ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888870

Spatial Frequency-Aware Self-Distillation for Weakly-Supervised Semantic Segmentation

Jingyuan Fang
School of Computer Science and Technology
Shandong Jianzhu University
Jinan, P.R. China
2022110105@stu.sdjzu.edu.com

Yang Ning*
School of Computer Science and Technology
Shandong Jianzhu University
Jinan, P.R. China
ningyang20@sdjzu.edu.cn

Xiushan Nie
School of Computer Science and Technology
Shandong Jianzhu University
Jinan, P.R. China
niexsh@hotmail.com

***Abstract*—Weakly-supervised semantic segmentation (WSSS) aims to achieve pixel-level classification under image-level supervision. Recent class activation map (CAM)-based methods seek to expand foreground activation while suppressing background. However, they often overlook the uncertainty of CAM, where non-salient activation in some regions complicates semantic classification. These regions are typically dismissed as noise, resulting in inappropriate activations due to inadequate regularization. To resolve this, we introduce a Spatial Frequency-Aware Self-Distillation strategy (SFS). Firstly, to enhance the perception of high-frequency spatial information in uncertain regions, we propose a boundary self-distillation and uncertain region reconstruction strategy, which captures high-frequency boundary information and fine-grained spatial context in these regions. Secondly, to enhance the discrimination of low-frequency semantic features, we propose a contrastive attention mechanism that guides the Vision Transformer (ViT) to focus more on the foreground, thereby improving the distinction between foreground and background. Finally, our SFS demonstrates outstanding performance on both the VOC 2012 and COCO 2014 datasets, attributed to its superior spatial frequency perception capabilities. The code is available at https://github.com/fjoybest/SFS.**

***Index Terms*—weakly-supervised semantic segmentation, self-distillation, spatial frequency awareness.**

I. Introduction

Weakly-supervised semantic segmentation (WSSS) methods leverage image-level annotations to achieve pixel-level classification [1]–[5]. Current WSSS methods typically utilize class activation maps (CAMs) to derive initial object localization [6]–[8]. After refinement, these CAMs are used as pseudo-labels to guide the segmentation network [9]–[12].

However, CAMs often activate only the most salient regions, leading to under-activation [13]–[15]. Conversely, networks with strong semantic associations frequently experience over-activation of the background [16], [17]. While existing methods address under-activation and over-activation separately [18], [19], they often overlook the uncertainty of CAM, which may be a common factor of these issues.

*Corresponding author.

This work is supported in part by the Shandong Provincial Natural Science Foundation for Distinguished Young Scholars (ZR2021JQ26), National Natural Science Foundation of China (62176141), Shandong Provincial Natural Science Foundation (ZR2020QF029), Major science and technology innovation project of Shandong Province (2021CXGC11204), the Natural Science Foundation of Shandong Province (No. ZR202103010201).

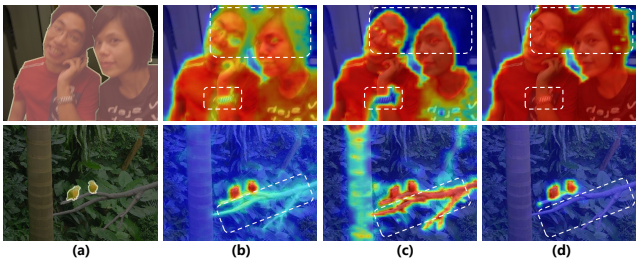

Fig. 1. CAMs visualization for illustrating our motivation. (a) Image and ground truth. (b) Intermediate layer of the network. (c) Network lacking uncertainty regularization. (d) Our SFS. White dashed boxes highlight the uncertain regions, which are primarily indicated by the yellow-green areas.

The uncertainty of CAM refers to the condition where the activation values are neither significantly high nor low, rendering it challenging to accurately classify these regions as foreground or background. As shown in Fig. 1(b), uncertainty is notably prevalent in the network's intermediate layers. Current WSSS methods often discard these uncertain regions as noise during regularization or pseudo-label optimization [18], [20]. We analyze that insufficient regularization of uncertain regions may cause over-activation or under-activation in these methods. For example, as shown in Fig. 1, regions such as "hair" and "watches" exhibit uncertainty in the intermediate layers. The lack of regularization for these uncertain regions results in their incorrect classification as background by deeper network layers, causing under-activation. Conversely, "branches" that exhibit uncertainty in the intermediate layers are subject to excessive activation in later layers, leading to over-activation of background.

Inspired by the learning strategy of self-distillation [21], [22], we introduce a Spatial Frequency-Aware Self-Distillation Network (SFS) to guide the correct division of uncertain regions. As demonstrated in Fig. 1(b), uncertainty is predominantly located at the boundaries between different semantic categories. To enhance the network's boundary and high-frequency spatial awareness, we propose a high-frequency spatial context miming (HSCM) strategy, which includes high-frequency boundary self-distillation (HBS) and uncertainty

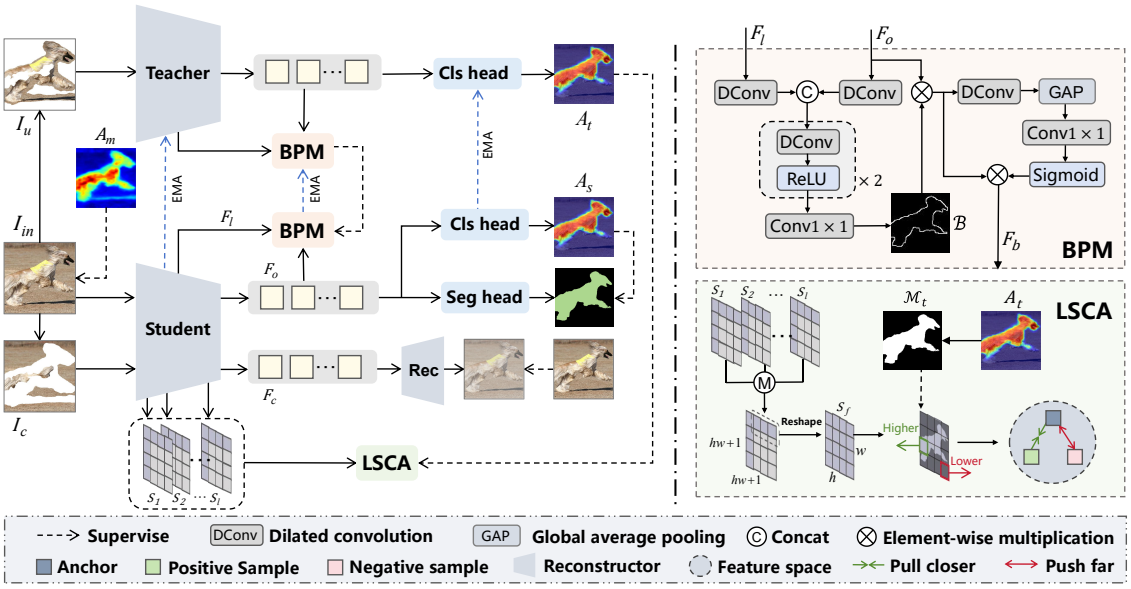

Fig. 2. Overall Architecture of the proposed Spatial Frequency-Aware Self-Distillation Network (SFS). The teacher network updates its parameters using Exponential Moving Averages (EMA).

spatial reconstruction (USR), providing the network with fine-grained spatial boundary representations of uncertain regions. Additionally, to alleviate uncertainty by enhancing semantic discrimination, we introduce a Low-Frequency Semantic-Aware Contrastive Attention Mechanism (LSCA), which directs multi-head attention to focus on low-frequency semantic distinctions between foreground and background. Finally, as shown in Fig. 1(d), our SFS, leveraging spatial frequency awareness, effectively guides the partitioning of uncertain regions and prevents inappropriate activations.

The main contributions of this paper can be summarized as follows:

- We propose a high-frequency spatial context mining (HSCM) strategy to achieve self-distillation of high-frequency boundary information and fine-grained spatial inference for uncertain regions.
- We introduce a low-frequency semantic-aware contrastive attention (LSCA) mechanism to enhance low-frequency semantic discrimination and reduce semantic uncertainty.
- Our method significantly enhances pseudo-label accuracy and achieves new WSSS state-of-the-art performance on VOC 2012 and COCO 2014 datasets.

II. Methods

The proposed SFS is based on the Vision Transformer (ViT) [23] architecture, with an overview shown in Fig. 2. This section begins with the generation of a CAM-based uncertain region mask and data augmentation. We then introduce our HSCM and LSCA strategies. Finally, we integrate proposed strategies into an end-to-end WSSS framework.

A. Prerequisites

1) Uncertain Region Mask Generation: Due to the higher uncertainty typically present in the intermediate layer of the network, we utilize CAMs from this layer as auxiliary pseudo-labels to distinguish between certain and uncertain regions.

Specifically, input image I_{in} is processed through the ViT encoder to extract intermediate features $F_m \in \mathbb{R}^{h\times w\times C}$, where $h \times w$ represents the spatial dimension after reshaping the sequence of hw tokens into a two-dimensional format, and C denotes channel dimension. An auxiliary classifier then facilitates the acquisition of classifier weights $W_{cls} \in \mathbb{R}^{K\times C}$, where K represents the number of categories. Then the intermediate layer CAM $A_m \in \mathbb{R}^{h\times w\times K}$ is generated by:

$$A_m = norm(ReLu(F_m W_{cls}^T)), \tag{1}$$

where $norm(\cdot)$ denotes $Min\text{-}Max$ normalization scaling A_m to $[0, 1]$, and $ReLu(\cdot)$ is used to eliminate negative values.

Subsequently, by introducing foreground and background thresholds (θ_l, θ_h) [18], [20], the uncertain region mask $\mathcal{M}_u \in \mathbb{R}^{h\times w}$ is obtained as follows:

$$\mathcal{M}_u^{i,j} = \begin{cases} 1, & if\ \theta_l < Max(A_t^{i,j,:}) < \theta_h, \\ 0, & otherwise, \end{cases} \tag{2}$$

where (i, j) represents the spatial indices, and $Max(\cdot)$ yields the maximum value along the channel dimension.

2) Uncertainty-Guided Data Augmentation: To improve the teacher network's attention to uncertain regions, we mask the input images to contain only these uncertain parts. Additionally, we mask the uncertain regions of the input images, enabling the network to infer the uncertainty spatial details based on the certain representations. Uncertain image I_u and certain image I_c can be obtained by:

$$I_u = UP(\mathcal{M}_u) \otimes I_{in},\ I_c = UP(1 - \mathcal{M}_u) \otimes I_{in}, \tag{3}$$

where $UP(\cdot)$ upsamples the mask to the spatial size of I_{in}. $\otimes$ denotes element-wise multiplication.

B. High-Frequency Spatial Context Mining

1) High-Frequency Boundary Self-Distillation: Given that uncertainty often arises at the boundaries of different semantics, we designed a High-Frequency Boundary Self-Distillation (HBS) strategy to more accurately divide uncertain regions by extracting high-frequency boundary contexts.

Specifically, input image I_{in} is input into the student network, where the encoder generates output features F_o and extracts shallow features F_l. Subsequently, F_o and F_l are fed into a boundary perception module (BPM). Inspired by [24], [25], we design the novel BPM module based on convolution to extract potential high-frequency information. As illustrated in Fig. 2, BPM initially fuses F_l and F_o to generate a boundary map $\mathcal{B}$, which is then combined with F_o to produce boundary-aware features F_b.

To enhance the teacher's fine-grained representation and provide boundary supervision of uncertain regions, we input uncertain image I_u into the teacher network, obtaining boundary-aware features $F_b^{'}$. With the uncertain region mask $\mathcal{M}_u$, HBS performs self-distillation of high-frequency boundary information through the following loss function:

$$\mathcal{L}_{hbs} = \frac{1}{|M|}\sum_{i\in M}(F_{b,i} - F_{b,i}^{'})^2, \tag{4}$$

where M represents the set of feature pixel indices in $\mathcal{M}_u$ with a response value of 1.

2) Uncertainty Spatial Reconstruction: To enhance the network's spatial perception of uncertain regions, we propose an Uncertainty Spatial Reconstruction (USR) strategy to infer uncertain areas based on salient representations, thereby improving uncertainty fine-grained spatial awareness. Specifically, image I_c containing only certain regions is input into the student network to extract features, which are then fed into a U-shaped reconstructor [24]. The reconstruction loss is defined as follows:

$$\mathcal{L}_{usr} = |RC(F_c) - I_{in}|_1, \tag{5}$$

where $RC(\cdot)$ represents the reconstructor and $|\cdot|_1$ denotes the L_1 norm.

C. Low-frequency Semantic-aware Contrastive Attention

To enhance low-frequency semantic learning and improve semantic discrimination, we propose Low-frequency Semantic-aware Contrastive Attention (LSCA), which guides the transformer attention to focus more on the foreground, ensuring precise object segmentation.

Specifically, for the i-th block in ViT, the multi-head attention averaged across all heads is denoted as $S_i \in \mathbb{R}^{(hw+1)\times(hw+1)}$, where $hw+1$ consists of hw patch tokens and one class token. As shown in Fig. 2, by averaging the attention across l blocks, we derive semantic attention $S \in \mathbb{R}^{(hw+1)\times(hw+1)}$. Then, we extract the attention corresponding to the class token from S to obtain foreground-aware attention $S_f \in \mathbb{R}^{1\times hw}$, which captures the attention between the class token and the other patch tokens.

Since class tokens embody foreground semantics [18], [26], S_f reflects the responses of patch tokens to the target foreground. To enhance foreground perception and reduce background activation, we reshape S_f to $\mathbb{R}^{h\times w}$ and utilize the CAM A_t from the teacher network to supervise S_f. Specifically, the foreground mask $\mathcal{M}_t \in \mathbb{R}^{h\times w}$ is first derived from $A_t \in \mathbb{R}^{h\times w\times K}$ as follows:

$$\mathcal{M}_t^{i,j} = \begin{cases} 1, & if\ Max(A_t^{i,j,:}) > \theta_h, \\ 0, & otherwise, \end{cases} \tag{6}$$

where (i, j) represents the spatial indices, and $Max(\cdot)$ yields the maximum value along the channel dimension.

To guide the attention blocks focus more on foreground semantics, we minimize the distance between the class token and foreground tokens while increasing the distance to background tokens. To prevent negative optimization from incorrect sample classification, we use the certain region mask $1 - M_u$ to exclude noisy samples. Finally, the contrastive attention loss $\mathcal{L}_{lsca}$ is defined as follows:

$$\mathcal{L}_{lsca} = -\sum_{i,j}(1 - \mathcal{M}_u^{i,j})\mathcal{M}_t^{i,j} log(S_f^{i,j}), \tag{7}$$

where (i, j) indexes the spatial locations.

D. Weakly-Supervised Training of SFS

Following a common practice, we utilize multi-label soft margin loss $\mathcal{L}_{cls}$ to train the classifiers and pixel-wise cross-entropy loss $\mathcal{L}_{seg}$ for the segmentation decoder. To enhance the diversity of the ViT and mitigate over-smoothing, we employ an affinity loss $\mathcal{L}_{aff}$ [18]. For the above common losses, we use $\mathcal{L}_{com}$ to represent them. $\mathcal{L}_{com}$ can be expressed as:

$$\mathcal{L}_{com} = \lambda_1\mathcal{L}_{cls} + \lambda_2\mathcal{L}_{aff} + \lambda_3\mathcal{L}_{seg}. \tag{8}$$

After introducing the proposed losses, the overall loss function for our SFS network is detailed as follows:

$$\mathcal{L}_{sfs} = \mathcal{L}_{com} + \lambda_4\mathcal{L}_{hbs} + \lambda_5\mathcal{L}_{usr} + \lambda_6\mathcal{L}_{lsca}, \tag{9}$$

where $\{\lambda_i\}_{i=1}^6$ represent the weighting factors.

Based on the end-to-end WSSS framework, our SFS enhances the network's spatial frequency awareness, improves CAM activation precision by reducing uncertainty, and refines segmentation outcomes.

III. Experiments

A. Experimental Setup

1) Dataset and Evaluation Metric: We validate our method on two benchmarks for WSSS, *i.e.*, VOC 2012 [27] and COCO 2014 [28], using mean Intersection over Union (mIoU) as the evaluation metric. The VOC 2012 dataset consists of 20 foreground classes and a background class. Following the common practice of previous works [18], [20], [22], we use the augmented SBD dataset [29], which includes 10,582 images for training, 1,449 for validation, and 1,456 for testing. The COCO 2014 dataset includes 80 categories and a background class, with 10,582 images for training, 1,449 for validation, and 1,456 for testing.

TABLE I
QUALITY OF CAM PSEUDO-LABELS ON THE PASCAL VOC 2012 *train* AND *val* DATASETS, EVALUATED USING MIOU AS THE METRIC.

Method	Backbone	*train*	*val*
ViT-PCM (ECCV'22) [30]	ViT-B	67.7	66.0
AFA (CVPR'22) [20]	MiT-B1	68. 7	66.5
ToCo (CVPR'23) [18]	ViT-B	73.6	72.3
CPAL (CVPR'24) [31]	ResNet38	75.8	-
DuPL (CVPR'24) [17]	ViT-B	76.0	74.1
SFS (ours)	ViT-B	**78.0**	**76.7**

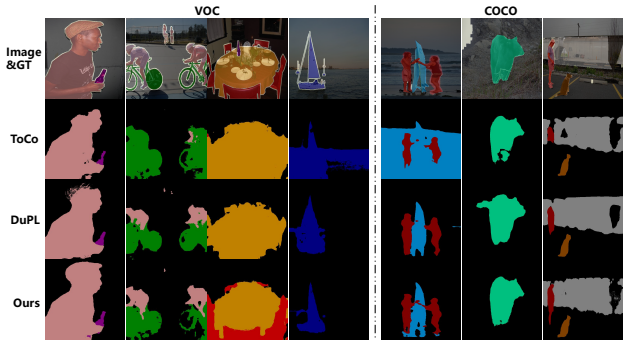

Fig. 3. Qualitative segmentation results of ToCo [18] , DuPL [17], and our proposed SFS.

TABLE II
SEGMENTATION RESULTS ON THE VOC *val* AND *test* DATASETS AND THE COCO *val* DATASET. $\mathcal{M}$ AND $\mathcal{S}$ DENOTE MULTI-STAGE METHODS AND SINGLE-STAGE METHODS, RESPECTIVELY.

Method	Type	Backbone	VOC		COCO
			val	*test*	*val*
MCTformer (CVPR'22) [16]	$\mathcal{M}$	ResNet38	71.9	71.6	42.0
FPR (ICCV'23) [35]		ResNet38	70.0	70.6	43.9
ACR (CVPR'23) [36]		ResNet38	71.9	71.9	45.3
SSC (TIP'24) [37]		ResNet101	72.7	72.8	38.1
CPAL (CVPR'24) [31]		ResNet38	72.5	72.9	42.9
MCTformer+ (TPAMI'24) [19]		ResNet38	74.0	73.6	45.2
AFA (CVPR'22) [20]	$\mathcal{S}$	MiT-B1	66.0	66.3	38.9
TSCD (AAAI'23) [21]		MiT-B1	67.3	67.5	40.1
ToCo (CVPR'23) [18]		ViT-B	71.1	72.2	42.0
DuPL (CVPR'24) [17]		ViT-B	73.3	72.8	44.6
SFS (ours)		ViT-B	**75.0**	**74.8**	**45.8**

TABLE III
ABLATION STUDY OF PROPOSED MODULES. CAM AND *Seg* REPRESENT THE MIOU OF CAM AND SEGMENTATION RESULTS, RESPECTIVELY.

baseline	LSCA	HBS	USR	CAM	*Seg*
✓				71.9	69.7
✓	✓			74.1	71.1
✓	✓	✓		76.5	73.1
✓	✓	✓	✓	76.7	73.8

2) Implementation Details: We use ViT-B [23] pretrained on ImageNet [32] as the encoder. To improve patch and class token correlation, we add two cross-attention layers after the encoder [22]. The student network is optimized with AdamW, starting with a base learning rate of 6e-5 and subsequently decaying following a cosine schedule. Images are randomly cropped to 448×448, and we adopt the multi-crop and data augmentation strategies described in [33]. The background threshold is set to (0.25, 0.7). The loss weight factors $\{\lambda_i\}_{i=1}^{6}$ are set as $(1.0, 0.2, 0.1, 0.1, 0.2, 0.1)$. We train the models for 20,000 iterations on VOC 2012 and 80,000 iterations on COCO 2014 with a batch size of 8. During testing, we employ multi-scale testing and Conditional Random Fields (CRF) for post-processing [34].

B. Comparison With State-of-the-Arts

1) Quality of Pseudo Labels: As shown in Table I, we assesse the quality of the CAMs pseudo masks generated by SFS and other recent competitors. The results demonstrate that our SFS achieves mIoU scores of 78.0% and 76.7% on the VOC *train* and *val* sets, respectively, closely approaching the ground truth and surpassing other methods.

2) Segmentation Performance: Table II compares the segmentation performance of our SFS against other WSSS methods on the VOC and COCO datasets. SFS surpasses other single-stage state-of-the-art methods by 1.7% and 2.0% mIoU on the VOC *val* and *test* sets, and by 1.2% mIoU on the COCO *val* set. SFS achieves superior segmentation accuracy while omitting complex training procedures, surpassing the state-of-the-art multi-stage methods.

Qualitative segmentation results on VOC and COCO are visualized in Fig. 3, which demonstrate that SFS provides clearer and more accurate multi-object boundaries, as well as improved foreground-background distinction.

C. Ablation Studies

As shown in Table III, we perform an ablation study of the proposed strategies on VOC *val* dataset. ViT with common losses $\mathcal{L}_{com}$ (Sec. II-D) is used as the baseline. It is notable that all ablation experiments are conducted without post-processing. The results demonstrate that our LSCA, HBS, and USR strategies collectively enhance both CAM and segmentation performance. The integration of these strategies equips the network with both spatial high-frequency and low-frequency semantic awareness, leading to improved performance.

IV. CONCLUSION

In this work, we introduce a novel Spatial Frequency-Aware Self-Distillation (SFS) strategy for accurately dividing uncertain regions. Firstly, we design a self-distillation strategy to extract boundary high-frequency information and then apply deterministic representations to reconstruct these uncertain regions, enhancing the network's spatial high-frequency perception. Secondly, we propose a semantic-aware contrastive attention to focus more on the low-frequency foreground semantic and distinguish it from the background. Finally, Combining these strategies, our SFS method achieves superior spatial context perception of both high and low frequencies, outperforming other state-of-the-art methods on the VOC and COCO datasets.

REFERENCES

[1] L. Xu, M. Bennamoun, F. Boussaid, S. An, and F. Sohel, "An improved approach to weakly supervised semantic segmentation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 1897–1901.

[2] L. Chen, W. Wu, C. Fu, X. Han, and Y. Zhang, "Weakly supervised semantic segmentation with boundary exploration," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*. Springer, 2020, pp. 347–362.

[3] L. Xu, M. Bennamoun, F. Boussaïd, S. An, and F. Sohel, "An improved approach to weakly supervised semantic segmentation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1897–1901.

[4] Z. Xie and H. Lu, "Exploring category consistency for weakly supervised semantic segmentation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2609–2613.

[5] W. Wu, T. Dai, X. Huang, F. Ma, and J. Xiao, "Image augmentation with controlled diffusion for weakly-supervised semantic segmentation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 6175–6179.

[6] J. Lee, E. Kim, and S. Yoon, "Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4071–4080.

[7] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5267–5276.

[8] X. Li, T. Zhou, J. Li, Y. Zhou, and Z. Zhang, "Group-wise semantic mining for weakly supervised semantic segmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 3, 2021, pp. 1984–1992.

[9] S. Lee, M. Lee, J. Lee, and H. Shim, "Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5495–5505.

[10] Z. Chen, T. Wang, X. Wu, X.-S. Hua, H. Zhang, and Q. Sun, "Class re-activation maps for weakly-supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 969–978.

[11] J. Lee, S. J. Oh, S. Yun, J. Choe, E. Kim, and S. Yoon, "Weakly supervised semantic segmentation using out-of-distribution data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 897–16 906.

[12] J. Li, Z. Jie, X. Wang, X. Wei, and L. Ma, "Expansion and shrinkage of localization for weakly-supervised semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 037–16 051, 2022.

[13] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2209–2218.

[14] P.-T. Jiang, Y. Yang, Q. Hou, and Y. Wei, "L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 886–16 896.

[15] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 275–12 284.

[16] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu, "Multi-class token transformer for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4310–4319.

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

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

[19] L. Xu, M. Bennamoun, F. Boussaid, H. Laga, W. Ouyang, and D. Xu, "Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, 2024.

[20] L. Ru, Y. Zhan, B. Yu, and B. Du, "Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 846–16 855.

[21] R. Xu, C. Wang, J. Sun, S. Xu, W. Meng, and X. Zhang, "Self correspondence distillation for end-to-end weakly-supervised semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3045–3053.

[22] J. He, L. Cheng, C. Fang, Z. Feng, T. Mu, and M. Song, "Progressive feature self-reinforcement for weakly supervised semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 2085–2093.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[24] Y. Sun, S. Wang, C. Chen, and T.-Z. Xiang, "Boundary-guided camouflaged object detection," *arXiv preprint arXiv:2207.00794*, 2022.

[25] L. Chen, L. Gu, D. Zheng, and Y. Fu, "Frequency-adaptive dilated convolution for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3414–3425.

[26] Z. Chen, J. Ding, L. Cao, Y. Shen, S. Zhang, G. Jiang, and R. Ji, "Category-aware allocation transformer for weakly supervised object localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6643–6652.

[27] M. Everingham and J. Winn, "The pascal visual object classes challenge 2011 (voc2011) development kit," *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep*, vol. 8, 2011.

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[29] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *2011 international conference on computer vision*. IEEE, 2011, pp. 991–998.

[30] S. Rossetti, D. Zappia, M. Sanzari, M. Schaerf, and F. Pirri, "Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation," in *European conference on computer vision*. Springer, 2022, pp. 446–463.

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

[32] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "Imagenet-21k pretraining for the masses," *arXiv preprint arXiv:2104.10972*, 2021.

[33] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[34] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[35] L. Chen, C. Lei, R. Li, S. Li, Z. Zhang, and L. Zhang, "Fpr: False positive rectification for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1108–1118.

[36] H. Kweon, S.-H. Yoon, and K.-J. Yoon, "Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 329–11 339.

[37] T. Chen, Y. Yao, X. Huang, Z. Li, L. Nie, and J. Tang, "Spatial structure constraints for weakly supervised semantic segmentation," *IEEE Transactions on Image Processing*, 2024.