AUT Journal of Electrical Engineering
10.22060/EEJ.2024.23490.5616

# Leveraging Swin Transformer for Local-to-Global Weakly Supervised Semantic Segmentation

Rozhan Ahmadi[1], Shohreh Kasaei[2]

[1] Masters of Computer Engineering, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran. (roz.ahmadi@sharif.edu)
[2] Professor of Artificial Intelligence, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran. (kasaei@sharif.edu)

**Abstract:**

Recent advancements in Weakly Supervised Semantic Segmentation have highlighted the use of image-level class labels as a form of supervision. Many methods use pseudo-labels from Class Activation Maps to address the limited spatial information in class labels. However, Class Activation Maps generated from Convolutional Neural Networks are often led to focus on prominent features, making it difficult to distinguish foreground objects from their backgrounds. While recent studies show that features from Vision Transformers are more effective in capturing the scene layout than Convolutional Neural Networks, the use of hierarchical Vision Transformers has not been widely studied in Weakly Supervised Semantic Segmentation. This work introduces "SWTformer" and explores the effect of Swin Transformer's local-to-global view on improving the accuracy of initial seed Class Activation Maps. SWTformer-V1 produces Class Activation Maps solely based on patch tokens as its input features. SWTformer-V2 enhances this process by integrating a multi-scale feature fusion mechanism and employing a background-aware mechanism that refines the accuracy of localization maps, resulting in better differentiation between objects. Experiments on the Pascal VOC 2012 dataset demonstrate that compared to state-of-the-art models, SWTformer-V1 achieves 0.98% mAP higher in localization accuracy and generates initial localization maps that are 0.82% mIoU higher in accuracy while relying solely on the classification network. SWTformer-V2 enhances the accuracy of the seed Class Activation Maps by 5.32% mIoU. Code available at: https://github.com/RozhanAhmadi/SWTformer

**Keywords:** Weakly Supervised Semantic Segmentation, Class Activation Map, Hierarchical Vision Transformer, Image-level label

## 1. Introduction

Semantic segmentation is an important task in computer vision where every pixel in an image is classified. Although advancements in fully supervised learning have highly improved results in this area, manually annotating images at the pixel level is labor-intensive and expensive.

In recent years, weakly supervised semantic segmentation (WSSS) has emerged as a solution to lower annotation costs. This method trains segmentation models using weak labels (bounding boxes [1], scribble [2, 3], points [4], and image-level labels). Image-level labels are the most commonly used in WSSS due to their ease of annotation, despite lacking detailed spatial information about objects. To tackle this, many approaches use a three-step pipeline. These labels are involved in generating seed Class Activation Maps (CAMs) from an image classification model, which highlight key object parts [5]. These seeds are then refined to create pseudo-labels, which are used to train a fully supervised segmentation network. The success of this approach heavily relies on the quality of the initial seed CAMs, making them a critical focus of research efforts.

Convolutional neural networks (CNNs) are commonly used for WSSS but struggle with capturing complete object regions, Fig. 1 (a), due to their limited local perception. In comparison to CNNs, Vision Transformers (ViTs) are able to capture  long-range dependencies for a more global understanding of scenes, Fig. 1 (b). However, switching from CNNs to ViTs can result in losing fine details while gaining better coverage of large objects. Hierarchical Vision Transformers (HVTs) combine the strengths of both CNNs and ViTs by generating feature maps at multiple resolutions. This allows them to capture both local and global context effectively, Fig. 1 (c), making them suitable for accurate multi-scale object localization. Despite their potential, HVTs have not yet been applied in WSSS.

This research presents a new method, SWTformer, to explore the validity of this concept. SWTformer-V1, which utilizes Swin Transformer [6] as its backbone classifier network, is supervised by image-level labels. This presents challenges since Swin Transformer relies on patch tokens instead of class tokens

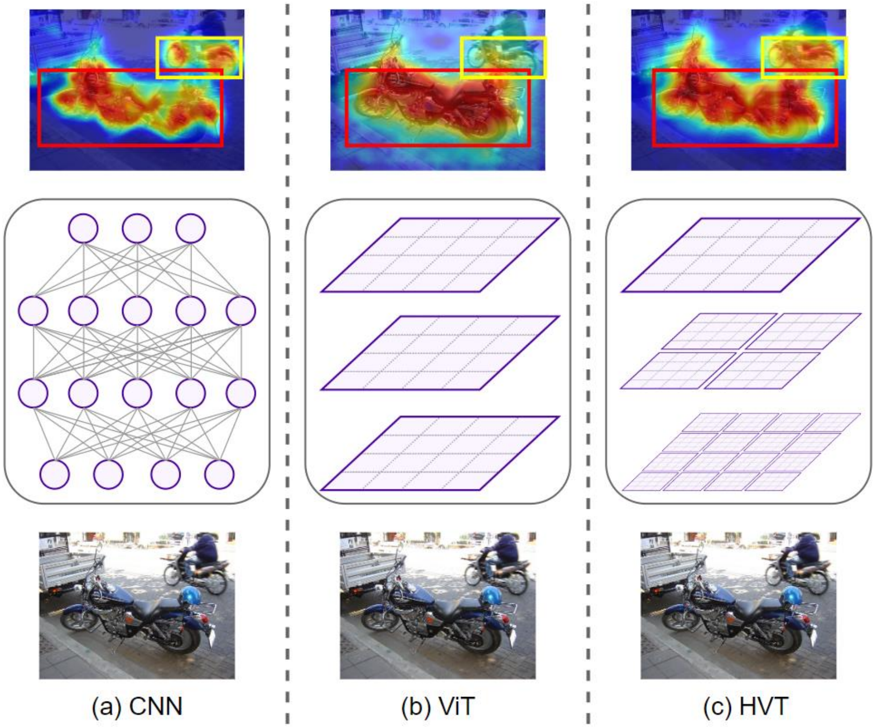

Fig. 1. Class activation maps generated by a (a) CNN (Resnet-50), (b) ViT (DeiT-S) and (c) HVT (Swin-T). Red and yellow boxes indicate the large and small scale objects relative to the image size.

commonly used in WSSS. The shifted window mechanism of Swin Transformer also requires careful tuning. In order to generate more accurate CAMs, previous ViTs have benefited from Attention Roll-Out [7] which is a mechanism that aggregates attention maps derived from the layers. Attention Roll-Out facilitates a more nuanced analysis of the attention flow present within the network. Despite being effective, this method is not directly applicable to Swin Transformer due to its shifted window mechanism and hierarchical multi-scale features. To address these challenges and enhance contextual understanding, SWTformer-V2 proposes a multi-scale feature fusion module within a background-aware refinement mechanism. This aims to produce more accurate localization masks with improved discrimination across objects.

The main contributions of this work are:

- Introducing SWTformer, the first hierarchical transformer-based solution for generating initial Class Activation Maps (CAMs) in weakly supervised semantic segmentation (WSSS). This approach addresses the limitations of CNNs' local receptive fields and ViTs' global scene views.
- Presenting SWTformer-V1, which utilizes the Swin Transformer as a backbone for classification and initial CAM generation using only patch tokens.
- Developing SWTformer-V2 to overcome the challenges of applying Attention Roll-Out to the Swin Transformer architecture, proposing a solution that incorporates hierarchical feature fusion and a background-aware refinement mechanism.
- Validating the effectiveness of the proposed methods through extensive experiments on the Pascal VOC 2012 dataset.

## 2. RELATED WORK

### 2-1- Vision Transformers

In recent years, Vision Transformers (ViTs) have significantly revolutionized the field of computer vision. ViT [8] is a deep learning model that transforms an input image into a sequence of patch tokens plus a class token that represents the entire image, and analyzes the visual data using multi-head self-attention blocks. This self-attention mechanism allows ViT to capture global information and long-range dependencies in the data. DeiT [9] builds on ViT by introducing new data augmentation methods and a distillation token. Although ViTs succeed in capturing global context, they have limitations in capturing local details. Conformer [10] addresses ViT’s limitation in capturing local details by combining a CNN branch with a ViT branch, although this integration requires significant training adjustments and computational resources. Hierarchical Vision Transformers (HVTs), namely T2T [11] and PVT [12], provide an effective solution by bringing the strengths of ViTs and CNNs together. Their pipeline starts from fine-grained local details and moves towards long-range global dependencies. Swin Transformer [6] utilizes a novel patch merging module and a shifted window self-attention mechanism. This approach allows smaller groups of patches to be mixed together, enabling the model to capture long-range feature dependencies more accurately.

### 2-2- Weakly Supervised Semantic Segmentation with CNNs

Recent studies on WSSS mostly use image-level labels for supervision and rely on Class Activation Maps (CAM) to localize objects. Methods that utilize a CNN as the classification backbone, generate the seed localization maps by calculating a CAM for each class through a weighted combination of the feature maps in the last layer of a CNN [5]. While CAMs are capable of visualizing the most discriminative regions of an image, they have limitations in comprehensively activating objects and distinguishing them from the background. Post-processing methods such as PSA [13] and IRN [14] have further refined the initial CAMs through iterative seed region growing. As the performance of WSSS is heavily dependent on the quality of the initial CAMs, various techniques have been explored to improve the accuracy of the initial activation maps. These methods include adversarial erasing [15-17], cross-affinity extraction modules and contrastive learning [18-21], prototype-based learning [22-24], attention mechanism [25, 26] and self-supervised learning [23, 27]. Recent research [28-33], pioneered by CLIMS [34], has also explored using language models such as CLIP [35] in order to extract further context from an image by matching the corresponding label prompts in the CLIP embedding space. With the emergence of SAM [36] in the field of full-supervised semantic segmentation, some recent works such as [37-40] have investigated the effect of combining the features extracted from this model with conventional methods in WSSS. Knowledge distillation is another popular field that has been utilized in SeCo [41] to mitigate the issue of frequent object co-occurrence in images.

### 2-3- Weakly Supervised Semantic Segmentation with ViTs

With Vision Transformers (ViTs) making significant progress in various tasks, some recent works have utilized them for WSSS. AFA [42] proposes refining initial pseudo labels using global semantic affinity learned from self-attention. MCTformer [43] replaces ViT’s singular class token with multiple tokens, each corresponding to a particular semantic class. It also employs patch affinity learned from attention maps to refine the initial CAMs. ViT-PCM [44] proposes an end-to-end CAM independent framework relying on ViT's spatial characteristics. ToCo [45] addresses the over-smoothing issues of ViTs by using the model's intermediate knowledge to supervise its output features. TransCam [46] adopts Conformer [10] by proposing to use the attention weights of the ViT branch to refine the CAMs generated from the CNN branch. A recent work, CTI [47] has focused on class tokens and proposed infusion methods to improve CAM consistency within classes. DuPL [48] integrates two ViT subnets to provide supervision for one another while also developing regularization on discarded regions.

It is worth mentioning that the majority of these studies rely on class tokens, inspired by observations made in DINO [49] that class tokens and their attention to patch tokens contain useful knowledge regarding the semantic layout of a scene. Hierarchical Vision Transformers are a rather recent development in the field of vision transformers and have not yet been introduced to WSSS. HVTs are expected to capture scene layout more effectively than CNNs and ViTs. However, their specific impact in the context of WSSS remains an open area for research.

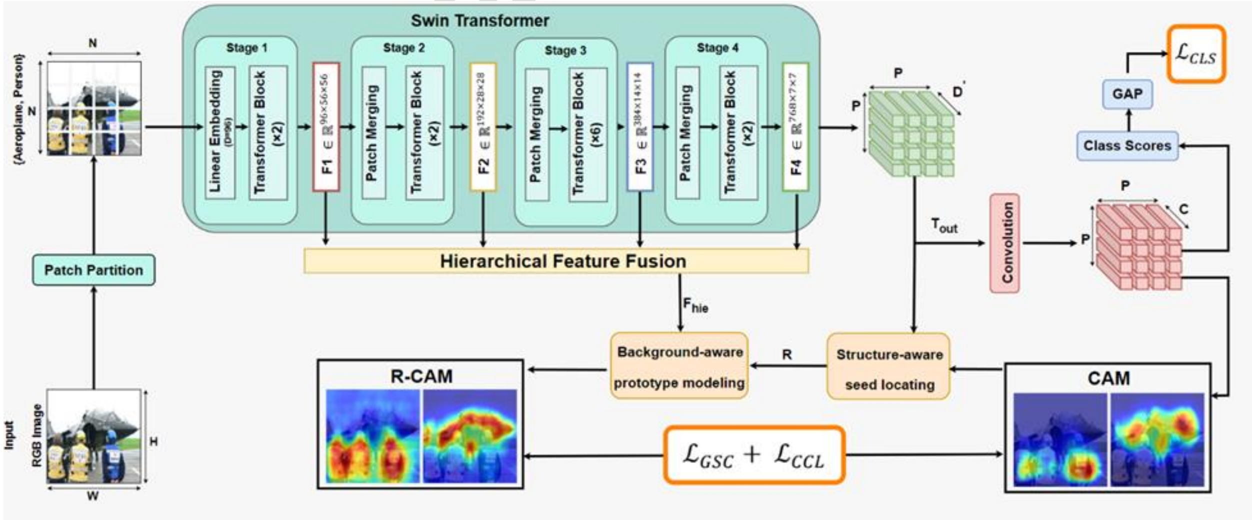

Fig. 2. An overview of the proposed SWTformer (V2). The backbone is the Swin-T version of the Swin Transformer and the training of the model is optimized by the *CLS*, *GSC* and *CCL* loss functions. The "Structure-aware seed locating" and "Background-aware prototype modeling" modules are adopted from SIPE [23] with modifications.

## 3. PROPOSED METHOD

### 3-1- Overview

This paper introduces SWTformer, illustrated in Fig. 2, a novel framework that utilizes the Swin Transformer as the classifier backbone to generate initial localization maps for WSSS. Moreover, Swin Transformer's multiscale contextual information is utilized through a novel Hierarchical Feature Fusion (HFF) module within a background-aware prototype exploration mechanism based on SIPE [23].

### 3-2- Generating Class Activation Maps from Patch Tokens

Unlike traditional ViTs, Swin Transformer (SWTformer-V1) utilizes only patch tokens without class tokens. Inspired by [43], SWTformer-V1 incorporates a CAM module to generate activation seeds and class scores from Swin's output patch tokens for classifier training. Swin encodes an input image $I \in \mathbb{R}^{3\times H\times W}$ by partitioning it into $N\times N$ patches and later projecting them into tokens $T \in \mathbb{R}^{D\times N\times N}$ with $D$ being the embedding dimension. A patch merging module connects subsequent transformer blocks, doubling the embedding dimension and halving the patch size, resulting in an output token sequence of $T_{out} \in \mathbb{R}^{D^{'}\times N\times N}$, where $D^{'} = 8D$ and $P = N/8$. To generate CAMs for $C$ classes, $T_{out}$ is converted to a to a $2D$ feature map $F_{out} \in \mathbb{R}^{C\times P\times P}$. Since $F_{out}$ may contain negative values, a ReLU function is applied to $F_{out}$, followed by a feature normalization function. This process results in feature maps $C_{out} \in \mathbb{R}^{C\times P\times P}$ and can be summarized as Eq. (1)

$$\begin{aligned} F_{out} &= \mathrm{Conv}_{\mathrm{2D}}(T_{out}) \\ C_{out} &= \mathrm{Norm}(\mathrm{ReLU}(F_{out})) \end{aligned} \quad (1)$$

$C_{out}$ is then upsampled to the size of the original image, producing the initial class activation maps $M \in \mathbb{R}^{C\times H\times W}$.

### 3-3- Multi-label Classification Training

For multi-label classification, global average pooling is applied to $F_{out}$, generating class scores $s_c \in \mathbb{R}^C$ for $c \in \{1,..,C\}$ semantic categories. The classification loss, $L_{CLS}$, is then calculated by averaging over multi-label soft margin losses over all $C$ classes as Eq. (2),

$$L_{CLS} = \frac{1}{C}\sum_{i=1}^{C} \hat{s}_c \log(s_c) + (1-\hat{s}_c)\log(1-s_c). \tag{2}$$

where $s_c$ and $\hat{s}_c$ represent the predicted score and its corresponding ground-truth, respectively. This loss helps with the training optimization of the classifier by using image-level labels as supervision.

### 3-4- Hierarchical Feature Fusion

SWTformer-V1 uses patch tokens to compute class scores and generate CAMs, but its hierarchical structure makes combining attention maps from intermediate layers challenging. To address this, SWTformer-V2 introduces a Hierarchical Feature Fusion (HFF) module, which leverages Swin's multiscale contextual information, instead of combining attention maps to learn semantic patch affinity. In the field of deep learning, leveraging feature maps from both the final and intermediate layers of a hierarchical network is a well-established strategy [50], [51]. Inspired by this, the suggested approach takes advantage of the distinct characteristics of information captured at different stages of the network. Shallow layers, which are closer to the input data, are capable of identifying low-level granular local features such as edges, texture and color. On the other hand, deep layers, as the network hierarchy is ascended, are capable of recognizing more abstract, high-level features and complex patterns. By fusing feature maps from both shallow and deep layers, the model can harness a comprehensive range of information, from simple to complex patterns. This fusion does not add significant computational overhead and maximizes semantic knowledge from all four transformer blocks. Fig. 3 illustrates the proposed hierarchical feature fusion method. The HFF module extracts the output patch features from all four transformer blocks and concatenates them in two stages to maximize the semantic knowledge obtained. The upsampling in this module is achieved through bilinear interpolation, while the downsampling is performed using a convolutional layer. This module is specifically designed to be compatible with the Swin Transformer and outputs a new feature map $F_{hie}$ that contains the combined local to global semantic contexts of the scene.

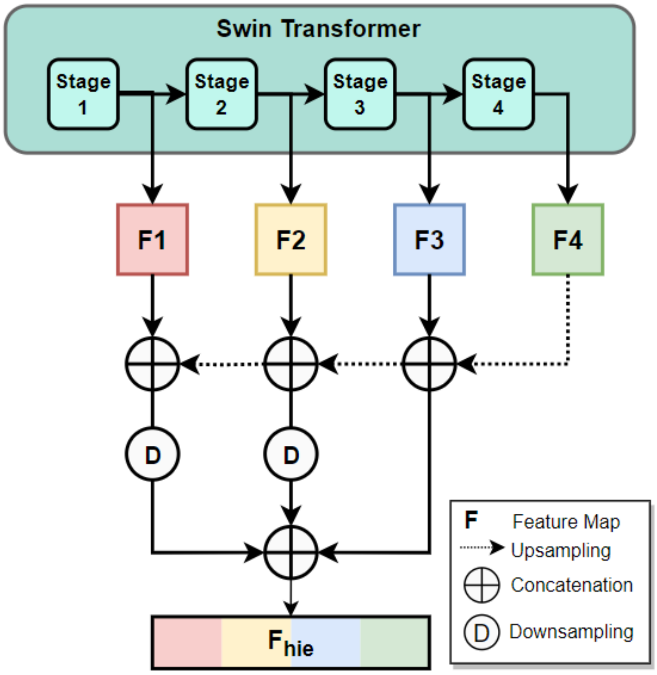

Fig. 3. Illustration of the proposed hierarchical feature fusion (HFF) module in SWTformer.

### 3-5- Background-aware Prototype Exploration

SWTformer-V2 builds the HFF module on SWTformer-V1 by employing it in a background-aware prototype exploration mechanism. The primary goal of SWTformer-V2 is to refine the initially generated CAMs from SWTformer-V1, enabling the model to create more comprehensive object regions and accurately distinguish foreground objects from the background. In order to leverage the semantic knowledge encapsulated in $F_{hie}$ for enhancing the initially generated CAMs and generating masks with more comprehensive object activation, SWTformer-V2 adapts and modifies the utilization strategy demonstrated by SIPE [23], an architecture based on a CNN. Given $M \in \mathbb{R}^{C\times H\times W}$ for $C$ foreground classes, to enhance the model's awareness of the background, an activation map $M_B$ is estimated as Eq. (3)

$$M_B = 1 - \max_{1\le c\le C} M_c \tag{3}$$

$M_B$ is then concatenated to the initial foreground CAMs, making $M \in \mathbb{R}^{(C+1)\times H\times W}$.

In the next step, $T_{out}$ and $M$ are input to a modified version of the structure-aware seed locating module from SIPE [23]. This module generates seed maps $R \in \mathbb{R}^{(C+1)\times P\times P}$ for each class category $C+1$, including the background. The module operates by calculating the cross-token semantic affinity map $S_c$ from $T_{out}$ to capture each token's spatial structure. It then compares the similarity between each token's spatial structure in $S_c$ with the class activations in $M$ and assigns that token the class label to which it has the most structural similarity. In contrast to the original method, SWTformer-V2 produces the cross-token semantic similarity map $S_c$ by calculating the cosine similarity as Eq. (4)

$$S_c(T_{out}) = \left|\mathrm{Cos}\,Sim(T_{out}, T_{out})\right| = \left|\frac{T_{out}\cdot(T_{out})^T}{\left\|T_{out}\right\|\left\|(T_{out})^T\right\|}\right|, \tag{4}$$

where $\cdot$ denotes the dot product. The use of the absolute value of the similarity is motivated by experiments showing that even a negative value similarity between two tokens represents a high structural correlation between them.

In the final step of this framework, the generated seed maps $R$ and the hierarchical feature $F_{hie}$ are passed to the background-aware prototype modeling module from SIPE [23]. This module first creates image-specific prototypes $P_c$ for all $C+1$, classes, which are equivalent to the centroid of $R$ for each class in the feature space of $F_{hie}$. Lastly, the refined image-specific CAMs, R-CAMs, are generated from the correlation calculated between $P_c$ and $F_{hie}$.

To ensure consistency between the initial CAM and the refined R-CAM, the utilization of a normalization loss is suggested by the original paper as Eq. (5),

$$L_{GSC} = \frac{1}{C+1}\left\| CAM - R\_CAM \right\|_1 . \tag{5}$$

SWTformer-V2 proposes to use a Class-wise Contrastive Loss ($CCL$) in addition to the $GSC$ loss. The $CCL$ loss aims to enhance the generation of comprehensive initial CAMs at each step, building on the R-CAMs generated in the previous step. It achieves this by optimizing the model to minimize the distance between the representations of similar classes and maximize the distance between representations of dissimilar classes, represented in CAM and R-CAM. The $CCL$ loss is calculated as Eq. (6)

$$L_{CCL} = \frac{1}{2}\left[ (\frac{2}{3} \times \mathrm{Cos}\, Sim(CAM, R\_CAM))^2 + (1 - \mathrm{Cos}\, Sim(CAM, R\_CAM))^2 \right]. \tag{6}$$

In summary, the overall loss for optimizing the model training includes the $CLS$, $GSC$ and $CCL$ loss functions as Eq. (7)

$$L_{Total} = L_{CLS} + L_{GSC} + L_{CCL} . \tag{7}$$

# 4. EXPERIMENTS

## 4-1- Dataset

The proposed method is evaluated on the PASCAL VOC 2012 [52] dataset, a widely used benchmark for image classification and segmentation, particularly in WSSS. The dataset consists of 20 classes, 1,464 training images, 1,449 for validation, and 1,456 for testing. Additionally, an augmented set of 10,582 images is added from [53] for training, following standard practices in semantic segmentation.

### 4-2- Evaluation Metrics

The mean Average Precision (mAP) metric is used to evaluate the classifier model's localization accuracy. Additionally, mean Intersection-over-Union (mIoU) is employed to measure the accuracy of the generated class activation maps.

### 4-3- Implementation Details

SWTformer is built with Swin-T [6] pre-trained on ImageNet [54] as its backbone. Images are cropped to 224 × 224 for training, and data augmentation is done following [55]. The model is trained using the AdamW optimizer with a batch size of 16 on two Nvidia T4 GPUs. Seed maps are equivalent to the refined CAMs, R-CAM.

### 4-4- Experimental Results

#### 4-4-1 Improvement on object localization

The main objective of this study is to investigate the impact of using the Swin Transformer as the classification backbone for WSSS in localizing objects supervised by image-level labels and generating CAMs. Table 1 compares the localization accuracy of the Swin Transformer used in SWTformer with DeiT-S [9], which is commonly employed in state-of-the-art WSSS methods using a vision transformer as the backbone. Specifically, the localization results of DeiT-S utilized in MCTformer [43] are considered for comparison. The results show that using Swin-T outperforms DeiT-S as a backbone for WSSS by 0.98%, demonstrating the effectiveness of Swin’s local-to-global view in localizing objects.

In terms of computational complexity, fine-tuning Swin-T took approximately one hour for both versions of SWTformer, converging in 30 epochs. In contrast, fine-tuning DeiT-S required twice as long, about two hours and 60 epochs to converge, demonstrating Swin-T's advantage in both localization accuracy and efficiency.

Table 1. COMPARISON OF OBJECT LOCALIZATION ON PASCAL VOC 2012 DATASET. † DENOTES OUR IMPLEMENTATION.

| Method | Backbone | mAP (%) |
|---|---|---|
| MCTformer-V1† [43] | DeiT-S | 95.62 |
| MCTformer-V2† [43] | DeiT-S | 95.47 |
| SWTformer-V1 | Swin-T | 96.49 |
| **SWTformer-V2** | **Swin-T** | **96.60** |

### 4-4-2 Evaluation of seed localization maps

Given that the generation of seed CAMs is the most crucial step in WSSS, this study aimed to propose a framework for utilizing the Swin Transformer in this process. A comparison of the proposed method with other state-of-the-art methods is presented in Table 2, demonstrating the mIoU accuracy of the seed maps. To ensure an accurate comparison and evaluate the effectiveness of the Swin Transformer, results from other approaches that rely solely on the backbone are utilized. The comparison reveals that SWTformer V1 achieves an average of 0.82% mIoU accuracy higher than other methods, demonstrating the method’s comparable performance. Furthermore, SWTformerV2 improves upon SWTformer-V1 by 5.32% in mIoU, therefore demonstrating the effectiveness of the strategies proposed to address the limitation of using attention maps from the Swin Transformer for refinement.

Table 2. EVALUATION OF THE INITIAL SEED LOCALIZATION MAPS (SEED) ON THE PASCAL VOC 2012 TRAIN SET IN TERMS OF MIOU (%).

| Method | Backbone | Seed |
|---|---|---|
| PSA [13] | VGG-16 | 48.00 |
| IRN [14] | ResNet-50 | 48.30 |
| SEAM [26] | ResNet-38 | 47.43 |
| SC-CAM [27] | ResNet-101 | 50.90 |
| SIPE [23] | ResNet-50 | 50.10 |
| MCTformer-V1 [43] | DeiT-S | 47.20 |
| MCTformer-V2 [43] | DeiT-S | 48.51 |
| TransCAM [46] | Conformer | 51.70 |
| **SWTformer-V1** | **Swin-T** | **49.84** |
| **SWTformer-V2** | **Swin-T** | **55.16** |

### 4-4-3 Qualitative results

The effectiveness of the proposed approach is further confirmed through various qualitative evaluations of the model's performance. Fig. 4 visualizes refined seed class activation maps (R-CAM) generated by SWTformer for various categories.

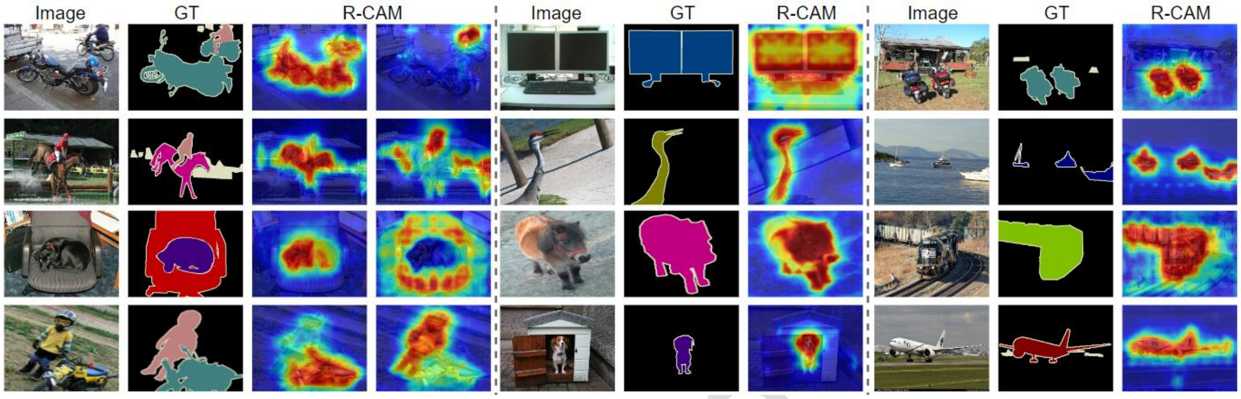

Fig. 4. Qualitative results of the class activation maps generated by SWTformer on PASCAL VOC 2012 train set. Images contain singular or multiple class labels.

### 4-5- Ablation Studies

The training procedure of the proposed model is optimized by the sum of three loss functions, termed $CLS$ loss, $GSC$ loss, and $CCL$ loss. An analysis of the impact of each of these loss functions on the enhancement of SWTformer is presented in Table 3. Experiments demonstrate that the simultaneous use of these three losses results in the best accuracy. Specifically, $CLS$ is responsible for classification, $GSC$ provides consistency between the two sets of CAMs, and $CCL$ further balances these modules.

Table 3. ABLATION STUDY ON THE EFFECTIVENESS OF THE PROPOSED LOSS FUNCTIONS ON THE ACCURACY OF THE SEED MAP.

| $L_{CLS}$ | $L_{GSC}$ | $L_{CCL}$ | mIoU (%) |
|---|---|---|---|
| ✓ | | | 49.84 |
| ✓ | ✓ | | 54.58 |
| ✓ | ✓ | ✓ | **55.16** |

## 5. Conclusion

This paper introduces SWTformer, a novel approach that uses the Swin Transformer as a backbone for weakly supervised semantic segmentation (WSSS). SWTformer-V1 effectively captures both local details and global structure through the Swin Transformer's hierarchical flow. However, due to the challenges introduced by the Swin Transformer's shifted window and multi-scale feature mechanisms, which limit direct use of the transformer's attention flow for refining activation maps (a common approach in non-hierarchical strategies), SWTformer-V2 introduces a hierarchical feature fusion module to capture multi-scale semantic knowledge. It also refines activation maps through a modified background-aware mechanism. SWTformer outperforms state-of-the-art transformers in object localization and yields comparable results to other approaches in generating seed activation maps. The strategies introduced in SWTformer-V2 further enhance this framework, refining initial activation maps to cover object regions more comprehensively. Future work aims to further leverage Swin Transformer's attention mechanisms to unlock its full potential for refined activation map generation and object localization.

## 6. REFERENCES

[1] K.H. J. Dai, and J. Sun, Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation, in:  Proceedings of the IEEE/CVF International Conference on Computer Vision, 2015, pp. 1635–1643.
[2] J.D. D. Lin, J. Jia, K. He, and J. Sun, Scribblesup: Scribble-supervised convolutional networks for semantic segmentation, in:  Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 3159–3167.
[3] L.a.Z. Wu, Zhun and Fang, Leyuan and He, Xingxin and Liu, Qiang and Ma, Jiayi and Chen, Hao, Sparsely annotated semantic segmentation with adaptive gaussian mixtures, in:  Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15454--15464.
[4] O.R. A. Bearman, V. Ferrari, and L. Fei-Fei, What's the point: Semantic segmentation with point supervision, in:  Proceedings of the European conference on computer vision, 2016, pp. 549–565.
[5] A.K. B. Zhou, A. Lapedriza, A. Oliva, and A. Torralba, Learning deep features for discriminative localization, in:  Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
[6] Y.L. Z. Liu, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B.Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in:  Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[7] S.A.a.W. Zuidema, Quantifying attention flow in transformers, in:  arXiv preprint arXiv:2005.00928, 2020.

[8] L.B. A. Dosovitskiy, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.

[9] M.C. H. Touvron, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, Training data-efficient image transformers & distillation through attention, in: International conference on machine learning, 2021, pp. 10347–10357.

[10] Z.a.H. A. Peng, Wei and Gu, Shanzhi and Xie, Lingxi and Wang, Yaowei and Jiao, Jianbin and Ye, Qixiang, Conformer: Local features coupling global representations for visual recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 367-376.

[11] Y.C. L. Yuan, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, Tokens-to-token vit: Training vision transformers from scratch on imagenet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 558–567.

[12] E.X. W. Wang, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 568–578.

[13] J.A.a.S. Kwak, Learning pixel-level semantic affinity with imagelevel supervision for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4981–4990.

[14] S.C. J. Ahn, 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.

[15] J.F. Y. Wei, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, Object region mining with adversarial erasing: A simple classification to semantic segmentation approach, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 1568–1576.

[16] E.K. J. Lee, S. Lee, J. Lee, and S. Yoon, Ficklenet: Weakly and semisupervised semantic image segmentation using stochastic inference, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5267–5276.

[17] T. Chen, X. Jiang, G. Pei, Z. Sun, Y. Wang, Y. Yao, Knowledge Transfer with Simulated Inter-Image Erasing for Weakly Supervised Semantic Segmentation, in: Proceedings of the European conference on computer vision, Springer, 2025, pp. 441-458.

[18] M.Z. T. Zhou, F. Zhao, and J. Li, Regional semantic contrast and aggregation for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4299– 4309.

[19] L.a.Z. Wu, Zhun and Ma, Jiayi and Wei, Yunchao and Chen, Hao and Fang, Leyuan and Li, Shutao, Modeling the Label Distributions for Weakly-Supervised Semantic Segmentation, in: arXiv preprint arXiv:2403.13225, 2024.

[20] J. Fan, Z. Zhang, T. Tan, C. Song, J. Xiao, Cian: Cross-image affinity net for weakly supervised semantic segmentation, in: Proceedings of the AAAI conference on artificial intelligence, 2020, pp. 10762-10769.

[21] G. Sun, W. Wang, J. Dai, L. Van Gool, Mining cross-image semantics for weakly supervised semantic segmentation, in: Proceedings of the European conference on computer vision, Springer, 2020, pp. 347-365.

[22] Z.F. Y. Du, Q. Liu, and Y. Wang, Weakly supervised semantic segmentation by pixel-to-prototype contrast, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4320–4329.

[23] L.Y. Q. Chen, J.-H. Lai, and X. Xie, Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4288–4298.

[24] F. Tang, Z. Xu, Z. Qu, W. Feng, X. Jiang, 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.

[25] J.H. T. Wu, G. Gao, X. Wei, X. Wei, X. Luo, and C. H. Liu, Embedded discriminative attention mechanism for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16765–16774.

[26] J.Z. Y. Wang, 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. 12275–12284.

[27] Q.W. Y.-T. Chang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, and M.H. Yang, Weakly-supervised semantic segmentation via sub-category exploration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8991–9000.

[28] M.C. Y. Lin, W. Wang, B. Wu, K. Li, B. Lin, H. Liu, and X. He, Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15305–15314.

[29] S.a.Z. Deng, Wei and Xie, Jinheng and Shen, Linlin, Qa-clims: Question-answer cross language image matching for weakly supervised semantic segmentation, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 5572--5583.

[30] B.a.H. Murugesan, Rukhshanda and Bhattacharya, Rajarshi and Ben Ayed, Ismail and Dolz, Jose, Prompting classes: exploring the power of prompt class learning in weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 291--302.

[31] B.a.Y. Zhang, Siyue and Wei, Yunchao and Zhao, Yao and Xiao, Jimin, Frozen CLIP: A Strong Backbone for Weakly Supervised Semantic Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3796--3806.

[32] L.a.O. Xu, Wanli and Bennamoun, Mohammed and Boussaid, Farid and Xu, Dan, Learning multi-modal class-specific tokens for weakly supervised dense object localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19596--19605.

[33] L. Zhu, X. Wang, J. Feng, T. Cheng, Y. Li, B. Jiang, D. Zhang, J. Han, WeakCLIP: Adapting CLIP for Weakly-Supervised Semantic Segmentation, International Journal of Computer Vision, (2024) 1-21.

[34] X.H. J. Xie, K. Ye, and L. Shen, Clims: Cross language image matching for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 483–4492.

[35] J.W.K. A. Radford, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning(PMLR), 2021, pp. 8748–8763.

[36] E.M. A. Kirillov, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al, Segment anything, in: arXiv preprint arXiv:2304.02643, 2023, pp. 1–30.

[37] Z.L. W. Sun, Y. Zhang, Y. Zhong, and N. Barnes, An alternative to wsss? an empirical study of the segment anything model (sam) on weakly-supervised semantic segmentation problems, in: arXiv preprint arXiv:2305.01586, 2023.

[38] P.-T.J.a.Y. Yang, Segment anything is a good pseudo-label generator for weakly supervised semantic segmentation, in: arXiv preprint arXiv:2305.01275, 2023.

[39] Z.M. T. Chen, R. Li, and W.-l. Chao, Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation, in: arXiv preprint arXiv:2305.05803, 2023.

[40] H.a.Y. Kweon, Kuk-Jin, From SAM to CAMs: Exploring Segment Anything Model for Weakly Supervised Semantic Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 19499--19509.

[41] Z.a.F. Yang, Kexue and Duan, Minghong and Qu, Linhao and Wang, Shuo and Song, Zhijian, Separate and conquer: Decoupling co-occurrence via decomposition and representation for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3606--3615.

[42] Y.Z. L. Ru, 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. 16846–16855.

[43] W.O. L. Xu, M. Bennamoun, F. Boussaid, and D. Xu, Multiclass token transformer for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4310–4319.

[44] D.Z. S. Rossetti, M. Sanzari, M. Schaerf, and F. Pirri, Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation, in: Proceedings of the European conference on computer vision, 2022, pp. 446–463.

[45] H.Z. L. Ru, 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.

[46] Z.M. R. Li, Z. Zhang, J. Jang, and S. Sanner, Transcam: Transformer attentionbased cam refinement for weakly supervised semantic segmentation, in: Elsevier Journal of Visual Communication and Image Representation, 2023, pp. 103800.

[47] S.-H.a.K. Yoon, Hoyong and Kim, Hyeonseong and Yoon, Kuk-Jin, Class Tokens Infusion for Weakly Supervised Semantic Segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3595--3605.

[48] Y.a.Y. Wu, Xichen and Yang, Kequan and Li, Jide and Li, Xiaoqiang, 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.

[49] H.T. Mathilde Caron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the European conference on computer vision, 2021.

[50] P.D. T.-Y. Lin, R. Girshick, K. He, B. Hariharan, and S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[51] T.C.a.L. Mo, Swin-fusion: swin-transformer with feature fusion for human action recognition, in: Springer Neural Processing Letters, 2023, pp. 11109–11130.

[52] L.V.G. M. Everingham, C. K. Williams, J. Winn, and A. Zisserman, Andrew, The pascal visual object classes (voc) challenge, in: Springer International journal of computer vision, 2010, pp. 303–338.

[53] P.A. B. Hariharan, L. Bourdev, S. Maji, and J. Malik, Semantic contours from inverse detectors, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2011, pp. 991–998.

[54] W.D. J. Deng, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, Imagenet: A largescale hierarchical image database, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[55] E.K. J. Lee, 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.