SemiDAViL: Semi-supervised Domain Adaptation with Vision-Language Guidance for Semantic Segmentation

Hritam Basak*, Zhaozheng Yin
Dept. of Computer Science, Stony Brook University, NY, USA
*hbasak@cs.stonybrook.edu

1. Overview

The supplementary material of SemiDAViL provides a detailed analysis of the superiority of our proposed vision-language attention mechanism in section 2, studies the class statistics and analyzes the influence of DyCE loss for tail classes in section 3, detailed comparison with other class-balancing losses in section 4, qualitative analysis of SemiDAViL on segmentation tasks in section 5, and detailed comparison with SemiVL [5] in section 6. Finally, we discuss some of the potential future directions in section 7.

2. Effectiveness of Dense Language Guidance

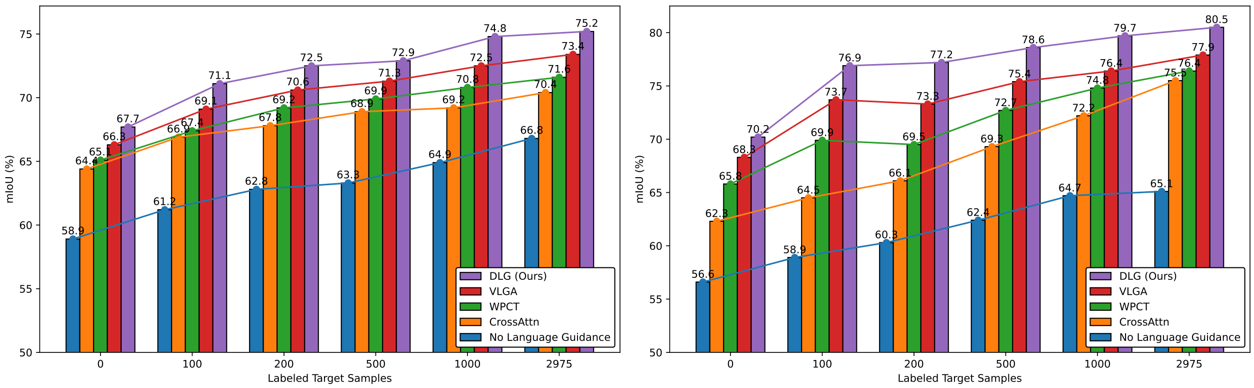
To evaluate the effectiveness of our proposed vision-language attention mechanism using DLG, we perform a comparative analysis of DLG with some of the existing attention mechanisms. Figure 1 highlights the effectiveness of our proposed Dense Language Guidance (DLG) mechanism compared to previous vision-language integration approaches. DLG achieves state-of-the-art performance across all label settings in both GTA5 → Cityscapes and Synthia → Cityscapes benchmarks. For instance, in the zero-label setting, DLG attains a mean Intersection-over-Union (mIoU) of 67.7% (GTA5) and 70.2% (Synthia), outperforming the Word-Pixel Correlation Tensor (WPCT) [6] by +2.6% and +4.4%, respectively, and Vision-Language Guided Attention (VLGA) (which is used in SemiVL [5]) by +1.4% and +1.9%. As the number of labeled target samples increases, DLG maintains consistent improvements, with mIoU gains of up to +3.6% over WPCT and +2.6% over VLGA at 2975 labeled samples. These results underscore DLG's superior ability to fuse vision and language features effectively across various levels of supervision.

DLG's improved performance arises from its balanced treatment of vision and language features in its attention mechanism. Unlike prior methods that treat language features merely as attention weights, DLG transforms both visual and textual features into key-value pairs and treats them equivalently during cross-modal interaction. This enables a richer fusion process where attended vision and language features are integrated symmetrically, producing a true multimodal representation that retains information from both modalities. By normalizing and applying attention across both vision and language axes, DLG ensures comprehensive integration, avoiding the dominance of one modality over the other. The result is a robust mechanism capable of capturing nuanced cross-modal dependencies, which directly translates to improved generalization and segmentation performance, particularly in low-label scenarios where prior approaches struggle. Thus DLG sets a new benchmark for robust VL guidance in SSDA.

3. Class Imbalance Analysis

We further evaluate the extent of class imbalance in Synapse and Cityscapes datasets in Figure 2 and Figure 3, respectively along with the performance improvements brought by the proposed dynamic CE (DyCE) loss. The plotted results in Figure 2 demonstrate a strong inverse correlation between class frequency and the efficacy of our DyCE loss in improving segmentation performance under SSL settings with 20% annotations. For rare classes (e.g., Ga, Es, LAG, and RAG), which exhibit frequencies below 1%, DyCE achieves substantial improvements, with URPC [9] and UA-MT [11] backbones delivering gains exceeding 50% in some cases. Conversely, for more frequently occurring classes like Li and St, DyCE exhibits modest gains, highlighting its ability to address class imbalance by prioritizing underrepresented categories. This trend underscores the robustness of DyCE in improving representation learning for minority classes, a critical challenge in medical image segmentation tasks.

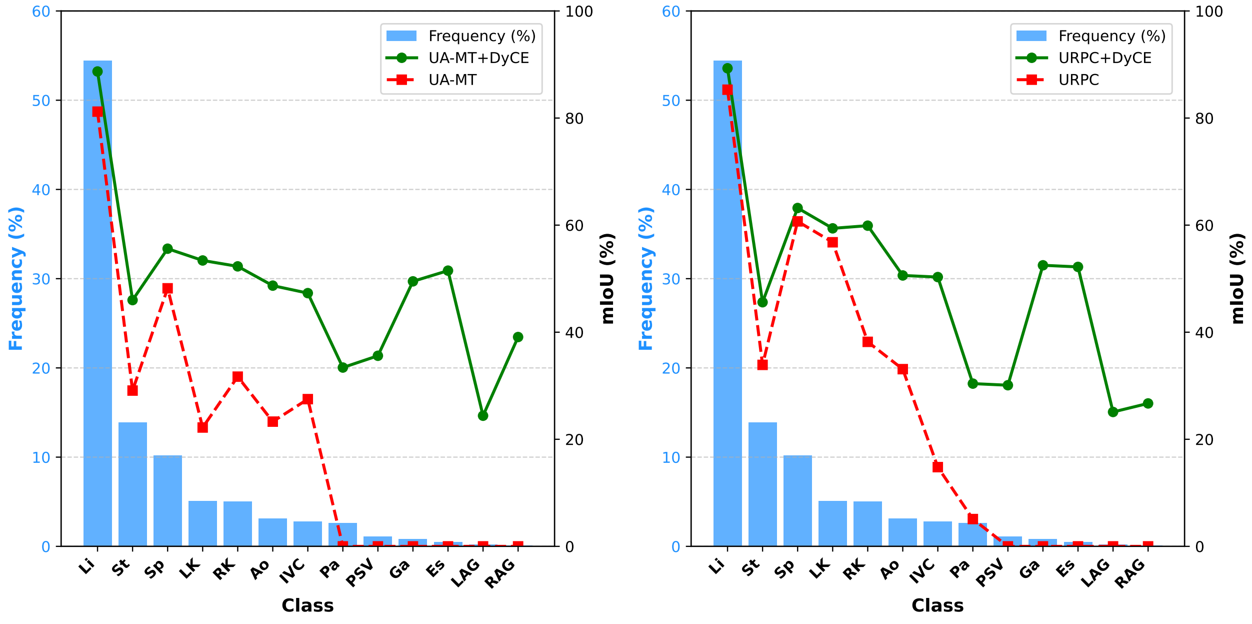
The Cityscapes results in Figure 3 highlight the significant performance gains achieved for rare classes, such as Train, Motorcycle, and Rider, which collectively represent less than 1% of the dataset. Notably, these classes exhibit improvements of up to 2.3% (Rider) in the GTA5→Cityscapes scenario and 2.05% (Rider) in the Synthia→Cityscapes scenario. Similarly, other infrequent classes, such as Bus, Truck, and Wall, also experience sub-



(a) GTA5→Cityscapes

(b) Synthia→Cityscapes

Figure 1. Comparative analysis of multiple vision-language attention mechanisms: our Dense Language Guidance (DLG), VLGA [5], WPCT [6], and Cross-attention [2] on (a) **GTA5→Cityscapes** and (b) **Synthia→Cityscapes** using different labeled target annotations.



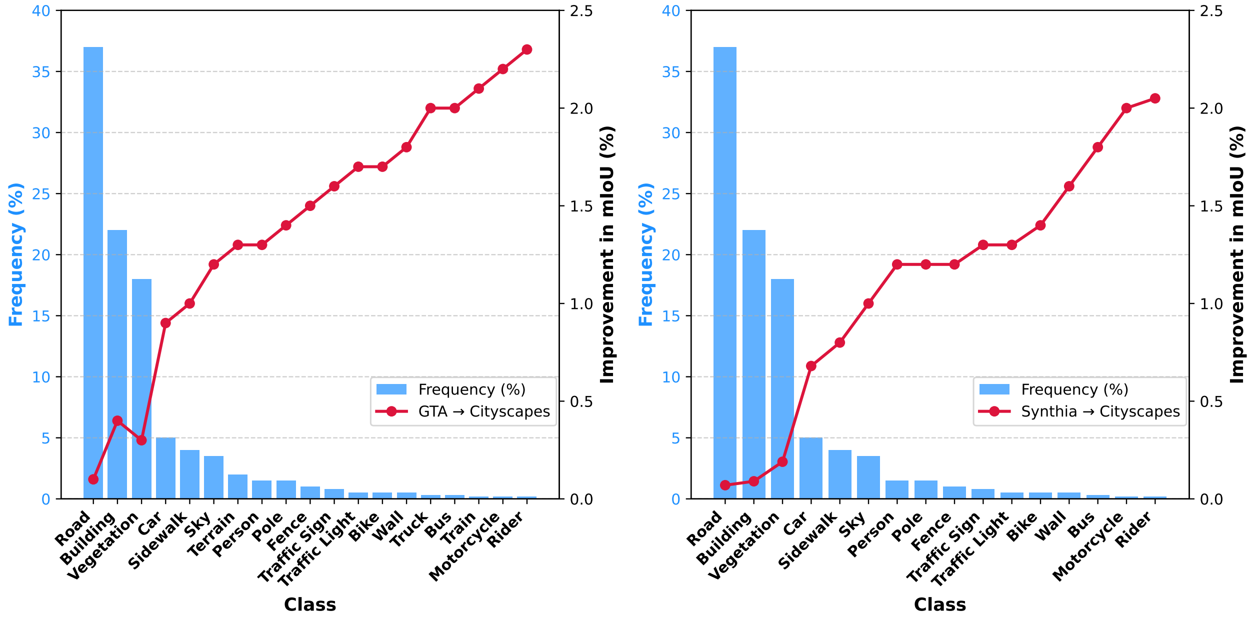
(a) Improvement on UA-MT

(b) Improvement on URPC

Figure 2. Class distribution of the **Synapse** medical dataset and percentage improvement achieved by our DyCE loss in (a) UA-MT [11] and (b) URPC [9] networks using 20% labeled data in SSL setting.

stantial boosts, ranging from 1.8% to 2% across both domain adaptation settings. This improvement underscores DyCE's effectiveness in addressing domain shift challenges for low-frequency classes, where traditional methods often struggle. Conversely, the results for common classes like Road and Building, which dominate the dataset's frequency distribution, show relatively limited improvements of 0.1% to 0.4% for GTA5→Cityscapes and 0.07% to 0.09% for Synthia→Cityscapes. Classes with moderate frequency, such as Car, Sidewalk, and Vegetation, also show consistent but smaller gains, further confirming DyCE's capability to prioritize and balance underrepresented classes while maintaining stable performance for dominant categories. This class-specific focus allows the model to achieve improved overall segmentation quality without compromising accuracy in the more frequent classes.

Overall, DyCE demonstrates superior efficacy in addressing class imbalance by dynamically prioritizing underrepresented tail classes, resulting in significant performance gains for rare categories while maintaining stability for dominant ones, establishing its critical role in advancing segmentation tasks where tail-class accuracy is crucial.



(a) Improvement on **GTA5→Cityscapes**

(b) Improvement on **Synthia→Cityscapes**

Figure 3. Class distribution of the Cityscapes dataset and percentage improvements achieved by our DyCE loss in (a) **GTA5→Cityscapes** and (b) **Synthia→Cityscapes** scenarios under the SSDA setting with 500 target annotations.

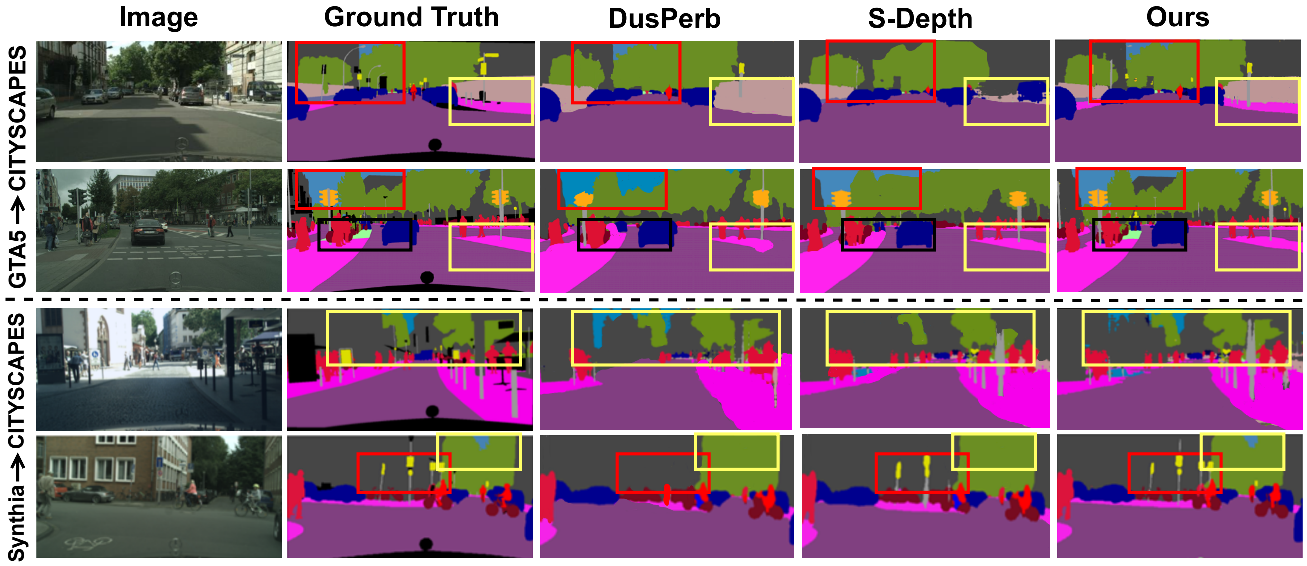


Figure 4. Qualitative comparison of SemiDAViL with previous state-of-the-art method, DusPerb [10], S-Depth [4] on 100-labeled target data on **GTA5→Cityscapes** and **Synthia→Cityscapes** adaptation settings.

4. Comparison with Class-balancing Losses

In this section, we present a comparative analysis of our proposed Dynamic Cross Entropy (DyCE) loss against several state-of-the-art loss functions designed to address class imbalance in semantic segmentation tasks. The methods under comparison include Cross Entropy (CE) [12] , Weighted Cross Entropy (WCE) [1], Focal Loss (FL) [8], Dual Cross Entropy (DCE) [7], and Dual Focal Loss (DFL) Hossain et al. [3].

Cross Entropy serves as a baseline, calculating the logarithmic difference between predicted and actual class distributions. Weighted Cross Entropy introduces class-specific

Table 1. Performance comparison of our DyCE loss with previous loss functions to address class imbalance. We report 19-class and 16-class mIoU scores for the **GTA5→Cityscapes** and **Synthia→Cityscapes** settings, respectively across 0, 100, 200, 500, 100, and 2975 (100%) labeled target images. Our results are **highlighted** whereas the previous-best and second-best results are marked in red and blue.

GTA5→Cityscapes						
Loss	**Labeled Target Samples**					
	0	**100**	**200**	**500**	**1000**	**2975**
CE [12]	66.9	70.3	71.6	72.1	73.9	74.4
WCE [1]	64.3	69.4	70.5	71.8	72.7	73.3
DCE [7]	67.1	70.5	71.6	72.3	73.3	74.7
FL [8]	67.0	70.4	71.8	72.2	74.1	74.9
DFL [3]	67.3	70.6	71.9	72.2	74.0	74.8
DyCE (ours)	**67.7**	**71.1**	**72.5**	**72.9**	**74.8**	**75.2**
Synthia→Cityscapes						
CE [12]	69.5	74.9	76.8	77.7	79.2	79.6
WCE [1]	67.2	73.1	74.4	75.3	77.6	78.9
DCE [7]	69.7	75.2	76.8	78.1	79.1	79.7
FL [8]	69.8	75.4	76.9	78.0	79.3	79.7
DFL[3]	69.9	75.7	77.0	78.2	78.9	79.8
DyCE (ours)	**70.2**	**76.9**	**77.2**	**78.6**	**79.7**	**80.5**

weights to emphasize underrepresented categories. Focal Loss modulates the loss contribution of well-classified examples, allowing the model to focus on challenging instances. Dual Cross Entropy incorporates a regularization term to balance positive and negative predictions. Dual Focal Loss combines adaptive scaling with regularization to achieve more balanced gradient propagation. Our proposed DyCE loss builds upon these foundations, introducing a dynamic, gradient-adaptive mechanism to optimize convergence and segmentation performance across varying levels of supervision.

Table 1 establishes the preeminence of DyCE in mitigating class imbalance across varying degrees of target annotations. In the **GTA→Cityscapes** domain adaptation scenario, DyCE achieves a 19-class mIoU of 67.7% in the absence of labeled target samples, scaling up to 75.2% under full supervision (2975 labels). For **Synthia→Cityscapes**, DyCE exhibits analogous dominance, attaining a 16-class mIoU of 70.2% without annotations and 80.5% under complete supervision. These findings underscore DyCE's adaptability and robustness across supervision levels. Unlike WCE, which employs static weighting, and DCE, which incorporates a regularization term, DyCE dynamically accentuates harder-to-classify categories without succumbing to gradient attenuation. This design enables it to surpass state-of-the-art methods such as DFL and FL, particularly under annotation-scarce regimes.

CE establishes a baseline yet falters in addressing class imbalance due to uniform loss weighting. WCE refines this by imposing fixed penalties on minority classes; however, static weights remain suboptimal in evolving scenarios. FL introduces dynamic scaling tailored to class difficulty, excelling in emphasizing hard-to-classify examples, but its susceptibility to vanishing gradients impedes training convergence. DCE mitigates this gradient vanishing by regularizing negative class predictions, albeit at the cost of disproportionately penalizing false negatives. DFL synthesizes FL's adaptive scaling with DCE's regularization, ensuring balanced gradient propagation. DyCE advances this paradigm through a gradient-adaptive mechanism that dynamically modulates loss and penalizes misclassifications more rigorously, especially for underrepresented classes. This innovative construct ensures optimal convergence and superior segmentation performance, as evidenced by DyCE's consistent outperformance across all experimental scenarios in Table 1.

5. Qualitative Results

We provide a qualitative comparison of our proposed Semi-DAViL with the previous best methods, DusPerb [10], S-Depth [4], and the available ground truth label in Figure 4. Experiments are performed using 100-labeled target samples on **GTA5→Cityscapes** and **Synthia→Cityscapes** settings. Upon closer examination, the proposed method consistently produces segmentation masks that more closely align with the ground truth than the other two methods. In particular, our method demonstrates superior performance in accurately delineating object boundaries and capturing fine details for semantically confusing classes (e.g., sidewalk vs. wall, rider vs bike, traffic light vs. vegetation, wall vs. sky, etc.), owing to the strong semantic prior from vision-language initialization and dense language guidance using multimodal attention. This suggests that the proposed method has a better capacity for understanding and interpreting the relationship between linguistic descriptions and visual features, resulting in more accurate and refined segmentation outputs. Moreover, our approach shows a marked improvement for tail classes (e.g., rider, motorcycle, wall, etc.) in precisely segmenting intricate shapes and maintaining object integrity. This improvement can be attributed to our proposed class-balancing DyCE loss, which dynamically prioritizes imbalanced and underperforming classes. Previous SoTA methods like [4, 10] fall short in these two major aspects, leading to suboptimal performance.

6. Comparison with SemiVL

Although they may appear similar at first glance, our work is fundamentally different from SemiVL [5]. Whereas SemiVL targets semi-supervised semantic segmentation with a focus on label efficiency—employing a language-guided decoder that leverages frozen CLIP predictions and dataset-specific class definitions—our method, SemiDAViL, pioneers semi-supervised domain adaptation. We address the domain shift challenge by integrating a Dense Language Guidance (DLG) module that fuses fine-grained visual and textual embeddings for robust, pixel-level semantic alignment across domains. Furthermore, our approach tackles class imbalance through a novel Dynamic Cross-Entropy (DyCE) loss that reweights minority classes during training. In addition, our pseudo-labeling strategy synergistically combines consistency regularization with dense language embeddings to refine predictions, while our language guidance utilizes detailed captions for both content and spatial positioning rather than fixed class definitions. Together, these innovations enable SemiDAViL to effectively bridge the semantic gap between source and target domains, setting it apart as the first language-guided semi-supervised DA method for semantic segmentation.

To further validate the differences quantitatively, we perform experiments under both semi-supervised domain adaptation (SSDA) and semi-supervised learning (SSL) settings using varying numbers of labeled target samples. Under the SSDA scenario—specifically adapting from GTA or synthetic data to Cityscapes—our method consistently achieves higher mIoU scores across all target sample sizes. For example, when adapting from Syn. to Cityscapes, our method attains 76.9, 77.2, 78.6, and 79.7 mIoU with 100, 200, 500, and 1000 labeled samples respectively, outperforming SemiVL by up to 5.5 mIoU. This performance gain is primarily attributed to our integration of domain adaptation components, such as the Dense Language Guidance (DLG) module, which fuses visual and textual features to better align semantic representations across domains, and the novel Dynamic Cross-Entropy (DyCE) loss that rebalances class distributions to mitigate source bias. In contrast, SemiVL, which lacks explicit domain adaptation mechanisms, tends to overfit to the abundant source labels, resulting in a suboptimal adaptation to the target domain. Moreover, even under the SSL setting (Cityscapes→Cityscapes), where domain shift is not a factor, our method still outperforms SemiVL (81.6 vs. 80.4 mIoU at 1000 labeled samples), indicating that our approach enhances feature localization and pseudo-label refinement through a more robust consistency training framework. These detailed experimental results confirm that our method addresses domain shift more effectively and improves the overall semantic segmentation performance under limited annotation.

7. Future Works

While SemiDAViL demonstrates strong performance in mitigating class imbalance and enhancing segmentation accuracy, there remain areas that offer opportunities for further refinement. The reliance on pre-trained vision-language models like CLIP may pose challenges in adapting to domains where such resources are limited or less aligned with the data. The dense multimodal attention and dynamic loss modulation, while effective, could benefit from optimizations to ensure scalability across larger datasets and real-time applications. The use of off-the-shelf captioning models in Dense Language Guidance (DLG) highlights the importance of high-quality linguistic features, which could be further enhanced for greater robustness. Addressing these aspects could unlock even broader applications and performance improvements for this promising approach.

Table 2. Quantitative comparison with SemiVL [5] on SSDA and SSL settings.

Type	**Method**	**Labeled Target Sample**			
		100	**200**	**500**	**1000**
SSDA (GTA→City.)	SemiVL [5]	68.5	69.9	70.6	71.7
	Ours	71.1	72.5	72.9	74.8
SSDA (Syn.→City.)	SemiVL [5]	71.4	72.3	74.3	75.1
	Ours	**76.9**	**77.2**	**78.6**	**79.7**
SSL (City.→City.)	SemiVL [5]	76.2	77.9	80.2	80.4
	Ours	**77.1**	**78.2**	**81.4**	**81.6**

References

[1] Yuri Sousa Aurelio, Gustavo Matheus De Almeida, Cristiano Leite de Castro, and Antonio Padua Braga. Learning from imbalanced data sets with weighted cross-entropy function. *Neural processing letters*, 50:1937–1949, 2019. 3, 4

[2] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 2

[3] Md Sazzad Hossain, John M Betts, and Andrew P Paplinski. Dual focal loss to address class imbalance in semantic segmentation. *Neurocomputing*, 462:69–87, 2021. 3, 4

[4] Lukas Hoyer, Dengxin Dai, Qin Wang, Yuhua Chen, and Luc Van Gool. Improving semi-supervised and domain-adaptive semantic segmentation with self-supervised depth estimation. *International Journal of Computer Vision*, 131 (8):2070–2096, 2023. 3, 4

[5] Lukas Hoyer, David Joseph Tan, Muhammad Ferjad Naeem, Luc Van Gool, and Federico Tombari. Semivl: Semi-supervised semantic segmentation with vision-language guidance. In *European Conference on Computer Vision*, pages 257–275. Springer, 2025. 1, 2, 4, 5

[6] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 1, 2

[7] Xiaoxu Li, Liyun Yu, Dongliang Chang, Zhanyu Ma, and Jie Cao. Dual cross-entropy loss for small-sample fine-grained vehicle classification. *IEEE Transactions on Vehicular Technology*, 68(5):4204–4212, 2019. 3, 4

[8] T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017. 3, 4

[9] Xiangde Luo, Guotai Wang, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Dimitris N Metaxas, and Shaoting Zhang. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency. *Medical Image Analysis*, 80:102517, 2022. 1, 2

[10] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7236–7246, 2023. 3, 4

[11] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part II 22*, pages 605–613. Springer, 2020. 1, 2

[12] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018. 3, 4