ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation of China (62192783, 62276128, 62406140), Young Elite Scientists Sponsorship Program by China Association for Science and Technology (2023QNRC001), the Key Research and Development Program of Jiangsu Province under Grant (BE2023019) and Jiangsu Natural Science Foundation under Grant (BK20221441, BK20241200). The authors would like to thank Huawei Ascend Cloud Ecological Development Project for the support of Ascend 910 processors.

ETHICS STATEMENT

All authors have read and agree to abide by the ICLR Code of Ethics. This work does not involve interventions with human participants or personally identifiable information. We use only publicly available datasets under their original licenses and follow the terms of use. Potential risks and our mitigations are summarized below:

- **Privacy & Security.** We do not collect or release any personal data. When showing qualitative examples, all images/videos are from public datasets; any sensitive content is filtered.
- **Bias & Fairness.** We report results on multiple benchmarks and provide detailed settings to facilitate external auditing. We acknowledge possible dataset biases and encourage follow-up evaluation on broader demographics and domains.
- **Dual Use / Misuse.** The method could be misused to enable undesired large-scale labeling or surveillance. To reduce misuse, we release only research artifacts (code/configs) and exclude any tools for scraping or re-identifying individuals.
- **Legal Compliance.** We comply with licenses of all third-party assets (code, models, and datasets) and cite their sources. Any additional third-party terms are respected.
- **Research Integrity.** We document preprocessing, training recipes, and evaluation protocols; random seeds and hyperparameters are provided to enable reproducibility.

Where applicable, institutional review information is withheld for double-blind review and can be provided after acceptance.

REPRODUCIBILITY STATEMENT

We include training and evaluation details in the main paper and Appendix. Concretely: (i) all hyperparameters, optimization settings, and compute budgets; (ii) full data preprocessing and splits; (iii) code structure with scripts to reproduce the main tables and figures; (iv) checkpoints and logs for the primary models.

REFERENCES

Qi Bi, Jingjun Yi, Hao Zheng, Haolan Zhan, Yawen Huang, Wei Ji, Yuexiang Li, and Yefeng Zheng. Learning frequency-adapted vision foundation model for domain generalized semantic segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 94047–94072, 2024.

Prithvijit Chattopadhyay, Kartik Sarangmath, Vivek Vijaykumar, and Judy Hoffman. Pasta: Proportional amplitude spectrum training augmentation for syn-to-real domain generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19288–19300, 2023.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022.

Junhyeong Cho, Gilhyun Nam, Sungyeon Kim, Hunmin Yang, and Suha Kwak. Promptstyler: Prompt-driven style generation for source-free domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15702–15712, 2023.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Zhitong Gao, Bingnan Li, Mathieu Salzmann, and Xuming He. Generalize or detect? towards robust semantic segmentation under multiple distribution shifts. *Advances in Neural Information Processing Systems*, 37:52014–52039, 2024.

Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of field robotics*, 37(3):362–386, 2020.

Tianrui Guan, Divya Kothandaraman, Rohan Chandra, Adarsh Jagan Sathyamoorthy, Kasun Weerakoon, and Dinesh Manocha. Ga-nav: Efficient terrain segmentation for robot navigation in unstructured outdoor environments. *IEEE Robotics and Automation Letters*, 7(3):8138–8145, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yong Guo, David Stutz, and Bernt Schiele. Robustifying token attention for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17557–17568, 2023.

Christopher J Holder and Muhammad Shafique. On efficient real-time semantic segmentation: a survey. *arXiv preprint arXiv:2206.08605*, 2022.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Tommie Kerssies, Niccolò Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your ViT is Secretly an Image Segmentation Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, and Kwanghoon Sohn. Pin the memory: Learning to generalize semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4350–4360, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Jiashi Li, Xin Xia, Wei Li, Huixia Li, Xing Wang, Xuefeng Xiao, Rui Wang, Min Zheng, and Xin Pan. Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios. *arXiv preprint arXiv:2207.05501*, 2022.

Changhan Liu, Xunzhi Xiang, Zixuan Duan, Wenbin Li, Qi Fan, and Yang Gao. Don't need retraining: A mixture of detr and vision foundation models for cross-domain few-shot object detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pp. 4990–4999, 2017.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Chenbin Pan, Wenbin He, Zhengzhong Tu, and Liu Ren. Dino-r1: Incentivizing reasoning capability in vision foundation models. *arXiv preprint arXiv:2505.24025*, 2025.

Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the european conference on computer vision (ECCV)*, pp. 464–479, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pp. 102–118. Springer, 2016.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10765–10775, 2021.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Quan Tang, Chuanjian Liu, Fagui Liu, Jun Jiang, Bowen Zhang, CL Philip Chen, Kai Han, and Yunhe Wang. Rethinking feature reconstruction via category prototype in semantic segmentation. *IEEE Transactions on Image Processing*, 2025.

Qiang Wan, Zilong Huang, Jiachen Lu, Gang Yu, and Li Zhang. Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation. In *The eleventh international conference on learning representations*, 2023.

Jian Wang, Chenhui Gou, Qiman Wu, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. Rtformer: Efficient design for real-time semantic segmentation with transformer. *Advances in neural information processing systems*, 35:7423–7436, 2022.

Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28619–28630, June 2024.

Qishuai Wen and Chun-Guang Li. Rethinking decoders for transformer-based semantic segmentation: A compression perspective. *Advances in Neural Information Processing Systems*, 37: 49806–49833, 2024.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Jiacong Xu, Zixiang Xiong, and Shankar P. Bhattacharyya. Pidnet: A real-time semantic segmentation network inspired from pid controller, 2022.

Zhengze Xu, Dongyue Wu, Changqian Yu, Xiangxiang Chu, Nong Sang, and Changxin Gao. Sctnet: Single-branch cnn with transformer semantic information for real-time segmentation. *arXiv preprint arXiv:2312.17071*, 2023.

Guoyu Yang, Yuan Wang, Daming Shi, and Yanzhong Wang. Golden cudgel network for real-time semantic segmentation, 2025. URL https://arxiv.org/abs/2503.03325.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. *CoRR*, abs/1808.00897, 2018. URL http://arxiv.org/abs/1808.00897.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, Jingyu Wang, and Wenbing Tao. Perception r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*, 2025.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

Seokju Yun, Seunghye Chae, Dongheon Lee, and Youngmin Ro. Soma: Singular value decomposed minor components adaptation for domain generalizable representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 25602–25612, June 2025.

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

Xin Zhang and Tan Robby T. Mamba as a bridge: Where vision foundation models meet vision language models for domain-generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International conference on machine learning*, pp. 27378–27394. PMLR, 2022a.

Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2582–2593, 2022b.

Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *CVPR*, 2022c.

A THE USE OF LLMS

We used ChatGPT-4o to polish our manuscript, using the following prompt:

```
I want you to act as an expert in scientific writing.  I will
provide you with some paragraphs in English and your task is
to improve the spelling, grammar, clarity, conciseness, and
overall readability of the text provided, while breaking down long
sentences, reducing repetition and increasing logic.  You should use
artificial intelligence tools, such as natural language processing,
rhetorical knowledge, and your expertise in effective scientific
writing techniques to reply.  Provide the output as a table in
a readable mode.  The first column is the original sentence, the
second column is the sentence after editing, and the third column
provides explanation of your edits and reasons.  Please edit the
following text in a scientific tone:
```

B APPENDIX

B.1 ABLATION FOR BOUNDARY ACCURACY

To evaluate boundary prediction accuracy, we tested our model under the GTA5 $\rightarrow$ Cityscapes setting. We compared QPrompt-R1 with the RTSS methods PIDNet, Next-ViT, and the DGSS methods Mask2Former and REIN. .The 1px B-mIoU and 3px B-mIoU metrics measure boundary mIoU at 1 and 3 pixels from the ground-truth boundaries, respectively. As shown in Table 10, our model performs similarly to Mask2Former and REIN, but without multi-scale features or pixel encoders, instead using a lightweight transposed-conv upsampler. This results in speed improvement with higher FPS. These demonstrate that our approach effectively balances efficiency and boundary accuracy, with no substantial loss in segmentation quality.

Table 10: Performance about Boundary accuracy.

Method	**1px B-mIoU**	**3px B-mIoU**	**mIoU**	**FPS**
PIDNet	25.4	29.3	45.7	46
Next-ViT	17.1	31.6	50.1	57
Mask2Former	43.5	45.8	63.7	11
REIN	44.1	46.5	66.4	10
QPrompt-R1	42.7	45.3	66.1	54

B.2 GRQA APPLY TO GENERAL SEGMENTATION

Since the GRQA algorithm leverages mutual supervision between queries for optimization, it can be applied to improve query-based models for general semantic segmentation, not just domain-generalized task. We conduct additional in-domain testing on the Citys$\rightarrow$Citys setting. As shown in Table 11, we found that GRQA still provides performance improvement in general semantic segmentation.

B.3 PERFORMANCE COMPARISON WITH PROMPTED SAM MODELS

We also compared our model with prompt-based methods, specifically FastSAM (Zhang et al., 2023) and Grounded SAM(Ren et al., 2024). Since SAM produces class-agnostic masks, we tested them under the open-vocabulary semantic segmentation setting, where we used text prompts that

Table 11: Performance comparison **general semantic segmentation** (Citys→Citys)

Method	**Citys→Citys (mIoU)**
QPrompt	79.2
+GRQA	80.4
Mask2Former	82.4
+GRQA	83.1

include class labels to make the predicted masks class-specific. The experiments were conducted on Cityscapes, and as shown in Table 12, our model outperforms both FastSAM and Grounded SAM, achieving significantly higher mIoU and FPS. Additionally, we observed that the speed bottleneck of FastSAM lies in mask classification, rather than in the "everything" mode used for class-agnostic mask prediction.

Table 12: Performance Comparison with Prompted SAM Models.

Method	mIoU (Citys)	FPS
FastSAM	32.6	< 1
Grounded SAM	36.7	< 1
Ours	66.1	54