

# 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

## Abstract

Domain Adaptation (DA) and Semi-supervised Learning (SSL) converge in Semi-supervised Domain Adaptation (SSDA), where the objective is to transfer knowledge from a source domain to a target domain using a combination of limited labeled target samples and abundant unlabeled target data. Although intuitive, a simple amalgamation of DA and SSL is suboptimal in semantic segmentation due to two major reasons: (1) previous methods, while able to learn good segmentation boundaries, are prone to confuse classes with similar visual appearance due to limited supervision; and (2) skewed and imbalanced training data distribution preferring source representation learning whereas impeding from exploring limited information about tailed classes. Language guidance can serve as a pivotal semantic bridge, facilitating robust class discrimination and mitigating visual ambiguities by leveraging the rich semantic relationships encoded in pre-trained language models to enhance feature representations across domains. Therefore, we propose the first language-guided SSDA setting for semantic segmentation in this work. Specifically, we harness the semantic generalization capabilities inherent in vision-language models (VLMs) to establish a synergistic framework within the SSDA paradigm. To address the inherent class-imbalance challenges in long-tailed distributions, we introduce class-balanced segmentation loss formulations that effectively regularize the learning process. Through extensive experimentation across diverse domain adaptation scenarios, our approach demonstrates substantial performance improvements over contemporary state-of-the-art (SoTA) methodologies. Code is available: [GitHub](#).

## 1. Introduction

The remarkable progress in deep learning has significantly enhanced the performance of visual understanding tasks, including image classification [7], object detection [103], and, more recently, semantic segmentation [61, 71]. These advancements have been particularly notable when a wealth

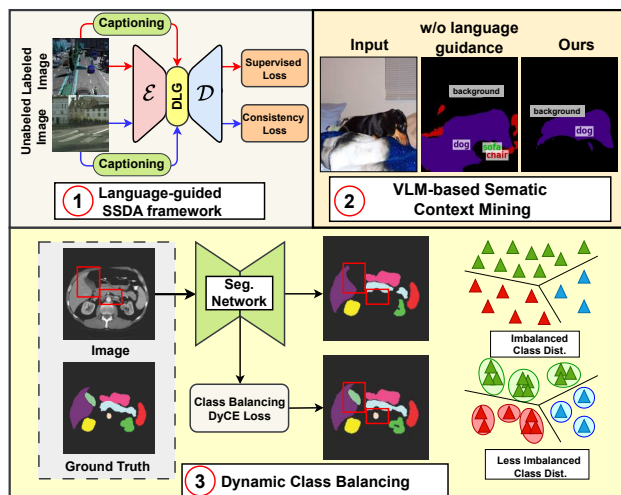


Figure 1. Major contributions of SemiDAViL: (1) We propose the first language-guided SSDA framework for semantic segmentation, (2) Utilizing spatial context via dense language guidance (DLG) improves segmentation performance, (3) Our proposed DyCE loss dynamically reweighs imbalanced class distributions, resulting in precise segmentation of **minority classes**.

of labeled training data is available. However, as noted in [68], their performance degrades precipitously when confronted with annotation-scarce environments, especially in the context of semantic segmentation, where dense pixel-wise annotations are essential. Furthermore, these sophisticated models exhibit substantial vulnerability when tasked with generalizing across domains characterized by significant distributional shifts [30, 36] - a challenge particularly evident in real-world applications where models trained on synthetic data must maintain robust performance in naturalistic settings, such as autonomous navigation systems [19, 66]. This inherent limitation in cross-domain generalization has catalyzed the emergence of two pivotal research paradigms: Domain Adaptation (DA) and Semi-supervised Learning (SSL).

The confluence of DA and SSL has given rise to Semi-supervised Domain Adaptation (SSDA), a hybrid approach

that strategically leverages three distinct data streams: comprehensively labeled source domain data, sparsely labeled target domain samples, and a wealth of unlabeled target domain instances [42, 65]. While SSDA holds intuitive appeal for real-world applications, existing methods encounter critical limitations when applied to semantic segmentation tasks. Specifically, (1) despite achieving accurate segmentation boundaries, current approaches [52, 92] often suffer from *misclassification among visually similar classes*, due to restricted supervision within the target domain; (2) the SSDA framework tends to over-prioritize source domain features, driven by abundant source labels, while generating *error-prone pseudo-labels for target data*, which hampers adaptation performance; (3) class-imbalance, a common issue in real-world datasets, exacerbates these challenges, limiting effective exploration and representation of *minority (tail) classes* in the target domain.

To address the identified SSDA challenges, we augment the SSDA paradigm with vision-language (VL) guidance using VLMs (e.g., CLIP [60]) to enrich semantic representation, leveraging their large-scale image-caption pretraining. By incorporating VLM features into a global-local context exploration module, we mitigate *misclassification among visually similar classes*.

To tackle the over-reliance on source features, we introduce a joint embedding space guided by language priors, enhancing instance separability and *reducing domain bias*, unlike traditional divergence-based alignment methods [41, 94]. Finally, to *combat class imbalance*, we design a tailored cross-entropy loss that dynamically reweighs minority classes, thereby facilitating more equitable exploration and representation of tail classes in the target domain. Specifically, our contributions can be summarized as:

1. **Language-Guided SSDA Framework:** We pioneer the first language-guided SSDA framework for semantic segmentation by harnessing the rich semantic knowledge encoded in pre-trained Vision-Language Models (VLMs). Our novel attention-based fusion mechanism seamlessly integrates visual features with dense language embeddings, establishing a robust semantic bridge between source-target domains while providing enhanced contextual understanding.
2. **Enhanced Feature Localization:** Recognizing that VL pre-training primarily operates at the image level, we address the critical challenge of feature localization in semantic segmentation through targeted fine-tuning. To mitigate the risks of overfitting and semantic knowledge degradation inherent in limited-annotation scenarios, we develop a sophisticated consistency regularization framework that preserves the rich semantic representations acquired during pre-training.
3. **Adaptive Class-Balanced Loss:** To tackle class imbalance in a limited annotation scenario, we introduce a

Dynamic Cross-Entropy (DyCE) loss formulation that dynamically calibrates the learning emphasis on tail classes. This innovative, plug-and-play loss mechanism demonstrates broad applicability across various class-imbalanced learning scenarios.

4. **State-of-the-Art Performance:** Through detailed evaluation across diverse domain-adaptive and class-imbalanced segmentation benchmarks, our methodology demonstrates superior performance and robustness, consistently surpassing contemporary state-of-the-art approaches by significant margins.

## 2. Related Works

### 2.1. Semi-supervised Domain Adaptation

Recent advances in Semi-Supervised Domain Adaptation (SSDA) for semantic segmentation have focused on utilizing limited labeled target data and abundant unlabeled data to bridge the domain gap at the pixel level [1, 2]. Early approaches like MME [65] and ASDA [58] used entropy minimization for feature alignment, but their classification-centric strategies struggled with fine-grained segmentation tasks, leading to suboptimal boundary delineation. To address this, SSL-based methods such as DECOTA [91] and SS-ADA [89] employed teacher-student frameworks with consistency constraints, generating pseudo-labels for unlabeled target data. However, these methods faced issues with noisy pseudo-labels, particularly for minority and boundary classes. More recent methods have explored novel directions: S-Depth [32] leverages self-supervised depth estimation as an auxiliary task to enhance feature learning, while DSTC [24] introduces a domain-specific teacher-student framework that dynamically adapts to target domain characteristics. IIDM [23] proposes an innovative inter-intra-domain mixing strategy to address domain shift and limited supervision simultaneously. However, these methods often struggle with two critical limitations: class confusion due to limited supervision and skewed data distribution favoring source domain representations.

### 2.2. Vision Language Model

Vision-Language Models (VLMs), like CLIP and its extensions [34, 50, 60, 99], leverage large-scale image-text pre-training for semantic segmentation via a shared embedding space that aligns visual and textual features. Initial zero-shot methods, such as MaskCLIP [98] and GroupViT [87], struggled with boundary precision due to reliance on high-level features. Later, fine-tuned models like OpenSeg [25] and LSeg [40] improved segmentation accuracy using labeled data and text embeddings. Techniques such as ZegFormer [17] and OVSeg [46] utilize frozen CLIP features for mask proposal classification, while ZegCLIP [100] aligns dense visual-text embeddings in a streamlined man-

ner. Recently, SemiVL [33] has shown that language cues can enhance semantic insights and mitigate class confusion; however, their application in domain adaptation remains under-explored. The recent LIDAPS model [53] and follow-up works [20, 37, 84] apply language guidance for domain bridging in panoptic segmentation but rely on manual thresholding for pseudo-mask filtering and a complex multi-stage training process. Despite their improvements, these methods still misclassify tail classes (e.g., *fence*, *bike*, *wall*), posing critical risks for applications like autonomous driving, where errors in identifying such objects can lead to severe consequences.

### 2.3. Class-Imbalance Handling

Class imbalance significantly hinders real-world semantic segmentation, as small object classes often appear less frequently and cover fewer pixels than dominant background classes, unlike balanced datasets like CIFAR-10/100, ImageNet, and Caltech-101/256 [15, 26, 38]. Data-level methods like oversampling/undersampling adjust sampling probabilities for minority classes but struggle in dense tasks due to uneven class distribution [64, 101]. Algorithmic strategies such as class-weighted losses address bias by penalizing rare classes more [47], but treating small object classes equally often leads to instability [69, 88].

In unsupervised domain adaptation (UDA) for segmentation, common strategies include data-level adjustments using source domain frequencies [29–31], and adaptive weighting based on target statistics [88], but these are computationally costly for dense predictions. Approaches that relax pseudo-label filtering for rare classes still inherit source biases, causing misclassifications [102]. Most UDA methods prioritize data-level sampling [56], overlooking the synergy of combining data and algorithmic approaches, which remains impractical and lacks generalizability for diverse tasks [68, 76, 82].

## 3. Proposed Method

In our SSDA setting, we utilize image-label pair from the source domain  $\mathcal{D}^{Sr}$ :  $\{(x_i^{Sr}, y_i^{Sr})\}_{i=1}^{\mathbb{N}^{Sr}}$ , a limited set of labeled target samples  $\{(x_i^{TrL}, y_i^{TrL})\}_{i=1}^{\mathbb{N}^{TrL}}$ , and a large pool of unlabeled target data  $\{(x_i^{TrU})\}_{i=1}^{\mathbb{N}^{TrU}}$ , where  $\mathbb{N}^{TrU} \gg \mathbb{N}^{TrL}$ . Our proposed SemiDAViL framework effectively tackles the challenges of SSDA by leveraging VL-pretrained encoders (subsection 3.1) for enriched semantic representation learning from  $Sr \cup TrL \cup TrU$ , addressing the issue of misclassification among visually similar classes. We incorporate dense semantic guidance from language embeddings (subsection 3.2) to enhance instance separability and reduce domain bias. Consistency-regularized SSL (subsection 3.3) mitigates over-reliance on source features, while the class-balancing DyCE loss (subsection 3.4)

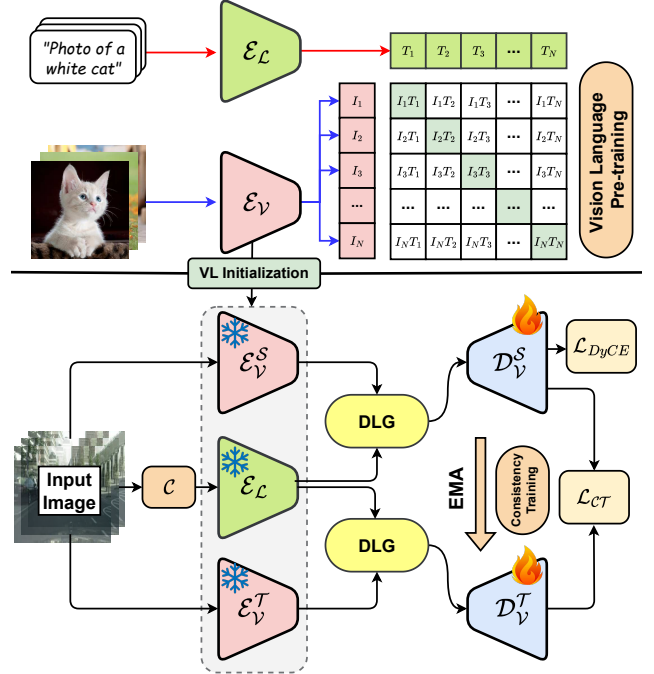


Figure 2. Overview of **SemiDAViL**: We leverage Vision-Language (VL) Pre-training (top) to initialize the language encoder  $\mathcal{E}_L$  and vision encoders  $\mathcal{E}_V^{\{S, T\}}$  in a semi-supervised setting (bottom), where  $S$  and  $T$  denote the student and teacher branches, respectively. To bridge image-level VL features for dense pixel-level tasks, we utilize a captioning model  $\mathcal{C}$  to generate text descriptions of images and a Dense Language Guidance (DLG) module. The framework is trained with a supervised loss  $\mathcal{L}_{DyCE}$  for labeled data and a consistency loss  $\mathcal{L}_{CT}$  for unlabeled data.

combats class imbalance by reweighting tail classes. The overall architecture of SemiDAViL is outlined in Figure 2.

### 3.1. Vision-Language Pre-training

Previous regularization-based SSDA methods have shown effectiveness in semi-supervised semantic segmentation by enforcing stable predictions on unlabeled data. However, as discussed in section 2, they often struggle with distinguishing visually similar classes, especially when only a limited set of labeled target samples  $\{(x_i^{TrL}, y_i^{TrL})\}$  is available. The primary issue arises due to the lack of diverse semantic coverage, leading to errors in class discrimination. To address this, we leverage Vision-Language Models (VLMs) like CLIP [60], which are trained on large-scale image-text datasets,  $\mathcal{D}_{clip} = \{(x, t)\}$ , where  $x$  and  $t$  are images and their associated captions. CLIP consists of a vision encoder  $\mathcal{E}_V$  and a language encoder  $\mathcal{E}_L$ , optimized jointly using a contrastive loss:

$$\mathcal{L}_{contrast} = - \sum_{i=1}^N \log \frac{\exp(\langle \mathcal{E}_V(x_i), \mathcal{E}_L(t_i) \rangle / \tau)}{\sum_{j=1}^N \exp(\langle \mathcal{E}_V(x_i), \mathcal{E}_L(t_j) \rangle / \tau)}, \quad (1)$$

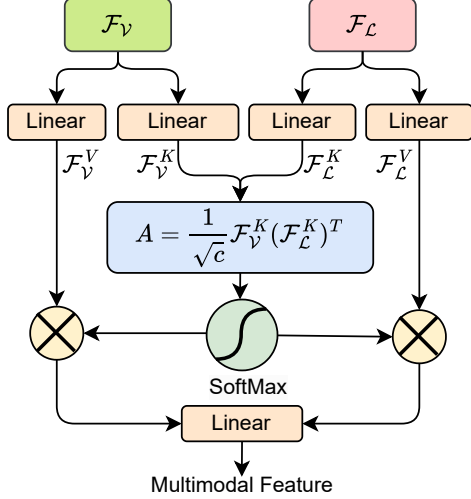


Figure 3. Overall architecture of our proposed DLG module: it is based on dense similarity maps of the vision and text embeddings. More details are provided in [subsection 3.2](#).

where  $\langle \cdot, \cdot \rangle$  denotes the cosine similarity and  $\tau$  is a temperature parameter. This objective aligns the visual and textual embeddings into a shared semantic space, learning robust, class-agnostic representations that generalize across diverse classes. Unlike conventional ImageNet pre-training that relies on manually annotated labels, CLIP’s training with web-crawled image-caption pairs allows it to capture richer semantics without restricting to a fixed set of categories.

To mitigate the limited semantic knowledge in standard consistency training in our SSDA framework, we initialize our (student-teacher) segmentation encoders  $\mathcal{E}_V^{\{S, T\}}$  with CLIP’s pre-trained vision encoder  $\mathcal{E}_V$ , rather than using an ImageNet-trained backbone. This transfer of rich semantic priors enables enhanced feature extraction and better semantic differentiation (as found in [Table 3](#) and well supported in [\[33\]](#)), particularly for visually ambiguous classes, leading to more robust segmentation performance.

### 3.2. Dense Language Guidance (DLG)

Most prior VLM methods employ a standard attention mechanism for multi-modal feature integration [\[16, 49, 96\]](#), i.e., features from two modalities (query and key) generate an attention matrix to aggregate vision features based on language-derived weights. However, this approach only utilizes the language feature to compute attention scores, without directly incorporating it into the fused output, effectively treating the result as a reorganized single-modal vision feature. Consequently, the output vision feature dominates the decoder, leading to a substantial loss of language information. Based on our empirical findings (provided in [supplementary file](#), and well supported in [\[49\]](#)), we argue that while generic attention effectively processes value inputs, it fails to fully exploit query features for deep cross-

modal interaction, resulting in insufficient fusion of vision and language modalities.

To address this, we utilize Dense Language Guidance (DLG) that transforms both the vision and language features into key-query pairs and treats them equally, as shown in [Figure 3](#). First, visual features  $\mathcal{F}_V \in \mathbb{R}^{h \times w \times c}$  with  $h \times w$  dimension and  $c$  channels for image  $\mathcal{X}$  are extracted through  $\mathcal{E}_V^1$  with frozen weight  $\phi_V$ :  $\mathcal{F}_V \leftarrow \mathcal{E}_V(\mathcal{X}; \phi_V^*)$ . To further utilize language embeddings  $\mathcal{F}_L \in \mathbb{R}^{n_L \times c}$ , we extract text description with  $n_L$  tokens for  $\mathcal{X}$  using off-the-shelf captioning model  $\mathcal{C}$ , followed by CLIP-initialized language encoder  $\mathcal{E}_L$  with frozen weights  $\phi_L$ :  $\mathcal{F}_L \leftarrow \mathcal{E}_L(\mathcal{C}(\mathcal{X}); \phi_L^*)$ . This is followed by projection of  $\mathcal{F}_{\{V, L\}}$  to key-value pairs using linear layers:  $\mathcal{F}_{\{V, L\}}^{\{K, V\}} \leftarrow \text{Linear}(\mathcal{F}_{\{V, L\}})$ . Next, multi-modal key values are used to generate an attention matrix  $\mathcal{A} \in \mathbb{R}^{n_L \times h \times w}$ :

$$\mathcal{A} = \frac{1}{\sqrt{c}} \mathcal{F}_V^K \cdot (\mathcal{F}_L^K)^T \quad (2)$$

Instead of applying attention to a single modality as in conventional methods, we normalize across both dimensions and compute cross-attention on vision and language features. Specifically, we employ a SoftMax activation followed by attention over  $\mathcal{F}_V^V$  and  $\mathcal{F}_L^V$  to generate language-attended vision features and vision-attended language features, respectively, ensuring balanced and comprehensive feature fusion:

$$\begin{aligned} \mathcal{F}_V^A &= \text{SoftMax}[\mathcal{A}] \mathcal{F}_V^V \\ \mathcal{F}_L^A &= \text{SoftMax}[\mathcal{A}] \mathcal{F}_L^V \end{aligned} \quad (3)$$

Finally, these two attended feature maps are combined to generate a true multimodal feature representation  $\mathcal{F}_M \in \mathbb{R}^{n_L \times h \times w}$ :  $\mathcal{F}_M = \mathcal{F}_V^A \cdot (\mathcal{F}_L^A)^T$ , which is the basis of our VL-guided SSDA pipeline. This is thereafter passed to the student-teacher decoders  $\mathcal{D}_V^{\{S, T\}}$  for consistency training.

### 3.3. Consistency Training (CT)

To effectively utilize labeled and unlabeled data in our SSDA setting, we utilize a student-teacher network [\[70\]](#) for consistency training. Specifically, for unlabeled target data  $\{(x_i^{Tru})_{i=1}^{N^{Tru}} \in \mathcal{D}^{Tru}$ , we obtain two multimodal features  $\mathcal{F}_M^S, \mathcal{F}_M^T$  (from DLG), and pass them through two identical but differently initialized decoders  $\{\mathcal{D}_V^S, \mathcal{D}_V^T\}$  with trainable parameters  $\{\theta_V^S, \theta_V^T\}$ , respectively, and enforce their predictions to be consistent:

$$L_{CT} = \frac{1}{N^{Tru}} \sum_{x_i \in \mathcal{D}^{Tru}} \sum_{p=1}^{h \times w} \mathbb{1}_{\max(y_p^S) \geq Th} CE(y_p^S, y_p^T) \quad (4)$$

<sup>1</sup> student-teacher encoders  $\mathcal{E}_V^{\{S, T\}}$  represented as  $\mathcal{E}_V$  for simplicity.

where  $y_p^S$  and  $y_p^T$  are the  $p^{\text{th}}$  pixel prediction from student and teacher model:  $y^m \leftarrow \mathcal{D}_V^m(\mathcal{F}_M; \theta_V^m)$ ;  $m = \{S, T\}$ ,  $Th$  is a threshold to exclude noisy pseudo-labels in  $L_{CT}$ . To further utilize labeled target data  $\{\mathcal{D}^{Sr} \cup \mathcal{D}^{Tr\mathcal{L}}\}$ , we can employ a supervised CE loss between ground truth  $y$  and student prediction  $y^S$ :

$$L_S = \frac{1}{\mathbb{N}^{Sr} + \mathbb{N}^{Tr\mathcal{L}}} \sum_{x_i \in \{\mathcal{D}^{Sr} \cup \mathcal{D}^{Tr\mathcal{L}}\}} \sum_{c=0}^{N_C} CE(y_c^S, y_c), \quad (5)$$

where  $N_C$  represents the number of classes. However,  $L_S$  might incur suboptimal performance due to inherent class imbalance, as discussed in [subsection 2.3](#) and evident in previous methods in [Table 4](#), [Table 5](#). We propose a dynamic CE loss to alleviate this shortcoming, as detailed in [subsection 3.4](#). The student branch is updated using a combined consistency and DyCE loss, whereas the teacher model is updated using an exponential moving average (EMA) of the student parameters:

$$\theta_V^T(t) \leftarrow \alpha \theta_V^T(t-1) + (1-\alpha) \theta_V^S(t) \quad (6)$$

where  $t$  is step number,  $\alpha$  is the momentum coefficient [\[28\]](#).

### 3.4. Class-balanced Dynamic CE Loss (DyCE)

The Cross-Entropy (CE) loss measures the difference between predicted probabilities and ground truth labels by computing a negative log-likelihood for each class, averaged over all instances in the mini-batch:

$$L_{CE} = -\frac{1}{S} \sum_{i=1}^S \sum_{c=0}^{N_C} y_{i,c} \log p_{i,c}, \quad (7)$$

where  $y_{i,c}$  and  $p_{i,c}$  are the GT and predicted probability for class  $c$ ,  $S$  is the batch size. Taking the gradient of  $L_{CE}$  for each instance, we have:

$$\frac{\partial L_{CE}}{\partial p_{i,c}} = \begin{cases} -\frac{1}{S} \frac{1}{p_{i,c}}, & \text{if } y_{i,c} = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

Hence, CE loss only updates the gradient for the target class per instance, using a uniform weight of  $-\frac{1}{S}$ . This leads to two key problems in large imbalanced datasets: (1) equal weighting across classes overlooks class imbalance, treating frequent and rare classes the same; (2) the gradient magnitude becomes vanishingly small as  $N$  scales to millions, causing ineffective updates (gradient vanishing). While recent studies have proposed reweighting schemes to address the class imbalance issue [\[59, 73\]](#), they fail to tackle the core problem of diminished gradient magnitudes (refer to supporting evidence in [supplementary file](#)), limiting the optimization efficiency in dense segmentation tasks.

To address this, we propose a Dynamic CE (DyCE) loss that dynamically adjusts the weighting of gradients based

on the class distribution within each mini-batch, addressing the persistent class imbalance issue that remains even after discarding simple instances. The key idea is to adaptively align the gradient contributions to the real-time class distribution at every training step. This is formalized as:

$$L_{\text{DyCE}} = -\frac{1}{f_H^\omega} \sum_{c=0}^{N_C} \frac{1}{f_c^{(1-\omega)}} \sum_{i=1}^S \mathbb{1}_{i \in H} y_{i,c} \log p_{i,c}, \quad (9)$$

where  $f_c = \sum_{i \in H} y_{i,c}$  is the total count of class  $c$  in the mined subset  $H$  which consists of  $h\%$  hardest instances from the batch,  $f_H = |H|$  is the count of instances in subset  $H$ . The loss computation involves four key steps: (1) computing the standard CE loss for each sample; (2) creating a subset  $H$  from the batch, similar to [\[85\]](#); (3) assigning dynamic class weights  $\frac{1}{f_c^{(1-\omega)}}$ , inversely proportional to the mined class frequency; and (4) scaling the loss by a volume weight  $-\frac{1}{f_H^\omega}$ , which adjusts for the batch size and mined subset size. The hyperparameter  $\omega \in (0, 1)$  acts as a weight-balancing factor, balancing the influence of instance-level and class-level weighting. The gradient of DyCE loss is:

$$\frac{\partial L_{\text{DyCE}}}{\partial p_{i,c}} = \begin{cases} -\frac{1}{f_H^\omega} \frac{1}{f_c^{(1-\omega)}} p_{i,c} & \text{if } y_{i,c} = 1, \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

Here  $\frac{1}{f_H^\omega} \frac{1}{f_c^{(1-\omega)}} \geq \frac{1}{S}$  as compared to [Equation 8](#), as  $S \geq f_H \geq f_c$  and hence the vanishing gradient issue is resolved.

## 4. Experimental Results

### 4.1. Dataset Description

We evaluate our proposed SSDA method on a segmentation task by adapting from two synthetic datasets, GTA5 [\[62\]](#) and SYNTHIA [\[63\]](#), to the real-world Cityscapes dataset [\[14\]](#). The Cityscapes dataset consists of 2,975 training images and 500 validation images, all manually annotated with 19 classes. Since the test set annotations are not publicly available, we evaluate on the validation set, and tune the hyper-parameters on a small subset of the training set, following previous works [\[4, 32\]](#). GTA5 provides 24,966 training images, and we consider the 19 classes that overlap with Cityscapes. The SYNTHIA dataset includes 9,400 fully labeled images, and we evaluate results based on the 16 classes it shares with Cityscapes.

Furthermore, to validate our DyCE loss's effectiveness, we evaluate on an extremely imbalanced medical dataset, Synapse [\[39\]](#). The Synapse dataset comprises 30 CT scans covering 13 different organs (i.e., foreground classes): spleen (Sp), right and left kidneys (RK/LK), gallbladder (Ga), esophagus (Es), liver (Li), stomach (St), aorta (Ao), inferior vena cava (IVC), portal and splenic veins (PSV), pancreas (Pa), and right and left adrenal glands

Table 1. Quantitative comparison of our proposed method with existing unsupervised domain adaptation (UDA), semi-supervised learning (SSL), and semi-supervised domain adaptation (SSDA) methods on **GTA5**  $\rightarrow$  **Cityscapes** benchmark. We report 19-class mIoU scores on the Cityscapes validation set across 0, 100, 200, 500, 1000, and 2975 (100%) labeled target images. Our results are **highlighted** whereas the previous best and second-best results are marked in **red** and **blue**.

Type	Methods	Labeled Target Samples					
		0	100	200	500	1000	2975
Supervised	DeepLabV2 [8]	-	43.0	48.3	54.8	58.3	66.1
UDA	PRoDA [97]	57.5	-	-	-	-	-
	DaFormer [29]	56.0	-	-	-	-	-
	CONFETI [45]	<b>62.2</b>	-	-	-	-	-
	DIGA [67]	<b>62.7</b>	-	-	-	-	-
SSL	CowMix [22]	-	50.8	54.8	61.7	64.8	-
	ClassMix [57]	-	54.4	58.6	62.1	64.3	-
	CPS (S) [10]	-	55.0	59.5	63.0	65.7	-
	CPS (E) [10]	-	<b>55.3</b>	<b>60.0</b>	<b>63.6</b>	<b>66.3</b>	-
	DusPerb [92]	-	<b>61.8</b>	<b>66.7</b>	<b>68.4</b>	<b>72.1</b>	-
SSDA	ASS [86]	-	54.2	56.0	60.2	64.5	69.1
	DECOTA (S) [91]	-	60.7	61.8	64.2	66.3	69.2
	DECOTA (E) [91]	-	61.3	62.3	64.7	67.0	69.9
	DLDM [9]	-	61.2	60.5	64.3	66.6	69.8
	SSDDA [13]	-	60.1	62.9	65.7	66.8	-
	DSTC (S) [24]	-	64.5	65.8	69.2	70.3	71.9
	DSTC (E) [24]	-	65.2	66.4	<b>70.0</b>	<b>70.9</b>	<b>72.6</b>
	S-Depth [32]	-	<b>66.1</b>	<b>67.3</b>	69.9	70.5	71.7
	IIDM [23]	-	<b>69.5</b>	<b>70.0</b>	<b>70.6</b>	<b>72.8</b>	<b>73.3</b>
	<b>Ours w/o DyCE</b>	<b>66.9</b>	<b>70.3</b>	<b>71.8</b>	<b>72.1</b>	<b>73.9</b>	<b>74.7</b>
	<b>Ours w/ DyCE</b>	<b>67.7</b>	<b>71.1</b>	<b>72.5</b>	<b>72.9</b>	<b>74.8</b>	<b>75.2</b>

(RAG/LAG). In this dataset, foreground voxels make up only 4.37% of the entire dataset, with 95.63% background, and the right adrenal gland contributes a mere 0.14% of foreground, whereas liver consists of 53.98% foreground, underscoring the severe class imbalance. Following the setup of [76], we split the dataset into 20 scans for training, 4 for validation, and 6 for testing.

## 4.2. Implementation Details

Following SemiVL [33], we utilize ViT-B/16 vision encoder [18] and a Transformer text encoder [74], both initialized with CLIP pre-training [60] and generate dense embeddings following [98]. The initial learning rate is set to  $10^{-4}$ , decaying exponentially with a factor of 0.9. We set the weight decay to  $2 \times 10^{-4}$  and momentum to 0.9. Following [35], we use BLIP-2 [43] as our off-the-shelf captioning model  $\mathcal{C}$  for all domains.  $Th$ ,  $\alpha$  is set to 0.95, 0.999, following [28, 33]. Following [23, 72], source images are resized to  $760 \times 1280$  and target images to  $512 \times 1024$ , followed by random cropping to  $512 \times 512$ . SemiDAViL is trained for 40k iterations which takes  $\sim 15$  hours on an NVIDIA RTX4090 GPU using Python environment.

## 4.3. Findings and Comparison with SoTA

Our proposed SemiDAViL framework demonstrates significant gains on both GTA5 $\rightarrow$ Cityscapes and

Table 2. Quantitative comparison of our method with existing UDA, SSL, and SSDA methods on **Synthia**  $\rightarrow$  **Cityscapes** benchmark. We report 16-class mIoU scores on the Cityscapes validation set and follow the same settings as in Table 1. DACS++<sup>†</sup> represents implementation of DACS [72] from UDA to SSDA.

Type	Methods	Labeled Target Samples					
		0	100	200	500	1000	2975
Supervised	DeepLabV2 [8]	-	53.0	58.9	61.0	67.5	70.8
UDA	DaCS [72]	54.8	-	-	-	-	-
	PRoDA [97]	<b>62.0</b>	-	-	-	-	-
	DaFormer [29]	58.8	-	-	-	-	-
	DIGA [67]	<b>67.9</b>	-	-	-	-	-
SSL	CowMix [22]	-	61.3	-	-	-	-
	ClassMix [57]	-	<b>61.4</b>	<b>67.6</b>	<b>72.3</b>	<b>73.1</b>	-
	DMT [21]	-	59.7	-	-	-	-
	DusPerb [92]	-	<b>68.4</b>	<b>71.4</b>	<b>74.2</b>	<b>76.1</b>	-
SSDA	ASS [86]	-	62.1	64.8	69.8	73.0	77.1
	ComplexMix [12]	-	70.6	-	-	-	75.6
	DACS++ <sup>†</sup> [72]	-	64.9	67.7	71.3	72.8	74.4
	DLDM [9]	-	68.4	-	-	-	-
	SSDDA [13]	-	70.6	71.8	72.6	74.0	-
	ALFSA [81]	-	68.9	<b>73.5</b>	<b>77.5</b>	<b>79.0</b>	<b>79.9</b>
	SLA [95]	-	63.7	-	-	-	-
	S-Depth [32]	-	<b>72.4</b>	<b>73.5</b>	75.4	76.3	77.1
	IIDM [23]	-	<b>74.2</b>	<b>76.4</b>	<b>77.0</b>	<b>78.8</b>	<b>79.2</b>
	<b>Ours w/o DyCE</b>	<b>69.5</b>	<b>74.9</b>	<b>76.8</b>	<b>77.7</b>	<b>79.2</b>	<b>79.6</b>
<b>Ours w/ DyCE</b>	<b>70.2</b>	<b>76.9</b>	<b>77.2</b>	<b>78.6</b>	<b>79.7</b>	<b>80.5</b>	

Table 3. Ablation experiments using three different SSDA settings on **GTA5** $\rightarrow$ **Cityscapes** and **Synthia** $\rightarrow$ **Cityscapes** to identify the contribution of individual components: Consistency Training (CT), Dynamic Cross-Entropy loss (DyCE), Vision-Language Pre-training (VLP), and DenseLanguage Guidance (DLG).

Components	GTA5 $\rightarrow$ Cityscapes				Synthia $\rightarrow$ Cityscapes							
	CT	DyCE	VLP	DLG	100	200	500	1000	100	200	500	1000
✓	-	-	-	-	54.5	58.2	62.3	64.6	60.2	65.3	71.5	72.0
✓	✓	-	-	-	63.3	64.5	65.9	68.2	68.7	70.1	72.2	74.7
✓	-	✓	-	-	65.6	66.8	69.1	69.9	71.9	73.0	74.7	75.9
✓	-	✓	✓	-	70.3	71.6	72.1	73.9	74.9	76.8	77.7	79.2
✓	✓	✓	✓	-	<b>71.1</b>	<b>72.5</b>	<b>72.9</b>	<b>74.8</b>	<b>76.9</b>	<b>77.2</b>	<b>78.6</b>	<b>79.7</b>

Synthia $\rightarrow$ Cityscapes benchmarks (Table 1, Table 2). We compare against state-of-the-art UDA, SSL, and SSDA techniques, highlighting its robustness with varying levels of target annotations.

In the **GTA5** $\rightarrow$ **Cityscapes** scenario, (A) using only 100 labeled target samples, SemiDAViL achieves 71.1% mIoU, outperforming the previous best, IIDM [23], by 1.6%. The advantage grows with 200 labeled samples, where we attain 72.5% mIoU, showcasing our framework’s strength in leveraging limited annotations through language-guided features; (B) in the fully unsupervised setting, our method achieves 67.7% mIoU, a 5% improvement over the previous best, DIGA [67], owing to our language-guided joint embedding, which provides more robust semantic alignment than divergence-based methods like DaFormer [29]; and (C) with 2975 labeled samples, our model reaches 75.2%

Table 4. Class-wise performance evaluation of our proposed method (with and without the proposed class-balancing DyCE loss), and comparison with the existing class-balanced UDA and SSDA methods. We report 19-class and 16-class mIoU scores on the **GTA5**  $\rightarrow$  **Cityscapes** and **Synthia**  $\rightarrow$  **Cityscapes** settings, respectively with 100 labeled target samples. The segmentation performance of tailed classes significantly improves by incorporating our DyCe loss in both settings. Our results are **highlighted** whereas the previous-best and second-best results are marked in **red** and **blue**. Please refer to **supplementary file** for detailed class distribution and improvement analysis.

		GTA5 $\rightarrow$ Cityscapes																				
Type	Methods	Target Labels	Road	Sidewalk	Building	Walls	Fence	Pole	T-Light	T-sign	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
Supervised	UniMatch [92]	2975	97.2	79.3	90.6	36.5	52.1	56.7	64.2	72.1	91.1	59.0	93.6	77.5	53.5	93.4	73.8	79.8	67.8	49.6	71.2	71.5
	U2PL [79]		97.6	82.1	90.7	39.3	53.4	58.0	64.2	74.0	90.9	60.2	93.7	77.0	49.8	93.7	64.8	77.9	47.9	51.2	73.6	70.5
UDA	CADA [90]	0	91.3	46.0	84.5	34.4	29.7	32.6	35.8	36.4	84.5	43.2	83.0	60.0	32.2	83.2	35.0	46.7	0.0	33.7	42.2	49.2
	IAST [54]		<b>93.8</b>	<b>57.8</b>	85.1	39.5	26.7	26.2	43.1	34.7	84.9	32.9	<b>88.0</b>	62.6	29.0	87.3	39.2	49.6	<b>23.2</b>	34.7	39.6	51.5
	DACS [72]		89.9	39.7	<b>87.9</b>	30.7	<b>39.5</b>	38.5	46.4	52.8	<b>88.0</b>	<b>44.0</b>	<b>88.8</b>	67.2	35.8	84.5	<b>45.7</b>	50.2	0.0	27.3	34.0	52.1
	Shallow [4]		91.9	48.9	<b>86.0</b>	38.6	28.6	34.8	45.6	43.0	86.2	42.4	87.6	65.6	<b>38.6</b>	86.8	38.4	48.2	0.0	46.5	<b>59.2</b>	53.5
	ProDA+distill [44]		87.8	56.0	79.7	<b>46.3</b>	<b>44.8</b>	<b>45.6</b>	<b>53.5</b>	<b>53.5</b>	<b>88.6</b>	<b>45.2</b>	82.1	<b>70.7</b>	<b>39.2</b>	<b>88.8</b>	45.5	<b>59.4</b>	1.0	<b>48.9</b>	<b>56.4</b>	<b>57.5</b>
	CPSL+distill [44]		<b>92.3</b>	<b>59.9</b>	84.9	<b>45.7</b>	29.7	<b>52.8</b>	<b>61.5</b>	<b>59.5</b>	87.9	41.5	85.0	<b>73.0</b>	35.5	<b>90.4</b>	<b>48.7</b>	<b>73.9</b>	<b>26.3</b>	<b>53.8</b>	<b>53.9</b>	<b>60.8</b>
SSDA	ALFSA [81]	100	<b>95.9</b>	<b>71.5</b>	<b>87.4</b>	<b>39.9</b>	<b>39.0</b>	<b>44.6</b>	<b>52.6</b>	<b>60.4</b>	<b>89.1</b>	<b>50.7</b>	<b>91.3</b>	<b>73.1</b>	<b>48.3</b>	<b>91.3</b>	<b>55.3</b>	<b>63.7</b>	<b>26.3</b>	<b>55.8</b>	<b>68.7</b>	<b>63.4</b>
	SS-ADA+UniMatch [89]		<b>96.4</b>	<b>75.0</b>	<b>89.2</b>	<b>43.7</b>	<b>45.1</b>	<b>53.3</b>	<b>58.2</b>	<b>68.8</b>	<b>90.7</b>	<b>55.4</b>	<b>93.8</b>	<b>75.8</b>	<b>49.7</b>	<b>91.6</b>	54.6	<b>67.4</b>	<b>43.6</b>	47.2	<b>69.4</b>	<b>66.8</b>
	SS-ADA+U2PL [89]		<b>96.5</b>	<b>75.5</b>	<b>89.7</b>	<b>47.1</b>	<b>47.7</b>	<b>55.3</b>	<b>60.6</b>	<b>68.1</b>	<b>90.6</b>	<b>55.3</b>	<b>92.1</b>	<b>77.4</b>	<b>52.5</b>	<b>92.5</b>	<b>67.1</b>	<b>67.8</b>	<b>41.2</b>	<b>49.9</b>	<b>70.8</b>	<b>68.3</b>
	Ours w/o DyCE		<b>97.3</b>	<b>80.2</b>	<b>89.5</b>	<b>50.2</b>	<b>49.2</b>	<b>58.3</b>	<b>62.3</b>	<b>69.3</b>	<b>90.6</b>	<b>57.7</b>	<b>93.2</b>	<b>78.6</b>	<b>53.9</b>	<b>92.9</b>	<b>68.4</b>	<b>67.9</b>	<b>47.9</b>	<b>56.5</b>	<b>70.9</b>	<b>70.3</b>
	Ours w/ DyCE		<b>98.1</b>	<b>80.9</b>	<b>90.3</b>	<b>51.9</b>	<b>51.5</b>	<b>60.2</b>	<b>62.5</b>	<b>71.1</b>	<b>90.9</b>	<b>59.5</b>	<b>93.2</b>	<b>79.9</b>	<b>56.8</b>	<b>93.2</b>	<b>71.6</b>	<b>68.1</b>	<b>49.1</b>	<b>58.6</b>	<b>71.4</b>	<b>71.1</b>
		Synthia $\rightarrow$ Cityscapes																				
Supervised	UniMatch [92]	2975	97.5	82.1	91.2	52.4	53.0	60.7	66.3	75.3	92.3	-	94.1	79.9	57.5	94.4	-	82.1	-	57.9	74.5	75.7
	U2PL [79]		97.5	81.7	90.0	36.9	50.9	56.8	59.9	71.7	91.6	-	93.1	76.5	43.5	93.6	-	75.4	-	45.2	72.1	71.0
UDA	FADA [77]	0	84.5	40.1	83.1	4.8	0.0	34.3	20.1	27.2	84.8	-	84.0	53.5	22.6	85.4	-	43.7	-	26.8	27.8	45.2
	IAST [54]		81.9	41.5	83.3	17.7	<b>4.6</b>	32.3	30.9	28.8	83.4	-	85.0	65.5	30.8	86.5	-	38.2	-	33.1	52.7	49.8
	DACS [72]		80.6	25.1	81.9	21.5	<b>2.6</b>	37.2	22.7	24.0	83.7	-	<b>90.8</b>	67.6	<b>38.3</b>	82.9	-	38.9	-	28.5	47.6	48.3
	Shallow [4]		<b>90.4</b>	<b>51.1</b>	83.4	3.0	0.0	32.3	25.3	31.0	84.8	-	85.5	59.3	30.1	82.6	-	<b>53.2</b>	-	17.5	45.6	48.4
	ProDA+distill [44]		87.8	<b>45.7</b>	<b>84.6</b>	<b>37.1</b>	0.6	<b>44.0</b>	<b>54.6</b>	<b>37.0</b>	<b>88.1</b>	-	84.4	<b>74.2</b>	24.3	88.2	-	<b>51.1</b>	-	<b>40.5</b>	45.6	<b>55.5</b>
	CPSL+distill [44]		87.2	43.9	<b>85.5</b>	<b>33.6</b>	0.3	<b>47.7</b>	<b>57.4</b>	<b>37.2</b>	<b>87.8</b>	-	<b>88.5</b>	<b>79.0</b>	<b>32.0</b>	<b>90.6</b>	-	49.4	-	<b>50.8</b>	<b>59.8</b>	<b>57.9</b>
SSDA	SS-ADA+U2PL [89]	100	<b>91.0</b>	<b>62.0</b>	<b>86.7</b>	<b>38.9</b>	<b>33.4</b>	<b>53.6</b>	<b>58.9</b>	<b>69.0</b>	<b>91.0</b>	-	<b>92.5</b>	<b>73.9</b>	<b>44.6</b>	<b>92.3</b>	-	<b>69.3</b>	-	<b>37.3</b>	<b>67.2</b>	<b>66.4</b>
	SS-ADA+UniMatch [89]		<b>97.1</b>	<b>79.4</b>	<b>90.2</b>	<b>49.8</b>	<b>49.8</b>	<b>56.9</b>	<b>58.2</b>	<b>72.2</b>	<b>91.6</b>	-	<b>93.4</b>	<b>78.1</b>	<b>53.3</b>	<b>92.8</b>	-	<b>69.1</b>	-	<b>48.4</b>	<b>72.1</b>	<b>72.0</b>
	Ours w/o DyCE		<b>98.5</b>	<b>78.9</b>	<b>91.6</b>	<b>52.7</b>	<b>52.8</b>	<b>62.3</b>	<b>63.9</b>	<b>74.3</b>	<b>91.5</b>	-	<b>94.4</b>	<b>79.7</b>	<b>56.9</b>	<b>93.1</b>	-	<b>76.6</b>	-	<b>55.5</b>	<b>74.9</b>	<b>74.9</b>
	Ours w/ DyCE		<b>98.9</b>	<b>80.9</b>	<b>92.2</b>	<b>57.6</b>	<b>56.2</b>	<b>63.8</b>	<b>67.1</b>	<b>76.7</b>	<b>91.9</b>	-	<b>95.9</b>	<b>80.6</b>	<b>59.9</b>	<b>93.8</b>	-	<b>78.9</b>	-	<b>59.8</b>	<b>76.6</b>	<b>76.9</b>

mIoU, surpassing the previous best, IIDM by 1.9%, attributed to the effective handling of class imbalance via the DyCE loss.

For **Synthia** $\rightarrow$ **Cityscapes**, similar gains are observed against the previous best methods like IIDM [23] and ALFSA [81], demonstrating the strength of our dense language guidance and adaptive DyCE loss. Competing methods, such as DACS++ [72], attempt to address class imbalance via pseudo-label refinement, but their reliance on multi-stage training and hyperparameter tuning limits scalability. Unlike these, our end-to-end solution with DyCE loss adaptively re-weights based on class distribution, yielding balanced learning without the need for manual adjustments.

Overall, our framework consistently outperforms across all benchmarks and supervision levels, demonstrating its robustness and scalability. Detailed class-wise analysis is provided in [subsection 4.5](#).

#### 4.4. Ablation Experiments

We conduct ablation experiments to analyze the effectiveness of each component in our proposed framework: Consistency Training (CT), Dynamic Cross-Entropy loss (DyCE), Vision-Language Pre-training (VLP), and Dense Language Guidance (DLG) in [Table 3](#). **(A)** Starting with CT as our baseline, we observe moderate performance with mIoU scores of 54.5% and 60.2% on GTA5 and Synthia respectively with 100 labeled samples. **(B)** The addition of our DyCE loss significantly improves performance by addressing class imbalance issues, boosting the mIoU by

8.8% (to 63.3%) and 8.5% (to 68.7%) respectively. This, along with the findings in [Table 4](#) and [Table 5](#) demonstrates the effectiveness of the proposed dynamic loss function to address class imbalance. **(C)** When incorporating VLP without DyCE, we achieve better results than DyCE alone, with mIoU improvements of 11.1% (to 65.6%) and 11.7% (to 71.9%) over the baseline. This demonstrates the effectiveness of leveraging semantic knowledge from pre-trained vision-language models. **(D)** The addition of DLG further enhances performance substantially, reaching 70.3% and 74.9% mIoU, as it enables fine-grained semantic understanding through dense language embeddings. **(E)** Finally, our full model combining all components achieves the best performance across all settings, with notable improvements of 16.6% (to 71.1%) and 16.7% (to 76.9%) over the baseline with 100 labeled samples. The consistent performance gains across different label ratios demonstrate the complementary nature of our proposed components in addressing the challenges of SSDA for semantic segmentation.

#### 4.5. Detailed Analysis on the DyCE Loss

We provide a comprehensive evaluation of our proposed method (with and without the DyCE loss) against existing UDA and SSDA methods that use various solutions on the class-imbalance problem on the GTA5  $\rightarrow$  Cityscapes and Synthia  $\rightarrow$  Cityscapes benchmarks, using 100 labeled target samples in [Table 4](#). **(A)** Without DyCE, our model already achieves competitive mIoU scores, but the addition of DyCE enhances class separability by adaptively re-

Table 5. Comparative performance improvement by incorporating our proposed DyCE loss to address **class imbalance** in 20% labeled Synapse dataset in SSL setting. Please refer to **supplementary file** for detailed class-distribution and improvement analysis.

Type	Methods	Average						Average DSC <sup>†</sup> of Each Class									
		DSC <sup>†</sup>	ASD <sup>↓</sup>	Sp	RK	LK	Ga	Es	Li	St	Ao	IVC	PSV	PA	RAG	LAG	
Supervised	V-Net [55]	62.1	10.3	84.6	77.2	73.8	73.3	38.2	94.6	68.4	72.1	71.2	58.2	48.5	17.9	29.0	
General SSL	UA-MT [93]	20.3	71.7	48.2	31.7	22.2	0.0	0.0	81.2	29.1	23.3	27.5	0.0	0.0	0.0	0.0	
	URPC [51]	25.7	72.7	60.7	38.2	56.8	0.0	0.0	85.3	33.9	33.1	14.8	0.0	5.1	0.0	0.0	
	CPS [11]	33.6	41.2	62.8	55.2	45.4	35.9	0.0	91.1	31.3	41.9	49.2	8.8	14.5	0.0	0.0	
	SS-Net [83]	35.1	50.8	62.7	67.9	60.9	34.3	0.0	89.9	20.9	61.7	44.8	0.0	8.7	4.2	0.0	
	DePL [78]	36.3	36.0	62.8	61.0	48.2	54.8	0.0	90.2	36.0	42.5	48.2	10.7	17.0	0.0	0.0	
Class-balanced SSL	Adsh [27]	35.3	39.6	55.1	59.6	45.8	52.2	0.0	89.4	32.8	47.6	53.0	8.9	14.4	0.0	0.0	
	CRest [80]	38.3	22.9	62.1	64.7	53.8	43.8	8.1	85.9	27.2	54.4	47.7	14.4	13.0	18.7	4.6	
	SimiS [6]	40.1	33.0	62.3	69.4	50.7	61.4	0.0	87.0	33.0	59.0	57.2	29.2	11.8	0.0	0.0	
	ACISSMIS [3]	33.2	43.8	57.4	53.8	48.5	46.9	0.0	87.8	28.7	42.3	45.4	6.3	15.0	0.0	0.0	
	CLD [48]	41.1	32.2	62.0	66.0	59.3	61.5	0.0	89.0	31.7	62.8	49.4	28.6	18.5	0.0	5.0	
	DHC [73]	48.6	10.7	62.8	69.5	59.2	66.0	13.2	85.2	36.9	67.9	61.5	37.0	30.9	31.4	10.6	
A&D [76]	60.9	2.5	85.2	66.9	67.0	52.7	62.9	89.6	52.1	83.0	74.9	41.8	43.4	44.8	27.2		
SSL+Our DyCE loss	UA-MT[93]+DyCE	48.1(+27.9)	10.9(-60.8)	55.6(+7.4)	52.3(+20.6)	53.4(+31.2)	49.5(+49.5)	51.5(+51.5)	88.7(+7.5)	46.0(+16.9)	48.7(+25.4)	47.3(+19.8)	35.6(+35.6)	33.4(+33.4)	39.1(+39.1)	24.4(+24.4)	
	URPC[51]+DyCE	48.9(+23.2)	10.1(-62.6)	63.2(+2.5)	59.9(+21.7)	59.4(+2.6)	52.5(+52.5)	52.2(+52.2)	89.3(+4.0)	45.6(+11.7)	50.6(+17.5)	50.3(+35.5)	30.1(+30.1)	30.4(+25.3)	26.7(+26.7)	25.1(+25.1)	
	DePL[78]+DyCE	51.8(+15.6)	9.2(-26.1)	65.9(+3.1)	64.2(+3.2)	65.1(+16.9)	59.5(+4.7)	58.2(+58.2)	90.5(+0.3)	41.7(+5.7)	54.9(+12.4)	55.6(+7.4)	32.2(+21.5)	35.8(+18.8)	26.9(+26.9)	23.3(+23.3)	
	DHC[73]+DyCE	57.9(+9.3)	6.4(-4.3)	77.5(+14.7)	72.5(+3.0)	64.3(+5.1)	70.3(+4.3)	53.2(+40.0)	88.7(+3.5)	44.8(+7.9)	79.1(+11.2)	67.8(+6.3)	42.0(+5.0)	38.2(+7.3)	34.8(+3.4)	20.1(+9.5)	
A&D[76]+DyCE	65.5(+4.6)	5.8(-0.7)	87.3(+2.1)	78.5(+11.6)	72.4(+5.4)	65.1(+12.4)	64.7(+1.8)	90.9(+1.3)	55.3(+3.2)	84.1(+1.1)	77.5(+2.6)	47.3(+5.5)	49.8(+6.4)	48.4(+3.6)	30.1(+2.9)		

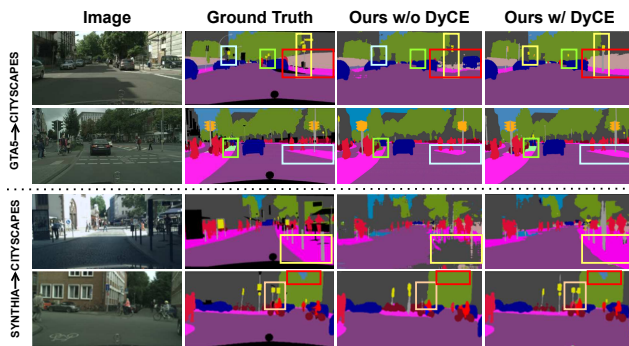


Figure 4. Qualitative segmentation performance of SemiDAViL with and without DyCE loss on 100 labeled target data.

weighting underrepresented classes based on their occurrence, addressing inherent dataset imbalances. **(B)** The integration of DyCE loss shows significant improvements across a variety of challenging classes, particularly in tailed categories with high intra-class variation such as *Fence*, *Wall*, *Terrain*, *Rider*, and *Pole*. We provide a qualitative visualization of improvement in segmentation performance by incorporating DyCE loss in Figure 4. For GTA5→Cityscapes, the performance on tail classes like *Wall* (51.9%), *Fence* (51.5%), and *Pole* (60.2%) shows substantial improvement over previous state-of-the-art methods SS-ADA+U2PL [89] (47.1%, 47.7%, and 55.3% respectively) and SS-ADA+UniMatch [89] (43.7%, 45.1%, and 53.3% respectively). **(C)** We observe a significant improvement over UDA approaches like [4, 44, 72]. **(D)** The overall mIoU rises from 70.3% to 71.4% by integrating DyCE loss, showing consistent performance gains across all label splits. Similar improvements are also evident in Synthia→Cityscapes, demonstrating the robust superiority of our language-guided and class-balanced approach.

To further validate the effectiveness of DyCE loss, we evaluate it on a more challenging scenario: Synapse medical dataset with severe class imbalance (95.63% background, 4.37% foreground: foreground classes vary from

53.98% to 0.14%). As summarized in Table 5, we highlight the significant impact of our DyCE loss as a plug-in enhancement across various SSL methods [51, 75, 76, 78, 93]. **(A)** For the lowest-performing general SSL method in Table 5, i.e., UA-MT [93], its mDSC leaps from 20.3% to 48.1%, with minority classes like *Gallbladder* improving from 0.0% to 49.5%, showcasing DyCE’s impressive adaptive weighting. Similarly, URPC’s mDSC increases from 25.7% to 48.9%, whereas the previous best general SSL-based DePL sees a boost from 36.3% to 51.8%, underscoring DyCE’s capability to mitigate severe class imbalance by prioritizing underrepresented organs (e.g., *Right Adrenal Gland* from 0.0% to 26.9%). **(B)** Recent SoTA of balanced SSL like DHC and A&D also benefit, with DHC’s DSC rising from 48.6% to 57.9%, and A&D reaching 65.5% (up from 60.9%). Notable gains include improvements in *Gallbladder* (DHC: 66.0% to 70.3%) and *Left Adrenal Gland* (A&D: 27.2% to 30.1%), demonstrating DyCE’s effectiveness as a plug-in loss across class-imbalanced datasets. Further experimental findings, qualitative and quantitative analysis are provided in the **supplementary material**.

## 5. Conclusion

In this work, we introduced SemiDAViL, a novel SSDA framework that leverages vision-language guidance and a dynamic loss formulation to address key challenges in domain-adaptive semantic segmentation. Through comprehensive experiments on several SSDA and SSL benchmarks, our method demonstrated consistent improvements over state-of-the-art techniques, especially in low-label regimes and class-imbalanced scenarios. The integration of vision-language pre-training, dense language embeddings, and the proposed DyCE loss contributes to discriminative feature extraction, better handling of minority classes, and enhanced semantic understanding. Overall, SemiDAViL sets a new benchmark in SSDA, showcasing strong generalizability across diverse domain shifts and label constraints.

## Acknowledgment

The authors greatly appreciate the financial support from the NSF project CMMI-2246673.

## References

- [1] Hritam Basak and Zhaozheng Yin. Quest for clone: Test-time domain adaptation for medical image segmentation by searching the closest clone in latent space. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 555–566. Springer, 2024. 2
- [2] Hritam Basak and Zhaozheng Yin. Forget more to learn more: Domain-specific feature unlearning for semi-supervised and unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 130–148. Springer, 2025. 2
- [3] Hritam Basak, Sagnik Ghosal, and Ram Sarkar. Addressing class imbalance in semi-supervised image segmentation: A study on cardiac mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 224–233. Springer, 2022. 8
- [4] Adriano Cardace, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Shallow features guide unsupervised domain adaptation for semantic segmentation at class boundaries. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1160–1170, 2022. 5, 7, 8
- [5] Baixu Chen, Junguang Jiang, Ximei Wang, Pengfei Wan, Jianmin Wang, and Mingsheng Long. Debaised self-training for semi-supervised learning. *Advances in Neural Information Processing Systems*, 35:32424–32437, 2022. 8
- [6] Hao Chen, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Marios Savvides, and Bhiksha Raj. An embarrassingly simple baseline for imbalanced semi-supervised learning. *arXiv preprint arXiv:2211.11086*, 2022. 8
- [7] Leiyu Chen, Shaobo Li, Qiang Bai, Jing Yang, Sanlong Jiang, and Yanming Miao. Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22):4712, 2021. 1
- [8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 6
- [9] Shuaijun Chen, Xu Jia, Jianzhong He, Yongjie Shi, and Jianzhuang Liu. Semi-supervised domain adaptation based on dual-level domain mixing for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11018–11027, 2021. 6
- [10] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2613–2622, 2021. 6
- [11] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2613–2622, 2021. 8
- [12] Ying Chen, Xu Ouyang, Kaiyue Zhu, and Gady Agam. Semi-supervised domain adaptation for semantic segmentation. *arXiv preprint arXiv:2110.10639*, 2021. 6
- [13] Ying Chen, Xu Ouyang, Kaiyue Zhu, and Gady Agam. Semi-supervised dual-domain adaptation for semantic segmentation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 230–237. IEEE, 2022. 6
- [14] 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*, pages 3213–3223, 2016. 5
- [15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3
- [16] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021. 4
- [17] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 2
- [18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 6
- [19] Özgür Er kent and Christian Laugier. Semantic segmentation with unsupervised domain adaptation under varying weather conditions for autonomous vehicles. *IEEE Robotics and Automation Letters*, 5(2):3580–3587, 2020. 1
- [20] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. A simple recipe for language-guided domain generalized segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23428–23437, 2024. 3
- [21] Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Dmt: Dynamic mutual training for semi-supervised learning. *Pattern Recognition*, 130:108777, 2022. 6
- [22] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. *BMVC*, 2020. 6
- [23] Weifu Fu, Qiang Nie, Jialin Li, Yuhuan Lin, Kai Wu, Jian Li, Yabiao Wang, Yong Liu, and Chengjie Wang. Iidm:

- Inter and intra-domain mixing for semi-supervised domain adaptation in semantic segmentation, 2024. 2, 6, 7
- [24] Yuan Gao, Zilei Wang, and Yixin Zhang. Delve into source and target collaboration in semi-supervised domain adaptation for semantic segmentation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 2, 6
- [25] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 2
- [26] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech 256, 2022. 3
- [27] Lan-Zhe Guo and Yu-Feng Li. Class-imbalanced semi-supervised learning with adaptive thresholding. In *International conference on machine learning*, pages 8082–8094. PMLR, 2022. 8
- [28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 5, 6
- [29] Lukas Hoyer, Dengxin Dai, and Luc 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*, pages 9924–9935, 2022. 3, 6
- [30] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *European conference on computer vision*, pages 372–391. Springer, 2022. 1
- [31] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11721–11732, 2023. 3
- [32] 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. 2, 5, 6
- [33] 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. 3, 4, 6
- [34] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hananeh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Open-clip, 2021. 2
- [35] Tarun Kalluri, Bodhisattwa Prasad Majumder, and Manmohan Chandraker. Tell, don’t show!: Language guidance eases transfer across domains in images and videos. *arXiv preprint arXiv:2403.05535*, 2024. 6
- [36] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. *Advances in neural information processing systems*, 33:3569–3580, 2020. 1
- [37] Young-Eun Kim, Yu-Won Lee, and Seong-Wan Lee. Lc-sm: Language-conditioned masked segmentation model for unsupervised domain adaptation. *Pattern Recognition*, 148:110201, 2024. 3
- [38] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL <http://www.cs.toronto.edu/kriz/cifar.html>, 5(4):1, 2010. 3
- [39] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 5
- [40] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2023. 2
- [41] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3918–3930, 2020. 2
- [42] Jichang Li, Guanbin Li, Yemin Shi, and Yizhou Yu. Cross-domain adaptive clustering for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2505–2514, 2021. 2
- [43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 6
- [44] Ruihuang Li, Shuai Li, Chenhang He, Yabin Zhang, Xu Jia, and Lei Zhang. Class-balanced pixel-level self-labeling for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11593–11603, 2022. 7, 8
- [45] Tianyu Li, Subhankar Roy, Huayi Zhou, Hongtao Lu, and Stéphane Lathuilière. Contrast, stylize and adapt: Unsupervised contrastive learning framework for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4869–4879, 2023. 6
- [46] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yanan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 2
- [47] T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017. 3
- [48] Yiqun Lin, Huifeng Yao, Zezhong Li, Guoyan Zheng, and Xiaomeng Li. Calibrating label distribution for class-imbalanced barely-supervised knee segmentation. In *International Conference on Medical Image Computing and*

- Computer-Assisted Intervention*, pages 109–118. Springer, 2022. 8
- [49] Chang Liu, Henghui Ding, Yulun Zhang, and Xudong Jiang. Multi-modal mutual attention and iterative interaction for referring image segmentation. *IEEE Transactions on Image Processing*, 32:3054–3065, 2023. 4
- [50] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096, 2022. 2
- [51] 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. 8
- [52] Jie Ma, Chuan Wang, Yang Liu, Liang Lin, and Guanbin Li. Enhanced soft label for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1185–1195, 2023. 2
- [53] Elham Amin Mansour, Ozan Unal, Suman Saha, Benjamin Bejar, and Luc Van Gool. Language-guided instance-aware domain-adaptive panoptic segmentation. *arXiv preprint arXiv:2404.03799*, 2024. 3
- [54] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 415–430. Springer, 2020. 7
- [55] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 8
- [56] Khan Muhammad, Tanveer Hussain, Hayat Ullah, Javier Del Ser, Mahdi Rezaei, Neeraj Kumar, Mohammad Hiji, Paolo Bellavista, and Victor Hugo C de Albuquerque. Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):22694–22715, 2022. 3
- [57] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1369–1378, 2021. 6
- [58] Can Qin, Yizhou Wang, and Yun Fu. Robust semi-supervised domain adaptation against noisy labels. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4409–4413, 2022. 2
- [59] Shoumeng Qiu, Xianhui Cheng, Hong Lu, Haiqiang Zhang, Ru Wan, Xiangyang Xue, and Jian Pu. Subclassified loss: Rethinking data imbalance from subclass perspective for semantic segmentation. *IEEE Transactions on Intelligent Vehicles*, 2023. 5
- [60] 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*, pages 8748–8763. PMLR, 2021. 2, 3, 6
- [61] Wenqi Ren, Yang Tang, Qiyu Sun, Chaoqiang Zhao, and Qing-Long Han. Visual semantic segmentation based on few/zero-shot learning: An overview. *IEEE/CAA Journal of Automatica Sinica*, 2023. 1
- [62] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 5
- [63] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 5
- [64] Manisha Saini and Seba Susan. Tackling class imbalance in computer vision: a contemporary review. *Artificial Intelligence Review*, 56(Suppl 1):1279–1335, 2023. 3
- [65] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8050–8058, 2019. 2
- [66] Manuel Schwonberg, Joshua Niemeijer, Jan-Aike Termöhlen, Nico M Schmidt, Hanno Gottschalk, Tim Fingscheidt, et al. Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving. *IEEE Access*, 11:54296–54336, 2023. 1
- [67] Fengyi Shen, Akhil Gurram, Ziyuan Liu, He Wang, and Alois Knoll. Diga: Distil to generalize and then adapt for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15866–15877, 2023. 6
- [68] Wei Shen, Zelin Peng, Xuehui Wang, Huayu Wang, Jiazhong Cen, Dongsheng Jiang, Lingxi Xie, Xiaokang Yang, and Qi Tian. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 45(8):9284–9305, 2023. 1, 3
- [69] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 3
- [70] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 4

- [71] Hans Thisanke, Chamli Deshan, Kavindu Chamith, Sachith Seneviratne, Rajith Vidanaarachchi, and Damayanthi Herath. Semantic segmentation using vision transformers: A survey. *Engineering Applications of Artificial Intelligence*, 126:106669, 2023. 1
- [72] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1379–1389, 2021. 6, 7, 8
- [73] Thanh-Dat Truong, Ngan Le, Bhiksha Raj, Jackson Cothren, and Khoa Luu. Freedom: Fairness domain adaptation approach to semantic scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19988–19997, 2023. 5
- [74] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 6
- [75] Haonan Wang and Xiaomeng Li. Dhc: Dual-debiased heterogeneous co-training framework for class-imbalanced semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 582–591. Springer, 2023. 8
- [76] Haonan Wang and Xiaomeng Li. Towards generic semi-supervised framework for volumetric medical image segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 6, 8
- [77] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *European conference on computer vision*, pages 642–659. Springer, 2020. 7
- [78] Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. Debiased learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14647–14657, 2022. 8
- [79] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4248–4257, 2022. 7
- [80] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10857–10866, 2021. 8
- [81] Lu Wen, Yuanyuan Xu, Zhenghao Feng, Jiliu Zhou, Luping Zhou, and Yan Wang. Semi-supervised domain adaptation for semantic segmentation via active learning with feature- and semantic-level alignments. *IEEE Transactions on Intelligent Vehicles*, 2024. 6, 7
- [82] Wenbo Qi, Jiafei Wu and SC Chan. Gradient-aware for class-imbalanced semi-supervised medical image segmentation. *ECCV*, 98:1–86, 2024. 3
- [83] Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 34–43. Springer, 2022. 8
- [84] Yao Wu, Mingwei Xing, Yachao Zhang, Yuan Xie, and Yanyun Qu. Clip2uda: Making frozen clip reward unsupervised domain adaptation in 3d semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8662–8671, 2024. 3
- [85] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Bridging category-level and instance-level semantic image segmentation. *arXiv preprint arXiv:1605.06885*, 2016. 5
- [86] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8068–8078, 2022. 6
- [87] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 2
- [88] Hongliang Yan, Zhetao Li, Qilong Wang, Peihua Li, Yong Xu, and Wangmeng Zuo. Weighted and class-specific maximum mean discrepancy for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 22(9):2420–2433, 2019. 3
- [89] Weihao Yan, Yeqiang Qian, Yueyuan Li, Tao Li, Chunxiang Wang, and Ming Yang. Ss-ada: A semi-supervised active domain adaptation framework for semantic segmentation. *arXiv preprint arXiv:2407.12788*, 2024. 2, 7, 8
- [90] Jinyu Yang, Weizhi An, Chaochao Yan, Peilin Zhao, and Junzhou Huang. Context-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 514–524, 2021. 7
- [91] Luyu Yang, Yan Wang, Mingfei Gao, Abhinav Shrivastava, Kilian Q Weinberger, Wei-Lun Chao, and Ser-Nam Lim. Deep co-training with task decomposition for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8906–8916, 2021. 2, 6
- [92] 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. 2, 6, 7
- [93] 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. 8
- [94] Lei Yu, Wanqi Yang, Shengqi Huang, Lei Wang, and Ming Yang. High-level semantic feature matters few-shot unsu-

- pervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11025–11033, 2023. [2](#)
- [95] Yu-Chu Yu and Hsuan-Tien Lin. Semi-supervised domain adaptation with source label adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24100–24109, 2023. [6](#)
- [96] Jing Zhang, Yingshuai Xie, Weichao Ding, and Zhe Wang. Cross on cross attention: Deep fusion transformer for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):4257–4268, 2023. [4](#)
- [97] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang 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*, pages 12414–12424, 2021. [6](#)
- [98] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. [2](#), [6](#)
- [99] Jinghao Zhou, Li Dong, Zhe Gan, Lijuan Wang, and Furu Wei. Non-contrastive learning meets language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11028–11038, 2023. [2](#)
- [100] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023. [2](#)
- [101] Zheng Zhou, Change Zheng, Xiaodong Liu, Ye Tian, Xiaoyi Chen, Xuexue Chen, and Zixun Dong. A dynamic effective class balanced approach for remote sensing imagery semantic segmentation of imbalanced data. *Remote Sensing*, 15(7):1768, 2023. [3](#)
- [102] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. [3](#)
- [103] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. [1](#)