All-pairs Consistency Learning for Weakly Supervised Semantic Segmentation

Weixuan Sun[1,2,3], Yanhao Zhang[5], Zhen Qin[2], Zheyuan Liu[1], , Lin Cheng[4], Fanyi Wang[5], Yiran Zhong[2,3], Nick Barnes[1]
[1]Australian National University, [2]OpenNLPLab, [3]Shanghai AI Lab,
[4]Xiamen University, [5]OPPO Research Institute.

Abstract

*In this work, we propose a new transformer-based regularization to better localize objects for Weakly supervised semantic segmentation (WSSS). In image-level WSSS, Class Activation Map (CAM) is adopted to generate object localization as pseudo segmentation labels. To address the partial activation issue of the CAMs, consistency regularization is employed to maintain activation intensity invariance across various image augmentations. However, such methods ignore pair-wise relations among regions within each CAM, which capture context and should also be invariant across image views. To this end, we propose a new all-pairs consistency regularization (*ACR*). Given a pair of augmented views, our approach regularizes the activation intensities between a pair of augmented views, while also ensuring that the affinity across regions within each view remains consistent. We adopt vision transformers as the self-attention mechanism naturally embeds pair-wise affinity. This enables us to simply regularize the distance between the attention matrices of augmented image pairs. Additionally, we introduce a novel class-wise localization method that leverages the gradients of the class token. Our method can be seamlessly integrated into existing WSSS methods using transformers without modifying the architectures. We evaluate our method on PASCAL VOC and MS COCO datasets. Our method produces noticeably better class localization maps (67.3% mIoU on PASCAL VOC train), resulting in superior WSSS performances.*

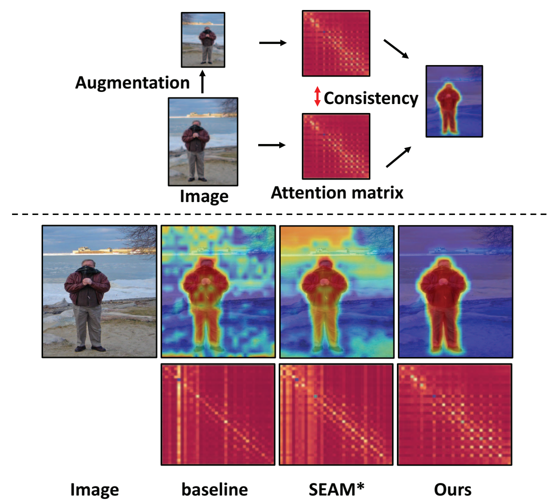

Figure 1. **Top:** conceptual illustration of the proposed ACR. Given two views of the same image by *e.g.*, resizing & flipping, we regularize the consistency between the corresponding positions of the two self-attention matrices, employing two types of invariant consistencies, *i.e. Region affinity consistency* and *Region activation consistency*. **Bottom:** object localization and their corresponding attention matrices, all results are obtained based on vision transformer with only class labels. Baseline: the model is trained with only classification loss. Other WSSS models, e.g., SEAM [60], only perform learning with activation consistency. SEAM*, a transformer variant of SEAM. Ours: trained with our ACR shows the benefit of including affinity consistency. Our approach can effectively localize targeted objects.

1. Introduction

Weakly supervised semantic segmentation (WSSS) aims to relieve the laborious and expensive process of pixel-wise labeling with different types of weak labels including image-level labels [23, 2, 19, 66, 52], points [4], scribbles [57, 36, 54] and bounding boxes [13, 40, 31, 39, 51]. Image-level WSSS is particularly challenging as it uses only class labels to supervise pixel-wise predictions without any location prior. An essential step of image-level WSSS is to obtain class-wise localization maps, i.e., seeds, which provide object localization to generate pseudo segmentation labels. Previous WSSS methods[59, 60, 7, 27, 2, 46, 48, 1] generally rely on Class Activation Maps (CAMs) [74] based on the Convolutional Neural Networks. Although significant research has been undertaken to improve CAMs, it still suffers from incomplete and inaccurate activation. These issues are caused by the supervision gap between the image tags and pixel-wise segmentation supervision since the classification network is indifferent to pixel-wise activation and only requires a sufficient average pooled value.

Existing work [60, 72, 15] uses augmentation invariant consistency to refine CAMs, where they consider *region activation consistency* which forces the absolute class activation values to be consistent between augmented views. Although such regularization has been demonstrated to be effective such as [60, 72, 15], activation consistency can only discover activation in novel views but non-activated regions and background noise cannot be solved through contextual relations. Thus, we propose to also maintain pair-wise consistency across the views, termed *region affinity consistency*. Specifically, we look at the relations between regions within each image and compare these relations across views. In Fig. 1, this implies that the relation intensities, *e.g*. between the person and sky, should stay invariant to augmentations between two views. Our motivation is that affinity is a manner of context encoding and context has been demonstrated to be essential for pixel-wise predictions [72, 68, 60]. Thus, every region in an image is encouraged to have the same relationships with all other regions as the augmented view, rather than simply the same value (such as SEAM [60]). So both targeted and non-targeted objects are reinforced by affinity consistency. Samples in Fig. 1 validate our motivations. The attention matrices of baseline and SEAM are distracted by specific tokens (bright columns) which are not desired since an region is either targeted or non-targeted, while our method captures better object shapes (Diagonal grid patterns show that the targeted and non-targeted image regions are clearly distinguished).

Our method, named All-pairs Consistency Regularization (ACR), uses a vision transformer to simultaneously enforce region activation consistency and region affinity consistency. Transformer-based models have achieved great success in various tasks [14, 63, 42, 75, 41, 61, 58, 38]. As the core of the transformer, we find that the self-attention matrices can be naturally used to regularize our two consistencies without requiring additional affinity computation. Specifically, given an image that is split into $h \times w = n$ patch tokens, an attention matrix $A \in \mathbb{R}^{(n+1)\times(n+1)}$ is generated in the self-attention module. Its first row encodes relations between the class token and the patch tokens, such a class-to-patch attention can be reshaped to an $h \times w$ map showing potential object activation [5, 66, 9, 50] for the *region activation consistency*. Additionally, the patch-to-patch attention $A\texttt{[1:,1:]} \in \mathbb{R}^{n\times n}$ encodes pair-wise relations among all pairs of patch tokens that can be used for the *region affinity consistency*. During classification training, we input the image and its augmented view into a Siamese vision transformer to obtain attention matrices for the two views respectively. For the augmented attention matrix, we introduce a novel approach to restore the original spatial order inverting the transformation. Therefore we can directly regularize the corresponding positions of the attention matrices across two views to enforce the two consistencies.

The attention-based consistency is class agnostic, therefore, we cannot directly obtain a class-wise localization for the downstream WSSS task. Further, simply transplanting the CNN-based CAM [74] to transformers relies on the output features (*i.e*. patch tokens), but extensive noise is observed [66, 44, 53]. To this end, we propose a new class localization generation method for vision transformers with a single class token. Thanks to our consistency regularization during training, the attention matrices encode rich class-wise object information. We find that the class-wise gradients of the class-to-patch attention $\bigtriangledown A\texttt{[0,1:]} \in \mathbb{R}^{n}$ already provide decent class-wise object localization. We additionally leverage the patch-to-patch attention $A\texttt{[1:,1:]} \in \mathbb{R}^{n\times n}$ to refine our class-wise localization maps and generate segmentation seeds. We note that our training regularization and seed generation method can be seamlessly integrated into the vision transformer networks.

To summarize, our main contributions are:

- We propose All-pairs Consistency Regularization (ACR) for wsss. It ensures affinity consistency as well as activation consistency during the classification training, which leads to better initial seeds for wsss.
- We propose to leverage the self-attention structure of the vision transformer to regularize the two types of consistencies, which can be directly used on vision transformers without modifications. To enforce the regularization, we propose a technique to re-align the spatial orders of the two views' self-attention matrices that inverts the effects of a broad range of image transformations.
- We propose to use the gradients to generate accurate class-wise localization maps from a single class token, and further refine it with the learned region affinity.

The proposed method generates significantly improved class-wise localization maps compared to all previous WSSS methods and leads to state-of-the-art performance on PASCAL VOC and MS COCO.

2. Related Work

Various WSSS methods are proposed to avoid laborious pixel-wise annotation. The adopted weak labels include image-level labels [2, 40, 60, 7, 70, 69, 32, 72, 49], scribbles [36], points [4], and bounding boxes [13, 31, 39]. We mainly focus on image-level methods in this review. Existing image-level WSSS methods generally rely on CAMs [74] as initial seeds to generate pseudo segmentation labels. Various solutions are proposed to refine the CAMs.

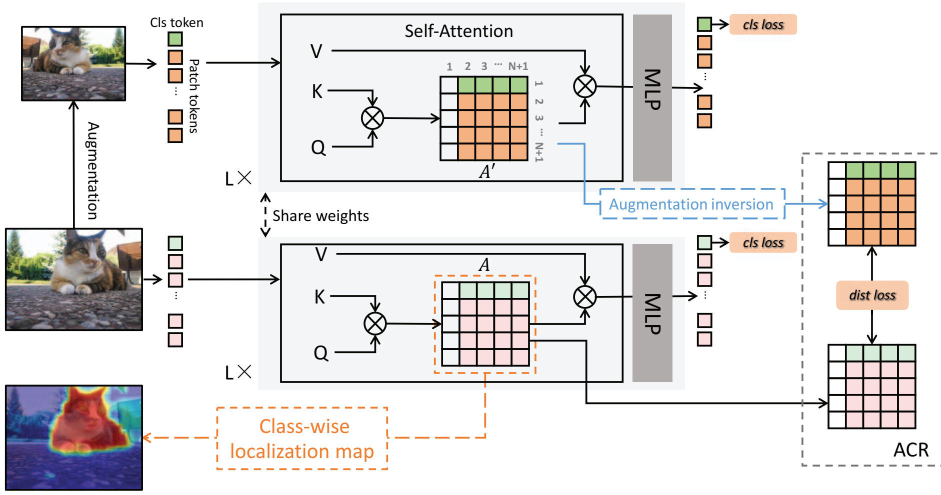

Figure 2. An overview of ACR. An image is augmented to a novel view then the augmented pair is input into a Siamese vision transformer (two branches share weights), consisting of L successive transformer blocks. The class token (green) is used to make classification predictions. In each self-attention matrix, class-to-patch attention (green) encodes region activation and patch-to-patch attention (pink) encodes region affinity. We propose regularizing the distance between two views' self-attention matrices to enforce ACR. Our class localization map is generated using the self-attention matrix as shown in the bottom dashed orange box detailed in Fig. 3.

Consistency Regularization. Different types of consistencies are proposed to refine the initial seeds for WSSS. [72] studies CAMs consistency from complementary patches of the same image. [73] explores the consistency between two parallel classifiers which tries to increase the distinction between the CAMs and merge the two-branch outputs to obtain complete CAMs. Further, foreground-background contrastive [11, 64] and intra-class contrastive [48] are proposed to refine the localization accuracy. [47, 17] introduce feature consistency across paired images from the same class to mine more regions. Finally, [15] proposes a prototype-based metric learning methodology, that enforces feature-level consistencies in both inter-view and intra-view regularizations. A similar method to ACR is SEAM proposed in [60]. However, it only enforces CAM invariant consistency across augmentations but does not consider affinity consistency, i.e., the CAM values should be the same across different augmented image views.

Learning Affinity Refinement. Pair-wise affinity is often adopted in WSSS to refine the initial seeds. [65] uses an auxiliary saliency detection task to learn the affinity. [60, 72] adopts the low-level feature maps from a CNN network to generate affinity that preserves detailed context information. [2, 1] propose to learn a network to discriminate paired pixels from the reliable seeds of CAMs. Then they use the learned network to guide random walk propagation to refine CAMs. In the transformer era, affinity is inherently encoded in the self-attention module. [44] adopts reliable seeds of CAMs to directly supervise the affinity of the self-attention to capture object shapes. [66] adopts multiple class tokens to generate class-wise localization maps and also uses the affinity from the self-attention to refine the maps.

In summary, existing WSSS methods disregard the consistency of the affinity across views, i.e., In this work, we explore leveraging the self-attention mechanism to enforce such consistency.

3. Method

In this section, we first outline the key design choices for the proposed regularization, then present our ACR training framework. Fig. 2 outlines our framework. Our two forms of regularization are applied to a vision transformer [14] without modifying the network structure. In Section 3.3, we detail our approach in obtaining the class localization maps from the network gradients.

3.1. Overview

We base our design on the vision transformer [14, 50, 61, 58, 38], as existing work [66, 44, 53] has demonstrated that better activation is obtained. Compared to CNNs [60, 72, 6, 47, 11, 52], transformers explicitly encode region dependencies among all tokens with self-attention layers. Such characteristics naturally suit our need for modeling the two forms of consistencies without introducing extra modules. Specifically, we use class-to-patch attention to achieve region activation consistency, which sets our method apart from existing CNN-based work [60, 72]. Moreover, although previous work [60, 72] involves extra dedicated modules that model the affinity within each image, they do not use such a concept for regularizing the consistency across multiple views. We instead directly leverage the patch-to-patch attention to achieve region affinity consistency.

3.2. Attention Consistency Regularization

Here, we present the design of ACR, with notation following [14]. We split the input image into $n = h \times w$, (height by width) non-overlapping patches and flatten them to a sequence of n tokens. A class token is inserted to form the input sequence $T \in \mathbb{R}^{(n+1)\times d}$ where d is the embedding dimension. The class token attends to all patch tokens and is used for classification prediction. Within each transformer block, we obtain attention matrix $A \in \mathbb{R}^{(n+1)\times(n+1)}$ by $\text{softmax}(QK^T/\sqrt{d})$ [56], where $Q, K \in \mathbb{R}^{(n+1)\times d}$ are the query and key matrices projected from T.

During classification training, we augment the input image I directly to a novel view I' by a randomly selected transformation. Then we input the two views into a Siamese vision transformer to obtain two attention matrices A and A'. As discussed, self-attention encodes region activation and region affinity simultaneously, we calculate the distance between the two matrices to enforce our attention consistency regularization. To handle matrices that are not spatially equivalent after augmentations, we propose a method that rearranges the order of the tokens accordingly. We introduce the two proposed regularization terms, as well as the token-rearranging method in detail below.

Region Activation Consistency encourages the network to generate object localization that is invariant to transformations. Consider the first row of the attention matrix A, we can extract the class-to-patch attention $A\texttt{[0,1:]} \in \mathbb{R}^{1\times n}$. As discussed in [20, 8, 53, 66, 5], $A\texttt{[0,1:]}$ can be reshaped and normalized to a class-agnostic objectness map $M \in \mathbb{R}^{h\times w}$ as the class token is used for classification. Thus, given the attention matrix A and its augmented view's attention matrix A', we regularize the activation across two views by comparing the class-to-patch attention:

$$L_{\text{act}} = \|A\texttt{[0,1:]} - f^{-1}(A'\texttt{[0,1:]})\|_1, \tag{1}$$

where f^{-1} is an inverse transformation to restore the spatial ordering of the tokens after the image has undergone an augmentation such as flip. So $f^{-1}A'$ and A have the same spatial ordering of tokens, but different values, since the image transformation also alters pixel ordering within each of the patches themselves, leading to altered features. In other words, we do not invert the embeddings of the tokens but only their relative positions. The inversion ensures that we can regularize the corresponding positions of the two attention matrices. The class token attends to all patches, so n image patch tokens correspond to $A\texttt{[0,1:]} \in \mathbb{R}^{1\times n}$. In training, we calculate the ℓ_1 loss between corresponding areas of the two attention matrices to enforce region activation regularization.

Region Affinity Consistency encourages pair-wise relations between image regions to be invariant to transformations. Given attention matrix A and its augmented view attention A', and considering that $A\texttt{[1:,1:]} \in \mathbb{R}^{n\times n}$, encodes the affinity between all patch tokens, the affinity consistency regularization is formulated as:

$$L_{\text{aff}} = \|A\texttt{[1:,1:]} - f^{-1}(A'\texttt{[1:,1:]})\|_1. \tag{2}$$

During training, we measure ℓ_1 loss between the corresponding pair-wise patch tokens of the two attention matrices to enforce region affinity regularization.

Transformation Inverse and Optimization Objective. Image augmentation changes the appearance and the relative positions of the patch tokens . Thus, the attention matrices from two views may not be spatially equivalent, which prohibits direct distance calculation. To address this, we introduce a transformation to invert the image augmentation of the attention matrix in terms of token ordering. Note that we only consider token ordering in this section and omit the transformation that is applied inside each image patch as we only aim to restore the original spatial information, not the embedding. This operation is shown in the dashed blue box in Fig. 2 and denoted as f^{-1} in equation 1, 2,

we present the details of the inversion in this section.

In practice, we use spatial transformations including resize, flip, and rotation. Resize does not affect token ordering so we can simply resize the attention matrix back to the original size. Image flip and rotation can be performed and inverted by general matrix operations. Given a patched input image $X \in \mathbb{R}^{h\times w}$ without considering the transformation inside each patch, flip is a permutation operation and rotation can be considered as a transpose followed by a flip. So the augmented image can be formulated:

$$\text{flip:} \quad X' = P_h X P_w, \tag{3}$$
$$\text{rotation:} \quad X' = P_h X^T P_w, \tag{4}$$

where $P_h \in \mathbb{R}^{h\times h}$ and $P_w \in \mathbb{R}^{w\times w}$ are permutation matrices in x and y directions respectively. Let A' be the attention matrix of X', then the inversion of A' can be written as:

$$f^{-1}(A') \quad = C^T(P_w \otimes P_h^T)A'(P_w \otimes P_h^T)^T C, \tag{5}$$

where $\otimes$ is Kronecker product. $C \in \mathbb{R}^{n\times n}$ is a commutation matrix for rotation and an identity matrix when flipping. Here, we omit the class token for simplicity. Note that such a formulation enables inversion of a wide range of possible image transformations that can be described by

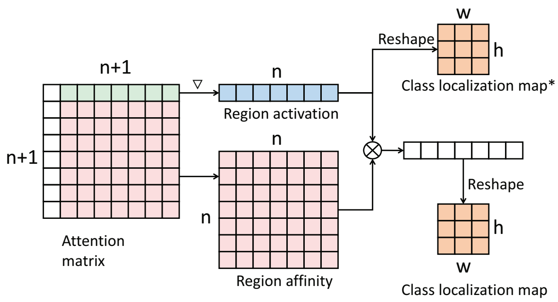

Figure 3. An overview of our class-wise localization map generation framework. We use the gradients of the class-to-patch attention (the blue vector) to generate a class localization map*. Further, we use the learned region affinity (the pink block) to refine the class localization map. Sample visualizations of the learned region affinity are shown in Fig. 4.

permutation matrices, though many may not be helpful augmentations. Please refer to the supplementary material for detailed derivations and discussions.

In summary, A and $f^{-1}(A')$ have the same token ordering according to equation 5. Hence, we can directly calculate the distance between the two attention matrices to apply ACR. Our optimization objective is the combination of the two-view classification and the consistency losses:

$$L = L_{\text{cls}} + \alpha L_{\text{act}} + \beta L_{\text{aff}}. \tag{6}$$

where α, β are hyperparameters.

3.3. Gradient-based Transformer Class Localization Map

At test time, the object activation provided by the class-to-patch attention $A\texttt{[0,1:]} \in \mathbb{R}^{1\times n}$ is class-agnostic [5, 66]. To obtain class-wise localizations for the downstream WSSS task, a naive solution is to directly transplant the CAM [74] method into the transformer, by using the average pooled patch tokens instead of the class token to produce classification predictions. However, in line with existing works [53, 44, 66], we find that this achieves poor results (Table 5). Another approach [66] uses multiple class-specific tokens to generate class-wise seeds. However, this requires modifying the transformer architecture and computational complexity grows with the number of classes. Inspired by recent transformer interpretability work [8, 9], we introduce a gradient-based approach. Different from [8, 9] which incorporate gradients with the attention values or the network relevances [3], we empirically find that the gradients can directly provide accurate localization information and construct our gradient-based transformer class localization methods.

In Fig. 3, given the class-to-patch attention matrix $A\texttt{[0,1:]} \in \mathbb{R}^{1\times n}$ (the green vector) and target class c, we calculate gradients by back-propagating the classification score y^c, formulated as $\bigtriangledown A^c\texttt{[0,1:]} := \partial y^c/(\partial A\texttt{[0,1:]})$ (the blue vector). Intuitively, $\bigtriangledown A^c\texttt{[0,1:]}$, *i.e.* class-wise gradients of the class-to-patch attention, represent patch tokens' contributions to the final classification scores. Then we remove negative values and reshape it to $h \times w$ to obtain the class localization map* shown in Fig. 3. We empirically find that averaging the multi-layer gradients performs well. Given a transformer with l successive layers, the localization map for class c is defined as:

$$M^c = \frac{1}{l}\sum_i^l \bigtriangledown A_i^c\texttt{[0,1:]}. \tag{7}$$

Affinity Refinement. Inspired by [60, 66], we further adopt the learned patch-wise affinity $A\texttt{[1:,1:]} \in \mathbb{R}^{n\times n}$ (the pink matrix) to refine the activation maps as shown in Fig 3. Thanks to our Region affinity consistency, better context is encoded in self-attention modules. We visualize the patch-wise affinity in Fig. 4. Note that the baseline model is trained with only classification loss without our regularization, the generated affinity (the third column) is distracted by specific patch tokens, leading to noisy seeds (the second column). Our affinity (the fifth column) can capture better object contexts and generate integral localization. Formally, our class localization map for class c is defined as:

$$M^c = \left(\frac{1}{l}\sum_i^l \bigtriangledown A_i^c\texttt{[0,1:]}\right) \times \left(\frac{1}{l}\sum_i^l A_i\texttt{[1:,1:]}\right). \tag{8}$$

Then, $M^c \in \mathbb{R}^{1\times n}$ can be reshaped to $h\times w$ and normalized to obtain our final class localization maps, i.e., initial seeds.

Our class-wise localization maps provide accurate and dense object coverage. The reasons are two-fold. First, the region activation consistency encourages the class token to attend to accurate object localization as shown in row 4 of Fig. 5. Second, affinity consistency regularization encourages the network to capture precise pair-wise affinity, such affinity propagates the localized pixels to cover comprehensive object regions, as shown in row 5 of Fig. 5.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate our method on the PASCAL VOC [16] and MS COCO [37] datasets. The official PASCAL VOC has 20 objects classes and one background class, with 1,446 training, 1,449 validation, and 1,456 testing images. Following common practice in WSSS, we use an augmented training set consisting of 10,582 images with annotations from [21]. MS COCO 2014 is much more challenging than PASCAL VOC. It contains 81 classes including background with 80k training and 40k validation images.

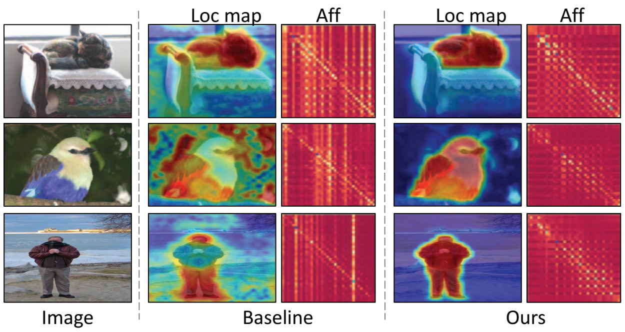

Figure 4. Class localization maps (Loc map) and pair-wise affinity of patch tokens (Aff). Our method can capture better context encoding and generate accurate localization maps. Baseline: the model is trained with only classification loss. The attention matrices are down-sampled for readability.

Table 1. Performance comparison of WSSS methods on MS COCO. w/ saliency: the method adopts extra saliency information. Best number is in bold.

Methods	Venue	w/ saliency	Val
AuxSegNet [65]	ICCV2021	✓	33.9
EPS [18]	CVPR2022	✓	35.7
L2G [24]	CVPR2022	✓	44.2
Wang et al. [59]	IJCV2020		27.7
Ru et al. [44]	CVPR2022		38.9
SEAM [60]	CVPR2020		31.9
CONTA [71]	NeurIPS2020		32.8
CDA [46]	ICCV2021		33.2
Ru et al. [43]	IJCV2022		36.2
URN [34]	AAAI2022		41.5
MCTformer [66]	CVPR2022		42.0
ESOL [33]	NeurIPS2022		42.6
SIPE [11]	CVPR2022		43.6
RIB [28]	NeurIPS2020		43.8
ACR			**45.0**

Implementation Details. We adopt ViT-hybrid-B [14]. Training images are resized and cropped to 384×384. For semantic segmentation, following previous WSSS methods [66, 2, 1, 27], we use DeepLabV2 [10] with a ResNet101 [22] backbone as the segmentation model. During segmentation inference, we use multi-scale testing and adopt CRFs [26] for post-processing. Detailed implementation details are presented in the supplementary material.

4.2. Comparison with State-of-the-art

4.2.1 MS COCO

Table 1 shows segmentation results on MS COCO. We achieve a segmentation mIoU of 45%, which surpasses existing methods with a clear margin. Notably, this result does not rely on any extra saliency information but outperforms all previous WSSS methods including the ones with saliency. MS COCO is a bigger dataset with more semantic classes and complex images that include multiple objects. This result indicates that saliency may hinder WSSS approaches' ability to scale to complex scenes, hence we do not incorporate saliency into our approach. Our result demonstrates that ACR is able to generate reliable class localization maps in challenging scenes. We report per-class results of MS COCO in the supplementary material.

Table 2. Performances of the initial Seeds and pseudo segmentation labels on PASCAL VOC *train* set. *(s)*: methods that rely on saliency to generate seeds. ACR*: our localization maps without affinity refinement. Our seeds outperform previous non-salient methods by a significant margin.

Methods	Seed	w/ saliency	Pseudo
EDAM (CVPR2021) [62]	52.8	✓	68.1
ReCAM (CVPR2022) [12]	54.8	✓	70.9
L2G (CVPR2022) [24]	56.2	✓	71.9
EPS (ECCV2020) [18]	69.4 *(s)*	✓	71.6
Du et al.(CVPR2022) [15]	**70.5** *(s)*	✓	**73.3**
PSA (CVPR2018) [2]	48.0		61.0
SEAM (CVPR2020) [60]	55.4		63.6
CDA (ICCV2021) [46]	55.4		67.7
AdvCAM (CVPR2021) [29]	55.6		68.0
CPN (ICCV2021) [72]	57.4		67.8
Ru et al. (CVPR2022) [44]	–		68.7
SIPE (CVPR2022) [11]	58.6		–
Du et al.(CVPR2022) [15]	61.5		69.2
MCTformer (CVPR2022) [66]	61.7		69.1
ACR*	59.4		–
ACR	**67.3**		**70.8**

4.2.2 PASCAL VOC

Seed Performance. We report mIoU for the class localization maps in Table 2, including the performances with and without affinity refinement. As shown, without affinity refinement, ACR* still outperforms most existing non-salient methods (59.4% mIoU). Our ACR achieves significantly improved initial seeds, which shows the efficacy of the proposed ACR. Without the assistance of saliency, previous best [66] also adopts transformer affinity to refine the seed, ACR outperforms it by 5.2%. We show qualitative results in Fig. 5. Further, Fig. 6 shows seeds on complex scenes with multiple objects, ACR learns precise affinity to facilitate complete object shapes with precise boundaries.

Pseudo Label Performance. The last column of Table 2 shows the pseudo segmentation label performances. Following common practice, we adopt PSA [2] to process the activation maps (seed) into pixel-wise pseudo segmentation labels. We empirically find that PSA is easily affected by false positive samples, i.e., over-activation. To avoid over-activation, we use ACR* to train the PSA network. Then, the trained PSA network will refine the ACR seeds (67.3%) into pseudo labels. As shown, our method achieves notably improved pseudo labels.

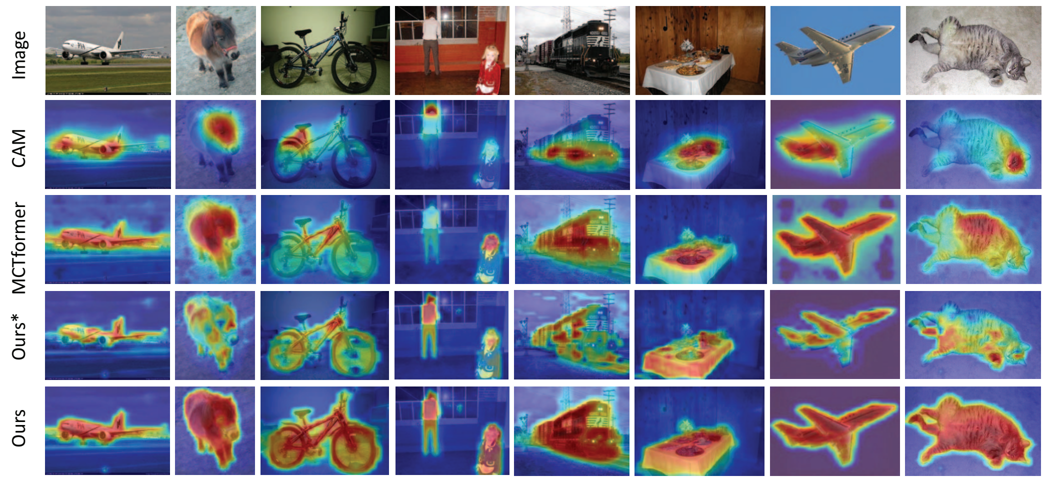

Figure 5. Visualization samples of the class localization maps of different methods. CAM: Class Activation Methods [74]. MCTformer: class localization maps of [66] which also adopt transformer attention refinement. Ours*: our class localization maps without affinity refinement. Ours: our final class localization maps with affinity refinement.

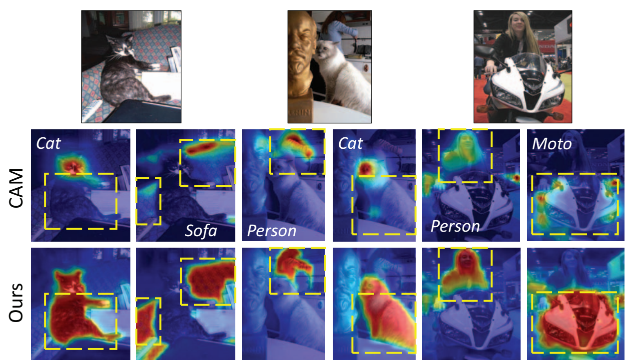

Figure 6. Visualization samples of our class localization maps with multiple classes. ACR can discriminate accurate boundaries between connected objects and localize complete shapes.

Table 3. Performance comparison of WSSS methods on PASCAL VOC 2012 *val* and *test* sets. w/ saliency: the method adopts extra saliency information. Best numbers are in bold.

Methods	Venue	w/ saliency	Val	Test
NSRM [67]	CVPR2021	✓	70.4	70.2
EDAM [62]	CVPR2021	✓	70.9	70.6
EPS [32]	CVPR2021	✓	71.0	71.8
DRS [25]	AAAI2021	✓	71.2	71.4
L2G [24]	CVPR2022	✓	72.1	71.7
Du et al. [15]	CVPR2022	✓	**72.6**	**73.6**
PSA [2]	CVPR2018		61.7	63.7
SEAM [60]	CVPR2020		64.5	65.7
CDA [46]	ICCV2021		66.1	66.8
ECS-Net [48]	ICCV2021		66.6	67.6
Du et al. [15]	CVPR2022		67.7	67.4
CPN [72]	ICCV2021		67.8	68.5
AdvCAM [29]	CVPR2021		68.1	68.0
Kweon et al. [27]	ICCV2021		68.4	68.2
ReCAM [12]	CVPR2022		68.5	68.4
SIPE [11]	CVPR2022		68.8	69.7
URN [34]	AAAI2022		69.5	69.7
ESOL [33]	NeurIPS2022		69.9	69.3
PMM [35]	ICCV2021		70.0	70.5
MCTformer[1] [66]	CVPR2022		70.6	70.3
VWL-L [43]	IJCV2022		70.6	70.7
Lee et al. [30]	CVPR2022		70.7	70.1
ACR			**71.2**	**70.9**

Semantic Segmentation Performance. Table 3 shows semantic segmentation results on PASCAL VOC. ACR achieves competitive results of 71.2% and 70.9% on *val* and *test* sets respectively, which outperform previous non-salient methods. Fig. 7 shows that the segmentation model trained with our pseudo labels can produce accurate and complete predictions. We report per-class results of PASCAL VOC in the supplementary material.

[1]Xu et al. [66] report 71.9 (*val*) and 71.7(*test*), but we are unable to reproduce these results with their provided code and seeds. We instead report our reproduced performances using their official implementation at `https://github.com/xulianuwa/MCTformer`.

4.3. Ablation Studies

Effectiveness of ACR. We propose to simultaneously regularize region activation and region affinity during the classification training. We ablate the two regularization terms in Table 4. First, we observe that region affinity can significantly improve seed quality even in the baseline, which validates the contextual encoding ability of the vision

Figure 7. Segmentation results on the PASCAL VOC *val* set.

Table 4. Ablation analysis of the two proposed consistency regularization. Act Regu: region activation consistency regularization. Aff Regu: region affinity consistency regularization. aff: whether to use affinity refinement during seed generation.

Act Regu	Aff Regu	w/o aff	w/ aff
		51.1	57.7
✓		54.8	63.5
	✓	55.4	64.9
✓	✓	**59.4**	**67.3**

Table 5. Analysis of different seeds generation methods.

Methods	Seed
CAM [74]	44.0
TS-CAM [20]	40.1
Grad-CAM [45]	50.5
Generic [8]	52.5
Generic [8] + aff	59.6
ACR	**67.3**

transformer. By introducing the two regularization terms, we observe that they contribute noticeable improvements to the performance respectively. We achieve superior results with both regularization terms, leading to an overall 15.8% mIoU increase over the vanilla transformer baseline (51.1%) , which demonstrates the effectiveness of ACR.

Different Seeds Generation Methods. To fully utilize the structure of the vision transformer, we integrate the class-to-patch gradients with region affinity to generate class localization maps as seeds. In Table 5, we ablate different seeds generation methods with models trained using our ACR. We first integrate the conventional CAM [74] method into the vision transformer, which achieves only 44%, potentially because the context aggregated in the class token is not used, and the global receptive field may spread noise. We further test various network visualization methods including TS-CAM [20], Grad-CAM [45], and Generic [8]. Notably, we refine the outputs of Generic [8] and observe a performance boost, which shows that the region affinity refinement can also be integrated with other methods for a performance increase. Ultimately, ACR achieves the best result, demonstrating the effectiveness of our seed generation method.

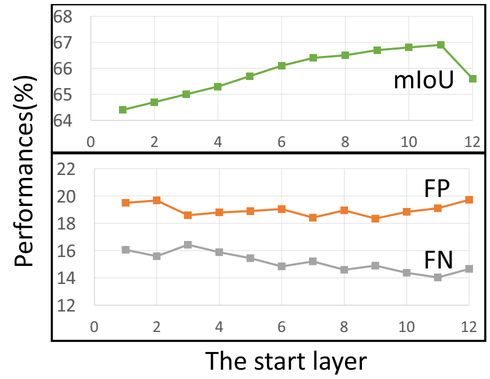

Figure 8. Performance in mIoU (%), false positive (FP), and false negative (FN) of the initial seeds generated by averaging over transformer layers. The horizontal axis represents which layer we start to obtain seeds.

Different Vision Transformer Backbones. in Table 6, we compare ACR with the previous best localization maps generated by MCTformer [66] using the same vision transformer backbone, i.e., Deit-S [55], which has substantially fewer parameters and lower complexity compared to ViT-hybrid-B. Compared to ACR, MCTformer produces better localization maps without affinity refinement (58.2 vs 56.8) since it uses multiple class tokens which require more complexity, while we only rely on a single one. However, our model is more benefited with the affinity refinement (61.7 vs 63.4). This is because our ACR learns better pair-wise affinity which leads to more integral object localization. Moreover, we integrate our ACR during the MCTformer training (Table 6: MCTformer + ACR). As shown, ACR improves MCTformer by 2.2 mIoU without affinity and 0.7 mIoU with affinity. In summary, it demonstrates that ACR can work with different transformer backbones and existing transformer-based WSSS methods as well.

Different Layers of CAM generation. We obtain the class localization maps by averaging the outputs of successive transformer layers. Following [60], we report mIoU, false positive (FP) and false negative (FN) of the localization maps when we fuse from different layers. FP indicates over-activation and FN indicates under-activation. As shown in Fig. 8, the mIoU tends to increase and FN tends to decrease when reducing the number of layers used, and both values are saturate when only the last two layers are involved. This indicates that early layers may contain unhelpful low-level noise, and with only the last two layers, we can obtain the best object completeness. Further, our seeds are generally over-activated as the FP is consistently higher than the FN. It indicates that the incompleteness issue is effectively mitigated by ACR. However, current pseudo generation methods [2, 1] are designed for under-

Table 6. Evaluation of class localization maps on Deit-S backbones. aff: whether to use affinity refinement.

Method	Backbone	w/o aff	w/ aff
MCTformer [66]	Deit-S	58.2	61.7
MCTformer + ACR	Deit-S	60.4 (↑ 2.2)	62.4 (↑ 0.7)
ACR	Deit-S	56.8	63.4

activated seeds, which might be the reason that our pseudo label improvement is not as significant as our class localization maps. A compatible solution for over-activation is expected in the future and it would potentially improve the segmentation results of ACR even further.

5. Conclusion

In this paper, we propose a simple yet effective training framework to generate better class localization maps from transformer named ACR. We exploit two types of consistencies during the classification training, i.e., region-wise activation consistency and region affinity consistency. The self-attention mechanism of the transformers is leveraged to simultaneously regularize two consistencies. We show that ACR can learn precise object localization by only one single class token as well as accurate pair-wise affinity to extract object extent. Our class localization maps significantly outperform previous methods and lead to state-of-the-art performances. ACR can be seamlessly integrated with the vision transformer network without any extra modification, which can further facilitate other transformer-based tasks.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2209–2218, Long Beach, CA, USA, 2019. Computer Vision Foundation / IEEE. 1, 3, 6, 8

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4981–4990. Computer Vision Foundation / IEEE, 2018. 1, 2, 3, 6, 7, 8

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 5

[4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European Conference on Computer Vision (ECCV)*, pages 549–565, 2016. 1, 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2, 4, 5

[6] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Mixup-cam: Weakly-supervised semantic segmentation via uncertainty regularization. In *British Machine Vision Conference (BMVC)*, Virtual Event, UK, 2020. BMVA Press. 3

[7] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8991–9000. Computer Vision Foundation / IEEE, 2020. 1, 2

[8] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers, 2021. 4, 5, 8

[9] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization, 2021. 2, 5

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):834–848, 2017. 6

[11] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua 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 (CVPR)*, pages 4288–4298, June 2022. 3, 6, 7

[12] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 969–978, 2022. 6, 7

[13] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1635–1643, 2015. 1, 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3, 4, 6

[15] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4329, 2022. 2, 3, 6, 7

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010. 5

[17] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Cian: Cross-image affinity net for weakly supervised semantic segmentation. *ArXiv e-prints*, 2018. 3

[18] Junsong Fan, Zhaoxiang Zhang, and Tieniu Tan. Employing multi-estimations for weakly-supervised semantic segmentation. In *European Conference on Computer Vision (ECCV)*, pages 332–348, 2020. 6

[19] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 1

[20] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization, 2021. 4, 8

[21] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *IEEE International Conference on Computer Vision (ICCV)*, pages 991–998. IEEE, 2011. 5

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[23] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7014–7023. Computer Vision Foundation / IEEE, 2018. 1

[24] Peng-Tao Jiang, Yuqi Yang, Qibin Hou, and Yunchao 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 (CVPR)*, pages 16886–16896, June 2022. 6, 7

[25] Beomyoung Kim, Sangeun Han, and Junmo Kim. Discriminative region suppression for weakly-supervised semantic segmentation. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1754–1761, 2021. 7

[26] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 109–117, 2011. 6

[27] Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6994–7003, October 2021. 1, 6, 7

[28] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 2021. 6

[29] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4071–4080. Computer Vision Foundation / IEEE, 2021. 6, 7

[30] Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16897–16906, 2022. 7

[31] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation, 2021. 1, 2

[32] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5495–5505. Computer Vision Foundation / IEEE, 2021. 2, 7

[33] Jinlong Li, Zequn Jie, Xu Wang, Xiaolin Wei, and Lin Ma. Expansion and shrinkage of localization for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2209.07761*, 2022. 6, 7

[34] Yi Li, Yiqun Duan, Zhanghui Kuang, Yimin Chen, Wayne Zhang, and Xiaomeng Li. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. *arXiv preprint arXiv:2112.07431*, 2021. 6, 7

[35] Yi Li, Zhanghui Kuang, Liyang Liu, Yimin Chen, and Wayne Zhang. Pseudo-mask matters in weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6964–6973, October 2021. 7

[36] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3159–3167. Computer Vision Foundation / IEEE, 2016. 1, 2

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 5

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 3

[39] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6913–6922. Computer Vision Foundation / IEEE, 2021. 1, 2

[40] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1742–1750, 2015. 1, 2

[41] Zhen Qin, Xiaodong Han, Weixuan Sun, Dongxu Li, Lingpeng Kong, Nick Barnes, and Yiran Zhong. The devil in linear transformer. *arXiv preprint arXiv:2210.10340*, 2022. 2

[42] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran

Zhong. cosformer: Rethinking softmax in attention. In *International Conference on Learning Representations*, 2022. 2

[43] Lixiang Ru, Bo Du, Yibing Zhan, and Chen Wu. Weakly-supervised semantic segmentation with visual words learning and hybrid pooling. *International Journal of Computer Vision*, 130(4):1127–1144, 2022. 6, 7

[44] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo 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*, pages 16846–16855, 2022. 2, 3, 5, 6

[45] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016. 8

[46] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7004–7014, 2021. 1, 6, 7

[47] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. *arXiv preprint arXiv:2007.01947*, 2020. 3

[48] Kunyang Sun, Haoqing Shi, Zhengming Zhang, and Yongming Huang. Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7283–7292, October 2021. 1, 3, 7

[49] Weixuan Sun, Zheyuan Liu, Yanhao Zhang, Yiran Zhong, and Nick Barnes. An alternative to wsss? an empirical study of the segment anything model (sam) on weakly-supervised semantic segmentation problems. *arXiv preprint arXiv:2305.01586*, 2023. 2

[50] Weixuan Sun, Zhen Qin, Hui Deng, Jianyuan Wang, Yi Zhang, Kaihao Zhang, Nick Barnes, Stan Birchfield, Lingpeng Kong, and Yiran Zhong. Vicinity vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 3

[51] Weixuan Sun, Jing Zhang, and Nick Barnes. 3d guided weakly supervised semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1

[52] Weixuan Sun, Jing Zhang, and Nick Barnes. Inferring the class conditional response map for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2878–2887, 2022. 1, 3

[53] Weixuan Sun, Jing Zhang, Zheyuan Liu, Yiran Zhong, and Nick Barnes. Getam: Gradient-weighted element-wise transformer attention map for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2112.02841*, 2021. 2, 3, 4, 5

[54] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *European Conference on Computer Vision (ECCV)*, pages 507–522, 2018. 1

[55] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pages 10347–10357. PMLR, 2021. 8

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[57] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7158–7166. Computer Vision Foundation / IEEE, 2017. 1

[58] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer, 2021. 2, 3

[59] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1354–1362. Computer Vision Foundation / IEEE, 2018. 1, 6

[60] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12275–12284. Computer Vision Foundation / IEEE, 2020. 1, 2, 3, 5, 6, 7, 8

[61] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 2, 3

[62] Tong Wu, Junshi Huang, Guangyu Gao, Xiaoming Wei, Xiaolin Wei, Xuan Luo, and Chi Harold Liu. Embedded discriminative attention mechanism for weakly supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16765–16774. Computer Vision Foundation / IEEE, 2021. 6, 7

[63] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[64] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 989–998, June 2022. 3

[65] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, Ferdous Sohel, and Dan Xu. Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6984–6993, October 2021. 3, 6

[66] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 9

[67] Yazhou Yao, Tao Chen, Guo-Sen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang. Non-salient region object mining for weakly supervised semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2623–2632. Computer Vision Foundation / IEEE, 2021. 7

[68] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *ArXiv e-prints*, 2018. 2

[69] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019. 2

[70] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 12765–12772, 2020. 2

[71] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2009.12547*, 2020. 6

[72] Fei Zhang, Chaochen Gu, Chenyue Zhang, and Yuchao Dai. Complementary patch for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7242–7251, October 2021. 2, 3, 6, 7

[73] Tianyi Zhang, Guosheng Lin, Weide Liu, Jianfei Cai, and Alex Kot. Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[74] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929. Computer Vision Foundation / IEEE, 2016. 1, 2, 5, 7, 8

[75] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2