github.com/vsitzmann/awesome-implicit-representations

12. Chen, A., Xu, Z., Wei, X., Tang, S., Su, H., & Geiger, A. (2023). Dictionary Fields: Learning a Neural Basis Decomposition. ACM Transactions on Graphics, 42(4), 1–12. https://doi.org/10.1145/3592135

13. Kupyn O, Martyniuk T, Wu J, Wang Z. DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE; 2019. p. 8877–86.

14. Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T. Simple Baselines for Image Restoration. In: Computer Vision - ECCV 2022. Switzerland: Springer; 2022. p. 17–33.

15. Zamir SW, Arora A, Khan S, Hayat M, Fahad Shahbaz Khan, Ming-Hsuan Yang. Restormer: Efficient Transformer for High-Resolution Image Restoration. arXiv.org. 2022;

16. Zhang, K., Ren, W., Luo, W., Lai, W.-S., Stenger, B., Yang, M.-H., & Li, H. (2022). Deep Image Deblurring: A Survey. International Journal of Computer Vision, 130(9), 2103–2130. https://doi.org/10.1007/s11263-022-01633-5

17. Müller, T., Evans, A., Schied, C., & Keller, A. (2022). Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics, 41(4), 1–15. https://doi.org/10.1145/3528223.3530127

18. Chugunov, I., Zhang, Y., Xia, Z., Zhang, X., Chen, J., Heide, F.: The implicit values of a good hand shake: Handheld multi-frame neural depth refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2852–2862 (2022)

19. https://github.com/NVlabs/tiny-cuda-nn/tree/master accessed on 9.11.2023