Learning Multi-Modal Class-Specific Tokens for Weakly Supervised Dense Object Localization

Lian Xu[1], Wanli Ouyang[2], Mohammed Bennamoun[1], Farid Boussaid[1], and Dan Xu[3]
[1]The University of Western Australia [2]Shanghai AI Laboratory
[3]Hong Kong University of Science and Technology
{lian.xu,mohammed.bennamoun,farid.boussaid}@uwa.edu.au,
wanli.ouyang@sydney.edu.au, danxu@cse.ust.hk

Abstract

Weakly supervised dense object localization (WSDOL) relies generally on Class Activation Mapping (CAM), which exploits the correlation between the class weights of the image classifier and the pixel-level features. Due to the limited ability to address intra-class variations, the image classifier cannot properly associate the pixel features, leading to inaccurate dense localization maps. In this paper, we propose to explicitly construct multi-modal class representations by leveraging the Contrastive Language-Image Pre-training (CLIP), to guide dense localization. More specifically, we propose a unified transformer framework to learn two-modalities of class-specific tokens, i.e., class-specific visual and textual tokens. The former captures semantics from the target visual data while the latter exploits the class-related language priors from CLIP, providing complementary information to better perceive the intra-class diversities. In addition, we propose to enrich the multi-modal class-specific tokens with sample-specific contexts comprising visual context and image-language context. This enables more adaptive class representation learning, which further facilitates dense localization. Extensive experiments show the superiority of the proposed method for WSDOL on two multi-label datasets, i.e., PASCAL VOC and MS COCO, and one single-label dataset, i.e., OpenImages. Our dense localization maps also lead to the state-of-the-art weakly supervised semantic segmentation (WSSS) results on PASCAL VOC and MS COCO. [1]

1. Introduction

Fully supervised dense prediction tasks have achieved great success, which however comes at the cost of expensive pixel-level annotations. To address this issue, recent works have investigated the use of weak labels, such as image-level labels, to generate dense object localization maps as pseudo labels for those tasks. For the weakly supervised object localization (WSOL) task, most methods evaluate localization results on the bounding-box level and a few recent methods [5] evaluate on the pixel level. We use WSDOL to focus on the pixel-level evaluation, which is critical for downstream dense prediction tasks such as WSSS.

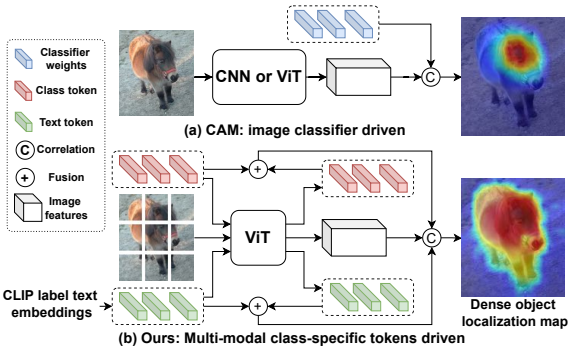

Figure 1. (a) CAM exploits the correlation between the image classifier and the pixel features. (b) We propose to construct multi-modal class-specific tokens to guide dense object localization.

Previous works have exploited Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) [7] for WSDOL with image-level labels [21, 40]. These methods have generally relied on Class Activation Mapping (CAM) [48], which generates class-specific localization maps by computing the correlation between the class-specific weight vectors of the image classifier and every pixel feature vector. However, image classifiers generally have a limited ability to address the intra-class variation, let alone at the pixel level. This thus leads to inaccurate dense localization results. In the conventional fully supervised learning paradigm, the image classification model aims to convert images to numeric labels, ignoring the context of the labels. Hence, it tends to learn the pattern that maximizes the inter-class differences but disregards the intra-class diversities. This largely restricts the model's ability of

[1] https://github.com/xulianuwa/MMCST

understanding semantic objects.

Recently, Vision-Language (VL) models have attracted much attention. In particular, CLIP, a representative VL model, pre-trained on 400 million image-text pairs that are readily available publicly, has been successfully applied to a number of downstream tasks, due to its strong generalization ability. CLIP introduces a contrastive representation learning method that constrains an image to match its related text while dis-matching the remaining texts from the same batch, in a multi-modal embedding space. This enables the model to perceive the differences across images, thus facilitating it to better discriminate intra-class samples.

Motivated by these observations, we propose to leverage the strong representations of visual concepts encoded by the pre-trained CLIP language model to guide the dense object localization. More specifically, we extract the class-related text embeddings by feeding the label prompts to the pre-trained CLIP language model. As shown in Figure 1, we propose a unified transformer framework which includes multi-modal class-specific tokens, *i.e.,* class-specific visual tokens and class-specific textual tokens. The class-specific visual tokens aim to capture visual representations from the target image dataset, while the class-specific textual tokens take the rich language semantics from the CLIP label text embeddings. These two modalities of class-specific tokens, with complementary information, are jointly used to correlate pixel features, contributing to better dense localization.

In order to construct more adaptive class representations, which can better associate the sample-specific local features for dense localization, we propose to enhance the global multi-modal class-specific tokens with sample-specific contextual information. To this end, we introduce two designs: (**i**) at the feature level, we use the sample-specific visual context to enhance both the class-specific visual and textual tokens. This is achieved by combining these global tokens with their output local counterparts which aggregate the patch tokens of the image through the self-attention layers; (**ii**) at the loss level, we introduce a regularization contrastive loss to encourage the output text tokens to match the CLIP image embeddings. This allows the CLIP model to be better adapted to our target datasets. Moreover, due to its image-language matching pre-training objective, the CLIP image encoder is learned to extract the image embeddings that match the CLIP text embeddings of their corresponding image captions. We thus argue that through this contrastive loss, the rich image-related language context from the CLIP could be implicitly transferred to the text tokens, which are more beneficial for guiding the dense object localization, compared to the simple label prompts.

In summary, the contribution of this work is three-fold:

- We propose a new WSDOL method by explicitly constructing multi-modal class representations in a unified transformer framework.
- The proposed transformer includes class-specific visual tokens and class-specific textual tokens, which are learned from different data modalities with diverse supervisions, thus providing complementary information for more discriminative dense localization.
- We propose to enhance the multi-modal global class representations by using sample-specific visual context via the global-local token fusion and transferring the image-language context from the pre-trained CLIP via a regularization loss. This enables more adaptive class representations for more accurate dense localization.

The proposed method achieved the state-of-the-art results on PASCAL VOC 2012 (72.2% on the test set) and MS COCO 2014 (45.9% on the validation set) for WSSS.

2. Related Work

Weakly supervised dense object localization. Most existing methods rely on CAM which generates class-specific localization maps based on the correlation between the class weights of the image classifier and the pixel features. However, CAM produces inaccurate localization maps where only the most discriminative object regions are activated. Most previous works are based on CNNs and they have generally focused on improving image feature learning in various ways, such as manipulating training images [20, 32], modifying the classification architecture [35, 41], and designing new losses [4, 8, 14]. Recently, ViTs have made breakthroughs in many computer vision tasks, outperforming CNNs, due to their self-attention based processing blocks that enable long-dependence modeling and cater for different data modalities. Recent ViT-based methods, TS-CAM [10] and MCTformer [40] have exploited the transformer attention [12] to generate localization maps, but they still need to integrate CAM into ViTs to achieve a decent localization performance. For these CAM-based methods, one problem that has not been well investigated, is that the class weights used to associate pixel features are the global class representations for the entire dataset. Thus, they cannot properly address the complex pixel-level intra-class variations across samples. Chen *et al.* [3] proposed the Image-Specific CAM (IS-CAM) using the image-specific class prototypes constructed by applying the masked-average pooling on the image feature maps, and imposed a consistency loss between the IS-CAM and the raw CAM to enhance the feature learning for better localization.

In contrast to all these CAM-based methods, we propose to explicitly construct multi-modal class representations for dense object localization.

Vision-language models. Recent progress in vision-language pre-training with large-scale datasets has provided a rich source of transferable information. In particular, a representative VL model, *i.e.,* CLIP, brought a significant

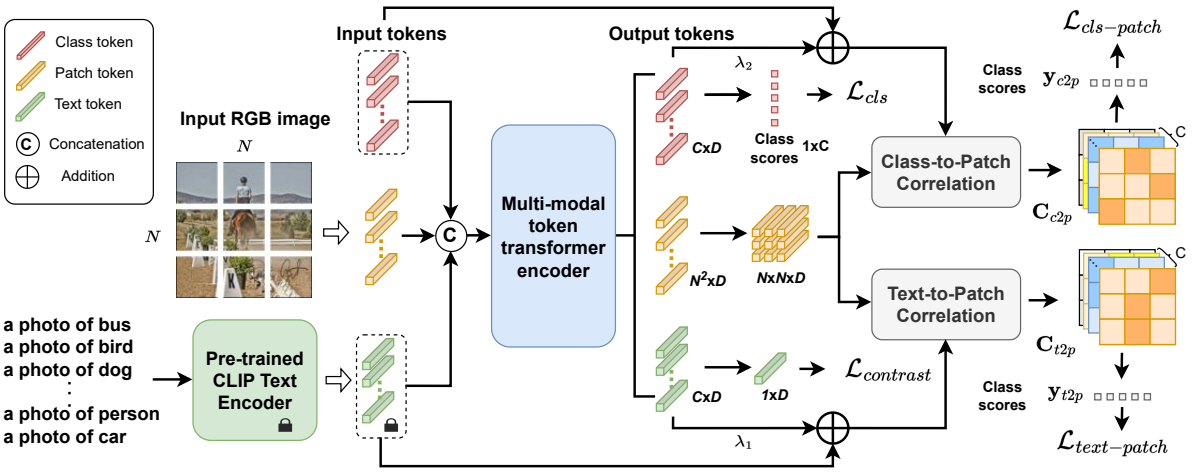

Figure 2. An overview of the proposed transformer framework. An input RGB image is first split and then embedded into patch tokens. The proposed transformer additionally includes class-specific visual tokens (*i.e.,* input class tokens) and class-specific textual tokens (*i.e.,* input text tokens), in which the input text tokens are initialized by the pretrained CLIP text embeddings of the class-related prompts. The output class tokens are averaged to predict class scores. We fuse the input and output class tokens to correlate the output patch tokens. This generates the class-to-patch correlation maps, which are then globally pooled to produce class scores. Similarly for the text tokens. We also apply a contrastive regularization loss (Figure 4) on the averaged output text tokens to transfer the rich VL context from CLIP.

insight that one can use natural language to connect the visual concept, thus allowing a flexible prediction space for a wide knowledge transfer. The pre-trained CLIP model has been used in a variety of open-world visual tasks. Gu *et al.* [11] proposed to distill knowledge from the pre-trained CLIP image and text encoders to learn open-vocabulary object detectors. Li *et al.* [24] used the pre-trained text embeddings as semantic label representations to correlate visual features for semantic segmentation. Similar language-driven segmentation methods have been proposed for 2D images [29] and 3D point clouds [30]. In contrast to these methods which have dense supervision, we investigate a pixel-text matching problem in a weakly supervised setting.

Xie *et al.* [37] proposed a CLIP-based WSDOL method, which constrains the CAM activated object regions and the background image regions to match and dis-match the corresponding label prompt in the CLIP embedding space, respectively. It also relies on the additionally pre-defined text descriptions to suppress the co-occurring backgrounds. In contrast, without introducing new visual concepts, we propose to directly use the CLIP label text embeddings to infer dense object localization maps, and further refine them with sample-specific contexts, achieving better results in both WSDOL and WSSS (see Table 1 and Table 3).

3. Method

Overview. We propose a unified transformer framework to construct multi-modal class-specific tokens for weakly supervised dense object localization. As illustrated in Figure 2, an input RGB image is split into patches and then embedded into a sequence of patch tokens. Additionally, the proposed transformer has two sets of class-specific tokens, *i.e.,* class-specific visual and textual tokens. The class-specific visual tokens are all initialized by the pre-trained weights of the class token of ViT, and the class-specific textual tokens are initialized by the pre-trained CLIP text embeddings of the class-related prompts and kept constant during training. With the added positional embeddings, these three types of tokens are concatenated and serve as input to the proposed transformer. The input tokens go through consecutive masked self-attention based transformer encoding layers. The output class tokens are averaged along the channel dimension to predict class scores. The input and output class/text tokens are fused and then correlated with the output patch tokens, generating the class-to-patch and text-to-patch correlation maps. These two correlation maps are then globally pooled to produce class scores, respectively.

During training, the class predictions are supervised by the image-level ground-truth labels via the classification loss. In addition, a batch-softmax contrastive loss is used to encourage the averaged output text token to match the CLIP image embedding. This allows the image-language context transfer, further enabling the class representations to be more sample-adaptive for better localization. During inference, a multi-modal correlation map is generated by fusing the two correlation maps, which are further refined by the transformer attention as suggested in [40], generating the final dense object localization maps.

3.1. Multi-modal class-specific token learning

In order to construct multi-modal class representations to guide dense object localization, we propose a multi-modal

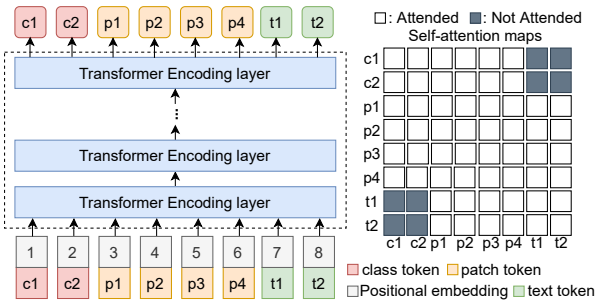

Figure 3. The detailed structure of the proposed multi-modal token transformer encoder.

token transformer framework. As illustrated in Figure 2, an input RGB image is first split into $N \times N$ patches, which are then embedded into a sequence of N^2 patch tokens. We additionally create C class-specific visual tokens and C class-specific textual tokens, where C is the number of classes. With the added learnable positional embeddings, these three types of tokens are first concatenated and then fed into the transformer encoder (Figure 3). In order to allow class tokens and text tokens to fully interact with patch tokens but not competing with each other, we use the masked self-attention as shown in Figure 3 (right). Finally, the transformer encoder outputs three types of tokens accordingly.

Class-specific visual token learning. Inspired by MCTformer [40], we use C class tokens to learn class-specific representations. The output class tokens $\mathbf{T}_{cls}^{out} \in \mathbb{R}^{C \times D}$ from the proposed multi-modal token transformer encoder are processed by channel-wise averaging, producing C class scores $\mathbf{y}_{cls} \in \mathbb{R}^{C}$. The class scores are supervised by the image-level ground-truth labels $\mathbf{y} \in \mathbb{R}^{C}$ using the multi-label soft margin loss ($MLSM$):

$$\mathcal{L}_{cls} = MLSM(\mathbf{y}_{cls}, \mathbf{y}) = - \frac{1}{C} \sum_{i=1}^{C} \mathbf{y}^i \log \sigma(\mathbf{y}_{cls}^i) + (1 - \mathbf{y}^i) \log(1 - \sigma(\mathbf{y}_{cls}^i)). \quad (1)$$

Class-specific textual token learning. CLIP presents a novel way of representing visual concepts by using textual prompts, which provide highly complementary information to the class representations learned from only visual data. It thus exhibits a great potential to better facilitate the dense object localization. This motivates us to construct C class-specific textual tokens by leveraging the rich semantics learned by the pre-trained CLIP language model. More specifically, we create C text prompts using the template of "a photo of [CLS]", where [CLS] refers to each class name in the label set of a given dataset. These text prompts are then fed into the pre-trained CLIP text encoder, generating C text embeddings. We use the C pre-trained label text embeddings to initialize the input text tokens, which are kept constant during training to retain the powerful representation ability of the pre-trained CLIP.

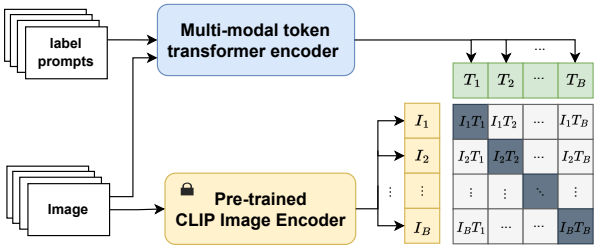

Figure 4. The regularization loss. The proposed multi-modal token transformer is encouraged to produce text tokens to match the image embeddings of the pre-trained CLIP image encoder, enabling image-language context transfer.

3.2. Sample-specific local contexts

To construct class-specific tokens tailored for each sample, we propose to exploit sample-specific local contexts to complement the global input class-specific tokens.

Visual context. Inspired by the context-aware prompting [29], we also use the visual context to refine the text embeddings. In contrast to [29] which employs an additional cross-attention module to model the interaction between textual and visual features, the inherent self-attention blocks of the proposed transformer encoder enables the text tokens to aggregate information from patch tokens, capturing the sample-specific visual context. Intuitively, the text embeddings of simple prompts (*i.e.,* "a photo of [CLS]"), after being refined by the sample-specific visual context, could represent more accurate image-related text. Given that the input text tokens $\mathbf{T}_{txt}^{in} \in \mathbb{R}^{C \times D}$ are global textual representations for each label while the output text tokens $\mathbf{T}_{txt}^{out} \in \mathbb{R}^{C \times D}$ capture sample-specific representations, we can obtain the enhanced class-specific textual tokens by fusing the input and output text tokens: $\mathbf{T}_{txt} = \mathbf{T}_{txt}^{in} + \lambda_1 \cdot \mathbf{T}_{txt}^{out}$. Similarly, the enhanced class-specific visual tokens are obtained by $\mathbf{T}_{cls} = \mathbf{T}_{cls}^{in} + \lambda_2 \cdot \mathbf{T}_{cls}^{out}$, where λ_1 and λ_2 are two learnable weights.

Image-language context. To further ensure that the output text tokens can capture meaningful sample-specific context, we propose to impose a contrastive loss on the output text tokens by leveraging the pre-trained CLIP image model. More specifically, as illustrated in Figure 4, given a batch of B input image-text pairs (the input text for each image is the same, *i.e,* C label prompts), the proposed transformer encoder outputs B text tokens. These B text tokens are used to compute the similarity matrix $\mathbf{S} \in \mathbb{R}^{B \times B}$ with B visual embeddings by feeding the input images to the pre-trained CLIP image encoder. When the output text token and the CLIP image embedding are from the same image, they form a positive pair. Otherwise, they form a negative pair. Such pair construction is reasonable in that: (**i**) The CLIP image embedding of an image matches its corresponding text description which commonly contains richer information than image-level labels. Thus, even two images with the same

class label generally have distinct CLIP image embeddings. (**ii**) Although the input text tokens are the same for each image, they are progressively refined through interactions with patch tokens at every layer of our model. This enables the output text tokens to represent specific input image content, *i.e.,* different images have different output text tokens. Therefore, the B scores in the diagonal of the similarity matrix are encouraged to be maximized and the other $B^2 - B$ similarity scores are minimized. This is implemented by computing the cross-entropy classification loss between the similarity matrix $\mathbf{S}$ and the ground-truth labels, *i.e.,* an identity matrix $\mathbf{I} \in \mathbb{R}^{B\times B}$:

$$\mathcal{L}_{contrast} = CrossEntropy(\mathbf{S}, \mathbf{I}). \tag{2}$$

The benefit is two-fold: (**i**) encouraging the output text tokens to match the pre-trained CLIP image embeddings, allows us to better adapt the pre-trained CLIP model to our target datasets. This also implicitly transfers richer image-related language contexts to the output text tokens, which are more beneficial for guiding the dense object localization, compared to the simple label prompts; (**ii**) through the batch-softmax cross-entropy loss on the similarity matrix, each text token is also constrained to be discriminative from other text tokens in the same batch, thereby enhancing the sample-specific characteristic of the text tokens.

3.3. Image-label supervised multi-modal class-to-patch correlation learning

Given the multi-modal class-specific tokens, *i.e.,* $\mathbf{T}_{txt}$ and $\mathbf{T}_{cls}$, the class-specific dense localization can be inferred by computing the correlation between the class-specific tokens and the patch tokens. More specifically, the output patch tokens from the proposed multi-modal token transformer encoder are first linearly projected and then transposed to 2D feature maps $\mathbf{F} \in \mathbb{R}^{N\times N\times D}$, where D is the feature dimension. The text-to-patch correlation maps $\mathbf{C}_{t2p} \in \mathbb{R}^{N\times N\times C}$ and the class-to-patch correlation maps $\mathbf{C}_{c2p} \in \mathbb{R}^{N\times N\times C}$ can be computed as: $\mathbf{C}_{t2p} = \mathbf{F}\mathbf{T}_{txt}^{\top}$ and $\mathbf{C}_{c2p} = \mathbf{F}\mathbf{T}_{cls}^{\top}$, respectively.

In contrast to recent language-driven dense prediction works [24, 29] where pixel-level ground-truth labels are used to supervise the learning of the class-to-patch correlations, only image-level labels are available in our case. Based on the assumption that the weighted sum of a class-to-patch correlation map can be regarded as the global correlation between the class and the whole image, we relax this class-to-patch correlation problem into a class-to-image correlation problem, which can thus be formulated into a classification task. Given that the commonly used global average pooling (GAP) and global max pooling inevitably either over-estimate or under-estimate the size of the object regions when generating the global correlation scores, respectively [16], we thus adopt the Global Weighted Ranking Pooling (GWRP) method [16]. Compared to GAP which assigns the same weight to each patch for aggregation, GWRP assigns different weights according to the ranking of the correlation scores of all patches for each class:

$$G_c(\mathbf{X}_\star) = \frac{1}{Z(d)} \sum_{j=1}^{N^2} d^{j-1} X_\star^{r_j,c}, \tag{3}$$

where $\mathbf{X}_\star \in \mathbb{R}^{N^2\times C}$ are the flattened correlation maps; $\mathbf{y}_\star = G(\mathbf{X}_\star) \in \mathbb{R}^C$ are the aggregated class correlation scores; r_j is the index of the ranking, *i.e.,* for a class c, $X_\star^{r_1,c} > X_\star^{r_2,c} > \cdots > X_\star^{r_{N^2},c}$; $Z(d) = \sum_{j=1}^{N^2} d^{j-1}$, d is a decay parameter. Therefore, the aggregated class-to-patch and text-to-patch correlation scores can be obtained as:

$$\mathbf{y}_{c2p} = G(\mathbf{X}_{c2p}), \tag{4}$$

$$\mathbf{y}_{t2p} = G(\mathbf{X}_{t2p}). \tag{5}$$

These class scores can then be supervised by the ground-truth image-level labels $\mathbf{y}$ via the classification loss:

$$\mathcal{L}_{class-patch} = MLSM(\mathbf{y}_{c2p},\ \mathbf{y}), \tag{6}$$

$$\mathcal{L}_{text-patch} = MLSM(\mathbf{y}_{t2p},\ \mathbf{y}). \tag{7}$$

The total objective loss function for training the proposed multi-modal token transformer is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cls}+\mathcal{L}_{contrast}+\mathcal{L}_{text-patch}+\mathcal{L}_{class-patch} \tag{8}$$

3.4. Class-specific dense localization inference

Once the training of the proposed multi-modal token transformer is completed, the correlations between two modalities of class-specific tokens and the patch tokens are computed to produce two maps, *i.e.,* text-to-patch correlation maps $\mathbf{C}_{t2p}$ and class-to-patch correlation maps $\mathbf{C}_{c2p}$, respectively. We further combine these two maps via an element-wise sum to obtain the multi-modal class-specific dense localization maps $\mathbf{C}_{mm} \in \mathbb{R}^{N\times N\times C}$:

$$\mathbf{C}_{mm} = \mathbf{C}_{t2p} + \mathbf{C}_{c2p}. \tag{9}$$

Moreover, similar to MCTformer [40], the multi-class token based transformer can produce class-specific attention maps $\mathbf{A}_{c2p}$ and a patch-level pairwise affinity map $\mathbf{A}_{p2p}$ from the transformer self-attention maps. Given their different localization mechanisms, the transformer self-attention maps can well complement the proposed multi-modal class-specific localization maps. We thus perform an element-wise multiplication ($\circ$) on these two maps to produce the fused maps $\hat{\mathbf{M}} \in \mathbb{R}^{N\times N\times C}$, which are further refined by the patch-level pairwise affinity map:

$$\hat{\mathbf{M}} = \mathbf{C}_{mm} \circ \mathbf{A}_{c2p}, \tag{10}$$

$$\mathbf{M}(i,j,c) = \sum_{k=1}^{N}\sum_{l=1}^{N} \mathbf{A}_{p2p}(i,j,k,l) \cdot \hat{\mathbf{M}}(k,l,c), \tag{11}$$

where $\mathbf{M} \in \mathbb{R}^{N \times N \times C}$ are the refined fused maps, which are then up-sampled to the original size of the input image before being normalized via the min-max normalization, to generate the final dense object localization maps.

4. Experiments

4.1. Experimental settings

Datasets. We evaluated the proposed method on three datasets including two multi-label datasets, *i.e.,* PASCAL VOC 2012 [9] and MS COCO 2014 [26], and one single-label dataset, *i.e.,* OpenImages [5]. **PASCAL VOC 2012** has 20 foreground object classes and one background class. It is split into three subsets, *i.e.,* training (*train*), validation (*val*) and test sets including 1,464, 1,449, and 1,456 images, respectively. Following the common practice [18, 19], the training set was augmented to 10,582 images by adding the data from [13]. **MS COCO 2014** has 80 foreground object classes and one background class. It has around 80K training images and 40K images for evaluation. **OpenImages** has 37,319 images in total and 100 object classes. It is split into a training set with 29,819 images, a validation set with 2,500 images and a test set with 5,000 images.

Evaluation metrics. For the multi-label datasets, we followed the common practice [18, 20, 21] to use the mean Intersection-over-Union (mIoU) to evaluate the multi-label dense localization maps on the *train* set, and used mIoU to evaluate WSSS results on the *val* and *test* sets. Results on the PASCAL VOC *test* set were obtained from the online official evaluation server. For the single-label dataset, we focused on the pixel-level evaluation and followed the prior works [5, 52] which used the peak Intersection-over-Union (pIoU) and the pixel average precision (PxAP) to evaluate the single-label dense localization maps on the *test* set.

Implementation details. We built the proposed transformer using ViT-base as backbone. We used the Adam optimizer with the initial learning rate of 5e-4 and a batch size of 32. We trained our network for 60 epochs on PASCAL VOC and MS COCO and 10 epochs on OpenImages. For the VL model, we used the publicly available pre-trained CLIP model (ViT-B/16). To generate pseudo masks for WSSS, we followed prior works [20, 21, 32, 44] to use IRNet [1] to post-process our dense localization maps. For semantic segmentation, we use ResNet38-based Deeplab-V1. More details can be found in the Supplementary Materials.

4.2. Comparison with State-of-the-art of WSDOL

Multi-label dense localization. Table 1 reports the evaluation results of the generated multi-label dense localization maps, which are commonly used as seeds to generate pseudo masks in WSSS. The proposed method achieved mIoUs of 66.3% and 40.9% on the train sets of PASCAL VOC 2012 and MS COCO 2014, respectively, outperforming the state-of-the-art methods by significant margins.

Table 1. Evaluation of the generated multi-class multi-label dense localization maps in terms of mIoU (%) on the PASCAL VOC 2012 and MS COCO 2014 *train* sets. Cls.: Classification. † denotes the reproduced result by [18], ‡ denotes the reproduced result by [44], and * denotes our reproduced result.

Method	Cls. Backbone	VOC	COCO
CAM (CVPR16) [48]	ResNet50	48.8	33.5†
SEAM (CVPR20) [34]	ResNet38	55.4	25.1‡
RIB (NeurIPS21) [18]	ResNet50	56.5	36.5
AdvCAM (CVPR21) [19]	ResNet38	55.6	37.2
CLIMS (CVPR22) [37]	ResNet50	56.6	-
SIPE (CVPR22) [4]	ResNet50	58.6	-
W-OoD (CVPR22) [21]	ResNet50	59.1	-
Du *et al.* (CVPR22) [8]	ResNet38	61.5	-
TS-CAM (ICCV21) [10]	ViT-small	41.3	-
MCTformer (CVPR22) [40]	ViT-small	61.7	-
MCTformer (CVPR22) [40]	ViT-base	62.3*	-
Ours	ViT-base	**66.3**	**40.9**

Table 2. Evaluation of multi-class single-label dense localization on the OpenImages test set.

Method	Cls. backbone	pIoU	PxAP
CAM (CVPR16) [48]	ResNet50	43.0	58.2
HAS (ICCV17) [31]	ResNet50	41.9	55.1
ACoL (CVPR18) [47]	ResNet50	41.7	56.4
SPG (ECCV18) [46]	ResNet50	41.8	55.8
ADL (CVPR19) [6]	ResNet50	42.1	55.0
CutMix (ICCV19) [43]	ResNet50	42.7	57.6
PAS (ECCV20) [2]	ResNet50	-	60.9
IVR (ICCV21) [15]	ResNet50	-	58.9
Zhu *et al.* (CVPR22) [52]	ResNet50	49.7	65.4
CREAM (CVPR22) [38]	ResNet50	-	64.7
Zhu *et al.* (ECCV22) [51]	ResNet50	52.2	67.7
Ours	ViT-base	**57.6**	**73.3**

Single-label dense localization. We also evaluated the effectiveness of the proposed method for single-label dense object localization. We used the recently proposed OpenImages dataset [5], which has more challenging background context than the commonly used ones for weakly supervised object localization. As shown in Table 2, the proposed method attained a pIoU of 57.6% and a PxAP of 73.3%, achieving better results than the state-of-the-art methods.

4.3. Comparison with State-of-the-art of WSSS

PASCAL VOC 2012. As shown in Table 3, the proposed method achieved mIoUs of 72.2% and 72.2% on the *val* and *test* sets of PASCAL VOC 2012, respectively, outperforming other WSSS methods only using image-level labels.

MS COCO 2014. Table 4 also shows that, on a more challenging dataset, *i.e.,* MS COCO 2014, the proposed method attained the best mIoU of 45.9%, which is significantly better than those relying on saliency maps, achieving the new state-of-the-art result. These results clearly demonstrate the effectiveness and the good generalization ability of the proposed method for WSSS.

Table 3. Performance comparison of the state-of-the-art WSSS methods on the PASCAL VOC 2012 *val* and *test* sets. Seg.: DeepLab version. Sup.: supervision; I: image-level ground-truth labels; S: off-the-shelf saliency maps; L: pre-trained VL model.

Method	Backbone	Seg.	Sup.	Val	Test
EDAM (CVPR21) [36]	ResNet101	V2	I+S	70.9	70.6
EPS (CVPR21) [23]	ResNet101	V1	I+S	71.0	71.8
Yao *et al.* (CVPR21) [41]	ResNet101	V2	I+S	68.3	68.5
AuxSegNet (ICCV21) [39]	ResNet38	V1	I+S	69.0	68.6
Li *et al.* (CVPR22) [25]	ResNet101	V2	I+S	72.0	72.9
Du *et al.* (CVPR22) [8]	ResNet101	V2	I+S	**72.6**	**73.6**
RCA (CVPR22) [50]	ResNet38	V2	I+S	72.2	72.8
L2G (CVPR22) [14]	ResNet38	V1	I+S	72.0	73.0
AdvCAM (CVPR21) [19]	ResNet101	V2	I	68.1	68.0
ECS-Net (ICCV21) [33]	ResNet38	V1	I	66.6	67.6
Kweon *et al.* (ICCV21) [17]	ResNet38	V1	I	68.4	68.2
CDA (ICCV21) [32]	ResNet38	V1	I	66.1	66.8
Zhang *et al.* (ICCV21) [45]	ResNet38	V1	I	67.8	68.5
MCTformer (CVPR22) [40]	ResNet38	V1	I	71.9	71.6
AMN (CVPR22) [22]	ResNet101	V2	I	70.7	70.6
W-OoD (CVPR22) [21]	ResNet38	V1	I	70.7	70.1
SIPE (CVPR22) [3]	ResNet38	V1	I	68.2	69.7
Yoon *et al.* (ECCV22) [42]	ResNet38	V1	I	70.9	71.7
CLIMS (CVPR22) [37]	ResNet101	V2	I+L	69.3	68.7
Ours	ResNet38	V1	I+L	**72.2**	**72.2**

Table 4. Performance comparison of state-of-the-art WSSS methods on the MS COCO 2014 *val* set in terms of mIoU (%).

Method	Backbone	Seg.	Sup.	Val
EPS (CVPR21) [23]	VGG16	V2	I+S	35.7
RCA (CVPR22) [50]	VGG16	V2	I+S	36.8
AuxSegNet (ICCV21) [39]	ResNet38	V1	I+S	33.9
L2G (CVPR22) [14]	ResNet101	V2	I+S	**44.2**
Kweon *et al.* (ICCV21) [17]	ResNet38	V1	I	36.4
CDA (ICCV21) [32]	ResNet38	V1	I	33.2
AdvCAM (CVPR21) [19]	ResNet101	V2	I	44.4
MCTformer (CVPR22) [40]	ResNet38	V1	I	42.0
Li *et al.* (CVPR22) [25]	ResNet101	V2	I	44.7
AMN (CVPR22) [22]	ResNet101	V2	I	44.7
SIPE (CVPR22) [3]	ResNet38	V1	I	43.6
Yoon *et al.* (ECCV22) [42]	ResNet38	V1	I	44.8
Ours	ResNet38	V1	I+L	**45.9**

4.4. Ablation studies

Effect of learning multi-modal class-specific tokens. As shown in Table 5, without text tokens, only relying on the multi-class token transformer attention results in a dense localization mIoU of 57.5%; By learning the multi-class tokens to correlate patch tokens using image-level supervision with GAP, a significant gain of 4.6% was achieved in mIoU. GWRP outperforms GAP, improving the dense localization result to 62.7%. By additionally learning class-specific textual tokens, the resulting multi-modal correlation maps driven by the global/input multi-modal class-specific tokens, attained an better mIoU of 64.1%, compared to that of the multi-modal correlation maps driven by the local/output multi-modal class-specific tokens. The fusion of the global and local multi-modal class-specific tokens leads to the best dense localization maps with an mIoU of 66.3%. These results demonstrate the effectiveness of the proposed method of learning multi-modal class-specific tokens in producing accurate class-specific dense localization maps.

Table 5. Evaluation of the class-specific dense localization maps generated by learning correlations (cor.) between class-specific tokens and patch tokens on the PASCAL VOC 2012 *train* set.

Configuration	Pooling	mIoU
Multi-class token attention	-	57.5
+ class-to-patch correlation	GAP	62.1
+ class-to-patch correlation (C2P cor.)	GWRP	62.7
+ Multi-modal (C2P + text-to-patch) cor. (global)	GWRP	64.1
+ Multi-modal cor. (local)	GWRP	63.3
+ Multi-modal cor. (global + local)	GWRP	**66.3**

Table 6. Evaluation of the class-specific dense localization by different regularization methods for sample-specific local context learning on the PASCAL VOC 2012 *train* set. CE: cross-entropy.

Visual context	Image-language context		mIoU
	Prior knowledge	Regularization loss	
✗	-	-	64.1
✓	-	-	64.8
✓	CLIP caption embed.	L1	63.7
✓	CLIP caption embed.	Batch-contrast CE	65.1
✓	CLIP image embed.	Batch-contrast CE	**66.3**

Figure 5 visualizes the generated dense object localization maps in different configurations. The multi-class token transformer attention only localizes partial object regions and includes many noises, such as the dog and the background in the third and second rows of Figure 5 (b), respectively. Learning the multi-class tokens to correlate patch tokens leads to more complete object localization maps, such as the sofa in the third row. This also results in several falsely activated regions near the object, such as the chair in the second rows of Figure 5 (c). Adding class-specific textual tokens leads to improved localization maps (Figure 5 (d)) with expanded true object regions and reduced falsely activated backgrounds. Further incorporating sample-specific context produces the best localization maps (Figure 5 (e)) which are close to the ground-truth. This demonstrates the advantage of the proposed method in generating dense localization maps for objects of different scales in various challenging scenarios.

Effect of sample-specific context learning. As shown in Table 6, without sample-specific context learning, the learned multi-modal class-specific tokens result in the class-specific localization maps with an mIoU of 64.1%. Incorporating the visual context information aggregated by the output text tokens leads to an improved mIoU of 64.8%. To further learn image-language context to enhance the text token learning, we investigated two regularization methods: (**i**) Leveraging the pre-trained CLIP text embeddings of image captions. More specifically, we used a pre-trained image captioning model, ClipCap [27], to generate captions for each image. By minimizing the distance between the text

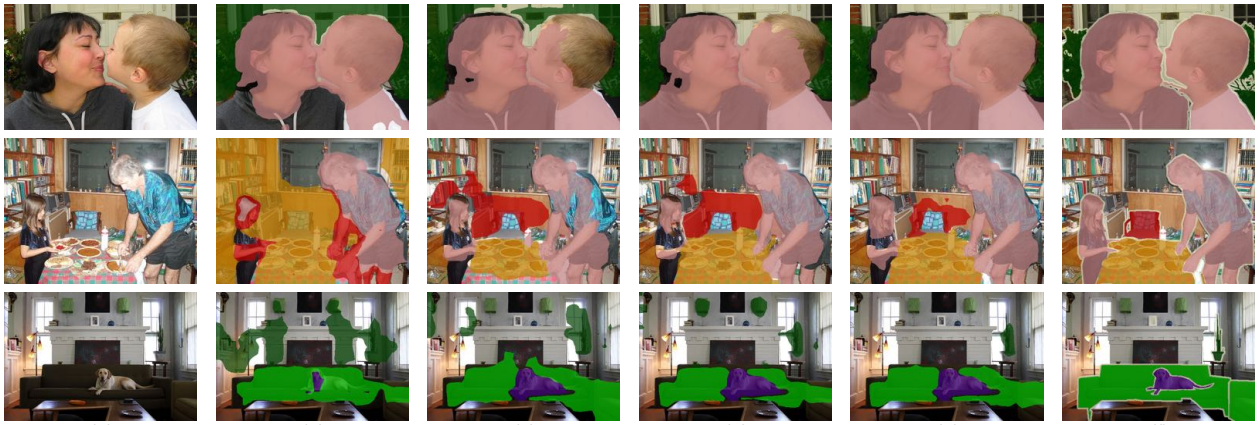

Figure 5. Visualization of the multi-label dense object localization results using different configurations of the proposed method on the PASCAL VOC 2012 *train* set. (a) Input; (b) MCTAttn (Multi-class token attention); (c) MCTAttn + C2P-CMap (class-to-patch correlation maps); (d) MCTAttn + MM-CMap (multi-modal correlation maps, *i.e.,* class-to-patch correlation maps + text-to-patch correlation maps); (e) MCTAttn + MM-CMap + sample-specific context; (f) Ground-truth.

Table 7. Evaluation of the class-specific dense localization maps by using different prompting methods for constructing class-specific textual tokens on the PASCAL VOC 2012 *train* set.

Prompt	mIoU
`a photo of [class]`	**66.3**
Prompt ensembling	65.5
CoOp (learning-based prompts) [49]	66.0

tokens and the pre-trained CLIP caption embeddings via L1 loss, the localization performance drops by 1.1%. Using the batch-softmax cross-entropy loss yields a slightly improved localization mIoU of 65.1%. L1 loss on the features allows to transfer all information of the generated image captions including useful and interfering contexts that could reduce the discriminative ability of the class-specific textual tokens, thus degrading the localization performance. In contrast, the batch-softmax cross-entropy loss on the feature similarities explicitly separates each text token from other text tokens in a same batch, which facilitates the sample-specific token learning for better localization. (**ii**) Leveraging the pre-trained CLIP image embeddings. By applying the batch-softmax cross-entropy loss on the similarity matrix of our text tokens and the pre-trained CLIP image embeddings, the class-specific localization attains a performance gain of 1.5% with the best mIoU of 66.3%. Compared to learning from the generated image captions, the batch-contrastive image-text matching loss enables the text tokens to better align with the visual features, thus leading to better dense localization results.

Effects of different prompts. We used different prompting methods to construct the class-specific textual tokens and evaluated their effects in the resulting class-specific dense localization maps. More specifically, we investigated three types of prompting methods including (**i**) the most commonly used prompt, *i.e.,* "a photo of [class]"; (**ii**) Prompt ensembling, which ensembles 8 best prompts over the embedding space via averaging as suggested in [28]; (**iii**) A learning-based prompting method, CoOp [49]. As shown in Table 7, these three prompting methods lead to comparable dense localization results. Among them, the general prompt template "a photo of [class]" produces the best localization mIoU. We speculate this is due to the general prompt template better preserving the class-discriminative language priors from the pre-trained CLIP model, compared to ensembling several prompts or learning extra parameters.

5. Conclusion

We proposed a new weakly supervised dense object localization method by explicitly constructing multi-modal class representations via a transformer-based framework. The proposed transformer includes class-specific visual and textual tokens by learning visual information from the target image data and exploiting language information from the pre-trained CLIP model, respectively. These diverse supervisions enable the multi-modal class-specific tokens to provide complementary information for more discriminative dense localization. By further incorporating the sample-specific contexts from visual context and image-language context, more adaptive multi-modal class-specific tokens can be learned to facilitate better dense localization. Our superior WSDOL results on two multi-label datasets, (*i.e.,* PASCAL VOC and MS COCO) and one single-label dataset, (*i.e.,* OpenImages), demonstrate the effectiveness of the proposed method. We also report state-of-the-art WSSS results on PASCAL VOC and MS COCO.

Acknowledgment. This research is supported in part by Australian Research Council Grant DP210101682, DP210102674, IH180100002, the Early Career Scheme of the Research Grants Council (RGC) of the Hong Kong SAR under grant No. 26202321 and HKUST Startup Fund No. R9253.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019. 6

[2] Wonho Bae, Junhyug Noh, and Gunhee Kim. Rethinking class activation mapping for weakly supervised object localization. In *ECCV*, 2020. 6

[3] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *CVPR*, 2022. 2, 7

[4] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. In *CVPR*, 2022. 2, 6

[5] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *CVPR*, 2020. 1, 6

[6] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *CVPR*, 2019. 6

[7] 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. In *ICLR*, 2021. 1

[8] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In *CVPR*, 2022. 2, 6, 7

[9] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep*, 2007:1–45, 2012. 6

[10] 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. In *ICCV*, 2021. 2, 6

[11] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 3

[12] Saurav Gupta, Sourav Lakhotia, Abhay Rawat, and Rahul Tallamraju. Vitol: Vision transformer for weakly supervised object localization. In *CVPR*, 2022. 2

[13] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011. 6

[14] 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 *CVPR*, 2022. 2, 7

[15] Jeesoo Kim, Junsuk Choe, Sangdoo Yun, and Nojun Kwak. Normalization matters in weakly supervised object localization. In *ICCV*, 2021. 6

[16] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016. 5

[17] 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 *ICCV*, 2021. 7

[18] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. *NeurIPS*, 2021. 6

[19] Jungbeom Lee, Eunji Kim, Jisoo Mok, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly supervised semantic segmentation and object localization. *PAMI*, 2022. 6, 7

[20] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *CVPR*, 2021. 2, 6

[21] Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. Weakly supervised semantic segmentation using out-of-distribution data. In *CVPR*, 2022. 1, 6, 7

[22] Minhyun Lee, Dongseob Kim, and Hyunjung Shim. Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds. In *CVPR*, 2022. 7

[23] 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 *CVPR*, 2021. 7

[24] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 3, 5

[25] Jing Li, Junsong Fan, and Zhaoxiang Zhang. Towards noiseless object contours for weakly supervised semantic segmentation. In *CVPR*, 2022. 7

[26] 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 *ECCV*, 2014. 6

[27] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 7

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 8

[29] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022. 3, 4, 5

[30] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *ECCV*, 2022. 3

[31] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017. 6

[32] Yukun Su, Ruizhou Sun, Guosheng Lin, and Qingyao Wu. Context decoupling augmentation for weakly supervised semantic segmentation. In *ICCV*, 2021. 2, 6, 7

[33] Kunyang Sun, Haoqing Shi, Zhengming Zhang, and Yongming Huang. Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps. In *ICCV*, 2021. 7

[34] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR*, 2020. 6

[35] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *CVPR*, 2018. 2

[36] 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 *CVPR*, 2021. 7

[37] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: Cross language image matching for weakly supervised semantic segmentation. In *CVPR*, 2022. 3, 6, 7

[38] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Rui-Wei Zhao, Tao Zhang, Xuequan Lu, and Shang Gao. Cream: Weakly supervised object localization via class re-activation mapping. In *CVPR*, 2022. 6

[39] 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 *ICCV*, 2021. 7

[40] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7

[41] 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 *CVPR*, 2021. 2, 7

[42] Sung-Hoon Yoon, Hyeokjun Kweon, Jegyeong Cho, Shinjeong Kim, and Kuk-Jin Yoon. Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. In *ECCV*, 2022. 7

[43] 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 *ICCV*, 2019. 6

[44] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In *NeurIPS*, 2020. 6

[45] Fei Zhang, Chaochen Gu, Chenyue Zhang, and Yuchao Dai. Complementary patch for weakly supervised semantic segmentation. In *ICCV*, 2021. 7

[46] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 6

[47] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *CVPR*, 2018. 6

[48] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 1, 6

[49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 8

[50] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *CVPR*, 2022. 7

[51] Lei Zhu, Qian Chen, Lujia Jin, Yunfei You, and Yanye Lu. Bagging regional classification activation maps for weakly supervised object localization. In *ECCV*, 2022. 6

[52] Lei Zhu, Qi She, Qian Chen, Yunfei You, Boyu Wang, and Yanye Lu. Weakly supervised object localization as domain adaption. In *CVPR*, 2022. 6