

# SESC<sup>2</sup>: Seeking Essential Semantics in Changing Contexts for Semi-Supervised Segmentation

THIS supplementary material presents additional experimental results that are omitted from the main paper due to the space limit. In this supplementary material, we provide:

- **1.** A comparison between the proposed method and Transformer-based methods. (in Sec.-A);
- **2.** An analysis of whether the proposed method increases training difficulty. (in Sec.-B);
- **3.** A comparison of the proposed Dual-Cleaning Semantic Enhancement with similar methods. (in Sec.-C);
- **4.** Analysis about the Capacity of the Class Memory Unit. (in Sec.-D)
- **5.** Performance analysis of the proposed method and background attention map via CAM. (in Sec.-E)
- **6.** Feasibility analysis of VCA as a data augmentation strategy. (in Sec.-E);
- **7.** Performance validation of the proposed method on large model-based methods. (in Sec.-G);
- **8.** Hyperparameter analysis of the consistency regularization loss. (in Sec.-H);

## A. Comparison with Transformer-Based Methods.

For modeling the features of object to achieve robust reasoning, some works [1]–[3] rely on the adaptive capacity of self-attention. However, self-attention is highly sensitive to the training data distribution and tends to suffer from overfitting or abnormal context dependency, particularly under limited labeled data, making it difficult to learn stable semantic representations. To further evaluate the advantages of our method over transformer-based methods under minimal amount of labeled data, we conduct comparative experiments against several state-of-the-art self-attention-based SSSS methods (SemiCVT [2], AllSpark [1], and S<sup>4</sup>Former-Base [3]) on the VOC and Cityscapes datasets with 1/16 and 1/8 label splits. We adopt UniMatch with ResNet-101 as the baseline to integrate our strategy. For fair comparison, we also replace the original CNN backbone of UniMatch with the self-attention-based SegFormer to construct a self-attention variant of UniMatch (denoted as **UniMatch<sup>†</sup>**). As shown in Tab. I, our method consistently outperforms self-attention-based approaches under low-supervision settings, with particularly notable gains in the 1/16 split. Furthermore, compared to the SegFormer-based UniMatch variant, our method also achieves significant performance improvements. Furthermore, we integrate the proposed SESC<sup>2</sup> into UniMatch<sup>†</sup> to verify the performance improvement of the proposed method over the Transformer-based method. As shown in Tab. I, after the integration of SESC<sup>2</sup>, the performance of UniMatch<sup>†</sup> is significantly improved under the condition of small labeled data amount, for example, the mIoU of VOC 1/16 partition

TABLE I  
COMPARISON BETWEEN OUR METHOD AND TRANSFORMER-BASED METHODS UNDER EXTREMELY LIMITED LABELED DATA CONDITIONS (VOC 2012 AND CITYSCAPES: 1/16 AND 1/8 SPLITS).

| Method.Venues                        | Pascal VOC   |              | Cityscapes   |              |
|--------------------------------------|--------------|--------------|--------------|--------------|
|                                      | 1/16(92)     | 1/8(183)     | 1/16(186)    | 1/8(372)     |
| SemiCVT-[CVPR'23]                    | 68.56        | 71.26        | 72.19        | 75.41        |
| UniMatch <sup>†</sup> -[ICCV'23]     | 75.38        | 77.64        | 76.42        | 78.23        |
| AllSpark-[CVPR'24]                   | 76.07        | 78.40        | 78.33        | 79.24        |
| S <sup>4</sup> Former-Base-[CVPR'24] | 75.82        | 77.40        | 76.70        | 79.21        |
| <b>UniMatch + Ours</b>               | <b>77.34</b> | <b>78.55</b> | <b>78.58</b> | <b>79.55</b> |
| <b>UniMatch<sup>†</sup> + Ours</b>   | <b>77.88</b> | <b>79.13</b> | <b>78.70</b> | <b>79.85</b> |

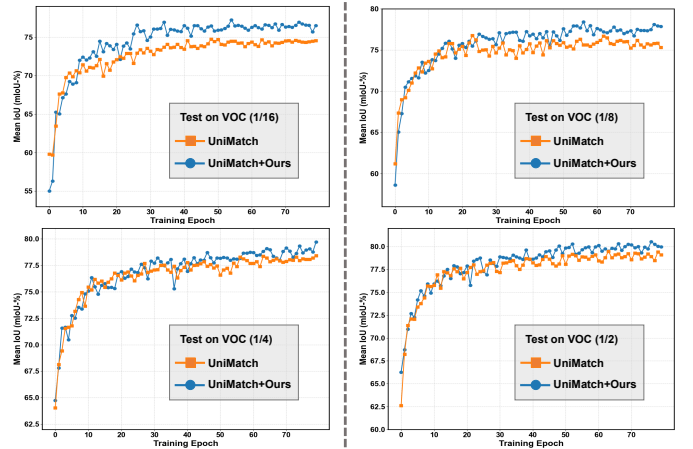


Fig. 1. On the UniMatch baseline, the variation of accuracy (mIoU) with training epochs of the model using and not using our method.

is improved by 2.50%. These results reveal the limitations of self-attention mechanisms under scarce annotation scenarios and further demonstrate the superiority of our method in capturing object-related semantic features for robust reasoning in semi-supervised settings.

## B. Impact of Proposed Method on Training Process

Our visual context alteration involves dynamic variations in image background, which may raise concerns about whether such changes could destabilize the training process, especially during the early stages of training. To investigate this, as shown in Fig. 1, we analyze the training mIoU curves of UniMatch with and without our method, using ResNet-50 and ResNet-101 as backbones. As illustrated, models trained with our strategy do not exhibit noticeable performance fluctuations in the early training stages, and they demonstrate improved convergence stability across multiple data split settings.

The robustness of our training process stems from two intrinsic design principles: First, Auxiliary Regularization

TABLE II  
QUANTITATIVE ANALYSIS FOR THE PERFORMANCE OF DCSE.

| Method               | Pascal VOC   |              | Cityscapes   |              | Param(M) |
|----------------------|--------------|--------------|--------------|--------------|----------|
|                      | 1/16(92)     | 1/8(183)     | 1/16(186)    | 1/8(372)     |          |
| with SA [4]          | 76.73        | 77.59        | 77.63        | 78.57        | 0.0011   |
| with CA [5]          | 76.66        | 77.41        | 77.32        | 78.36        | 0.0328   |
| with CBAM [6]        | 76.97        | 78.10        | 78.22        | 78.94        | 0.0329   |
| with DANet [7]       | 77.04        | 78.05        | 78.16        | 79.06        | 0.3283   |
| <b>with Our DCSE</b> | <b>77.34</b> | <b>78.55</b> | <b>78.58</b> | <b>79.55</b> | 0.0227   |

**Mechanism:** Our framework does not impose VCA as the sole training objective. Instead, it treats the transformed contexts as an auxiliary regularization term. This formulation anchors the optimization process on the primary supervision from original images, ensuring that the fundamental feature learning remains stable. Similar to the established CutMix strategy, VCA acts as a strong regularizer that prevents overfitting to specific background biases without disrupting the core gradient descent direction. Second, **Preservation of Semantic Fidelity:** A critical feature of VCA is that content exchange is strictly constrained to foreground entities of the same semantic category. This constraint ensures structural validity and semantic consistency, as objects of the same class inherently share compatible visual patterns. Consequently, the generated samples remain within the valid semantic manifold, effectively avoiding the introduction of label ambiguity or destructive noise that typically destabilizes the early training.

### C. Comparison between Our DCSE and Related Methods

In the Semantic Invariant Recombination stage, we introduce the Dual-Cleaning Semantic Enhancement (DCSE) module. Its primary objective is to purify the feature representation by strictly filtering out residual background noise while enhancing intrinsic object attributes. To validate the effectiveness of DCSE, we benchmark it against several established attention mechanisms: Spatial Attention (SA) [4], Channel Attention (CA) [5], and hybrid methods like DANet [7] and CBAM [6] that incorporate both perspectives. The comparative results are presented in Tab. II. As evidenced in Tab. II, standard mechanisms like CA and SA yield suboptimal performance compared to our method. The fundamental limitation of these approaches is their 'indiscriminate' nature: SA tends to focus on high-frequency spatial regions regardless of whether they belong to the foreground or background, often amplifying background clutter. Similarly, CA aggregates global statistics, which inevitably mixes valid object semantics with background noise, leading to contaminated feature descriptors. Furthermore, while DANet and CBAM attempt to fuse local and global contexts, they perform feature re-calibration on the entire noisy feature map. Without an explicit mechanism to distinguish and suppress background interference, these hybrid methods risk propagating or even intensifying the residual noise inherent in the context-altered features, thereby limiting their performance upper bound. In contrast, the superiority of DCSE lies in its 'Dual-Cleaning' mechanism. By strategically reusing the background attention map  $\mathcal{A}_{bg}$  (derived from the CCCA stage) as a semantic filter, DCSE explicitly guides both

TABLE III  
THE INFLUENCE OF THE CAPACITY SIZE OF CLASS MEMORY UNIT ON MODEL PERFORMANCE.

| Module                   | Capacity:25 | Capacity:50 | Capacity:75 |
|--------------------------|-------------|-------------|-------------|
| <b>Pascal VOC (1/16)</b> | 76.79       | 77.34       | 77.33       |
| <b>Pascal VOC (1/4)</b>  | 79.42       | 79.91       | 79.85       |
| <b>Cityscapes (1/16)</b> | 78.15       | 78.58       | 78.42       |
| <b>Cityscapes (1/4)</b>  | 79.33       | 79.55       | 79.91       |

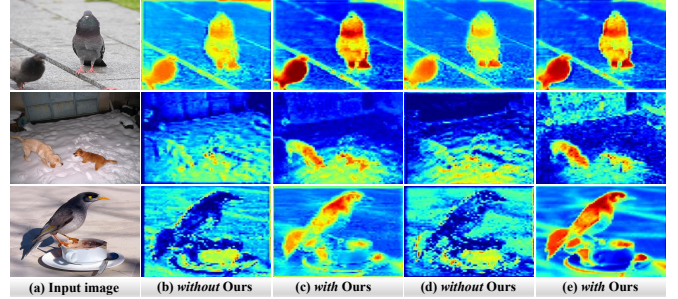


Fig. 2. Differences in feature visualization with and without our method on UniMatch. (b) and (c) use ResNet-50 as the backbone, while (d) and (e) use ResNet-101 as the backbone.

channel and spatial attention to bypass background regions. This ensures that the enhancement is exclusively concentrated on the foreground entities, effectively resolving the issue of feature re-contamination and resulting in a more robust and discriminative representation without a significant increase in the number of parameters.

### D. Analysis about the Capacity of the Class Memory Unit

The storage capacity of the Class Memory Unit is a critical hyperparameter that governs the trade-off between sample diversity and computational efficiency. To determine the optimal setting, we conducted sensitivity analysis on the capacity size using the UniMatch baseline (ResNet-101) under two representative protocols: 1/16 (low-data regime) and 1/4 (high-data regime). The quantitative results are summarized in Tab. III. We observe that a constrained capacity (e.g., 25) fails to provide a sufficiently diverse pool of candidates, thereby limiting the variability of content exchange and degrading performance. Conversely, while expanding the capacity to 75 yields slight performance gains (particularly in the 1/4 partition), it exhibits diminishing returns while incurring higher memory overhead. Consequently, to strike an optimal balance between recognition accuracy and resource consumption, we set the capacity to 50 for all subsequent experiments.

### E. Analysis of the Proposed Method and $\mathcal{A}_{bg}$ via CAM

To more intuitively verify the effectiveness of our method, we utilize Grad-CAM [8] to visualize the intermediate features of the backbone network in UniMatch with and without our method. As shown in Fig. 2 (b) and (d) in the second row, pet dogs are typically associated with lawns or indoor environments, so the presence of snow as an uncommon background factor interferes with the model's reasoning process. In contrast, after incorporating our method, UniMatch is able to

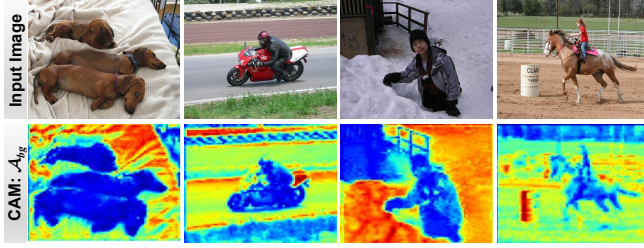


Fig. 3. Background attention maps  $\mathcal{A}_{bg}$  visualized using Grad-CAM.

suppress the adverse influence of the unfamiliar environment and accurately locate and highlight the core regions of the target, minimizing interaction with the background. Similarly, as shown in Fig. 2 (b) and (d) in the third row, the presence of a coffee cup—an unusual background element—significantly disrupts the model’s recognition and reasoning of the bird. However, after applying our method, the model’s attention to the bird is noticeably enhanced. These results further demonstrate that our method helps semi-supervised semantic segmentation models focus on extracting object-related features that are less affected by background context, thereby improving the model’s generalization capability. Furthermore, we utilize CAM to visualize the background attention maps  $\mathcal{A}_{bg}$  generated by the CCCA module. As evident in Fig. 3, these maps exhibit an activation pattern distinct from standard object attention: high-response regions accurately cover diverse contextual backgrounds, while foreground object areas are effectively suppressed. This precise ‘background masking’ capability confirms that the CCCA module successfully captures the global context distribution, providing a critical structural prior for the subsequent DCSE module to filter out background noise and purify foreground features.

#### F. Effectiveness of VCA as Data Augmentation

The proposed Visual Content Alteration (VCA) strategy (isolated from the SIR module) functions inherently as a data augmentation technique. Consequently, it is essential to benchmark its efficacy against established augmentation paradigms, specifically CutMix [9] and ClassMix [10]. In our experiments, we employ UniMatch as the baseline framework. For CutMix, we adhere to the implementation protocol described in [1], while for ClassMix and our VCA, we integrate them into the UniMatch pipeline following an identical protocol to ensure a fair comparison. The quantitative results are detailed in Tab. IV. As observed, applying CutMix results in a performance degradation. This can be attributed to its naive rectangular masking strategy, which disregards object boundaries and violates contextual logic, thereby introducing semantic conflicts that confuse the model’s representation learning. In contrast, ClassMix performs mixing based on semantic masks, which preserves object structure and yields moderate improvements. However, its stochastic selection of exchange categories often fails to prioritize the primary subject, potentially diverting the model’s focus to irrelevant background elements. Conversely, our method strategically defines the exchange target as the dominant semantic category (determined by pixel occupancy). This ensures that the augmentation consistently

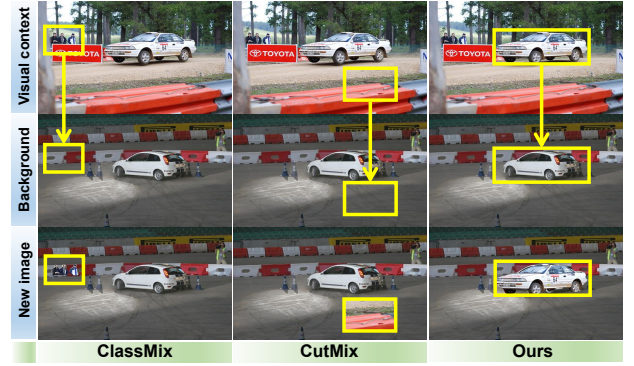


Fig. 4. The differences between our Visual Context Alteration and CutMix and ClassMix in terms of the execution process.

TABLE IV  
THE DIFFERENCE IN MODEL PERFORMANCE IMPROVEMENT BETWEEN OUR VISUAL CONTEXT ALTERATION AND CUTMIX AND CLASSMIX.

| Method        | Pascal VOC   |              | Cityscapes   |              |
|---------------|--------------|--------------|--------------|--------------|
|               | 1/16(92)     | 1/8(183)     | 1/16(186)    | 1/8(372)     |
| CutMix [9]    | 75.21        | 77.17        | 76.60        | 77.95        |
| ClassMix [10] | 75.95        | 77.11        | 77.04        | 78.31        |
| <b>Ours</b>   | <b>77.34</b> | <b>78.55</b> | <b>78.58</b> | <b>79.55</b> |

centers on the principal foreground entity, preserving semantic integrity while enhancing robustness.

#### G. Impact of Our Method on the Visual Foundation Model

With the rise of visual foundation models, large-scale models such as SAM and DINO have shown strong performance in semi-supervised semantic segmentation. To demonstrate that our method remains effective in this setting, we integrate it into two baselines: PS-SAM [11] (leveraging SAM for structural priors) and UniMatch-v2 [12] (using the large-scale pretrained DINO-v2 backbone). For comparison, as illustrated in Fig. 5, the proposed visual context alteration (VCA) and semantic invariance Recombination (SIR) are incorporated into their trainable backbones, yielding **PS-SAM\*** and **UniMatch-v2\***.

As shown in Tab. V, compared with PS-SAM, PS-SAM\* achieves mIoU improvements of 2.18% and 1.14% on VOC under the 1/16 and 1/8 partition settings, respectively. These results indicate that although SAM provides accurate structural priors, it does not encode category information, and class discrimination still relies on additional semi-supervised training modules. Consequently, when dealing with large amounts of unlabeled data, the model can generate relatively precise shapes but remains vulnerable to unfamiliar contexts, leading to misclassifications. UniMatch-v2, which employs DINO-v2 as its backbone, benefits from self-supervised contrastive learning to acquire rich feature representations. However, when labeled data features dominate, the contrastive learning of DINO-V2 tends to overfit the contextual features of labeled data, resulting in limited generalization to unlabeled data. In contrast, UniMatch-v2\* with our strategy achieves mIoU improvements ranging from 1.24% to 2.48%.

The above results indicate that although existing methods [11], [12] benefit from the strong modeling capacity of vision foundation models and achieve promising results,

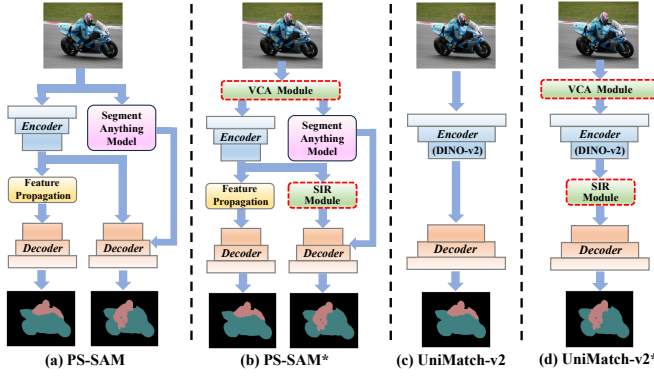


Fig. 5. The structural diagrams of PS-SAM and UniMatch-v2 with PS-SAM\* and UniMatch-v2\* that integrate our method. The red box indicates our method. We modify the channel parameters of the SIR module to align with the configurations of PS-SAM and UniMatch-v2.

TABLE V  
THE INFLUENCE OF OUR METHOD ON PS-SAM AND UniMatch-v2  
BASED ON THE VISUAL FOUNDATION MODEL.

| Method              | Pascal VOC   |              | Cityscapes   |              |
|---------------------|--------------|--------------|--------------|--------------|
|                     | 1/16(92)     | 1/8(183)     | 1/16(186)    | 1/8(372)     |
| PS-SAM              | 77.14        | 81.05        | 77.87        | 79.36        |
| <b>PS-SAM*</b>      | <b>79.32</b> | <b>82.19</b> | <b>79.23</b> | <b>80.48</b> |
| UniMatch-v2         | 79.03        | 85.52        | 80.61        | 81.91        |
| <b>UniMatch-v2*</b> | <b>81.51</b> | <b>86.76</b> | <b>82.62</b> | <b>83.14</b> |

their performance remains suboptimal when labeled data are extremely scarce (e.g., 1/16 and 1/8 splits). By incorporating our SESC<sup>2</sup>, the models achieve further improvements, demonstrating that our method effectively complements large-scale models by guiding them to focus more on object-related semantic cues rather than relying excessively on contextual priors, thereby enhancing generalization on unlabeled data.

#### H. Analysis of the loss functions and hyperparameter $\alpha$

We evaluate the contribution of the visual context alteration loss ( $\mathcal{L}_{vca}$ ) and background constraint loss ( $\mathcal{L}_{bg}$ ) in Fig. 6. We chose UniMatch with a ResNet-100 backbone as the baseline model. As can be seen from Fig. 6, removing either component results in a distinct performance drop compared to the full model on the 1/16 splits. Specifically, the absence of  $\mathcal{L}_{vca}$  hampers the learning of invariant semantics amidst context changes, while omitting  $\mathcal{L}_{bg}$  fails to guide background suppression in the CCCA module, thereby degrading feature purification. These observations corroborate that both auxiliary loss terms are complementary and essential for maximizing the model's robustness in semi-supervised settings.

Furthermore, we investigate the sensitivity of the hyperparameter  $\alpha$ , which controls the weight of the loss  $\mathcal{L}_{vca}$  relative to the primary segmentation objective. We evaluate the model's performance on the Pascal VOC and Cityscapes (1/16 split) by varying  $\alpha$  within the range of 0.1, 0.25, 0.5, 0.75, 1.0, with results summarized in Tab. VI. The results indicate a trade-off: performance improves as  $\alpha$  increases from 0.1, peaking at  $\alpha=0.5$  (77.34% on VOC), before declining at higher values (0.75-1.0). Lower values fail to provide sufficient regularization, whereas excessive weights dominate the optimization and distract from the primary task. Consequently, we adopt  $\alpha=0.5$

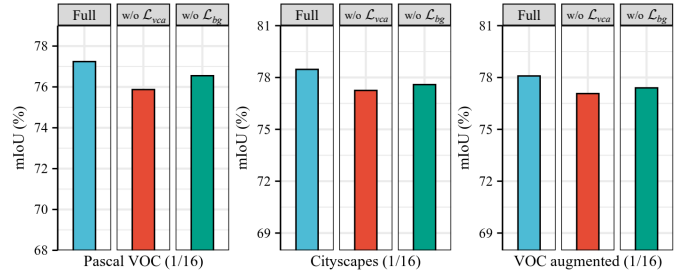


Fig. 6. Quantitative analysis of the proposed visual context alteration loss  $\mathcal{L}_{vca}$  and background constraint loss ( $\mathcal{L}_{bg}$ ).

TABLE VI  
THE EFFECT OF DIFFERENT  $\alpha$  ON MODEL (UniMatch+ResNet-101).

| Dataset                  | $\alpha=0.1$ | $\alpha=0.25$ | $\alpha=0.5$ | $\alpha=0.75$ | $\alpha=1.0$ |
|--------------------------|--------------|---------------|--------------|---------------|--------------|
| <b>Pascal VOC (1/16)</b> | 76.49        | 76.87         | <b>77.34</b> | 77.13         | 76.65        |
| <b>Cityscapes (1/16)</b> | 76.82        | 77.79         | <b>78.55</b> | 78.25         | 78.08        |

as the optimal setting to balance consistency regularization with supervised learning.

#### REFERENCES

- [1] H. Wang, Q. Zhang, Y. Li, and X. Li, "Allspark: Reborn labeled features from unlabeled in transformer for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3627–3636, June 2024.
- [2] H. Huang, S. Xie, L. Lin, R. Tong, Y.-W. Chen, Y. Li, H. Wang, Y. Huang, and Y. Zheng, "Semcvt: Semi-supervised convolutional vision transformer for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11340–11349, June 2023.
- [3] X. Hu, L. Jiang, and B. Schiele, "Training vision transformers for semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4007–4017, June 2024.
- [4] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, vol. 28, Curran Associates, Inc., 2015.
- [5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [6] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
- [7] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [9] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [10] V. Olsson, W. Tranheden, J. Pinto, and L. Svensson, "Classmix: Segmentation-based data augmentation for semi-supervised learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1369–1378, January 2021.
- [11] H. Hao, W. Liu, and W. Xing, "Prior-structure driven weakly-supervised learning for fine-grained human parsing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 1, pp. 461–476, 2025.
- [12] L. Yang, Z. Zhao, and H. Zhao, "Unimatch v2: Pushing the limit of semi-supervised semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 4, pp. 3031–3048, 2025.