

eX-ViT: A Novel eXplainable Vision Transformer for Weakly Supervised Semantic Segmentation

Lu Yu^a, Wei Xiang^{b,*}, Juan Fang^c, Yi-Ping Phoebe Chen^b, Lianhua Chi^b

^a*College of Science and Engineering, James Cook University, Cairns, QLD 4878, Australia*

^b*School of Computing, Engineering and Mathematical Sciences, La Trobe University, Melbourne, VIC 3086, Australia*

^c*College of Information Technology, Beijing University of Technology, Beijing 100124, China*

Abstract

Recently vision transformer models have become prominent models for a range of vision tasks. These models, however, are usually opaque with weak feature interpretability, making their predictions inaccessible to the users. While there has been a surge of interest in the development of post-hoc solutions that explain transformer decisions, these methods can not be broadly applied to different transformer architectures, as rules for interpretability have to change accordingly based on the heterogeneity of data and model structures. Moreover, there is no method currently built for an intrinsically interpretable transformer, which is able to explain its reasoning process and provide a faithful explanation. To close these crucial gaps, we propose a novel vision transformer dubbed the eXplainable Vision Transformer (eX-ViT), an intrinsically interpretable transformer model that is able to jointly discover robust interpretable features and perform the prediction. Specifically, eX-ViT is composed of the Explainable Multi-Head Attention (E-MHA) module, the Attribute-guided Explainer (AttE) module and the self-supervised attribute-guided loss. The E-MHA tailors explainable attention weights that are able to learn semantically interpretable representations from local patches in terms of model decisions with noise robustness. Meanwhile, AttE is proposed to encode discriminative attribute features for the target object through diverse attribute discovery, which constitutes faithful evidence for the model's predictions. In addition, a self-supervised attribute-guided loss is developed for our eX-ViT, which aims at learning enhanced representations through the attribute discriminability mechanism and attribute diversity mechanism, to localize

*Corresponding author

Email addresses: lu.yu@my.jcu.edu.au (Lu Yu), w.xiang@latrobe.edu.au (Wei Xiang), fangjuan@bjut.edu.cn (Juan Fang), phoebe.chen@latrobe.edu.au (Yi-Ping Phoebe Chen), L.Chi@latrobe.edu.au (Lianhua Chi)

diverse and discriminative attributes and generate more robust explanations. As a result, we can uncover faithful and robust interpretations with diverse attributes through the proposed eX-ViT. To verify and evaluate our method, we apply the eX-ViT to several weakly supervised semantic segmentation (WSSS) tasks, since these tasks typically rely on accurate visual explanations to extract object localization maps. Particularly, the obtained explanation results obtained via eX-ViT are regarded as pseudo segmentation labels to train WSSS models. Comprehensive simulation results demonstrate that the proposed eX-ViT shows comparable performance to the supervised baselines, and outperforms the state-of-the-art black-box methods using only image-level labels in accuracy and interpretability.

Keywords: Explainable, Attention map, Transformer, Weakly supervised

1. Introduction

Over the last few years, transformer models have attracted increasing attention with encouraging results in a number of challenging domains, such as natural language processing, vision or graphs [1]. The Multi-Head Attention (MHA) and Multi-Layer Perceptron (MLP) modules in transformers effectively model global representations without convolution [2]. This framework is proved to be effective to capture long-range dependencies for image recognition. Despite their excellent performance, most transformer architectures are usually expressed as black boxes [3]. To be concrete, transformers can achieve impressive results but lack an explanation of the results due to the large number of parameters and the complex interactions between modules. Since transformers have high applicability to high-risk decision-making domains such as healthcare and autonomous driving, there is a strong necessity for gaining insights in model's decisions [4]. Besides, the great popularity of transformer models necessitates the explainable solutions to their inference process. An interpretable solution is able to aid in debugging the models and identifying crucial modules for downstream tasks [5].

Explainable Artificial Intelligence (XAI) is an emerging sub-field of AI pursuing to capture the properties that have influence over the decision of a model [6]. Depending on the phases where predictions and explanations are performed, these methods can be categorized into two types: intrinsically explainable models and post-hoc explanation methods. Unlike post-hoc models, explainable models aim to directly incorporate interpretability in the structure of the models, thus revealing the intrinsic reasoning process of the models. Several previous studies have pointed out

that explainable models outperform post-hoc methods in faithfulness and stability [7]. Unfortunately, little work has been done so far in the field of explainable transformers. Pan et al. [8] developed an interpretability-aware method to drop less informative patches for transformers to reduce computational cost. Caron et al. [9] proposed a self-supervised approach based on vision transformers and concluded that the attentions maps contain features about the semantic information of the image. However, the explicit expressive features were not explored to obtain faithful explanations w.r.t. model decisions.

In contrast to intrinsically explainable models, post-hoc explanation methods aim to obtain explanations by approximating a trained model and its predictions. For instance, Zhou et al. [10] introduced a technique called Class Activation Mapping (CAM) for identifying informative areas w.r.t. the model decisions by a certain type of CNNs which do not have fully-connected layers. Various studies have been made on post-hoc frameworks to uncover the predictions of the transformers. Brunner et al. [11] recently developed a method to visualize the extraction of the attention heads. Abnar et al. [12] attempted to build elaborate explanations across layers of the transformer. Hao et al. [5] developed a self-attention attribution mechanism to explain the information flow inside transformers. However, it is known to be unfaithful and unreliable since irrelevant tokens or less important attention heads often become highlighted. Another option is to employ the layer-wise relevance propagation (LRP) method to interpret transformer models. Chefer et al. [13] proposed a layer-wise relevance propagation (LRP) method to compute different relevance of the attention heads throughout the transformer. Voita et al. [14] used LRP to calculate relevance for each attention head in each transformer layer. However, the LRP mechanism was done in a limited way, in which different relevance rules are subject to the heterogeneity of different data and model structures [7]. Moreover, post-hoc methods cannot faithfully uncover the decision making process of the trained models, and the explanations can be easily impaired by different input schemes (e.g., perturbations or transformations). Therefore, post-hoc methods are considered to be more fragile, sensitive, and less faithful.

Against the above background, this paper aims to design the so-called eXplainable Vision Transformer (eX-ViT) with the inherent attribute of explainability and high performance. Specifically, the eX-ViT comprises the Explainable Multi-Head Attention (E-MHA) module, which can inherently provide interpretable attention maps that align with informative input patterns with noise robustness. Furthermore, the Attribute-guided Explainer (AttE) module is integrated into the eX-

ViT, to learn discriminative attribute features for the target object. Intuitively, we assume each object is made up of several attributes, which could be basic elements including color, shape, and texture, or higher-level local features such as body parts. Our key idea is to decompose the feature representation into a set of learnable attribute features for the target object, capable of capturing diverse and discriminative object features. Besides, a novel attribute-guided loss is designed to promote the learning process inside AttE in a self-supervised manner. More precisely, this loss implicitly adds the regularization to force the representations to focus on various attributes of each target class through the attribute discriminability mechanism and attribute diversity mechanism. We then verify and evaluate our method on several weakly supervised semantic segmentation (WSSS) tasks, as they typically rely on accurate visual explanations to extract object localization maps. Particularly, the obtained interpretation results obtained via eX-ViT are regarded as pseudo segmentation labels to train WSSS models. To the best of our knowledge, this is the first work to develop an intrinsically explainable vision transformer for WSSS tasks. In summary, the major contributions of this paper are:

- We propose a novel eXplainable Vision Transformer (eX-ViT), which provides faithful and robust explanations with model-inherent interpretability. Specifically, the proposed eX-ViT is able to provide explainable representations with comparable or better performance than state-of-the-art transformers (e.g., Swin Transformer [15] and PVT [16]);
- We propose a novel Explainable Multi-Head Attention (E-MHA) module, which, as a basic building block of the eX-ViT, has two key attributes. That is, it provides model-inherent explainable attention maps that align with the informative input patterns and is robust to noise;
- We propose the Attribute-guided Explainer (AttE) module, which is integrated into the eX-ViT to recognize diverse and discriminative object attribute features for the target object with only image-level labels through diverse attribute discovery;
- We propose the attribute-guided loss function, which enables the self-supervised learning in the proposed eX-ViT, capable of not only learning explanations that are faithful to the model’s predictions, but also resulting in more robust feature representations across data transformations;

- Comprehensive simulation results demonstrate that the proposed eX-ViT is comparable to the supervised baselines, and outperforms the state-of-the-art transformers in accuracy and interpretability.

The remainder of the paper is organized as follows. Section 2 briefly describes some recent related works on vision transformers, XAI techniques for transformers, and weakly supervised semantic segmentation methodologies. Section 3 presents the explainable architecture, i.e., eX-ViT, and introduces its main modules. Experimental results and discussions are presented in Section 4, followed by concluding remarks drawn in Section 5.

2. Related work

2.1. Transformers for vision

Transformer-based models have recently been introduced to vision tasks and achieved remarkable progress. One of purely transformer-based models is the ViT [2], which has exhibited impressive performance without convolution. However, the ViT is inferior to CNNs when trained on a mid-sized dataset. DeiT [17] and T2T-ViT [18] addressed this issue by designing efficient transformer-based backbones. Liu et al. [15] proposed the Swin Transformer to compute the local attention within a window for object detection tasks. Wang et al. [16] designed the PVT, which is another architecture that has a feature hierarchy with multiple stages and multiple resolutions for various vision tasks. Moreover, there are some research studies with modified ViT architectures that benefit downstream vision tasks such as semantic segmentation. However, most of existing designs focus on efficient and effective frameworks for downstream tasks without considering interpretability. Thus these methods tend to be less faithful to the users. Recently, Wei et al. [19] proposed a transformer-based framework that generates semantic-aware CAMs for the weakly supervised object localization task. However, TS-CAM only leverages the class-agnostic feature maps and ignores the intrinsic semantic attributes for the target class. In this paper, we aim to address these issues by proposing the so-called eX-ViT, which exploits explainable features that are robust to noise and provides faithful explanations.

2.2. Explainable artificial intelligence (XAI) for transformers

There are mainly two sub-fields of explainable techniques: intrinsically explainable models and post-hoc explanation methods. Unlike post-hoc models, intrinsically explainable models aim to

directly incorporate interpretability in the structure of the models, thus revealing the intrinsic reasoning process of the models. Several previous studies have pointed out that explainable models outperform post-hoc methods in faithfulness and stability [7]. Unfortunately, little work has been done so far in the field of explainable transformers. Pan et al. [8] proposed an interpretability-aware method to drop less informative patches for transformers to reduce computational cost. However, the obtained interpretations were not complete due to the information loss. Recently, Caron et al. [9] utilized a self-supervised approach called DINO based on vision transformers and concluded that the attentions maps contain features about the semantic information of the image. But the expressive features were not explored to obtain faithful explanations. Different from the majority of previous studies, we attempt to build the first explainable transformer architecture with the objective of learning interpretable features.

In terms of post-hoc explanation approaches, there are a variety of recent studies that explore the explainability for transformers. Chefer et al. [13] proposed a layer-wise relevance propagation (LRP) method by introducing a relevancy propagation rule that is applicable to both positive and negative contributions. This approach, however, is not able to provide the interpretation for attention modules besides self-attention. Voita et al. [20] found that only a small portion of heads in transformers have a salient impact on the model decisions. Jawahar et al. [21] proved that syntactic patterns are learned in lower blocks, while semantic patterns are gained through higher layers. Abnar et al. [12] proposed to combine the attention scores across multiple layers, but this method fails to distinguish between positive and negative attributions. Raghu et al. [22] analyzed the internal representations of vision transformers, and found out that they tend to learn more uniform representations across all layers. Recently, Chefer et al. [23] also proposed a generic approach to explain transformers including bi-modal ones. However, most of the existing post-hoc methods tend to be fragile, sensitive, and less faithful. Since they cannot faithfully uncover the decision making process of the trained models, and the explanations can be easily impaired by different input schemes (e.g., perturbations or transformations).

2.3. Weakly supervised semantic segmentation

Compared to supervised learning methods, WSSS aims at training models with weak labels such as bounding boxes and image-level labels. As the cornerstone of WSSS, most of existing designs in terms of WSSS tasks use the Class Activation Mapping (CAM) to extract object localization

maps and approximate the segmentation mask. Despite the encouraging results, CAM suffers from the issue with incomplete object activation [1]. To address this drawback, several approaches have been proposed as the CAM expansions to remove the most discriminative parts of CAM and discover more complete object localization maps. Ahn et al. [24] utilized the AffinityNet model to learn the semantic affinity between pixels, then propagated the CAM activations several times to revise CAMs. Yao et al. [25] accumulated the activated regions to discover the objects in the non-salient regions. Li et al. [26] proposed a group-wise learning task to discover similar semantic regions, forming more reliable pseudo ground-truth masks. Papandreou et al. [27] adopted an Expectation-Maximization architecture to learn pseudo segmentation masks. Peng et al. [28] provided a method to train a CNN with discrete constraints and regularization priors for WSSS learning with favorable performance. The above methods are commonly based on CNNs, which reveals the inherent drawback of convolution. Xu et al. [1] introduced transformer attentions to learn class-specific localization maps. Ru et al. [3] adopted the semantic affinity in self-attentions in transformers to produce more integral pseudo labels for WSSS. However, there is still a large gap between fully supervised semantic segmentation and previous transformer-based WSSS methods. In our work, we propose a transformer-based model to extract explanation features to localize class-specific feature maps. We attempt to build a novel transformer architecture with the objective of learning interpretable representations in a self-supervised manner to narrow the supervision gap.

3. Method

This section details our proposed network architecture, i.e., the eX-ViT. First, we introduce the overall architecture. Second, we describe the intuition and design of the E-MHA and discuss several important properties of the E-MHA. Furthermore, The AttE is proposed to integrate into our eX-ViT to decompose the attention maps into features of attributes through diverse attribute discovery, and a self-supervised attribute-guided loss is adopted to learn robust semantic representations via the attribute diversity mechanism and attribute discriminability mechanism, which constitutes faithful evidence for the model’s predictions.

3.1. Architecture of the eX-ViT

The overall architecture of our proposed eX-ViT is depicted in Fig. 1. In particular, the eX-ViT is a siamese network, which comprises two branches for a pair of input images (two data

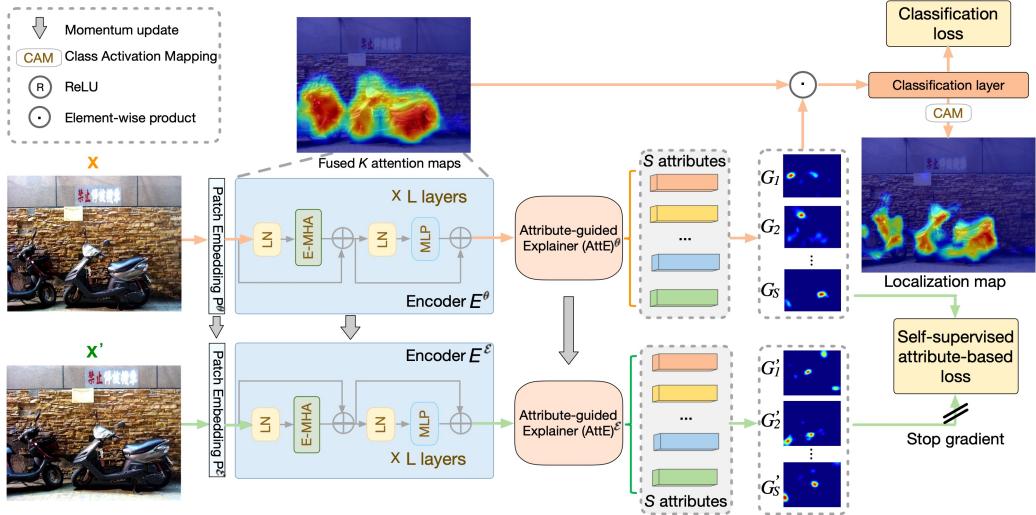


Figure 1: Illustration of the proposed eXplainable Vision Transformer (eX-ViT) architecture. \mathbf{x} and \mathbf{x}' are two different random transformations of an input image. We use a transformer backbone as the encoder to extract feature maps, the backbone contains consecutive L encoding layers with Explainable Multi-Head Attention (E-MHA) as the attention block. θ is the trainable module, while \mathcal{E} is an exponential moving average of θ . The Attribute-guided Explainer (AttE) is proposed atop of the encoder to decompose the attention maps into features of attributes through diverse attribute discovery, so as to facilitate the generation of more faithful and robust interpretations. We also design a self-supervised attribute-guided loss function for our eX-ViT, which aims at learning robust semantic representations via the attribute diversity mechanism and attribute discriminability mechanism.

augmentations from an original input) to learn interpretable attention maps in a self-supervised manner. Each branch comprises a transformer encoder with L transformer layers consisting of the novel Explainable Multi-Head Attention (E-MHA) module, and the Attribute-guided Explainer (AttE) module atop of it. Following MoCo [29], the parameters \mathcal{E} in the lower branch uses a momentum update with the upper θ . Empirically, the proposed architecture can conveniently replace the backbone networks in existing methods for WSSS tasks.

3.2. Explainable Multi-Head Attention (E-MHA)

In this section, we introduce our novel Explainable Multi-Head Attention (E-MHA) module as shown in Fig. 2, which consists of H parallel heads. Specifically, given an input feature map $\mathbf{X} \in \mathbb{R}^{T \times d}$ where T is the spatial size and d is the feature dimension, each head H_h holds an explainable attention weight $\mathbf{A}_h \in \mathbb{R}^{N \times d}$ (N is the spatial size of \mathbf{A}_h) that represents relative

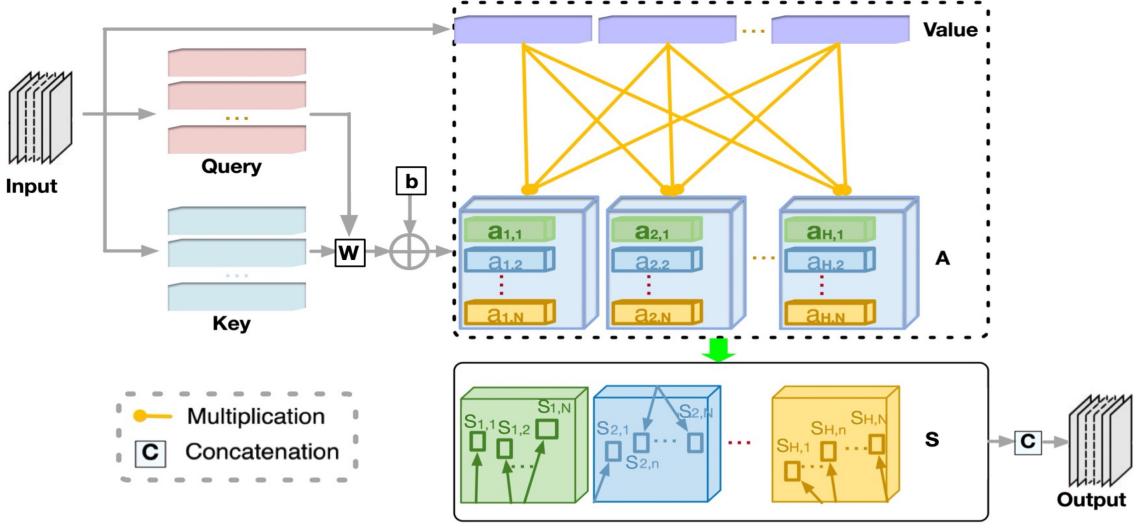


Figure 2: The architecture of Explainable Multi-Head Attention (E-MHA). Compared to MHA, the E-MHA dynamically aligns attention weights with the most discriminative input patterns during optimization, so as to be explainable and robust to noise.

importance of input features. That is, \mathbf{A}_h aims to learn explainable features for the output through the proposed E-MHA module.

In particular, we structure this section around two crucial attributes of the E-MHA module:

Noise robustness: The E-MHA is computed as a dynamic alignment between the input tokens and the attention weight. When the module is optimized, the attention weight is driven to focus on the most discriminative and class-related patterns from the input tokens. Instead of directly removing the irrelevant noises from the input image, we adopt a dynamic alignment mechanism in E-MHA to extract discriminative features from the input, thus reducing the noise information gradually and then preserving the key input patterns. This favorable attribute is empirically verified in section Section 4.3.1.

Inherent explainability: Given input \mathbf{X} , the E-MHA aims to learn the attention weight that maximises the alignment between input tokens and the attention weight. During the optimization process, maximising this alignment means encoding the projected input values as eigenvectors of the attention weight. As a result of this property, the model-inherent attention weight is learned with the discriminative input patterns and thus directly used to explain model decisions without needing any external tools.

First, given input \mathbf{X} , the projected key, query and value are obtained as follows

$$\mathbf{Q} = \mathbf{X}\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^V, \quad (1)$$

where $\mathbf{W}^Q \in \mathbb{R}^{d \times d}$, $\mathbf{W}^K \in \mathbb{R}^{d \times d}$, and $\mathbf{W}^V \in \mathbb{R}^{d \times d}$ are trainable transform matrices.

Second, the self-attention operation is constructed by

$$\mathbf{W} = \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}, \quad (2)$$

where the obtained matrix \mathbf{W} implies how much attention is paid on each token.

Third, the attention weight \mathbf{A} is defined as follows

$$\mathbf{A} = \ell(\mathbf{W} + \mathbf{b})^T, \quad (3)$$

where \mathbf{b} is a trainable bias term, which is introduced as an initial alignment for the input patterns. $\ell(\cdot)$ is a non-linear function that scales the L2 norm of its input, i.e., $\ell(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$ and $\|\ell(\mathbf{x})\| \leq 1$.

In what follows, the self-attention feature \mathbf{S} is formally expressed as

$$\mathbf{S} = \mathbf{A}^T \mathbf{V}, \quad (4)$$

According to Eq. (3), $\|\mathbf{A}\| \leq 1$. Therefore \mathbf{S} in Eq. (4) is upper-bounded as follows

$$\mathbf{S} = \|\mathbf{A}\| \|\mathbf{V}\| \cos(\mathbf{A}, \mathbf{V}) \leq \|\mathbf{V}\|. \quad (5)$$

When Eq. (5) is optimized, the attention weight \mathbf{A} is proportional to \mathbf{V} . In order to achieve maximal output, \mathbf{A} is driven to align with the discriminative features in \mathbf{V} , instead of the uninformative noise. Therefore, \mathbf{S} can only achieve this upper bound if all possible solutions of $\mathbf{v} \in \mathbf{V}$ are encoded as eigenvectors in the weight \mathbf{A} . This maximisation suggests with the attention weight \mathbf{A} , we will obtain an inherently explainable decomposition of input patterns.

Overall, the computation in layer l is expressed as

$$\begin{aligned} \mathbf{S}^l &= \text{E} - \text{MHA}(\text{LN}(\mathbf{F}^{l-1})), \\ \mathbf{Z}^l &= \mathbf{S}^l + \mathbf{F}^{l-1}, \\ \mathbf{F}^l &= \text{MLP}(\text{LN}(\mathbf{Z}^l)) + \mathbf{Z}^l, \end{aligned} \quad (6)$$

where $\text{LN}(\cdot)$ is the LayerNorm layer, $\text{MLP}(\cdot)$ denotes the multi-layer perceptron layer, and \mathbf{F}^l is the output of layer l .

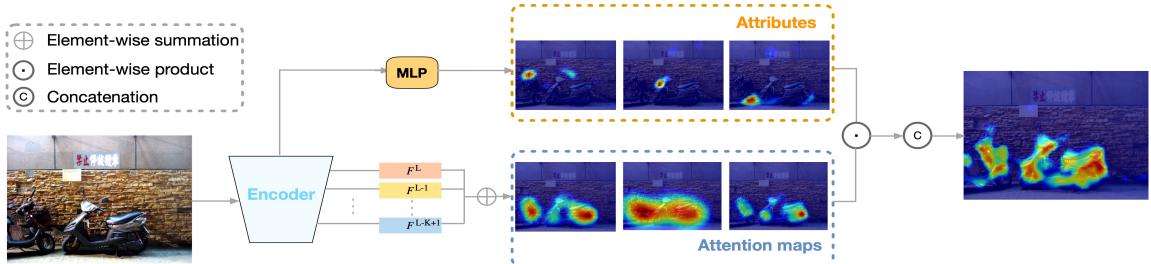


Figure 3: Illustration of Attribute-guided Explainer (AttE). We aggregate the interpretable attention maps from the last K transformer layers to generate a fused attention map with good precision on the complete object context information. The attribute features are regarded as the complement information to better guide the localization of the object context, thus producing robust attribute features in a weakly supervised manner.

Our key motivation of E-MHA is to dynamically align its attention weights with the discriminative patterns from input values while reducing the impact of noise. The cascade transformer layers in the encoder enable the model to suppress the noise information gradually and learn discriminative input patterns. As a result, the model is able to discover robust representations from the input image. With the attributes of noise robustness and inherent explainability, E-MHA produces the transformer attention map which inherently provides an explainable combination of contributions from discriminative input patterns w.r.t. the model predictions.

3.3. Attribute-guided Explainer (AttE)

Although the proposed E-MHA provides the intuitive process for explainable feature learning, it is non-trivial to obtain intrinsically interpretable representations that benefit the WSSS tasks without additional regularization. Inspired by the pixel-wise prediction scheme used in semantic segmentation frameworks to localize objects [30], we propose the Attribute-guided Explainer (AttE) module for our eX-ViT with the objective of decomposing the attention map into attribute features based on the diverse attribute discovery. By which, the learned feature maps can be viewed as a set of attributes at a granular level that capture more complete object information.

Given that the transformer structure tends to learn more uniform representations across all layers, we propose to utilize the transformer attention maps from the last layer in eX-ViT's encoder, to learn a set of trainable attribute features. Concretely, to model the context attention, given an input feature map $\mathbf{F} \in \mathbb{R}^{H \times W \times d}$ produced by the encoder E^θ 's last layer, we first calculate a set

of spatial feature maps that capture the relative importance of all HW locations as follows

$$\mathbf{W}_{i,j} = A_\phi(\mathbf{F}), \quad \forall \{i, j\} \in H \times W, \quad (7)$$

where $A_\phi(\cdot)$ is implemented by a 2-layer MLP block, with one hidden layer followed by a LN layer and the ReLU activation layer. $\mathbf{W}_{i,j} \in \mathbb{R}^{c \times H \times W}$ is the obtained embedding with the hidden dimension c . We will investigate the influence of c on the model performance in Section 4.3.4.

Furthermore, we apply the ℓ_2 -norm function to $\mathbf{W}_{i,j}$ along the channel dimension, which is formally expressed as

$$\overline{\mathbf{W}}_{i,j} = \frac{\mathbf{W}_{i,j}}{\|\mathbf{W}_{i,j}\|_2}, \quad (8)$$

where $\|\cdot\|_2$ denotes the L2 norm, $\overline{\mathbf{W}}_{i,j}$ is the normalized representation indicating which spatial features to emphasize or suppress.

Subsequently, $\overline{\mathbf{W}}$ is sliced into S groups on the channel dimension, i.e., $(\overline{\mathbf{W}}_1, \overline{\mathbf{W}}_2, \dots, \overline{\mathbf{W}}_S)$, where $\overline{\mathbf{W}}_s \in \mathbb{R}^{\frac{c}{S} \times H \times W}$ stands for the feature map of the s -th attribute, S is the total number of attributes. To this end, we can apply $\overline{\mathbf{W}}_s$ of attribute s to the feature \mathbf{F} by

$$\mathbf{G}_s = \overline{\mathbf{W}}_s \odot \mathbf{F}, \quad (9)$$

where \odot is the element-wise product, $\mathbf{G} = [\mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_S]$ is the final output with attribute features. By this means, each feature map \mathbf{F} is projected into S attribute representations that explicitly reveal which pixels are related to the attribute s . Likewise, we follow the same procedure described from Eq. (7) to Eq. (9), the attribute representation \mathbf{G}' of the second augmented input can be generated accordingly with the momentum encoder $E^\mathcal{E}$. And our $AI^\mathcal{E}$ is also the exponential moving average of the trained AI^θ .

In summary, the output of AttE can be seen as the decomposed contributions for individual attributes. By this means, our model is able to encode semantically explainable features for the target object in an explicit manner, which facilitates the learning of complete object context information. Moreover, we elaborately design our attribute-guided loss function to guide the learning of AttE, which will be presented in next subsection.

3.4. Attribute-guided loss function

A challenging problem for typical vision transformers is that they are not intrinsically interpretable due to lack of the representational power. A recent work, DINO [9] utilized a self-supervised

approach based on vision transformers and concluded that the attention maps contain features about the semantic information of the image. However, the semantic attributes of the target object were not explored to obtain faithful explanations. In our work, we propose to improve model's interpretability by regularizing its representations with the attribute-guided loss function, i.e., the global-level attribute-guided loss $\mathcal{L}_{\text{global}}$, the local-level attribute discriminability loss \mathcal{L}_{dis} loss and the attribute diversity loss \mathcal{L}_{div} . On one hand, the $\mathcal{L}_{\text{global}}$ encourages the predicted attribute features to approximate the target object, which ensures the faithfulness of the global representations. On the other hand, the \mathcal{L}_{dis} and \mathcal{L}_{div} aim to localize fine-grained attributes through the attribute discriminability mechanism and attribute diversity mechanism, thus enabling the robust feature learning.

Since higher layers discover high-level concepts like objects or scenes, we propose to fuse transformer attention maps from the last K encoder layers to achieve good accuracy on the complete object context information. Hence, given the obtained attention map \mathbf{F}^l in l -th encoder layer, the fused attention map is expressed as

$$\hat{\mathbf{F}} = \frac{1}{K} \sum_l^K \mathbf{F}^l, \quad (10)$$

where $\hat{\mathbf{F}}$ is the fused transformer attention map. By this means, we aggregate cascaded encoder blocks to produce reliable attention map for the complete object localization. As the aggregated attention map $\hat{\mathbf{F}}$ is attribute-agnostic, we propose to couple it with the attribute features \mathbf{G} to generate the attribute-guided attention map. The process is defined as follows

$$\mathbf{M} = \hat{\mathbf{F}} \odot \mathbf{G}, \quad (11)$$

where \mathbf{M} represents the final output of the attribute-guided feature map.

Based upon \mathbf{M} , the global-level attribute-guided loss $\mathcal{L}_{\text{global}}$ is computed by the multi-label soft margin loss

$$\mathcal{L}_{\text{global}} = \frac{1}{C} \sum_{c=1}^C (y^c \log(\hat{y}_c) + (1 - y^c) \log(1 - \hat{y}_c)), \quad (12)$$

where the prediction \hat{y}_c is obtained by feeding the feature map \mathbf{M} into a classification layer. By optimizing the $\mathcal{L}_{\text{global}}$, the interpretable features are gathered as a summation of the important scores of all attribute features, which ensures the faithfulness of the explanations.

In addition, in order to improve the ability of network to learn diverse and discriminative attribute representations for the target object, we propose the local-level attribute-guided loss through

the attribute discriminability mechanism and attribute diversity mechanism in a self-supervised manner. Intuitively, the attribute discriminability mechanism aims to make attribute features consistent discriminative between two types of input views, while the attribute diversity mechanism enables the model to learn the effective decomposition with the attribute diversity. Formally, the attribute discriminability loss \mathcal{L}_{dis} is defined by

$$\begin{aligned}\mathcal{L}_{\text{dis}} &= |d - \sum_{s=1}^S d_s|, \\ d &= \ell(g(\mathbf{G}), g(\mathbf{G}')), \\ d_s &= \ell(g(\mathbf{G}_s), g(\mathbf{G}'_s)),\end{aligned}\tag{13}$$

where $g(\cdot)$ is the generalized mean pooling. And we adopt the normalized Mean Square Error in BYOL [31] as the $\ell(\cdot)$ function to calculate distance between two features. As can be seen from Eq. (13), d is leveraged to minimize the difference between attribute features, while d_s is used to guarantee the consistency between \mathbf{G} and \mathbf{G}' for each individual attribute. Empirically, this attribute discriminability loss function \mathcal{L}_{dis} is able to facilitate the model to discover discriminative class-specific attributes and obtain more comprehensive localization maps. Meanwhile, we introduce the attribute diversity loss \mathcal{L}_{div} is formally defined by

$$\mathcal{L}_{\text{div}} = \frac{1}{S(S-1)} \sum_{i=1, j=1}^S \sum_{i \neq j}^S \frac{\langle \mathbf{G}_i, \mathbf{G}_j \rangle}{\|\mathbf{G}_i\|_2 \|\mathbf{G}_j\|_2},\tag{14}$$

The intuition behind the \mathcal{L}_{div} is to make attribute features to the maximally independent from each other, so as to make attribute features focus on different discriminative object regions.

Overall, the loss function for the proposed eX-ViT is given below

$$\mathcal{L} = \mathcal{L}_{\text{global}} + \alpha \mathcal{L}_{\text{dis}} + \beta \mathcal{L}_{\text{div}},\tag{15}$$

where $\mathcal{L}_{\text{global}}$ is the multi-label soft margin loss. α and β are the coefficient of \mathcal{L}_{dis} and \mathcal{L}_{div} , respectively.

As a result, our attribute-guided loss promotes the learning of attribute features. The global-level loss $\mathcal{L}_{\text{global}}$ ensures a faithful transformer model, while the \mathcal{L}_{dis} and \mathcal{L}_{div} enable discriminative and robust attribute features. The effectiveness of the loss function is further verified in the experimental section.

4. Experiments

In this section, we first introduce the experimental settings including datasets and implementation details. Second, we evaluate the efficiency of our proposed eX-ViT and compare it with the recent state-of-the-art methods. Third, we conduct a series of ablation studies to discover the performance contribution from different modules in our framework.

4.1. Setup

4.1.1. Datasets

We conduct experiments on PASCAL VOC 2012 dataset [32] and MS COCO 2014 dataset [33]. PASCAL VOC 2012 dataset includes 20 object classes and one background class for the semantic segmentation task. Following the common experimental configuration from SBD [34], we adopt the augmented dataset which contains three subsets, training, validation, and testing sets, each having 10582, 1449, and 1464 images, respectively. MS COCO 2014 dataset uses 81 classes, its training and validation sets have 82081 images and 40137 images respectively. Note image-level labels are only used during training and ground-truth bounding box annotations are solely used during test. In line with previous works [3], we report the mean Intersection-over-Union (mIoU) to evaluate the performance of our proposed model.

4.1.2. Implementation details

We use PyTorch for implementation and conduct experiments. The encoder parameters are pre-trained on ImageNet [35], while other parameters are initialized as suggested in [36]. During training, we use the AdamW optimizer. For the transformer encoder E^θ , the initial learning rate is set to be 5×10^{-5} , which is further decayed via a polynomial schedule. The learning rate for the rest of the parameters is 5×10^{-4} . For the training on the PASCAL VOC 2012 dataset, the batch size is set as 16, and the training process lasts 40k iterations. On MS COCO 2014 dataset, we trained the models for 80k iterations with a batch size of 8. For data augmentation, we used random scaling with a range of [0.5,2.0], random horizontally flipping, and random cropping.

The default hyper-parameters are set as follows. For encoders E^θ and $E^\mathcal{E}$, it contains 12 layers, 6 heads within each E-MHA, and the hidden dimension is set to 384. Empirically, we set α and β in Eq. (15) as 0.5 and 1.0 respectively throughout this paper. In line with previous works, we use the ResNet38 [37] as the backbone for semantic segmentation. At test time, since our eX-ViT is a

Table 1: mIoU (%) of localization maps on the PASCAL VOC 2012 training set.

Method	Local. Maps	+denseCRF
(CVPR'20) SCE [39]	50.9	55.3
(CVPR'20) SEAM [38]	55.4	56.8
(CVPR'21) EDAM [40]	52.8	58.2
(CVPR'21) AdvCAM [41]	55.6	62.1
(ICCV'21) ECS-Net [42]	56.6	58.6
(ICCV'21) CSE [43]	56.0	62.8
(CVPR'22) SIPE [4]	58.6	64.7
(CVPR'22) ReCAM [44]	56.6	-
(Ours) eX-ViT	59.1	65.3

siamese network, only one branch is needed. Following the common practice in prior studies [38], we use multi-scale testing and CRFs to obtain pseudo segmentation results.

4.2. Comparison with state-of-the-arts

4.2.1. Comparison on localization maps

We first evaluate the qualitative results of mIoU on localization maps, where those pseudo semantic segmentation results are obtained by applying PSA [24] on the localization maps produced by our proposed eX-ViT. Table 1 reports the results of our proposed method as well as other recent state-of-the-art approaches on the PASCAL VOC 2012 training set. As can be seen from the table, SIPE [4] achieves the state-of-the-art result with a mIoU of 58.6%. eX-ViT outperforms all compared methods in terms of both metrics. Concretely, the results show that our eX-ViT improves the mIoU to 59.1%. We also conduct experiments based on eX-ViT with denseCRF post-processing, and the gain becomes up to 65.3%. Fig. 4 shows visual comparison of object localization maps on PASCAL VOC 2012 training set. As shown in Fig. 4, the fused class-specific attribute-guided localization map can effectively capture the discriminative features within the object context of the target class with more useful clues. As a result, the fused localization map by use of our eX-ViT brings notable visual improvements to produce complete localization maps.

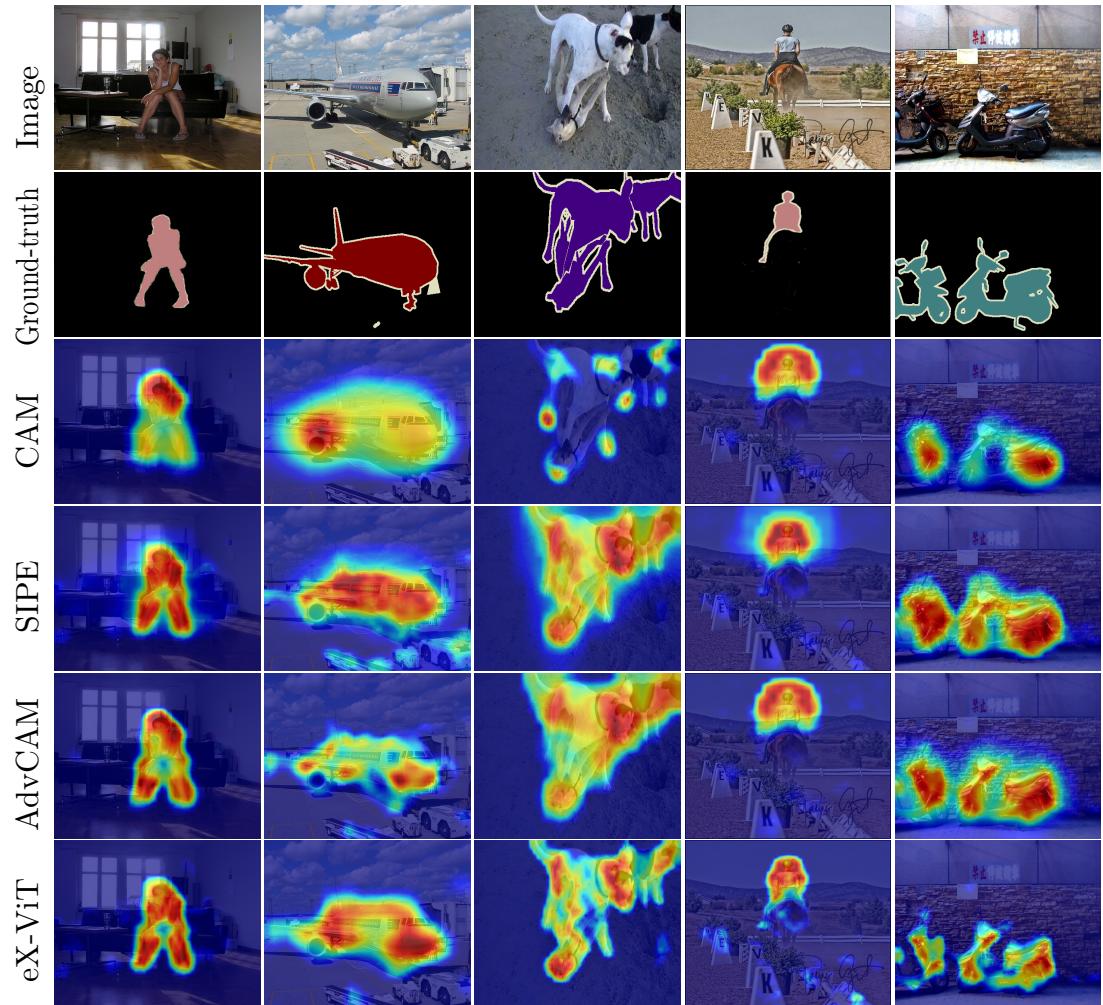


Figure 4: Visual comparison of localization maps generated by different methods on PASCAL VOC 2012 training set. From top to down: original image, ground-truth, CAM [10], SIPE [4], AdvCAM [41] and our eX-ViT.

Table 2: Performance comparison of state-of-the-art methods in terms of mIoU (%) on the PASCAL VOC 2012 validation and test sets. *Sup.* indicates supervision type. \mathcal{F} : full supervision; \mathcal{I} : image-level labels; \mathcal{S} : saliency maps.

Method	<i>Sup.</i>	Backbone	validation	test
Fully-supervised methods				
(TPAMI'18) DeepLab [45]	\mathcal{F}	ResNet101	77.6	79.7
(PR'19) WideResNet38 [37]	\mathcal{F}	WR38	80.8	82.5
(NeurIPS'21) Segformer [46]	\mathcal{F}	MiT-B1	78.7	-
Weakly-supervised methods				
(ECCV'20) MCIS [47]	$\mathcal{I} + \mathcal{S}$	ResNet101	66.2	66.9
(ICCV'21) AuxSegNet [48]	$\mathcal{I} + \mathcal{S}$	ResNet38	69.0	68.6
(CVPR'21) NSROM [25]	$\mathcal{I} + \mathcal{S}$	ResNet101	70.4	70.2
(CVPR'21) EPS [49]	$\mathcal{I} + \mathcal{S}$	ResNet101	70.9	70.8
(CVPR'21) EDAM [40]	$\mathcal{I} + \mathcal{S}$	ResNet101	70.9	70.6
(ICCV'19) OAA+ [50]	$\mathcal{I} + \mathcal{S}$	ResNet101	65.2	66.4
(TPAMI'21) A2GNN [51]	\mathcal{I}	ResNet101	66.8	67.4
(TPAMI'22) LIID [52]	\mathcal{I}	ResNet101	66.5	67.5
(CVPR'20) SEAM [38]	\mathcal{I}	ResNet38	64.5	65.7
(AAAI'20) RRM [53]	\mathcal{I}	ResNet38	62.6	62.9
(NeurIPS'20) CONTA [54]	\mathcal{I}	ResNet38	66.1	66.7
(ICCV'21) CDA [55]	\mathcal{I}	ResNet38	66.1	66.8
(ICCV'21) ECS-Net [42]	\mathcal{I}	ResNet38	66.6	67.6
(ICCV'21) CSE [43]	\mathcal{I}	ResNet38	68.4	68.2
(CVPR'21) AdvCAM [41]	\mathcal{I}	ResNet101	68.1	68.0
(NeurIPS'21) RIB [56]	\mathcal{I}	ResNet101	68.3	68.6
(CVPR'22) SIPE [4]	\mathcal{I}	ResNet101	68.8	69.7
(CVPR'22) ReCAM [44]	\mathcal{I}	ResNet101	68.5	68.4
(CVPR'22) Ru et al. [3]	\mathcal{I}	MiT-B1	66.0	66.3
(Ours) eX-ViT	\mathcal{I}	ResNet38	71.2	71.1



Figure 5: Qualitative segmentation results on the validation set of PASCAL VOC 2012. From top to down: original image, ground-truth, SIPE [4] and our eX-ViT.

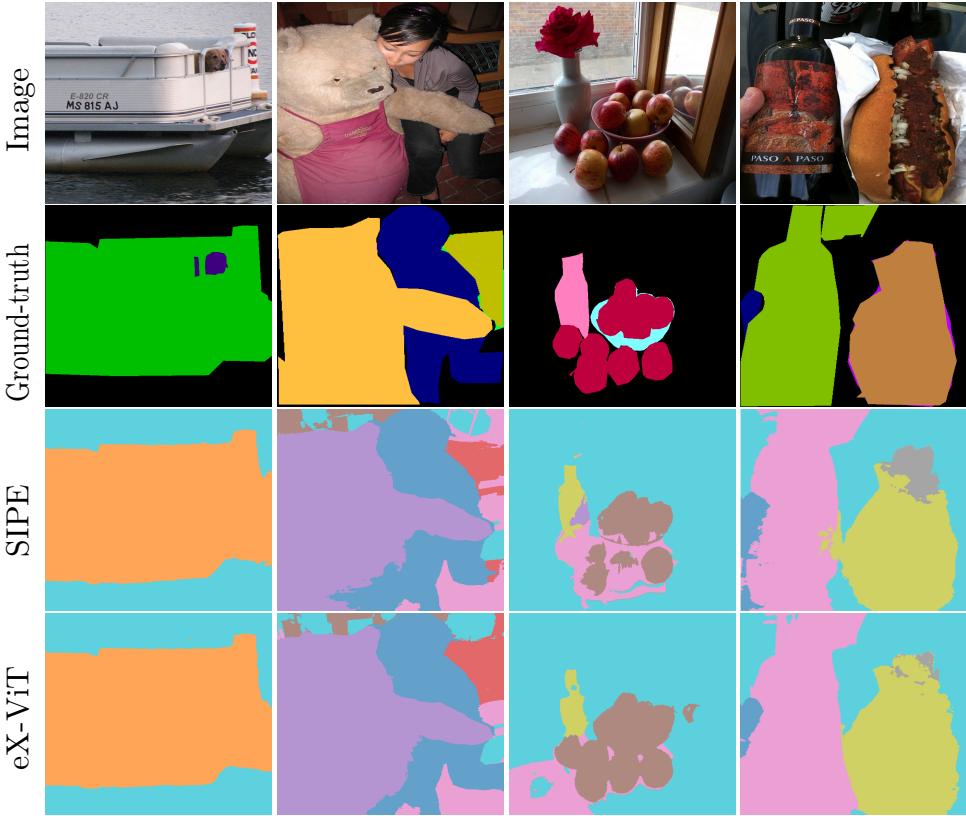


Figure 6: Qualitative segmentation results on the validation set of MS COCO 2014. From top to down: original image, ground-truth, SIPE [4] and our eX-ViT.

4.2.2. Comparison on segmentation results

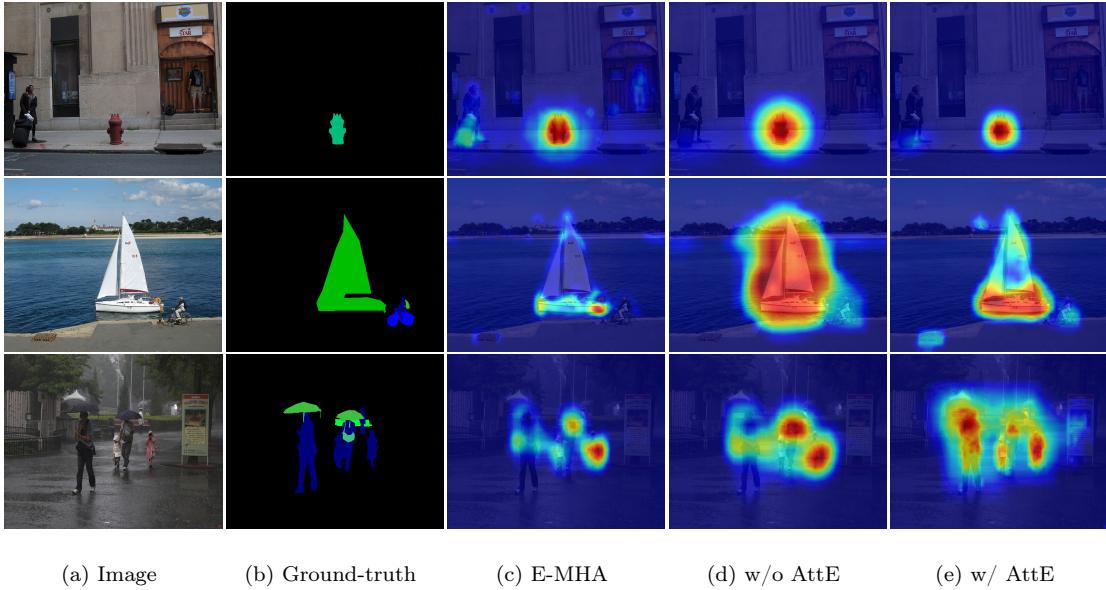
The comparison results among the fully-supervised and weakly supervised state-of-the-art methods on PASCAL VOC 2012 validation and test sets are reported in Table 2. We use the generated pseudo semantic segmentation labels to train DeepLabV2 [45]. Among the compared methods, the eX-ViT is able to remarkably improve the segmentation performance using only image-level labels on the validation and test sets, respectively. As can be observed, compared to the fully-supervised methods, the eX-ViT is able to obtain comparable performance with 71.2% mIoU on validation set and 71.1% mIoU on test set. Compared with the recent state-of-the-art weakly supervised models, e.g., EPS [49] and EDAM [40] which use both additional saliency maps and image-level labels as supervision signals, eX-ViT still shows superior performance. The qualitative segmentation results on the validation set are shown in Fig. 5. Based on our model, DeepLabV2 can produce accurate

Table 3: Performance comparison of state-of-the-art WSSS methods in terms of mIoU (%) on the MS COCO 2014 validation set. *Sup.* indicates supervision type. \mathcal{I} : image-level labels; \mathcal{S} : saliency maps.

Method	<i>Sup.</i>	Backbone	mIoU (%)
(CVPR’18) DSRG [57]	$\mathcal{I} + \mathcal{S}$	VGG16	26.0
(ICCV’21) AuxSegNet [48]	$\mathcal{I} + \mathcal{S}$	ResNet38	33.9
(CVPR’21) EPS [49]	$\mathcal{I} + \mathcal{S}$	ResNet101	35.7
(NeurIPS’20) CONTA [54]	\mathcal{I}	ResNet101	33.4
(CVPR’20) SEAM [38]	\mathcal{I}	ResNet38	31.9
(ICCV’21) CSE [43]	\mathcal{I}	ResNet38	36.4
(ICCV’21) CDA [55]	\mathcal{I}	ResNet38	33.2
(CVPR’22) SIPE [4]	\mathcal{I}	ResNet101	40.6
(CVPR’22) ReCAM [44]	\mathcal{I}	ResNet101	39.4
(CVPR’22) Ru et al. [3]	\mathcal{I}	MiT-B1	38.9
(Ours) eX-ViT	\mathcal{I}	ResNet38	42.9

and complete object segmentation results in various challenging scenarios, including different object scales and multiple objects.

In Table 3 we report the semantic segmentation results on the MS COCO 2014 dataset. It is observed that methods with the supervision of saliency maps such as DSRG [57] and AuxSegNet [48] do not provide results comparable or superior to the WSSS methods with only image-level labels. The poor performance is caused by the limitation of saliency maps generated by pre-trained models. Instead, our method that leverages image-level labels achieves a segmentation mIoU of 42.9% with ResNet38 backbone, which surpasses most recent state-of-the-art WSSS methods including SEAM [38], CSE [43], and CDA [55] by a large margin. Several qualitative segmentation results are shown in Fig. 6. These results confirm the effectiveness of our model, which is consistent with our intuition. Specifically, our eX-ViT remarkably improves the overall performance with the indispensable block of E-MHA and the AttE module. Adding these modules explicitly encourages eX-ViT to gain insightful clues on the complete object scene, and boost the model efficiency on producing accurate and complete object boundaries.



(a) Image

(b) Ground-truth

(c) E-MHA

(d) w/o AttE

(e) w/ AttE

Figure 7: Visualization results on the MS COCO 2014 validation set.

4.2.3. Comparison on interpretability

To compare our method with other explainable methods, we also adopt another two common metrics, i.e., average precision (AP) and average recall (AR). Which are commonly used in the literature to measure the interpretability [58]. We evaluate our method using the Deit backbone [59] and conduct the weakly-supervised image segmentation experiments, which is in line with earlier work [23]. The quantitative results are shown in Table 4. We can see that our model clearly surpasses the ViT model which contains the raw attentions, it reveals that our MAXNet achieves an AP of 15.7%, and an AR of 22.3%. We also observe that the post-hoc interpretability methods such as Rollout [12], GradCAM [60] and partial LRP [61] do not obtain faithful results compared to the counterparts. Which is possibly caused due to the extensive noises introduced by gradients or propagation rules.

Fig. 7 shows three cases of visualization results along with their ground truth segmentation label maps. Compared to the original CAM without AttE, attention maps produced with our model perform well on precisely locating both small and large objects with more complete object boundaries. This verifies our intuition with the design of eX-ViT and suggests that our proposed model is effective on learning comprehensive features for complete target objects.

Table 4: Performance comparison of various methods on the MS COCO validation set.

Method	AP	AP_medium	AP_large	AR	AR_medium	AR_large
(ICCV'17) GradCAM [60]	2.3	2.3	4.7	5.5	5.9	10.7
(ACL'19) Partial LRP [14]	4.7	8.0	5.1	10.4	19.9	8.0
(ICLR'20) ViT [2]	5.6	9.6	6.9	11.7	21.8	10.8
(ACL'20) Rollout [12]	0.1	0.1	0.2	0.4	0.1	0.9
(CVPR'21) Trans. attribution [13]	7.2	10.4	12.4	13.4	21.0	19.4
(ICCV'21) Chefer et al. [23]	13.1	14.4	24.6	19.3	23.9	33.2
(Ours) eX-ViT	15.7	15.3	26.5	22.3	24.3	36.1

4.3. Ablation studies

This section presents ablation studies to analyze the contributions of each component in our eX-ViT, including the transformer encoder with the proposed Explainable Multi-Head Attention (E-MHA), the Attribute-guided Explainer (AttE), the global-level attribute-guided loss function $\mathcal{L}_{\text{global}}$, the local-level attribute discriminability loss function \mathcal{L}_{dis} , and the attribute diversity loss function \mathcal{L}_{div} .

4.3.1. Effectiveness of E-MHA

It is an intuition that improved transformer attention mechanism in E-MHA will improve model’s ability to generate pseudo segmentation labels. In order to verify this idea, we simply apply CAM to the last K transformer encoder layers, denoted as E-MHA-CAM. Table 5 reports the mIoU results of the pseudo labels generated by CAM with the backbone of ResNet38, ResNet50, Conformer-S [62], and the encoder E^θ in our proposed eX-ViT. As can be seen, the backbone with the E-MHA module shows superior performance to its CNN counterparts. Specifically, E-MHA-CAM achieves the mIoU of 52.31%, which is a significant gain of +4.92%, +4.01% over ResNet38-CAM and ResNet50-CAM, respectively. By comparing the E-MHA with the recent state-of-the-art architecture, i.e., Conformer-S [62], we find that our proposed E-MHA still achieves a promising result. In details, compared to CrossFormer-S [62] which explicitly uses multi-scale representations with convolutions to localize object details, E-MHA-CAM achieves the best mIoU of 52.31%, which is 0.61% points higher than CrossFormer-S-CAM. The performance improvement shows that exploiting the most

Table 5: Performance comparison of various methods in terms of mIoU (%) on the PASCAL VOC 2012 training set.

Method	mIoU(%)
(CVPR'19) ResNet50-CAM (author?) [63]	48.30
(CVPR'20) ResNet38-CAM [38]	47.43
(ICCV'21) Conformer-S-CAM [62]	51.70
(Ours) E-MHA-CAM	52.31

Table 6: Effect of the contributions from various modules in terms of mIoU (%) on the PASCAL VOC training set.

E^θ	$\mathcal{L}_{\text{global}}$	\mathcal{L}_{dis}	\mathcal{L}_{div}	training	validation
✓				44.72	50.20
✓	✓			53.71	55.43
✓		✓		51.25	54.63
✓			✓	52.13	55.50
✓	✓	✓		55.27	58.10
✓	✓		✓	58.08	59.82
✓	✓	✓	✓	59.13	61.20

frequent and robust features by use of E-MHA is highly effective for WSSS tasks which require discriminative features to localize instances.

4.3.2. Effectiveness of the AttE and attribute-guided loss

Table 6 gives an ablation study of each component in our proposed eX-ViT. We consider the first row as a baseline, where the results of the object localization maps are obtained via the CAM approach. As is observed from the table, the baseline can be further improved to 53.71% and 55.43% on the training and validation set, respectively by using the attribute features obtained via AttE to refine the learned transformer attention from the eX-ViT. Empirically, with attribute-guided discriminability loss \mathcal{L}_{dis} the pseudo segmentation label maps can be improved by +6.53% compared to the baseline on the PASCAL VOC training set (51.25% vs. 44.72%) even without the global supervision $\mathcal{L}_{\text{global}}$. Moreover, the \mathcal{L}_{dis} further improves the mIoU to 54.63% on the validation set. By incorporating the attribute diversity loss function \mathcal{L}_{div} to explicitly regularize

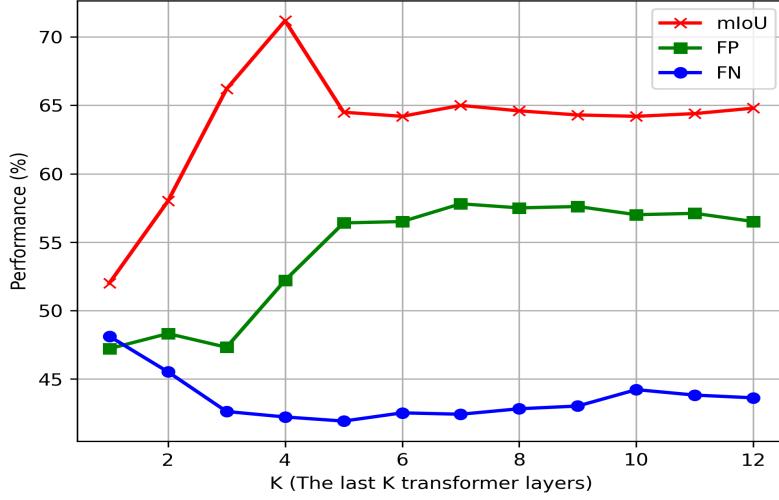


Figure 8: Evaluation of object localization maps generated by fusing the class-specific attentions from the last K transformer layers in eX-ViT’s encoder E^θ in terms of false positives (FP), false negatives (FN) and mIoU. The larger FP and FN values denote have more over-activated pixels, while the higher mIoU value indicates the generated localization maps have less over-activated pixels and more complete object coverage.

the attribute structure of the feature space, our full model gains promising results. Particularly, the results in Table 6 indicate that the proposed model performs better with the diversity constraint \mathcal{L}_{div} on the local consistency, which brings +4.37% and 4.39% mIoU improvements on the training and validation sets, respectively compared to the global-level loss. This also confirms our theory that improving the diversity among attributes promotes the learning of comprehensive localization maps.

4.3.3. Influence of the number of fused transformer layers

We further explore the impact of the number of fused transformer layers in Eq. (10) on the PASCAL VOC training set. Following the common practice in the prior work [38], we adopt three metrics to evaluate the performance, i.e., false positives (FP), false negatives (FN), and mIoU. The larger FP and FN values denote higher degrees of over-activated and under-activated areas, respectively. In Fig. 8, we compare the performance of the model variants using different numbers of the fused transformer layers. As is observed, when fusing layers with more than 10, we obtain localization maps with a larger FN value, which suggests the generated localization maps have

Table 7: The influence of the number of attributes in terms of mIoU (%) on the PASCAL VOC and MS COCO 2014 validation sets.

c	S	PASCAL VOC 2012 <i>val</i>	MS COCO 2014 <i>val</i>
512	8	69.42	40.31
	16	69.03	38.92
256	4	63.48	36.69
	8	71.23	41.23
	16	70.29	42.92
128	4	68.63	37.25
	8	68.56	38.79

more over-activated pixels and less complete activation coverage. This is mainly due to the limited ability of lower layers to encode high-level representations. By reducing the number of fused layers from the encoder E^θ , the performance of predicted localization maps become much better, i.e., more complete activation coverage (lower FN value) or fewer over-activated regions (lower FP value). Overall, the evaluation results indicate that using the last three attention layers can achieve the best mIoU of 71.2% with lower FN and FP values. Therefore we set $K = 3$ throughout the paper.

4.3.4. Influence of the number of attributes

The attribute-guided scheme allows the model to encode richer semantics into each attribute feature at a granular level. In order to discover the most suitable S concerning different datasets, we conduct extensive experiments to compare the performance of the model variants with different settings of hidden dimension c and the number of attributes S . As shown in Table 7, when $c = 128$, the model learns weaker representations for both datasets. In contrast, the performance becomes much better when the hidden dimension is enlarged to be 512. However, as the number of attributes increases to 16, the model exhibits a poor mIoU accuracy. In the end, we find that $c = 256$ achieves consistently superior performance across a range of attribute numbers. The best performance is achieved when $S = 8$ on the PASCAL VOC 2012 *val* set, and $S = 16$ on the MS COCO 2014 *val* set. These observations suggest that images in MS COCO 2014 tend to contain more local features that are discriminative for object localization.

5. Conclusion

In this paper, we proposed the eX-ViT, a new explainable vision transformer for weakly supervised semantic segmentation. In our framework, a novel Explainable Multi-Head Attention (E-MHA) module is proposed to produce discriminative feature representations with inherent explainability and noise robustness. Which is achieved by optimizing the dynamic alignment between the input tokens and attention weights. Moreover, a new Attribute-guided Explainer (AttE) module is designed to decompose the attention maps into the contribution of each individual attribute, empowering the feature representation with a set of attribute maps at a granular level. Based on AttE, we develop a self-supervised attribute-based loss to guide the learning of attribute features with the attribute discriminability mechanism and attribute diversity mechanism, which promotes the generation of diverse and discriminative object attributes. Extensive experiments were presented to demonstrate that the eX-ViT surpasses the state-of-the-art CNNs and transformers on two well-known benchmarks. We hope that the eX-ViT’s superior performance on WSSS tasks will inspire future research on the exploitation of the explainability of transformers.

References

- [1] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaïd, D. Xu, Multi-class token transformer for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1–19.
- [2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: Proceedings of the International Conference on Learning Representations (ICLR), 2020, pp. 1–21.
- [3] L. Ru, Y. Zhan, B. Yu, B. Du, Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1–17.
- [4] Q. Chen, L. Yang, J. Lai, X. Xie, Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1–11.

- [5] Y. Hao, L. Dong, F. Wei, K. Xu, Self-attention attribution: Interpreting information interactions inside transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 12963–12971.
- [6] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, et al., Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
- [7] Y. Wang, X. Wang, Self-interpretable model with transformation equivariant interpretation, in: M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, J. W. Vaughan (Eds.), Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2021, pp. 2359–2372.
- [8] B. Pan, R. Panda, Y. Jiang, Z. Wang, R. Feris, A. Oliva, Ia-red\$^2\$: Interpretability-aware redundancy reduction for vision transformers, in: M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, J. W. Vaughan (Eds.), Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2021, pp. 24898–24911.
- [9] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 9630–9640.
- [10] B. Zhou, A. Khosla, Á. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2921–2929.
- [11] G. Brunner, Y. Liu, D. Pascual, O. Richter, M. Ciaramita, R. Wattenhofer, On identifiability in transformers, in: Proceedings of 8th International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 2020, pp. 1–35.
- [12] S. Abnar, W. H. Zuidema, Quantifying attention flow in transformers, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 4190–4197.
- [13] H. Chefer, S. Gur, L. Wolf, Transformer interpretability beyond attention visualization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 782–791.

- [14] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, I. Titov, Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL), Florence, Italy, 2019, pp. 5797–5808.
- [15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 9992–10002.
- [16] W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 548–558.
- [17] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: Proceedings of the 38th International Conference on Machine Learning (ICML), 2021, pp. 10347–10357.
- [18] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, S. Yan, Tokens-to-token vit: Training vision transformers from scratch on imagenet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 538–547.
- [19] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, Q. Ye, TS-CAM: token semantic coupled attention map for weakly supervised object localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 2866–2875.
- [20] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, I. Titov, Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5797–5808.
- [21] G. Jawahar, B. Sagot, D. Seddah, What does BERT learn about the structure of language?, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL), Florence, Italy, 2019, pp. 3651–3657.

- [22] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, A. Dosovitskiy, Do vision transformers see like convolutional neural networks?, in: M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, J. W. Vaughan (Eds.), Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2021, pp. 12116–12128.
- [23] H. Chefer, S. Gur, L. Wolf, Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 387–396.
- [24] J. Ahn, S. Kwak, Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 4981–4990.
- [25] Y. Yao, T. Chen, G. Xie, C. Zhang, F. Shen, Q. Wu, Z. Tang, J. Zhang, Non-salient region object mining for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2623–2632.
- [26] X. Li, T. Zhou, J. Li, Y. Zhou, Z. Zhang, Group-wise semantic mining for weakly supervised semantic segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 1984–1992.
- [27] G. Papandreou, L. Chen, K. P. Murphy, A. L. Yuille, Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1742–1750.
- [28] J. Peng, H. Kervadec, J. Dolz, I. B. Ayed, M. Pedersoli, C. Desrosiers, Discretely-constrained deep network for weakly supervised segmentation, *Neural Netw.* 130 (2020) 297–308.
- [29] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 9729–9738.
- [30] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (4) (2017) 640–651.

- [31] J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap your own latent - A new approach to self-supervised learning, in: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 1–14.
- [32] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. M. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, *Int. J. Comput. Vis.* 111 (1) (2015) 98–136.
- [33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: common objects in context, in: Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 2014, pp. 740–755.
- [34] B. Hariharan, P. Arbelaez, L. D. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, in: D. N. Metaxas, L. Quan, A. Sanfeliu, L. V. Gool (Eds.), Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Barcelona, Spain, 2011, pp. 991–998.
- [35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, *Int. J. Comput. Vis.* 115 (3) (2015) 211–252.
- [36] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1026–1034.
- [37] Z. Wu, C. Shen, A. van den Hengel, Wider or deeper: Revisiting the resnet model for visual recognition, *Pattern Recognit.* 90 (2019) 119–133.
- [38] Y. Wang, J. Zhang, M. Kan, S. Shan, X. Chen, Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 12272–12281.
- [39] Y. Chang, Q. Wang, W. Hung, R. Piramuthu, Y. Tsai, M. Yang, Weakly-supervised semantic segmentation via sub-category exploration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 8988–8997.

- [40] T. Wu, J. Huang, G. Gao, X. Wei, X. Wei, X. Luo, C. H. Liu, Embedded discriminative attention mechanism for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16765–16774.
- [41] J. Lee, E. Kim, S. Yoon, Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4071–4080.
- [42] K. Sun, H. Shi, Z. Zhang, Y. 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), Montreal, QC, Canada, 2021, pp. 7263–7272.
- [43] H. Kweon, S. Yoon, H. Kim, D. Park, K. 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), Montreal, QC, Canada, 2021, pp. 6974–6983.
- [44] Z. Chen, T. Wang, X. Wu, X. Hua, H. Zhang, Q. Sun, Class re-activation maps for weakly-supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1–14.
- [45] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Trans. Pattern Anal. Mach. Intell.* 40 (4) (2018) 834–848.
- [46] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, Segformer: Simple and efficient design for semantic segmentation with transformers, in: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2021, pp. 1–18.
- [47] G. Sun, W. Wang, J. Dai, L. V. Gool, Mining cross-image semantics for weakly supervised semantic segmentation, in: A. Vedaldi, H. Bischof, T. Brox, J. Frahm (Eds.), Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 2020, pp. 347–365.

- [48] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaïd, F. Sohel, D. Xu, Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 6964–6973.
- [49] S. Lee, M. Lee, J. Lee, H. Shim, Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5495–5505.
- [50] P. Jiang, Q. Hou, Y. Cao, M. Cheng, Y. Wei, H. Xiong, Integral object mining via online attention accumulation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 2070–2079.
- [51] B. Zhang, J. Xiao, J. Jiao, Y. Wei, Y. Zhao, Affinity attention graph neural network for weakly supervised semantic segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* 1 (1) (2021) 1–15.
- [52] Y. Liu, Y. Wu, P. Wen, Y. Shi, Y. Qiu, M. Cheng, Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (3) (2022) 1415–1428.
- [53] B. Zhang, J. Xiao, Y. Wei, M. Sun, K. Huang, Reliability does matter: An end-to-end weakly supervised semantic segmentation approach, in: Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 2020, pp. 12765–12772.
- [54] D. Zhang, H. Zhang, J. Tang, X. Hua, Q. Sun, Causal intervention for weakly-supervised semantic segmentation, in: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 1–12.
- [55] Y. Su, R. Sun, G. Lin, Q. Wu, Context decoupling augmentation for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 6984–6994.
- [56] J. Lee, J. Choi, J. Mok, S. Yoon, Reducing information bottleneck for weakly supervised semantic segmentation, in: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2021, pp. 27408–27421.

- [57] Z. Huang, X. Wang, J. Wang, W. Liu, J. Wang, Weakly-supervised semantic segmentation network with deep seeded region growing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 7014–7023.
- [58] A. Khosla, A. S. Raju, A. Torralba, A. Oliva, Understanding and predicting image memorability at a large scale, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 2390–2398.
- [59] M. Guillaumin, D. Küttel, V. Ferrari, Imagenet auto-annotation with segmentation propagation, *Int. J. Comput. Vis.* 110 (3) (2014) 328–348.
- [60] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618–626.
- [61] A. Binder, G. Montavon, S. Lapuschkin, K. Müller, W. Samek, Layer-wise relevance propagation for neural networks with local renormalization layers, in: A. E. P. Villa, P. Masulli, A. J. P. Rivero (Eds.), *Proceedings of the International Conference on Artificial Neural Networks*, Barcelona, Spain, 2016, pp. 63–71.
- [62] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, Q. Ye, Conformer: Local features coupling global representations for visual recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 357–366.
- [63] J. Ahn, S. Cho, S. Kwak, Weakly supervised learning of instance segmentation with inter-pixel relations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 2209–2218.