

Combining DINOv3 and Mask2Former for Urban Streetscapes Instance Segmentation

Yinghao Chen
yinghao.chen@chalmers.se

Mohammad Albarham
albarham@chalmers.se

Abstract—This paper presents a unified instance segmentation framework (DINOv3-Mask2Former) for urban streetscapes, integrating a self-supervised Vision Transformer (DINOv3) backbone with a Mask2Former head via an enhanced ViT Adapter. Experiments on the Mapillary Vistas dataset validate the feasibility of this approach, showing that DINOv3’s transferable features can be adapted for structured urban scene understanding. The model achieves strong instance segmentation performance on large objects while identifying key challenges in small and occluded instance recognition, underscoring the potential of self-supervised transformers for scalable urban perception.

Index Terms—instance segmentation, urban streetscapes, DINOv3, Mask2Former, Mapillary Vistas

I. INTRODUCTION

Understanding urban environments depends on large-scale, fine-grained, and up-to-date data on streetscapes due to their complexity and dynamics. Traditional approaches [1], such as manual field surveys, long-term observations, and remote sensing, are expensive, labor-intensive, and often geographically constrained. Advances in computer vision and the proliferation of scalable street-level imagery from open geospatial platforms now enable AI-driven and data-driven urban analysis.

Despite the abundance of street-level imagery, translating raw visual data into actionable urban insights remains a major interdisciplinary challenge. Existing studies have shown the feasibility of deep learning-based approaches such as Mask2Former for segmentation, VGG-based networks for scene classification, and perceptual models for human-related analysis [2]. However, these fragmented, task-specific pipelines limit adaptability and transferability. To address this limitation, our project aims to implement and evaluate a unified instance segmentation framework built on a single self-supervised vision transformer backbone to explore an interdisciplinary frontier.

To achieve this, our model integrates a recently released ViT backbone (DINOv3 [3]) with a ViT Adapter [4] and a Mask2Former head [5], forming a cohesive architecture that captures both global context and local details for robust segmentation of complex urban scenes.

The paper reviews related work (Section II), presents the methodology (Section III), and discusses experimental results (Section IV) before the conclusion (Section V).

II. RELATED WORK

Instance segmentation detects and delineates individual objects in an image, combining the goals of object detection and semantic segmentation. The task has evolved from

region-based CNNs to unified transformer and self-supervised architectures. Early approaches such as Mask R-CNN [6] established strong baselines, yet relied heavily on hand-crafted anchors and region proposals. More recent methods adopt query-based transformer formulations; notably, Mask2Former [5] unifies instance, semantic, and panoptic segmentation as a single mask prediction problem, and has become a common, high-performing head for dense prediction. The Vision Transformer (ViT) [7] introduced token-based global attention, enabling large-scale pretraining and broad image understanding. Self-supervised learning (SSL) frameworks, particularly the DINO family culminating in DINOv3 [3], further advance this paradigm by learning transferable visual representations through teacher–student self-distillation without labeled data.

A recent attempt has investigated its adaptation to domain-specific segmentation tasks. SegDINO [8] applies a DINOv3 pretrained ViT with a MLP head for medical and natural image segmentation, achieving promising results in these relatively homogeneous domains.

III. METHOD

Our methodology centers on adapting a DINOv3-ViT backbone with a custom ViT Adapter and a Mask2Former head, forming a unified segmentation model (DINOv3-Mask2Former) evaluated on real-world urban data.

A. Data and data pre-processing

We use the Mapillary Vistas dataset [9] for training and evaluation. This large-scale benchmark contains 25,000 high-resolution street-level images collected worldwide under diverse weather, lighting, and geographic conditions. Each image is densely annotated with 124 semantic and 72 instance-level classes, representing typical urban elements. Instance segmentation on this dataset is challenging due to large appearance variations, frequent occlusions, cluttered backgrounds, and diverse object scales, and spatial context.

We experimented with multiple dataset splits for training and validation (9,000/1,000 and 6,000/800), and due to resource constraints, finalized a subset of 3,000 images for training and 600 images for validation, focusing on 12 categories with instance-level annotations. The selected classes are assigned unique Label IDs throughout training and evaluation, as detailed in Table I.

All images and masks are preprocessed using the `AutoImageProcessor`, which resizes inputs to 512×512 , normalize them, and formats `pixel_values`,

TABLE I: Selected instance classes from the Mapillary Vistas.

Label ID	Class Name
0	Crosswalk – Plain
1	Driveway
2	Person
3	Bicyclist
4	Street Light
5	Pole
6	Traffic Sign Frame
7	Bicycle
8	Bus
9	Car
10	Motorcycle
11	Truck

`mask_labels`, and `class_labels` for the DINOv3-Mask2Former model. Nearest-neighbor interpolation is used during both preprocessing and postprocessing to preserve alignment between predictions and ground-truth masks. Fig. 1 illustrates an example from the dataset, showing the original image, its binary mask, and the annotated overlay.

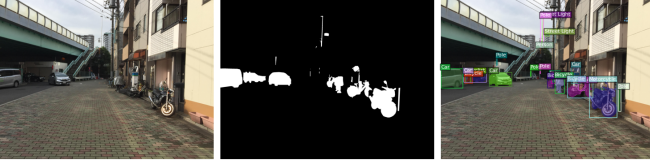


Fig. 1: Example case: original, masked, and annotated image.

B. Model architecture exploration

We explored several architectural variants before settling on the final configuration. The main components and tested modifications are summarized below:

- **DINOv3 Variants:** We experimented with ViT-S (21M parameters), ViT-S+ (29M parameters), and ViT-B (86M parameters) DINOv3 pre-trained models. Performance differences in preliminary tests were minimal; we selected ViT-S for efficiency.
- **Adapter Variants:** Two configurations were tested:
 - *Basic:* Uses 1×1 convolutions to project DINOv3’s uniform 384-channel features into multi-scale dimensions [96, 192, 384, 768] at strides {8, 16, 32, 32}. Features are extracted from layers [2, 5, 8, 11] and reshaped from (B, N, C) to (B, C, H, W) format.
 - *Enhanced:* Replaces 1×1 with 3×3 convolutions for larger receptive fields, adds Squeeze-and-Excitation (SE) blocks for channel attention, and implements top-down feature fusion via bilinear upsampling and 1×1 projections to improve multi-scale consistency.

C. Final model architecture design

Our final proposed architecture, depicted in Fig. 2, thus integrates the DINOv3 ViT-S (backbone), our Enhanced ViT Adapter (Neck), and the Mask2Former (head).

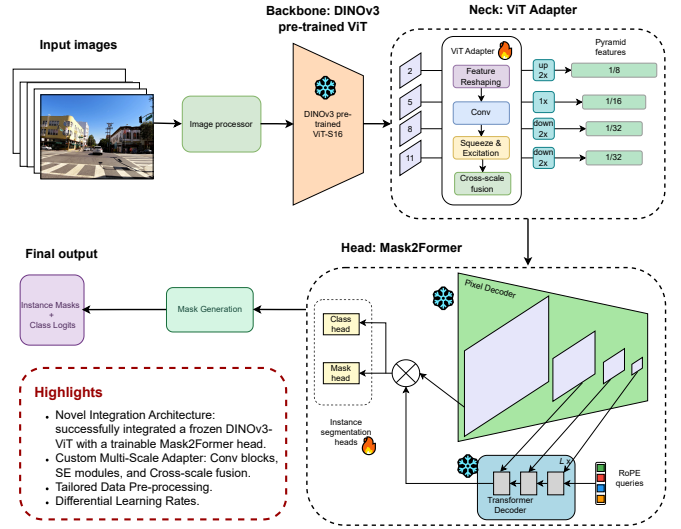


Fig. 2: Proposed instance segmentation architecture.

• Backbone (DINOv3 ViT-S):

Employs the `dinov3-vits16-pretrain-lvd1689m`. Hidden states from layers 2, 5, 8, and 11 are extracted to capture multi-level features.

• Neck (Enhanced ViT Adapter):

Processes the high-dimensional features from the selected backbone layers and projects them to multi-scale dimensions expected by the Mask2Former head. Our enhancements include:

- *3×3 Convolutions:* Replace 1×1 layers to enhance spatial context capture.
- *SE Blocks:* Apply channel attention after each projection to emphasize informative features.
- *Cross-Scale Fusion:* Enable top-down feature fusion via bilinear upsampling and 1×1 convolutions for better multi-scale consistency.

This enhanced adapter forms a more expressive and flexible bridge between the general-purpose backbone and the task-specific segmentation head.

• Head (Mask2Former):

Initialized from `facebook/mask2former-swin-large-mapillary-vistas-semantic`, it takes the adapted multi-scale features and performs query-based instance segmentation, predicting class labels and masks.

D. Training details

We train our model for 7 epochs with an effective batch size of 4 with 4 gradient accumulation steps, which takes around 12 hours.

- **Parameter Freezing:** To fine-tune efficiently under GPU constraints, we freeze the entire DINOv3 backbone along with Mask2Former’s pixel and transformer decoders, training only the custom adapter and segmentation head. Although full fine-tuning could improve performance,

this configuration enables faster convergence and reduced computational cost.

- **Loss Function:** Following [5], we use total loss that is a weighted sum of classification loss ($\mathcal{L}_{\text{cls}}$), binary cross-entropy mask loss ($\mathcal{L}_{\text{mask}}$), and dice loss ($\mathcal{L}_{\text{dice}}$), applied to matched query predictions. The specific formulation used is:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} \quad (1)$$

In our implementation, the coefficients are set to $\lambda_{\text{cls}} = 2.0$, $\lambda_{\text{mask}} = 5.0$, and $\lambda_{\text{dice}} = 5.0$, following the standard Mask2Former configuration.

- **Optimizer (differential learning rate):** We use an AdamW optimizer with a base learning rate of 1×10^{-4} and weight decay of 0.05. We apply a differential learning rate strategy: the adapter receives a $10\times$ multiplier and the prediction heads $5\times$.
- **Scheduler:** The learning rate is controlled by a cosine schedule with a 5% warmup phase over the total training steps. Gradient clipping is applied with a max norm of 1.0.

IV. RESULTS AND DISCUSSION

A. Training performance

The training loss curves for both adapter configurations are shown in Fig. 3. Both model configurations share the same training implementation and hyperparameter settings to ensure fair comparison, as detailed in Section III-D. The enhanced adapter (magenta curve) achieves a consistently lower and more stable loss compared to the basic adapter (olive curve), indicating improved feature alignment and faster convergence.

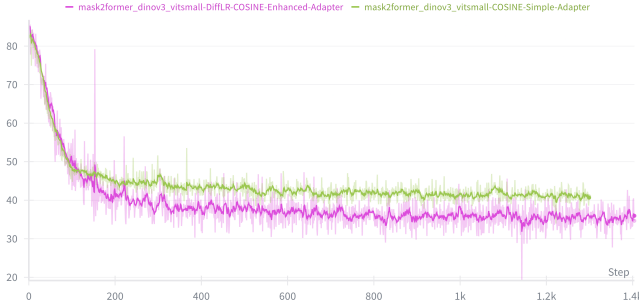


Fig. 3: Training loss curves for the basic and enhanced adapter model configuration.

B. Quantitative evaluation

We evaluate the model with the enhanced adapter on the validation set using standard metrics: mean Average Precision (mAP) for instance segmentation, mean Intersection over Union (mIoU) for semantic quality, Precision-Recall (PR) curves for per-class analysis and mean Pixel accuracy.

Table II summarizes the key performance metrics of our DINOv3-Mask2Former approach, yielding moderate performance. After limited training (7 epochs), the model achieves an overall mAP of 0.095 and mIoU of 0.473, demonstrating its

capability to capture coarse instance shapes and their semantic categories. A clear scale-dependent performance pattern is observed: the mAP for large objects substantially exceeds that for medium and especially small objects. Also, we achieved a mean pixel accuracy of 94.8% for all classes.

TABLE II: Summary of Key Performance Metrics.

Metric	Value
mIoU	0.473
mAP (overall)	0.095
mAP (small, area < 32 ² px)	0.004
mAP (medium, 32 ² ≤ area < 96 ² px)	0.045
mAP (large, area ≥ 96 ² px)	0.298
Mean Pixel accuracy	94.8%

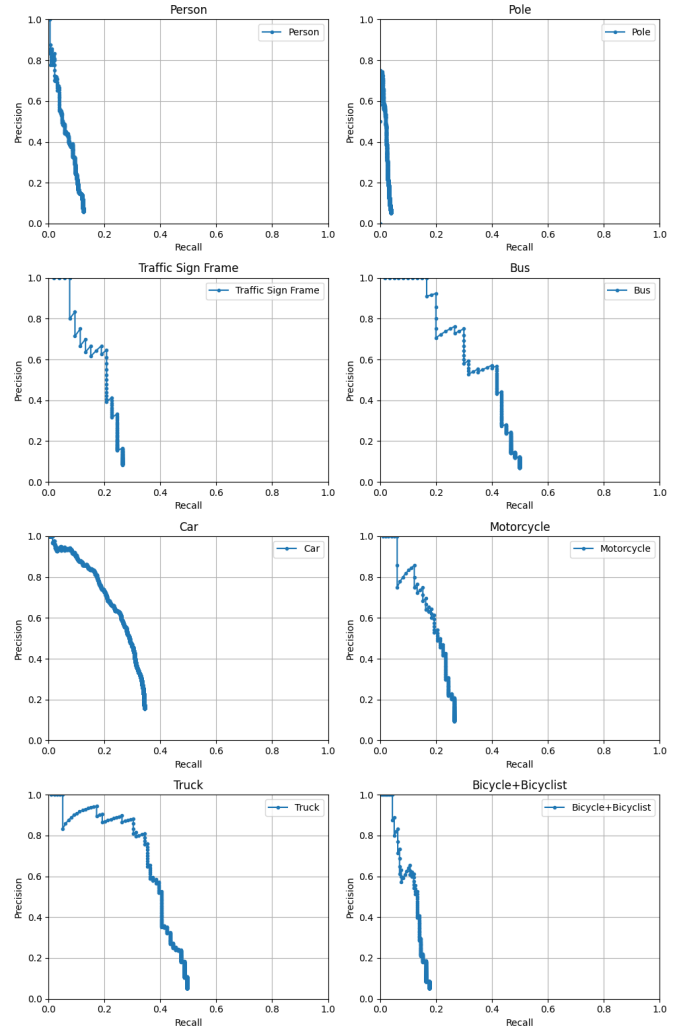


Fig. 4: Per class (selected classes) PR curves at IoU=0.50.

The per-class results PR curves (Fig. 4) further reinforce this pattern, highlighting strong performance in vehicle classes ("Truck", "Car", "Bus") but poor detection of smaller or thinner objects. The PR curves show steep precision drops at IoU=0.50 for small or thin objects.

This observed pattern indicates that the model performs reliably on large-scale entities, such as vehicles, while struggling with smaller or fine-grained targets due to limited resolution and feature granularity.

C. Qualitative evaluation

Fig. 5 compares ground-truth annotations with model predictions across representative urban scenes. In both commercial and residential scenarios, the predicted masks align well with dominant object boundaries, particularly for vehicles and pedestrians, illustrating effective generalization under varying illumination and spatial layouts.

However, the model occasionally merges multiple instances (e.g., grouping two bicycles or a group of people together) or splits a single instance (e.g., one car detected as two), especially in cluttered or occluded regions. Slender structures such as poles and street lights exhibit incomplete or fuzzy contours, while small objects are often under-segmented compared to large, well-defined targets.



(a) Urban street (commercial).



(b) Mixed-use street (residential).

Fig. 5: Comparison between ground-truth annotations and model predictions. Each subfigure shows the ground truth (left) and corresponding prediction (right).

D. Summary of findings

Our analysis reveals several consistent patterns:

- Segmentation of small objects remains challenging; large objects are segmented much more accurately. Pronounced class imbalance leads to significantly lower mAP and IoU for rare and small categories compared to dominant classes.
- The model sometimes merges multiple objects into one instance or splits a single object into multiple instances, particularly in crowded or occluded scenes.
- The coexistence of high pixel accuracy and low mAP indicates strong category-level recognition but limited instance-level differentiation.

V. CONCLUSION

This study presented a unified instance segmentation framework (DINOv3-Mask2Former) for urban streetscapes that integrates a SSL ViT backbone with a Mask2Former head via an enhanced ViT Adapter. The results demonstrate that transferable features from DINOv3 can be adapted for structured urban scene understanding.

Experiments on the Mapillary Vistas dataset validated the feasibility of this approach, achieving moderate segmentation accuracy after limited training. The model performs well on large and visually distinct objects but remains challenged by small, thin, or occluded instances. Another contributor to lower instance-level metrics is instance merging, where adjacent pedestrians are grouped as a single instance—producing visually coherent but evaluation-wise penalized results. These findings highlight both the potential and current limitations of applying self-supervised transformers to fine-grained urban perception.

Future work will focus on improving class balance through adaptive sampling or loss reweighting, employing higher-capacity backbones, and leveraging larger or domain-specific datasets. Such extensions are expected to enhance small-object sensitivity, mitigate class imbalance, and further generalize the framework for urban analytics applications.

REFERENCES

- [1] W. H. Whyte, *The Social Life of Small Urban Spaces*. Washington, D.C.: The Conservation Foundation, 1980.
- [2] Y. Hou, M. Quintana, M. Khomiakov, W. Yap, J. Ouyang, K. Ito, Z. Wang, T. Zhao, and F. Biljecki, “Global streetscapes — a comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 215, pp. 216–238, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0924271624002612>
- [3] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski, “Dinov3,” 2025. [Online]. Available: <https://arxiv.org/abs/2508.10104>
- [4] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, “Vision transformer adapter for dense predictions,” 2023. [Online]. Available: <https://arxiv.org/abs/2205.08534>
- [5] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, “Masked-attention mask transformer for universal image segmentation,” 2022. [Online]. Available: <https://arxiv.org/abs/2112.01527>
- [6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” 2018. [Online]. Available: <https://arxiv.org/abs/1703.06870>
- [7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021. [Online]. Available: <https://arxiv.org/abs/2010.11929>
- [8] S. Yang, H. Wang, Z. Xing, S. Chen, and L. Zhu, “Segdino: An efficient design for medical and natural image segmentation with dino-v3,” 2025. [Online]. Available: <https://arxiv.org/abs/2509.00833>
- [9] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, “The mapillary vistas dataset for semantic understanding of street scenes,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5000–5009.