Escaping the CAM Shadow: Uncertainty-Guided Reliable Learning for Weakly Supervised Semantic Segmentation

Luyao Chang[1,2], **Leiting Chen**[1,2], **Chen Yang**[1,2], **Chuan Zhou**[1,2*]

[1]School of Information and Software Engineering, University of Electronic Science and Technology of China, China
[2]Key Laboratory of Intelligent Digital Media Technology of Sichuan Province, China
changluyao16@gmail.com, richardchen@uestc.edu.cn, chasey.grad26@std.uestc.edu.cn, zhouchuan@uestc.edu.cn

Abstract

Weakly supervised semantic segmentation (WSSS) suffers from an inherent mismatch between coarse image-level annotations and dense pixel-level predictions. To bridge this gap, existing methods primarily focus on generating refined class activation maps (CAM) as pseudo-labels. However, we argue that this focus is insufficient as it overlooks a critical component: the segmentation decoder. The decoder is typically trained through superficial alignment of predictions with pseudo-labels in the logit space. Given the noisy nature of such labels, this naive supervision leads to error accumulation and limits performance. To address this, we propose an Uncertainty-Guided Reliable Learning (UGRL) framework that exerts dual control to reshape the learning process, achieving robust supervision that escapes the CAM shadow. The cornerstone of UGRL is a prototype-driven uncertainty modeling module that estimates the reliability of class-wise supervision. The modeled uncertainty enables two synergistic control mechanisms. First, it adaptively modulates classification and segmentation losses, encouraging the model to learn from more trustworthy signals. Second, it guides the structuring of the decoder's feature space. Rather than relying solely on superficial alignment, UGRL enforces deeper representation alignment by applying contrastive learning on reliable pixels. This enables rich semantic transfer to fine-grained segmentation details. Extensive experiments on PASCAL VOC and MS COCO demonstrate that our method surpasses other state-of-the-art WSSS methods.

Code — https://github.com/YL616/UGRL

Introduction

Weakly supervised semantic segmentation (WSSS) aims to generate dense predictions using weak supervision, such as points (Bearman et al. 2016), scribbles (Lin et al. 2016; Vernaza and Chandraker 2017), bounding boxes (Dai, He, and Sun 2015; Lee et al. 2021), and image-level labels (Ru et al. 2022; Yang et al. 2025a). Among these, image-level labels are considered particularly valuable yet most challenging, as they require the lowest annotation cost, but provide no spatial localization information, making it difficult to infer object boundaries or regions directly from the supervision.

Typically, the WSSS pipeline can be divided into three steps. It first trains a classification network using image-level labels to generate class activation maps (CAM) (Zhou et al. 2016). Then CAM is refined as pseudo-labels, which are finally leveraged to train a segmentation model in a fully supervised manner (Kwon et al. 2024). However, this multi-stage paradigm introduces additional complexity and reduces efficiency. Moreover, being commonly built on convolutional neural networks, these methods often struggle to capture global contextual relationships, thereby hindering the activation of complete object regions (Gao et al. 2021).

The advent of Vision Transformers (ViT) (Dosovitskiy et al. 2020) has provided a powerful alternative, as their self-attention mechanism excels at modeling long-range dependencies. Building on this advantage, many studies have proposed single-stage WSSS frameworks (Ru et al. 2022; Xu et al. 2023; Wu et al. 2024; Xu et al. 2025), which jointly generate pseudo-labels and optimize the segmentation head within an end-to-end training process. These Transformer-based methods have shown great promise in capturing more complete object regions (Ru et al. 2022). Subsequent works have further refined this paradigm, with some focusing on enhancing attention localization or integrating multi-scale cues to improve pseudo-label quality (Ru et al. 2023; Wu et al. 2024; Yang et al. 2025b), while others aim to mitigate challenges inherent to ViT, such as weak inductive bias and oversmoothness (Jang et al. 2024; Yang et al. 2025a).

Despite these promising advancements, current methods still suffer from a flawed supervisory pipeline rooted in two fundamental and interconnected issues. First, they are built on an implicit assumption of uniform data quality, treating all supervisory signals derived from image-level labels with uniform confidence. This assumption overlooks the inherent ambiguity and uncertainty present in the data. As illustrated in Fig. 1(a), not all classes correspond to visual evidence of equal saliency or clarity. Forcing the model to learn from visually atypical classes with the same intensity as from canonical ones paradoxically introduces noise and degrades the fidelity of the resulting CAM. Second, this flawed supervision is then ineffectively transferred to the segmentation decoder. Both multi-stage and single-stage methods face a chronic 'CAM-to-segmentation' performance gap (Rong et al. 2023). This is because conventional pixel-wise supervision operates solely in the logit space, enforcing only a

*Corresponding author.

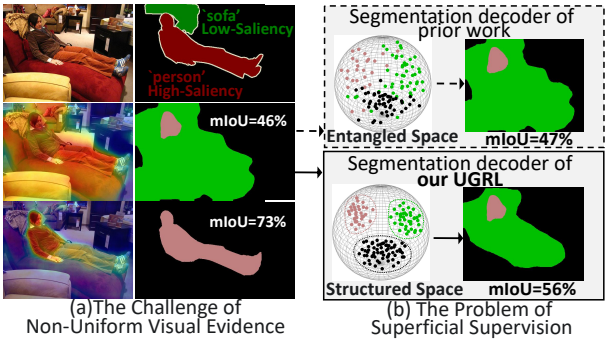

Figure 1: Our motivation. (a) Treating low-saliency class ('sofa') with equal confidence as high-saliency one ('person') introduces noise and degrades pseudo-label quality. (b) Prior work (Xu et al. 2025) mimics pseudo-labels at the logit level, resulting in entangled features and suboptimal segmentation. Our UGRL imposes reliable semantic constraints to structure the feature space, yielding better performance.

superficial pattern mimicry of the noisy pseudo-labels. As shown in Fig. 1(b), this process fails to impose structural constraints on the decoder's feature space. Consequently, the learned representations lack the intra-class compactness and inter-class separability required for precise segmentation, leading to a fundamental disconnect between the encoder's rich semantics and the decoder's final output.

To address these limitations, we propose an Uncertainty-Guided Reliable Learning (UGRL) framework that exerts dual control to reshape the learning process, achieving robust supervision that escapes the CAM shadow. The cornerstone of UGRL is our Prototype-driven Uncertainty Modeling (PUM) module, which establishes a principled way to quantify supervision reliability. By constructing a global prototype base in a hyperdimensional space, PUM estimates the uncertainty of each class-wise signal for every image. This semantically-grounded uncertainty then serves as a guiding signal for a dual control mechanism that reshapes the entire learning process. First, our Uncertainty-guided Loss Modulation (ULM) module leverages the estimated uncertainty to adaptively re-weight loss contributions, encouraging the model to learn from more trustworthy signals. Second, our Reliable Semantic Enhancement (RSE) module reshapes the geometry of the feature space within the decoder. By operating on a dynamic pool of low-uncertainty pixels, RSE imposes direct structural constraints on the pixel embeddings, fostering the intra-class compactness and inter-class separability required for high-fidelity segmentation.

The main contributions of our work are as follows:

- We propose a novel Uncertainty-Guided Reliable Learning (UGRL) framework to rectify the WSSS pipeline by tackling both the indiscriminate trust in noisy source signals and the superficial alignment of the decoder.
- We introduce the Prototype-driven Uncertainty Modeling (PUM) module that provides a principled and effective uncertainty estimate for a more discerning learning process, moving beyond simple confidence scores.
- We propose a dual-control mechanism that leverages estimated uncertainty to simultaneously re-weight supervisory losses via our Uncertainty-guided Loss Modulation (ULM) module and reshape the decoder's feature space via our Reliable Semantic Enhancement (RSE) module.
- Extensive experiments on PASCAL VOC and MS COCO benchmarks validate the effectiveness of our UGRL.

Related Work

Weakly Supervised Semantic Segmentation

Weakly supervised semantic segmentation (WSSS) with image-level labels typically leverages Class Activation Maps (CAM) to achieve dense predictions. Due to its classification nature, standard CAM tends to activate only the most discriminative object parts. Therefore, many works focused on enhancing the quality of CAM. Early research often employed heuristic strategies such as erasing and accumulation (Zhang et al. 2021; Yoon et al. 2022), auxiliary tasks (Su et al. 2021; Du et al. 2022), and CRF (Krähenbühl and Koltun 2011) to expand activation regions to expand activation regions and refine boundaries. The advent of Vision Transformer (ViT (Dosovitskiy et al. 2020) and its variant (Xie et al. 2021a) offers new potential for advancing WSSS, as the self-attention mechanism naturally captures the long-range dependencies required for integral object localization. AFA (Ru et al. 2022) learns inter-pixel affinities directly from attention maps to refine CAM. SeCo (Yang et al. 2024) contrasts class tokens from full-image labels with those from cropped uncertain regions. PCRE (Xu et al. 2025) gradually learns a faithful mask over the target region. However, these methods focus primarily on CAM refining, while the crucial step of effectively transferring semantic information to the segmentation decoder remains underexplored.

Uncertainty Estimation in Deep Learning

Uncertainty in deep learning typically refers to the confidence level associated with a model's predictions, which can be a statistical representation to indicate the model's understanding of scenes (Kendall and Gal 2017). Due to its potential for reliability evaluation and performance enhancement, uncertainty estimation has received considerable attention across various tasks (Liu et al. 2025; Li et al. 2025). Common uncertainty estimation techniques include Monte Carlo dropout (Gal and Ghahramani 2016), conformal prediction (Angelopoulos, Bates et al. 2023), and evidential deep learning (Sensoy, Kaplan, and Kandemir 2018). Recent research (Ni et al. 2023) shows that constructing hyperdimensional prototypes achieves competitive performance while offering significant speedups. The concept of uncertainty has also been introduced into WSSS. URN (Li et al. 2022) simulates noisy response variations by scaling prediction maps, while other methods (He et al. 2024; Lin et al. 2023a) estimate uncertainty directly from CAM scores. However, these methods typically estimate uncertainty in logit space, which is sensitive to noise and lacks semantic robustness. In contrast, we model uncertainty in a hyperdimensional feature space and propagate it to guide the whole learning process, a holistic strategy not explored by previous WSSS works.

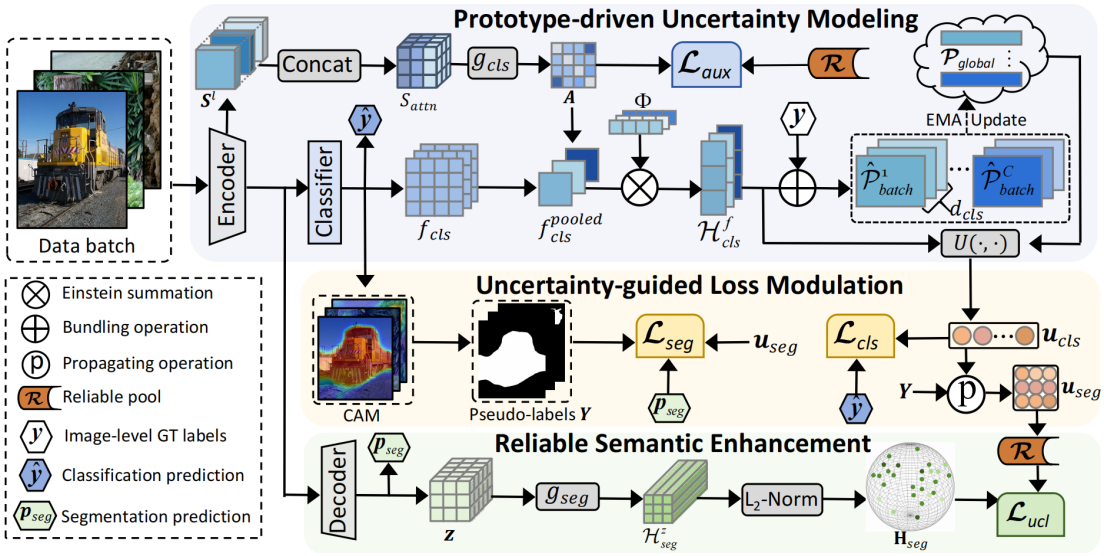

Figure 2: Overview of our UGRL. The framework exerts dual control over the learning process through three key stages: (1) PUM: Multi-head attention-weighted class representations are projected into a hyperdimensional space and then bundled into prototypes to quantify class-wise uncertainty. (2) ULM: Class-wise uncertainty directly modulates the classification loss and is also propagated to the pixel level to adaptively re-weight the segmentation loss. (3) RSE: The decoder's pixel embeddings are projected into a separate hyperspherical space and structured via an uncertainty-guided contrastive learning objective.

Methodology

The pipeline of our UGRL framework is illustrated in Fig. 2. It is composed of three key modules: Prototype-driven Uncertainty Modeling (PUM), Uncertainty-guided Loss Modulation (ULM), and Reliable Semantic Enhancement (RSE).

Prototype-driven Uncertainty Modeling

To challenge the assumption that all image-level labels provide equally trustworthy guidance, we propose the Prototype-driven Uncertainty Modeling (PUM) module. The core function of PUM is to construct a set of stable, dataset-wide class prototypes within a hyperdimensional space. These prototypes then serve as semantic anchors, allowing us to quantify class-wise uncertainty for each image by measuring its feature similarity to these anchors.

Hyperdimensional Prototype Representation Learning. We utilize a Transformer backbone to extract features for constructing class-wise prototypes. Standard global pooling causes semantic dilution and is suboptimal for this task. To address this, we introduce a semantic affinity matrix $A \in \mathbb{R}^{hw \times hw}$ to guide pooling, where hw is the flattened spatial size. Inspired by (Ru et al. 2022), A is derived from the rich relational information embedded in the self-attention maps across multiple Transformer layers. Specifically, for each of the L layers, we extract n attention maps and stack them to form a layer-specific spatial tensor $\boldsymbol{S}^{(l)} \in \mathbb{R}^{hw \times hw \times n}$. These tensors are then concatenated to create a fused multi-level representation: $S_{attn} = \text{Concat}(\boldsymbol{S}^{(1)}, \ldots, \boldsymbol{S}^{(L)})$. Based on S_{attn}, we use a lightweight MLP g_{cls} to compute the semantic affinity matrix A as follows:

$$A = g_{cls}(S_{attn} + S_{attn}^{\top}), \tag{1}$$

where $S_{attn}^{\top}$ is the transpose of S_{attn} to ensure symmetry.

This semantic affinity matrix A then guides the aggregation of class-specific features. First, a classification head is applied to the backbone's final feature maps to produce class activation maps $f_{cls} \in \mathbb{R}^{hw \times hw \times C}$, where C is the number of semantic classes. Then, the matrix A is used to perform a weighted spatial pooling on f_{cls}, yielding class representation vectors $f_{cls}^{pooled} \in \mathbb{R}^{C}$. To endow these representations with a robust geometric structure (Chen et al. 2025), we employ a matrix Φ to project them into a d_{cls}-dimensional space $\mathbf{H}_{cls}$. This process yields the class-wise hypervector representation $\mathcal{H}_{cls}^{f} \in \mathbb{R}^{C \times d_{cls}}$ as follows:

$$\phi(f_{cls}) = \Phi^{(C \times d_{cls})} \otimes f_{cls}^{pooled\ (C)} = \mathcal{H}_{cls}^{f\ (C \times d_{cls})}, \tag{2}$$

where $\otimes$ denotes the Einstein summation notation to retain the channel dimension. For a given mini-batch B, we then compute the prototype $\hat{\mathcal{P}}_{batch}^{c}$ for each class c by bundling the corresponding hypervectors from all images in the batch that contain that class. Inspired by (Levy and Gayler 2008), this bundling is implemented as a simple yet effective element-wise summation, formally expressed as:

$$\hat{\mathcal{P}}_{batch}^{c} = \bigoplus_{x \in B, y_x^c = 1}^{dim = d_{cls}} \phi(f_{cls}^{x}) \tag{3}$$

where $y_x^c \in \{0,1\}$ is ground-truth indicator of whether class c is present in image x, and $\bigoplus$ represents bundling operation.

Uncertainty Estimation. A naive aggregation that relies solely on the current batch may yield noisy and unstable prototypes sensitive to batch-to-batch variations. To improve stability and promote the learning of representations that

generalize well across the dataset, we employ an Exponential Moving Average (EMA) update strategy. The global prototype for class c is updated as:

$$\mathcal{P}^c_{global} \leftarrow \eta \cdot \mathcal{P}^c_{global} + (1-\eta) \cdot \hat{\mathcal{P}}^c_{batch}\ , \tag{4}$$

where $\eta \in [0, 1)$ is a hyperparameter controlling the momentum of the update rate. This update is applied throughout training to accumulate dataset-wide semantic knowledge.

With these refined prototypes $\boldsymbol{\mathcal{P}}_{global}$ serving as stable semantic anchors, we can estimate class-wise uncertainty for each sample. The core intuition is that a reliable prediction should be geometrically close to its corresponding class prototype, while a large distance signifies high uncertainty, suggesting the class is visually atypical or ambiguous. In $\mathbf{H}_{cls}$, this geometric proximity is well-measured by cosine similarity. We thus define function $U(\cdot, \cdot)$ to measure uncertainty score for class c as the cosine distance between a hypervector $\mathcal{H}^c_{cls}$ and its corresponding global prototype $\mathcal{P}^c_{global}$:

$$u^c_{cls} = U(\mathcal{H}^c_{cls},\ \mathcal{P}^c_{global}) = 1 - \frac{\langle\ \mathcal{H}^c_{cls}\ ,\ \mathcal{P}^c_{global}\ \rangle}{\|\mathcal{H}^c_{cls}\|_2 \cdot \|\mathcal{P}^c_{global}\|_2}. \tag{5}$$

Uncertainty-guided Loss Modulation

The uniform weighting inherent in conventional classification losses renders the learning process suboptimal, as it accords equal importance to both reliable supervisory signals and those from ambiguous or atypical samples. To rectify this, we dynamically modulate the loss based on the estimated uncertainty scores, facilitating learning from reliable supervision. Specifically, we employ a simple exponential decay function to transform $\boldsymbol{u}_{cls}$ into a confidence weight, which is then applied to the multi-label soft margin loss:

$$\mathcal{L}_{cls} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{e}^{\frac{1}{\alpha \cdot u^c_{cls}}} (y^c \log(\hat{y}^c) + (1-y^c)\log(1-\hat{y}^c)), \tag{6}$$

where α is a temperature hyperparameter, and $\hat{y}^c$ denotes the predicted probability for class c.

Similarly, we extend this uncertainty-guided principle to the segmentation loss. Following standard practice, we generate pixel-level pseudo-labels $\boldsymbol{Y} \in \mathbb{R}^{H \times W}$ based on CAM, where $H \times W$ denotes the image size. However, segmentation training with these pseudo-labels is susceptible to the noise and spatial inaccuracies inherited from the raw CAM.

To mitigate this, we modulate the standard pixel-wise Cross-Entropy loss based on the modeled uncertainty. First, we construct a pixel-wise uncertainty map, $\boldsymbol{u}_{seg} \in \mathbb{R}^{H \times W}$. The uncertainty for each pixel (i, j) is assigned the class-wise uncertainty of its designated pseudo-label, retrieved from the vector $\boldsymbol{u}_{cls} \in \mathbb{R}^C$:

$$u^{i,j}_{seg} = u_{cls}[Y_{i,j}]. \tag{7}$$

This process effectively propagates the class-wise uncertainty to pixel-wise, which can directly guide the segmentation loss. Subsequently, we incorporate $\boldsymbol{u}_{seg}$ into the segmentation objective to dynamically focus training on more trustworthy pseudo-labeled regions. The final segmentation loss is defined as:

$$\mathcal{L}_{seg} = \frac{1}{H \cdot W}\sum_{i=1}^{H}\sum_{j=1}^{W} \mathrm{e}^{\frac{1}{\beta \cdot u^{i,j}_{seg}}} \cdot Y_{i,j} \log(p^{i,j}_{seg}), \tag{8}$$

where β is a temperature hyperparameter and $p^{i,j}_{seg}$ denotes the predicted probability of the decoder at pixel (i, j).

Reliable Semantic Enhancement

Although the loss $\mathcal{L}_{seg}$ improves robustness by incorporating uncertainty, its operational domain is confined to logit space. Such supervision focuses on superficial alignment, forcing the predictions to approximate the pseudo-labels without imposing any explicit structural constraints on feature representations. However, the quality of these features is paramount for high-fidelity segmentation. A well-structured feature space, characterized by high intra-class compactness and inter-class separability, is essential to overcoming the limitations of CAM-based supervision in WSSS. To instill this deeper semantic structure, we propose the Reliable Semantic Enhancement (RSE) module. This module goes beyond superficial logit alignment to learn a more robust feature manifold, which is crucial for generating complete object masks and refining sharp segmentation boundaries.

Semantic Metric Space Construction. To acquire a feature representation rich in both semantic context and fine-grained spatial detail, our decoder sources multi-scale feature maps from various stages of the encoder backbone. These features are fused to produce a unified feature map $\boldsymbol{z} \in \mathbb{R}^{H \times W \times D_2}$, where D_2 is the feature dimension. To facilitate effective metric learning, we employ a non-linear projection head g_{seg} that maps $\boldsymbol{z}$ into a dedicated hyperdimensional space $\mathbf{H}_{seg}$. This yields the embedding $\mathcal{H}^z_{seg} = g_{seg}(z)$, where $\mathcal{H}^{\boldsymbol{z}}_{seg} \in \mathbb{R}^{N \times d_{seg}}$ and $N = H \times W$ denotes the number of spatial locations. Crucially, this design decouples the representational requirements of the segmentation task from those of metric learning. To render $\mathbf{H}_{seg}$ an effective space for semantic measurement, we apply L_2 normalization to all pixel embeddings as follows:

$$\mathcal{H}^i_{seg} \leftarrow \frac{\mathcal{H}^i_{seg}}{\|\mathcal{H}^i_{seg}\|_2 + \epsilon}\ , \quad \text{for } i = 1, \ldots, N. \tag{9}$$

where ϵ is a small constant. This normalization constrains the embeddings to the surface of a unit hypersphere, enabling cosine similarity to serve as a reliable proxy for semantic proximity in the high-dimensional space.

Uncertainty-guided Contrastive Learning. Ideally, pixel-level embeddings belonging to the same semantic class should be closer, while those from different classes should remain well separated in $\mathbf{H}_{seg}$. The pseudo-labels $\boldsymbol{Y} \in \mathbb{R}^{H \times W}$ provide the foundational supervision to achieve this. However, accepting these labels wholesale is problematic, as $\boldsymbol{Y}$ inherits the noise and spatial ambiguity of CAM, inevitably including mislabeled pixels. Applying contrastive loss indiscriminately in this context may reinforce incorrect relationships and degrade the learned feature space. To mitigate this, we propose the Uncertainty-guided

Contrastive Learning mechanism, designed to selectively structure the feature space by learning exclusively from the most reliable subset. For each image, we construct a reliable pool $\mathcal{R}$ by ranking all pixels based on their corresponding uncertainty scores $\boldsymbol{u}_{seg}$ in ascending order and selecting the top-K pixels with the lowest scores. For each anchor pixel i in $\mathcal{R}$, we define its positive set $\mathcal{R}^+$ and negative set $\mathcal{R}^-$ as:

- $\mathcal{R}^+$ contains all other pixels j in $\mathcal{R}$ that share the same pseudo-label, i.e., $Y_j = Y_i$.
- $\mathcal{R}^-$ contains all pixels j in $\mathcal{R}$ with a different pseudo-label, i.e., $Y_j \neq Y_i$.

For each anchor i in the reliable set $\mathcal{R}$, the loss is defined as:

$$\ell_i = -\log \frac{\sum\limits_{j\in\mathcal{R}^+} \exp\left(s(\mathcal{H}^i_{seg}\cdot\mathcal{H}^j_{seg})/\tau\right)}{\sum\limits_{j\in\mathcal{R}^+} \exp\Big(s(\mathcal{H}^i_{seg}\cdot\mathcal{H}^j_{seg})/\tau\Big) + \sum\limits_{j\in\mathcal{R}^-} \exp\Big(s(\mathcal{H}^i_{seg}\cdot\mathcal{H}^j_{seg})/\tau\Big)}, \tag{10}$$

where $s(\cdot)$ denotes cosine similarity, and τ is a temperature hyperparameter.By iterating all anchors in the reliable pool $\mathcal{R}$, the final uncertainty-guided contrastive loss is given by:

$$\mathcal{L}_{ucl} = \frac{1}{|\mathcal{R}|}\sum_{i\in\mathcal{R}} \ell_i \,, \tag{11}$$

where $|\mathcal{R}|$ is the number of reliable anchors.

Overall Training Objective

To further enhance performance, we adopt the affinity loss from (Ru et al. 2022), but restrict the computation of affinity matrix A to pixels within the reliable pool $\mathcal{R}$, formulated as:

$$\mathcal{L}_{aux} = \frac{1}{|\mathcal{R}^+|}\sum_{i\in\mathcal{R}^+} \text{sigmoid}(A^i) + \frac{1}{|\mathcal{R}^-|}\sum_{j\in\mathcal{R}^-}\big(1-\text{sigmoid}(A^j)\big) \tag{12}$$

The overall loss of our UGRL is formulated as:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1\mathcal{L}_{seg} + \lambda_2\mathcal{L}_{ucl} + \lambda_3\mathcal{L}_{aux}, \tag{13}$$

where λ_1, λ_2, and λ_3 are the weight factors.

Experiments

Experimental Settings

Datasets and Evaluation Metrics. Experiments are conducted on PASCAL VOC 2012 (Everingham et al. 2010) and MS COCO 2014 (Lin et al. 2014) datasets. The PASCAL VOC 2012 dataset consists of 21 semantic classes. Following common practice (Zhao et al. 2024; Yang et al. 2025a), we use the augmented SBD set (Hariharan et al. 2011), which includes 10,582 training images, 1,449 validation images, and 1,464 testing images. The MS COCO 2014 dataset contains 81 classes, with 82,081 images for training and 40,137 images for validation. Note that we only use image-level labels for annotation. We adopt the mean Intersection-Over-Union (mIoU) as the evaluation metric.

Implementation Details. We adopt the ImageNet-1K pretrained Mix Transformer (MiT) (Xie et al. 2021b) as our backbone. For the segmentation decoder, we employ the MLP decoder head (Xie et al. 2021b), which fuses multi-level feature maps for prediction with simple MLP layers.

Method	Backbone	Train	Val
Multi-stage WSSS methods			
CLIP-ES (Lin et al. 2023b)	RN101	70.5	75.0
CPAL (Tang et al. 2024)	RN101	71.9	-
BECO (Rong et al. 2023)	MiT-B2	-	73.0
CTI (Yoon et al. 2024)	RN101	73.7	-
PCSS (Kwon et al. 2024)	RN38	73.2	-
Single-stage WSSS methods			
AFA (Ru et al. 2022)	MiT-B1	68.7	66.5
ToCo (Ru et al. 2023)	ViT-B	73.6	72.3
DuPL (Wu et al. 2024)	ViT-B	76.0	74.1
DIAL (Jang et al. 2024)	ViT-B	75.2	73.1
PCRE (Xu et al. 2025)	ViT-B	77.6	76.3
FFR (Yang et al. 2025a)	ViT-B	-	76.4
UGRL (Ours)	MiT-B1	**78.5**	**77.9**

Table 1: Comparison of CAM pseudo labels. Evaluation is conducted on the PASCAL VOC 2012 train and val set and reported in mIoU (%).

The entire network is trained using the AdamW optimizer on two NVIDIA RTX 4090 GPUs. The initial learning rate for the backbone is set to 6×10^{-5} and decayed using a polynomial schedule. The learning rate for other parameters is set to 10 times that of the backbone. The weight decay is set to 0.01. We apply simple data augmentation strategies, including random scaling, random horizontal flipping, and random cropping to a fixed size of 512×512. The batch size is set to 8. For the PASCAL VOC 2012 dataset, we train the model for 20,000 iterations, with the first 2,000 iterations warmed up for the classification branch. For the MS COCO 2014 dataset, the total number of training iterations is 80,000, with the first 5,000 iterations for warm-up.

Comparison with State-of-the-arts

Evaluation of Pseudo-labels. We begin by evaluating the quality of the pseudo-labels generated by our UGRL framework, comparing them against current state-of-the-art methods on the PASCAL VOC 2012 train and validation sets. The competing methods are categorized into two groups: multi-stage approaches, which typically involve complex pipelines of initial seed generation followed by iterative refinement or expansion; and single-stage approaches, which generate pseudo-labels from CAMs within an end-to-end training process. As presented in Table 1, UGRL achieves 78.5% mIoU on the VOC train set and 77.9% on the validation set, consistently outperforming all other multi-stage and single-stage counterparts. This superior performance at the foundational pseudo-labeling stage demonstrates that our uncertainty-guided learning paradigm is highly effective, enabling UGRL to produce high-quality pseudo-labels by effectively suppressing supervision noise.

Figure 3 qualitatively compares the CAMs generated by our UGRL against the strong baseline PCRE. The comparison reveals that UGRL produces CAMs that are more accurate and maintain higher fidelity to the shapes of objects.

Evaluation of Segmentation Results. Table 2 presents a comprehensive comparison of our proposed UGRL against

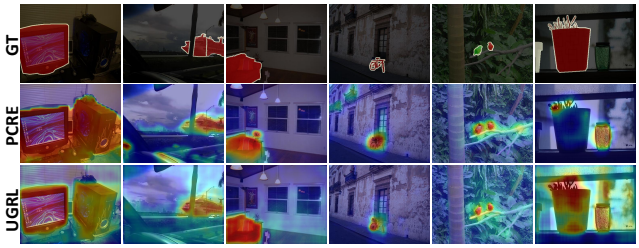

Figure 3: Visualization comparison of CAMs on VOC.

Method	*Sup.*	Backbone	VOC		COCO
			Val	Test	Val
Multi-stage WSSS methods					
L2G (Jiang et al. 2022)	$\mathcal{I}+\mathcal{S}$	RN101	72.1	71.7	44.2
SANCE (Li et al. 2022)	$\mathcal{I}+\mathcal{S}$	RN101	72.0	72.9	44.7
HSC (Wu et al. 2023)	$\mathcal{I}+\mathcal{S}$	RN101	73.6	74.5	–
CLIP-ES (Lin et al. 2023b)	$\mathcal{I}+\mathcal{L}$	RN101	72.7	72.8	45.4
CPAL (Tang et al. 2024)	$\mathcal{I}+\mathcal{L}$	RN101	74.5	74.7	46.3
MCTformer (Xu et al. 2022)	$\mathcal{I}$	RN38	71.9	71.6	42.0
BECO (Rong et al. 2023)	$\mathcal{I}$	MiT-B2	73.7	73.5	45.1
CTI (Yoon et al. 2024)	$\mathcal{I}$	RN101	74.1	73.2	45.5
SFC (Zhao et al. 2024)	$\mathcal{I}$	RN101	71.2	72.5	46.8
PCSS (Kwon et al. 2024)	$\mathcal{I}$	RN38	72.3	73.0	45.7
Single-stage WSSS methods					
1Stage (Ara. and Roth 2020)	$\mathcal{I}$	RN38	62.7	64.3	-
AFA (Ru et al. 2022)	$\mathcal{I}$	MiT-B1	66.0	66.3	38.9
TSCD (Xu et al. 2023)	$\mathcal{I}$	MiT-B1	67.3	67.5	40.1
ToCo (Ru et al. 2023)	$\mathcal{I}$	ViT-B	71.1	72.2	42.3
DuPL (Wu et al. 2024)	$\mathcal{I}$	ViT-B	73.3	72.8	44.6
DIAL (Jang et al. 2024)	$\mathcal{I}$+$\mathcal{L}$	ViT-B	74.5	74.9	44.4
M-SEE (Yang et al. 2025b)	$\mathcal{I}$	ViT-B	74.9	74.8	45.8
PCRE (Xu et al. 2025)	$\mathcal{I}$	ViT-B	75.5	75.9	47.2
FFR (Yang et al. 2025a)	$\mathcal{I}$	ViT-B	74.8	74.5	46.8
UGRL (Ours)	$\mathcal{I}$	MiT-B1	**77.4**	**77.1**	**48.2**

Table 2: Semantic segmentation results on VOC 2012 and COCO 2014 datasets. *Sup.* denotes supervision type. $\mathcal{I}$: image-level labels; $\mathcal{S}$: saliency maps; $\mathcal{L}$: language.

state-of-the-art multi-stage and single-stage methods on VOC COCO. The results demonstrate the superiority of our framework. Specifically, UGRL achieves state-of-the-art performance with 77.4% mIoU on the VOC validation set, 77.1% on the test set, and 48.2% on the COCO validation set. These scores represent significant improvements over the previous best methods by +2.5%, +1.6%, and +2.1%, respectively. Notably, even when compared with methods that incorporate language supervision via text prompts, UGRL consistently achieves superior segmentation performance. This finding underscores the critical importance of effectively modeling and leveraging the inherent uncertainty within the visual modality itself, thereby substantiating the core design principles of our framework.

Fig. 4 visualizes the segmentation results of our UGRL against a strong baseline PCRE on VOC and COCO. The results demonstrate that UGRL achieves more accurate delineation of challenging object categories and consistently produces more complete masks with sharper boundaries.

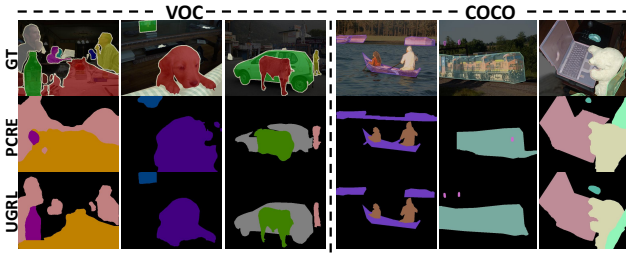

Figure 4: Visualization comparison of segmentation results on VOC and COCO.

Ablation Studies and Analysis

Efficacy of Key Components. The quantitative ablation results of UGRL are summarized in Table 3. We adopt MiT-B1 with auxiliary loss $\mathcal{L}_{aux}$ as the baseline. Setting I validates the effectiveness of our class-wise uncertainty modeling (PUM-c). Guiding the classification loss $\mathcal{L}cls$ with this uncertainty improves CAM and segmentation mIoU by 6.7%, respectively. Setting II further demonstrates the benefit of propagating this uncertainty down to the pixel level (PUM-p) to guide the segmentation loss $\mathcal{L}_{seg}$. The performance boost confirms the importance of trustworthy supervision. Setting III applies contrastive learning without uncertainty guidance. Even when using raw pseudo-labels as the sole supervisory signal, it improves performance and narrows the gap between CAM and segmentation. This powerfully substantiates our claim that directly structuring the decoder's feature space is a critical, and often overlooked, step for enhancing final segmentation quality. Finally, Setting IV represents the full UGRL framework, integrating all components and achieving the best performance.

	PUM-c	$\mathcal{L}_{cls}$	PUM-s	$\mathcal{L}_{seg}$	$\mathcal{L}_{ucl}$	M	Seg.
Base						65.3	63.4
I	✓	✓				69.7	68.2
II	✓	✓	✓	✓		72.5	70.4
III	✓	✓		✓	✓	74.8	74.0
IV	✓	✓	✓	✓	✓	77.9	77.4

Table 3: Ablation study of UGRL components on VOC val set. 'M' denotes the mIoU (%) of CAM performance, and 'Seg.' denotes the mIoU (%) of segmentation performance.

Efficacy of high-dimensional projection. To understand the benefits of high-dimensional projection for prototype learning, we probe the structure of the learned prototype space by visualizing the semantic similarity matrix, which is constructed based on the pair-wise cosine distances between prototypes. Fig. 5(a) reveals the result without this projection. The diagonal is perfectly correlated as expected, but the off-diagonal regions exhibit significant similarity, indicating that prototypes of distinct classes remain semantically entangled and are not effectively separated. In contrast, the similarity matrix produced by our UGRL, shown in Fig. 5(b), displays a clean structure with strong diagonal and off-diagonal values consistently near zero. This confirms that

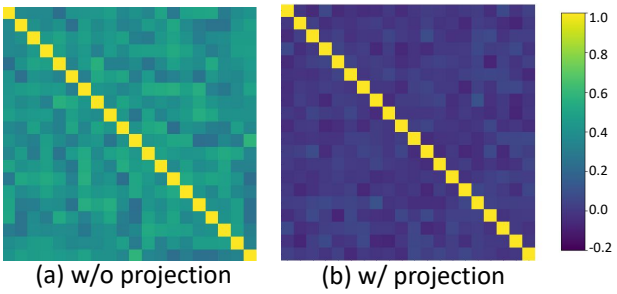

Figure 5: Visualization of the semantic similarity matrix. Classes follow the order of the VOC 2012 dataset.

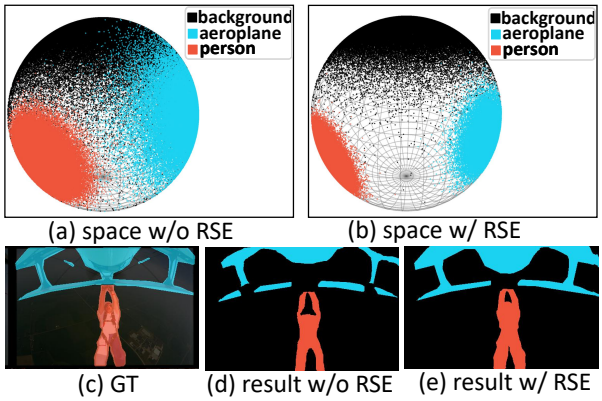

Figure 6: Qualitative analysis of RSE on the embedding space and segmentation results.

our method successfully learns a set of nearly orthogonal class prototypes. This orthogonality, which creates a well-structured and semantically disentangled manifold, is fundamental to the success of our method, as it ensures a reliable and robust foundation for uncertainty estimation.

Further Analysis

Analysis of RSE. To intuitively demonstrate the effectiveness of our RSE, Fig. 6 provides a qualitative comparison in both the embedding space and the segmentation output. As shown in Fig. 6(a), the model trained without RSE produces a chaotic feature space where embeddings from different classes are entangled. This lack of semantic separability leads to poor segmentation performance as shown in Fig. 6(d). In contrast, the incorporation of RSE constructs a well-structured feature space characterized with high intra-class compactness and inter-class separability, as visualized in Fig. 6(b). This high-quality representation enables accurate segmentation shown in Fig. 6(e). These observations validate that the RSE module serves not just as an incremental refinement but as a fundamental mechanism that enforces semantic structure critical for reliable segmentation.

Sensitivity Analysis. We analyze the impact of dimensionality in the two hyperspaces, $\mathbf{H}_{cls}$ and $\mathbf{H}_{seg}$, on both segmentation accuracy and computational cost (FLOPs). The results are presented in Fig. 7. For the prototype space

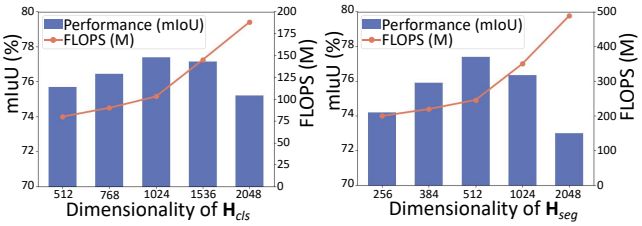

Figure 7: Impact of dimensionality selection on performance and computation on VOC. Left: $\mathbf{H}_{cls}$. Right: $\mathbf{H}_{seg}$.

$\mathbf{H}_{cls}$, where the initial feature dimension of the encoder is 512, we observe that the performance improves as the dimension increases. The mIoU peaks when the dimension reaches 1024, but then declines, while the computational cost (FLOPs) increases monotonically. This suggests that expanding beyond the initial feature dimension is crucial for capturing complex inter-class relationships; however, further increases lead to diminishing returns and increased computational overhead. A similar trend is evident for the pixel-embedding space $\mathbf{H}_{seg}$. Performance reaches its maximum at a dimension of 512. This analysis validates our choice of $d_{cls} = 1024$ and $d_{seg} = 512$, as these values represent the optimal trade-off among representational capacity, segmentation accuracy, and computational efficiency.

We further explore the impact of reliable pool size, controlled by K, on segmentation performance. As shown in Table. 4, performance on both benchmarks first improves and then declines as K increases. This suggests that a small K limits performance due to insufficient sample diversity for contrastive learning, while a large K introduces unreliable pseudo-labels that degrade the feature space.

Dataset	K (Percentage of reliable pixels)					
	5%	20%	35%	50%	65%	80%
VOC	75.8	76.3	**77.4**	77.0	75.2	74.2
COCO	44.1	47.3	**48.2**	46.9	45.5	42.7

Table 4: Analysis on the size of reliable pool $\mathcal{R}$. We vary the percentage K of selected low-uncertainty pixels and report the corresponding mIoU (%) on the VOC and COCO val set.

Conclusion

In this paper, we moved beyond the conventional focus on CAM refinement in WSSS to address a more fundamental issue: the absence of a reliability principle in the supervision pipeline. We argue that both the indiscriminate trust in noisy source signals and the superficial alignment of decoder stem from this deficiency, creating a performance bottleneck. To address this, we proposed the Uncertainty-Guided Reliable Learning (UGRL) framework, which instills reliability as a supervisory principle. Experiments on benchmarks demonstrate the effectiveness of UGRL. Although our method introduces some computational overhead, the proposed learning paradigm offers a promising direction for future research in various weakly-supervised learning tasks.

Acknowledgments

This study was supported by the Natural Science Foundation of Sichuan, China (No. 2023NSFSC0468, No. 2023NSFSC0031).

References

Angelopoulos, A. N.; Bates, S.; et al. 2023. Conformal prediction: A gentle introduction. *Foundations and trends® in machine learning*, 16(4): 494–591.

Bearman, A.; Russakovsky, O.; Ferrari, V.; and Fei-Fei, L. 2016. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, 549–565.

Chen, L.; Wang, J.; Mortlock, T.; Khargonekar, P.; and Al Faruque, M. A. 2025. Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22306–22316.

Dai, J.; He, K.; and Sun, J. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1635–1643.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, Y.; Fu, Z.; Liu, Q.; and Wang, Y. 2022. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4320–4329.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, 1050–1059.

Gao, W.; Wan, F.; Pan, X.; Peng, Z.; Tian, Q.; Han, Z.; Zhou, B.; and Ye, Q. 2021. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2886–2895.

Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *International conference on computer vision*, 991–998.

He, J.; Cheng, L.; Fang, C.; Feng, Z.; Mu, T.; and Song, M. 2024. Progressive feature self-reinforcement for weakly supervised semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 2085–2093.

Jang, S.; Yun, J.; Kwon, J.; Lee, E.; and Kim, Y. 2024. Dial: Dense image-text alignment for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, 248–266.

Jiang, P.-T.; Yang, Y.; Hou, Q.; and Wei, Y. 2022. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16886–16896.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Krähenbühl, P.; and Koltun, V. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24.

Kwon, H.; Jeong, J.; Yoon, S.-H.; and Yoon, K.-J. 2024. Phase concentration and shortcut suppression for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, 293–312.

Lee, J.; Yi, J.; Shin, C.; and Yoon, S. 2021. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the conference on computer vision and pattern recognition*, 2643–2652.

Levy, S. D.; and Gayler, R. 2008. Vector symbolic architectures: A new building material for artificial general intelligence. In *Artificial General Intelligence 2008*, 414–418. IOS Press.

Li, S.; Xu, X.; He, C.; Shen, F.; Yang, Y.; and Tao Shen, H. 2025. Cross-Modal Uncertainty Modeling With Diffusion-Based Refinement for Text-Based Person Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(3): 2881–2893.

Li, Y.; Duan, Y.; Kuang, Z.; Chen, Y.; Zhang, W.; and Li, X. 2022. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 1447–1455.

Lin, D.; Dai, J.; Jia, J.; He, K.; and Sun, J. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3159–3167.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755.

Lin, Y.; Chen, M.; Wang, W.; Wu, B.; Li, K.; Lin, B.; Liu, H.; and He, X. 2023a. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15305–15314.

Lin, Y.; Chen, M.; Wang, W.; Wu, B.; Li, K.; Lin, B.; Liu, H.; and He, X. 2023b. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15305–15314.

Liu, X.; Yang, R.; Wang, S.; Li, W.; Chen, J.; and Zhu, J. 2025. Uncertainty-Instructed Structure Injection for Generalizable HD Map Construction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22359–22368.

Ni, Y.; Chen, H.; Poduval, P.; Zou, Z.; Mercati, P.; and Imani, M. 2023. Brain-inspired trustworthy hyperdimensional computing with efficient uncertainty quantification. In *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, 01–09.

Rong, S.; Tu, B.; Wang, Z.; and Li, J. 2023. Boundary-enhanced co-training for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19574–19584.

Ru, L.; Zhan, Y.; Yu, B.; and Du, B. 2022. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16846–16855.

Ru, L.; Zheng, H.; Zhan, Y.; and Du, B. 2023. Token contrast for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3093–3102.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 31.

Su, Y.; Sun, R.; Lin, G.; and Wu, Q. 2021. Context decoupling augmentation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7004–7014.

Tang, F.; Xu, Z.; Qu, Z.; Feng, W.; Jiang, X.; and Ge, Z. 2024. Hunting attributes: Context prototype-aware learning for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3324–3334.

Vernaza, P.; and Chandraker, M. 2017. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7158–7166.

Wu, Y.; Li, X.; Dai, S.; Li, J.; Liu, T.; and Xie, S. 2023. Hierarchical Semantic Contrast for Weakly Supervised Semantic Segmentation. In *IJCAI*, 1542–1550.

Wu, Y.; Ye, X.; Yang, K.; Li, J.; and Li, X. 2024. Dupl: Dual student with trustworthy progressive learning for robust weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3534–3543.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021a. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021b. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

Xu, L.; Ouyang, W.; Bennamoun, M.; Boussaid, F.; and Xu, D. 2022. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4310–4319.

Xu, R.; Wang, C.; Sun, J.; Xu, S.; Meng, W.; and Zhang, X. 2023. Self correspondence distillation for end-to-end weakly-supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3045–3053.

Xu, X.; Zhang, P.; Huang, W.; Shen, Y.; Chen, H.; Lin, J.; Li, W.; He, G.; Xie, J.; and Lin, S. 2025. Weakly Supervised Semantic Segmentation via Progressive Confidence Region Expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9829–9838.

Yang, Z.; Fu, K.; Duan, M.; Qu, L.; Wang, S.; and Song, Z. 2024. Separate and conquer: Decoupling co-occurrence via decomposition and representation for weakly supervised semantic segmentation. In *Proceedings of the conference on computer vision and pattern recognition*, 3606–3615.

Yang, Z.; Zhao, X.; Wang, X.; Zhang, Q.; and Xiao, J. 2025a. FFR: Frequency Feature Rectification for Weakly Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 30261–30270.

Yang, Z.; Zhao, X.; Yao, C.; Zhang, Q.; and Xiao, J. 2025b. M-SEE: A multi-scale encoder enhancement framework for end-to-end Weakly Supervised Semantic Segmentation. *Pattern Recognition*, 162: 111348.

Yoon, S.-H.; Kweon, H.; Cho, J.; Kim, S.; and Yoon, K.-J. 2022. Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. In *European conference on computer vision*, 326–344.

Yoon, S.-H.; Kwon, H.; Kim, H.; and Yoon, K.-J. 2024. Class tokens infusion for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3595–3605.

Zhang, F.; Gu, C.; Zhang, C.; and Dai, Y. 2021. Complementary patch for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7242–7251.

Zhao, X.; Tang, F.; Wang, X.; and Xiao, J. 2024. Sfc: Shared feature calibration in weakly supervised semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 7525–7533.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2921–2929.