

A TRAINING ALGORITHM

Algorithm 1 summarizes the overall training strategy of our proposed framework. We take the source data $\{(x_s, y_s)\}$ and CLIP-based segmenter G_s as the input of our method. After N_1 training iterations for each batch of source data, we obtain the optimized PIN modules. Then we utilize them to generate target features. We use them to fine-tune the target segmenter G_t within N_2 iterations.

Algorithm 1 Adaptation algorithm

Input: Source data $\{(x_s, y_s)\}$, source segmenter G_s , Target domain description $\{\text{TrgDesc}\}$
Output: target segmenters $\{G_t\}$

- 1: Feed $\{\text{TrgDesc}\}$ into E_{txt} to extract $\{\text{TrgEmb}\} = E_{\text{txt}}(\{\text{TrgDesc}\})$;
- 2: Feed $\{x_s\}$ into E_{img} to extract $f_s = E_{\text{img}}(\{x_s\})$;
- 3: Transform f_s into $\{f_{s \rightarrow t}\}$ with PIN modules for each target domain;
- 4: Generate visual meta-nodes \mathbf{Q}_v with $\{\mu\}$ and $\{\sigma\}$ in PIN modules;
- 5: Construct hybrid cross-modality graph \mathcal{G}_h with \mathbf{Q}_v and $\{\text{TrgEmb}\}$;
- 6: Calculate the extreme features \mathbf{P}_v on the boundary of visual feature distribution ranges;
- 7: Mine graph motifs \mathcal{M} from \mathcal{G}_h ;
- 8: **for** $i = 1$ to N_1 **do**
- 9: Calculate the similarity sim of language-vision edges within each motif;
- 10: Calculate the motif matching loss L_{match} ;
- 11: Calculate the directional loss L_{dis} ;
- 12: Calculate the contrastive loss L_{con} ;
- 13: Update the PIN modules with overall loss L_{total} ;
- 14: **end for**
- 15: **for** $j = 1$ to N_2 **do**
- 16: Feed $\{x_s\}$ into E_{img} to extract $f_s = E_{\text{img}}(\{x_s\})$;
- 17: Transform f_s into $\{f_{s \rightarrow t}\}$ with trained PIN modules for each target domain;
- 18: Calculate the cross-entropy loss with $\{f_{s \rightarrow t}\}$ and source labels $\{y_s\}$;
- 19: Fine-tune the target segmenters $\{G_t\}$ for each target domain with $\{f_{s \rightarrow t}\}$ and y_s
- 20: **end for**
- 21: **return** target segmenters $\{G_t\}$

B MOTIVATION

The motivation for proposing a new graph motif-based adaptation method stems from the challenges faced in zero-shot domain adaptive semantic segmentation, where the goal is to transfer knowledge from a source domain to a target segmenter without access to target domain data. Existing methods that transform source features to the target domain using language-driven approaches tend to coarsely align language features to global features. This results in the sub-optimal performance of cross-domain feature alignment.

To overcome these issues, the new method focuses on balancing efficiency and effectiveness in feature alignment. It introduces a graph motif structure that is based on domain-wise image feature distributions. By adjusting the angle between language-vision directed edges, the method pulls visual features toward the language feature center, achieving a more precise cross-modality feature alignment without excessive computational demands. Additionally, the incorporation of directional and contrastive losses helps to mitigate the mode-collapse during feature stylization, stabilizing the learning process and enhancing the robustness of the adaptation.

C LIMITATIONS

Although our work has achieved state-of-the-art performance, it is not without limitations. Firstly, our method is not end-to-end. The segmented training mode complicates the pipeline, potentially restricting its applicability in real-world applications. Additionally, our approach necessitates training a distinct target segmenter for each target domain, thereby constraining its domain generalization capabilities.

Table 6: Comparison of our proposed method with different loss weights.

λ_{match}	λ_{dis}	λ_{con}	mIoU
0.01	0.01	0.01	41.13
0.05	0.05	0.05	41.52
0.1	0.05	0.05	41.80
0.15	0.05	0.05	41.43
0.05	0.1	0.05	39.73
0.05	0.05	0.1	40.29
0.05	0.1	0.1	38.48

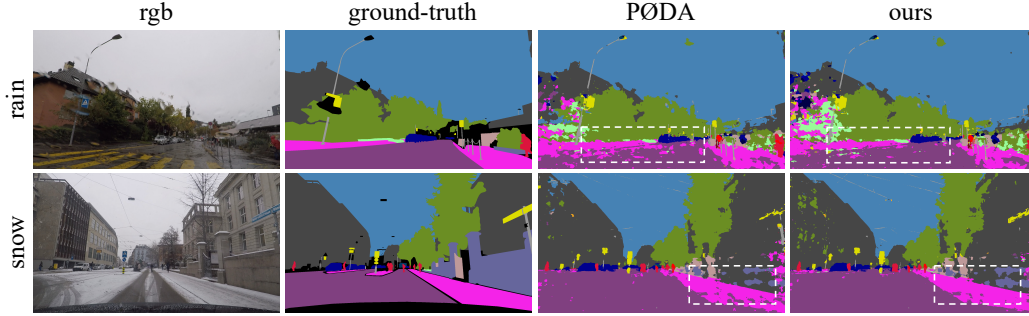


Figure 4: Qualitative comparison results on the task of Cityscapes→ACDC.

D ABLATION ON THE LOSS WEIGHTS

We evaluate our proposed method with different loss weights λ_{match} , λ_{dis} , and λ_{con} on the adaptation task of Cityscapes→ACDC. The contribution of different optimization objectives to PIN modules optimization can be balanced by adjusting the loss weights. Notably, relatively large λ_{dis} and λ_{con} are harmful to the adaptation performance, so we need to set them carefully. When λ_{match} , λ_{dis} , and λ_{con} are set to 0.1, 0.05, and 0.05, our method achieves the best performance. We experimentally determined the weights and applied them to all of our experiments.

E QUALITATIVE VISUALIZATION RESULTS

Fig. 4 shows more qualitative comparison results on the adaptation task from source to the subsets of ACDC. It shows that our method achieves more accurate pixel-level segmentation results. The segmentation results prove the effectiveness of our method.