



## 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**

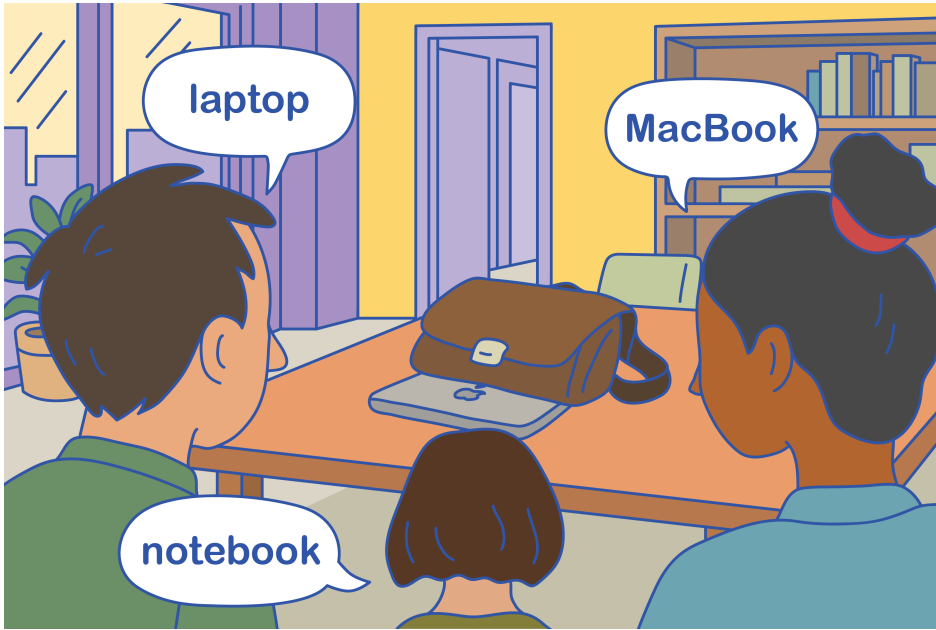
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <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.

# 1

## Introduction

**H**uman visual understanding of scenes is remarkable in its efficiency, flexibility and robustness. People can identify objects rapidly and often effortlessly, even when those objects are partially hidden [1, 2, 3], unusually shaped or oriented [4, 5, 6], poorly lit [7, 8] or unfamiliar in form [5, 9]. For example, we can recognize a car from just its rear bumper that peeks out from behind a hedge. When encountering a stylized, abstract sculpture of a cat in a museum, visitors can still identify it as a cat despite never having seen one rendered in that form. One might recognize their friend’s backpack in a cluttered room simply by the worn corner and a distinctive sticker on it. A child can point out an electric scooter as a “bike” even though it is very different from the traditional bicycle they learned to identify first, or a person can recognize a pan on a stovetop even if it is partially obscured by rising steam and surrounded by cooking utensils. While these examples demonstrate our ability to generalize from incomplete or unfamiliar appearances, human understanding is also remarkably flexible in how we refer to objects. People often use different names even when referring to the same item. For instance, three different adults might describe the same device as a “laptop”, a “notebook”, a “MacBook” or simply a “computer” depending on context, habit or personal background.

This impressive ability suggests that visual understanding relies on a complex interplay of perceptual and cognitive processes. Theories of object recognition often emphasize different underlying mechanisms and representations. Some propose recognition-by-components [9, 10, 11], while others focus on feature-based matching or exemplar-driven processes [12, 13, 14]. However, rather than assuming a single explanation, some researchers argue that object recognition is likely to be based on multiple partially distinct but interacting systems [15, 16]. These systems may be optimized for different goals such as recognizing a known object under occlusion, identifying a new category exemplar or distinguishing fine-grained details within familiar items.



**Figure 1.1:** Even when partially occluded, a MacBook laptop is easily recognized. Not just by its shape and color, but also by other objects in its context. At the same time, different individuals may refer to the same item in different ways: one might call it a laptop, another a MacBook and yet another a notebook. Illustration: Guanyan Wu and Osman Ülger.

In recent decades, the field of computer vision has made significant efforts to meet human visual understanding capabilities using machine learning and deep learning. Convolutional neural networks (CNNs), attention-based architectures, multimodal learning and other deep architectures have enabled substantial progress in image classification, object detection, and segmentation tasks. CNNs, such as AlexNet, ResNet or UNet [17, 18, 19], exploit local spatial structure through hierarchical feature extraction, allowing models to detect edges, textures and object parts, making them well-suited for tasks like image classification and object detection. Attention-based architectures, such as Vision Transformers (ViT) [20] and DEtection TRansformer (DETR) [21], introduce global context modeling, enabling better handling of spatially dispersed features and improving performance in dense prediction tasks like segmentation and object tracking. Multi-modal learning approaches, like CLIP [22] and BLIP [23], align visual and textual representations, enabling zero-shot classification, open-vocabulary recognition, and more human-like grounding of object semantics. Other deep architectures, such as diffusion models, have introduced new capabilities in image generation. These images may in turn be used as powerful priors and adapted to visual scene understanding tasks.

These models can now identify thousands of object categories with near-human accuracy

thanks to large-scale datasets and end-to-end training strategies. However, such models mostly excel under standard conditions in a constrained problem setting, where many parameters - such as the types of attributes to detect, the number of objects and the object categories - are fixed or known in advance. For example, DETR uses a fixed set of object queries, each designed to detect at most one object, and it is trained on a closed vocabulary like the 91 object categories of COCO [24]. As a result, it cannot recognize novel categories at test time unless they were part of the training set, and it always outputs the same number of detection slots, regardless of how many objects are actually present. Furthermore, in supervised setting, most models assume large amounts of annotated data that require manual labor, while also risking the introduction of annotation bias. For instance, in large-scale datasets like Visual Genome [25], which contain thousands of object categories, semantic ambiguity often arises during annotation potentially causing inconsistent labeling. For example, a “sofa” might also be labeled as a “couch” depending on the annotator. Similarly, a small handbag could be annotated as either a “purse” or a “bag”, even though those terms may reflect subtle functional or cultural distinctions. Such inconsistencies introduce noise into the training data and can degrade model performance, especially in fine-grained or open-vocabulary settings. This ambiguity also affects model evaluation, since standard metrics typically assume a single correct label per instance. Hence, a model that predicts “sofa” when the ground truth says “couch” is penalized, even if both are semantically equivalent.

These shortcomings conflict with the richness, scalability and adaptability of human perception, often zero- or few-shot in terms of learning. This, in turn, creates a gap between what models are capable of in constrained settings and what they are capable of in the real world.

This thesis investigates methods that aim to improve the open-endedness of perception systems, making them more useful and applicable for real-world use. To this end, we explore graph-based models, as well as multimodal learning, for the tasks of edge prediction, detection and semantic/instance segmentation in images, video or 3D LiDAR point clouds.

## 1.1 Thesis Outline and Research Questions

This thesis is divided into two parts. Part I explores how alternative scene representations - considering temporal dynamicness and object-object relations - can enhance classification, detection and segmentation with graph neural networks. Part II focuses primarily on the linguistic side of perception, with an emphasis on open-ended perception with automatically generated scene vocabularies.

In Part I, the following research questions are addressed.

**Research Question 1:** *How can we effectively predict multi-relational edge labels in temporally-dynamic videos, i.e. where entities may enter and/or exit?*

Graph Neural Networks (GNN) are a common technique for visual recognition and understanding. A common assumption with GNNs, however, is that entities remain static throughout video sequences, while in practice, it is common for objects to enter and exit scenes over time. In turn, their visual relations and interactions too evolve dynamically, prompting the need for GNNs that can handle such temporal dynamics. While previous works in research foci, such as skeleton-based action recognition [26, 27] and object-action relations in video [28, 29], have investigated dealing with the temporal dynamics of a graph, there is typically one set of entities that remains present in the scene over time. In **Chapter 2**, we introduce the task of future state multi-relational edge label prediction in temporally-dynamic graphs and propose a Multi-task Temporally-Dynamic Graph Neural Network (MTD-GNN) model. This graph network is centered around a factorized spatio-temporal graph attention layer, inspired by static graph attention.

Addressing this new task, where multiple relations are predicted as edge labels in graphs, inspired follow-up research into whether prior knowledge about such object-object relations can enhance standard tasks such as detection or instance segmentation, prompting the following research question.

**Research Question 2:** *How can relational prior knowledge be leveraged for detection and instance segmentation in images?*

Understanding the relationships between objects give humans a remarkable ability to perceive and reason about the world. In circumstances where a particular object is small, obscured or has a different shape and color from previously seen forms of it, a person can still reason about the object's semantics by its surrounding context of other objects. For example, imagine seeing a small,

unclear object on a beach next to a coconut tree. In that context, you are unlikely to interpret it as a bowling ball, and instead, would more reasonably assume it is a fallen coconut. To investigate how useful such object-object relations in neural networks, **Chapter 3** studies its effects on object detection and instance segmentation by leveraging relational prior knowledge automatically extracted from existing datasets. To this end, we propose a Relational Prior-based Feature Enhancement Model (RP-FEM), a graph transformer that enhances object proposal features using relational priors. The proposed architecture operates on top of scene graphs obtained from initial proposals and aims to concurrently learn relational context modeling for object detection and instance segmentation.

While we investigate the benefits of object relationships for visual understanding, previous work and our contributions still rely on predefined object categories and annotated relational structures. In contrast, the second part of this thesis moves beyond such predefined semantics and investigates how visual systems can operate in open-ended settings, where the set of relevant object categories is unknown beforehand and must be inferred from the visual or contextual cues themselves.

Part II explores the following research questions.

**Research Question 3:** *How can relevant target class names be autonomously identified and subsequently used for semantic segmentation?*

Historically, machine learning research in visual understanding of scenes has focused on fixed vocabularies of target classes to recognize. Recently, more efforts are made to design models that have more open recognition capabilities. One class is Open-Vocabulary Segmentation (OVS) methods, which are capable of performing detection or semantic segmentation tasks without relying on a fixed vocabulary and, instead, can infer classes from an arbitrarily prompted list. While increasing the flexibility of the object classes to be recognized, OVS methods still require a human in the loop to specify the vocabulary based on the task or dataset at hand. **Chapter 4** introduces *Auto-Vocabulary Semantic Segmentation (AVS)*, a new setting for open-ended image understanding where the dependence on human input is eliminated. Our proposed method, *AutoSeg*, leverages vision-language modeling to enable local captioning. To evaluate on open-ended classes, we furthermore propose a new evaluation paradigm aided by a Large Language Model (LLM).

*AutoSeg*'s local captioning approach operates on 2D images. However, when images are taken under challenging lighting conditions, the method's reliance on images alone could be a liability. For safety-critical applications such as autonomous driving, this could lead to

dangerous situations. The next research question dives into the potential benefits of using geometric information from LiDAR to overcome said challenges.

**Research Question 4:** *How can relevant target class names be autonomously identified and subsequently used for LiDAR-based 3D semantic segmentation?*

Existing vision-language models are sensitive to challenging lightning conditions, which causes a degradation in captioning performance. In **Chapter 5**, we discuss a *LiDAR-based 3D Auto-Vocabulary Segmentation* method, *3D-AVS*, which leverages geometric information from LiDAR for improved vocabulary generation. Since LiDAR does not rely on color information, it can serve as a robust complementary modality when color pixels do not yield sufficient information for captioning. To leverage LiDAR, we propose a *Sparse Masked Attention Pooling (SMAP)*, which is trained to aggregate features from points visible in the image. During inference, the image remains unavailable and a group of point masks is generated based solely on geometric information. Finally, a vocabulary is generated from the point masks.

While this and the previous research question focused on improving open-ended visual understanding by automatically generating target vocabularies, the methods follow a two-step approach: first generating the vocabulary, then performing open-vocabulary segmentation. In the next research question, we explore a method that aims to achieve the same goal without generating a vocabulary as an intermediate step.

**Research Question 5:** *How can online video instance segmentation be tackled in a vocabulary-free manner, taking into account temporal semantic changes?*

The previous two research questions introduced frameworks for auto-vocabulary segmentation in 2D images and 3D LiDAR data. However, both relied on a two-step process: first generating a vocabulary, then using it to guide segmentation. In contrast, the method explored in **Chapter 6**, *Vocabulary-Free Online Video Instance Segmentation (FOVIS)*, also follows a two-step approach but reverses the order: it first segments class-agnostic masks and then classifies them open-ended using a vision-language model and an LLM. This decoupling increases flexibility, as it avoids committing to a fixed vocabulary upfront and allows for richer, context-aware labels based on the actual content of each scene. FOVIS is particularly suited to video, where scenes are dynamic and the set of relevant object classes can change over time. A vocabulary generated at the beginning of a video or from a single frame may fail to anticipate newly appearing or reappearing objects. In contrast, FOVIS can adapt to an online setting, continuously discovering and naming new objects as the scene unfolds.

## 1.2 Origins

This thesis is based on the following publications.

### Chapter 2

Multi-Task Edge Prediction in Temporally-Dynamic Video Graphs

Osman Ülger, Julian Wiederer, Mohsen Ghafoorian, Vasileios Belagiannis, Pascal Mettes

In British Machine Vision Conference (BMVC), 2022

*Contribution of authors*

Osman Ülger: all aspects.

Julian Wiederer: conceptualization, technical advice, formal analysis, revision.

Mohsen Ghafoorian: conceptualization, technical advice, formal analysis, revision.

Vasileios Belagiannis: technical advice, revision.

Pascal Mettes: conceptualization, technical advice, formal analysis, revision.

### Chapter 3

Relational Prior Knowledge Graphs for Detection and Instance Segmentation

Osman Ülger, Yu Wang, Ysbrand Galama, Sezer Karaoglu, Theo Gevers, Martin R. Oswald

In International Conference on Computer Vision (ICCV), Scene Graphs and Graph Representation Learning (SG2RL) Workshop, 2023

*Contribution of authors*

Osman Ülger: all aspects.

Yu Wang: technical advice, formal analysis, revision.

Ysbrand Galama: conceptualization, technical advice, formal analysis, revision.

Sezer Karaoglu: technical advice, formal analysis, revision.

Theo Gevers: conceptualization, technical advice, formal analysis, revision.

Martin R. Oswald: conceptualization, technical advice, formal analysis, revision.

### Chapter 4

Auto-Vocabulary Semantic Segmentation

Osman Ülger\*, Maksymilian Kulicki\*, Yuki Asano, Martin R. Oswald

In International Conference on Computer Vision (ICCV), 2025

*Contribution of authors*

Osman Ülger: all aspects.

Maksymilian Kulicki: all aspects.

Yuki Asano: conceptualization, technical advice, formal analysis, revision.

Martin R. Oswald: conceptualization, technical advice, formal analysis, revision.

## Chapter 5

Auto-Vocabulary Segmentation for LiDAR Points

Weije Wei\*, Osman Ülger\*, Fatemeh Karimi Nejadasl, Theo Gevers, Martin R. Oswald  
In Computer Vision and Pattern Recognition Conference (CVPR), 2025

*Contribution of authors*

Weije Wei: all aspects.

Osman Ülger: all aspects.

Fatemeh Karimi Nejadasl: conceptualization, technical advice, formal analysis, revision.

Theo Gevers: conceptualization, technical advice, formal analysis, revision.

Martin R. Oswald: conceptualization, technical advice, formal analysis, revision.

## Chapter 6

FOVIS: Vocabulary-Free Online Video Instance Segmentation

Osman Ülger, Sezer Karaoglu, Theo Gevers, Martin R. Oswald  
In Review, 2025

*Contribution of authors*

Osman Ülger: all aspects.

Sezer Karaoglu: conceptualization, technical advice, formal analysis, revision.

Theo Gevers: conceptualization, technical advice, formal analysis, revision.

Martin R. Oswald: conceptualization, technical advice, formal analysis, revision.

\*equal contribution