Semi-Supervised Semantic Segmentation via Derivative Label Propagation

Yuanbin Fu[1], **Xiaojie Guo**[1*]

[1]Tianjin University
yuanbinfu@tju.edu.cn, xj.max.guo@gmail.com

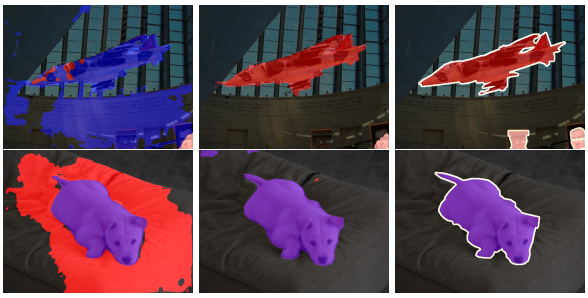
Abstract

Semi-supervised semantic segmentation, which leverages a limited set of labeled images, helps to relieve the heavy annotation burden. While pseudo-labeling strategies yield promising results, there is still room for enhancing the reliability of pseudo-labels. Hence, we develop a semi-supervised framework, namely DerProp, equipped with a novel derivative label propagation to rectify imperfect pseudo-labels. Our label propagation method imposes discrete derivative operations on pixel-wise feature vectors as additional regularization, thereby generating strictly regularized similarity metrics. Doing so effectively alleviates the ill-posed problem that identical similarities correspond to different features, through constraining the solution space. Extensive experiments are conducted to verify the rationality of our design, and demonstrate our superiority over other methods.

Code — https://github.com/ForawardStar/DerProp/
Data — http://host.robots.ox.ac.uk/pascal/VOC/voc2012/
Extended Version — https://arxiv.org/abs/2508.02254v1/

Introduction

Semantic segmentation (SS) aims at segmenting the complete objects that belong to different semantic categories, which is instrumental across various application scenarios, such as autonomous driving (Qu et al. 2021; Tu et al. 2025b; Feng et al. 2021; Wang et al. 2025; Han et al. 2025b), medical imaging (Wang et al. 2022a; Li et al. 2020; Xing et al. 2024, 2022), and robotic navigation (Sanderson and Matuszewski 2022; Tzelepi and Tefas 2021; Lu et al. 2025; Li et al. 2023; Han et al. 2025a; Dong et al. 2023; Fan et al. 2024; Tu et al. 2024a). With the rapid development of deep learning techniques, the accuracy of semantic segmentation has been evidently boosted, benefiting from their capacity to directly learn mappings from raw images to segmentation maps. Unfortunately, manually annotating dense segmentation maps is time-consuming and laborious, presenting a significant barrier to the widespread collection of high-quality data. To relieve such annotation burden, semi-supervised semantic segmentation (SSSS) has garnered substantial interest within the research community. Hence, the goal of SSSS is to attain the high segmentation accuracy, with the benefit of solely employing a small amount of manually annotated samples alongside a large number of unlabeled ones.



w/o DLP w/ DLP GT

Figure 1: Visual comparisons between with and without our proposed derivative label propagation (DLP).

Over the past decades, a plethora of SSSS methodologies have emerged, which predominantly follow a pipeline of generating pseudo-labels by deep neural networks for those unlabeled samples. In particular, existing SSSS methods can be roughly divided into: 1) mean teacher based methods (Hu et al. 2021; Jin, Wang, and Lin 2022; Liu et al. 2022; Xu et al. 2022) that accumulate the network parameters updated at different training iterations to produce pseudo-labels, 2) consistency regularization based methods (Ke et al. 2022; Yang et al. 2022; Yuan et al. 2021; Sun et al. 2024) that enforce the predictions corresponding to weakly and strongly augmented (Lin et al. 2024b,a) inputs to be consistent, and 3) confidence thresholding based methods (Liu et al. 2022; Sun et al. 2024) that set thresholds to preserve reliable pseudo-labels and suppress unreliable pseudo-labels. However, existing methods still suffer from producing imperfect pseudo-labels, hindering further accuracy improvements. To be more specific, the pseudo-labeling process inevitably suffers from noises/outliers, *i.e.*, incorrect labels, which severely misguide the training optimization direction. Consequently, networks trained on noisy pseudo-labels may generate noisier labels in subsequent iterations, leading to the error accumulation problem.

For the sake of improving the performance of SSSS, the pseudo-labels ought to be further rectified so that their re-

*Corresponding Author

liability can be guaranteed. But, it is challenging because there are no accurate dense labels available for supervising the rectification process. In other words, the ground truth conversions from wrong pseudo-labels to desired true labels is unknown. To address this issue, it is crucial to excavate valuable information inherent within the data itself or the learning process. Therefore, based on the principle that pixels with high similarities are highly likely to share the same class label while pixels with low similarities may differ in semantic category, the label propagation technique (Sun et al. 2024; Stojnic et al. 2025; Papadopoulos, Weber, and Torralba 2021; Zhang et al. 2020) can be leveraged to rectify the misclassified pixels, based on semantic similarities with neighboring pixels. This strategy, however, raises a critical question: *how can we reliably measure the similarities among pixels, so as to accurately identify and rectify the potential misclassified pixels in pseudo-labels?*

Contributions. We answer the above question by developing a semi-supervised framework, named DerProp. It adopts a simple yet effective derivative label propagation (DLP) method, which imposes the discrete derivative operation on the pixel-wise feature vectors as additional constraints, and supervises the similarities among pixels using our proposed derivative loss. Specifically, besides supervising/regularizing the similarities between the original feature vectors, the similarities between the derivative feature vectors should also be properly regularized. By doing so, the ill-posed problem that the same similarity scores may correspond to multiple (wrong) solutions, can be alleviated. For example, consider a simple case: $[1, 1, 1]$ and $[2 + \sqrt{3}, 1, 0]$ are the 3-dimensional feature vectors that can correctly represent the semantics of two different pixels. The cosine similarity between $[1, 1, 1]$ and $[2 + \sqrt{3}, 1, 0]$, is the same as that between another pair of vectors, *i.e.*, $[1, 1, 1]$ and $[2-\sqrt{3}, 1, 0]$, both being $\sqrt{2}/2$. But, only $[2 + \sqrt{3}, 1, 0]$ is the correct one. Using $\sqrt{2}/2$ as ground truth to supervise the similarity between these two pixels, encounters a high risk of generating the incorrect $[2 - \sqrt{3}, 1, 0]$ to represent the pixel, leading to misclassification for it. To resolve such ill-posedness, derivative-based features can be employed as additional constraints for restricting the solution space.

Therefore, we propose to calculate the similarities with respect to original feature vectors together with additional derivative feature vectors. Theoretical analysis is provided to formally reveal how discrete derivative operators alleviate the ill-posedness. Further, to mitigate the computational expense of high-order (≥ 2) discrete derivative operations, we introduce a sparsity term to approximate the regularization on the high-order derivative-based features. As shown in Fig. 1, our DLP can evidently improve the segmentation performance through introducing additional regularization.

Our contributions can be summarized as:

- We propose a novel derivative label propagation method that performs discrete derivative operations on the feature vectors. It can alleviate the ill-posed problem that identical similarities may result in different features.
- We prove that high-order discrete derivative operations on feature vectors make the problem well-posed, while the sparsity regularization on the 2-order derivative-based features can provide a well approximation.
- We conduct extensive experiments to quantitatively and qualitatively verify the effectiveness of our design on several benchmark datasets, and demonstrate our superiority over other state-of-the-art methods.

Related Work

Fully-Supervised Semantic Segmentation. The fully-supervised segmentation can be traced back to Fully Convolutional Networks (FCNs) (Shelhamer, Long, and Darrell 2017), pioneering the usgae of end-to-end deep neural networks. Subsequent advancements in enlarging the receptive fields are addressed through Atrous Spatial Pyramid Pooling (ASPP) (Chen et al. 2017). HRNet (Sun et al. 2019) enhances the high-resolution representations through aggregating the up-sampled representations. Zhang *et al.* (Zhang et al. 2018) explored the influence of global contextual information through designing a context encoding module. SegNet (Badrinarayanan, Handa, and Cipolla 2015) maps the low-resolution features to high-resolution features. BiSeNet (Yu et al. 2018) accelerates the running speed using a spatial branch with small-stride convolutions. Despite achieving promising performance, these methods require manually annotated dense segmentation labels for all training samples, which are expensive to collect.

Semi-Supervised Semantic Segmentation. For the mean teacher based methods, AEL (Hu et al. 2021) incorporates a confidence bank that dynamically tracks per-category performance. GTA-Seg (Jin, Wang, and Lin 2022) explicitly disentangles the effects of pseudo-labels on the feature extraction and segmentation prediction pathways. Several recent works (Liu et al. 2022; Xu et al. 2022) also introduced novel extensions to the mean teacher model. Another technological route is the consistency learning based methods, enforcing the predictions corresponding to perturbed inputs to be consistent. As representatives, the idea of (Ke et al. 2022) involves extracting pseudo masks on unlabeled data coupled with multi-task segmentation consistency enforcement. ST++ (Yang et al. 2022) applies strong data augmentations to unlabeled images, which mitigates noisy label overfitting and decouples teacher-student predictions. The distribution-specific batch normalization by (Yuan et al. 2021) addresses the problem of large distribution disparities caused by strong augmentation. Besides the aforementioned methods, many researchers have explored the influences of confidence thresholding. For instance, CorrMatch (Sun et al. 2024) generates a binary map by thresholding confidence scores, explicitly identifying trustworthy pixels in pseudo-labels.

Despite achieving promising results, the quality of pseudo-labels continues to limit the accuracy. We overcome this limitation through a novel derivative label propagation method that rectifies unreliable pseudo-labels.

Methodology

The labeled training set is: $\mathcal{D}_l := \{(\mathbf{X}_{nl} \in \mathbb{R}^{3\times M_{nl}}, \mathbf{Y}_{nl} \in \mathbb{R}^{C\times M_{nl}}, nl := 1, 2, ..., N_l\}$, and the unlabeled set is: $\mathcal{D}_u :=$

$\{\mathbf{X}_{nu} \in \mathbb{R}^{3\times M_{nu}}, nu := 1, 2, ..., N_u\}$, where N_l and N_u are the number of labeled and unlabeled samples, respectively. The total pixel amount of the nl-th labeled and nu-th unlabeled image, and the number of classes, are M_{nl}, M_{nu} and C, respectively. Note that images are represented as flattened 2D matrices in this paper, *i.e.*, $\mathbb{R}^{\text{channel}\times\text{pixel}}$, for notational simplicity, originating from their original 3D structure, *i.e.*, $\mathbb{R}^{\text{channel}\times\text{height}\times\text{width}}$, where pixel = height $\times$ width. During training, weak and strong augmentations are performed on $\mathbf{X}_{nu}$, to obtain $\mathbf{X}^w_{nu}$ and $\mathbf{X}^s_{nu}$, respectively. The predictions from $\mathbf{X}^w_{nu}$, *e.g.*, $\mathbf{P}^w_{nu} := \text{Softmax}(\mathcal{G}(\mathbf{X}^w_{nu}))$, can supervise the predictions from $\mathbf{X}^s_{nu}$, *e.g.*, $\mathbf{P}^s_{nu} := \text{Softmax}(\mathcal{G}(\mathbf{X}^s_{nu}))$, where $\mathcal{G}(\cdot)$ is the segmentation network.

Derivative Label Propagation

Generally, the label propagation can be described as:

$$\widetilde{\mathbf{L}}^w_{nu} := \mathbf{L}^w_{nu}\mathbf{S}_{nu}, \tag{1}$$

where $\mathbf{L}^w_{nu} := \mathcal{G}(\mathbf{X}^w_{nu}) \in \mathbb{R}^{C\times M_{nu}}$ represents the class logits, and $\widetilde{\mathbf{L}}^w_{nu} \in \mathbb{R}^{C\times M_{nu}}$ is the rectified logits after performing the label propagation. The rectified pseudo-labels can thus be obtained by: $\widetilde{\mathbf{P}}^w_{nu} := \text{Softmax}(\widetilde{\mathbf{L}}^w_{nu})$. $\mathbf{S}_{nu} \in \mathbb{R}^{M_{nu}\times M_{nu}}$ is the similarity matrix that models the similarities among pixels. The i-th row and j-th column in $\mathbf{S}_{nu}$ is calculated by $\mathcal{S}(\mathbf{v}_{nu,i}, \mathbf{v}_{nu,j})$, where $\mathcal{S}(\cdot,\cdot)$ means the cosine similarity, and $\mathbf{v}_{nu,i} \in \mathbb{R}^D$ is the feature vector representing the semantics of the i-th pixel. For concise expression, we omit the subscript nu in some following paragraphs, *e.g.*, $M_{nu} \to M$, $\mathbf{S}_{nu} \to \mathbf{S}$, and $\mathbf{v}_{nu,i} \to \mathbf{v}_i$.

However, solely supervising the similarity between the original D-dimensional feature vectors during training, may suffer from the ill-posed problem that identical similarities correspond to different solutions (including those incorrect/biased solutions). Hence, to make the problem well-posed so that extra incorrect solutions can be avoided, we propose to further impose the discrete derivative operation:

$$\Delta^q\mathbf{v}(i) := \Delta^{q-1}\mathbf{v}(i+1) - \Delta^{q-1}\mathbf{v}(i), \tag{2}$$

where $\Delta^q\mathbf{v} \in \mathbb{R}^{D_q}$ represents the D_q-dimensional derivative feature vectors calculated through the q-order discrete derivative operation ($q \in \{1, 2, ..., D-1\}$). Since the last element of the q-order derivative features (*i.e.*, the D_q-th element) has no subsequent element to subtract from, the length of the q-order derivative vector is one less than the $(q-1)$-order derivative vector, *i.e.*, $D_q = D_{q-1} - 1$. $\Delta^q\mathbf{v}(i)$ means the i-th element of the vector $\Delta^q\mathbf{v}$. The values in original features $\mathbf{v}$ will be normalized so that $||\mathbf{v}||_1 = 1$, where $||\cdot||_1$ is L1 norm. When the order of the discrete derivative operation is zero, we have: $\Delta^0\mathbf{v} = \mathbf{v}$. Rather than the spatial dimension of features, our proposed derivative operations are actually performed along the channel dimension. Different feature channels typically represent different visual patterns or semantics, *e.g.*, texture, shape, class. By analyzing the relative changes between adjacent channels, our DLP reflects the variation across these different semantics. The physical interpretation of this process is: the feature vector of an image pixel represents its coordinates in a feature space, and the similarity between two feature vectors measures the consistency between the orientation of two pixels in this feature space. Hence, our DLP, which addresses the ill-posed problem according to the following theorem, can complement the spatial information modeled by CNN and vanilla label propagation operation.

Theorem 1 (Well-Posedness)**.** *Assuming the feature vectors are L1-normalized, i.e., $||\mathbf{v}||_1 = 1$, there exists a unique solution for the pixel-wise feature vectors $\mathbf{v}_i$ ($i := 1, 2, ..., M$), only if: 1) $\mathcal{S}(\mathbf{v}_i, \mathbf{v}_j) = s_{i,j}$, and 2) $\mathcal{S}(\Delta^q\mathbf{v}_i, \Delta^q\mathbf{v}_j) = s^q_{i,j}, \forall q \in \{1, 2, ..., D-1\}$. Among them, $s_{i,j}$ and $s^q_{i,j}$ are the supervisions on the similarities with respect to original feature vectors and q-order derivative feature vectors, respectively.*

Unfortunately, $\forall q \in \{1, 2, ..., D-1\}$, sequentially performing the q-order discrete derivative operation is computationally expensive. The computational cost for regularizing the derivatives from 1-order to q-order together with original non-derivative features, increases linearly as q, since it requires computing q+1 similarity matrices. Hence, we employ a sparsity term for regularizing the high-order ($q > 1$) derivative-based features:

$$\begin{aligned}\mathcal{L}^{Der} :=& ||\mathbf{S} - \mathbf{S}^{GT}||_1 + ||\Delta^1\mathbf{S} - \Delta^1\mathbf{S}^{GT}||_1 \\ &+\eta||\Delta^2\mathbf{V}||_1,\end{aligned} \tag{3}$$

where η is a balancing weight. Given $||\mathbf{v}_i||_1 = 1$, the cosine similarity satisfies: $\mathcal{S}(\mathbf{v}_i, \mathbf{v}_j) := \mathbf{v}_i^\top\mathbf{v}_j / ||\mathbf{v}_i||_1||\mathbf{v}_j||_1 = \mathbf{v}_i^\top\mathbf{v}_j$. Thus, the i-th row and j-column in $\mathbf{S} := \mathbf{V}^\top\mathbf{V} \in \mathbb{R}^{M\times M}$ and $\Delta^1\mathbf{S} := (\Delta^1\mathbf{V})^\top(\Delta^1\mathbf{V}) \in \mathbb{R}^{M\times M}$, are $\mathcal{S}(\mathbf{v}_i, \mathbf{v}_j)$ and $\mathcal{S}(\Delta^1\mathbf{v}_i, \Delta^1\mathbf{v}_j)$, respectively. As shown in Fig. 2, the feature $\mathbf{V}$ is calculated by: $\mathbf{V} := \mathcal{J}(\mathbf{F}4)$, where $\mathcal{J}(\cdot)$ is the projection layer (the orange rectangle in Fig. 2), and $\mathbf{F}4$ means the features generated by the 4-th block of our backbone, *i.e.*, ResNet101, with image $\mathbf{X}$ as input. The i-th column in $\mathbf{V} \in \mathbb{R}^{D\times M}$, $\Delta^1\mathbf{V} \in \mathbb{R}^{(D-1)\times M}$ and $\Delta^2\mathbf{V} \in \mathbb{R}^{(D-2)\times M}$, are $\mathbf{v}_i \in \mathbb{R}^D$, $\Delta^1\mathbf{v}_i \in \mathbb{R}^{D-1}$, and $\Delta^2\mathbf{v}_i \in \mathbb{R}^{D-2}$, respectively. $\mathbf{S}^{GT}$ and $\Delta^1\mathbf{S}^{GT}$ are the ground truth similarity matrices, the i-th row and j-th column of which are $s_{i,j}$ and $s^1_{i,j}$, respectively.

We can infer from **Theorem** 2 that, minimizing $||\Delta^2\mathbf{V}||_1$, is equivalent to minimizing $||\Delta^q\mathbf{S} - \mathbf{0}||_1$ since $(2^{q-2}||\Delta^2\mathbf{V}||_1)^2$ is always greater than $||\Delta^q\mathbf{S} - \mathbf{0}||_1$, where $\mathbf{0}$ is an all-zero matrix.

Theorem 2 (Boundness)**.** *Assuming feature vectors are bounded, i.e., $||\mathbf{v}||_1 \le 1$, $\forall q \in \{2, ..., D-1\}$, the similarity matrix $\Delta^q\mathbf{S}$ satisfies: $||\Delta^q\mathbf{S} - \mathbf{0}||_1 \le (2^{q-2}||\Delta^2\mathbf{V}||_1)^2$.*

Please refer to our extended version for the proof of **Theorem** 1 and **Theorem** 2. Since the similarities with respect to high-order ($q > 1$) derivative-based features are enforced to be 0 for all pixels, their functionality in indicating the pixel-level semantic similarities, is suppressed. The key role of our sparsity regularization on these high-order derivative-based features, is to address the ill-posedness according to **Theorem** 1. We emphasize that, rather than theoretically solving the weakness of cosine similarity itself, we aim at mitigating the ill-posedness by restricting the solution space with

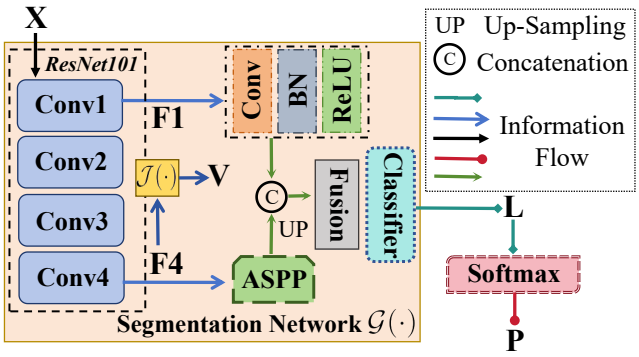


Figure 2: Network architecture of $\mathcal{G}(\cdot)$. $\mathbf{F}1$ and $\mathbf{F}4$ are the features generated by the 1-st and 4-th block of ResNet101, with image $\mathbf{X}$ as input. $\mathbf{L}$ and $\mathbf{P}$ denote the class logits and final segmentation results, respectively.

derivative operations. We do not rely on original or derivative features alone, but using them as joint constraints instead. Thus, the probability that different pairs of original or derivative feature vectors satisfy the same cosine similarities is evidently lower. Even the ill-posedness is theoretically associated with cosine similarity itself, our strategy can compensate for the intrinsic limitation of cosine similarity, thereby alleviating the ill-posedness effectively in practice. Hence, we perform the label propagation for rectifying the pseudo-labels of those unlabeled samples:

$$\widetilde{\mathbf{L}}^w_{nu} := \mathbf{L}^w_{nu}(\mathbf{S}^w_{nu} + \Delta^1\mathbf{S}^w_{nu}), \tag{4}$$

where $\mathbf{S}^w_{nu} := \mathbf{V}^{w\top}_{nu}\mathbf{V}^w_{nu}$, and $\mathbf{V}^w_{nu}$ is projected by: $\mathbf{V}^w_{nu} := \mathcal{J}(\mathbf{F}4^w_{nu})$ corresponding to the input $\mathbf{X}^w_{nu}$. The pseudo-labels can thus be obtained by:

$$\overline{\mathbf{P}}_{nu} := (1-\eta_{ep})\mathbf{P}^w_{nu} + \eta_{ep}\widetilde{\mathbf{P}}^w_{nu}, \tag{5}$$

where $\mathbf{P}^w_{nu} := \text{Softmax}(\mathbf{L}^w_{nu})$, and $\widetilde{\mathbf{P}}^w_{nu} := \text{Softmax}(\widetilde{\mathbf{L}}^w_{nu})$. η_{ep} ramps up as ep/EP to address the poor performance of $\widetilde{\mathbf{P}}^w_{nu}$ at the early training period. Note that beyond the standard backbone pre-training (pre-trained ResNet101 (He et al. 2016) in this paper), we do not require any extra pre-training phase. As defined in Eq. (5), the weight η_{ep} is used to blend the original and rectified pseudo-labels to adjust the influence of generated pseudo-labels, which implicitly acts as a built-in pre-training. Compared to the vanilla label propagation, our DLP makes the learning process more stable and less susceptible to spurious correlations caused by noises or data distribution shifts. The data augmentations for unlabeled images can be viewed as noise injection, and the consistent performance gains in Tabs. 1 and 2 across all labeled data proportions demonstrate our capability to improve pseudo-labels in the presence of noises.

Network Architecture

As illustrated in Fig. 2, our segmentation network adopts the pre-trained ResNet101 as backbone. Following previous works (Sohn et al. 2020; Yang et al. 2023; Sun et al. 2024; Wang et al. 2021; Tu et al. 2024b; Zhang et al. 2025; Tu et al. 2025a), for $\mathcal{G}(\cdot)$, the features from the 1-st and 4-th block of ResNet101 are fed into a Conv-BN-ReLU layer and ASPP module (Chen et al. 2017), respectively. The 1-st and 4-th blocks are complementary, exploring the low-level details and high-level semantics, respectively. In contrast, intermediate network layers exhibit feature redundancy with adjacent layers besides increasing computational costs, failing to substantially improve accuracy. Thus, we solely fuse the 1-st and 4-th blocks to balance accuracy and efficiency. Then, the fused features are processed by a classifier to output the class logits, *e.g.*, $\mathbf{L}_{nl}$, $\mathbf{L}^w_{nu}$, or $\mathbf{L}^s_{nu}$ are output by $\mathcal{G}(\mathbf{X}_{nl})$, $\mathcal{G}(\mathbf{X}^w_{nu})$, and $\mathcal{G}(\mathbf{X}^s_{nu})$, corresponding to inputs $\mathbf{X}_{nl}$, $\mathbf{X}^w_{nu}$, or $\mathbf{X}^s_{nu}$, respectively.

Drawing inspiration from ensemble learning principles (Krogh and Vedelsby 1994; Ueda and Nakano 1996; Breiman 2001; Wang et al. 2023; Bai et al. 2024), biases/interferences of a single training moment can be compensated by aggregating complementary information from multiple diverse sources. Hence, we aggregate knowledge across different training moments to yield better results than those derived from a single training moment. Specifically, we maintain a momentum network $\mathcal{G}^m(\cdot)$ corresponding to $\mathcal{G}(\cdot)$ that is trained by back-propagation. The parameters of our momentum network are updated by accumulating the parameters of its corresponding back-propagation network:

$$\mathbf{\Theta}^m_{(ep)} := (\mathbf{\Theta}^m_{(ep-1)} + \mathbf{\Theta}^b_{(ep)})/2, \tag{6}$$

where $\mathbf{\Theta}^m_{(ep)}$ and $\mathbf{\Theta}^b_{(ep)}$ are the parameters of the momentum network $\mathcal{G}^m(\cdot)$ and back-propagation network $\mathcal{G}(\cdot)$, after the ep-th epoch, respectively. In training, the pseudo-labels are generated by the back-propagation network. In testing, only the momentum network $\mathcal{G}^m(\cdot)$ is executed, while the back-propagation network $\mathcal{G}(\cdot)$ is abandoned.

Loss Function

Our overall loss $\mathcal{L}$ consists of: cross entropy loss $\mathcal{L}^{CE}$, KL divergence loss $\mathcal{L}^{KL}$, and derivative loss $\mathcal{L}^{Der}$:

$$\mathcal{L} := \lambda^{CE}\mathcal{L}^{CE} + \lambda^{KL}\mathcal{L}^{KL} + \lambda^{Der}\mathcal{L}^{Der}, \tag{7}$$

where λ^{CE}, λ^{KL}, and λ^{Der} are balancing weights.

Specifically, denote by l^{CE} the pixel-wise cross-entropy loss function, we have: $\mathcal{L}^{CE} := \sum_{nl:=1}^{N_l} l^{CE}(\mathbf{P}_{nl}, \mathbf{Y}_{nl}) + \sum_{nl:=1}^{N_l} l^{CE}(\widetilde{\mathbf{P}}_{nl}, \mathbf{Y}_{nl}) + [\sum_{nu:=1}^{N_u} l^{CE}(\mathbf{P}^s_{nu}, \overline{\mathbf{P}}_{nu}) + \sum_{nu:=1}^{N_u} l^{CE}(\widetilde{\mathbf{P}}^s_{nu}, \overline{\mathbf{P}}_{nu})]/2$. Among them, $\mathbf{P}_{nl} := \text{Softmax}(\mathbf{L}_{nl})$, and $\mathbf{P}^s_{nu} := \text{Softmax}(\mathbf{L}^s_{nu})$, with $\mathbf{L}_{nl} := \mathcal{G}(\mathbf{X}_{nl})$, and $\mathbf{L}^s_{nu} := \mathcal{G}(\mathbf{X}^s_{nu})$. $\widetilde{\mathbf{P}}_{nl}$ and $\widetilde{\mathbf{P}}^s_{nu}$ are obtained by: $\text{Softmax}(\mathbf{L}_{nl}(\mathbf{S}_{nl} + \Delta^1\mathbf{S}_{nl}))$ and $\text{Softmax}(\mathbf{L}^s_{nu}(\mathbf{S}^s_{nu} + \Delta^1\mathbf{S}^s_{nu}))$, respectively. Denote by $\mathbf{V}_{nl}$ and $\mathbf{V}^s_{nu}$ the features projected by $\mathcal{J}(\cdot)$ respectively corresponding to inputs $\mathbf{X}_{nl}$ and $\mathbf{X}^s_{nu}$, we have: $\mathbf{S}_{nl} := \mathbf{V}^\top_{nl}\mathbf{V}_{nl}$ and $\mathbf{S}^s_{nu} := \mathbf{V}^{s\top}_{nu}\mathbf{V}^s_{nu}$. Different from $\mathcal{L}^{CE}$, the KL divergence loss is used only for unlabeled samples: $\mathcal{L}^{KL} := \sum_{nu:=1}^{N_u} \text{KL}(\mathbf{P}^s_{nu}, \overline{\mathbf{P}}_{nu})$, where $\text{KL}(\cdot,\cdot)$ is Kullback-Leibler divergence function. Following previous SSSS methods (Yang et al. 2023; Sun et al. 2024), a binary map selecting high-confident pixels in pseudo-labels, is introduced for both $\mathcal{L}^{CE}$ and $\mathcal{L}^{KL}$.

For the derivative loss, we create the ground truth similarity through: $\mathbf{S}_{nl}^{GT} := \mathbf{Y}_{nl}^{\top}\mathbf{Y}_{nl}$, $\mathbf{S}_{nu}^{GT} := \mathbf{V}_{nu}^{w\top}\mathbf{V}_{nu}^{w}$, and $\Delta^1\mathbf{S}_{nu}^{GT} := (\Delta^1\mathbf{V}_{nu}^{w})^{\top}(\Delta^1\mathbf{V}_{nu}^{w})$, where $\mathbf{Y}_{nl} \in \mathbb{R}^{C\times M_{nl}}$ is the manually annotated one-hot segmentation label, and $\mathbf{V}_{nu}^{w} := \mathcal{J}(\mathbf{F}4_{nu}^{w})$ is the projected features corresponding to $\mathbf{X}_{nu}^{w}$. Thus, the derivative loss is formulated as:

$$\begin{aligned}\mathcal{L}^{Der} := \sum_{nl:=1}^{N_l} (||\mathbf{S}_{nl} - \mathbf{S}_{nl}^{GT}||_1 + ||\Delta^1\mathbf{S}_{nl} - \mathbf{S}_{nl}^{GT}||_1 \\ +\eta||\Delta^2\mathbf{V}_{nl}||_1) + \sum_{nu:=1}^{N_u} (||\mathbf{S}_{nu}^{s} - \mathbf{S}_{nu}^{GT}||_1 + \\ ||\Delta^1\mathbf{S}_{nu}^{s} - \Delta^1\mathbf{S}_{nu}^{GT}||_1 + \eta||\Delta^2\mathbf{V}_{nu}^{s}||_1).\end{aligned} \tag{8}$$

Experimental Validation

Implementation Details

We implement our network using PyTorch library on two V100 GPUs. In detail, the segmentation network $\mathcal{G}(\cdot)$ is randomly initialized and trained for 80 epochs using SGD optimizer, the input size of which is 321×321. λ^{CE}, λ^{KL}, and λ^{Der} are set to be 0.5, 0.25, and 0.5, respectively. η is set to be 0.5. The learning rate, momentum and weight decay of the SGD optimizer are set to 0.001, 0.9 and 0.0001, respectively. The batch size is set to be 4, and the total number of training epochs is 80.

Datasets & Evaluation Metrics

Two datasets are employed in this paper, which include: 1) the commonly used Pascal VOC 2012 (Mottaghi et al. 2014) having 1,464 training samples with 21 categories, and 1,449 testing samples, and 2) the urban scene understanding dataset Cityscapes (Cordts et al. 2016) consisting of 2,975 training images and 500 validating images.

As for the evaluation metrics, we report mean Intersection-over-Union (mIoU) on the Pascal VOC 2012 validation set using original images, consistent with prior work (Chen et al. 2021; French et al. 2020; Liu et al. 2022; Sun et al. 2024). For Cityscapes, following (Chen et al. 2021; Wang et al. 2022b; Yang et al. 2023), we evaluate via sliding window with fixed-size crops and compute mIoU on these crops. All results are obtained by using single-scale inference on the standard validation set.

Ablation Study

To assess the influences of our proposed components, we compare our method with the following alternatives: 1) **w/o DLP**. In this setting, we train the segmentation network without using our proposed derivative label propagation. The predictions from the weakly augmented inputs are taken as the pseudo-labels for supervising the predictions from the corresponding strongly augmented images, without being rectified through Eq. (5). The derivative loss $\mathcal{L}^{Der}$ formulated in Eq. (8) is also omitted; 2) **w/o** $\mathcal{L}^{Der}$. For this alternative, the pseudo-labels are rectified through Eq. (5), but the loss $\mathcal{L}^{Der}$ for regularizing the similarity matrix is abandoned; 3) **w/o momentum**. We do not maintain any momentum network. In testing, the back-propagation network $\mathcal{G}(\cdot)$ trained after the last epoch, is executed. In this setting, the pseudo-labels are rectified through Eq. (5), and the loss $\mathcal{L}^{Der}$ is introduced. It is worth mentioning that, for both **w/o DLP** and **w/o** $\mathcal{L}^{Der}$, the momentum network is not used. For fair comparison, the performance of all the alternatives are evaluated using the same random seed, and the checkpoints saved at the same epoch. The network architectures and all other hyper-parameters are also kept to be the same. Note that under the fully supervised scenario with all samples being labeled, *i.e.*, **Full (1464)**, no pseudo-labels are required in training. Consequently, the **w/o DLP** variant is not applicable and its results are not reported for this data setting.

As given in Tab. 1 and Tab. 2, omitting our proposed derivative label propagation induces significant performance degradation across all labeled data proportions, *i.e.*, 1/16(92), 1/8(183), 1/4(366), and 1/2(732). This underscores DLP's efficacy in enhancing segmentation accuracy by rectifying pseudo-labels. Moreover, abandoning the derivative loss $\mathcal{L}^{Der}$ consistently reduces accuracy, confirming the necessity of this loss function for effective regularization. We also evaluate the performance of eliminating the momentum network in testing, which yields measurable performance deterioration. It highlights the role in maintaining satisfactory segmentation accuracy.

Fig. 3 provides visual results to reveal that our method exhibits precise boundary delineation and structural integrity, particularly for small objects.

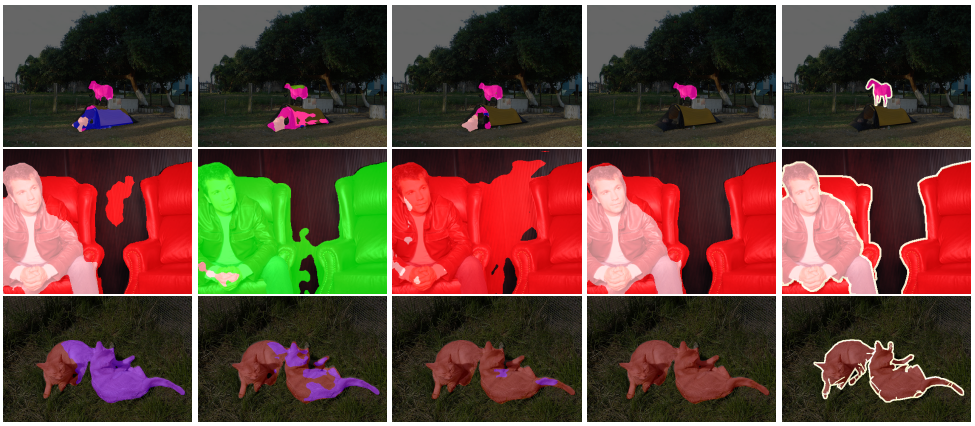
Comparisons with other State-of-the-arts

We compare with recent semi-supervised methods including: UniMatch (Yang et al. 2023), AllSpark (Wang et al. 2024), RankMatch (Mai et al. 2024), CorrMatch (Sun et al. 2024), MGCT (Hu et al. 2025), RCC (Mai, Sun, and Wu 2025), ScaleMatch (Lv and Zhang 2025). Note that previous semi-supervised semantic segmentation methods use an input size of 801×801 on the Cityscapes dataset, while our method uses a smaller size of 321×321. Hence, for fair comparison, we re-train other competitors, *i.e.*, UniMatch, CorrMatch, and ScaleMatch, with 321×321 inputs.

As given in Tab. 1, our method achieves state-of-the-art performance across all labeled data proportions, *i.e.*, 1/16(92), 1/8(183), 1/4(366), and 1/2(732), surpassing recent competitors by significant margins. This confirms that our proposed DLP remains effective with different number of labeled and unlabeled samples. For example, DerProp outperforms CorrMatch (76.4%) and ScaleMatch (76.1%) by around 1%, which demonstrates our exceptional robustness to extreme label scarcity. While methods like AllSpark use larger input size, *i.e.*, 513×513, and larger amount of parameters, our DerProp still achieves superior accuracy with smaller 321×321 inputs without introducing additional parameters, highlighting its computational efficiency. As for the Cityscapes dataset, it can be inferred from Tab. 2 that, our DerProp consistently dominates all settings, achieving 62.1% mIoU at 1/4 labeled data (744 images), surpassing ScaleMatch (61.6%) and UniMatch (60.8%). Notably, our method outperforms CorrMatch, which also uses correlation maps for label propagation. It demonstrates that the derivative operations act as effective constraints for restricting the

Method	Year	Size	Param.	1/16(92)	1/8(183)	1/4(366)	1/2(732)	Full(1464)
UniMatch	CVPR'23	321^2	59.5M	75.2	77.2	78.8	79.9	81.2
DDFP	CVPR'24	513^2	59.5M	74.9	78.0	79.5	81.2	81.9
AllSpark	CVPR'24	513^2	89.3M	76.0	78.4	79.7	80.7	82.1
RankMatch	CVPR'24	513^2	-	75.5	77.6	79.8	80.7	**82.2**
CorrMatch	CVPR'24	321^2	59.5M	76.4	78.5	79.4	80.6	81.8
MGCT	TMM'25	321^2	-	75.2	76.7	76.9	-	-
RCC	AAAI'25	513^2	-	75.3	77.9	79.8	81.0	82.1
ScaleMatch	AAAI'25	321^2	-	76.1	78.6	79.6	80.7	81.8
w/o DLP	-	321^2	59.5M	66.6	77.8	78.3	79.8	-
w/o $\mathcal{L}^{Der}$	-	321^2	59.5M	72.9	77.0	77.5	78.8	81.0
w/o momentum	-	321^2	59.5M	75.1	78.0	79.6	78.8	81.9
DerProp (Ours)	-	321^2	59.5M	**77.6**	**78.7**	**80.5**	**81.3**	**82.2**

Table 1: Results on Pascal VOC 2012 in terms of mIoU (%), with ResNet101 as the backbone. Best results are **Bolded**.



w/o DLP | w/o $\mathcal{L}^{Der}$ | w/o momentum | Ours | GT

Figure 3: Visual results of ablation studies on Pascal VOC 2012.

solution space. Moreover, as shown in Fig. 4, DerProp eliminates undesired misclassifications by CorrMatch.

Discussion on the Derivative Losses

For $Q \in \{0, 1, ..., D-1\}$, we give the general form of derivative loss as:

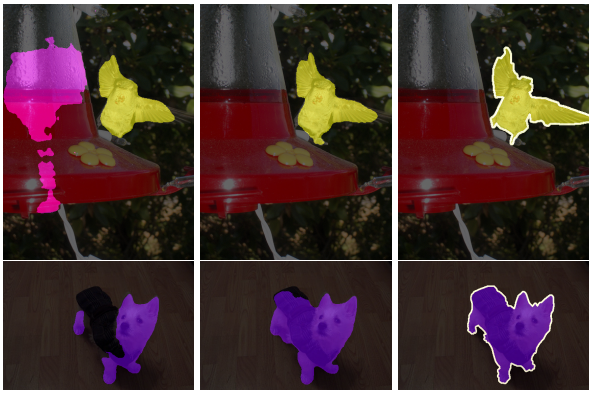
$$\begin{aligned}\mathcal{L}_Q^{Der} := &\sum_{nl:=1}^{N_l}[\sum_{q:=0}^{Q}||\Delta^q\mathbf{S}_{nl} - \mathbf{S}_{nl}^{GT}||_1 \\ &+\eta||\Delta^{Q+1}\mathbf{V}_{nl}||_1] \\ &+\sum_{nu:=1}^{N_u}[\sum_{q:=0}^{Q}||\Delta^q\mathbf{S}_{nu}^s - \Delta^q\mathbf{S}_{nu}^{GT}||_1 \\ &+\eta||\Delta^{Q+1}\mathbf{V}_{nu}^s||_1],\end{aligned} \tag{9}$$

where $\Delta^q\mathbf{S}_{nu}^{GT} := (\Delta^q\mathbf{V}_{nu}^w)^\top(\Delta^q\mathbf{V}_{nu}^w) \in \mathbb{R}^{M\times M}$, $\forall q \in \{1, 2, ..., D-1\}$, and $\Delta^q\mathbf{V}_{nu}^w$ is the q-order derivative features projected by $\mathcal{J}(\cdot)$ corresponding to the nu-th weakly augmented image $\mathbf{X}_{nu}^w$. $\Delta^q\mathbf{S}_{nu}^s$ is calculated by $(\Delta^q\mathbf{V}_{nu}^s)^\top(\Delta^q\mathbf{V}_{nu}^s)$, and $\Delta^q\mathbf{V}_{nu}^s$ is the q-order derivative features corresponding to the nu-th strongly augmented image $\mathbf{X}_{nu}^s$. When $Q := 1$, the general form of derivative loss in Eq. (9), say $\mathcal{L}_1^{Der}$, is consistent with the loss in Eq. (8). Based on this, we can evaluate the results of altering the supervisions on similarity matrices with respect to derivative features. All the alternatives also use the same random seed, and the checkpoints saved at the same epoch.

As given in Tab. 3, the impact of varying the parameter Q in Eq. (9) on segmentation accuracy is evident. A clear trend emerges: increasing Q results in a significant degradation in performance. Consequently, $\mathcal{L}_1^{Der}$ ($Q = 1$) demonstrates superior effectiveness compared to other alternatives with higher Q values. Furthermore, we observe a consistent accuracy drop across all evaluated derivative losses, *i.e.*, $\mathcal{L}_0^{Der}$, $\mathcal{L}_1^{Der}$, $\mathcal{L}_2^{Der}$, and $\mathcal{L}_3^{Der}$, when the sparsity regularization terms, *i.e.*, $\eta||\Delta^{Q+1}\mathbf{V}_{nl}||_1$ and $\eta||\Delta^{Q+1}\mathbf{V}_{nu}^s||_1$ in Eq. (9), are omitted (see the results of **w/o sparsity** in Tab. 3). This pronounced decline validates the efficacy of the proposed sparsity regularization in constraining the solution space and enhancing segmentation accuracy.

Method	Year	Size	1/16(186)	1/8(372)	1/4(744)
UniMatch	CVPR'23	321^2	58.6	58.7	60.8
CorrMatch	CVPR'24	321^2	59.3	59.9	60.2
ScaleMatch	AAAI'25	321^2	59.7	60.5	61.6
w/o DLP	-	321^2	59.1	60.4	60.9
w/o $\mathcal{L}^{Der}$	-	321^2	57.1	60.5	61.0
w/o momentum	-	321^2	57.6	60.4	61.6
DerProp (Ours)	-	321^2	**60.1**	**61.8**	**62.1**

Table 2: Results on Cityscapes in terms of mIoU (%), with ResNet101 as the backbone. Best results are **Bolded**.



CorrMatch Ours GT

Figure 4: Visual comparisons on Pascal VOC 2012.

Method	w/o sparsity	w/ sparsity
$\mathcal{L}_0^{Der}$	74.1	75.1
$\mathcal{L}_1^{Der}$ (Ours)	**77.0**	**77.6**
$\mathcal{L}_2^{Der}$	76.0	76.9
$\mathcal{L}_3^{Der}$	76.4	76.7

Table 3: Results of different derivative losses on Pascal VOC 2012 in terms of mIoU (%). All the results in this table are obtained under the setting of 1/16 (92), *i.e.*, the amount of labeled images is 92, which accounts for 1/16 of total training samples. Best results are **Bolded**.

Method	1/16 (92)	1/8 (183)
Eq. (10)	76.1	78.6
Eq. (11)	76.4	77.1
Eq. (12)	76.8	78.5
Eq. (2) (Ours)	**77.6**	**78.7**

Table 4: Results of different derivative operations on Pascal VOC 2012 in terms of mIoU (%). Best results are **Bolded**.

Discussion on the Derivative Operations

In this section, we evaluate the results of replacing our proposed derivative operation formulated in Eq. (2) with the following three different operations:

$$\Delta^q \mathbf{v}(i) := (\Delta^{q-1}\mathbf{v}(i+2) - \Delta^{q-1}\mathbf{v}(i))/2, \tag{10}$$

$$\Delta^q \mathbf{v}(i) := \Delta^{q-1}\mathbf{v}(i+1) + \Delta^{q-1}\mathbf{v}(i), \tag{11}$$

$$\begin{aligned}\Delta^q \mathbf{v}(i) :=& \Delta^{q-1}\mathbf{v}(i+1) + \\ & \Delta^{q-1}\mathbf{v}(i-1) - 2\Delta^{q-1}\mathbf{v}(i).\end{aligned} \tag{12}$$

As given in Tab. 4, our proposed derivative operation in Eq. 2 consistently outperforms all the other alternatives. To be more specific, Eq. (10) achieves suboptimal results compared with our proposed derivative operation, due to its sensitivity to feature noises. Compared with our method, the wider receptive field (from the i-th to $(i+2)$-th element of features vectors) of Eq. (10) introduces instability when capturing high-frequency semantic transitions. The summation operation formulated in Eq. (11) performs the worst among all the alternatives, as it fails to capture feature variations. It solely accumulates adjacent feature values, contradicting the derivative's purpose of modeling semantic differences between adjacent elements of feature vectors. Moreover, Eq. (12) shows marginal improvement over Eq. (11), but remains inferior to Eq. (10). Its operation over-constrains the solution space, suppressing subtle class boundaries. To ensure the fairness, we also use the checkpoints saved at the same epoch for all the alternatives.

Generally, our derivative operation in Eq. (10), optimally balances noise robustness and discriminative capacity. According to **Theorem** 1, our subtraction calculation between adjacent elements, effectively preserves the semantics of pixels while mitigating the ill-posedness.

Conclusion

In this work, we proposed DerProp, a novel semi-supervised semantic segmentation framework that addresses the challenge of unreliable pseudo-labels through derivative label propagation (DLP). We impose discrete derivative operations on pixel-wise features, which introduces additional regularization on similarity metrics. As a profit, our DLP effectively alleviates the ill-posed problem, thereby alleviating the ill-posedness. Extensive experiments were conducted on Pascal VOC 2012 and Cityscapes dataset to demonstrate the state-of-the-art performance of our DLP.

References

Badrinarayanan, V.; Handa, A.; and Cipolla, R. 2015. Segnet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv: 1505.07293.

Bai, Y.; Zhao, H.; Lin, Z.; Kale, A.; Gu, J.; Yu, T.; Kim, S.; and Fu, Y. 2024. Advancing vision-language models with adapter ensemble strategies. In *EMNLP*, 15702–15720.

Breiman, L. 2001. Random Forests. *Mach. Learn.*, 45(1): 5–32.

Chen, L.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv: 1706.05587*.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, 2613–2622.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 3213–3223.

Dong, Y.; Wang, Y.; Fan, L.; Ding, X.; and Huang, Y. 2023. High-resolution feature representation driven infrared small-dim object detection. In *PRCV*, volume 14436, 315–327. Springer.

Fan, L.; Wang, Y.; Hu, G.; Li, F.; Dong, Y.; Zheng, H.; Lin, C.; Huang, Y.; and Ding, X. 2024. Diffusion-based continuous feature representation for infrared small-dim target detection. *TGRS*, 62: 1–17.

Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Gläser, C.; Timm, F.; Wiesbeck, W.; and Dietmayer, K. 2021. Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges. *TITS*, 22(3): 1341–1360.

French, G.; Laine, S.; Aila, T.; Mackiewicz, M.; and Finlayson, G. D. 2020. Semi-supervised semantic segmentation needs strong, varied perturbations. In *BMVC*.

Han, Y.; Guo, Q.; Pan, L.; Liu, L.; Guan, Y.; and Yang, M. 2025a. Dynfocus: dynamic cooperative network empowers llms with video understanding. In *CVPR*, 8512–8522.

Han, Y.; Wang, H.; Hu, Y.; Gong, Y.; Song, X.; and Guan, W. 2025b. Content-aware balanced spectrum encoding in masked modeling for time series classification. In *AAAI*, 17059–17067.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hu, H.; Wei, F.; Hu, H.; Ye, Q.; Cui, J.; and Wang, L. 2021. Semi-supervised semantic segmentation via adaptive equalization learning. In *NeurIPS*, 22106–22118.

Hu, K.; Chen, X.; Chen, Z.; Zhang, Y.; and Gao, X. 2025. Multi-perspective pseudo-label generation and confidence-weighted training for semi-supervised semantic segmentation. *TMM*, 27: 300–311.

Jin, Y.; Wang, J.; and Lin, D. 2022. Semi-supervised semantic segmentation via gentle teaching assistant. In *NeurIPS*.

Ke, R.; Avilés-Rivero, A. I.; Pandey, S.; Reddy, S.; and Schönlieb, C. 2022. A three-stage self-training framework for semi-supervised semantic segmentation. *TIP*, 31: 1805–1815.

Krogh, A.; and Vedelsby, J. 1994. Neural network ensembles, cross validation, and active learning. In *NIPS*, 231–238.

Li, K.; Yu, L.; Wang, S.; and Heng, P. 2020. Towards cross-modality medical image segmentation with online mutual knowledge distillation. In *AAAI*, 775–783.

Li, M.; Zhang, Y.; Ma, X.; Qu, Y.; and Fu, Y. 2023. Bev-dg: cross-modal learning under bird's-eye view for domain generalization of 3d semantic segmentation. In *ICCV*, 11598–11608.

Lin, J.; Wu, Y.; Wang, Z.; Liu, X.; and Guo, Y. 2024a. Pair-id: a dual modal framework for identity preserving image generation. *ISPL*, 31: 2715–2719.

Lin, J.; Zhao, G.; Xu, J.; Wang, G.; Wang, Z.; Dantcheva, A.; Du, L.; and Chen, C. 2024b. Difftv: identity-preserved thermal-to-visible face translation via feature alignment and dual-stage conditions. In *ACM MM*, 10930–10938.

Liu, Y.; Tian, Y.; Chen, Y.; Liu, F.; Belagiannis, V.; and Carneiro, G. 2022. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *CVPR*, 4248–4257. IEEE.

Lu, J.; Wang, H.; Xu, Y.; Wang, Y.; Yang, K.; and Fu, Y. 2025. Representation potentials of foundation models for multimodal alignment: a survey. *arXiv preprint arXiv:2510.05184*.

Lv, L.; and Zhang, L. 2025. Scalematch: multi-scale consistency enhancement for semi-supervised semantic segmentation. In *AAAI*, 5910–5918.

Mai, H.; Sun, R.; and Wu, F. 2025. Relaxed class-consensus consistency for semi-supervised semantic segmentation. In *AAAI*, 6045–6053.

Mai, H.; Sun, R.; Zhang, T.; and Wu, F. 2024. Rankmatch: exploring the better consistency regularization for semi-supervised semantic segmentation. In *CVPR*, 3391–3401.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 891–898.

Papadopoulos, D. P.; Weber, E.; and Torralba, A. 2021. Scaling up instance annotation via label propagation. In *ICCV*, 15344–15353.

Qu, Z.; Jin, H.; Zhou, Y.; Yang, Z.; and Zhang, W. 2021. Focus on local: detecting lane marker from bottom up via key point. In *CVPR*, 14122–14130.

Sanderson, E.; and Matuszewski, B. J. 2022. Fcn-transformer feature fusion for polyp segmentation. In *MIUA*, volume 13413, 892–907.

Shelhamer, E.; Long, J.; and Darrell, T. 2017. Fully convolutional networks for semantic segmentation. *TPAMI*, 39(4): 640–651.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.; Cubuk, E. D.; Kurakin, A.; and Li, C. 2020. Fixmatch: simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*.

Stojnic, V.; Kalantidis, Y.; Matas, J.; and Tolias, G. 2025. Lposs: label propagation over patches and pixels for open-vocabulary semantic segmentation. In *CVPR*, 9794–9803.

Sun, B.; Yang, Y.; Zhang, L.; Cheng, M.; and Hou, Q. 2024. Corrmatch: label propagation via correlation matching for semi-supervised semantic segmentation. In *CVPR*, 3097–3107.

Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; and Wang, J. 2019. High-resolution representations for labeling pixels and regions. arXiv preprint arXiv: 1904.04514.

Tu, R.-C.; Ji, Y.; Jiang, J.; Kong, W.; Cai, C.; Zhao, W.; Wang, H.; Yang, Y.; and Liu, W. 2025a. Global and local semantic completion learning for vision-language pre-training. *TPAMI*.

Tu, R.-C.; Ma, Z.-A.; Lan, T.; Zhao, Y.; Huang, H.; and Mao, X.-L. 2024a. Automatic evaluation for text-to-image generation: Task-decomposed framework, distilled training, and meta-evaluation benchmark. *arXiv preprint arXiv:2411.15488*.

Tu, R.-C.; Mao, X.-L.; Liu, J.-Y.; Ma, Z.-A.; Lan, T.; and Huang, H. 2025b. Prospective layout-guided multi-modal online hashing. *TIP*, 34: 5935–5947.

Tu, R.-C.; Sun, W.; Jin, Z.; Liao, J.; Huang, J.; and Tao, D. 2024b. Spagent: Adaptive task decomposition and model selection for general video generation and editing. *arXiv preprint arXiv:2411.18983*.

Tzelepi, M.; and Tefas, A. 2021. Semantic scene segmentation for robotics applications. In *IISA*, 1–4.

Ueda, N.; and Nakano, R. 1996. Generalization error of ensemble estimators. In *ICNN*, 90–95.

Wang, H.; Li, W.; Xi, Y.; Hu, J.; Chen, H.; Li, L.; and Wang, Y. 2023. Ift: image fusion transformer for ghost-free high dynamic range imaging. *arXiv preprint arXiv:2309.15019*.

Wang, H.; Lu, J.; Zhang, Y.; and Fu, Y. 2025. Outlier-aware post-training quantization for image super-resolution. In *ICCV*, 16175–16184.

Wang, H.; Tian, Q.; Li, L.; and Guo, X. 2021. Image demoiréing with a dual-domain distilling network. In *ICME*.

Wang, H.; Zhang, Q.; Li, Y.; and Li, X. 2024. Allspark: reborn labeled features from unlabeled in transformer for semi-supervised semantic segmentation. In *CVPR*, 3627–3636.

Wang, T.; Lu, J.; Lai, Z.; Wen, J.; and Kong, H. 2022a. Uncertainty-guided pixel contrastive learning for semi-supervised medical image segmentation. In *IJCAI*, 1444–1450.

Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; and Le, X. 2022b. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *CVPR*, 4238–4247.

Xing, Z.; Ye, T.; Yang, Y.; Liu, G.; and Zhu, L. 2024. Segmamba: long-range sequential modeling mamba for 3d medical image segmentation. In *MICCAI*, volume 15008, 578–588.

Xing, Z.; Yu, L.; Wan, L.; Han, T.; and Zhu, L. 2022. Nestedformer: nested modality-aware transformer for brain tumor segmentation. In *MICCAI*, volume 13435, 140–150.

Xu, H.; Liu, L.; Bian, Q.; and Yang, Z. 2022. Semi-supervised semantic segmentation with prototype-based consistency regularization. In *NeurIPS*.

Yang, L.; Qi, L.; Feng, L.; Zhang, W.; and Shi, Y. 2023. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *CVPR*, 7236–7246.

Yang, L.; Zhuo, W.; Qi, L.; Shi, Y.; and Gao, Y. 2022. St++: make self-training work better for semi-supervised semantic segmentation. In *CVPR*, 4258–4267.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: bilateral segmentation network for real-time semantic segmentation. In *ECCV*, volume 11217, 334–349.

Yuan, J.; Liu, Y.; Shen, C.; Wang, Z.; and Li, H. 2021. A simple baseline for semi-supervised semantic segmentation with strong data augmentation*. In *ICCV*, 8209–8218.

Zhang, H.; Dana, K. J.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; and Agrawal, A. 2018. Context encoding for semantic segmentation. In *CVPR*, 7151–7160.

Zhang, Y.; Deng, B.; Jia, K.; and Zhang, L. 2020. Label propagation with augmented anchors: a simple semi-supervised learning baseline for unsupervised domain adaptation. In *ECCV*, volume 12349, 781–797.

Zhang, Y.; Ma, X.; Bai, Y.; Wang, H.; and Fu, Y. 2025. Accessing vision foundation models via imagenet-1k. In *ICLR*.