

Instance Segmentation for Urban Streetscapes

Yinghao Chen, Mohammad Albarham

Chalmers University of Technology
Electrical Engineering Department



CHALMERS
UNIVERSITY OF TECHNOLOGY

Introduction

Recent advances in computer vision and the growing availability of large-scale street-level imagery have opened new opportunities for understanding urban environments. To explore this interdisciplinary frontier, we focus on instance segmentation for urban streetscapes.

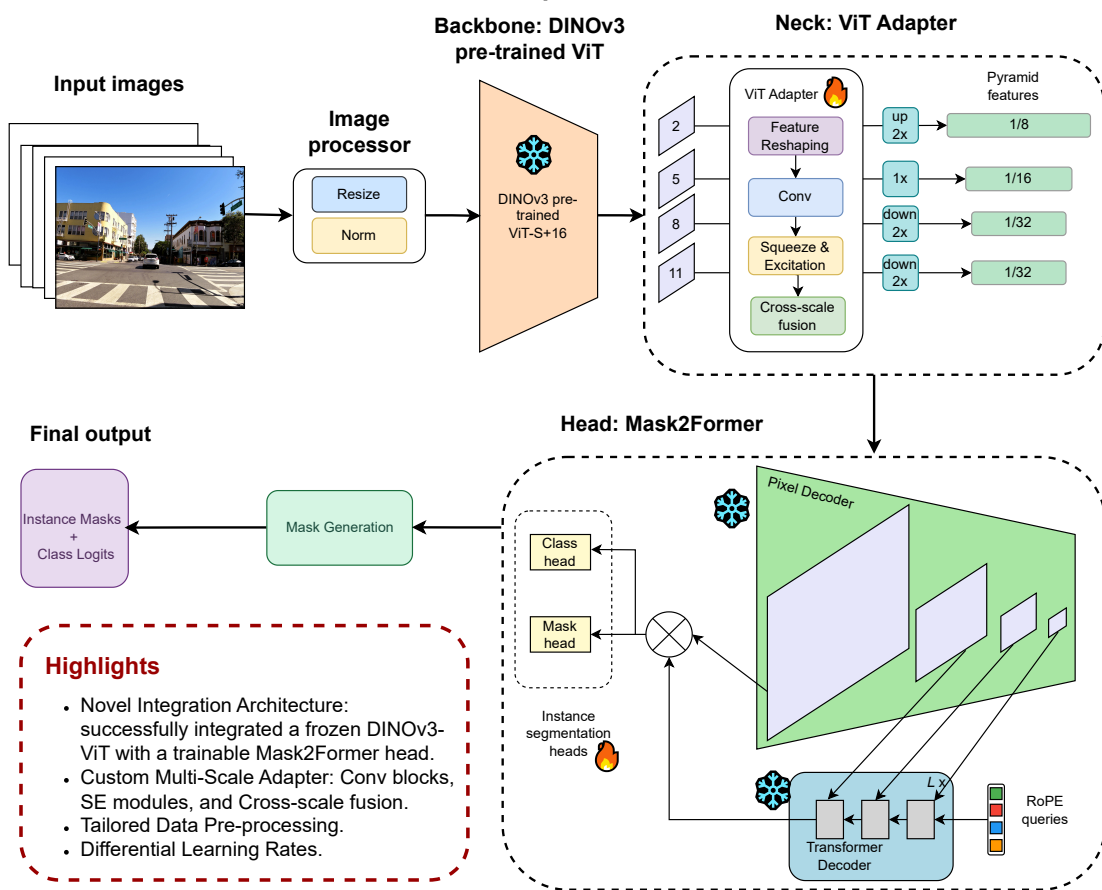


Figure 1: Proposed instance segmentation architecture.

We proposed an architecture that integrates three state-of-the-art components: [DINOv3-ViT](#)[5], [ViT Adapter](#)[1], and [Mask2Former](#)[2].

- **Backbone:** We replaced the standard backbone in Mask2Former with a DINOv3 ViT-S+16, and kept it FROZEN.
- **Neck:** We implemented a multi-scale ViT Adapter.
- **Head:** We trained the Mask2Former head.
- **Other:** We implemented the data pre-processing and the final mask generation logic.

Theoretical background

- **DINOv3 pretrained ViT model (Backbone):** Performed as a universal and high-quality feature extractor and frozen to [reduce the trainable parameter count for downstream blocks](#).
- **ViT Adapter (Neck):** Functioned as a multi-scale feature converter to [bridge the non-hierarchical output of the ViT backbone to the multi-scale feature pyramid](#).
- **Mask2Former (Head):** Chosen as a universal and efficient segmentation paradigm to [fully utilize DINOv3's superior dense features](#).

The architectures of the main components are presented as follows.

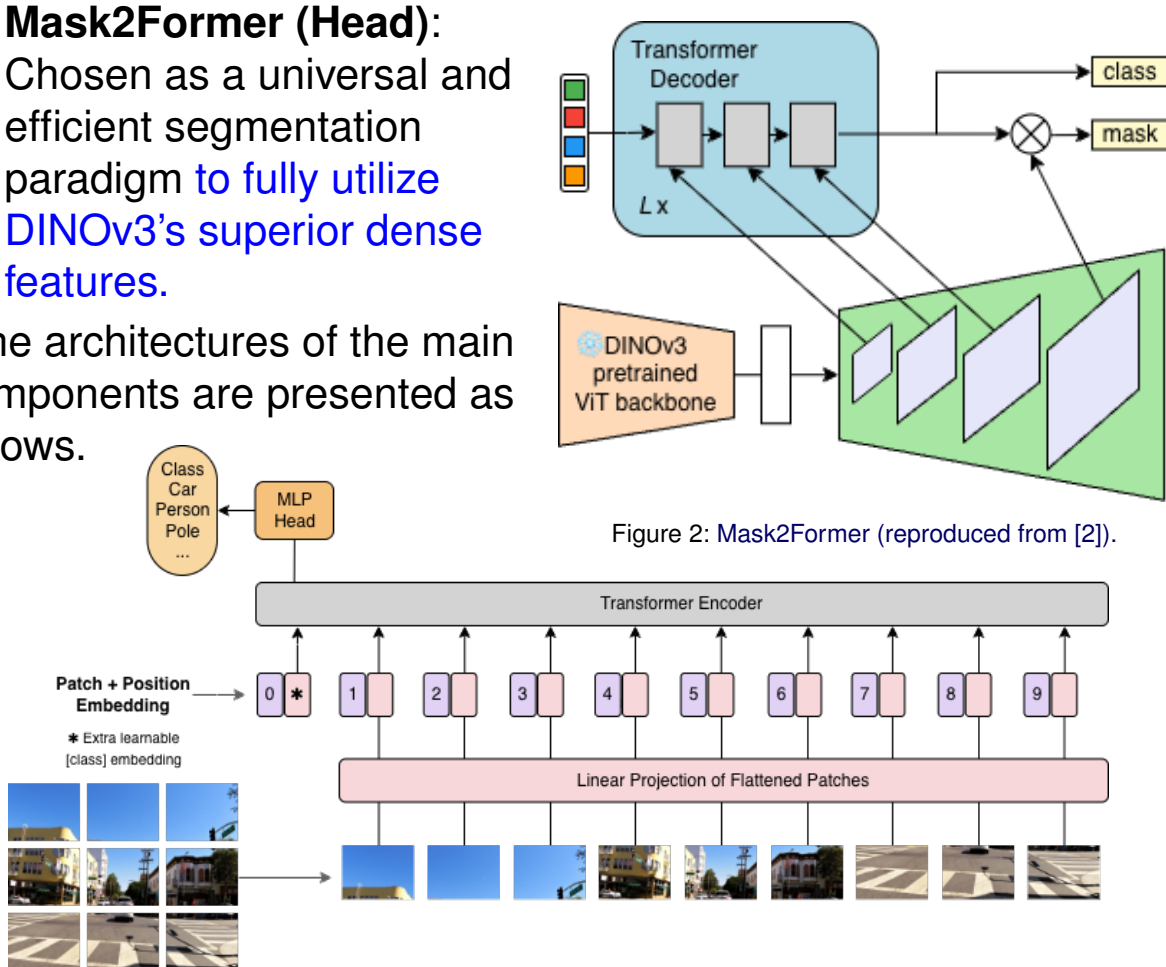


Figure 3: Vision Transformer (reproduced from [3]).

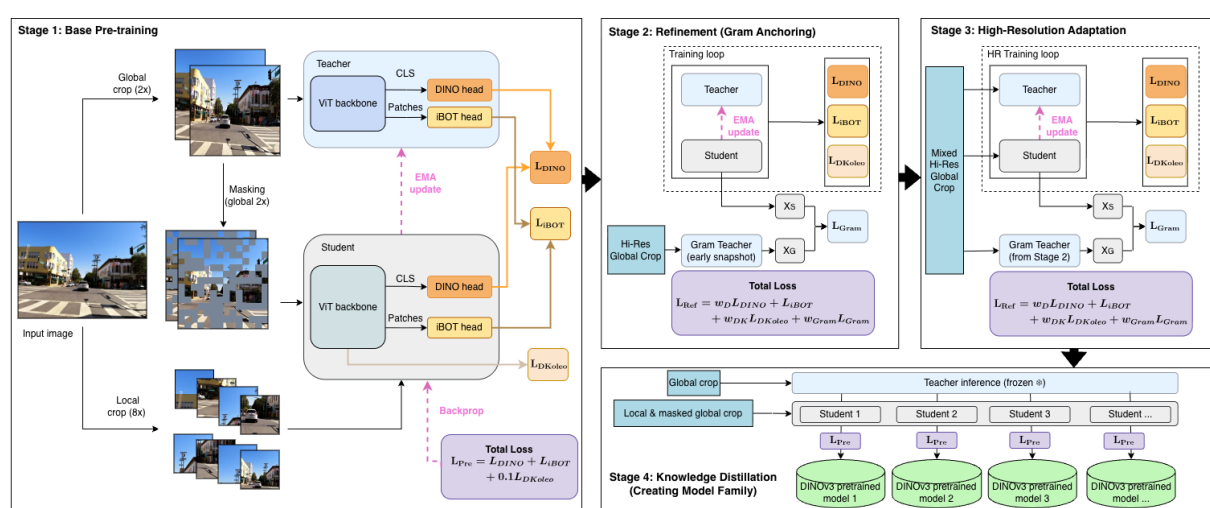


Figure 4: DINOv3 (adapted from [5]).

Dataset

► Mapillary Vistas Dataset [4]

- Large-scale: 25K high-resolution worldwide street-level
- Diverse: across diverse conditions (weather, season, and time of day), devices (mobile phones, action cameras, tablets, and professional rigs), and photographers of varying experience.
- Rich annotated: 124 object categories and 72 instance-specific classes



Figure 5: Examples of annotated images from Mapillary Vistas Dataset.

Results and Conclusions

Current best model:

- Evaluation methods: mean Average Precision (mAP), mean Intersection over Union (mIoU), and Precision-Recall (PR) Curves.
- Trained for **7 epochs** with an **effective training batch size of 16**.
- Optimizer: AdamW.
- Scheduler: Cosine with warmup.
- Current best result: mIoU = **0.47**; mAP_{small} = **0.002**, mAP_{medium} = **0.0045**, mAP_{large} = **0.298**.

► Total Loss function:

$$\mathcal{L}_{\text{total}} = 2\mathcal{L}_{\text{cls}} + 5\mathcal{L}_{\text{mask}} + 5\mathcal{L}_{\text{dice}}$$

where $\mathcal{L}_{\text{cls}}$ denotes the classification loss, $\mathcal{L}_{\text{mask}}$ the binary mask loss, and $\mathcal{L}_{\text{dice}}$ the dice coefficient loss.

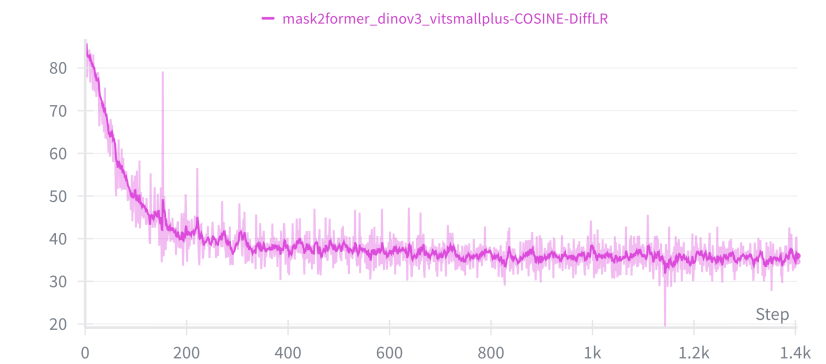


Figure 6: Train loss curve.

PR curves per class:

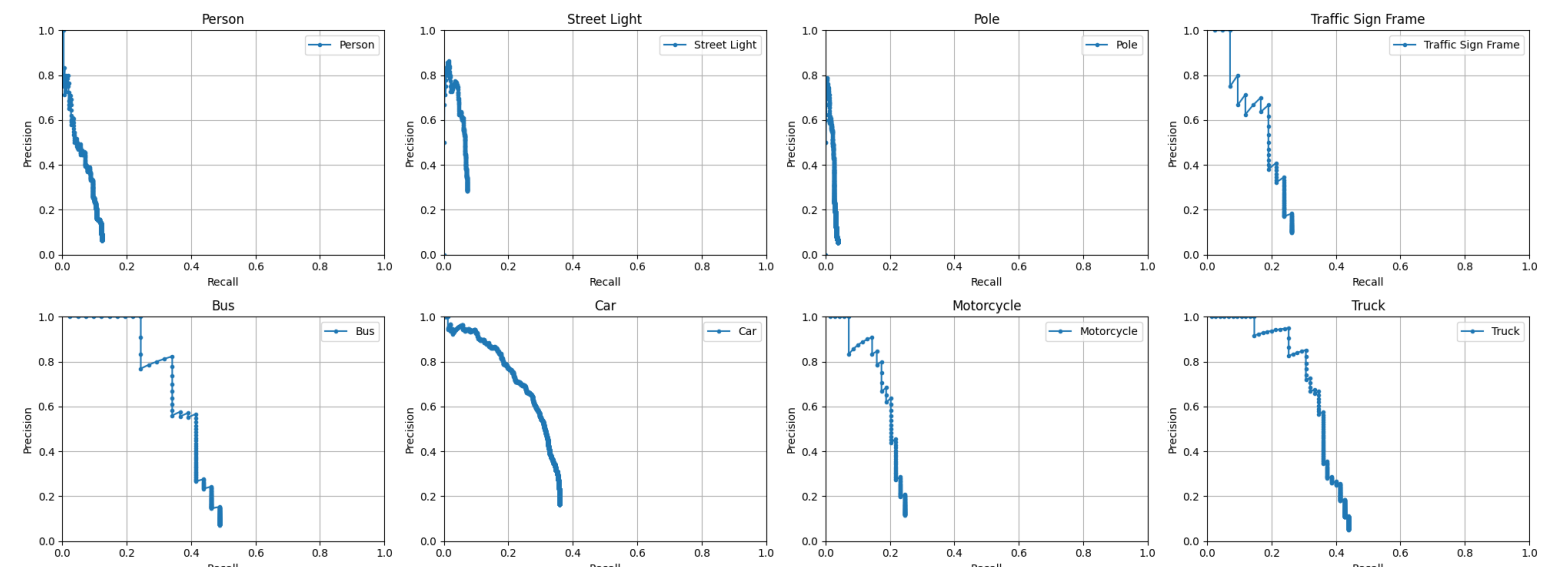


Figure 7: Per class (selected classes) Precision-Recall curves.

A few examples:



Figure 8: Examples of instance segmentation results from the current model.

The model achieves strong performance on frequent and large objects but reduced accuracy on small or rare classes. Future improvements include longer training and increasing samples for underrepresented classes.

References

- [1] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao. Vision transformer adapter for dense predictions, 2023.
- [2] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation, 2022.
- [3] 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.
- [4] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5000–5009, 2017.
- [5] 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.