Frequency-guided Iterative Bi-directional Exchange Network for Cross-Domain Few-Shot Segmentation

Supplementary Material

8. Overview

This supplementary document provides extended analyses, implementation details, and additional results. The materials are organized as:

- An overview of the datasets used in our experiments.
- Comprehensive ablation studies on loss weights and refinement iterations.
- A computational cost analysis that includes parameter counts.
- A complete summary of all hyperparameter settings.
- Additional qualitative results under the 1-shot and 5-shot settings.
- The full pseudocode of the proposed method for reproducibility.

9. Dataset Overview

As described in Section 5.1 of the main paper, we evaluated our approach on several datasets. This supplementary section provides additional analysis (Fig. 9), highlighting the specific challenges each dataset poses for cross-domain few-shot segmentation (CD-FSS).

Deepglobe [10] contains high-resolution satellite imagery. Its large-scale layouts, non-rigid terrain boundaries, and multi-scale texture patterns differ markedly from conventional visual scenes. The combination of wide spatial ranges and heterogeneous surface structures makes extracting coherent semantics from only a few annotated supports challenging.

FSS-1000 [21], in the natural-image domain, covers objects with diverse shapes, materials, and viewpoints. This broad appearance spectrum results in highly non-uniform classes, reducing the reliability of prototypes from limited support examples and increasing the chance of inconsistent query transfer.

ISIC [9] introduces a medical-imaging shift. Dermoscopic images exhibit specialized illumination, lesion pigmentation variations, and fine-grained boundary transitions. These clinically characteristic patterns differ from natural-image cues and make capturing subtle lesion structures difficult under sparse supervision.

Chest X-Ray [3, 17] presents grayscale radiographs with low-contrast organ boundaries and overlapping anatomical regions. The limited texture content and intensity homogeneity complicate structural reasoning, as distinguishing foreground and background regions relies on subtle cues absent in RGB-based priors.

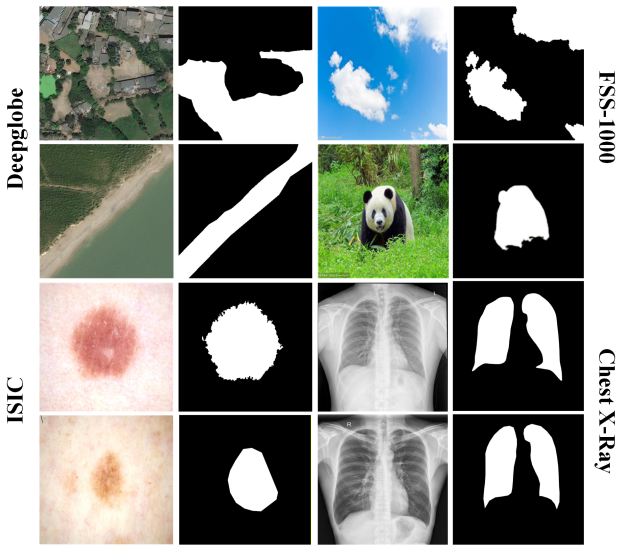

Figure 9. Overview of the four segmentation datasets used in our experiments—Deepglobe, FSS-1000, ISIC, and Chest X-Ray—covering satellite imagery, natural scenes, diverse objects, and medical radiographs.

Together, these datasets offer a challenging benchmark that highlights the effectiveness of FIBEN (Frequency-guided Iterative Bi-directional Exchange Network) in cross-domain and few-shot scenarios.

10. Loss Weight Ablation

As discussed in Section 4.4 of the main paper, the effect of the regularization weight λ_{reg} has already been analyzed. Here, we focus on the remaining loss weights, λ_{base} and λ_{iter}, through single-factor ablation experiments. The total loss is defined as:

$$\mathcal{L} = \lambda_{\text{base}}\mathcal{L}_{\text{base}} + \lambda_{\text{iter}}\mathcal{L}_{\text{iter}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}.$$

Here, $\mathcal{L}_{\text{base}}$ provides supervision for the primary segmentation task, $\mathcal{L}_{\text{iter}}$ encourages iterative alignment between support and query features, and $\mathcal{L}_{\text{reg}}$ acts as a regularization term to enhance generalization.

Effect of the Base Loss Weight. To examine how the primary supervision influences model performance, we vary $\lambda_{\text{base}} \in \{0.5, 0.8, 1.0\}$ while keeping $\lambda_{\text{iter}} = 0.1$ and $\lambda_{\text{reg}} = 0.5$. Increasing λ_{base} consistently improves performance in both the 1-shot and 5-shot settings. For instance, the 1-shot mIoU increases from 68.96 to 70.24, and the 5-shot mIoU from 70.56 to 73.77. This trend indicates that

placing greater emphasis on the base loss helps the model learn more robust and generalizable feature representations, particularly when multiple support examples are provided. Based on these observations, we set $\lambda_{\text{base}} = 1.0$ in all experiments. The results are reported in Table 7.

Effect of the Iterative-Refinement Loss Weight. Next, we investigate the impact of the auxiliary iterative-refinement loss by varying $\lambda_{\text{iter}} \in \{0.1, 0.2, 0.3\}$, while fixing $\lambda_{\text{base}} = 1.0$ and $\lambda_{\text{reg}} = 0.5$. The results are summarized in Table 8. In contrast to the base loss, increasing λ_{iter} gradually reduces performance: the 1-shot mIoU decreases from 70.24 to 68.97, and the 5-shot mIoU from 73.77 to 71.73. This indicates that assigning excessive weight to the iterative-refinement term can slightly degrade performance, as the model may overemphasize iterative alignment at the expense of optimizing the primary supervision. A lower weight provides a more appropriate balance among the base, iterative, and regularization losses, and we therefore adopt $\lambda_{\text{iter}} = 0.1$ in all experiments.

Table 7. Ablation of the base loss weight λ_{base} under 1-shot and 5-shot settings (with $\lambda_{\text{reg}} = 0.5$, $\lambda_{\text{iter}} = 0.1$, $T = 3$).

λ_{base}	1-shot mIoU	5-shot mIoU
0.5	68.96	70.56
0.8	69.46	71.92
1.0	**70.24**	**73.77**

Table 8. Ablation of the iterative-refinement loss weight λ_{iter} under 1-shot and 5-shot settings (with $\lambda_{\text{reg}} = 0.5$, $\lambda_{\text{base}} = 1.0$, $T = 3$).

λ_{iter}	1-shot mIoU	5-shot mIoU
0.1	**70.24**	**73.77**
0.2	69.69	72.56
0.3	68.97	71.73

11. Iterative Refinement and FSEM Analysis

Building upon the analysis in Section 5.4 of the main paper, which summarizes the average mIoU across different iteration numbers, Table 9 provides a dataset-level breakdown, revealing how the number of refinement steps influences performance with and without the FSEM (Frequency–Spatial Enhancement Module).

For all datasets, performance generally increases as the refinement depth increases, though the gain on FSS-1000 is extremely small due to its close similarity to the source domain. Adding FSEM brings the most noticeable improvement in the early refinement steps, after which the gap between the "with FSEM" and "without FSEM" settings gradually narrows as the iterative refinement converges. The benefit is particularly visible on texture- or structure-dependent datasets such as ISIC and Deepglobe, where frequency-guided cues strengthen.

Table 9. mIoU (%) over 1–7 BPFR iterations, without (w/o) and with (w/) frequency guidance (FSEM).

Iter	Chest X-Ray		ISIC		FSS-1000		Deepglobe		Average	
	w/o F	w/ F	w/o F	w/ F	w/o F	w/ F	w/o F	w/ F	w/o F	w/ F
1	78.76	81.70	62.72	66.62	80.25	80.28	44.81	47.19	66.64	68.95
2	79.52	82.30	64.02	67.74	80.30	80.37	46.02	48.32	67.47	69.68
3	80.75	82.54	65.92	68.80	80.57	80.60	47.33	49.02	68.64	70.24
4	81.32	82.62	66.67	68.86	80.70	80.72	48.42	49.09	69.28	70.32
5	82.16	82.97	67.32	69.02	80.75	80.81	49.02	49.00	69.81	70.45
6	82.62	83.02	68.21	69.12	80.83	80.87	49.19	49.27	70.21	70.57
7	82.87	83.12	68.97	69.15	80.90	80.93	49.30	49.39	70.51	70.65

FSS-1000, in contrast, shows only marginal improvement because its distribution is very close to the source domain PASCAL VOC 2012. Here, the frequency characteristics of the two domains are already highly aligned, reducing the domain discrepancies that FSEM is designed to correct. As a result, frequency-based cues offer limited additional benefit beyond the spatial refinements already captured by BPFR.

These findings suggest that frequency-based augmentation is particularly helpful under significant texture distortion or structural ambiguity, whereas its impact is reduced when the domain gap is mild and spatial cues dominate.

12. Model Parameter Analysis

Table 10 reports the number of trainable parameters for each model variant; FLOPs and inference time are omitted because they depend on implementation choices and hardware. The baseline contains 8.7 million parameters. Integrating the BPFR (Bi-directional Prototype–Feature Refinement) module increases the total to 11.8 million, reflecting a modest overhead despite the addition of iterative prototype refinement and support–query interaction. Incorporating the complete FSEM (Frequency–Spatial Enhancement Module), which consists of the DualSpectrumModulator, FrequencyGuidAttention, and SpatialSelfAttention blocks, expands the parameter count to 17.6 million. This increase primarily results from lightweight spectral-modulation and attention components, along with extra fusion parameters for the support and query branches. Overall, the full model remains compact at 17.6 million parameters, indicating that the observed performance gains are attributable to architectural design rather than substantial parameter growth.

13. Experimental Hyperparameters

Details of the experimental settings are provided in Section 5.2 of the main paper. Table 11 presents a comprehensive summary of all key hyperparameters used in our experiments. The table includes the weights assigned to

Table 10. Number of trainable parameters for different models.

Method	Params (M)
Baseline	8.7
Baseline + BPFR	11.8
Baseline + BPFR + FSEM (FIBEN)	**17.6**

each loss component, the number of iterative refinement steps during training, the learning rate, batch size, and other parameters relevant to the training procedure. These hyperparameters were carefully selected to ensure stable convergence and consistent performance across different datasets. Unless stated otherwise, the same hyperparameter configuration is applied to all datasets and experimental settings, ensuring both reproducibility of our results and a fair comparison between different models and setups.

14. Qualitative Results

5-Shot Setting. The quantitative results for the 5-shot setting are reported in Table 1 of the main paper. In this supplementary material, we present additional qualitative comparisons in Figure 11 across four datasets: Chest X-Ray, ISIC, FSS-1000, and Deepglobe. Each example includes five support images, one query image, and predictions generated by SSP* [13], IFA [29], and our method. On the Chest X-Ray dataset, our approach delineates lung boundaries more accurately, with fewer leakage artifacts, compared with the incomplete or imprecise masks produced by SSP* and IFA. For ISIC, the method reliably segments lesions with irregular shapes and varying sizes, demonstrating improved robustness over baseline models. On FSS-1000, it produces well-defined object contours and avoids the spurious activations common in SSP* and IFA outputs. For Deepglobe, our predictions remain stable in complex aerial scenes and suppress false positives more effectively than the baselines. These qualitative outcomes consistently align with the quantitative advantage reported in the main paper, further validating the effectiveness of our framework under the 5-shot setting.

Qualitative Visualization of FIBEN. Complementing the quantitative analysis in Section 5.4, Figure 10 provides qualitative visualizations on Chest X-Ray, ISIC, Deepglobe, and FSS-1000 under the 1-shot setting, illustrating how FSEM (Frequency–Spatial Enhancement Module) and BPFR (Bi-directional Prototype–Feature Refinement)contribute to mask quality. The baseline SSP* [13] produces numerous false positives and unstable boundaries, particularly in datasets with substantial appearance gaps such as Deepglobe and Chest X-Ray. This highlights that conventional feature representations are highly domain-sensitive and prone to semantic drift.

Introducing the FSEM improves feature stability by suppressing noisy responses and reducing artifacts caused by domain shifts. The resulting predictions become more compact and structurally coherent, especially on ISIC and FSS-1000, showing that FSEM effectively strengthens feature

Table 11. Summary of hyperparameters used in our experiments, including fine-tuning stage.

Hyperparameter	**Value / Setting**
Loss weights	
λ_{base}	1.0
λ_{iter}	0.1
λ_{reg}	0.5
Model parameters	
Number of iterative steps	3
Support shots (K)	1; optionally 5
Backbone	ResNet-50
DualSpectrumModulator: low height × width	16 × 16
DualSpectrumModulator: max phase	$\pi/6$
FrequencyGuidAttention: channel / reduction	1024 / 8
SpatialSelfAttention: channels / reduction ratio / down_kv	1024 / 16 / 2
Query merge conv kernel size	1
Fusion weights ($\lambda_{\text{fg}}/\lambda_{\text{bg}}$)	0.5 / 0.3
Foreground confidence threshold	0.7
Background confidence threshold	0.6
Top-K selection for prototypes	12
Attention temperature α	2.0
Stability constant ε	1×10^{-5}
Training settings (source domain)	
Learning rate	1×10^{-3} for all datasets
Optimizer	SGD (momentum=0.9, weight decay=5e-4)
Batch size (per dataset)	12
Total epochs	20
Iteration per step (BPFR)	3
Image resize and crop	400 × 400
Random seed	0; evaluation averaged over 5 seeds
Fine-tune settings (target domain)	
Learning rate (per dataset)	Deepglobe 5×10^{-4}, ISIC 5×10^{-4}, FSS-1000 5×10^{-4}, Chest X-Ray 1×10^{-5}
Optimizer	SGD (momentum=0.9, weight decay=5e-4)
Batch size	12
Total epochs	60; first 30 initial, last 30 adaptation
Iteration per step (BPFR)	3
Image resize and crop	400 × 400
Random seed	0; evaluation averaged over 5 seeds

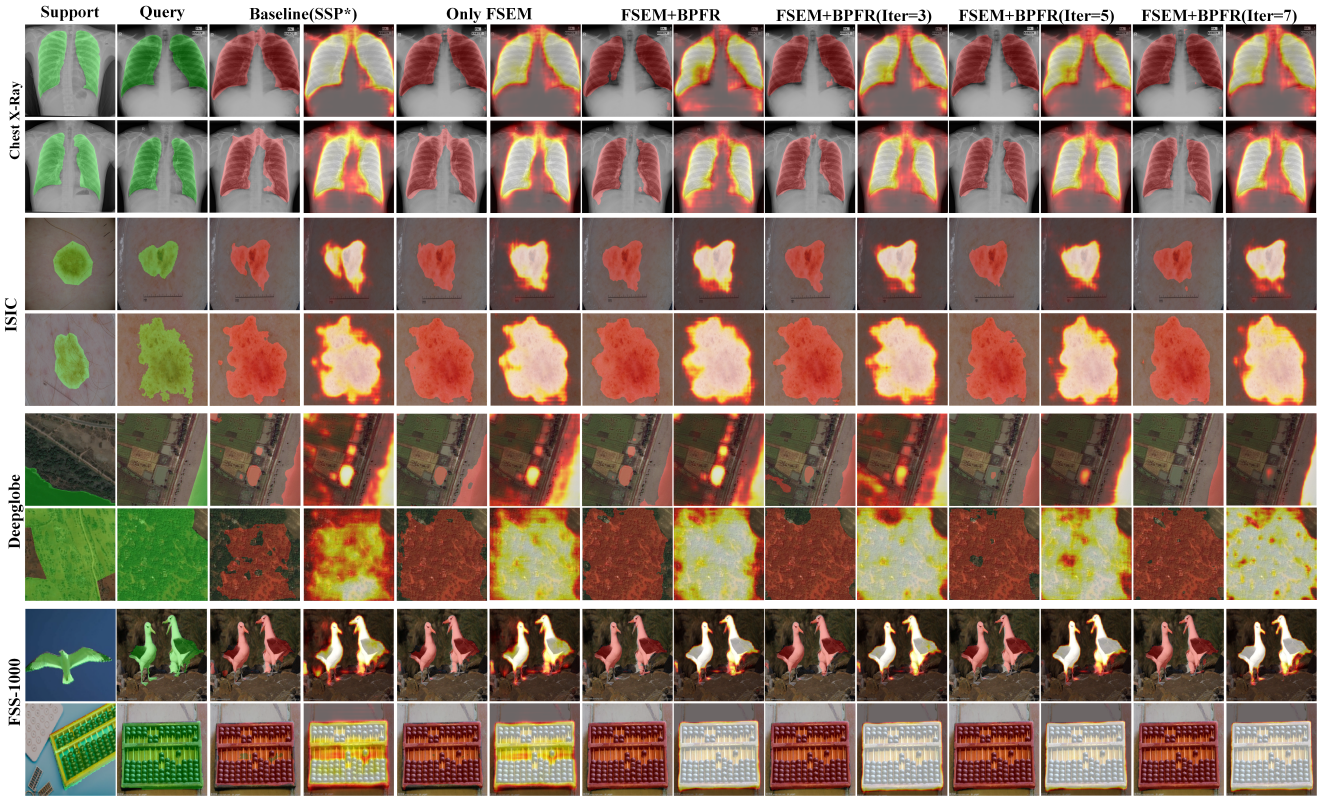

Figure 10. Qualitative comparison on Chest X-Ray, ISIC, Deepglobe, and FSS-1000. From left to right: Support, Query, Baseline (SSP*), Only FSEM, and our method with BPFR under different refinement iterations. Our components progressively reduce false positives and produce sharper, more accurate masks. * indicates our reproduced results.

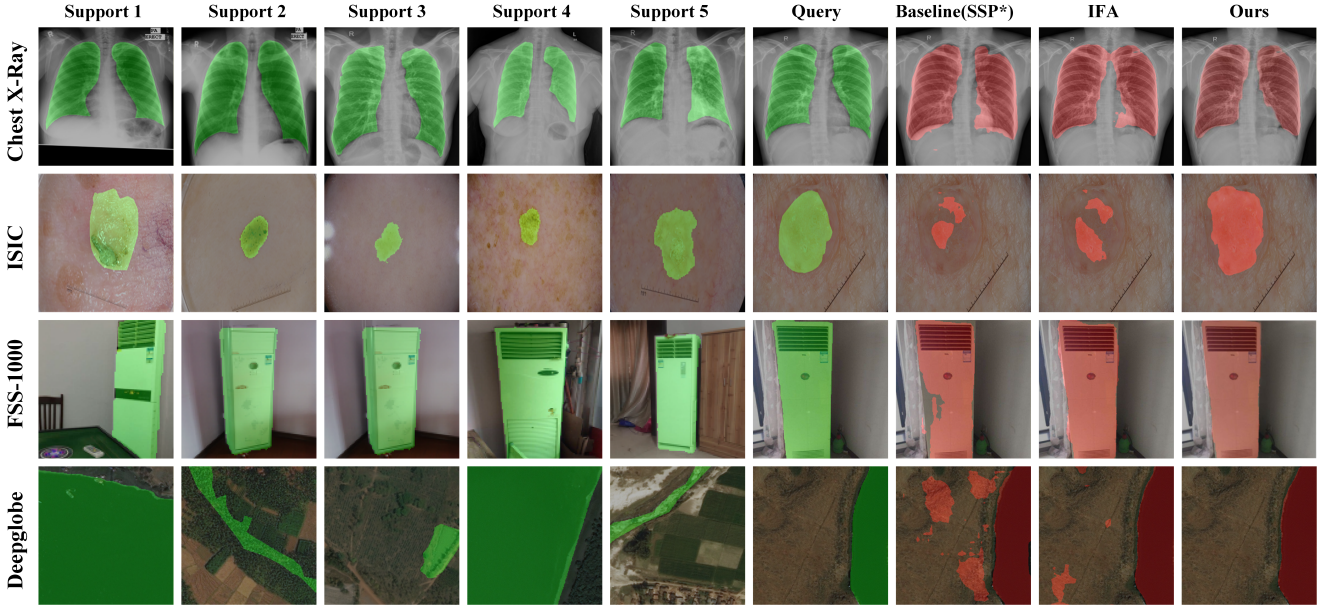

Figure 11. Qualitative results on Chest X-Ray, ISIC, FSS-1000, and Deepglobe. Each row shows five support images, one query image, and predictions from SSP*, IFA and our method. Our approach yields cleaner and more accurate segmentations with fewer false positives across all datasets. * indicates our reproduced results.

robustness. Incorporating the BPFR module further enhances segmentation quality by iteratively aligning support and query features. Gradually increasing the number of refinement iterations from one to seven consistently sharpens mask boundaries, reduces false positives, and improves spatial consistency. The iterative prototype–feature interactions systematically correct early mismatches and restore fine structural details.

Overall, the visualizations demonstrate that FSEM and BPFR provide complementary benefits. FSEM alleviates domain overfitting at the feature level, while BPFR stabilizes semantic matching through iterative refinement. Their combination enables FIBEN to produce sharper, cleaner, and more reliable segmentation masks across diverse cross-domain scenarios.

15. FIBEN Algorithm

Algorithm 1 summarizes the proposed Frequency-guided Iterative Bi-directional Exchange Network (FIBEN). Given a support image I_s with annotations M_s and a query image I_q, the algorithm refines feature representations and semantic prototypes to establish robust cross-domain correspondence. The model G_s is first trained on the labeled source domain, and a lightweight target-domain adaptation updates only later layers while freezing early backbone and batch normalization parameters to align with the target distribution; during training, the query mask M_q is used for loss computation, and T denotes the number of BPFR iterations.

The algorithm proceeds in three stages:
(1) Frequency–Spatial Enhancement Module (FSEM). Support and query features are enhanced via frequency modulation and spatial-channel attention, suppressing domain-specific noise and strengthening semantic structures.
(2) Initial Prototype Extraction. Foreground and background prototypes are aggregated from enhanced support features using the support mask, providing initial anchors for cross-instance matching.
(3) Bi-directional Prototype–Feature Refinement (BPFR). Prototypes and query features are iteratively updated in a bi-directional manner, fusing global and local cues via λ_{fg} and λ_{bg}. This progressively aligns semantic structures, corrects early mismatches, and improves boundary precision. Final predictions are obtained by computing similarity between refined prototypes and query features.

Overall, FIBEN combines frequency-guided feature enhancement with bi-directional iterative prototype–feature refinement, enabling stable and accurate segmentation across domains with significant distribution shifts. The lightweight target-domain adaptation further reduces residual domain gaps, making the framework both robust and reproducible for cross-domain few-shot segmentation.

Algorithm 1 Pipeline of the proposed Frequency-guided Iterative Bi-directional Exchange Network (1-shot setup, updated query feature iteration).

Input: Support image I_s, Query image I_q, Support mask M_s, Query mask M_q
Parameters: Source model G_s, number of BPFR iterations T
Output: Updated source model G_s (later used for target adaptation)
while epoch $<$ max_epoch **do**
  /* Frequency–Spatial Enhancement (FSEM) */
  Extract feature maps X_s, X_q via G_s
  Compute FFT: $\mathcal{F}(X) = Ae^{j\phi}$; apply amplitude and phase modulation
  $X_{\text{freq}} \leftarrow \mathcal{F}^{-1}(A'e^{j\phi'})$
  Compute descriptors u_a and u_p
  Generate channel attention CA and spatial guidance SG
  Obtain refined features:
  $X_{\text{attn}} = X_{\text{freq}} \odot (1 + CA) \odot (1 + SG)$
  Fuse multi-branch enhanced features and apply linear attention to obtain F_s, F_q
  /* Initial prototype extraction */
  for $c \in \{\text{fg}, \text{bg}\}$ **do**
    $P^0_{s,c} \leftarrow \text{MAP}(F_s, M_s)$
  end for
  /* Iterative Bi-directional Prototype–Feature Refinement (BPFR) */
  $t \leftarrow 0$
  while $t < T$ **do**
    /* Query-side refinement */
    Compute similarity maps $S^t_{q,c}$ and obtain preliminary mask M^{pre}_q
    Extract query prototypes using SPE (Self-guided Prototype Extraction):
    $(P^{g,t}_{q,fg}, P^{g,t}_{q,bg}, P^{l,t}_{q,bg}) \leftarrow \text{SPE}(F^t_q, M^{\text{pre}}_q)$
    Update query prototypes:
    $P^t_{q,fg} = \lambda_{fg} P^t_{s,fg} + (1 - \lambda_{fg}) P^{g,t}_{q,fg}$
    $P^t_{q,bg} = \lambda_{bg} P^{g,t}_{q,bg} + (1 - \lambda_{bg}) P^{l,t}_{q,bg}$
    Fuse features with prototypes to obtain F^{t+1}_q
    Update query features for next iteration: $F_q \leftarrow F^{t+1}_q$
    /* Support-side refinement */
    Compute similarity using updated query prototypes and obtain M^{pre}_s
    Extract support prototypes using SPE:
    $(P^{g,t}_{s,fg}, P^{g,t}_{s,bg}, P^{l,t}_{s,bg}) \leftarrow \text{SPE}(F_s, M^{\text{pre}}_s)$
    Update support prototypes:
    $P^{t+1}_{s,fg} = \lambda_{fg} P^t_{s,fg} + (1 - \lambda_{fg}) P^{g,t}_{s,fg}$
    $P^{t+1}_{s,bg} = \lambda_{bg} P^t_{s,bg} + (1 - \lambda_{bg}) P^{l,t}_{s,bg}$
    $t \leftarrow t + 1$
  end while
  /* Loss computation */
  Compute segmentation losses $\mathcal{L}_{base}, \mathcal{L}_{iter}$ and spectral regularizer $\mathcal{L}_{reg}$
  $\mathcal{L} = \lambda_{base}\mathcal{L}_{base} + \lambda_{iter}\mathcal{L}_{iter} + \lambda_{reg}\mathcal{L}_{reg}$
  Update parameters of G_s
end while