



UvA-DARE (Digital Academic Repository)

Visual understanding of dynamic scenes using object relationships and open vocabularies

Ülger, O.

Publication date
2026

[Link to publication](#)

Citation for published version (APA):

Ülger, O. (2026). *Visual understanding of dynamic scenes using object relationships and open vocabularies*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

7

Summary & Conclusions

7

In this thesis, we have investigated the potential of understanding dynamic scenes visually using object relationships and open vocabularies. In the first part of the thesis, we argued that object relationships can be beneficial to identify objects more efficiently by incorporating the context they appear in through object-object relationships, especially in circumstances where a conventional neural network might struggle to differentiate one class from the other. Such struggles could be related to the object being of a rare semantic category, small in size or occluded.

Our starting point was to explore how relationships can be incorporated to improve the understanding of visual scenes, specifically by treating the relationship between two objects as a predictable attribute. Because our focus is on dynamic scenes, both spatially and temporally, this requires methods that can handle their inherent variability and openness, such as the continuous appearance, disappearance and transformation of objects throughout a scene. In chapter two, we conducted a study into related work and found that existing literature in relationship prediction focused exclusively on scenarios where the graph is static in the number of nodes, each representing an object, throughout the video's spatio-temporal graph. Objects that enter or exit the video dynamically were modelled through padded nodes without explicitly addressing the evolution of the graph throughout time. To address this research gap, we have proposed a novel graph-based method, MTD-GNN, able to predict object relationships in temporally-dynamic graphs. MTD-GNN utilizes a factored spatiotemporal graph attention layer, where temporal and spatial node connections are attended separately in spatial or temporal neighbourhoods respectively. This in turn enabled information propagation across scene graphs even when the number of objects changes over time. Since objects can have multiple relationships with each other at once, we tackled a setting where multiple labels can be predicted for the same object pair.

In chapter three, we shifted our focus from predicting relationships to using them for the benefit of the downstream tasks object detection and instance segmentation. Research has shown that humans have a remarkable ability to perceive and reason about the world around them by understanding and utilizing the relationships between objects. Inspired by this human ability, we formulated a framework in which relational prior knowledge from an existing knowledge base would enhance the predictive capability of a conventional object detection and instance segmentation model. The proposed model, RP-FEM, updates the edge weights of a fully connected scene graph where nodes represent the latent features of object proposal. It does so by attending between the edge values of a relational prior knowledge graph and the edges of the scene graph, based on the features of the nodes the edges are connecting. By iteratively updating the scene graph's edge values and node features consecutively, the latent features of the nodes get updated. The extent to which each node's features are propagated to other nodes thereby became based on relational prior knowledge, allowing for more structured information propagation.

The second part of the thesis has focused on visual scene understanding with open vocabularies. The dynamicness of scenes is not only defined by how objects relate to each other in terms of relationships, but also by the vast semantic range to describe objects. Previous literature has mostly focused on predefined object categories, which limits the amount and specificity of semantic classes that can be covered after training. Dynamicness can also relate to the fact that the real world is open-ended and previously unseen objects may appear at any time at arbitrary numbers. Chapter four introduced the paradigm of the 'auto-vocabulary', a vocabulary automatically generated in zero-shot manner and specific to a scene. The proposed method, AutoSeg, includes an enhanced captioning pipeline that enables object-level captioning via unsupervised semantic grouping of vision-language features, resulting in exhaustive yet semantically precise scene-specific vocabularies. The generated vocabularies serve as effective target classes for an open-vocabulary segmentation method, thereby eliminating the need for manual specification of the target vocabulary. While our method was suitable to address the open-endedness of visual scenes effectively, it suffered from one major shortcoming, which is that its performance heavily relies on ideal lighting conditions. For safety-critical applications, such as autonomous driving, where we deemed AutoSeg most relevant, this could potentially create dangerous situations. Hence, in chapter five, we investigated how LiDAR can be leveraged to alleviate these issues. Our proposed method, 3D-AVS, leverages geometric information from LiDAR to improve the vocabulary generation independent of color information originating from image pixels. Features from points visible in the image are used to train a Sparse Masked Attention Pooling module, which enables the generation of a vocabulary from a point mask during inference while the image remains unavailable. By following this approach, the need to train a point-based vision-language model for direct captioning - something that is hard to achieve due to the absence of internet-

scale combined point and language data - was mitigated. Instead, the complimentary effect of the 3D data's semantic information on the image-level vocabulary was achieved by transferring knowledge from the 2D vision-language model to the 3D backbone.

In both chapter four and five, evaluation of open-ended classes was a crucial and challenging topic. The two proposed auto-vocabulary methods perform segmentation with arbitrary classes from a vast semantic space that roughly corresponds to the vision-language model's internet-scale training data. Its qualitative outputs were often accurate, or at the very least plausible. However, to compare our method to existing open-vocabulary segmentation methods, where vocabularies are fixed or specified by a human, it was required to translate our open-ended segmentation outputs to conventional performance metrics. This posed a challenge: open-ended predictions are not always directly comparable with annotated classes. In some cases, predictions of segments that our methods made were not annotated at all, resulting in the penalization of such complimentary predictions. To address this evaluation issue, in chapter four an LLM-aided evaluation protocol was proposed to create a most likely mapping between the predicted classes and annotated classes in order to calculate mIoU. In chapter five, a new Text-Point Semantic Similarity metric was proposed to directly measure the extent to which the vision-language embedding of a point aligns with the vision-language embedding of an auto-class. In essence, it was measuring not only how accurate our method was able to name objects, but also how imprecise some annotated classes were for certain types of objects.

The sixth chapter of the thesis aims to unify the problem settings of the two parts: videos that are dynamic throughout time and require an open vocabulary to be maintained. Tackling an online setting of videos, where frames are registered one at a time, the proposed Vocabulary-Free Online Video Instance Segmentation aims to directly label segments with open-ended classes without creating a vocabulary as an intermediate step. This was achieved by calculating a latent token specific to a region using a conditional vision tokenizer, indicated via a class-agnostic mask. The combination of temporal dynamicness and open vocabularies posed an additional challenge, namely that of semantic change over time. Since new objects may appear throughout the video, such new objects need to be segmented, requiring the model to run inference again. Since this can be costly for long videos with a high framerate, it is relevant to only do this when necessary, i.e. when a semantic change occurs. To this end, we proposed VLChange, a vision-language token-based change detector that triggers the segmentation only when semantic shifts occur.

7.1 Conclusions

The goal of this thesis was to study visual understanding of dynamic scenes through two lenses: the structure provided by relationships between objects and the flexibility open vocabularies. Together, these lines of work aim to bring machine learning models closer to the way people reason in changing environments, where both the set of objects and their meanings can evolve over time.

Chapter two developed MTD-GNN to predict relations in graphs where the set of entities changes as a video unfolds. Rather than fixing the node set and padding for missing objects, the model attends over links in space and in time separately while supporting multi-label relation prediction for each object pair. Experiments on CLEVRER and Action Genome showed that this factorized attention preserves information flow even as objects appear, disappear or interact in new ways. A practical lesson from this chapter is that relation inference remains sensitive to missed or false detections from the backbone. A promising next step is to couple detection and relation modeling more tightly so that relation evidence can help recover weak detections.

Chapter three turned from predicting relations to using them to improve detection and instance segmentation. RP-FEM injects relational prior knowledge from a knowledge base into the scene graph's edges formed by region proposals. By attending between prior edges and scene edges, and by iteratively updating edges and node features, the method amplifies features that are consistent with likely context and suppresses features that are unlikely to co-occur together. This has shown to reduce duplicate predictions, yield more reliable instance counts and give large gains for rare classes. The approach does, however, depend on the coverage and quality of the external knowledge base. Future work can alleviate that dependence by learning priors from text and video at larger scale and by allowing the model to question or refine priors when the scene provides strong counter evidence.

Chapters four and five addressed a different kind of dynamicness: the open-ended nature of semantics in the real world. Chapter four introduced AutoSeg, which builds a scene-specific vocabulary by grouping vision-language features and producing object-level captions, then uses that vocabulary as targets for open vocabulary segmentation. As open-ended predictions rarely match annotations one-to-one, making comparison with state-of-the-art challenging, the chapter also proposed LAVE. LAVE is a mapping protocol that aligns predicted names with labeled classes to compute standard metrics. The method achieved strong results in the zero-label setting while remaining competitive with methods that rely on provided vocabularies. Its main limitation is the reliance on image quality and lighting, which can weaken the model's ability to predict semantic clusters used for captioning.

Chapter five aimed to resolve that limitation by leveraging geometric information coming

from LiDAR data. Our solution, 3D-AVS, transfers knowledge from a 2D vision language model to a 3D backbone and uses a sparse masked attention pooling module to generate a vocabulary from point masks even when the image is not available. The work introduced the Text-Point-Semantic-Similarity metric to assess how well point features align with auto-generated classes, which captures semantic quality beyond mask overlap. The results showed that segmentations can be more semantically precise than labels of the human-annotated data set in some cases, while producing accurate masks. This chapter highlights the benefit of combining modalities, where points supply stable geometry when appearance is unreliable and language supplies useful labels when 3D annotations are scarce.

Chapter six tackled the theme of open-ended vocabularies in an online video setting. Our model, FOVIS, performs vocabulary-free instance segmentation on incoming frames by computing a conditional vision token for each region indicated by a class-agnostic mask. To keep inference efficient over long videos, the system uses VLChange to detect semantic shifts and only triggers segmentation when new objects with different semantic classes appear. A foreground grounding stage ensures that the model focuses on object regions rather than background to use class-agnostic masks effectively. The framework reaches state-of-the-art performance on videos with many unseen categories, showing the potential of modeling for an open-ended world.

There are two main take-aways across the six chapters. First, structure matters. Learning and utilizing relationships gives machine learning models *handles* to share evidence across objects, recover from ambiguity and act with fewer proposals. Second, modeling for an open-ended world matters. Automatically generated vocabularies or class labels let machine learning models name objects with what they perceive rather than what humans specify in advance. Evaluation is challenging due to the open-ended nature of predictions, but techniques such as LAVE and the TPSS metric can contribute to fairer comparisons.

An exciting path forward is to combine these ideas into a single architecture. A system that detects and tracks objects online, reasons over relations, builds and revises a scene-specific vocabulary across modalities and allocates computation based on detected change throughout time. This would better match the demands of perception models deployed in the real-world perception. This thesis provides building blocks and evidence that such integration is both feasible and useful.