MoRe: Class Patch Attention Needs Regularization for Weakly Supervised Semantic Segmentation

Zhiwei Yang[1,2,3], **Yucong Meng**[2,3], **Kexue Fu**[4], **Shuo Wang**[2,3*], **Zhijian Song**[1,2,3*]

[1]Academy for Engineering and Technology, Fudan University, Shanghai 200433, China
[2]Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, Shanghai 200032, China
[3]Shanghai Key Laboratory of Medical Imaging Computing and Computer Assisted Intervention, Shanghai 200032, China
[4]Shandong Computer Science Center (National Supercomputer Center in Jinan)
{zwyang21, ycmeng21}@m.fudan.edu.cn, {fukexue, shuowang, zjsong}@fudan.edu.cn

Abstract

Weakly Supervised Semantic Segmentation (WSSS) with image-level labels typically uses Class Activation Maps (CAM) to achieve dense predictions. Recently, Vision Transformer (ViT) has provided an alternative to generate localization maps from class-patch attention. However, due to insufficient constraints on modeling such attention, we observe that the Localization Attention Maps (LAM) often struggle with the artifact issue, i.e., patch regions with minimal semantic relevance are falsely activated by class tokens. In this work, we propose MoRe to address this issue and further explore the potential of LAM. Our findings suggest that imposing additional regularization on class-patch attention is necessary. To this end, we first view the attention as a novel directed graph and propose the Graph Category Representation module to implicitly regularize the interaction among class-patch entities. It ensures that class tokens dynamically condense the related patch information and suppress unrelated artifacts at a graph level. Second, motivated by the observation that CAM from classification weights maintains smooth localization of objects, we devise the Localization-informed Regularization module to explicitly regularize the class-patch attention. It directly mines the token relations from CAM and further supervises the consistency between class and patch tokens in a learnable manner. Extensive experiments are conducted on PASCAL VOC and MS COCO, validating that MoRe effectively addresses the artifact issue and achieves state-of-the-art performance, surpassing recent single-stage and even multi-stage methods. Code is available at *https://github.com/zwyang6/MoRe*.

Introduction

Weakly supervised semantic segmentation (WSSS) aims to achieve pixel-wise predictions with cheap annotations, such as points (Bearman et al. 2016), scribbles (Lin et al. 2016; Vernaza and Chandraker 2017), bounding boxes (Dai, He, and Sun 2015; Lee et al. 2021b), and image-level labels (Pinheiro and Collobert 2015; Yang et al. 2024a). It significantly reduces the annotation cost of fully supervised segmentation methods (Tang et al. 2023; Xu et al. 2024b) and has attracted increasing attention in recent years. Among all annotation forms, the image-level label is the most accessible
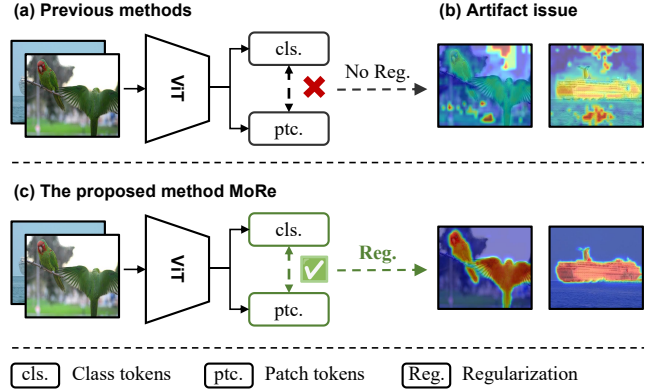
---

[*]Corresponding authors.

Figure 1: Our motivation. Localization Attention Maps (LAM) from ViT provide an alternative to CAM. (a) Since no regularization is conducted between class-patch attention, (b) LAM particularly suffers from the artifact issue. (c) We propose MoRe to tackle it and generate better LAM by regularizing attention among class-patch tokens.

while challenging, as it contains the least semantic information to localize objects. In this work, we focus on the WSSS paradigm with only image-level labels.

Typically, the pipeline of WSSS can be divided into three steps. It first trains a classification network to generate Class Activation Maps (CAM) (Zhou et al. 2016) with image-level labels. Then CAM is refined as pseudo labels, which are further leveraged to provide dense supervision to retrain a segmentation model (Ahn and Kwak 2018; Yang et al. 2024b). However, due to the limited supervision, CAM intends to activate only the most discriminative parts of objects, which dramatically impairs the performance of WSSS. Recently, Vision Transformer (ViT) (Dosovitskiy et al. 2020), famous for building long-range dependency, has been widely adopted in WSSS. Benefiting from the self-attention mechanism, several studies (Caron et al. 2021; He et al. 2020) have shown that attention maps between class and patch tokens can reliably highlight objects, offering a promising alternative for generating precise localization maps. Inspired by it, TS-CAM (Gao et al. 2021) directly extracts the Localization Attention Maps (LAM) from class-patch attention and leverages it to generate pseudo labels. However, since the

original ViT only contains a single-class token, TSCAM still struggles to perform class-specific dense predictions. To fill in this gap, MCTformer series (Xu et al. 2022, 2024a) integrates multi-class tokens into ViT and successfully generates class-specific LAM, which demonstrates the great potential of LAM in enhancing the performance of WSSS.

However, as demonstrated in (Sun et al. 2024; Darcet et al. 2023), in order to capture long-range dependencies, ViT inclines to aggregate global semantics in low-information patches. It causes unrelated patches to be frequently correlated with class tokens during attention, leading to numerous false activations of artifact patch tokens. Previous LAM-based WSSS methods barely regularize such attention. Consequently, this issue commonly exists in ViTs (such as ViT on ImageNet (Deng et al. 2009) or DeiT (Touvron et al. 2021)) and severally impairs the quality of LAM, as shown in Figure 1 (a, b). Importantly, unlike the notorious over-smoothness issue (Wang et al. 2022) of ViT that stems from the high similarity among patch tokens, the reported artifact issue is derived from the improper correlation between class-patch tokens. Although recent WSSS works have proposed to solve the over-smoothness (Ru et al. 2023; He et al. 2023), both LAM-based and CAM-based methods particularly overlook the artifact issue from improper class-patch attention, leaving this problem unaddressed.

In this work, we find that more regularization is necessary for constructing proper class-patch attention and propose MoRe to tackle the artifact issue. First, we propose the Graph Category Representation (GCR) module to implicitly regularize the relation among class-patch tokens. GCR views the class-patch attention as a novel directed graph representation (Wu et al. 2020). To this end, each token in ViT is regarded as a node and is further represented by the head and tail embeddings. To precisely model the correlation among class-patch entities, class-related neighbors are dynamically updated and additional edge embeddings are designed to parameterize the relation between heads and tails. Building upon this dynamic graph structure, the Graph Aggregation mechanism is designed to fully consider the knowledge from head, tail, and edge, allowing more reliable patches to aggregate class-related semantics into class tokens.

In addition, since CAM and LAM exhibit strong alignment in object localization, we design the Localization-informed Regularization (LIR) module to explicitly regularize class-patch attention. Although previous works propose to fuse LAM with CAM (Xu et al. 2022, 2024a), the fusion strategy simply combines them with a multiplying operation while no optimization is conducted during attention. Instead, LIR mines token relations from CAM and further leverages the supervision to facilitate the consistency between class-patch tokens in a learnable manner. Specifically, guided by the clues from CAM, a confident relation enhancement loss is designed to promote the correlation between class and confidently related patch tokens while suppressing unrelated artifacts. With the enhanced class tokens, we also formulate an uncertain relation enhancement loss to supervise more uncertain yet relevant regions being attended by class tokens. Finally, based on the GCR and LIR modules, MoRe is seamlessly incorporated into ViT-based WSSS and effec-tively tackles the artifact end-to-end, producing high-fidelity LAM and pseudo labels, as shown in Figure 1 (c).

The main contributions of our work are listed as follows:

- This work reports and addresses the artifact issue when generating LAM from class-patch attention. We find that more regularization is needed to constrain the improper attention and propose MoRe to achieve it.

- Two forms of regularization are designed. We design the Graph Category Representation (GCR) module to implicitly regularize the class-patch attention, which enhances the interaction of class-related information from a graph perspective. The Localization-informed Regularization (LIR) module is proposed to explicitly constrain the relation among class-patch tokens, which facilitates the class-patch consistency in a learnable manner.

- Extensive experiments are conducted on PASCAL VOC and MS COCO, demonstrating the efficacy of MoRe and the superior performance over recent single-stage methods and even sophisticated multi-stage techniques.
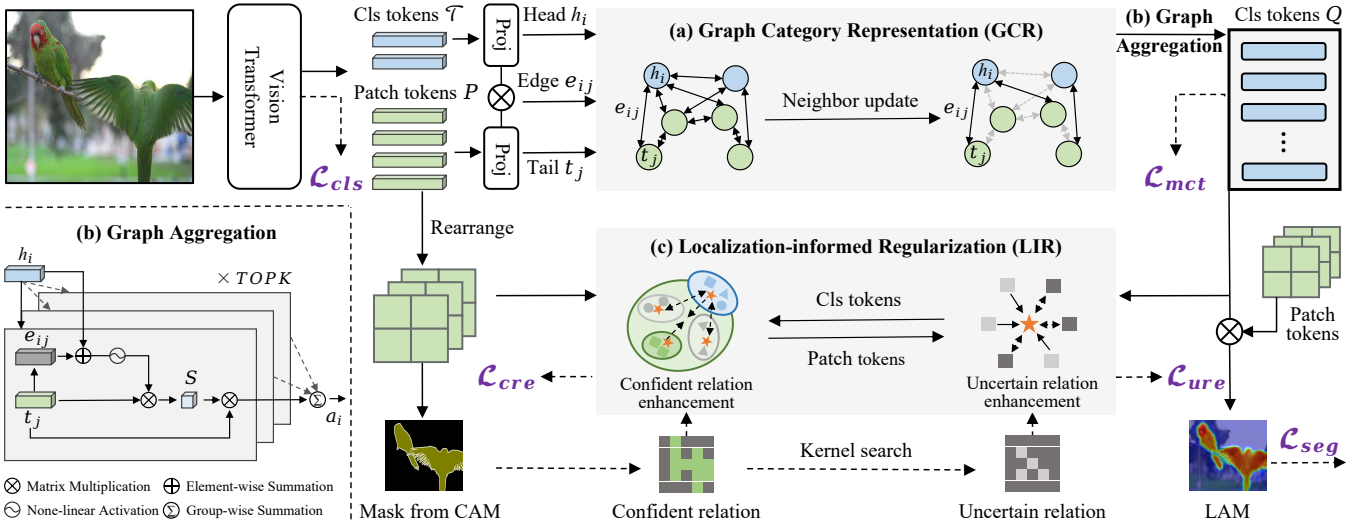
Related Work

Weakly Supervised Semantic Segmentation

Weakly supervised semantic segmentation (WSSS) with image-level labels particularly leverages Class Activation Maps (CAM) to achieve dense predictions. Due to the classification nature of CAM, CAM only activates the most discriminative parts of objects. To generate more complete CAM, considerable efforts have been paid with intriguing insights, such as prototype maintaining (Du et al. 2022; Tang et al. 2024), affinity propagation (Ru et al. 2022), and multi-modal supervision (Lin et al. 2022), etc. With the emergence of ViT, ViT-based WSSS has substantially mitigated this issue. AFA (Ru et al. 2022) learns affinity from attention and uses it to refine CAM. TSCAM (Gao et al. 2021) notices that class-patch attention maps in ViT can localize objects and generate more complete Localization Attention Maps (LAM). Inspired by it, MCTformer series (Xu et al. 2022, 2024a) integrates multi-class tokens and further generates class-specific LAM, which serves as a substitute for CAM. Despite these advancements, previous LAM-based methods overlook the artifact issue. In this work, we further explore the potential of LAM and intend to address this issue by regularizing the attention between class and patch tokens.

Graph Neural Networks

Graph neural networks (GNN) aim to model the representation with graph elements (e.g., nodes and edges) and approximation inference (Wu et al. 2020). With its ability to capture complex interactions within the entity topology, GNN has attracted widespread attention across various fields, such as 3D pose estimation (Mehraban et al. 2024) and medical image analysis (Zhou et al. 2019). For WSSS, most GNN-related works are based on convolutional neural networks (CNN). GraphNet (Pu et al. 2018) and A2GNN (Zhang et al. 2021) are proposed to construct graph structures based on superpixel or affinity under the supervision of bounding

Figure 2: Overview of our MoRe. The input image is sent to ViT encoder and generates multi-class and patch tokens. (a) We first send them into Graph Category Representation (GCR) module, which takes the class-patch attention as a directed graph with the projected entities head h_i, tail t_j, and learnable edge e_{ij}. (b) Then Graph Aggregation mechanism is designed to condense the related tail semantics into class tokens. (c) CAM is also generated from patches. It acts as the confident and uncertain relation supervision to the proposed Localization-informed Regularization (LIR) module with two objectives $\mathcal{L}_{cre}$ and $\mathcal{L}_{ure}$. Finally, LAM is generated from the similarity score map between class-patch tokens and is used to train a segmentation decoder.

boxes. GSM (Li et al. 2021) employs GNN to capture semantic dependencies among images within the same class. Although impressive, they typically rely on pre-defined relations to construct the CNN-based graph and ignore the directed interaction among entities. In this work, we build our graph on ViT following (Luo, Thost, and Shi 2024; Li et al. 2024), while uniquely highlighting a directed topology between class-patch tokens to mine their mutual information.

Methodology

Framework Overview

The pipeline of MoRe is illustrated in Figure 2. The input image space and classification label are defined as X and $\mathcal{Y} = \{1, 2, \ldots, C\}$, where C denotes the number of categories. Given a training batch (I, l) from image and label space, $I \in \mathbb{R}^{3 \times \mathcal{H} \times \mathcal{W}}$ is the training image and $l \in \mathcal{Y}$ is the class label. We inherit the multi-class token settings from (Xu et al. 2022) and send I into ViT, which generates the patch tokens $P \in \mathbb{R}^{L \times D}$ and multi-class tokens $\mathcal{T} \in \mathbb{R}^{C \times D}$. L, D are the number of patches and the embedding dimension. Building upon this, we first send $\mathcal{T}$ and P to the proposed Graph Category Representation (GCR) module, which views the class attention as a novel directed graph, shown in Figure 2 (a). Then Graph Aggregation mechanism is designed to fully condense the related patch semantics into class tokens, as shown in Figure 2 (b). With the aggregated class tokens $Q \in \mathbb{R}^{C \times D}$, the Localization-informed regularization (LIR) module is designed to further regularize the attention between class and patch tokens, as shown in Figure 2 (c). Finally, we generate LAM by calculating the cosine similarity score between class-patch tokens. LAM is leveraged as pseudo labels to train a decoder with segmentation loss L_{seg}. We also maintain a classification loss L_{cls} to generate

CAM, which provides localization relations for LIR module. More details are introduced in the subsequent sections.

Graph Category Representation

Category Graph Construction To suppress the artifacts from improper interaction between class-patch tokens, we view the attention as a directed graph structure and implicitly regularize the relation with learnable topological features. As shown in Figure 2 (a), given the multi-class tokens $\mathcal{T}$ and patches P from ViT, we adopt two linear projectors to transform them into heads $H \in \mathbb{R}^{C \times D}$ and tails $T \in \mathbb{R}^{L \times D}$, where heads model the correlation to patches while tails represent the contribution from patches to heads. Since artifacts commonly exist in low-information patches, we assume that those tokens should be dynamically excluded during attention. To this end, we customize the edge embeddings e_{ij} and select candidate neighbors to model the relation between head h_i and tail t_j. We first select the top-K cosine similarity scores to quantify the similarity between h_i and its most related patches, which are formulated as:

$$r_i = softmax(TOPK\{h_i^T t_j\}_{j=1}^L). \tag{1}$$

To exclude the unrelated artifact tokens, we propose updating neighbors with the tail candidates ranked by the similarity score r_i. The index of most related top-K tail neighbors to head h_i can be denoted as follows:

$$N_i = \{j \mid j \in \text{argmax}_{TOPK}\{h_i^T t_j\}_{j=1}^L\}. \tag{2}$$

Based on the updated neighbors, the edge embeddings maintained between head h_i and tail t_j are formulated as:

$$e_{ij} = r_{ij} t_j + (1 - r_{ij}) h_i, j \in N_i. \tag{3}$$

With the generated entities, we delineate the attention between class and patch tokens as a dynamic directed graph

$G = \{V, R, Z, E\}$, where V is the nodes corresponding to class and patch tokens, Z represents the head and tail. E denotes the edge. $R = \{(h, e, t) : (h, t) \in Z, e \in E\}$ generalizes head, tail, and edge, representing the directed information on the directed edge. Compared to the vanilla class attention, the topological graph structure effectively reasons the relation among class and patch tokens by considering the triplets of head, tail, and the learnable directed edge.

Graph Aggregation Mechanism Building upon the constructed directed graph G, we further propose the Graph Aggregation mechanism to aggregate reliable knowledge from patches into class tokens, as shown in Figure 2 (b). To model the different importance of knowledge propagated from tail to head, a weighting factor $S(h_i, e_{ij}, t_j)$ is designed to guarantee the propagation process, which is determined by the triplets of head, tail, and edge as:

$$S(h_i, e_{ij}, t_j) = \text{softmax}(t_j^T \sigma(h_i + e_{ij})), \qquad (4)$$

where $\sigma(\cdot)$ is the gated activation function $tanh(\cdot)$.

Then the knowledge from related neighbors to h_i is denoted as: $a_i = \sum_{j \in N_i} S(h_i, e_{ij}, t_j) t_j$. By representing relationships among triplets as edge-based knowledge, head nodes can effectively receive and capture signals from tail nodes. Finally, we generate the graph-regularized class tokens Q by fusing the knowledge with the initial heads. We introduce a bi-directional aggregating strategy to facilitate the information propagation:

$$Q = \delta_1 \left(w_1 \left(h_i + a_i \right) \right) + \delta_2 \left(w_2 \left(a_i \odot h_i \right) \right), \qquad (5)$$

where δ_1/δ_2 adopt LeakyReLU, w_1/w_2 are project matrix, and $\odot$ is the element-wise multiplication.

Localization-informed Regularization

Confident Relation Enhancement Since CAM holds a high alignment of localization objects with LAM, we further leverage such prior to generate confident relation M_c and uncertain relation M_u. They explicitly regularize the class-patch attention in a learnable manner, as shown in Figure 2 (c). Formally, we leverage $\mathcal{L}_{cls}$ to maintain a classification head to generate CAM $\in \mathbb{R}^{n_h \times n_w \times C}$ by projecting the classification matrix on rearranged P, where $n_h \times n_w$ is the feature size. We adopt a multi-threshold filtering strategy to refine CAM into reliable masks, which contain foreground, background, and uncertain regions:

$$M_a = \begin{cases} \text{argmax}(CAM_{i,j,:}), & \text{if max } (CAM_{i,j,:}) > \lambda_h, \\ 0, & \text{if max } (CAM_{i,j,:}) < \lambda_l, \\ 255, & \text{otherwise}, \end{cases}$$
$$(6)$$

where $M_a \in \mathbb{R}^{n_h \times n_w}$, thresholds $0 < \lambda_l < \lambda_h < 1$, 0 and 255 denote the index of background and uncertain regions.

To promote the correlation between class and related patch tokens while suppressing unrelated artifacts, we extract the reliable foreground regions from M_a and take it as the confident relation map $M_c \in \mathbb{R}^{n_h \times n_w}$ to guide the class-patch attention. Specifically, for patch token p_{ij} in P, we calculate the cosine similarity with class tokens Q. Then we leverage the class index on M_c as the relation supervision. If the class token $q_l \in Q$ shows the same class index

l with the pixel (i, j) on M_c, we view the correlation between class token q_l and patch token p_{ij} as positive while those with different classes are negative pairs. The contrast between class-patch tokens is achieved by:

$$\mathcal{L}_{cre} = -\frac{1}{N_c^+} \sum_{q_l \in Q_l} \sum_{p_{ij}^+ \in P_l^+} \log \frac{\exp(q_l^T p_{ij}^+/\tau)}{\sum_{p_{ij} \in P} \exp \left(q_l^T p_{ij}/\tau \right)},$$
$$(7)$$

where N_c^+ denotes the number of positive pairs between class tokens and patch tokens. Q_l is the class token set with class label l, P_l^+ is the patch token set that shows the same class index l with class token q_l. τ is the temperature factor to control the sharpness of contrast.

It is noted that the contrast benefits from two folds: 1) it explicitly improves the correlation between class and related patch tokens, which helps reliably activate patches and suppress the artifacts. 2) it directly promotes discrimination of class tokens, which further reduces the overlapping attention regions derived from ambiguity among class tokens.

Uncertain Relation Enhancement Due to the incompleteness nature of CAM, CAM-informed LAM still struggles with the issue that only most deterministic patches are activated by class tokens. To alleviate this problem, we further extract uncertain relation map $M_u \in \mathbb{R}^{n_h \times n_w}$ from M_a and leverage it to supervise more uncertain yet relevant regions being activated by class tokens. Since M_u is inevitably noisy, we design an uncertain mining kernel to search comparatively reliable patches. Specifically, given a $d \times d$ kernel centered at position (i, j), we send it to walk through both M_u and M_c. Only the uncertain patch token p_{ij} is selected when the number of both foreground and uncertain pixels in the kernel exceeds a certain proportion φ. The search process for uncertain yet relevant patch tokens is formulated as:

$$U = \{u_{ij} \in P_u : \text{kernel} \left(ij, M_c, M_u\right) / (d \times d) > \varphi\},$$
$$(8)$$

where P_u is the uncertain patch candidates whose class index on M_a is 255. $kernel(\cdot)$ is the kernel searching operator, and φ is the proportion value to select patch tokens.

With the enhanced class tokens Q and the selected uncertain patch tokens U, we make the class tokens act as class centroids and pull uncertain patch tokens correlated to them. Formally, given the input image with class label l, we view the class token q_l and the selected uncertain patch token u_{ij} as positive pairs and maximize the similarity between them. For class token q_{-l} whose class index is not l, we view the relation with u_{ij} as negative pairs and minimize the similarity. The process can be expressed as:

$$\mathcal{L}_{ure} = \frac{1}{N_u^+} \sum (1 - \text{cosim}(q_l^T u_{ij})) + \frac{1}{N_u^-} \sum \text{cosim}(q_{-l}^T u_{ij}),$$
$$(9)$$

where $q_l \in Q_l, q_{-l} \in Q_{-l}, u_{ij} \in U$, N_u^+/N_u^- denotes the number of positive/negative pairs, respectively. $cosim(\cdot)$ denotes the computation of cosine similarity.

Training Objectives

As shown in Figure 2, MoRe consists of two class-patch attention regularization losses, i.e., $\mathcal{L}_{cre}$ and $\mathcal{L}_{ure}$, and a

Method	Sup.	Net.	Seed	Mask
Multi-stage WSSS methods.				
RIB (Lee et al. 2021a)	$\mathcal{I}+\mathcal{S}$	RN101	56.5	70.6
L2G (Jiang et al. 2022)	$\mathcal{I}+\mathcal{S}$	RN101	-	71.9
†CLIMS (Xie et al. 2022)	$\mathcal{I}+\mathcal{L}$	RN38	56.6	70.5
†CLIP-ES (Lin et al. 2022)	$\mathcal{I}+\mathcal{L}$	RN101	70.8	75.0
†CPAL (Tang et al. 2024)	$\mathcal{I}+\mathcal{L}$	RN101	71.9	75.8
†PSDPM (Zhao et al. 2024b)	$\mathcal{I}+\mathcal{L}$	RN101	-	77.3
CDA (Su et al. 2021)	$\mathcal{I}$	RN101	58.4	66.4
ReCAM (Chen et al. 2022)	$\mathcal{I}$	RN101	54.8	70.0
†MCTformer (Xu et al. 2022)	$\mathcal{I}$	RN38	61.7	69.1
†LPCAM (Chen and Sun 2023)	$\mathcal{I}$	RN50	65.3	72.7
†MCTformer+ (Xu et al. 2024a)	$\mathcal{I}$	RN101	68.8	76.2
SFC (Zhao et al. 2024a)	$\mathcal{I}$	RN101	64.7	73.7
†CTI (Yoon et al. 2024)	$\mathcal{I}$	RN101	69.5	73.7
Single-stage WSSS methods.				
1Stage (Ara. and Roth 2020)	$\mathcal{I}$	RN38	-	66.9
†AFA (Ru et al. 2022)	$\mathcal{I}$	MiT-B1	65.0	68.7
†ViT-PCM (Rossetti et al. 2022)	$\mathcal{I}$	ViT-B	67.7	71.4
†ToCo (Ru et al. 2023)	$\mathcal{I}$	ViT-B	71.6	72.2
†DuPL (Wu et al. 2024)	$\mathcal{I}$	ViT-B	-	75.1
†**MoRe-CAM(Ours)**	$\mathcal{I}$	ViT-B	**76.9**	**79.7**
†**MoRe-LAM(Ours)**	$\mathcal{I}$	ViT-B	**77.0**	**80.0**

Table 1: Performance comparison of pseudo labels on PASCAL VOC train set. Sup. denotes the supervision type. $\mathcal{I}$: image-level labels. $\mathcal{S}$: saliency maps. $\mathcal{L}$: Language. †: methods based on Vision Transformer.

Method	Net.	VOC		COCO
		Val	Test	Val
Multi-stage WSSS methods.				
RIB (Lee et al. 2021a)	RN101	70.2	70.0	43.8
L2G (Jiang et al. 2022)	RN101	72.1	71.7	44.2
RCA (Zhou et al. 2022)	RN38	72.2	72.8	36.8
CDA (Su et al. 2021)	RN38	66.1	66.8	33.2
ESOL (Li et al. 2022)	RN101	69.9	69.3	42.6
†MCTformer (Xu et al. 2022)	RN38	71.9	71.6	42.0
†CLIP-ES (Lin et al. 2022)	RN101	72.2	72.8	45.4
†OCR (Cheng et al. 2023)	RN38	72.7	72.0	42.5
†BECO (Rong et al. 2023)	RN101	73.7	73.5	45.1
†MCTformer+ (Xu et al. 2024a)	RN38	74.0	73.6	45.2
†CTI (Yoon et al. 2024)	RN101	74.1	73.2	45.4
†CPAL (Tang et al. 2024)	RN101	74.5	74.7	46.3
Single-stage WSSS methods.				
1Stage (Ara. and Roth 2020)	RN38	62.7	64.3	-
SLRNet (Ru et al. 2022)	RN38	67.2	67.6	35.0
†AFA (Ru et al. 2022)	MiT-B1	66.0	66.3	38.9
†ViT-PCM (Rossetti et al. 2022)	ViT-B	70.3	70.9	-
†ToCo (Ru et al. 2023)	ViT-B	71.1	72.2	42.3
†DuPL (Wu et al. 2024)	ViT-B	73.3	72.8	44.6
†**MoRe(Ours)**	**ViT-B**	**76.4**	**75.0**	**47.4**

Table 2: Semantic segmentation results on PASCAL VOC and MS COCO in terms of mIoU(%). Net. denotes the backbone for single-stage methods or segmentation network for the multi-stage. †: methods based on Vision Transformer.

multi-label soft margin classification loss $\mathcal{L}_{cls}$. We incorporate $\mathcal{L}_{mct}$ to further promote the discrepancy of class tokens following (Xu et al. 2024a). Denoting the weight factors of α and β, the optimizing objective of MoRe is formulated as:

$$L_{MoRe} = L_{cls} + L_{mct} + \alpha L_{cre} + \beta L_{ure}. \quad (10)$$

In addition, MoRe trains segmentation decoder end-to-end. The loss L_{seg} for segmentation uses cross-entropy. Therefore, the overall loss for MoRe is: $L = L_{MoRe} + \gamma L_{seg}$. Following the prevalent settings in single-stage WSSS (Ru et al. 2023; Wu et al. 2024), regularization losses to improve CAM and segmentation masks are also adopted.

Experiments

Experimental Settings

Datasets and Evaluation Metrics The proposed method is evaluated on PASCAL VOC 2012 (Everingham et al. 2010) and MS COCO 2014 (Lin et al. 2014). PASCAL VOC contains 21 classes. Following (Ru et al. 2022, 2023), the augmented data with $10,582$ images are used for training, $1,449$ for validating, and $1,456$ for testing. MS COCO includes 81 classes. $82,081$ and $40,137$ images are used for training and validation. Mean Intersection-over-Union (mIoU) is the main evaluation metric. Following (Yang et al. 2024a), the confusion ratio is also adopted to validate the efficacy of suppressing false positives from artifact issues.

Implementation Details MoRe adopts ViT-B/16 pretrained on ImageNet (Ridnik et al. 2021) as encoder. The decoder in this work uses DeepLab-LargeFOV (Chen et al. 2017). Following the training settings (Wu et al. 2024; Xu

et al. 2024a; Hu et al. 2024), the AdamW optimizer with an initial learning rate $6e-5$ is used. The training images are augmented with random resized cropping into 448×448, random horizontal flipping, and color jittering. We maintain neighbors with $K = 392$ tail nodes. For kernel searching, we set uncertain pixels in M_u as 1 and foregrounds in M_c as 2. When the sum of both masks in kernel exceeds $\varphi = 1.2$ times the kernel size $d \times d$, the uncertain patch token is selected. The loss weight factors (α, β, γ) for our training objectives are $(0.2, 0.1, 0.12)$. All experiments are conducted on RTX 3090. More details can be found in *Appendix*.

Comparison with State-of-the-arts

Evaluation of Pseudo Masks The quantitative comparisons of localization seeds and pseudo masks are reported in Table 1. Without post-processing, MoRe generates seeds with 77.0% mIoU for LAM and 76.9% for CAM. Both seeds surpass most multi-stage methods with sophisticated refinements. With DenseCRF (Krähenbühl and Koltun 2011), our pseudo masks from LAM and CAM are further improved to 80.0% and 79.7%, respectively. With regularization on class-patch attention, MoRe shows the superiority of 3.8% over the LAM-based counterpart (Xu et al. 2024a) and 4.9% over single-stage SOTA (Wu et al. 2024). The qualitative comparisons are showcased in Figure 3, which shows the competence of MoRe to generate better LAM as well.

Performance of Segmentation Table 2 reports the segmentation comparisons of MoRe with other SOTA methods on VOC and COCO. MoRe achieves 76.4% mIoU on VOC val set, which noticeably outperforms both single-stage and
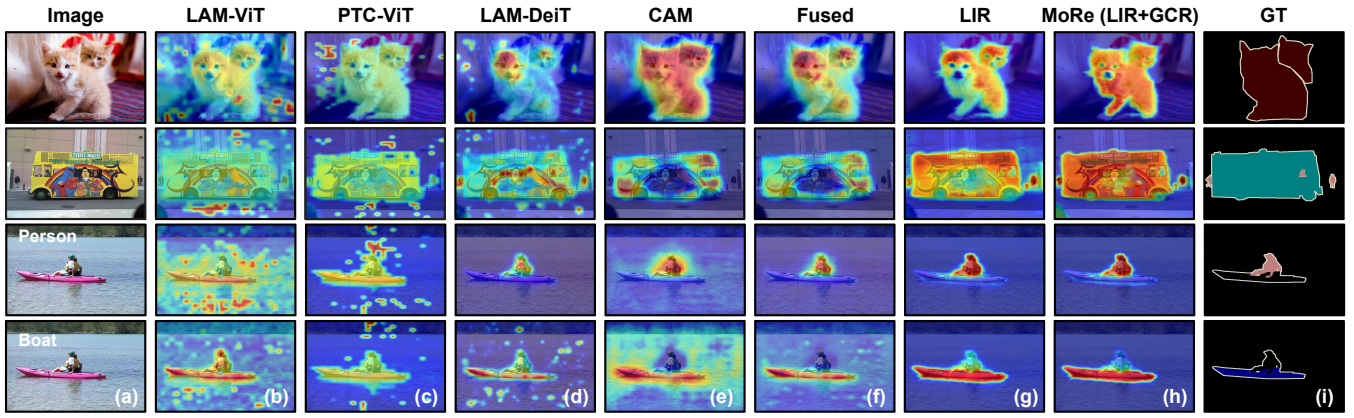
Figure 3: Visualization of LAM. (a) image. (b) LAM on ViT pretrained on ImageNet. (c) LAM with PTC loss (Ru et al. 2023) for tackling over-smoothness of ViT. (d) LAM on DeiT. (e) Patch CAM from MCTformer+ (Xu et al. 2024a). (f) Refined LAM with fusion strategy by multiplying both LAM and CAM. (g) LAM with our designed LIR module. (h) LAM with both our designed LIR and GCR modules. (i) Ground truth. More visualized results are showcased in *Appendix*.
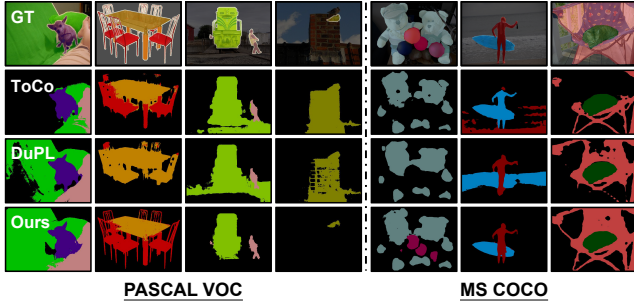

Figure 4: Segmentation visualization with SOTA single-stage methods (i.e., ToCo and DuPL) on VOC and COCO.

Conditions	$\mathcal{L}_{cre}$	$\mathcal{L}_{ure}$	GCR	Precision	Recall	mIoU
Baseline (ViT-B)				76.2	46.5	41.6
w/o $\mathcal{L}_{cre}$		✓	✓	75.3	80.1	65.0
w/o $\mathcal{L}_{ure}$	✓		✓	84.3	85.7	74.4
w/o GCR	✓	✓		81.6	85.5	72.1
MoRe	✓	✓	✓	**84.5**	**88.0**	**76.4**

Table 3: Ablation study of MoRe on VOC val set.

	ToCo	DuPL	MoRe(Ours)	
Aeroplane	80.6 (0.19)	77.9 (0.26)	$85.2_{+4.6}$	**(0.13)** -7.0%
Bird	68.4 (0.42)	81.7 (0.20)	$91.7_{+10.0}$	**(0.05)** -15.0%
Boat	45.4 (1.11)	58.7 (0.53)	$73.0_{+14.3}$	**(0.23)** -30.0%
TV monitor	63.1 (0.44)	45.5 (1.10)	$64.3_{+1.2}$	**(0.38)** -6.0%
Car	83.3 (0.11)	77.5 (0.20)	$83.7_{+0.4}$	**(0.10)** -1.0%
Sofa	43.8 (0.77)	53.9 (0.59)	$55.1_{+1.2}$	**(0.44)** -15.0%
Potted plant	56.5 (0.59)	60.7 (0.54)	$67.7_{+7.0}$	**(0.20)** -34.0%
Average	71.1 (0.32)	73.3 (0.31)	$76.4_{+3.1}$	**(0.22)** -9.0%

Table 4: Class-specific performance of IoU and confusion ratio (in bracket) with recent methods on VOC val set.

even sophisticated multi-stage methods by at least 3.1% and 1.9%. It achieves 75.0% on VOC test and 47.4% on COCO val set, which also gains improvement over other methods.

Figure 4 visualizes segmentation results on VOC and COCO. It shows that MoRe clearly differentiates categories and generates complete predictions with precise boundaries. For example, MoRe successfully segments chairs from a table and even captures the outline of chair arms, while the recent methods confuse the objects (case in column 2).

Ablation Studies and Analysis

Efficacy of Key Components Quantitative ablation results are detailed in Table 3. We leverage ViT/B pretrained on ImageNet as the baseline, which also inherits multi-class tokens supervised by objectives from (Xu et al. 2022) end-to-end. As reported, it cannot achieve satisfying predictions. The proposed $\mathcal{L}_{cre}$ in LIR module explicitly promotes the correlation between class and related patch tokens. $\mathcal{L}_{ure}$ reliably searches uncertain yet relevant patch tokens and ensures them being activated. Without supervision from $\mathcal{L}_{cre}$, the precision and mIoU heavily drop by 9.2% and 11.4%, which explains that $\mathcal{L}_{cre}$ not only suppresses artifacts, but also reduces the over-lapping attention region among class tokens by improving the discrimination. Without supervision from $\mathcal{L}_{ure}$, the recall result also drops from 88.0% to 85.7%. The proposed GCR module dynamically updates re-

lated neighbors with learnable edge information and controls the information propagated to class tokens. Without the graph structure GCR, the segmentation performance heavily decreases by 4.3%. The above results confirm that the proposed components effectively reduce artifacts and contribute to generating high-quality semantic predictions.

Effectiveness of Suppressing Artifacts Qualitative ablation results are further illustrated in Figure 3 to investigate the impact of proposed modules. As shown in Figure 3 (b), the ViT baseline typically suffers from the artifact issue and cannot achieve reasonable localization. The proposed LIR module directly maximizes the similarity between class and related patch tokens while minimizing the irrelevant. As seen from Figure 3 (g), LIR significantly suppresses the artifacts and activates more non-deterministic regions, which verifies the efficacy of $\mathcal{L}_{cre}$ and $\mathcal{L}_{ure}$, respectively. The proposed GCR module works by dynamically constructing

Multi-stage.	Train Time	GPU	Val	Test
CLIMS (Xie et al. 2022)	1068 mins	18.0 G	70.4	70.0
MCTformer+ (Xu et al. 2024a)	1496 mins	18.0 G	74.0	73.6
Single-stage.				
AFA (Ru et al. 2022)	554 mins	19.0 G	66.0	66.3
ToCo (Ru et al. 2023)	506 mins	17.9 G	71.1	72.2
DuPL (Wu et al. 2024)	508 mins	14.9 G	73.3	72.8
MoRe(Ours)	**372 mins**	**12.1 G**	**76.4**	**75.0**

Table 5: Efficiency of MoRe compared to others. The experiment is conducted on PASCAL VOC with RTX 3090.

related neighbors and precisely aggregating useful semantics into class tokens from a graph perspective. Figure 3 (h) demonstrates that MoRe models robust representation of categories and activates objects with higher confidence. It validates that GCR further improves the quality of LAM.

In addition, Table 4 reports the class-specific segmentation performance of MoRe compared to recent methods and introduces Confusion Ratio (FP/TP) to investigate the ability to tackle artifact issues. MoRe shows a significantly lower confusion ratio over other methods, such as potted plant $(-34.0\%$ to DuPL), boat(-30.0%), and higher IoU on all selected categories. For the average performance on VOC val set, MoRe achieves 0.22 confusion ratio, lower by at least -9.0% to recent SOTAs, which validates the competence of MoRe to reduce false positives from artifact issues.

Difference with Over-smoothness Over-smoothness of ViT is notorious for incurring false positives in WSSS. However, the reported artifact issue in our work shows different properties. To investigate it, the PTC loss (Ru et al. 2023) is incorporated in our ViT baseline, which is effective in tackling over-smoothness. As shown in Figure 3 (c), although LAM is regularized by PTC, artifacts still commonly exist and severally impair the quality. The reason lies in that over-smoothness arises from the high similarity among patch tokens while the artifact is derived from the improper relation between class-patch tokens. Previous methods barely pay efforts to regularize the class-patch attention, which impedes the development of LAM. By noticing this, MoRe regularizes such attention with GCR and LIR, which significantly improves the quality of LAM, as shown in Figure 3 (g,h).

Analysis of Training Efficiency Instead of designing sophisticated architectures, the proposed regularization modules can be seamlessly incorporated into ViT and effectively regularize the artifact. The training efficiency comparison is reported in Table 5. The recent LAM-based method (Xu et al. 2024a) cannot be implemented end-to-end, thus 1496 minutes and 18.0 GB GPU memory are taken to finish the WSSS workflow. In contrast, MoRe works in a single-stage manner and only needs 372 minutes and 12.1 GB GPU memory to finish the workflow, which significantly outperforms both multi-stage and single-stage counterparts.

Analysis of Hyper-parameters The analysis of important hyper-parameters in GCR and LIR, such as *top-K, uncertain selection threshold, searching kernel size, etc.*, is specifically discussed in *Appendix*.
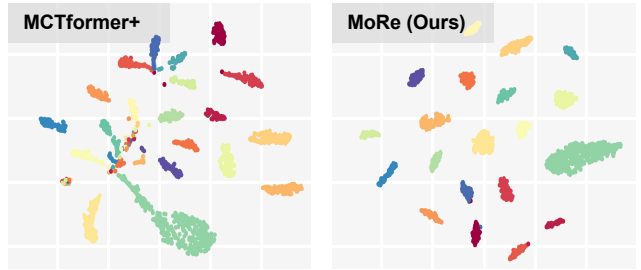

Figure 5: Multi-class token representation between MCTformer+ and MoRe on VOC train set, visualized with t-SNE.

Analysis of Category Representation Previous LAM-based methods perform in-distinctive category representation, as shown in Figure 3 (b,c,d). To further investigate the role of the proposed regularization modules, we extract the multi-class tokens from the last layer of ViT encoder and visualize the representation with t-SNE (Van der Maaten and Hinton 2008) on VOC train set, as shown in Figure 5. Although the LAM-based SOTA (Xu et al. 2024a) intends to separate class tokens by minimizing their similarity, it is found that regularizing class tokens alone is insufficient to model separated feature space. In contrast, MoRe enhances the representation with a powerful topological graph and promotes the correlation between class tokens and related patch semantics, which demonstrates the distinguished class representation. Considering LAM is directly generated from the class-patch attention, the comparison supports the better pseudo mask performance in Table 4, Figure 3, and verifies that MoRe builds a more robust class representation.

In addition, when constrained to class tokens alone, the competitor struggles with the incompleteness and artifact issues as well, as shown in Figure 3 (d). Although it intends to refine LAM with CAM, the fusion strategy with a simple multiplying operation fails to improve the representation of categories and is still incompetent to cover the objects adequately, as shown in Figure 3 (e,f). In contrast, MoRe mines class-patch token relations from CAM and leverages it to supervise the attention, which further enhances the representation of categories and benefits the quality of both LAM and CAM, as shown in Figure 3 (g,h) and Table 1, respectively.

Conclusion

This work identifies and addresses the artifact issue when generating LAM from class-patch attention. We find that more regularization is needed to constrain the improper attention and propose MoRe to regularize it. Specifically, we propose the Graph Category Representation (GCR) module to model the class-patch attention as a directed graph, which implicitly excludes irrelevant artifact patches and aggregates useful information into class tokens. Then the Localization-informed Regularization (LIR) is proposed to explicitly promote the correlation of class and related patch tokens, which further helps suppress the artifacts and guarantees uncertain regions being activated. Extensive experiments and analysis are conducted on PASCAL VOC and MS COCO datasets, validating the efficiency of MoRe in tackling the artifact issue and significantly improving the performance of WSSS.

Acknowledgments

References

Ahn, J.; and Kwak, S. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *CVPR*, 4981–4990.

Ara., N.; and Roth, S. 2020. Single-stage semantic segmentation from image labels. In *CVPR*, 4253–4262.

Bearman, A.; Russakovsky, O.; Ferrari, V.; and Fei-Fei, L. 2016. What's the point: Semantic segmentation with point supervision. In *ECCV*, 549–565. Springer.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *ICCV*, 9650–9660.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848.

Chen, Z.; and Sun, Q. 2023. Extracting Class Activation Maps From Non-Discriminative Features As Well. In *CVPR (CVPR)*, 3135–3144.

Chen, Z.; Wang, T.; Wu, X.; Hua, X.-S.; Zhang, H.; and Sun, Q. 2022. Class re-activation maps for weakly-supervised semantic segmentation. In *CVPR*, 969–978.

Cheng, Z.; Qiao, P.; Li, K.; Li, S.; Wei, P.; Ji, X.; Yuan, L.; Liu, C.; and Chen, J. 2023. Out-of-candidate rectification for weakly supervised semantic segmentation. In *CVPR*, 23673–23684.

Dai, J.; He, K.; and Sun, J. 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 1635–1643.

Darcet, T.; Oquab, M.; Mairal, J.; and Bojanowski, P. 2023. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, Y.; Fu, Z.; Liu, Q.; and Wang, Y. 2022. Weakly Supervised Semantic Segmentation by Pixel-to-Prototype Contrast. In *CVPR (CVPR)*, 4320–4329.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Gao, W.; Wan, F.; Pan, X.; Peng, Z.; Tian, Q.; Han, Z.; Zhou, B.; and Ye, Q. 2021. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *ICCV*, 2886–2895.

He, J.; Cheng*, L.; Fang, C.; Zhang, D.; Wang, Z.; and Chen, W. 2023. Mitigating Undisciplined Over-Smoothing in Transformer for Weakly Supervised Semantic Segmentation. *arXiv preprint arXiv:2305.03112*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9729–9738.

Hu, M.; Yuan, K.; Shen, Y.; Tang, F.; Xu, X.; Zhou, L.; Li, W.; Chen, Y.; Xu, Z.; Peng, Z.; et al. 2024. OphCLIP: Hierarchical Retrieval-Augmented Learning for Ophthalmic Surgical Video-Language Pretraining. *arXiv preprint arXiv:2411.15421*.

Jiang, P.-T.; Yang, Y.; Hou, Q.; and Wei, Y. 2022. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In *CVPR*, 16886–16896.

Krähenbühl, P.; and Koltun, V. 2011. Efficient inference in fully connected crfs with gaussian edge potentials. *NeurIPS*, 24.

Lee, J.; Choi, J.; Mok, J.; and Yoon, S. 2021a. Reducing information bottleneck for weakly supervised semantic segmentation. *NeurIPS*, 34: 27408–27421.

Lee, J.; Yi, J.; Shin, C.; and Yoon, S. 2021b. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2643–2652.

Li, J.; Chen, Y.; Chu, H.; Sun, Q.; Guan, T.; Han, A.; and He, Y. 2024. Dynamic Graph Representation with Knowledge-aware Attention for Histopathology Whole Slide Image Analysis. In *CVPR*, 11323–11332.

Li, J.; Jie, Z.; Wang, X.; Wei, X.; and Ma, L. 2022. Expansion and shrinkage of localization for weakly-supervised semantic segmentation. *NeurIPS, 35: 16037–16051*, 35: 16037–16051.

Li, X.; Zhou, T.; Li, J.; Zhou, Y.; and Zhang, Z. 2021. Groupwise semantic mining for weakly supervised semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1984–1992.

Lin, D.; Dai, J.; Jia, J.; He, K.; and Sun, J. 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 3159–3167.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755. Springer.

Lin, Y.; Chen, M.; Wang, W.; Wu, B.; Li, K.; Lin, B.; Liu, H.; and He, X. 2022. CLIP is Also an Efficient Segmenter:

A Text-Driven Approach for Weakly Supervised Semantic Segmentation. *arXiv preprint arXiv:2212.09506*.

Luo, Y.; Thost, V.; and Shi, L. 2024. Transformers over directed acyclic graphs. *NeurIPS*, 36.

Mehraban; Soroush; Adeli, V.; and Taati, B. 2024. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In *WACV*, 6920–6930.

Pinheiro, P. O.; and Collobert, R. 2015. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 1713–1721.

Pu, M.; Huang, Y.; Guan, Q.; and Zou, Q. 2018. GraphNet: Learning image pseudo annotations for weakly-supervised semantic segmentation. In *Proceedings of the 26th ACM international conference on Multimedia*, 483–491.

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

Rong, S.; Tu, B.; Wang, Z.; and Li, J. 2023. Boundary-Enhanced Co-Training for Weakly Supervised Semantic Segmentation. In *CVPR*, 19574–19584.

Rossetti, S.; Zappia, D.; Sanzari, M.; Schaerf, M.; and Pirri, F. 2022. Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. In *ECCV*, 446–463. Springer.

Ru, L.; Zhan, Y.; Yu, B.; and Du, B. 2022. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *CVPR*, 16846–16855.

Ru, L.; Zheng, H.; Zhan, Y.; and Du, B. 2023. Token contrast for weakly-supervised semantic segmentation. In *CVPR*, 3093–3102.

Su, Y.; Sun, R.; Lin, G.; and Wu, Q. 2021. Context decoupling augmentation for weakly supervised semantic segmentation. In *ICCV*, 7004–7014.

Sun, M.; Chen, X.; Kolter, J. Z.; and Liu, Z. 2024. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*.

Tang, F.; Xu, Z.; Huang, Q.; Wang, J.; Hou, X.; Su, J.; and Liu, J. 2023. DuAT: Dual-aggregation transformer network for medical image segmentation. In *PRCV*.

Tang, F.; Xu, Z.; Qu, Z.; Feng, W.; Jiang, X.; and Ge, Z. 2024. Hunting Attributes: Context Prototype-Aware Learning for Weakly Supervised Semantic Segmentation. In *CVPR*, 3324–3334.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vernaza, P.; and Chandraker, M. 2017. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *CVPR*, 7158–7166.

Wang, P.; Zheng, W.; Chen, T.; and Wang, Z. 2022. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962*.

Wu, Y.; Ye, X.; Yang, K.; Li, J.; and Li, X. 2024. DuPL: Dual Student with Trustworthy Progressive Learning for Robust Weakly Supervised Semantic Segmentation. In *CVPR (CVPR)*, 3534–3543.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Philip, S. Y. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1): 4–24.

Xie, J.; Hou, X.; Ye, K.; and Shen, L. 2022. Clims: Cross language image matching for weakly supervised semantic segmentation. In *CVPR*, 4483–4492.

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

Xu, L.; Ouyang, W.; Bennamoun, M.; Boussaid, F.; and Xu, D. 2022. Multi-class token transformer for weakly supervised semantic segmentation. In *CVPR*, 4310–4319.

Xu, Z.; Tang, F.; Chen, Z.; Zhou, Z.; Wu, W.; Yang, Y.; Liang, Y.; Jiang, J.; Cai, X.; and Su, J. 2024b. Polyp-Mamba: Polyp Segmentation with Visual Mamba. In *MICCAI*. Springer.

Yang, Z.; Fu, K.; Duan, M.; Qu, L.; Wang, S.; and Song, Z. 2024a. Separate and Conquer: Decoupling Co-occurrence via Decomposition and Representation for Weakly Supervised Semantic Segmentation. In *CVPR (CVPR)*, 3606–3615.

Yang, Z.; Meng, Y.; Fu, K.; Wang, S.; and Song, Z. 2024b. Tackling Ambiguity from Perspective of Uncertainty Inference and Affinity Diversification for Weakly Supervised Semantic Segmentation. *arXiv preprint arXiv:2404.08195*.

Yoon, S.-H.; Kwon, H.; Kim, H.; and Yoon, K.-J. 2024. Class Tokens Infusion for Weakly Supervised Semantic Segmentation. In *CVPR*, 3595–3605.

Zhang, B.; Xiao, J.; Jiao, J.; Wei, Y.; and Zhao, Y. 2021. Affinity attention graph neural network for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 8082–8096.

Zhao, X.; Tang, F.; Wang, X.; and Xiao, J. 2024a. Sfc: Shared feature calibration in weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7525–7533.

Zhao, X.; Yang, Z.; Dai, T.; Zhang, B.; and Xiao, J. 2024b. PSDPM: Prototype-based Secondary Discriminative Pixels Mining for Weakly Supervised Semantic Segmentation. In *CVPR (CVPR)*, 3437–3446.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*, 2921–2929.

Zhou, T.; Zhang, M.; Zhao, F.; and Li, J. 2022. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *CVPR*, 4299–4309.

Zhou, Y.; Graham, S.; Alemi Koohbanani, N.; Shaban, M.; Heng, P.-A.; and Rajpoot, N. 2019. Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images. In *ICCV workshops*, 0–0.