Segmenting Visuals With Querying Words: Language Anchors For Semi-Supervised Image Segmentation

Numair Nadeem[1], Saeed Anwar[2], Muhammad Hamza Asad[3], and Abdul Bais[1]

[1] University of Regina, Canada
nng794@uregina.ca, abdul.bais@uregina.ca
[2] The University of Western Australia, Australia
saeed.anwar@uwa.edu.au
[3] University Canada West, Canada
muhammadhamza.asad@ucanwest.ca

Abstract. Vision–Language Models (VLMs) provide rich semantic priors but are underexplored in Semi-supervised Semantic Segmentation. Recent attempts to integrate VLMs to inject high-level semantics overlook the semantic misalignment between visual and textual representations that arises from using domain-invariant text embeddings without adapting them to dataset- and image-specific contexts. This lack of domain awareness, coupled with limited annotations, weakens the model's semantic understanding by preventing effective vision-language alignment. As a result, the model struggles with contextual reasoning, shows weak intra-class discrimination, and confuses similar classes. To address these challenges, we propose the Hierarchical Vision–Language Transformer (HVLFormer), which achieves domain-aware and domain-robust alignment between visual and textual representations within a query-driven mask-transformer architecture. Firstly, we transform text embeddings from a pre-trained VLM into multi-scale, dataset-aware textual object queries that capture class semantics from coarse to fine granularity and enhance the model's semantic understanding. Next, these queries are refined using image-specific visual context, aligning global textual semantics with local scene structures and improving class discrimination. Finally, to achieve domain-robustness, we introduce cross-view and modal consistency regularization, which enforces prediction consistency within the mask-transformer architecture across augmented views. It ensures that language queries remain robust to intra-class variations and diverse visual scenes without overfitting to the small labeled set while maintaining stable vision–language alignment under perturbations during decoding. With limited training data, HVLFormer outperforms state-of-the-art methods on public benchmarks. Project page: numnz.github.io/HVLFormer.

Keywords: Semi-Supervised Learning · Semantic Segmentation · Vision Language Models · Query Driven Mask Transformer · Multi-Modal Learning

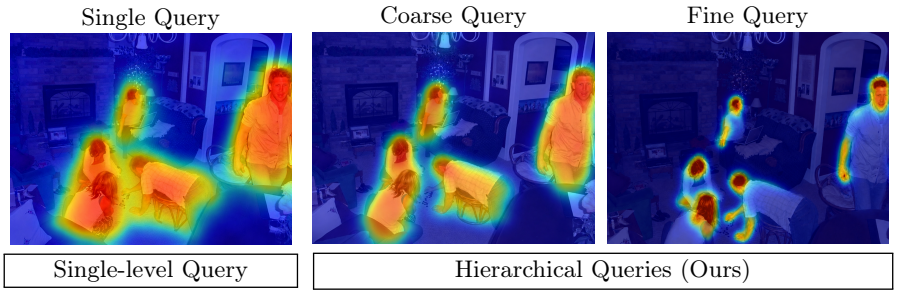

Fig. 1: Visualization of single-level and hierarchical "person" text queries on the same image. The single-level query fails to capture the human structure precisely, whereas the proposed multi-scale queries provide complementary semantic cues: the coarse query localizes the overall body structure precisely, while the fine query highlights boundaries and local body details. This demonstrates the benefit of hierarchical textual queries for precise class-region alignment under limited supervision.

1 Introduction

Developing semantic segmentation models that require minimal labeled data remains a significant challenge in computer vision, particularly in autonomous driving, agriculture, and medical imaging, where acquiring dense pixel-level annotations is expensive and time-consuming. Semi-supervised Semantic Segmentation (SSS) addresses this challenge by leveraging a small subset of labeled images alongside a large pool of unlabeled data, aiming to achieve strong performance without full supervision [5, 24, 25, 45, 49]. Typical strategies include adversarial training [23], and self-training [31, 42, 44], which aim to extract supervisory signals from unlabeled data.

Despite recent advances in achieving precise segmentation [13,43], SSS models still frequently misclassify, particularly in scenarios with significant intra-class variation, visually similar classes, and rare classes. For instance, confusing sofas with chairs often arises from the limited representation of rare categories within the small labeled subset [50]. Unlike unsupervised domain adaptation [26,27,34] and domain generalization [16], which benefit from fully labeled source data to establish a strong semantic understanding, SSS lacks sufficient labeled examples, resulting in weak feature learning and limited semantic understanding. Recent advances in VLMs [11,11,21] offer a promising direction, as they encode domain-invariant semantic representations that align visual and textual modalities within a shared embedding space [1]. Motivated by this, recent works integrated VLM-derived text embeddings to guide pixel-level predictions. However, their role in SSS remains underexplored, as most VLM-based segmentation methods either operate without dense annotations [2, 39, 52], resulting in limited segmentation precision, or rely on large-scale labeled datasets [7, 40, 54], which are expensive. SemiVL [13] partially bridges this gap by adapting CLIP [29] text embeddings through spatial fine-tuning and a language-aware decoder. In parallel, Pseudo-

SD [47] leverages pseudo-labels and text embeddings to fine-tune stable diffusion models for generating labeled synthetic images.

The approaches mentioned above pose two key issues: insufficient domain-awareness in text embeddings and their reliance on pixel-centric decoders, where language remains a weak auxiliary signal. While they emphasize domain invariance in text embeddings to provide global class semantics, they overlook domain awareness. We identify this as a key bottleneck in low-data regimes, where dataset- and image-specific cues are necessary to compensate for the model's weak semantic understanding. Since pre-trained VLMs are trained on generic web-scale corpora, their embeddings encode broad semantics but often miss the fine contextual cues required for precise feature learning. For example, *"chair"* and *"sofa"* may share overlapping meanings in generic corpora. However, within the ADE20K [51] dataset, they differ in shape and spatial context: chairs often appear around tables, whereas sofas are typically located in living areas. This issue is further amplified by using only one text embedding per class and by activating all class embeddings for every image. Under limited annotations, a single text embedding cannot capture how the same class may appear across different scales, textures, and contexts, leading to misclassifications and reduced performance on intra-class variation. Moreover, existing methods use text embeddings from all dataset classes to guide the segmentation process, regardless of whether those classes are present in the image. Such image-irrelevant text embeddings introduce noisy guidance. For instance, if no "bus" appears in an image, its embedding should not participate in mask generation, as it may still attend to bus-like visual structures and lead to misclassification. Overall, without adapting text embeddings to dataset- and image-specific semantics, the model receives generic and irrelevant guidance, resulting in blurred decision boundaries, confusion between correlated classes, and unstable representations, especially for rare classes. Beyond lacking domain awareness, these methods also fall short in how they integrate language with vision. Specifically, they rely on pixel-centric decoders that are primarily designed to convert visual features into mask predictions. Although language embeddings are incorporated through fusion mechanisms, the mask-generation process remains dominated by per-pixel vision features. We argue that this bias towards visual features keeps language as a weak auxiliary signal rather than an active semantic guide, leading to shallow vision–language alignment and limited extraction of semantics from VLMs.

To address these limitations, we propose the Hierarchical Vision–Language transFormer (HVLFormer), a unified mask-transformer-based framework in which language queries explicitly predict class-specific segmentation masks under limited supervision. By using language as an active semantic guide through the mask-decoder, HVLFormer reduces the model's bias towards visual features, which lack strong semantic discrimination when trained with limited pixel-level annotations. Specifically, HVLFormer uses domain-invariant text embeddings from pretrained VLMs [29] as object queries and makes them domain-aware through dataset- and image-specific contextual conditioning. This enables more

effective extraction of knowledge from pretrained VLMs and strengthens multi-modal alignment under limited supervision.

For this purpose, HVLFormer first introduces the *Hierarchical Textual Query Generation (HTQG)* module, which addresses the limitations of generic, single-class text embeddings. HTQG enriches class names with visual-semantic dataset attributes to generate dataset-aware textual embeddings; for example, in an urban rainy dataset, we initialize "car" embedding as "'a photo of a car on a wet road in rain within an urban scene'" instead of "'a photo of a car'", encouraging the pretrained VLM to focus on the intended dataset-specific class concept. It then transforms each class embedding into multi-scale textual queries that capture class semantics from coarse to fine granularity, improving robustness to intra-class variation and reducing confusion among rare or visually similar classes. Fig. 1 illustrates this effect: compared with a single-level "person" query, which imprecisely highlights person regions along with background noise, the proposed multi-scale queries provide complementary semantics. The coarse query captures the overall body structure, while the fine query highlights local details. HTQG also estimates the presence probability of each class in the input image and weights its corresponding text queries accordingly, suppressing semantic guidance from absent-class queries during mask prediction.

Next, the *Pixel-Text Refinement Module (PTRM)* injects image context from the pixel-decoder visual features into these textual queries, aligning textual semantics with local scene structure and making them more discriminative and spatially aware. The refined queries then interact with pixel features in the mask-decoder to form coherent semantic groups and predict segmentation masks. Finally, to compensate for the limited supervision available for learning stable pixel-text correspondences, we introduce *Cross-View and Modal Consistency Regularization (CMCR)*. It extracts supervisory signals from unlabeled data by enforcing prediction and pixel-text consistency across augmented views at every decoder layer. These signals provide stronger optimization guidance for HTQG and PTRM, helping the model learn reliable query generation and pixel-text refinement under limited labels. Since even small pixel-text misalignments can be amplified through iterative decoding, layer-wise consistency stabilizes overall optimization, reduces error propagation, and improves domain robustness. Our main contributions are summarized as follows:

- We introduce a language-driven SSS framework that explicitly leverages text embeddings from pre-trained VLMs as object queries and refines them with dataset- and image-specific context. By adapting global semantic priors to local visual distributions, the proposed domain-aware textual queries address weak feature learning and semantic understanding under limited supervision, while multimodal regularization ensures stable vision-language alignment during training.
- We develop a unified hierarchical vision–language architecture that enables rich interactions between domain-aware textual queries and multi-scale pixel features. Through progressive vision–language alignment, the framework bridges global linguistic semantics with dataset-specific visual representations, allow-

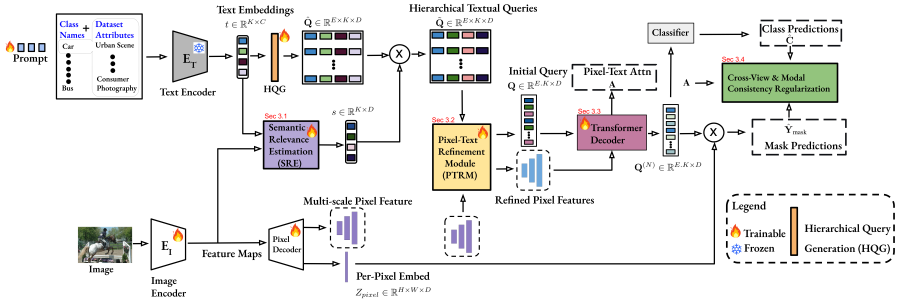

Fig. 2: Overview of HVLFormer: The pre-trained VLM encodes visual and textual features, generating dataset-aware text embeddings from class names and dataset attributes. The **HQG** module transforms them into multi-scale queries, with **SRE** weighing queries based on the presence of their corresponding classes in an image, refined by **PTRM** through pixel–text alignment. These queries guide the transformer decoder to predict class-specific masks, with each query responsible for generating a segmentation mask of its represented class, while **CMCR** enforces cross-modal consistency for stable vision–language alignment under limited supervision.

ing textual queries to inject rich semantic knowledge into the model. Reinforced by cross-view and modal consistency regularization, this process leads to stronger class discrimination, improved contextual stability, and superior label efficiency in low-annotation regimes.

2 Related Works

2.1 Semi-supervised Semantic Segmentation (SSS)

Early methods adopted adversarial frameworks [3,37] to extract supervisory signals from unlabeled data. However, these approaches often suffered from instability [43]. Subsequent work simplified the paradigm through entropy minimization [55] and consistency regularization [12, 41], which have become the foundation of modern approaches. Utilizing them, self-training strategies [13] generate pseudo-labels for unlabeled data and refine them iteratively while enforcing prediction consistency under augmentations such as CutOut and CutMix [46]. Similarly, cross pseudo supervision [3] further improved stability through co-training, and adaptive equalization Learning [14] biased CutMix towards under-represented classes. To address poor early-stage pseudo-label quality, ST++ [44] introduced staged pseudo-label augmentation to enable progressive refinement. Recently, methods like UniMatch [42] have adopted confidence-based filtering [43, 48], inspired by FixMatch [30], in which weak–strong augmentation pairs enforce consistency. These designs influenced vision–language extensions such as SemiVL [13], which incorporates CLIP [29] text embeddings as semantic regularizers. However, SemiVL [13] treats textual cues as single-level, domain-invariant embeddings, limiting their capacity to capture intra-class diversity and fine-grained semantics. In contrast, HVLFormer enhances the textual embeddings by

converting them into domain-aware, multi-scale per-class queries that effectively adapt to the local visual context. These refined queries serve as dynamic semantic anchors, improving same-class pixel grouping and multimodal alignment, with the CMCR enforcing stable multimodal alignment under sparse supervision.

2.2 Language-Driven Segmentation Frameworks

VLMs use large-scale image–text pretraining to construct a shared embedding space between visual and textual modalities, enabling generalization across diverse tasks such as image classification, retrieval, and open-vocabulary recognition. Extending the mentioned semantic segmentation paradigm, preliminary approaches such as MaskCLIP [8] and GroupViT [39] attempted to learn segmentation solely from image–caption pairs, without pixel-level annotations, and demonstrated that segmentation cues can emerge from contrastively trained encoders; however, their reliance on image–caption supervision produced coarse as well as noisy masks. To improve mask quality, ZegCLIP [54] aligns dense visual features with textual embeddings via learned decoders and attention mechanisms. Meanwhile, parameter-efficient tuning approaches, such as prompt tuning [15, 53], improve generalization with minimal supervision. However, both contrastive pretraining and prompt-tuning strategies remain limited by their reliance on coarse global semantics, lacking the fine-grained, image-specific alignment needed for precise dense prediction.

Lately, query-based transformer architectures have achieved strong performance in dense prediction. Mask2Former [4] aggregates multi-scale visual features from the vision encoder into spatially rich pixel features through the pixel decoder. Randomly initialized object queries interact with these features through masked cross-attention in a transformer decoder to group pixels of the same class. Building on this, TQDM [28] introduces a textual-query-driven framework for domain-generalized segmentation, utilizing CLIP-based text embeddings to initialize object queries and enhance the semantic clarity of pixel features by computing text-to-pixel attention before the transformer decoder. However, the one-way text-to-pixel attention restricts feedback from the visual to the textual space, preventing queries from adapting to image-specific variations before being fed into the transformer decoder. TQDM's focus on domain invariance is effective for domain generalization but ignores domain awareness—an essential factor in SSS. As a result, TQDM [28] adapts poorly to queries under SSS, widening the modality gap because generic textual semantics fail to anchor weak visual features, leading to unstable feature correspondence.

Moreover, the quality of these initial queries to the decoder is critical, as they determine how effectively the decoder groups pixels and propagates semantics across layers. Generic, domain-invariant queries such as *"bus"* and *"train"*—which share a broad "transport" concept but differ contextually in cityscapes [6]—lack the fine-grained cues required for accurate segmentation under limited training examples. In contrast, domain-aware queries provide stronger semantic grounding, enabling clearer and more stable pixel grouping during decoding. HVL-Former addresses this by introducing domain-awareness during text embedding

generation through dataset-specific prompting and by injecting visual context from pixel features into these queries. However, without explicit regularization, the scarcity of labeled data limits supervision signals, causing vision–language misalignments to amplify across decoder layers and degrade semantic consistency. To counter this, we enforce layer-wise consistency regularization across augmented views. This encourages textual queries and visual features to remain aligned throughout decoding; when their predictions or pixel-text correspondences become inconsistent across views, the consistency loss penalizes the mismatch and guides the model toward more stable multimodal alignment.

3 HVLFormer

HVLFormer, a vision-language segmentation framework, is built on a mask-classification backbone [4], with the pre-trained VLM's image encoder $\mathbf{E}_I$ that is fine-tuned on labeled images using standard cross-entropy loss and a text encoder $\mathbf{E}_T$ that remains frozen to preserve the pre-trained VLM's semantic space. For an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, $\mathbf{E}_I$ produces multi-scale feature maps that are fed to the HVLFormer pixel decoder, providing multi-scale pixel-features $z : \{z^e\}_{e=1}^{E}$ and dense per-pixel embeddings $\mathbf{Z}_{pixel} \in \mathbb{R}^{H\times W\times D}$. The textual queries from text embeddings, along with pixel features, are input to a mask-transformer decoder to generate segmentation masks.

To effectively align the pixel features with language semantics (textual queries), HVLFormer introduces a series of vision–language interaction modules. The HTQG module generates multi-scale, dataset-aware per-class textual queries, which are enriched by image-context in PTRM through cross-modal spatial interactions, and then produces segmentation masks after being processed by the transformer decoder [4]. Finally, CMCR enables domain-robust vision-language alignment through consistency regularization, thus forcing the model to effectively enrich textual queries with diverse image contexts. The overall pipeline is illustrated in Fig. 2.

3.1 Hierarchical Textual Query Generation (HTQG)

Learnable prompting architectures have demonstrated a strong ability to generate semantically coherent prompts [13]. Building on this capability, we employ a learnable dataset-specific prompt p_k to produce attribute-rich textual descriptions for each semantic class. The prompt p_k takes the *dataset attributes* with *class names* as inputs, ensuring that the generated text captions for each class capture scene-specific class semantics relevant to the dataset. This design is important because the visual appearance of a class can vary significantly across datasets. For example, in urban driving scenes (Cityscapes [6]), classes such as "car" and "road" are typically observed from a street-level perspective with strong geometric regularities, whereas in object-centric images (Pascal VOC [10]), the same classes appear under diverse backgrounds, scales, and viewpoints. By incorporating dataset attributes, the text encoder E_T generates dataset-aware textual

embeddings that better align with the visual distributions of each dataset. Details of the attributes used for each dataset are provided in the supplementary material.

For each class k, dataset aware textual prompt p_k is encoded using the frozen text encoder E_T, i.e. $\mathbf{t}_k = E_T(p_k) \in \mathbb{R}^C$.

However, in dense segmentation, single-level text embeddings fail to capture the full range of intra-class variations, an issue further amplified by limited labeled data. They struggle to represent both coarse structure and fine local details. This is particularly problematic for visually similar classes, where the key differences often lie in subtle details, leading to misclassification. In addition, applying all class-level queries to every image introduces noise from absent class queries into the pixel features. For example, in an indoor scene containing a "table" but no "chair", the chair query may respond to table legs or object boundaries because they share similar part-level structures. Under limited labels, the model lacks sufficient supervision to reliably learn subtle differences, causing irrelevant queries to weaken class-specific feature learning and produce imprecise masks.

To overcome these limitations, HTQG jointly performs Hierarchical Query Generation (HQG) and Semantic Relevance Estimation (SRE). HQG projects each class embedding $\mathbf{t}_k$ into multiple abstraction levels—from coarse to fine—to model semantics across scales and improve discrimination between similar or rare classes. Specifically, $\mathbf{t}_k$ is transformed by E distinct MLP heads, each aligned with e^{th} pixel feature $\{z^e\}_{e=1}^{E}$: $\hat{\mathbf{Q}} = \{\mathbf{q}_k^e = \mathrm{MLP}_e(\mathbf{t}_k)\}_{k=1,e=1}^{K,E} \in \mathbb{R}^{E\times K\times D}$. Following the pixel decoder's coarse-to-fine hierarchy, the heads generate queries at increasing abstraction levels of the same class rather than separate subclasses. Coarse queries encode global object structure, while finer queries capture texture and boundary details. During training, each query is refined only through its aligned feature map in PTRM, ensuring a coarse-to-fine hierarchy in queries. Furthermore, to prevent redundancy, a diversity regularization term: $\ell_{\text{div}} = \sum_{e\neq e'} \left\| \frac{(\mathbf{q}_k^e)^\top \mathbf{q}_k^{'e}}{\|\mathbf{q}_k^e\|\cdot\|\mathbf{q}_k^{'e}\|} \right\|_2^2$, encourages distinct, complementary semantics across scales.

Simultaneously, SRE estimates the likelihood of each class being present in the image, thereby suppressing irrelevant queries. For each class, a relevance score s_k is computed by projecting multi-scale pixel features from the pixel decoder into a shared space with text embeddings via a lightweight adapter as: $s_k = \sigma\left(\mathrm{MLP}\left([\phi(f_1, f_2, f_3), \mathbf{t}_k]\right)\right)$, where $\phi(f_1, f_2, f_3)$ represents fused multi-scale features from E_I, and $\sigma(\cdot)$ is a sigmoid activation. The MLP is trained using labeled images and a supervised loss separately. Final refined hierarchical queries are then obtained as $\tilde{\mathbf{Q}} = \{\tilde{\mathbf{q}}_k^e = s_k \cdot \mathbf{q}_k^e\}_{k=1,e=1}^{K,E}$. This yields dataset-context-aware, scale-adaptive textual queries that suppress irrelevant class semantics and encode dataset-specific features, improving vision-text alignment and ensuring robust semantic understanding under limited supervision.

3.2 Pixel-Text Refinement Module (PTRM)

PTRM injects image-specific context into hierarchical textual queries by integrating visual cues from pixel features z, such as structure, texture, and illumination, thereby enabling queries to capture scene-dependent variations. Simultaneously, class-level semantics from text enhance semantic coherence in pixel features. This bidirectional adaptation aligns textual and visual representations, enabling queries to dynamically respond to image context and yield spatially coherent, semantically consistent masks. Unlike conventional cross-attention, which relies solely on token-level similarity, PTRM introduces spatially guided conditioning that allows the image to determine where and how strongly each query should influence prediction. Consequently, each query is spatially distributed over the feature map and selectively strengthened in regions where the visual patterns match its class semantics while being suppressed in visually irrelevant regions.

At the pixel feature level e, textual queries $\tilde{\mathbf{Q}}^e : \{\tilde{\mathbf{q}}_k^e\}_{k=1}^K$, and the pixel feature map z^e are projected into a shared latent space to ensure semantic–spatial comparability, resulting in text T^e and visual V^e representations. A shallow fusion integrates semantic and spatial priors: $F^e = T^e + V^e$. This coupling creates a coarse alignment between class semantics and image-spatial structure, allowing textual information to roughly map onto visual layout cues.

To localize cross-modal importance, we compute average-pooled (ς) representations across all feature maps:

$$X^e = [\varsigma(T^e), \varsigma(V^e), \varsigma(F^e)]. \tag{1}$$

It highlights regions that are globally informative in either modality or their fusion. Two convolutional layers with ReLU and sigmoid activations applied to X^e then generate spatial attention weights W^e, which are decomposed into three attention maps, i.e., text-guided W_T^e, pixel-guided W_V^e, and fused attention W_F^e. These maps jointly regulate the bidirectional adaptation between modalities:

$$T^{'e} = T^e \odot W_V^e \odot W_F^e, \tag{2}$$

$$V^{'e} = V^e \odot W_T^e \odot W_F^e. \tag{3}$$

Here, W_V^e injects local visual cues such as texture, structure, and illumination into textual features, enriching them with image-conditioned semantics. Conversely, W_T^e transfers class-level priors from text to pixel features, thereby injecting semantic understanding into visual representations. Finally, W_F^e acts as a consensus gate, emphasizing regions of mutual confidence to ensure updates occur only where textual and visual information agree.

The final aligned textual queries and enriched pixel features are obtained by applying global average pooling to the spatially distributed text feature map $T^{'e} \in \mathbb{R}^{J \times H^e \times W^e}$, aligning them into queries $\mathbf{Q}^e \in \mathbb{R}^{K \times D}$, and using 1x1 convolution to reproject refined image features $V^{'e}$ from the shared latent space back to the decoder's original feature domain $z_{\text{out}}^e \in \mathbb{R}^{H^e \times W^e \times D}$, ensuring compatibility with subsequent decoding layers.

For all $'e'$, PTRM thus outputs enhanced pixel features z_{out} enriched with class semantics and refined, context-aware textual queries $\mathbf{Q} = \{\mathbf{Q}^e\}_{e=1}^{E}$, which are subsequently passed to the transformer decoder for final segmentation.

3.3 Transformer Decoder

Finally, the transformer decoder [4] iteratively refines the initially aligned textual queries with z_{out} through N masked attention layers and outputs refined queries $\mathbf{Q}^{(N)}$. These are projected into mask embeddings $\mathbf{M_E}$ to generate pixel-level masks with the per-pixel embeddings $\mathbf{Z}_{\text{pixel}}$, expressed as $\hat{\mathbf{Y}}_{\text{mask}} = \mathbf{M_E} \cdot \mathbf{Z}_{\text{pixel}}$. Each query embedding from $\mathbf{Q}^{(N)}$ is classified through a linear projection to produce class logits $\hat{\mathbf{C}}$. The final segmentation prediction is obtained by combining the pixel-level mask and the class scores as $\hat{\mathbf{Y}} = \hat{\mathbf{C}}^{\top} \cdot \hat{\mathbf{Y}}_{\text{mask}}$, where $\hat{\mathbf{C}}^{\top}$ denotes the transpose of $\hat{C}$.

3.4 Cross-View and Modal Consistency Regularisation (CMCR)

Limited labeled data restricts the model's exposure to diverse visual conditions, causing overfitting to the annotated subset and weakening its ability to generalize across varying image distributions. Such insufficient supervision also limits the model's ability to learn reliable vision–text mappings. In mask-transformer architectures, this insufficient supervision can amplify early misalignments between pixel features and textual queries during iterative decoding [28], leading to cumulative errors and reduced domain robustness.

To overcome this, we designed the CMCR, which extracts supervisory signals from unlabeled data and reinforces domain robustness. CMCR enforces prediction consistency across multiple augmented views of the same image, compelling the model to maintain stable vision–language alignment under appearance variations. Applied at every decoder layer, it prevents error accumulation and enables textual queries to adapt effectively to diverse image contexts under limited supervision, thus enhancing the model's semantic understanding. We adopt a three-view augmentation strategy comprising the original image I, a weakly augmented version I_w (e.g., Gaussian blur, slight brightness, or contrast changes), and a strongly augmented version I_s (e.g., Cutout, color jitter, and geometric distortions) following UniMatch [42]. The original and weak views preserve intrinsic domain characteristics—such as color distribution, structure, and contextual layout—while the strong view introduces appearance diversity. CMCR enforces consistency across all three views, i.e., original–weak and original–strong pairs, allowing the model to anchor on domain-stable cues while learning invariance to superficial perturbations. This balance ensures domain robustness without eroding domain awareness.

Given an unlabeled image I and its augmented counterparts (I_w, I_s), the transformer decoder at each stage $n \in \{1, \ldots, N\}$ produces a mask prediction $(\hat{\mathbf{Y}}^n, \hat{\mathbf{Y}}^n_w, \hat{\mathbf{Y}}^n_s)$, class predictions by each query $i \in \{1, \ldots, K.E\} : (\hat{\mathbf{C}}^n_i, \hat{\mathbf{C}}^n_{w,i}, \hat{\mathbf{C}}^n_{s,i})$, and pixel-text attention maps $(\mathbf{A}^n_i, \mathbf{A}^n_{w,i}, \mathbf{A}^n_{s,i})$. The CMCR objective enforces

Table 1: Comparison of HVLFormer with State-Of-The-Art (SOTA) SSS methods on Pascal VOC. The mIoU (%) is reported across varying labeled splits. The best results are shown in red, the previous SOTA in blue, and **bold** denotes variants of our method. †: reproduced by [13].

Methods	Backbones	1/115 92	1/58 183	1/29 366	1/14 732	1/7 1464	#Params
PseudoSeg [55]	R101	57.6	65.5	69.1	72.4	–	59.5
CPS [3]	R101	64.1	67.4	71.7	75.9	–	59.5
ST++ [44]	R101	65.2	71.0	74.6	77.3	79.1	59.5
U^2PL [36]	R101	68.0	69.2	73.7	76.2	79.5	59.5
PCR [38]	R101	70.1	74.7	77.2	78.5	80.7	59.5
ESL [22]	R101	71.0	74.0	78.1	79.5	81.8	59.5
LogicDiag [19]	R101	73.3	76.7	77.9	79.4	–	59.5
UniMatch V1 [42]	R101	75.2	77.2	78.9	79.9	81.2	59.5
3-CPS [18]	R101	75.7	77.7	80.1	80.9	82.0	59.5
ZegCLIP† [54]	ViT-B	69.3	74.2	78.7	81.0	82.0	88.0
ZegCLIP+Unimatch V1†	ViT-B	78.0	80.3	80.9	82.8	83.6	88.0
UniMatch V1† [42]	ViT-B	77.9	80.1	82.0	83.3	84.0	88.0
Pseudo-SD [47]	ViT-B	78.9	79.2	80.7	81.9	82.7	88.0
SemiVL [13]	ViT-B	84.0	85.6	86.0	86.7	87.3	88.0
Ours	ViT-B	**87.3**	**88.8**	**89.4**	**90.0**	**90.2**	107.0
Ours	EVA02-S	**84.8**	**86.2**	**87.1**	**88.5**	**89.3**	40.0
UniMatch V2 [43]	DINOv2-S	79.0	85.5	85.9	86.7	87.8	24.8
POS [32]	DINOv2-S	80.7	86.8	87.2	87.5	88.3	24.8
UniMatch V2 [43]	DINOv2-B	86.3	87.9	88.9	90.0	90.8	98.0
TQDM (baseline) [28]	SigLIP2	76.1	79.5	82.5	86.1	88.6	107.0
Ours	SigLIP2	**89.1**	**90.4**	**91.0**	**91.6**	**91.7**	109.0

prediction and pixel-text consistency across these augmented views for each query:

$$\ell_{\text{CMCR}} = \sum_{n=1}^{N}[\,\ell_{\text{mask}}(\hat{\mathbf{Y}}^n, \hat{\mathbf{Y}}_w^n, \hat{\mathbf{Y}}_s^n) + \sum_{i=1}^{K}(\,\ell_{\text{class}}(\hat{\mathbf{C}}_i^n, \hat{\mathbf{C}}_{w,i}^n, \hat{\mathbf{C}}_{s,i}^n) + \ell_{\text{align}}(\mathbf{A}_i^n, \mathbf{A}_{w,i}^n, \mathbf{A}_{s,i}^n))]. \tag{4}$$

Here, ℓ_{mask} enforces global segmentation consistency across augmented views, while ℓ_{class} and ℓ_{align} regularize query-specific class predictions and pixel–text correspondences. The details of these losses are provided in the supplementary. Our overall training objective combines segmentation(ℓ_{seg}), and regularization (ℓ_{reg}) from TQDM [28], along with diversity and CMCR regularization, expressed as

$$\ell_{\text{total}} = \ell_{\text{seg}} + \lambda_{\text{reg}}\ell_{\text{reg}} + \lambda_{\text{div}}\ell_{\text{div}} + \lambda_{\text{cmcr}}\ell_{\text{cmcr}}, \tag{5}$$

where $\{\lambda_{\text{reg}}, \lambda_{\text{div}}, \lambda_{\text{cmcr}}\}$ are weighting factors.

4 Experiments

4.1 Implementation Details

Network Architecture. Our model's vision encoder E_I [9] and text encoder E_T [35] are initialized using one of three backbones: CLIP [29], SigLip2 [33], or an EVA02-CLIP [11] with patch sizes of 16 or 14, respectively. We employ an 8-token learnable prompt p_k to generate attribute-rich textual descriptions per dataset-class. The hierarchical queries in HTQG are generated with $E = 3$ MLP heads; hence, there are three queries per class, strictly aligned with three feature maps from the pixel-decoder. The multi-scale pixel decoder is adopted from DN-DETR [17], with $M = 6$ layers. The transformer decoder follows

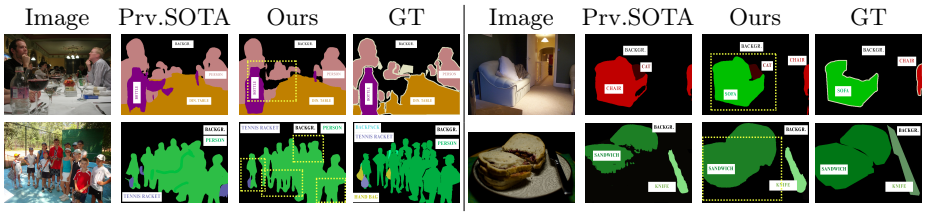

Fig. 3: The Pascal VOC with 92 labels (top) and COCO with 232 labels (bottom). HVLFormer correctly disambiguates visually similar classes, suppresses over-segmentation in cluttered regions, and accurately identifies multiple persons in dense crowds, demonstrating enhanced semantic understanding in low-label regime.

Mask2Former [4] with $N = 9$ masked attention layers. We adopt TQDM [28] as the baseline. We utilize mean Intersection over Union (mIoU) as a metric, while results are averaged over three random data sampling seeds.

Training. We benchmark on Pascal VOC [10], COCO [20], ADE20K [51], and Cityscapes [6] using two 24 GB NVIDIA GeForce RTX-3090 GPUs. Following SemiVL [13], each iteration processes a mixed batch of eight labeled and eight unlabeled images. Inputs are randomly cropped to 512×512 pixels, except for Cityscapes, where we use 801×801. The model is trained for 80, 10, 40, and 240 epochs on VOC, COCO, ADE20K, and Cityscapes, respectively, using AdamW [7]. The initial learning rates are set to 10^{-4} with a polynomial decay of 0.9. Linear warm-up [25] is applied for $t_{\text{warm}} = 1.5$k iterations. The weighting factors for each objective are set to $\{\lambda_{\text{reg}}, \lambda_{\text{div}}, \lambda_{\text{cmcr}}\} = \{1, 1, 5\}$, with a detailed hyperparameter study provided in the supplementary material. Image augmentation in the CMCR module follows Unimatch V2 [43], with weak augmentations that apply mild photometric changes, and strong augmentations that include color jitter, CutMix, and geometric distortions. Following [24], rare-class sampling is applied.

4.2 Comparison with the State-Of-The-Art Methods

We benchmark HVLFormer against leading SSS methods across four benchmarks. The model achieves SOTA performance with consistent gains over both VLM and non-VLM methods. By leveraging domain-aware and robust language semantics for segmentation mask generation, HVLFormer enhances model semantic understanding under low-label conditions, demonstrating larger improvements at lower labeled training ratios and strong label efficiency overall. Even with the shallower EVA02-S [11] backbone, HVLFormer (40M) outperforms all VLM-based SOTA, including SemiVL (88M) [13] and Pseudo-SD (88M) [47], highlighting its effectiveness in utilizing global textual priors for enhanced semantic understanding in low-label regimes. As shown in Fig. 3, HVLFormer produces precise segmentations, effectively distinguishing confusing or rare class pairs such as sofa–chair and avoiding background misclassifications like racket nets, demonstrating strong domain awareness, contextual reasoning, and semantic consistency.

Table 2: HVLFormer comparison on COCO under SSS.

Method	BB	1/512 (232)	1/256 (463)	1/128 (925)	1/64 (1.8k)	1/32 (3.7k)
UniM V1 [42]	XC65	31.9	38.9	44.4	48.2	49.8
LogicD. [19]	XC65	33.1	40.3	45.4	48.8	50.5
UniM V1† [42]	ViT-B	36.6	44.1	49.1	53.5	55.0
Pseudo-SD [47]	ViT-B	39.9	44.8	48.2	50.2	51.9
SemiVL [13]	ViT-B	50.1	52.8	53.6	55.4	56.5
Ours	ViT-B	**54.9**	**56.7**	**61.2**	**62.9**	**65.1**
Ours	E.02-S	**52.7**	**54.8**	**56.2**	**58.5**	**61.1**
POS [32]	D.v2-S	40.9	46.8	54.3	56.2	57.7
UniM V2 [43]	D.v2-B	47.9	55.8	58.7	60.4	63.3
TQDM [28]	SigLIP2	40.2	48.7	54.3	58.9	62.7
Ours	SigLIP2	**58.8**	**61.7**	**63.9**	**64.8**	**67.2**

Table 3: HVLFormer comparison on Cityscapes under SSS.

Method	BB	1/30 (100)	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1.4k)
P.Seg [55]	R101	61.0	–	69.8	72.4	–
UniM V1† [42]	ViT-B	73.8	76.6	78.2	79.1	79.6
Pseudo-SD [47]	ViT-B	76.1	78.5	79.5	80.5	81.5
SemiVL [13]	ViT-B	76.2	77.9	79.4	80.3	80.6
TQDM [28]	ViT-B	65.2	72.8	77.1	80.7	81.8
Ours	ViT-B	**78.1**	**83.0**	**84.6**	**84.8**	**85.1**
Ours	E.02-S	**77.5**	**81.3**	**83.7**	**84.0**	**84.3**
POS [32]	D.v2-S	40.9	81.4	82.7	83.0	83.2
UniMV2 [43]	D.v2-B	–	83.6	84.3	84.5	85.1
TQDM [28]	SigLIP2	69.1	75.5	79.2	81.6	83.9
Ours	SigLIP2	**79.9**	**84.4**	**85.0**	**85.2**	**85.4**

Comparisons on Pascal VOC. We vary the number of labeled images from 92 to 1,464 on the Pascal VOC dataset (10,582 training images; 1,464 labeled). Tab. 1 summarizes results across these splits, showing that HVLFormer consistently outperforms prior VLM-based methods, including Pseudo-SD [47] and SemiVL [13], with gains of up to +5.1 mIoU under the lowest labeled split. Compared with its baseline [28], our improvements range from +13.0 mIoU to +3.1 mIoU, confirming that conditioning domain-invariant queries with domain awareness leads to substantial benefits in low-annotation regimes. With only 14% labeled data, it achieves a new SOTA of 91.7 mIoU over Unimatch V2's [43] 90.8%.

Comparison on COCO. The COCO dataset (118k training images; 81 classes) introduces a higher degree of complexity for SSS due to the higher number of classes and fine-grained inter-class overlap. Despite this, HVLFormer achieves SOTA, with a +10.9 mIoU gain using only 232 labeled images and up to a +18.6 mIoU improvement over its baseline depicted in Tab 2. Its consistent gains across splits indicate reduced ambiguity in segmenting visually similar and rare classes, alongside improved modeling of intra-class variation. Extended performance comparisons are provided in supp.

Comparison on ADE20K. We present an even broader scene-parsing challenge using ADE20K (20.21k train images; 150 classes). Tab. 4 shows that HVLFormer achieves a +15.0 mIoU gain over its baselines with only 158 labels, while achieving a +4.8 mIoU over existing SOTA with 316 training labels, further validating its effectiveness across several classes and challenging datasets.

Comparison on Cityscapes. On Cityscapes (2.975k training images; 19 classes), HVLFormer outperforms SemiVL [13] by 3.7% with 100 labels and surpasses UniMatch V2 [43] by +0.8 mIoU with 186 labels, as shown in Tab. 3. While the overall gains are smaller compared to other benchmarks, this is due to the presence of numerous small and distant object classes (e.g., poles, traffic signs) that are not part of the captions in VLM pre-training.

4.3 Analysis

Ablation. Tab. 6 presents the ablation results of HVLFormer (SigLIP2) on Pascal VOC and COCO with the TQDM [28] baseline. Across both datasets, each component contributes progressively, with larger gains in low-label splits. This

Table 4: HVLFormer comparison on ADE20K under SSS.

Method	BB	1/128 (158)	1/64 (316)	1/32 (632)	1/16 (1.3k)	1/8 (2.5k)
AEL [14]	R101	–	–	28.4	33.2	38.0
UniMatch V1† [42]	ViT-B	18.4	25.3	31.2	34.4	38.0
Pseudo-SD [47]	ViT-B	–	28.1	33.3	35.2	39.3
SemiVL [13]	ViT-B	28.1	33.7	35.1	37.2	39.4
Ours	ViT-B	**35.2**	**39.7**	**46.6**	**47.5**	**51.0**
Ours	EVA02-S	**33.8**	**35.3**	**36.9**	**39.1**	**45.7**
UniM V2 [43]	D.v2-B	–	38.7	45.0	46.7	49.8
TQDM [28]	SigLIP2	25.3	31.8	39.9	43.2	48.1
Ours	SigLIP2	**40.3**	**45.1**	**49.7**	**52.4**	**54.5**

Table 5: Study on Semantic Relevance Estimator (SRE) with different variants on Pascal VOC and COCO.

Configuration	Pascal $mIoU_{92}$	COCO $mIoU_{232}$
No SRE	88.3	57.8
Fixed threshold ($s_k > 0.5$)	88.5	58.0
Soft weighing (weigh queries by s_k)	89.1	58.8

Table 6: Ablation on Pascal VOC and COCO under SSS. Low-label splits benefit more than higher-label splits.

Configuration	Pascal		COCO	
	$mIoU_{92}$	$mIoU_{1464}$	$mIoU_{232}$	$mIoU_{3700}$
Baseline	76.1	88.6	40.2	62.7
+HTQG (Dataset attr)	77.3	88.9	41.9	63.2
+HTQG (Dataset attr + HQG)	79.4	89.4	45.3	64.1
+HTQG (Dataset attr + HQG + SRE)	80.5	89.6	46.9	64.4
+PTRM	84.1	90.2	52.3	65.8
+CMCR (L_{mask})	85.8	90.8	54.1	66.1
+CMCR (L_{mask} + L_{class})	86.2	90.9	54.8	66.2
+CMCR (L_{mask} + L_{class} + L_{align})	89.1	91.7	58.8	67.2

Table 7: Comparison of prompting strategies for HTQG on Pascal VOC, COCO, and Cityscapes.

Prompting Method	Pascal $mIoU_{92}$	COCO $mIoU_{232}$	Cityscapes $mIoU_{100}$
Fixed prompt — Class name only	86.4	54.2	75.1
Fixed prompt — Class + dataset attribute	86.8	55.1	75.4
Learnable prompt — Class name only	88.2	57.9	78.3
Learnable prompt — Class + dataset attribute	89.1	58.8	79.9

trend indicates that HVLFormer's modules, designed to enhance semantic understanding through domain-robust and aware textual queries, provide greater benefits when supervision is limited, as additional labels naturally reduce the need for further guidance. Introducing HTQG (Dataset Attr) increases Pascal $mIoU_{92}$ by +1.2 and COCO $mIoU_{232}$ by +1.7, demonstrating that enriching class semantics with dataset-aware language priors improves performance. Extending this with HQG provides additional gains. Its stronger effect, especially on COCO, likely arises from the dataset's higher class diversity, where hierarchical query granularity helps reduce confusion among intraclass variation and similar classes. Adding SRE further improves results, depicting that filtering out absent class queries mitigates semantic drift. The PTRM module yields substantial gains, confirming that image-conditioned query refinement enhances cross-modal alignment and semantic understanding. Finally, incorporating the CMCR ℓ_{mask} and ℓ_{align} consistency terms produces further improvements, indicating that enforcing cross-view consistency and pixel-text regularization strengthens query domain robustness and stabilizes vision–language interactions. These are further supported by ℓ_{class}. Detailed ablation is provided in supp.

Comparison of Dataset-Aware Prompting. The comparison of different prompting strategies for HTQG on Pascal VOC and COCO is shown in Tab. 7. Fixed prompts using only class names perform the weakest, as they lack contextual understanding. Learnable prompts improve results by adapting to task-specific features. Significant boosts comes from adding dataset attributes, which supply the VLM with crucial contextual cues about scene types—such as general consumer photos in Pascal VOC and urban driving scenes in Cityscapes—helping

Table 8: Comparison of hierarchical levels and query enhancement methods on Pascal VOC and COCO datasets.

(a) Hierarchical level for HQG

E	Pascal				COCO
	IoU_{chair}	IoU_{sofa}	IoU_{tab}	$mIoU_{92}$	$mIoU_{232}$
1	47.2	61.1	74.5	87.3	56.0
2	58.4	67.8	78.6	88.4	57.6
3	63.7	72.3	80.4	89.1	58.8

(b) T.Query Enhancement

Enhancement Method	Pascal $mIoU_{92}$	COCO $mIoU_{232}$
Baseline (text → pixel)	86.8	56.0
C.Attn(text ↔ pixel)	87.2	56.8
PTRM	89.1	58.8

it generate domain-aware text embeddings. These enable the model to produce richer, context-sensitive language embeddings that capture intra-class diversity and complex visual semantics, leading to the best overall performance. The significant increase also indicates that rich embeddings provide HVLFormer semantics that better align with visual features.

Impact of SRE. The impact of different SRE variants on the framework is given in Tab. 5. Removing SRE reduces performance because irrelevant class queries introduce semantic noise. Using a fixed threshold ($s_k > 0.5$) slightly improves results by filtering unlikely class queries, but it risks discarding queries of class actually present in an image. In contrast, the soft weighting strategy achieves the best performance by adaptively scaling the class queries. Empirically on Pascal VOC, s_k ranges from 0 (absent) to 0.96 (high presence), with an average of 0.36, reflecting that most images contain only a subset of classes.

Analysis of HQG. Tab. 8(a) analyzes the effect of increasing query hierarchy levels (E). Performance consistently improves with deeper hierarchies, as multi-level queries capture semantics from coarse to fine granularity. The effect is more prominent on COCO, where the larger number of classes makes hierarchical modeling crucial for resolving fine-grained distinctions and intra-class variability. Increasing hierarchy levels also improve class-wise IoU, with HVLFormer showing particularly strong gains on visually similar or rare categories such as *sofa*, *chair*, and *dining table*, highlighting that multiscale queries encode coarse-to-fine features, necessary for enhanced semantic discrimination and contextual reasoning. Further analysis is provided in supp.

Effectiveness of Textual Query Enhancement. Our assumption that integrating image context into queries leads to more effective segmentation in data scarcity is confirmed by Tab. 8(b). PTRM achieves the best performance by incorporating pixel-level positional cues and local feature distributions in textual queries, enabling them to align more precisely with spatial structures. It surpasses the cross-attention variant.

5 Conclusion

We present a novel and unified framework for SSS. By integrating domain-aware semantics with domain-invariant embeddings from vision–language models as textual queries and refining them through HTQG, PTRM, and CMCR, HVLFormer achieves robust class discrimination as it effectively compensates for weak semantic understanding under minimal supervision. We have demonstrated SOTA performance across SSS benchmarks with rigorous ablation studies.

References

1. Abdelfattah, R., Guo, Q., Li, X., Wang, X., Wang, S.: Cdul: Clip-driven unsupervised learning for multi-label image classification. In: IEEE/CVF International Conference on Computer Vision. pp. 1348–1357 (2023) 2
2. Cha, J., Mun, J., Roh, B.: Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11165–11174 (2023) 2
3. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2613–2622 (2021) 5, 11
4. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299 (2022) 6, 7, 10, 12
5. Chi, H., Pang, J., Zhang, B., Liu, W.: Adaptive bidirectional displacement for semi-supervised medical image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4070–4080 (2024) 2
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3213–3223 (2016) 6, 7, 12
7. Ding, J., Xue, N., Xia, G.S., Dai, D.: Decoupling zero-shot semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11583–11592 (2022) 2, 12
8. Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., et al.: Maskclip: Masked self-distillation advances contrastive language-image pretraining. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10995–11005 (2023) 6
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 11
10. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (2010) 7, 12
11. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva-02: A visual representation for neon genesis. Image and Vision Computing **149**, 105171 (2024) 2, 11, 12
12. Guan, D., Huang, J., Xiao, A., Lu, S.: Unbiased subclass regularization for semi-supervised semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9958–9968 (2022) 5
13. Hoyer, L., Tan, D.J., Naeem, M.F., Van Gool, L., Tombari, F.: Semivl: Semi-supervised semantic segmentation with vision-language guidance. In: European Conference on Computer Vision. pp. 257–275. Springer (2024) 2, 5, 7, 11, 12, 13, 14
14. Hu, H., Wei, F., Hu, H., Ye, Q., Cui, J., Wang, L.: Semi-supervised semantic segmentation via adaptive equalization learning. In: Advances in Neural Information Processing Systems. pp. 22106–22118 (2021) 5, 14
15. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022) 6

16. Jia, Y., Hoyer, L., Huang, S., Wang, T., Gool, L.V., Schindler, K., Obukhov, A.: Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In: European Conference on Computer Vision. Springer (2024) 2
17. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13619–13627 (2022) 11
18. Li, Y., Wang, X., Yang, L., Feng, L., Zhang, W., Gao, Y.: Diverse cotraining makes strong semi-supervised segmentor. arXiv preprint arXiv:2308.09281 (2023) 11
19. Liang, C., Wang, W., Miao, J., Yang, Y.: Logic-induced diagnostic reasoning for semi-supervised semantic segmentation. In: IEEE/CVF International Conference on Computer Vision. pp. 16197–16208 (2023) 11, 13
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. pp. 740–755. Springer (2014) 12
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916 (2023) 2
22. Ma, J., Wang, C., Liu, Y., Lin, L., Li, G.: Enhanced soft label for semi-supervised semantic segmentation. In: IEEE/CVF International Conference on Computer Vision. pp. 1185–1195 (2023) 11
23. Mittal, S., Tatarchenko, M., Brox, T.: Semi-supervised semantic segmentation with high-and low-level consistency. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(4), 1362–1379 (2019) 2
24. Nadeem, N., Asad, M.H., Anwar, S., Bais, A.: Maskadapt: Unsupervised geometry-aware domain adaptation using multimodal contextual learning and rgb-depth masking. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5422–5432 (2025) 2, 12
25. Nadeem, N., Asad, M.H., Bais, A.: Robust uda for crop and weed segmentation: Multi-scale attention and style-adaptive techniques. In: European Conference on Computer Vision. pp. 284–302. Springer (2025) 2, 12
26. Nadeem, N., Asad, M.H., Bais, A.: CCUDA-SS: Cross-crop unsupervised domain adaptation for semantic segmentation. IEEE Transactions on AgriFood Electronics **4**(1), 334–350 (2026) 2
27. Nadeem, N., Asad, M.H., Ullah, M., Bais, A.: Geometry-guided plant stand count estimation in densely populated lentil fields. IEEE Transactions on AgriFood Electronics pp. 1–17 (2026) 2
28. Pak, B., Woo, B., Kim, S., Kim, D.h., Kim, H.: Textual query-driven mask transformer for domain generalized segmentation. In: European Conference on Computer Vision. pp. 1234–1245. Springer (2024) 6, 10, 11, 12, 13, 14
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, S., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR (2021) 2, 3, 5, 11
30. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: Advances in Neural Information Processing Systems. vol. 33, pp. 596–608 (2020) 5
31. Souly, N., Spampinato, C., Shah, M.: Semi supervised semantic segmentation using generative adversarial network. In: IEEE International Conference on Computer Vision. pp. 5688–5696 (2017) 2

32. Sun, R., Mai, H., Li, W., Chen, Y., Wang, Y.: Two losses, one goal: Balancing conflict gradients for semi-supervised semantic segmentation. In: IEEE/CVF International Conference on Computer Vision. pp. 20357–20367 (2025) 11, 13
33. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025) 11
34. Ullah, M., Ali, N., Nadeem, N., Bais, A.: Agmic: Agricultural masked image consistency for cross-domain segmentation. In: IEEE/CVF International Conference on Computer Vision Workshops. pp. 7139–7149 (October 2025) 2
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30 (2017) 11
36. Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., Le, X.: Semi-supervised semantic segmentation using unreliable pseudo-labels. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4248–4257 (2022) 11
37. Wu, L., Fang, L., He, X., He, M., Ma, J., Zhong, Z.: Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 8827–8844 (2023) 5
38. Xu, H., Liu, L., Bian, Q., Yang, Z.: Semi-supervised semantic segmentation with prototype-based consistency regularization. In: Advances in Neural Information Processing Systems. vol. 35, pp. 26007–26020 (2022) 11
39. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18134–18144 (2022) 2, 6
40. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2945–2954 (2023) 2
41. Yang, L., Zhao, Z., Qi, L., Qiao, Y., Shi, Y., Zhao, H.: Shrinking class space for enhanced certainty in semi-supervised learning. In: IEEE/CVF International Conference on Computer Vision. pp. 16141–16150 (2023) 5
42. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7236–7246 (2023) 2, 5, 10, 11, 13, 14
43. Yang, L., Zhao, Z., Zhao, H.: Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(4), 3031–3048 (2025) 2, 5, 11, 12, 13, 14
44. Yang, L., Zhuo, W., Qi, L., Shi, Y., Gao, Y.: St++: Make self-training work better for semi-supervised semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4268–4277 (2022) 2, 5, 11
45. Yuan, S., Zhong, R., Yang, C., Li, Q., Dong, Y.: Dynamically updated semi-supervised change detection network combining cross-supervision and screening algorithms. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024) 2
46. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: IEEE/CVF International Conference on Computer Vision. pp. 6022–6031 (2019) 5

47. Zhao, D., Zang, Q., Wang, S., Sebe, N., Zhong, Z.: Pseudo-sd: pseudo controlled stable diffusion for semi-supervised and cross-domain semantic segmentation. In: IEEE/CVF International Conference on Computer Vision. pp. 22393–22403 (2025) 3, 11, 12, 13, 14
48. Zhao, Z., Long, S., Pi, J., Wang, J., Zhou, L.: Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23705–23714 (2023) 5
49. Zhao, Z., Wang, Z., Wang, L., Yu, D., Yuan, Y., Zhou, L.: Alternate diverse teaching for semi-supervised medical image segmentation. In: European Conference on Computer Vision. pp. 227–243. Springer (2024) 2
50. Zhong, Z., Cui, J., Yang, Y., Wu, X., Qi, X., Zhang, X., Jia, J.: Understanding imbalanced semantic segmentation through neural collapse. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19550–19560 (2023) 2
51. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 633–641 (2017) 3, 12
52. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision. pp. 696–712. Springer (2022) 2
53. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022) 6
54. Zhou, Z., Lei, Y., Zhang, B., Liu, L., Liu, Y.: Zegclip: Towards adapting clip for zero-shot semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11175–11185 (2023) 2, 6, 11
55. Zou, Y., Zhang, Z., Zhang, H., Li, C.L., Bian, X., Huang, J.B., Pfister, T.: Pseudoseg: Designing pseudo labels for semantic segmentation. In: International Conference on Learning Representations (2021) 5, 11, 13