LangDA: Language-guided Domain Adaptive Semantic Segmentation

Chang Liu[1] **Saad Hossain**[1] **C Thomas**[2] **Kwei-Herng Lai**[2]
Raviteja Vemulapalli[2] **Sirisha Rambhatla**[1] **Alexander Wong**[1,2]
[1]University of Waterloo [2]Apple
{chang.liu,s42hossa,sirisha.rambhatla,alexander.wong}@uwaterloo.ca
{c.thomas,khlai,r_vemulapalli}@apple.com

Abstract

Pixel-level manual annotations are expensive and time-consuming to obtain for semantic segmentation tasks. Unsupervised domain adaptation (UDA), which outperforms direct zero-shot methods, adapts from a label-rich source domain to a target domain where labels are scarce or unavailable. Recent progress in foundational models has demonstrated the potential of large vision-language models (VLMs) in zero-shot segmentation and domain adaptive classification. However, the efficacy of VLMs in bridging domain gaps for semantic segmentation remains under-explored. To improve segmentation performance in UDA, we introduce a novel language-guided adaptation method (LangDA), which aligns image features with VLMs' domain-invariant text embeddings during training. We generate the text embeddings by using a captioning VLM to create image-specific textual descriptions, which are then passed to a frozen CLIP-based encoder. To the best of our knowledge, this is the first work to utilize text to align vision domains in unsupervised domain adaptation for semantic segmentation (DASS). Our proposed language-driven plug-and-play UDA approach achieved a 62.0% mean Jaccard index on the standard Synthia $\rightarrow$ Cityscapes benchmark, outperforming the current state-of-the-art by 0.9% with negligible parameter overheads.

1 Introduction

Training state-of-the-art neural networks for visual recognition requires large-scale annotated datasets, but collecting and annotating data at pixel-level is time-consuming and tedious. To reduce the need for manual annotation, unsupervised domain adaptation for semantic segmentation (DASS) methods train segmentation networks on an available labelled source domain and adapt to an unlabeled target domain [1–4] (Figure 1a). However, a notable performance gap still exists between UDA methods (that use unlabeled target data) and their supervised counterparts [2, 4, 5].

To enhance the efficacy in bridging domain gaps, this work introduces language in the DASS setting alongside conventional unlabeled target data. Our method leverages recent advancements in large-scale vision-language models (VLMs)[6–10], which have demonstrated remarkable performance and are now pivotal in computer vision research. Notably, architectures like CLIP [9] and LLaVA [10] have demonstrated the potential of natural language supervision in learning rich visual representations. Furthermore, recent studies have also successfully applied VLMs to zero-shot domain adaptation [11–13], (Figure 1b) in contexts where target domain data is unavailable.

Despite the successes of VLMs in zero-shot DA, VLMs' role in UDA has been overlooked. Given the remarkable success of VLMs in zero-shot DA[11–13] (Figure 1b), we hypothesize VLMs have significant untapped potential for advancing UDA methods. Our preliminary experiments reveal that incorporating VLMs to address domain shifts for UDA methods yields particularly promising results,

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

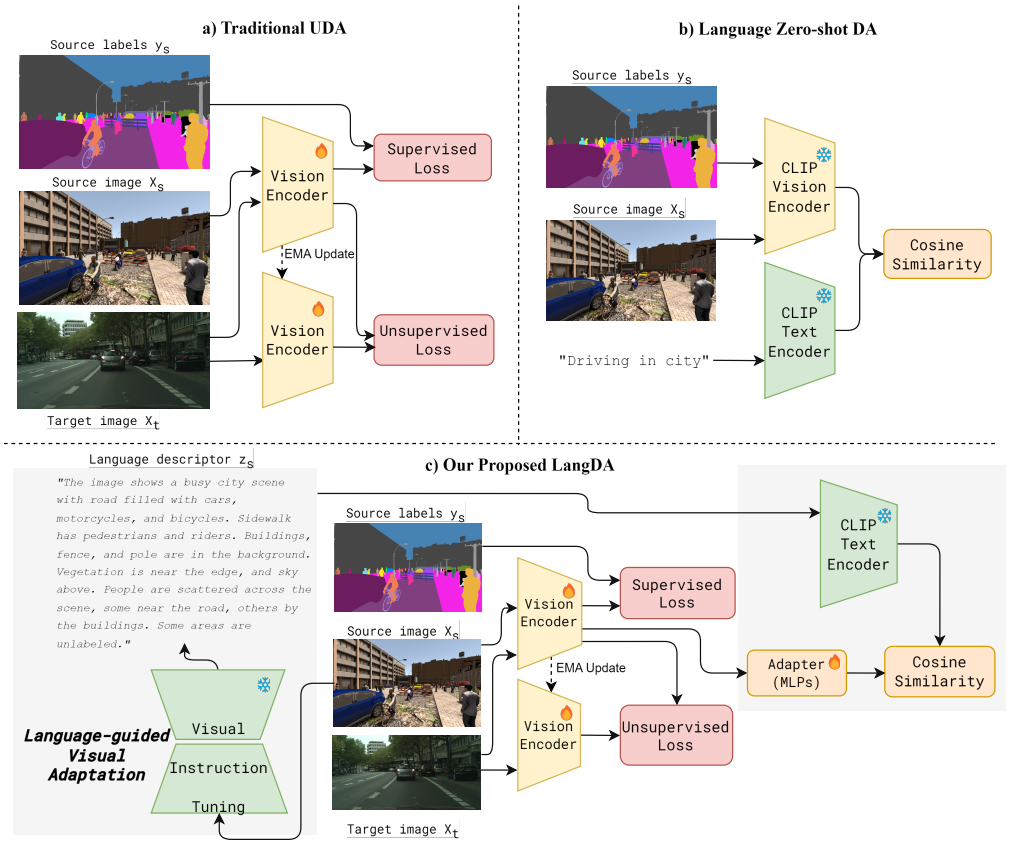

Figure 1: **(a)** In traditional UDA, a supervised source-trained model adapts an unsupervised target model to the unlabelled target images by updating the target model's weights with an exponential moving average. **(b)** Zero-shot adaption uses a prompt description instead of unlabeled target images to mitigate potential domain shifts. **(c)** Our proposed LangDA method generates image-level language descriptors to facilitate adaptation while introducing very few learnable parameters (namely adapters).

which evokes the question: Is the sole focus on traditional visual domain alignment in DASS still the optimal approach moving forward? In this paper, we investigate strategies to efficiently integrate prior linguistic knowledge derived from VLMs to facilitate model adaptation to unlabeled target visual domains, specifically for semantic segmentation tasks within the context of UDA.

2 Related Works

Unsupervised Domain Adaptation. In UDA (Figure 1a), a model trained on a labelled source domain is adapted to an unlabeled target domain. Most UDA approaches rely on discrepancy minimization[3, 14–17], adversarial training [3, 18], or self-training [2, 19–22]. However, the role of language in unsupervised domain adaptation remains largely unexplored. Our work seeks to address this gap by leveraging textual information from visual prompt tuning to bridge visual domain gaps in DASS.

Zero-shot Domain Adaptation with Language. In zero-shot domain adaption (Figure 1b), researchers facilitate domain transfer using a labelled source domain and a textual description of the unavailable target images. For instance, [13] and [11] utilize text embedding from the CLIP model to approximate the target visual domain. However, to the best of our knowledge, no existing work combines textual information with unlabeled visual target data to bridge the domain gap. Our research fills this blank by proposing a novel approach that leverages linguistic cues in conjunction with unlabeled target image data to mitigate visual domain discrepancy.

Feature Alignment To facilitate adaptation, researchers often employ feature alignment techniques to minimize discrepancies between various feature representations. To better leverage ImageNet's real-world high-level semantic classes, [2] regularizes the bottleneck features with ImageNet features. Furthermore, [13] redirects image-domain features towards text-domain features with a brief one-sentence description processed by a pre-trained CLIP model [9]. However, the potential of textual features in mitigating domain shifts between two visual domains remains largely unexplored. Our paper addresses this research gap by aligning source and target visual features with CLIP-encoded [9] language features. Overall, this work harnesses the rich semantic information embedded in linguistic representations to enhance visual adaptation tasks.

3 Methodology

Our primary objective is unsupervised domain adaptation for semantic segmentation (DASS). DASS involves classifying each pixel of an unlabeled target image $\mathcal{D}_\text{T}$ into K semantic categories given the labelled source data $\mathcal{D}_\text{S}$, where $\mathcal{D}_\text{T} = \{x_T^{(i)} \mid x_T^{(i)} \in \mathbb{R}^{H\times W\times 3}\}$ and $\mathcal{D}_\text{S} = \{(x_S^{(i)}, y_S^{(i)}) \mid x_S^{(i)} \in \mathbb{R}^{H\times W\times 3}, y_S^{(i)} \in \{0,1\}^{H\times W\times K}\}$. For adaptation between the visual domains, we employed online self-training (see Appendix B for details). To provide visual features with language guidance, we utilized an additional cross-domain vision-language alignment minimization objective with image-level textual descriptions.

Text Generation We use visual prompt instruction tuning, specifically, the captioning model LLaVA [10], to generate the set of image-level captions $\mathcal{C}_\text{S} = \{z_S^{(i)} \mid z_S^{(i)} \in \mathbb{R}^l\}$. We further obtain $\mathcal{C}_r = \{z_r^{(i)} \mid z_r^{(i)} \in \mathbb{R}^l, l \leq 77\}$ by refining $\mathcal{C}_\text{S}$ using the LLM Mistral-Large-2 [23] to include all the source class names from the ground truth segmentation masks and ensure the caption is less than 77 tokens, which is the maximum token length accepted by CLIP [9]. We pass the refined captions $\mathcal{C}_r$ to the frozen CLIP encoder E_CLIP [9] to obtain the set of text feature vectors $\mathcal{V}_\text{CLIP} = \{v_\text{CLIP}^{(i)} \mid v_\text{CLIP}^{(i)} = E_\text{CLIP}(z_r^{(i)})\}, v_\text{CLIP}^{(i)} \in \mathbb{R}^{512}\}$.

Prompt-guided Adaptation To align source image features $\mathcal{F}_\text{S}$=$\{f_\text{S}^{(i)} \mid f_\text{S}^{(i)}$=$E_{g_\theta}(x_S^{(i)})\}$ with domain-invariant textual features, we introduce an image-level minimization objective on the distance between CLIP textual features $\mathcal{V}_\text{CLIP}$ and the source feature $\mathcal{F}_\text{S}$.

We define the objective function as follows:

$$\mathcal{L}_p^{(i)}(f_\text{S}^{(i)}, v_\text{CLIP}^{(i)}) = 1 - \frac{f_\text{S}^{(i)} \cdot v_\text{CLIP}^{(i)}}{\|f_\text{S}^{(i)}\| \, \|v_\text{CLIP}^{(i)}\|} \,. \tag{1}$$

This CLIP-space cosine distance, previously employed in a similar manner in text-driven image editing by [24], guides the source features toward the text embedding.

The global target features are also implicitly guided towards the text feature space through the EMA model update in Equation (2).

The overall UDA loss $\mathcal{L}$ is a minimization problem of the weighted sum of the supervised loss, unsupervised loss and language-guided loss $\mathcal{L} = \mathcal{L}_S + \mathcal{L}_T + \lambda_\text{p}\mathcal{L}_\text{p}$.

4 Main Result

We report our results using the standard semantic segmentation metric, Jaccard similarity coefficient (mean Intersection over Union), as shown in Table 1. Notably, we observe a performance improvement of 0.9% in mIoU with the introduction of image-level textual guidance. This promising preliminary result highlights the efficacy of incorporating linguistic information in DASS.

Figure 2 illustrates the t-SNE visualizations of feature distributions. After integrating language-driven feature alignment, our method LangDA shows improved per-class clustering. For instance, in DAFormer's [2] t-SNE, the feature representations of walls (light orange) and traffic signs (rose pink) overlap in the image domain, likely due to traffic signs often visually appear in front of walls from driver's first-person view. On the other hand, walls and traffic signs are semantically distinguishable

in terms of language, contributing to enhanced segmentation mIoU for LangDA in Table 2. These findings demonstrate the promising capability of our proposed language-augmented approach in mitigating domain discrepancies and enhancing semantic segmentation performance in UDA settings.

Table 1: Comparison with state-of-the-art methods in UDA and Zero-shot DA. We performed our experiments on standard adaptation benchmark Synthia → Cityscapes. Note source only refers to lower bound DA baselines with no adaptation (i.e. training on source and evaluation on target). See Appendix A for implementation details.

Method	**Backbone**	**Unlabeled Target Data**	**Prompt Description**	**% mIoU↑**
Source only	ResNet-50			29.3
PODA [13]	ResNet-50		✓	29.5
ULDA [11]	ResNet-50		✓	30.8
Source only	ResNet-101			29.4
ADVENT [3]	ResNet-101	✓		41.2
CBST [25]	ResNet-101	✓		42.6
DACS [4]	ResNet-101	✓		48.3
CorDA [26]	ResNet-101	✓		55.0
ProDA [27]	ResNet-101	✓		55.5
DAFormer [2]	SegFormer	✓		61.1
LangDA (Ours)	SegFormer	✓	✓	**62.0** (+0.9%)

Table 2: Per-class performance on synthetic-to-real adaptation benchmark: Synthia→Cityscapes

Method	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Sky	Person	Rider	Car	Bus	M.bike	Bike	% mIoU ↑
PODA [13]	19.0	13.2	61.6	11.6	0.5	34.9	13.6	11.9	74.4	77.0	62.5	13.3	61.2	20.1	8.2	9.7	29.5
ADVENT [3]	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2
CBST [25]	68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	78.3	60.6	28.3	81.6	23.5	18.8	39.8	42.6
DACS [4]	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	90.8	67.6	38.3	82.9	38.9	28.5	47.6	48.3
CorDA [26]	**93.3**	61.6	85.3	19.6	5.1	37.8	36.6	42.8	84.9	90.4	69.7	41.8	85.6	38.4	32.6	53.9	55.0
ProDA [27]	**87.8**	45.7	84.6	37.1	0.6	44.0	54.6	37.0	**88.1**	84.4	74.2	24.3	**88.2**	51.1	40.5	45.6	55.5
DAFormer [2]	86.2	42.3	88.2	38.4	**8.6**	49.9	55.6	54.1	86.9	89.3	73.4	47.1	87.8	**57.3**	53.1	60.2	61.1
LangDA (Ours)	83.1	**43.5**	**88.8**	**43.7**	5.6	**51.5**	**57.8**	**57.4**	85.5	**92.5**	**74.9**	**49.7**	87.6	53.4	**56.3**	**61.7**	**62.0**

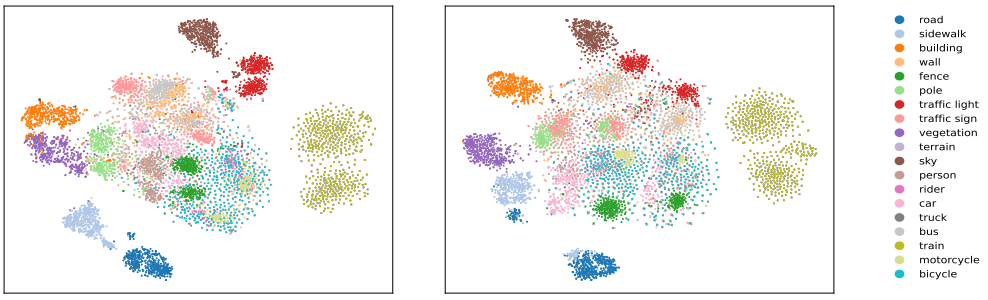

Figure 2: **Left**: DAFormer [2], adaptation using only visual images. **Right**: LangDA (Ours), adaptation using both visual images and textual descriptions. After aligning language and visual features, we observe more well-defined boundaries and improved class clustering in t-SNE.

5 Discussion

With guidance from image-wise captioning, this work shows a notable +0.9% improvement in mIoU and significantly more differentiable target domain features. Our result is a promising initial step in illuminating the potential of language-guided DASS. For future work, we seek to apply our method to multi-resolution adaptation [5], building upon existing single-resolution adaptation techniques [2]. Additionally, we will further investigate strategies for driving target features toward textual domains by concurrently aligning source and target image representations with their corresponding language representations.

References

[1] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 289–305.

[2] L. Hoyer, D. Dai, and L. Van Gool, "Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9924–9935.

[3] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2517–2526.

[4] W. Tranheden, V. Olsson, J. Pinto, and L. Svensson, "Dacs: Domain adaptation via cross-domain mixed sampling," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 1379–1389.

[5] L. Hoyer, D. Dai, and L. Van Gool, "Hrda: Context-aware high-resolution domain-adaptive semantic segmentation," in *European conference on computer vision*. Springer, 2022, pp. 372–391.

[6] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[7] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9579–9589.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[9] ——, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[10] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[11] S. Yang, Z. Tian, L. Jiang, and J. Jia, "Unified language-driven zero-shot domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 407–23 415.

[12] G. Kwon and J. C. Ye, "Clipstyler: Image style transfer with a single text condition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 062–18 071.

[13] M. Fahes, T.-H. Vu, A. Bursuc, P. Pérez, and R. De Charette, "Poda: Prompt-driven zero-shot domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 623–18 633.

[14] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *International conference on machine learning*. PMLR, 2017, pp. 2208–2217.

[15] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 443–450.

[16] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[17] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems*, vol. 17, 2004.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[19] L. Chen, Z. Wei, X. Jin, H. Chen, M. Zheng, K. Chen, and Y. Jin, "Deliberated domain bridging for domain adaptive semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 15 105–15 118, 2022.

[20] D. Kim, M. Seo, K. Park, I. Shin, S. Woo, I. S. Kweon, and D.-G. Choi, "Bidirectional domain mixup for domain adaptive semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1114–1123.

[21] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2. Atlanta, 2013, p. 896.

[22] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[23] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[24] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2085–2094.

[25] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 289–305.

[26] Q. Wang, D. Dai, L. Hoyer, L. Van Gool, and O. Fink, "Domain adaptive semantic segmentation with self-supervised depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8515–8525.

[27] P. Zhang, B. Zhang, T. Zhang, D. Chen, Y. Wang, and F. Wen, "Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 414–12 424.

[28] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[29] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[30] M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation, 2020.

[31] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

A Implementation Details

Datasets For the source domain, we use the Synthia dataset[28], which consists of 9,400 synthetic images with resolution 1280×760. For the target domain, we utilize the standard benchmark Cityscapes dataset [29], which includes 2,975 training images and 500 validation images at a resolution of 2048×1024.

Network Architecture We implement our model using the mmsegmentation framework [30], employing the DAFormer architecture with the MiT-B5 encoder [31], which generates a feature pyramid with channels $C = [64, 128, 320, 512]$. The DAFormer decoder is configured with 256 channels and dilation rates of 1, 6, 12, and 18. All encoders are pre-trained on ImageNet-1k [32].

B Self-training

Following [2, 4], we leverage labelled source $\mathcal{D}_\text{S}$ and unlabeled target images $\mathcal{D}_\text{T}$ to facilitate adaptation by training a student model g_θ and a teacher model h_θ. We first train the student segmentation network g_θ with a categorical cross-entropy supervised segmentation loss on the labelled source domain $\mathcal{D}_\text{S}$:

$$\mathcal{L}_S^{(i)} = -\sum_{j=1}^{H\times W}\sum_{c=1}^{C} y_S^{(i,j,c)} \log g_\theta(x_S^{(i)})^{(j,c)}$$

We then have a teacher network h_ϕ to generate pseudo-labels for unlabeled target domain data $\mathcal{D}_\text{T}$, using the argmax of the softmax output (note gradients are not backpropagated into the teacher).

$$p_T^{(i,j,c)} = \left[c = \arg\max_{c'} h_\phi(x_T^{(i)})^{(j,c')}\right],$$

A quality estimate for pseudo-labels is provided based on the ratio of pixels exceeding a confidence threshold τ in the softmax probability.

$$q_T^{(i)} = \frac{\sum_{j=1}^{H\times W}\left[\max_{c'} h_\phi(x_T^{(i)})^{(j,c')} > \tau\right]}{H \cdot W}.$$

These pseudo-labels and their confidence estimates are used to further train g_θ on the target domain to compute the unsupervised loss for the teacher model.

$$\mathcal{L}_T^{(i)} = -\sum_{j=1}^{H\times W}\sum_{c=1}^{C} q_T^{(i)} p_T^{(i,j,c)} \log g_\theta(x_T^{(i)})^{(j,c)}.$$

Pseudo-labels can be generated either online or offline. Follow [2] we opt for online ST due to its simplicity and single training stage, which is crucial for comparing and ablating network architectures. In online ST, the teacher network is updated as the exponentially moving average of the student network after each training step.

$$\phi_{t+1} \leftarrow \alpha\phi_t + (1-\alpha)\theta_t \tag{2}$$