

## Research paper

# RoadSitu: Leveraging road video frame extraction and three-stage transformers for situation recognition

Subhajit Chatterjee<sup>a</sup>, Hoorang Shin<sup>b</sup>, Joon-Min Gil<sup>a</sup>, Yung-Cheol Byun<sup>c,\*</sup>

<sup>a</sup> Department of Computer Engineering, Jeju National University, Jeju 63243, South Korea

<sup>b</sup> INUC Co., Ltd., #111 Hyupjae2-gil 37, Hanlim-eup Jeju, South Korea

<sup>c</sup> Department of Computer Engineering, Major of Electronic Engineering, Jeju National University, Institute of Information Science & Technology, Jeju 63243, South Korea

## ARTICLE INFO

## Keywords:

Machine learning  
Deep learning  
Road situation recognition  
Transformers  
Road video frame  
Video analysis

## ABSTRACT

Situation recognition is an crucial problem in scene understanding, activity understanding, and action reasoning as it provides a structured representation of the main activity depicted in the image. Semantic role labeling is crucial to situation recognition, which is challenging because a single action can have multiple meanings and purposes depending on its context. Understanding images beyond the highlighted actions requires inferences about the context of the scene, the objects, and their role in the captured event. Recently, situation recognition (SR) has been introduced, which jointly derives a collection of the action (activity), meaning-role, and noun (entities) pairs in the form of moving images. To label these frames as action frames, we must assign nouns (entities) to the role based on the content of the observed image. One of the main challenges is managing the complex dependencies between the assigned roles (nouns) and the predicted action, as the correct role assignment often depends on the accuracy of the action prediction. We introduce, RoadSitu, a road situation recognition that involves generating a structured summary of what is happening in a road scenario using an action and the semantic roles played by agents from a video frame. The action can describe a diverse set of situations, and the same agent can play various roles depending on the situation depicted in the video frame. Therefore, a situation recognition model needs to understand the context of each video frame and the visual-linguistic meaning of the semantic roles of that particular frame. One of the main challenges in this work is the complex task of annotating video frames with semantic roles and handling the structured dependencies between the assigned roles (nouns) and the predicted action (activity). Additionally, the sparsity of meaningful semantic information within road scenarios poses further difficulties. To overcome these challenges, we introduce a novel approach where action recognition and noun estimation work together interactively to form structured summaries of each situation. In experiments using a road video dataset obtained from a South Korean company, RoadSitu achieved significant improvements across various performance metrics, with a Top-1 verb accuracy of 43.46%, Top-5 verb accuracy of 72.48%, and value accuracy of 34.21%, outperforming baseline models such as GSRTR and JSL by 2.4% and 3.86% in Top-1 verb accuracy, respectively. These results demonstrate the effectiveness of our model in handling complex road scenarios.

## 1. Introduction

Image situation recognition goes beyond simply identifying the main action in an image. It aims to predict not only the key action describing the activity but also all the entities involved. Semantic roles act as a bridge, connecting individual objects to the action (activity) itself. These roles define the specific function each entity plays within the scene. By

predicting both the action and its associated roles, situation recognition becomes a structured prediction task. The final output is a comprehensive understanding of the scene, encompassing the action and the entities participating in it. Images capture our life events and moments, preserving them over time. These images often include a wide range of scenarios with several events, various objects, and their surroundings. They also often relate to several occurrences. While there are many uses

\* Corresponding author.

E-mail addresses: [subhajitchatterjee@stu.jejunu.ac.kr](mailto:subhajitchatterjee@stu.jejunu.ac.kr) (S. Chatterjee), [jason@inucreative.com](mailto:jason@inucreative.com) (H. Shin), [jmgil@jejunu.ac.kr](mailto:jmgil@jejunu.ac.kr) (J.-M. Gil), [ycb@jejunu.ac.kr](mailto:ycb@jejunu.ac.kr) (Y.-C. Byun).

<https://doi.org/10.1016/j.rineng.2024.103197>

Received 16 September 2024; Received in revised form 4 October 2024; Accepted 20 October 2024

Available online 13 November 2024

2590-1230/© 2024 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

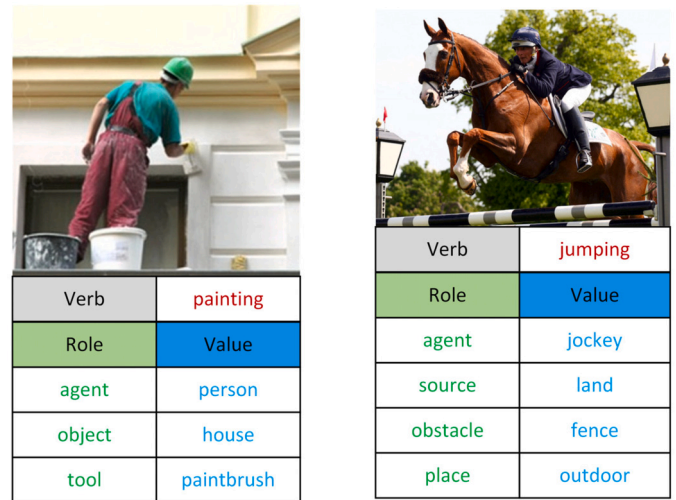
for extracting such rich and complicated information from images, such as understanding images and recovering visual content, situation recognition is the best option for providing a comprehensive description of this kind of scenario. Road situation recognition from road video is a very complicated scene or video-understanding task.

### 1.1. Research background

To comprehend a video sequence holistically, especially one that involves different actions and many entities, it is necessary to predict the verb as well as the associated roles for each verb. Holistic understanding of such a video sequence, especially one that involves multiple entities, requires predicting the action and the associated roles for the particular verb. For example, a human being after watching a video, it is easy to describe a details interpretation of the narrative. Answering such questions as who acted (agent), why they did it (purpose/goal), how they did it (method), and where they did it (location) might be helpful (multi-event understanding). An analysis of visual information leads to a conclusion based on visual reasoning. Image situation understanding from a single frame is an important computer vision technique applied in many different contexts to comprehend visual scenes. This work primarily focuses on using road video frame analysis to understand various road scenarios and different driving patterns. An assistance system that provides scene descriptions can be helpful for a better comprehension of driving patterns and increased safety in many everyday driving situations on the road. Computer vision emerges as a pivotal and rapidly advancing domain in machine learning. The field of computer vision has made tremendous progress over the past few years in image understanding, including categorizing objects [1,2], detecting objects [3], and even recognizing individual actions [4]. In contrast, classification by single-term is not sufficient to gain a deeper understanding of the image. For many real-world applications, it remains challenging to understand and comprehend image content in greater detail. Human visual understanding goes beyond simply identifying entities or the main action in an image. When we look at an image or video, we form impressions about the entire scene: what is happening, who is involved, the tools used, and so on. This requires a deeper analysis of the activity within the image or going beyond simply identifying objects; situation recognition strives for a deeper understanding of a scene. This detailed analysis is crucial for various applications. For instance, autonomous vehicles and robots, considered intelligent agents, need to interpret situations to make informed decisions about their actions in response to the environment.

### 1.2. Related work

Image situation recognition refers to the process of understanding and interpreting the contextual information present in an image. It involves identifying and analyzing various elements within an image to infer the situation or scenario depicted. This includes recognizing objects, activities, interactions, spatial relationships, and other relevant contextually important factors. The goal of image situation recognition is to enable machines to comprehend visual scenes like how humans do, allowing them to make informed decisions or take appropriate actions based on the perceived situation. This capability finds applications in various fields such as computer vision, autonomous systems, surveillance, image understanding, and human-computer interaction. Key components of image situation recognition include object detection and recognition, activity recognition, scene understanding, spatial reasoning, temporal reasoning, and contextual reasoning. Image recognition has achieved significant success with the development of deep neural networks [5–7]. Image situation recognition, which involves understanding and classifying the activities or events occurring within an image or a scene in a video, plays a crucial role in various computer vision applications. Situation Recognition was first proposed by [8,9], which is further generalized as an action classification task by



**Fig. 1.** Example of road situation recognition understanding the scene in the road. Two distinct circumstances are shown here. The task of road situation recognition is to predict the action (verb) and values of all associated semantic roles (agent, location, manner). In the left image, a car (agent) is driving reverse (verb) on road (location) in reverse direction (manner). In the left image, a person (agent) is painting (verb) a house (object) with a paintbrush (tool). In the right-side image of the agent jumping, the task is to identify who is jumping (the jockey), where the agent is jumping to and from (the land), what obstacle the agent is jumping over (the fence), and where the action is taking place (outdoors).

[10,11]. SR tasks include image captioning [12–14], another approach is scene graph generation [15,16], and human object interaction detection proposed by several authors [17–19] included in these research fields.

Yatskar et al. [9] first proposed the concept of image situation recognition as a way to achieve this more comprehensive understanding. This task involves analyzing both the verbs (actions) happening in a scene and the associated semantic roles, essentially labeling the image with “action frames.” An verb is contextualized by the roles associated with each verb-specific frame. To create these frames, you must assign values (nouns) to these roles depending on the image content. These frames then provide a structured way to access valuable semantic information, such as who performs the action, where it takes place, and what the potential outcome might be.

As illustrated in Fig. 1, situation recognition goes beyond identifying actions in images. It delves deeper into understanding the specific roles each object plays within the scene. These roles, like “agent” (person), “object” (object being painted), and “tool” (paintbrush) connect individual objects to the main verb (e.g., “painting”). However, this task presents significant challenges. Verbs can occur in diverse situations with varying agents, and different verbs require different roles. For instance, “jumping” has roles like “source,” “destination,” “agent,” “place,” and “obstacle”, while verb “brushing” in Fig. 1 involves roles like “agent”, “target”, and “tool”. Notably, the possible values for these roles can vary significantly even for the same action, as shown by the Fig. 1 image depicting “painting” and “jumping”. For the right side, “jumping”, it is necessary to recognize who is jumping (jockey), where the agent is jumping (land), what obstacles the agent is jumping over (fence), and where the action takes place (outside). This vast number of potential roles and values makes situation recognition a challenging reasoning task. Situation recognition, as exemplified in Figs. 1 and 2, goes beyond identifying the main action (verb) to predict the specific roles each object plays within the scene. Fig. 2 showcases this by illustrating how a model might predict the activity in images from the extreme left as “driving reverse” with associated roles like “agent” (car), “location” (road), “manner” (in reverse direction). Su et al. [20], explores the enhancement of traditional LSTM networks by in-

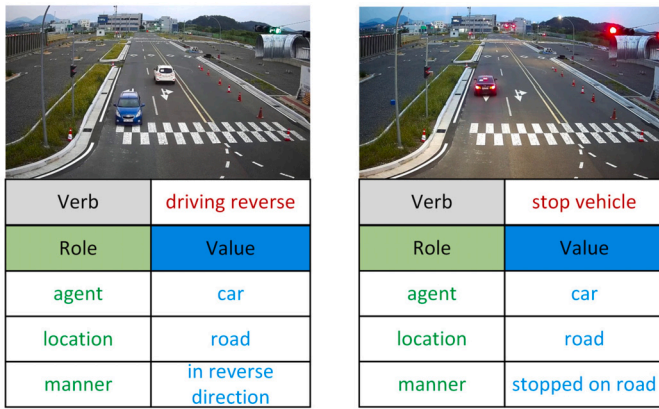


Fig. 2. An image situation recognition example, which goes beyond the salient action of the image. Two distinct circumstances are shown here. Predicting the action (verb) and values of all associated semantic roles (agent, object, tool, place) is the task of situation recognition. Given the right-side image of stop vehicle, the task is to identify what the object acting as driving (car), where the scene of the event (road), what the action being performed (stop vehicle) and describes how the driving is happening (stopped on road).

roducing dependent bidirectional RNNs [21], improving performance on sequence-based tasks. This study authors [22] propose a framework for unsupervised multi-modal neural machine translation, leveraging multiple input modalities to improve translation quality without parallel data. This work introduces TMPCA, a tree-structured multi-stage principal component analysis method, and discusses its theoretical foundations and applications in signal processing [23,24]. In the work [25], presents a tree-structured multi-linear PCA approach for efficient text classification, significantly reducing computational costs while maintaining classification accuracy. Another survey examines various types of recurrent neural networks, analyzing their memory retention capabilities and their applications in sequential data processing tasks [26].

Several studies on video processing and situation awareness are looking for ways to improve object tracking in complex scenarios. In task [27], we introduce GCEVT, which integrates global context integration for vehicle tracking in UAV videos and combines non-local blocks and conversion layers to improve tracking accuracy. Another approach uses fine-grained and time-sensitive features to improve the robustness of multi-object tracking systems [28]. Similarly, [29] provides a task-specific framework to improve performance by separating detection and re-identification from one-time multi-object tracking (MOT). Further research in [30] explores how finer-grained features can be used in one-time MOT of drone videos to resolve optimization conflicts between detection and ReID operations.

The WordNet [31] and FrameNet [32,33] offer a powerful tool for analyzing verbs and their roles in conveying situational activities within a text. Verb and noun predictions were made simultaneously in the early CRF-based situation identification methods through structured learning. It has been demonstrated, however, that training distinct models in two stages is preferable to sharing the visual representation for the two tasks. As a result, current attention-based [34], GNN-based [35], and RNN-based [36,37] techniques all predict the verb in the first stage and identify the corresponding semantic role in the second. In the realm of situation recognition, one well-known challenge is situation recognition from an image. The landscape of situation recognition research is expanding rapidly, encompassing various domains such as video analysis and both one-stage and two-stage predictions. Despite this growth, the field is still evolving and has not reached full maturity. Initially, efforts were concentrated on verb prediction, but there is now a shift towards activity prediction based on the context. In the following sections, we will delve into the nuances of different approaches in SR. An interesting work introduced by Pratt et al. [38] was grounded situation recognition (GSR). The advancement over situation recognition,

and grounded situation recognition goes a step further by grounding this understanding in real-world sensory data, thereby enabling more robust and contextually rich interpretations of the observed environment. To fully utilize this framework in visual scenario identification, the author proposed a Collaborative Glance-Gaze Transformer (CoFormer) [39], which consists of a Gaze transformer and a Gaze transformer. A task that involves creating bounding-box groundings of entities, entities involved in the activity with their responsibilities (e.g., agent, tool), and structured semantic summaries of photos defining the main activity. Building on the original image SR there are some datasets available, SituNet, imSitu, and Situations With Groundings (SWiG) dataset which adds bounding box (bbox) annotations for all visible semantic roles (63.9% of roles have bbox annotations). Hence, various models have been devised, employing techniques such as Conditional Random Fields (CRF) [40], Long Short-Term Memory networks (LSTMs) [38], Recurrent Neural Networks (RNNs) [34], Graph Neural Network (GNN) [41], and Gated Graph Neural Networks (GGNNs) [37] to capture the broader relationships between verbs and roles in the context of situation recognition. In recent times, the attention of researchers has largely shifted towards transformer-based models [42–46], particularly in the context of situation recognition tasks utilizing the SWiG dataset.

### 1.3. Comparison with existing works

To illustrate the contribution of the RoadSitu model, we compare it with state-of-the-art models in the field of situational awareness. We highlight ways to improve existing methods across several dimensions, including verb prediction, role prediction, and grounded noun estimation.

As can be seen from the comparison in Table 1, RoadSitu performs better than other models on a number of evaluation parameters. Specifically, RoadSitu's three-step conversion architecture efficiently and complementarily processes actions and accompanying name predictions. Compared to sequential processing in GSR and RGNN, where role prediction comes after verb prediction without any feedback in between, this is a major gain.

Additionally, unlike GSRFormer and CoFormer, which rely primarily on complex attention processes, RoadSitu presents an integrated strategy that uses interactions between verbs and nouns to increase prediction accuracy. Comparing this model to previous ones, the base roll accuracy is improved to 28.11%. The ClipSitu use CLIP models for conditional prediction, the underlying roll's performance is still inferior than RoadSitu's, underscoring the advantages of the model's unique construction for road video data.

### 1.4. Problem statement

To achieve human-like understanding of road events, SR entails identifying the salient action (or activity) in an image and detecting the appropriate semantic roles (e.g., agent and objects). This task is crucial in road scenarios for applications like traffic monitoring and autonomous driving, where safe and informed decision-making depends on precise scene perception. Recently, SR techniques work in two stages: first, they anticipate the action (activity), and then they identify the associated semantic roles.

- **Verb Classification Challenges:** In general, loss functions for object recognition are not sufficient for verb classification due to the large intra-class variation between activities and the high similarity between verbs. There are subtle differences in road conditions in actions such as “drive” and “stop”, which are difficult to understand and can lead to misclassification.
- **Autoregressive Role Prediction:** Traditional SR methods predict semantic roles (nouns) in an autoregressive manner, detecting each role in turn. This approach does not effectively model complex

**Table 1**  
Comparison of RoadSitu with state-of-the-art models.

Method	Model architecture	Role prediction method	Verb Prediction Accuracy (%)	Grounded Role Accuracy (%)
RoadSitu (Ours)	3-Stage Transformer	Interactive Role and Action Prediction	<b>43.46</b>	<b>28.11</b>
ClipSitu [46]	CLIP-based Transformer	Conditional Verb and Noun Prediction	41.32	26.40
GSRFormer [45]	Transformer with Semantic Attention Refinement	Noun-Based Action Refinement	41.06	26.04
CoFormer [39]	Glance and Gaze Transformers	Joint Verb and Noun Estimation	41.06	26.04
RGNN [47]	Graph Neural Network	Relational Graph for Role Predictions	39.60	25.03
GSR [38]	2-Stage Framework	Sequential Role Prediction	38.83	22.47

relationships between roles, such as how “agents” (e.g. cars) interact with “objects” (e.g. pedestrians) or “places” (e.g. roads).). As a result, these models struggle to capture the extensive conditional dependencies that exist between roles in a dynamic road environment.

- **Semantic Sparsity in Road Situations:** Road scenarios are very diverse in terms of possible objects, roles and interactions and rare in terms of available semantic data. This sparsity complicates the model’s ability to generalize and accurately predict the set of behaviors and roles that occur in real-world driving conditions.

Addressing these challenges requires an integrated approach that can simultaneously predict the verb and semantic roles while modeling the complex relationships between entities. Our work introduces such an approach, improving the performance of situation recognition models in road environments by using a three-stage transformer architecture that predicts activities and roles in a structured and complementary manner.

### 1.5. Contribution

This work focuses on using road video scene analysis to understand various driving patterns. An assistance system that is capable of providing accurate scene descriptions can be helpful for a better comprehension of driving behaviors and improve safety in many everyday driving situations. As a rapidly advancing field, computer vision plays a pivotal role in machine learning applications such as road situation recognition. In this study, we used the road video dataset to develop a computer system for road situation recognition. First, the model predicts an action (activity). Next, a transformer analyzes nouns (entities) and their relations by leveraging role features. Finally, a third transformer predicts the associated nouns for the roles linked to the predicted action. We implemented several modifications to improve our approach and demonstrated how to set up experiments with smaller video datasets. We also perform extensive annotation of video frames and semantic role labeling. The main contributions of this study are:

- We introduce a novel method where verb predictions and noun estimations are performed interactively and complement each other within a collaborative framework.
- Significant effort was invested in data preparation, including renaming video files and creating annotations for each frame. Using the Makesense.ai tool, we created detailed annotations by drawing bounding boxes around visual entities in the sampled frames. These annotated files were then used to generate the train.json, validation.json, and test.json files, essential for training, validating, and testing the model.
- Our comprehensive evaluation of the road video dataset demonstrates that the proposed method achieves state-of-the-art performance across all evaluation metrics.
- We illustrate the effectiveness of our proposed model through extensive experiments and detailed analyses with the road video data.

## 2. Methods

We assume discrete sets of verbs  $V$ , nouns  $N$ , and frames  $F$  for situation recognition. Each frame  $f \in F$ , associated with a set of semantic roles denoted by  $R \in R_s$ , is associated with a noun  $n_e \in N \cup \{\phi\}$ , where  $\phi$  specifies that the noun is either unknown or not applicable. The set of pairs of semantic roles and nouns are referred to as a realized frame,  $R_s = (R, n_e) : R \in R_s$ . Given an image, the task then is to predict  $S = (v, R_s)$ , where  $v \in V$  is the salient verb (action) corresponding to the image and  $R_s$  its corresponding realized frame. For example, consider the image in Fig. 1. The realized frame corresponding to the verb painting and jumping consists of four and five role-value pairs, respectively. i.e., left-side: (agent, person), (object, house), (tool, paintbrush), right-side: (agent, jockey), (source, land), (obstacle, fence), (place, outdoor). On the other hand, if we consider the purely road based video data frame example as Fig. 2. The realized frame corresponding to the verb driving reverse and stop vehicle consists of three role-value pairs, respectively. i.e., left-side: (agent, car), (location, road), (manner, in reverse direction), right-side: (agent, car), (location, road), (manner, stopped on road).

The common pipeline for SR resembles two processes: verb prediction and each role related to the predicted verb (activity). The accuracy of the predicted verb is very important because the estimation of the noun (entity) and corresponding roles depends entirely on the predicted verb. As long as the verb prediction is incorrect, the estimated noun is also incorrect, because the predicted verb determines a series of roles based on it and is used as a basis for estimating nouns. Moreover, verbs are difficult to predict since they are very abstract, and their situations vary considerably. Although verbal prediction is important and difficult, it has often been attempted naively, for instance using a single verb. Graph neural networks (GNN) use this neural network structure during training and inference to predict verbs and their role pairs.

In traditional methods, verbs can provide information about the predicted verbs to help predict roles, but the reverse is not possible. We solve this problem with a collaboration framework that uses a converter-based attention mechanism. Inspired by [39] propose a three-stage collaborative converter model consisting of a converter-encoder and a converter-encoder-decoder, as shown in Fig. 3. The transducer encoder is responsible for aggregating the characteristics of the image and predicting verbs using self-attention. Meanwhile, the transformer Decoder-Encoder focuses on the corresponding image area through self-attention and cross-attention and estimates the corresponding name and role.

As shown in diagram 3, the three-stage transformer configuration is shown in the updated schematic. This is a transformer-decoder-encoder, where the transformer-decoder estimates the role names associated with the projected actions (activities), and the transformer-encoder uses the role characteristics to further evaluate the nouns and their relation. We show how to predict action (activity) to the path from feature extraction to the expected final result is presented in a simple and systematic way in this step-by-step way.

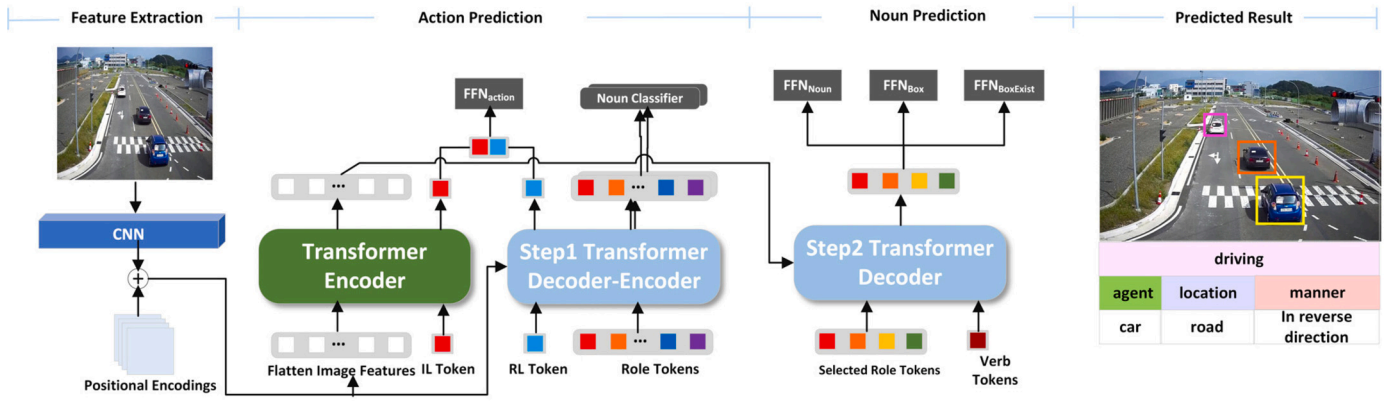


Fig. 3. This diagram shows the overall architecture of a three-stage transformers. Transformer-encoders predict verbs through transformer-decoder-encoders, which analyze nouns and their relations by using role features, while transformer-decoders estimate nouns for the roles associated with the predicted verbs.

**Table 2**  
Details of each class type in the dataset.

Sl no.	Defined class labels	Number of videos per class	Video length
Class 1	driving_reverse	100 videos	10 seconds
Class 2	driving_reverse(others)	139 videos	10 seconds
Class 3	object_falling	34 videos	10 seconds
Class 4	pedestrian	111 videos	10 seconds
Class 5	stop_vehicle	114 videos	10 seconds



Fig. 4. Examples of road scenarios categorized into five distinct classes: “driving reverse,” “driving reverse (others),” “object falling,” “pedestrian,” and “stop vehicle.” These classes represent diverse driving situations used in the dataset for road situation recognition.

## 2.1. Dataset overview

We collected the data from an industrial company to support our research. The dataset comprises 10-second road videos organized into five distinct folders with class labels. The dataset consists of a total of 498 videos, all of which exhibit no irregularities and were captured in accordance with various road scenarios. These videos are categorized into five road situations, namely *driving\_reverse*, *driving\_reverse(others)*, *object\_falling*, *pedestrian*, and *stop\_vehicle*. The distribution of videos and their corresponding video lengths is presented in Table 2. The dataset encompasses five classes, each reflecting different driving patterns and road scenarios. The classes are *driving\_reverse*, *driving\_reverse(others)*, *object\_falling*, *pedestrian*, and *stop\_vehicle*. The organization of these classes facilitates the exploration of diverse driving behaviors and the nature of car-related situations captured in the dataset. Fig. 4 presents examples of videos from each class.

The primary objective of our research is the comprehensive detection of different situations on the road for enhanced road safety. Our focus encompasses identifying lane presence, road obstructions, instances of objects falling onto the road, and the presence of pedestrians. Mainly, we aim to address scenarios where vehicles are stationary in improper lanes or are traveling in the wrong direction, contributing to potential road hazards. To facilitate this, we have structured our problem into five distinct classes: “*driving\_reverse*,” “*object\_falling*,” “*driving\_reverse(others)*,” “*pedestrian*,” and “*stop vehicle*.” Each class encapsulates a specific event or situation in the videos, and the target domain is what the model aims to distinguish during video analysis in the testing phase.

## 2.2. Data preparation and annotation for road situation recognition

In a 10-second video consisting of more than five events, frames are sampled at 1 frame per second, resulting in 10 frames,  $F = \{f_t\}_{t=1}^T$ , where  $T = 10$ . Each of these frames is annotated individually using the Makesense.ai tool, with a separate annotation task created for each frame in the video. The name of each frame corresponds to the verb associated with the event in that frame. For example, a frame named *drivingreverse\_0* indicates the verb “driving reverse” followed by the frame number. Annotation for class ‘driving reverse’, Verb: *drivingreverse*

Nouns:

- agent = car
- object = object
- location = road
- manner = driving in reverse

This process ensures that each frame is properly labeled with the relevant verb and nouns, facilitating accurate situation recognition in the video.

### 2.2.1. Details on verb and role declaration according to our dataset

Our dataset is designed for grounded situation recognition, where each image is annotated with a verb and its corresponding semantic roles. The annotations are structured similarly to the well-known ImageNet dataset [1], which is based on WordNet [48], a large lexical database of English. In our situation, we also keep the structure of assigning roles to verbs (action) in relation to visual data. The annotated images, as exhibited in Fig. 5, use bounding-box (bb) annotations to highlight important components for road scenario detection inside the regions of interest. The location (road) is represented by the cyan-colored ‘bb,’ while the agent (vehicle) is highlighted by the red-colored ‘bb’.

Situational recognition typically employs JavaScript Object Notation (JSON) files to arrange and store data in human- and machine-readable



**Fig. 5.** Comparison between the original and annotated images. The annotated image highlights the region of interest with bounding-box (bb) annotations in  $[x1, y1, x2, y2]$  format, used to identify the location for road situation recognition. The blue and orange lines indicate the key boundaries of the annotated area. For example, annotated image 1 describes cyan-colored ‘bb’ representing place (road) and annotated image 2 describes red-colored ‘bb’ representing agent (car).

formats. These files play a crucial role in displaying and transmitting structured information about diverse circumstances and events. JSON files are vital tools for organizing complicated information by efficiently encapsulating data linked to entities, activities, locations, timestamps, and other associated features. By establishing a hierarchy, JSON files allow for the clear organization of nested data, such as the relationships between verbs, roles, and their associated attributes. This organized format not only makes data management easier, but also allows event-based annotations to be sent easily, making JSON a significant resource for situational awareness and related fields.

- **Verb:** Each image in the dataset is paired with a verb that represents the key action taking place in the scene. This is reflected in the annotation file, where the “verb” field indicates the primary action associated with the image.
- **Frame:** The verb is linked to a frame that defines the set of semantic roles (e.g., agent, object, location) that are necessary to describe the action. The frame helps provide structure to the annotation by specifying the roles involved in the action.
- **Annotation:** The annotation for each role is provided in the bounding-box format, represented as  $[x1, y1, x2, y2]$ . In the annotation file, “bb” denotes the bounding-box groundings for each role. If no grounding is available for a specific role, the placeholder  $[-1, -1, -1, -1]$  is used to indicate the absence of a bounding box.

Additionally, The dataset annotations are organized in separate JSON files for training, validation, and testing. The structure of these files is shown in Figs. 6, 7, and 8. Fig. 6 illustrates the format of the train.json file, which includes the annotations used for training the model, such as verbs, roles, and bounding-box groundings for each image. Fig. 7 demonstrates the structure of the validation.json file, which is used to evaluate the model’s performance during training. Lastly, Fig. 8 presents the format of the test.json file, used for final evaluation and benchmarking. We have organized our data into three JSON files for efficient processing:

- **Train.json:** Used for model training.
- **Validation.json:** Used for evaluating the model’s performance during training.
- **Test.json:** Used for final evaluation and benchmarking of the model.

### 2.3. Background of the model

We present a specialized transformer model designed for road situation recognition on videos in Fig. 9. The self-attention function, computed over a series of spatiotemporal tokens taken from the input video, lies at the heart of this architecture. We suggest many techniques to

factorize our model along spatial and temporal dimensions, improving efficiency and scalability to handle the potentially large number of spatio-temporal tokens inherent in videos. We also present methods to train our model efficiently on our small industrial dataset. We encapsulated pre-trained image models and regularization algorithms to deal with small data during training. Notably, convolutional models have been a mainstay of the community’s efforts for several years, and in the last few years, several best practices have been identified for these models.

In the RoadSitu model, road video frames are processed as images, where flattened image features are extracted using a CNN backbone followed by a flattening operation. Verb prediction uses the output features from these transformers that correspond to Image-Looking (IL) and Role-Looking (RL) tokens. These features are first passed through an initial transformer model and then into a second transformer model. To predict verbs, both image and role tokens are produced from these transformers.

Following the second transformer’s prediction of the verb, a third transformer uses the image features processed by the second transformer to estimate the grounded nouns associated with the predicted verb. It can be seen in Fig. 9 that the transformers for verb prediction and noun estimation work collaboratively, complementing each other in an interactive and reciprocal manner. With the first transformer, features from flattened images are fed into a single encoder, which produces learnable image tokens. In the first transformer, the image token captures all the essential elements required to predict verbs; the second transformer refines these features with self-attention.

A decoder and an encoder are both included in the second transformer. The learnable role tokens, which stand for all potential role candidates, and the flattened image features are accepted by the decoder. Using these tokens, it pulls features related to a role from the image features. To capture the connections between nouns and their matching roles for verb prediction, the encoder then processes the role features and further refines the role tokens through self-attentions. Lastly, the third transformer is just a decoder that uses the aggregated picture features from the second transformer along with learnable tokens. The expected verb and the roles that go along with it are represented by these input tokens. These are used by the third transformer to derive role traits, which are subsequently used for grounded noun prediction.

### 2.4. Verb prediction

Verb prediction in RoadSitu begins by extracting flattened image features using a CNN backbone, which is then fed into the first transformer. This transformer is designed to predict the main activity or verb of the scene by aggregating the image features through self-attention mechanisms. The Glance transformer uses an image token to capture essential features for verb prediction, while the encoder within the transformer aggregates the image features for more accurate predictions. The second transformer further assists the verb prediction by analyzing potential

```
"drivingreverse_0.jpg": {
  "frames": [{"item": "n02958343", "place": "n03519981", "agent": "n03215508", "object": "n05810948"},
             {"item": "n02958343", "place": "n03519981", "agent": "n03215508", "object": "n05810948"},
             {"item": "n02958343", "place": "n03519981", "agent": "n03215508", "object": "n05810948"}],
  "width": 512,
  "verb": "drivingreverse",
  "bb": {"item": [-1, -1, -1, -1], "place": [145.15463917525773,108.20618556701031,176.16494845360825,233.56701030927832],
        "agent": [113.48453608247422,106.88659793814433,240.82474226804123,238.18556701030923], "object": [-1, -1, -1, -1]},
  "height": 512},
```

Fig. 6. Structure of the train.json file. This file contains the annotations used for training the model, including verbs, roles, and bounding-box groundings for each image in the training set.

```
"drivingreverse_70.jpg": {
  "bb": {"item": [-1, -1, -1, -1], "place": [154.39175257731958,107.54639175257732,172.2061855670103,232.90721649484533],
        "agent": [112.82474226804123,107.54639175257732,242.8041237113402,236.20618556701027], "object": [-1, -1, -1, -1]},
  "height": 512,
  "width": 512,
  "verb": "drivingreverse",
  "frames": [{"item": "n02958343", "place": "n03519981", "agent": "n03215508", "object": "n05810948"},
             {"item": "n02958343", "place": "n03519981", "agent": "n03215508", "object": "n05810948"},
             {"item": "n02958343", "place": "n03519981", "agent": "n03215508", "object": "n05810948"}]}
```

Fig. 7. Structure of the validation.json file. This file outlines the format for validation data, used to evaluate the model’s performance during the training process.

```
"drivingreverse_85.jpg": {
  "bb": {"item": [63.34020618556701,72.57731958762889,255.340206185567,226.96907216494844],
        "place": [77.1958762886598,109.52577319587628,205.1958762886598,112.16494845360826],
        "agent": [-1, -1, -1, -1], "object": [-1, -1, -1, -1]},
  "height": 512,
  "width": 512,
  "verb": "drivingreverse",
  "frames": [{"item": "n02958343", "place": "n03519981", "agent": "n03215508", "object": "n05810948"},
             {"item": "n02958343", "place": "n03519981", "agent": "n03215508", "object": "n05810948"},
             {"item": "n02958343", "place": "n03519981", "agent": "n03215508", "object": "n05810948"}]}
```

Fig. 8. Structure of the test.json file. This file includes the annotations for the test dataset, used for final model evaluation and benchmarking, with the same format as the training and validation sets.

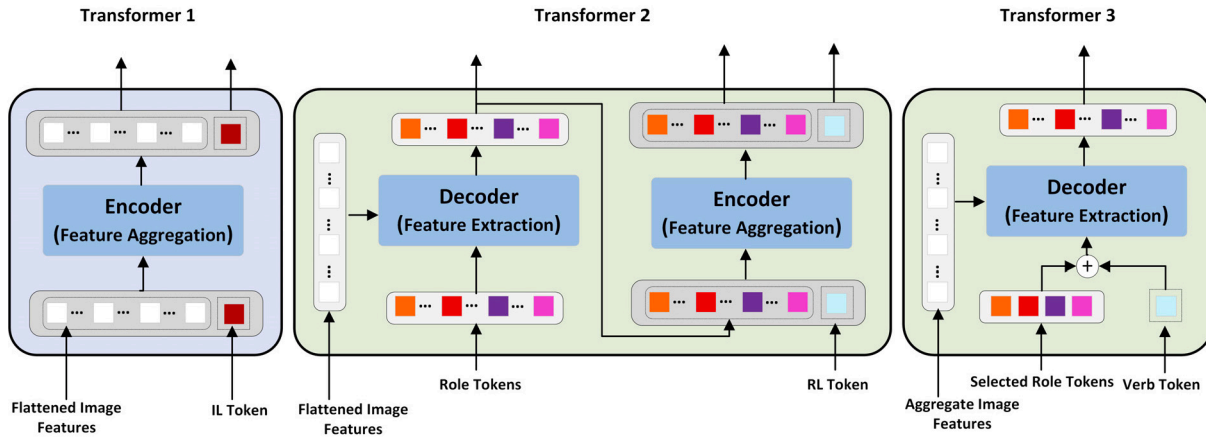


Fig. 9. Overview of the transformer architectures, composed of an encoder for feature extraction and a decoder for prediction, facilitates efficient processing of sequential data.

nouns and their relationships, providing noun-aware verb prediction. The combined efforts of the first and second transformers lead to enhanced verb prediction, with the predicted verb being used to refine the next stage of noun prediction.

### 2.5. Grounded noun prediction

Once the verb has been predicted, the third transformer takes over to estimate the nouns associated with the roles defined by the predicted verb. The third transformer focuses on grounded noun prediction by using frame-role queries. These queries are constructed by combining the learnable role token embeddings and the predicted verb token embedding, which enables the transformer to narrow down the noun candidates associated with the roles. Through self-attention in frame-roll queries and mutual attention between queries and aggregated image

features, the translator accurately captures names and their spatial relationships. The extracted role features are then fed into classifiers to predict the nouns, bounding boxes, and the existence of these boxes. This grounded noun prediction method ensures that the model can efficiently handle the spatial grounding of nouns corresponding to the predicted roles of the activity.

### 2.6. Training losses

In order for our model to be accurate, we employ multiple loss functions so that we can predict both actions (verbs) and roles (nouns) correctly. Specifically, we use:

- **Action Classification Loss (Cross-Entropy Loss):** This loss calculates the cross-entropy between the predicted action (verb) distri-

bution and the ground truth. By doing so, the model ensures that it predicts the action of the video accurately at every frame.

- **Noun Classification Losses (Cross-Entropy Losses):** For the noun estimation, three separate noun classification losses are applied, corresponding to different stages of the model:
  - The first noun loss is computed from the initial noun predictions, focusing on role tokens.
  - The second loss is computed using the role and entity relations to refine the noun predictions.
  - The final noun classification loss is computed during the final grounding process to ensure the correct entity is linked to the predicted action.
- **Bounding Box Existence Loss (Cross-Entropy Loss):** This loss ensures that the model correctly predicts whether a role has a corresponding bounding box (for grounded situations). Bounding boxes are important when some entities (such as backgrounds or occluded objects) are not visible.
- **Bounding Box Regression Losses (L1 Loss and GIoU Loss):** These losses are used for accurate bounding box predictions. The L1 loss measures the distance between the predicted box and the ground truth, while the Generalized Intersection over Union (GIoU) loss ensures that the predicted box overlaps sufficiently with the ground truth.

The total training loss is a linear combination of these individual losses, optimized together to achieve accurate action and noun predictions as well as precise bounding box placements.

## 2.7. Training RoadSitu

In our research, the RoadSitu training process to suit the specific needs of road situation classification using video data. RoadSitu model: Road Video dataset with verbs and roles annotated in each frame was used to train the RoadSitu model. This included, instantiating the model parameters consists of learnable embeddings for the encoders and decoders is proceeded with training them using multi-head self-attention and cross-attention mechanisms. On both verbs and nouns, we applied a label smoothing regularization to the model to allow them to generalize on our dataset better by blending factors specifically designed for our dataset. The optimizer used was AdamW, which facilitated effective gradient updates while mitigating overfitting through weight decay. Training involved a careful balance of loss coefficients, incorporating verb classification loss, noun classification loss, box existence prediction loss, and box regression losses to guide the model in learning robust representations for verb and noun predictions. Using a 16 GB GeForce RTX 4080 GPU for effective computation, our model was refined across 200 epochs with an 8-batch size during the training phase. To ensure thorough model evaluation, the dataset was divided into training, validation, and testing sets in a 70-15-15 ratio. Through the implementation of these tactics, we were able to develop an optimized model that can reliably capture and anticipate intricate road scenarios from video data.

## 3. Results

To verify the effectiveness of our model, we conducted extensive experiments on the road video dataset. The RoadSitu model was evaluated on this dataset, which was constructed by adding bounding-box annotations to each frame of the video. Each image was paired with a corresponding verb annotation and roles annotated for each frame.

### 3.1. Evaluation metric

To assess the performance of the RoadSitu model, we employed a comprehensive set of evaluation metrics designed to measure both the accuracy and the robustness of verb and noun predictions.

### Metric Details:

- **Verb Classification Accuracy:** This metric evaluates how accurately the model predicts the verb (action) for each frame in the video. It measures the proportion of correctly predicted verbs against the total number of frames, providing a direct measure of the model's ability to understand the primary actions occurring in the road scenes.
- **Noun Classification Accuracy:** This measures the accuracy of predicting nouns associated with the semantic roles in the scene. Since situation recognition involves identifying the roles of various entities (e.g., agent, object, location), this metric is crucial in assessing how well the model captures these relationships.
- **Bounding Box Precision and Recall:** For grounded noun prediction, we use precision and recall metrics to evaluate the model's performance in predicting the spatial locations of objects. Precision measures the proportion of correctly predicted bounding boxes out of all predictions made, while recall measures the proportion of correctly predicted bounding boxes out of all actual objects. These metrics help in understanding the model's ability to localize objects accurately within the frames.
- **Overall F1-Score:** To provide a balanced evaluation, we use the F1-score, which is the harmonic mean of precision and recall. This metric offers a single measure that balances both false positives and false negatives, offering a more holistic view of the model's performance in identifying verbs and their associated roles.

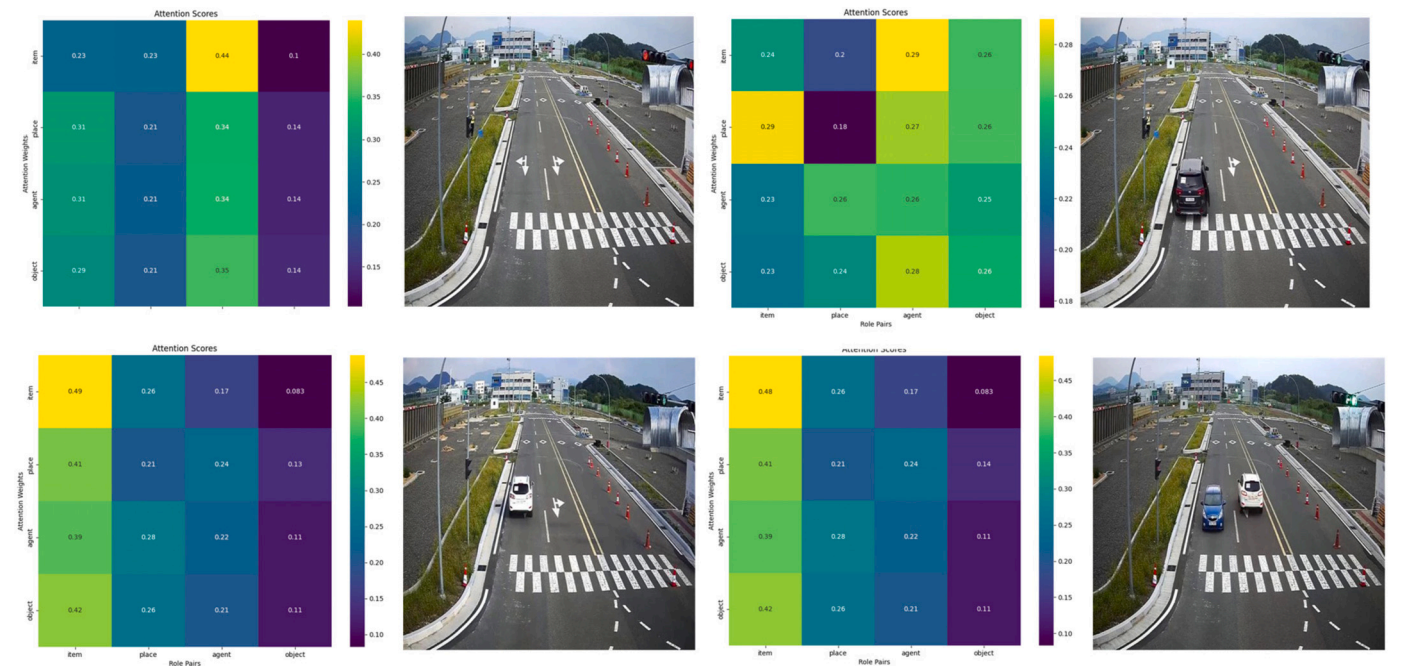
### Evaluation Settings:

- **Dataset Split:** The dataset was split into three subsets: 70% for training, 15% for validation, and 15% for testing. This distribution ensures that the model has sufficient data for learning while also having a separate set for performance evaluation.
- **Training and Validation:** During training, the model's performance was monitored using the validation set to ensure it generalizes well and does not overfit the training data. The best-performing model, in terms of validation metrics, was selected for testing.
- **Testing:** The final evaluation was conducted on the test set, which contains unseen data. This step provides an unbiased measure of the model's effectiveness in real-world scenarios by evaluating its ability to predict verbs and roles accurately across various road situations.

We used the ImageNet-pretrained ResNet-50 backbone [49] without the Feature Pyramid Network (FPN) [50]. The ResNet-50 backbone produces image features  $X_{img} \in \mathbb{R}^{c \times h \times w}$  where  $c = 2048$ . The hidden dimensions of each semantic role query, verb token, and image feature are 512 ( $d = 512$ ). The embedding dimensions for the verb and role tokens are set to 256 ( $d_v = d_r = 256$ ). We utilized learnable 2D embeddings for positional encodings, and the number of attention heads for all multi-head self-attention (MHSA) and multi-head attention (MHA) blocks was set to 8.

We used 2 fully connected layers with a ReLU activation function for the following components: the feed-forward network (FFN) blocks in the encoder and decoder, the verb classifier, the noun classifier, and the bounding box existence predictor. The hidden dimensions for these components were set to 2048, 2d, 2d, and 2d, respectively. The corresponding dropout rates were 0.15, 0.3, 0.3, and 0.2. The bounding box regressor consisted of 3 fully connected layers with ReLU activation and 2d hidden dimensions and used a dropout rate of 0.2. Label smoothing regularization [51] was applied to both verb and noun labels, with smoothing factors of 0.3 and 0.2, respectively.

The model was trained using the AdamW optimizer [52] with a learning rate of  $10^{-4}$  (and  $10^{-5}$  for the backbone), weight decay of  $10^{-4}$ , and  $\beta$  parameters set to  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . Max gradient clipping was set to 0.1, and the BatchNorm layers in the backbone were



**Fig. 10.** We present the attention scores that are calculated within RoadSitu’s attention layers visually. This visualization shows the attention scores between the semantic roles computed in the last decoding layer’s block. A column-wise sum of 1 is used to represent attention scores.

trained. The training process spanned 200 epochs with a batch size of 8, using a 16 GB GeForce RTX 4080 GPU. Training took approximately 8 hours to complete. The loss coefficients were set to  $\lambda_{\nu} = \lambda_n = 1$  and  $\lambda_{exits} = \lambda_{L1} = \lambda_{GloU} = 5$ . The dataset was divided into 70% for training, 15% for validation, and 15% for testing. All models were implemented using the PyTorch framework [53].

Fig. 10 shows four images representing different contexts for the verb “driving\_reverse”. This image highlights objects, agents, locations (place), and element (item) roles. In particular, when the verb “driving\_reverse” exists, the role of “place” (i.e., road) is strongly linked to that of “agent” (car) and “manner” (reverse direction). However, the first image in Fig. 10 focuses on the role of “place” where no verb appears. This suggests that the relationship between roles can be understood adaptively depending on the context of each image. On the other hand, the image in the right corner shows the two “agents” (car) and “manner” (reverse direction) and (right direction). This complex situation could lead to more unpredicted results in a situation classification by the model. Because of the context and intensity of the video frames from the road video data, it is important to understand the meaning inside the image.

As shown in Fig. 11, the RoadSitu model is evaluated on a street scene with no visible agents or objects. In such cases, the model has difficulty predicting the verb and its function. This scenario highlights one of the limitations of the model when faced with scenes without salient entities or behaviors. As the light blue box shows, even if the model correctly identifies a location, it cannot make any further predictions about the agent, object, or method.

This result demonstrates the model’s dependence on the presence of recognizable elements to make accurate predictions. Due to the lack of visible agents and objects (e.g., cars), the model cannot determine appropriate verbs and roles, requiring additional improvements in situations where scene context is minimal or ambiguous. These cases include incorporating additional situational cues or leveraging temporal information from previous frames to improve model performance in predicting road conditions when direct visual measurements are not available.

Fig. 12 illustrates the prediction results of our RoadSitu model on a sequence of frames from the road video test dataset. The figure presents

a progressive analysis of the model’s ability to recognize the verb and identify the associated roles in a road situation where a car is driving in reverse.

- **Frame-by-Frame Analysis:** The figure shows a sequence of frames where the car progressively moves in reverse direction. In the initial frames, the model struggles to detect the verb and the agent accurately, as indicated by the empty or incorrect values in the “Verb” and “Role” columns. As the sequence progresses, the model improves in its predictions, successfully identifying the verb as “driving reverse” and the agent as “car.” This illustrates how the model’s predictions evolve over time and context.
- **Correct Predictions:** As the frames advance, the model correctly identifies the verb “driving reverse” and the corresponding roles, including the agent (“car”) and location (“road”). This demonstrates the model’s ability to learn from the context and make accurate predictions as more visual information becomes available.
- **Incorrect Grounding and OOV Cases:** In some frames, sky blue boxes indicate incorrect grounding predictions or out-of-vocabulary (OOV) cases. For instance, the model fails to detect certain objects or assigns incorrect labels, which are marked with ‘OOV’ (out of vocabulary). These errors are highlighted to show the model’s limitations in handling ambiguous or less common scenarios within the dataset.
- **Role Prediction Challenges:** Incorrect or unpredicted role predictions are highlighted in red. In the early frames, the model has difficulty predicting certain roles, such as the “object” or “manner.” This indicates that while the model can identify the main action (verb) with some degree of accuracy, predicting specific roles, especially in complex or ambiguous contexts, remains challenging.

Overall, Fig. 12 illustrates the RoadSitu model’s ability to interpret complex road scenarios in a variety of environments and demonstrates the power of learning situational information to improve predictions as well as OOV cases and complex role predictions. Highlights the two limits of treatment. This analysis highlights the need for additional improvements to address diverse and complex real-world situations.

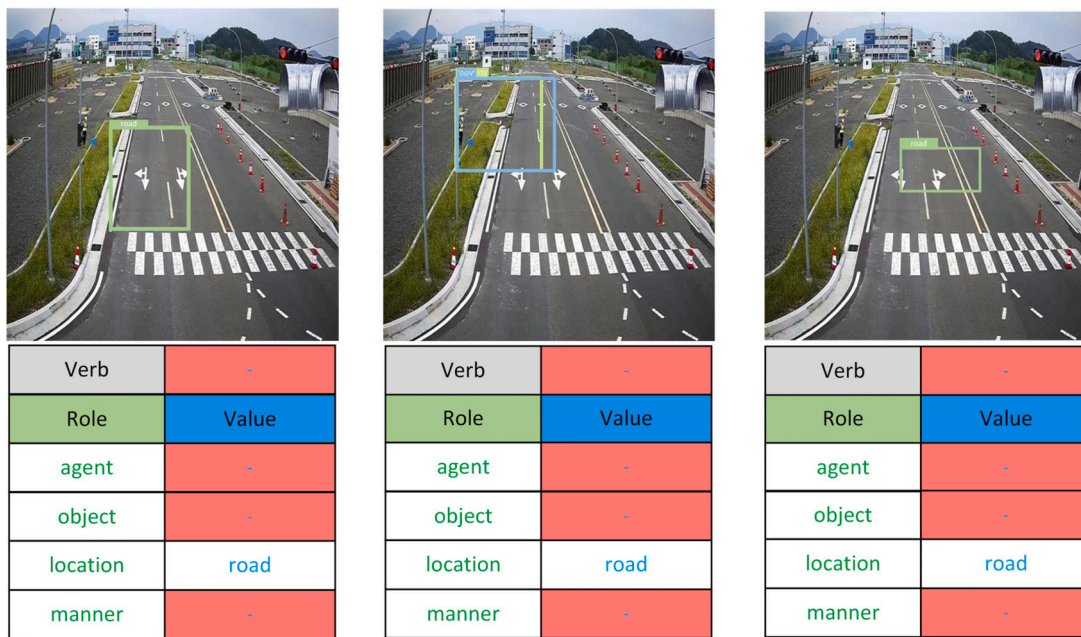


Fig. 11. Prediction results of the RoadSitu model on a road scene with no visible agent or object. The model is unable to predict the verb and corresponding roles due to the absence of detectable elements in the scene. Sky blue boxes indicate the location identified, but no further predictions are made for the agent, object, or manner.

Table 3

Performance comparison of RoadSitu with other state-of-the-art models on the road video dataset.

Model	Top-1 Verb Accuracy (%)	Top-5 Verb Accuracy (%)	Value Accuracy (%)	Grounded Role Accuracy (%)
RoadSitu (Ours)	43.46	72.48	34.21	28.11
SituFormer [44]	42.52	71.01	33.55	27.91
CoFormer [39]	42.31	71.98	33.87	27.78
GSRFormer [45]	41.84	71.03	32.69	27.51
GSRTR [43]	41.06	69.46	32.52	26.04
JSL [38]	39.60	67.71	31.18	25.03
ISL [38]	38.83	65.74	30.47	22.47

**Quantitative Evaluation on the Road Video Dataset:**

Table 3 presents a quantitative evaluation of the RoadSitu model compared to existing models such as GSRTR, JSL, and ISL on the road video dataset. The evaluation metrics used include Top-1 Verb Accuracy, Top-5 Verb Accuracy, Value Accuracy, and Grounded-Value Accuracy.

- **Top-1 and Top-5 Verb Accuracy:** Our RoadSitu model achieved a Top-1 Verb Accuracy of 43.46%, surpassing GSRTR’s 41.06% and other baseline models. The Top-5 Verb Accuracy of RoadSitu also showed significant improvement, reaching 72.48%, which indicates the model’s enhanced ability to predict the correct verb within the top five choices.
- **Value Accuracy:** RoadSitu demonstrated an increase in Value Accuracy, achieving 34.21%, and in Value-All Accuracy with 21.19%. These metrics evaluate the model’s performance in predicting the correct noun values associated with each semantic role. The improved accuracy highlights RoadSitu’s capability to effectively understand and label the various entities present in road scenarios.
- **Grounded-Value Accuracy:** RoadSitu further outperformed other models in terms of Grounded-Value Accuracy with scores of 28.11%. These metrics reflect the model’s ability to predict and localize entities correctly within the frames, confirming the model’s effectiveness in grounded situation recognition.

Overall, the RoadSitu model performs better on all evaluation metrics compared to GSRTR and other benchmark models. This demonstrates the robustness and effectiveness of the model in accurately de-

tecting and classifying complex road situations and further confirms the effectiveness of the proposed approach.

**4. Ablation study of RoadSitu**

In order to further validate the contribution of different components in our RoadSitu model, an ablation study was conducted in which key components were systematically removed or modified in order to assess their impact on performance. The results are summarized in Table 4.

- **w/o Stage 1 Transformer:** An initial stage in a transformer responsible for extracting features and predicting actions (verbs). Consequently, Top-1 Predicted Verb and Ground-Truth Verb accuracies dropped significantly, illustrating the need for early-stage processing to make accurate verb predictions.
- **w/o Stage 2 Transformer:** It was found that the Top-5 Predicted Verbs accuracy decreased significantly when the second transformer, which refines the relationship between roles and actions, was removed, showing the importance of this stage for capturing the relationship between actions and the roles they are associated with.
- **w/o Noun Classifiers on Transformer:** Without the noun classifiers, which predict the roles (nouns) in conjunction with the verb, we observed a significant reduction in the overall Value Accuracy and Ground-Truth Verb Accuracy. This indicates that the interaction between verb prediction and noun estimation is essential for accurate role identification.

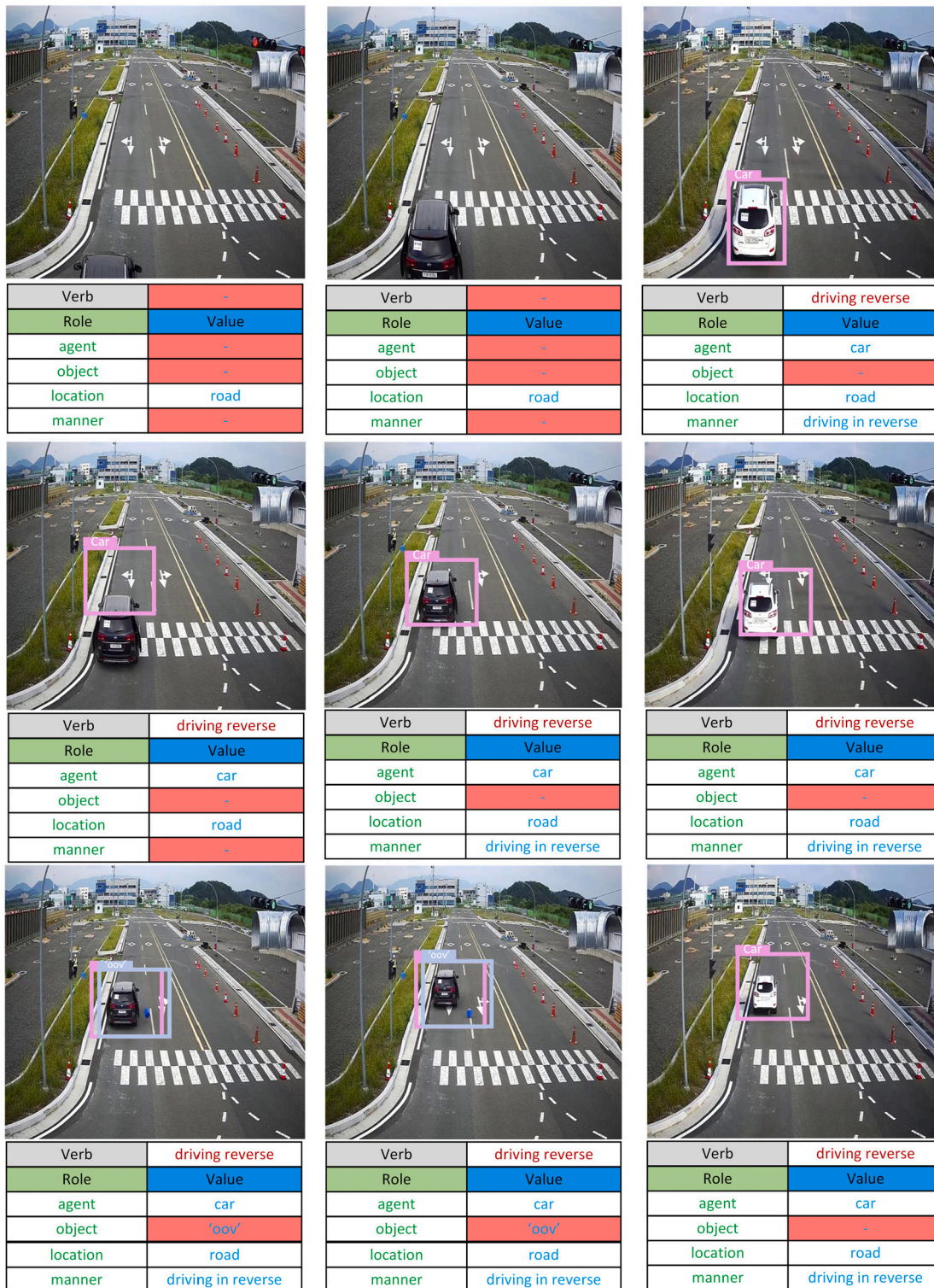


Fig. 12. Prediction results of our RoadSitu model on the Road Video test dataset. Sky blue boxes indicate incorrect grounding predictions or out-of-vocabulary cases. Incorrect or unpredicted role predictions are highlighted in red.

**Table 4**

Ablation study of RoadSitu on the road video dataset. The contributions of different components used in our model are evaluated.

Method	Top-1 Verb Accuracy (%)	Value Accuracy (%)	Top-5 Verb Accuracy (%)
w/o Stage 1 Transformer	41.46	32.21	69.89
w/o Stage 2 Transformer	40.12	30.24	67.17
w/o Noun Classifiers on Transformer	41.30	32.33	69.76
w/o Gradient Flow from Transformer	42.96	32.57	70.97
w/o Verb Token in Transformer	42.76	32.42	71.34
RoadSitu (Ours)	43.46	34.21	72.48

- **w/o Gradient Flow from Transformer:** Taking away the gradient flow between the stages of the transformer slightly reduced all performance metrics, suggesting that the communication between transformers is essential for predicting roles and actions accurately.
- **w/o Verb Token in Transformer:** In addition to reducing performance, the elimination of the verb token, which encodes the predicted action, also reduced performance. The model's inability to predict accurate verbs and roles without this token emphasizes the necessity of incorporating verb tokens for better context understanding.
- **Full Model (RoadSitu):** The full model achieved the highest accuracies across all metrics, demonstrating the importance of each component in the pipeline. Despite its abbreviated version, RoadSitu outperforms its ablated counterpart with a Top-of-the-Line Predicted Verb accuracy of 43.46% and a Ground-Truth Verb accuracy of 72.48%.

These results highlight the contribution of each component to the overall performance of RoadSitu, validating the effectiveness of the three-stage transformer model for road situation recognition.

## 5. Discussion

Situation Recognition (SR) models produce structured output, usually actions (activities) and their nouns (entities/roles), and play an important role in understanding videos/images. This is particularly useful for tasks that require accurate, interpretable, and actionable information. Representative examples include autonomous driving and surveillance. SR models like RoadSitu are optimized for domain-specific tasks, with a focus on producing highly accurate predictions about well-defined roles and behaviors.

In contrast, the rapid development of large multimodal language models (LLMs) has enabled AI models to generate human-like free-text descriptions for image and video inputs. Although the LLM offers flexibility and rich contextual explanations, its objectives differ from those of the SR model. In this section, we discuss the advantages and limitations of comparing SR models and multimodal LLMs and highlight how each approach addresses the unique requirements of video/image understanding.

### Situation Recognition vs. Multimodal LLMs:

- **Structured Output for Actionable Insights:** Unlike multimodal LLMs that produce free-form textual descriptions, situation awareness typically aims to produce structured outputs such as a single verb (action) or multiple nouns (entities/roles). This structure is particularly useful for applications that require clear, interpretable, and actionable information, such as autonomous driving and monitoring, where ambiguity can lead to costly or dangerous misunderstandings.
- **Efficiency in Domain-Specific Tasks:** Models of situational recognition are made to maximize performance in a particular domain, like a road scene. Compared to the more flexible and comprehensive LLM, the findings are more intensive and computationally efficient. This emphasis is especially helpful in situations where judgments must be made quickly, like real-time road condition analysis.

- **Focused Semantic Understanding:** Multi-model LLMs can provide a broader and more flexible explanation, but their general utility may reduce the precision of domain-specific operations. Situational awareness models like RoadSitu, on the other hand, are optimized to understand and predict the precise relationships between verbs and their entities in limited scenarios, enabling greater accuracy in specialized applications.

### Limitations of Image Situation Recognition:

**Limited Expressiveness:** A notable limitation of situation recognition models is their constrained output, typically limited to one verb and multiple nouns. While this structure ensures clarity and reduces ambiguity in the model's predictions, it restricts the ability to generate more nuanced or complex explanations that multimodal LLMs are capable of providing. This limitation can be particularly noticeable in scenarios where rich, descriptive text or more context is required to fully capture the complexity of a situation.

In summary, multimodal LLMs offer flexibility and richness in natural language generation, making them suitable for tasks that require detailed and descriptive outputs. In contrast, situation recognition models provide the structure and precision needed for tasks that require clear, interpretable outputs with minimal ambiguity. Both approaches have their unique advantages and are applicable depending on the specific requirements of the task at hand in the broader context of video/image understanding.

## 6. Conclusion

In this research, we propose a new transformer-based three-step model for road scenario detection from road video data, called RoadSitu. Our method effectively integrates noun estimate and action prediction into a complimentary framework. The model shows notable increases in accuracy and efficiency across all assessment criteria by utilizing the converter's robust feature extraction, attention mechanisms, and machine learning capabilities. Our findings demonstrate how crucial road situational awareness forecasts are to the analysis of traffic situations. The model's accuracy in identifying the semantic roles linked to commands like "stop vehicle" and "driving reverse" serves as evidence of this. More steady and accurate predictions have been made possible by the addition of multi-head internal attention and the normalization of earlier layers, particularly while handling a variety of intricate road scenarios. Extensive experiments on real road data sets confirm the effectiveness of the proposed method and find superior performance compared to other models in the field. These results suggest that RoadSitu can be used to improve road safety by enabling advanced driver assistance systems to better understand and respond to complex traffic scenarios. The experimental results on a challenging road video dataset have shown that our approach tries to achieve complex verb and role prediction, obtaining better relationships between the verb and associated semantic roles. However, from the experiments, it is evident that verb prediction and corresponding semantic roles remain a major challenge for road situation recognition due to the scarcity of the data, the complexity of the data and defining the proper roles. Despite these advances, further improvements are needed to enhance the robustness of the model in such scenarios. Future work will explore expanding the

dataset with annotation data and further refining the model's ability to generalize across a broader range of road conditions and video inputs. In future work, we plan to explore how pseudo-sentence generation and evaluation metrics like BLEU can be incorporated into our framework to provide a more nuanced evaluation of the model's outputs.

### CRedit authorship contribution statement

**Subhajit Chatterjee:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Hoorang Shin:** Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Data curation. **Joon-Min Gil:** Validation, Supervision, Project administration, Investigation, Funding acquisition, Formal analysis. **Yung-Cheol Byun:** Visualization, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgements

This research was supported by “Regional Innovation Strategy (RIS)” through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2023RIS-009), and this work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00405278).

### Data availability

The data that has been used is confidential.

### References

- [1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
- [2] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallocci, A. Kolesnikov, et al., The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale, *Int. J. Comput. Vis.* 128 (7) (2020) 1956–1981.
- [3] K.-H. Kim, S. Hong, B. Roh, Y. Cheon, M. Park, Pvanet: deep but lightweight neural networks for real-time object detection, arXiv preprint, arXiv:1608.08021.
- [4] Z. Zhao, H. Ma, S. You, Single image action recognition using semantic body part actions, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3391–3399.
- [5] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint, arXiv:1409.1556.
- [6] S. Maji, L. Bourdev, J. Malik, Action recognition from a distributed representation of pose and appearance, in: CVPR 2011, IEEE, 2011, pp. 3177–3184.
- [7] Z. Lu, L. Wang, Learning descriptive visual representation for image classification and annotation, *Pattern Recognit.* 48 (2) (2015) 498–508.
- [8] S. Gupta, J. Malik, Visual semantic role labeling, arXiv preprint, arXiv:1505.04474.
- [9] M. Yatskar, L. Zettlemoyer, A. Farhadi, Situation recognition: visual semantic role labeling for image understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5534–5542.
- [10] J. Carreira, A. Zisserman, Quo vadis, action recognition? A new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
- [11] D. Girish, V. Singh, A. Ralescu, Understanding action recognition in still images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 370–371.
- [12] J. Guo, J. Li, D. Li, A.M.H. Tiong, B. Li, D. Tao, S.C. Hoi, From images to textual prompts: zero-shot VQA with frozen large language models, arXiv preprint, arXiv:2212.10846.
- [13] L. Ke, W. Pei, R. Li, X. Shen, Y.-W. Tai, Reflective decoding network for image captioning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8888–8897.
- [14] W. Kim, B. Son, I. Kim, Vilt: vision-and-language transformer without convolution or region supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 5583–5594.
- [15] D. Xu, Y. Zhu, C.B. Choy, L. Fei-Fei, Scene graph generation by iterative message passing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5410–5419.
- [16] Y. Cong, W. Liao, H. Ackermann, B. Rosenhahn, M.Y. Yang, Spatial-temporal transformer for dynamic scene graph generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16372–16382.
- [17] G. Gkioxari, R. Girshick, P. Dollár, K. He, Detecting and recognizing human-object interactions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8359–8367.
- [18] J. Lim, V.M. Baskaran, J.M.-Y. Lim, K. Wong, J. See, M. Tistarelli, Enet: an efficient and reliable human-object interaction detection network, *IEEE Trans. Image Process.* 32 (2023) 964–979.
- [19] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, J. Feng, Ppdm: parallel point detection and matching for real-time human-object interaction detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 482–490.
- [20] Y. Su, C.-C.J. Kuo, On extended long short-term memory and dependent bidirectional recurrent neural network, *Neurocomputing* 356 (2019) 151–161.
- [21] Y. Su, Y. Huang, C.-C.J. Kuo, Dependent bidirectional RNN with extended-long short-term memory, 2018.
- [22] Y. Su, K. Fan, N. Bach, C.-C.J. Kuo, F. Huang, Unsupervised multi-modal neural machine translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10482–10491.
- [23] Y. Sua, R. Lina, C.-C.J. Kuo, On tree-structured multi-stage principal component analysis (tmpca) for text classification, arXiv preprint, arXiv:1807.08228.
- [24] Y. Su, R. Lin, C.-C.J. Kuo, Tree-structured multi-stage principal component analysis (tmpca): theory and applications, *Expert Syst. Appl.* 118 (2019) 355–364.
- [25] Y. Su, Y. Huang, C.-C.J. Kuo, Efficient text classification using tree-structured multi-linear principal component analysis, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 585–590.
- [26] Y. Su, C.-C.J. Kuo, et al., Recurrent neural networks and their memory behavior: a survey, *APSIPA Trans. Signal Inf. Process.* 11 (1) (2022) e26.
- [27] H. Wu, Z. He, M. Gao, Gcevt: learning global context embedding for vehicle tracking in unmanned aerial vehicle videos, *IEEE Geosci. Remote Sens. Lett.* 20 (2022) 1–5.
- [28] H. Wu, J. Nie, Z. Zhu, Z. He, M. Gao, Leveraging temporal-aware fine-grained features for robust multiple object tracking, *J. Supercomput.* 79 (3) (2023) 2910–2931.
- [29] H. Wu, J. Nie, Z. Zhu, Z. He, M. Gao, Learning task-specific discriminative representations for multiple object tracking, *Neural Comput. Appl.* 35 (10) (2023) 7761–7777.
- [30] H. Wu, J. Nie, Z. He, Z. Zhu, M. Gao, One-shot multiple object tracking in UAV videos using task-specific fine-grained features, *Remote Sens.* 14 (16) (2022) 3853.
- [31] G.A. Miller, Wordnet: a lexical database for English, *Commun. ACM* 38 (11) (1995) 39–41.
- [32] C.F. Baker, C.J. Fillmore, J.B. Lowe, The Berkeley framenet project, in: COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics, 1998.
- [33] C.J. Fillmore, C.R. Johnson, M.R. Petrucci, Background to framenet, *Int. J. Lexicogr.* 16 (3) (2003) 235–250.
- [34] A. Mallya, S. Lazebnik, Recurrent models for situation recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 455–463.
- [35] P. Vicol, M. Tapaswi, L. Castrejon, S. Fidler, Moviegraphs: towards understanding human-centric situations from videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8581–8590.
- [36] T. Cooray, N.-M. Cheung, W. Lu, Attention-based context aware reasoning for situation recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4736–4745.
- [37] R. Li, M. Tapaswi, R. Liao, J. Jia, R. Urtasun, S. Fidler, Situation recognition with graph neural networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4173–4182.
- [38] S. Pratt, M. Yatskar, L. Weihs, A. Farhadi, A. Kembhavi, Grounded situation recognition, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, Springer, 2020, pp. 314–332.
- [39] J. Cho, Y. Yoon, S. Kwak, Collaborative transformers for grounded situation recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19659–19668.
- [40] M. Yatskar, V. Ordonez, L. Zettlemoyer, A. Farhadi, Commonly uncommon: semantic sparsity in situation recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7196–7205.
- [41] M. Suhail, L. Sigal, Mixture-kernel graph attention network for situation recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10363–10372.
- [42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, vol. 30, 2017.
- [43] J. Cho, Y. Yoon, H. Lee, S. Kwak, Grounded situation recognition with transformers, arXiv preprint, arXiv:2111.10135.
- [44] M. Wei, L. Chen, W. Ji, X. Yue, T.-S. Chua, Rethinking the two-stage framework for grounded situation recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 2651–2658.

- [45] Z.-Q. Cheng, Q. Dai, S. Li, T. Mitamura, A. Hauptmann, Gsrformer: grounded situation recognition transformer with alternate semantic attention refinement, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 3272–3281.
- [46] D. Roy, D. Verma, B. Fernando, Clipsitu: effectively leveraging clip for conditional predictions in situation recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 444–453.
- [47] Y. Jing, J. Wang, W. Wang, L. Wang, T. Tan, Relational graph neural network for situation recognition, *Pattern Recognit.* 108 (2020) 107544.
- [48] C. Fellbaum, *Wordnet*, in: *Theory and Applications of Ontology: Computer Applications*, Springer, 2010, pp. 231–243.
- [49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
- [50] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
- [51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
- [52] I. Loshchilov, Decoupled weight decay regularization, arXiv preprint, arXiv:1711.05101.
- [53] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch, 2017.