Pattern Recognition 142 (2023) 109666

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

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 Mathmatical Sciences, La Trobe University, Melbourne, VIC 3086, Australia
[c] College of Information Technology, Beijing University of Technology, Beijing 100124, China

ARTICLE INFO

Article history:
Received 20 June 2022
Revised 22 March 2023
Accepted 30 April 2023
Available online 12 May 2023

Keywords:
Explainable
Attention map
Transformer
Weakly supervised

ABSTRACT

Recently vision transformer models have become prominent models for a multitude 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 model 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 with the self-supervised attribute-guided loss. The E-MHA tailors explainable attention weights that are able to learn semantically interpretable representations from tokens 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 predictions. Additionally, we have developed a self-supervised attribute-guided loss for our eX-ViT architecture, which utilizes both the attribute discriminability mechanism and the attribute diversity mechanism to enhance the quality of learned representations. As a result, the proposed eX-ViT model can produce faithful and robust interpretations with a variety of learned attributes. 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 explanation results obtained via eX-ViT are regarded as pseudo segmentation labels to train WSSS models. Comprehensive simulation results illustrate that our proposed eX-ViT model achieves comparable performance to supervised baselines, while surpassing the accuracy and interpretability of state-of-the-art black-box methods using only image-level labels.

1. Introduction

Over the last few years, transformer models have attracted increasing attention with encouraging results in a multitude 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]. The effectiveness of this framework lies in its ability to capture long-range dependencies. Despite their excellent performance, most transformer architectures are usually expressed as black boxes [3]. Specifically, the large number of parameters and complex interactions between modules make it challenging to provide explanations for the model predictions. Given the high applicability of transformers in high-risk decision-making domains, such as healthcare and autonomous

[☆] This work was supported by the Australian Government through the Australian Research Council's Discovery Projects Funding Scheme under Project DP220101634.

[*] Corresponding author.

E-mail addresses: lu.yu@my.jcu.edu.au (L. Yu), w.xiang@latrobe.edu.au (W. Xiang), fangjuan@bjut.edu.cn (J. Fang), phoebe.chen@latrobe.edu.au (Y.-P.P. Chen), L.Chi@latrobe.edu.au (L. Chi).

https://doi.org/10.1016/j.patcog.2023.109666

driving, there is a strong necessity for gaining insights into the model's decision-making process [4]. An interpretable solution is able to aid in debugging the models and identifying crucial features for downstream tasks.

Explainable Artificial Intelligence (XAI) is an emerging subfield of AI pursuing to capture the properties that have influence over the decision of a model [5]. 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. Several previous studies have pointed out that explainable models outperform post-hoc methods in faithfulness and stability [6]. Unfortunately, little work has been done so far in the field of explainable transformers. In order to leverage advantages of explainability, recent research efforts have been made to explore the possibility of building inherently explainable transformers [7]. However, the explicit expressive features were not explored to obtain faithful explanations w.r.t. model decisions.

Recently, transformers have shown promising results in weakly supervised semantic segmentation (WSSS) tasks [8]. The generation of pixel-level pseudo segmentation ground-truth labels based on image-level labels is a pivotal step for this task. Transformers employ MHA and MLP to effectively capture long-range semantic correlations, which play a critical role in localizing the target object. Despite the fact that different attention heads in the transformer can attend to diverse semantic areas of an image, it is still unclear how to correctly align these features with a particular semantic class [3]. One common issue among existing transformer-based works is the utilization of a token for each class, which often highlight the most discriminative region of an object instead of the entire object region [1].

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 for WSSS tasks. 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 WSSS tasks. 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., MCTFormer [1] and TransCAM [8]);
- 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 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 capturing local details. DeiT [9] addressed this issue by employing a strong image classifier as the teacher model to train data-efficient transformer models. Li et al. [8] designed the TransCAM, which explicitly utilizes the attention weights produced from the transformer to refine CAM results. Moreover, there are some research studies with modified ViT architectures that benefit downstream vision tasks such as semantic segmentation. However, most of the 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, Peng et al. [10] proposed the Conformer to aggregate both the convolutional operations and self-attention mechanisms into a unified framework. However, Conformer results in a more complicated design and additional computational cost. Xu et al. [1] added extra class tokens and enforced them learning the activation maps of different classes, it has limited ability to encode more information when it comes to a larger data set, e.g., COCO [11]. 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 [6]. Several previous studies have pointed out that explainable models outperform post-hoc methods in faithfulness and stability [6]. Unfortunately, little work has been done so far in the field of explainable transformers. Caron et al. [7] utilized a self-supervised approach called DINO based on vision transformers and concluded that the attention 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. [6] 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. Abnar and Zuidema [12] proposed to combine the attention scores across multiple layers, but this method fails to distinguish between positive and negative attributions. Recently, Chefer et al. [13] 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, The Class Activation Mapping (CAM) technique is widely used in the design of WSSS tasks to extract object localization maps and approximate the segmentation mask [14]. Despite the encouraging results, CAM suffers from the issue of 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. Chen et al. [15] designed the ReCAM, which a method that leverages CAM to extract pixels belonging to specific classes and subsequently incorporates them into a fully-connected layer along with the corresponding class label for further learning. Yuan et al. [16] proposed the multi-strategy contrastive learning framework to discover the similarity and dissimilarity of contrastive sample pairs. Lee et al. [17] learned pixel-level feedback by use of saliency map generated from the off-the-shelf detection model. Chen et al. [4] introduced several image-specific prototype features 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 the transformer attention 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 explainable 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 model 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 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 the encoder. Specifically, the parameters $\mathcal{E}$ in the lower branch use 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 $\boldsymbol{X} \in \mathbb{R}^{T\times d}$ where T is the spatial size and d is the feature dimension, each head H_h holds an explainable attention weight $\boldsymbol{A}_h \in \mathbb{R}^{N\times d}$ (N is the spatial size of $\boldsymbol{A}_h$.) that represents the relative importance of input features. That is, $\boldsymbol{A}_h$ 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 $\boldsymbol{X}$, the E-MHA aims to learn the attention weight that maximizes the alignment between input tokens and the attention weight. During the optimization process, maximizing 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 $\boldsymbol{X}$, the projected key, query and value are obtained as follows

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}^{Q}, \quad \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^{K}, \quad \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}^{V}, \tag{1}$$

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

Second, the self-attention operation is constructed by

$$\boldsymbol{W} = \frac{\boldsymbol{Q}\boldsymbol{K}^{T}}{\sqrt{d}}, \tag{2}$$

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

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

$$\boldsymbol{A} = f_{\theta}(\boldsymbol{W} + \boldsymbol{b})^{\mathrm{T}}, \tag{3}$$

where $\boldsymbol{b}$ is a trainable bias term, which is introduced as an initial alignment for the input patterns. $f_{\theta}(\cdot)$ is a non-linear function that

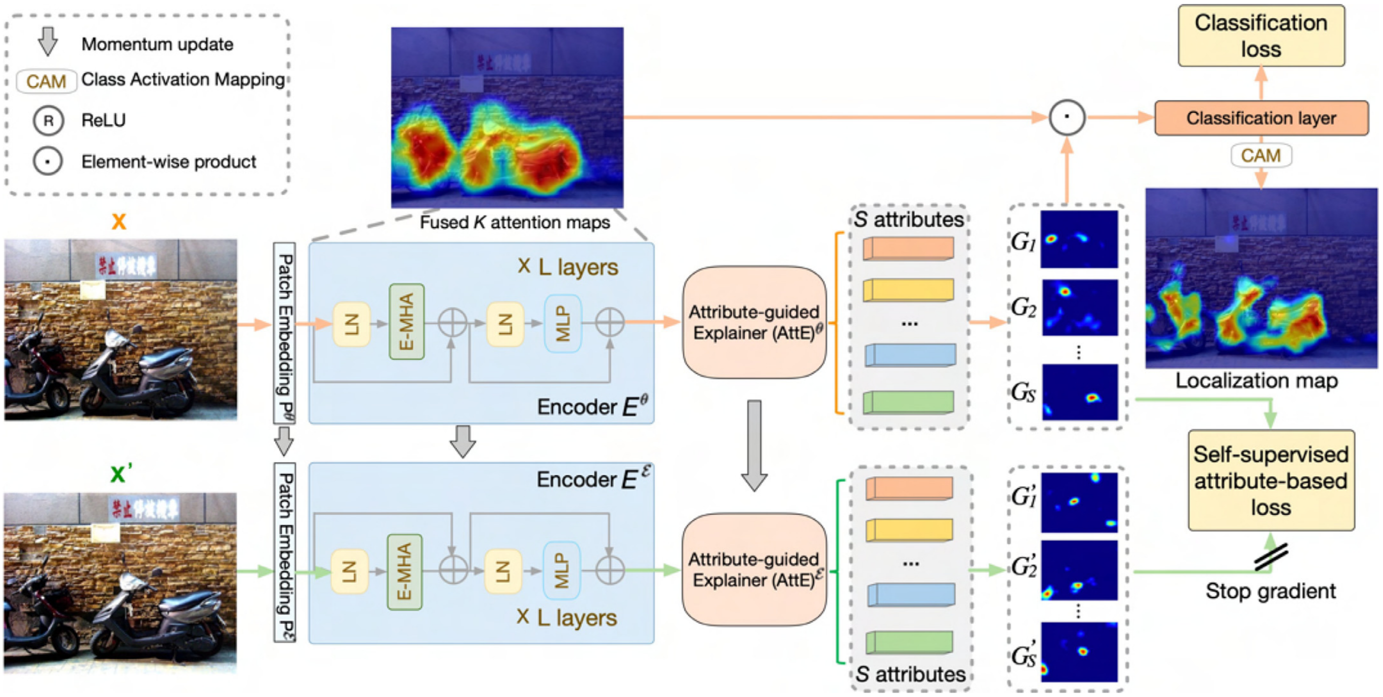


Fig. 1. Illustration of the proposed eXplainable Vision Transformer (eX-ViT) architecture. $\boldsymbol{x}$ and $\boldsymbol{x}'$ are two different random transformations of an input image. We use a transformer backbone as the encoder to extract feature maps, the backbone contains consecutive L encoding 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 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.

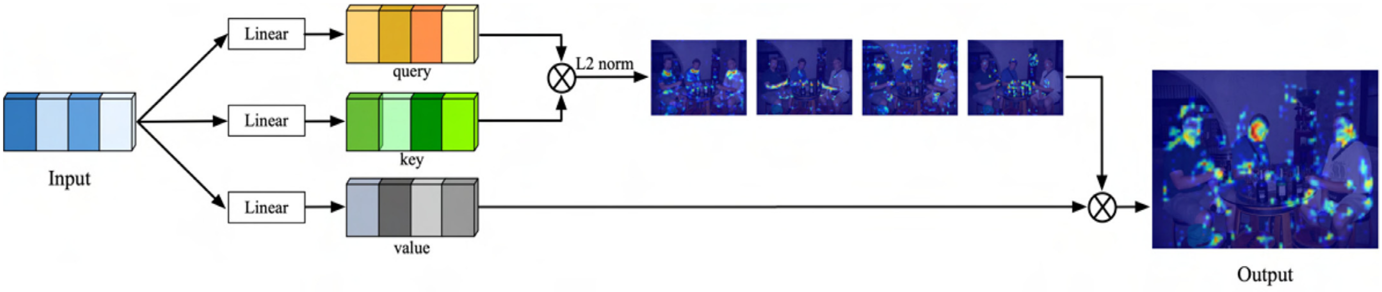


Fig. 2. The architecture of Explainable Multi-Head Attention (E-MHA). We use ⊗ to denote matrix multiplication.

scales the L2 norm of its input, i.e., $f_\theta(\boldsymbol{x}) = \frac{\boldsymbol{x}}{||\boldsymbol{x}||_2}$ and $||f_\theta(\boldsymbol{x})|| \le 1$. In our case, L2 norm is applied to the vector of $(\boldsymbol{W} + \boldsymbol{b})$.

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

$$\boldsymbol{S} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}, \tag{4}$$

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

$$\boldsymbol{S} = ||\boldsymbol{A}||\ ||\boldsymbol{V}||\cos(\boldsymbol{A}, \boldsymbol{V}) \le ||\boldsymbol{V}||. \tag{5}$$

Where both $\boldsymbol{A}$ and $\boldsymbol{V}$ are reshaped to a row-wise feature vector before applying the L2 norm function $||\cdot||$. When Eq. (5) is optimized, the attention weight $\boldsymbol{A}$ is proportional to $\boldsymbol{V}$. In order to achieve maximal output, $\boldsymbol{A}$ is driven to align with the discriminative features in $\boldsymbol{V}$, instead of the uninformative noise. Therefore, $\boldsymbol{S}$ can only achieve this upper bound if all possible solutions of $\boldsymbol{v} \in \boldsymbol{V}$ are encoded as eigenvectors in the weight $\boldsymbol{A}$. This maximization suggests with the attention weight $\boldsymbol{A}$, we will obtain an inherently explainable decomposition of input patterns.

Overall, the computation in layer l is expressed as

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

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

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.

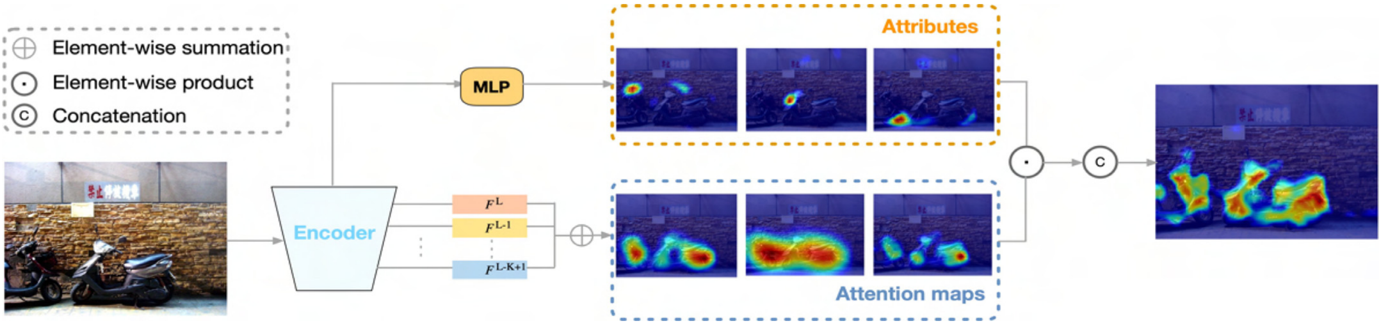


Fig. 3. Illustration of Attribute-guided Explainer (AttE). We aggregate the interpretable attention maps from the last K transformer layers to generate a fused attention map with good precision on the complete object context information. The attribute features are regarded as the complement information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

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, 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 (Fig. 3).

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 the feature map $\boldsymbol{F}^L \in \mathbb{R}^{H\times W\times d}$ produced by the encoder E^θ, we first calculate a set of spatial feature maps that capture the relative importance of all HW locations as follows

$$\boldsymbol{C}_{i,j} = f_\phi(\boldsymbol{F}^L), \quad \forall\{i,j\} \in H\times W, \tag{7}$$

where $f_\phi(\cdot)$ is implemented by a 2-layer MLP block, with one hidden layer followed by a LN layer and the ReLU activation layer. $\boldsymbol{C}_{i,j} \in \mathbb{R}^{H\times W\times c}$ is the obtained feature map with the channel 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 $\boldsymbol{C}_{i,j}$ along the channel dimension, which is formally expressed as

$$\bar{\boldsymbol{C}}_{i,j} = \frac{\boldsymbol{C}_{i,j}}{||\boldsymbol{C}_{i,j}||_2}, \tag{8}$$

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

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

$$\boldsymbol{G}_s = \bar{\boldsymbol{C}}_s \odot \boldsymbol{F}^L, \tag{9}$$

where $\odot$ is the element-wise product, and the $\bar{\boldsymbol{C}}_s$ is broadcast along the channel dimension to match the shape of $\boldsymbol{F}^L$. $\boldsymbol{G} = [\boldsymbol{G}_1, \boldsymbol{G}_2, \ldots, \boldsymbol{G}_S]$ is the final output that is concatenated along the channel dimension. By this means, each feature map $\boldsymbol{F}^L$ 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 Eqs. (7) to (9), the attribute representation $\boldsymbol{G}'$ of the second augmented input can be generated accordingly with the momentum encoder $E^{\mathcal{E}}$. And our $AttE^{\mathcal{E}}$ is also the exponential moving average of the trained $AttE^\theta$.

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. In our work, we propose to improve model 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 $\mathcal{L}_{\text{dis}}$ loss and the attribute diversity loss $\mathcal{L}_{\text{div}}$. 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 $\mathcal{L}_{\text{dis}}$ 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 feature map $\boldsymbol{F}^l$ in lth encoder layer, the fused attention map is expressed as

$$\hat{\boldsymbol{F}} = \frac{1}{K}\sum_{l}^{K}\boldsymbol{F}^l, \tag{10}$$

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

$$\boldsymbol{M} = \hat{\boldsymbol{F}} \odot \boldsymbol{G}, \tag{11}$$

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

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

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

where the prediction $\hat{y}_c$ is obtained by feeding the feature map $\boldsymbol{M}$ into a classification layer followed by a generalized mean pooling operation. 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, to improve the ability of network for learning 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 consistently 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 $\mathcal{L}_{\text{dis}}$ is defined by

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

where $g(\cdot)$ is the generalized mean pooling. And we adopt the normalized Mean Square Error as the $\ell(\cdot)$ function to calculate the 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 $\boldsymbol{G}$ and $\boldsymbol{G}^{'}$ for each individual attribute. Empirically, this attribute discriminability loss function $\mathcal{L}_{\text{dis}}$ 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 $\mathcal{L}_{\text{div}}$ is formally defined by

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

The intuition behind the $\mathcal{L}_{\text{div}}$ 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 $\mathcal{L}_{\text{dis}}$ and $\mathcal{L}_{\text{div}}$, 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 $\mathcal{L}_{\text{dis}}$ and $\mathcal{L}_{\text{div}}$ 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 [18] and MS COCO 2014 dataset [11]. PASCAL VOC 2012 dataset includes 20 object classes and one background class for the semantic segmentation task. Following the common experimental configuration from others, we adopt the augmented dataset which contains three subsets, training, validation, and testing sets, each having 10,582, 1449, and 1464 images, respectively. MS COCO 2014 dataset uses 81 classes, its training and validation sets have 82,081 images and 40,137 images respectively. Note that image-level labels are only used during training and ground-truth bounding box annotations are solely used during the inference time. In line with previous works [3], we report the mean Intersection-over-Union (mIoU) to evaluate the performance of our proposed model.

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

Method	Local. Maps	+denseCRF
(CVPR'20) SCE [21]	50.9	55.3
(CVPR'20) SEAM [20]	55.4	56.8
(CVPR'21) EDAM [22]	52.8	58.2
(CVPR'21) AdvCAM [23]	55.6	62.1
(ICCV'21) ECS-Net [24]	56.6	58.6
(ICCV'21) CSE [25]	56.0	62.8
(CVPR'22) SIPE [4]	58.6	64.7
(CVPR'22) ReCAM [15]	56.6	-
(Ours) eX-ViT	**59.1**	**65.3**

4.1.2. Implementation details

We use PyTorch for implementation and conduct experiments. The encoder parameters are pre-trained on ImageNet. 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 [19] as the backbone for semantic segmentation. At test time, only the branch with encoder E^{θ} is needed. Following the common practice in prior studies [20], 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 CAM in mIoU(%) on localization maps. 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 dense-CRF post-processing, and the gain becomes up to 65.3%. Fig. 4 shows visual comparisons of object localization maps on the 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 and precise localization maps.

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

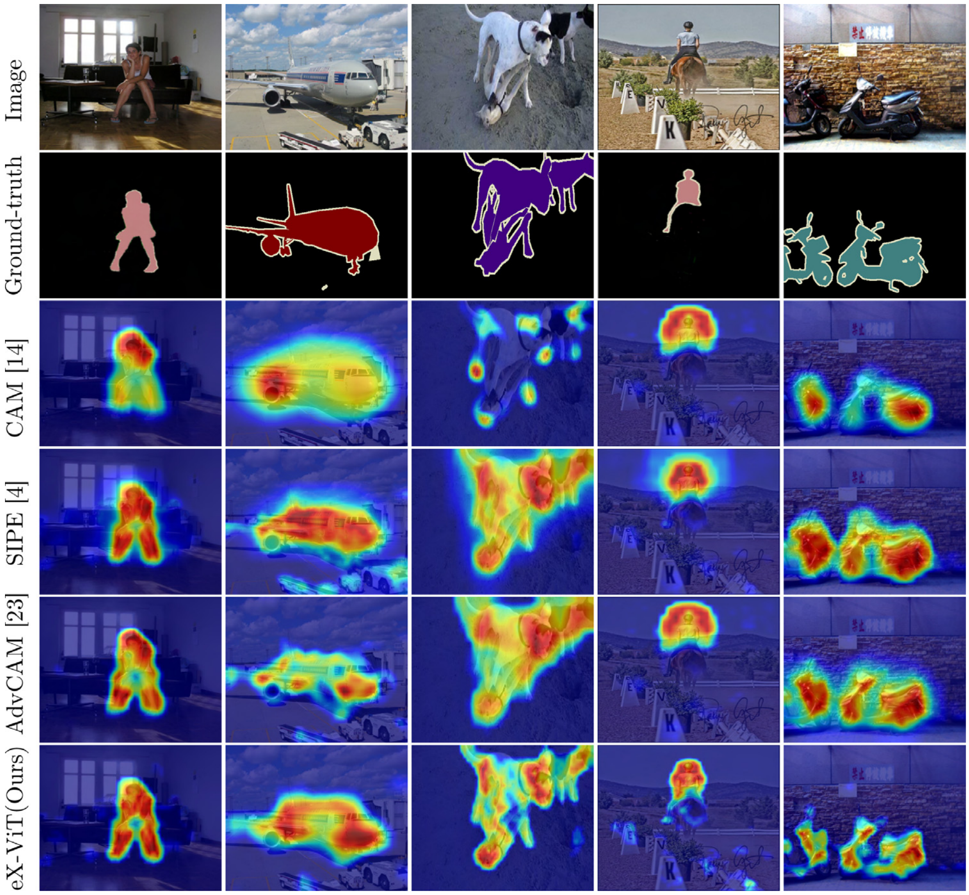


Fig. 4. Visual comparison of localization maps generated by different methods on PASCAL VOC 2012 training set. From top to down: original image, ground-truth, CAM [14], SIPE [4], AdvCAM [23] and our eX-ViT.

validation and test sets are reported in Table 2. 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 the validation set and 71.1% mIoU on the test set. Compared with the recent state-of-the-art weakly supervised models, e.g., EPS [17] and EDAM [22] that 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 and complete object segmentation results in various challenging scenarios, including different object scales and multiple objects.

Table 3 reports the semantic segmentation results on the MS COCO 2014 dataset. It is observed that methods with the supervision of saliency maps such as DSRG [37] and AuxSegNet [30] 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 [20], CSE [25], and MCTformer [1] 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

Table 2
Performance comparison of various methods in 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	*Sup.*	Backbone	validation	test
Fully-supervised methods				
(TPAMI'18) DeepLab [26]	$\mathcal{F}$	ResNet101	77.6	79.7
(PR'19) WideResNet38 [19]	$\mathcal{F}$	WR38	80.8	82.5
(NeurIPS'21) Segformer [27]	$\mathcal{F}$	MiT-B1	78.7	-
Weakly-supervised methods				
(CVPR'20) SEAM [20]	$\mathcal{I}$	ResNet38	64.5	65.7
(AAAI'20) RRM [28]	$\mathcal{I}$	ResNet101	66.3	66.5
(NeurIPS'20) CONTA [29]	$\mathcal{I}$	ResNet38	66.1	66.7
(ICCV'21) AuxSegNet [30]	$\mathcal{I}+\mathcal{S}$	ResNet38	69.0	68.6
(CVPR'21) EPS [17]	$\mathcal{I}+\mathcal{S}$	ResNet101	70.9	70.8
(CVPR'21) EDAM [22]	$\mathcal{I}+\mathcal{S}$	ResNet101	70.9	70.6
(ICCV'21) CDA [31]	$\mathcal{I}$	ResNet38	66.1	66.8
(ICCV'21) ECS-Net [24]	$\mathcal{I}$	ResNet38	66.6	67.6
(ICCV'21) CSE [25]	$\mathcal{I}$	ResNet38	68.4	68.2
(CVPR'21) AdvCAM [23]	$\mathcal{I}$	ResNet101	68.1	68.0
(NeurIPS'21) RIB [32]	$\mathcal{I}$	ResNet101	68.3	68.6
(TPAMI'21) A2GNN [33]	$\mathcal{I}$	ResNet101	66.8	67.4
(TPAMI'22) LIID [34]	$\mathcal{I}$	ResNet101	66.5	67.5
(CVPR'22) SIPE [4]	$\mathcal{I}$	ResNet101	68.8	69.7
(CVPR'22) ReCAM [15]	$\mathcal{I}$	ResNet101	68.5	68.4
(CVPR'22) Ru et al. [3]	$\mathcal{I}$	MiT-B1	66.0	66.3
(CVPR'22) MCTformer [1]	$\mathcal{I}$	ResNet38	71.9	71.6
(PR'22) Kho et al. [35]	$\mathcal{I}$	ResNet38	66.4	66.8
(PR'22) RRM-ResNet [36]	$\mathcal{I}$	ResNet101	69.3	69.2
(PR'23) MuSCLe [16]	$\mathcal{I}$	EfficientNet	66.6	68.8
TransCAM [8]	$\mathcal{I}$	ResNet38	69.3	69.6
(Ours) eX-ViT	$\mathcal{I}$	ResNet38	**71.2**	**71.1**

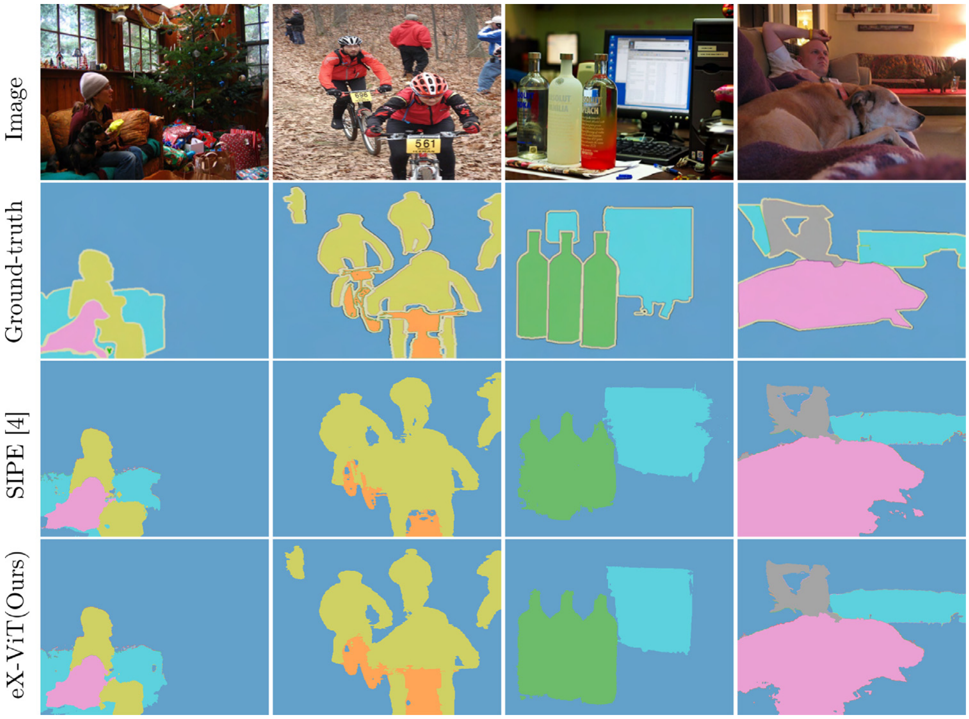


Fig. 5. Qualitative segmentation results on the validation set of PASCAL VOC 2012. From top to down: original image, ground-truth, SIPE [4] and our eX-ViT.

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

Method	*Sup.*	Backbone	mIoU (%)
CNN			
(CVPR'18) DSRG [37]	$\mathcal{I}+\mathcal{S}$	VGG16	26.0
(ICCV'21) AuxSegNet [30]	$\mathcal{I}+\mathcal{S}$	ResNet38	33.9
(CVPR'21) EPS [17]	$\mathcal{I}+\mathcal{S}$	ResNet101	35.7
(NeurIPS'20) CONTA [29]	$\mathcal{I}$	ResNet101	33.4
(CVPR'20) SEAM [20]	$\mathcal{I}$	ResNet38	31.9
(ICCV'21) CSE [25]	$\mathcal{I}$	ResNet38	36.4
(ICCV'21) CDA [31]	$\mathcal{I}$	ResNet38	33.2
(CVPR'22) ReCAM [15]	$\mathcal{I}$	ResNet101	39.4
(CVPR'22) SIPE [4]	$\mathcal{I}$	ResNet38	43.6
(NeurIPS'21) RIB [32]	$\mathcal{I}$	ResNet101	43.8
Transformer			
(CVPR'22) Ru et al. [3]	$\mathcal{I}$	MiT-B1	38.9
(CVPR'22) MCTformer [1]	$\mathcal{I}$	ResNet38	42.0
(Ours) eX-ViT	$\mathcal{I}$	ResNet38	**42.9**

efficiency in producing accurate and complete object boundaries. It is noted that both the RIB [32] and SIPE [4] outperform our proposed eX-ViT model on the COCO validation set. This is mainly because that vision Transformers are a relatively new model architecture for WSSS compared to their traditional CNNs counterparts. Therefore, ViTs still require further refinement and optimization to achieve the state-of-the-art performance. We hope that the eX-ViT's promising performance will inspire further research efforts to enhance ViTs' performance for WSSS tasks.

4.2.3. Comparison on interpretability

To compare our method with other explainable methods, we also adopt two common metrics, i.e., average precision (AP) and average recall (AR). Which are commonly used in the literature to measure interpretability. We evaluate our method using the DeiT backbone [9] and conduct the weakly-supervised image segmentation experiments, which is in line with earlier work [13]. 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 [38], and partial LRP [39] do not obtain faithful results compared to the counterparts. Which is possibly caused due to the extensive noises introduced by gradients or propagation rules.

Figure 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 in 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.

4.2.4. Analysis of misclassified examples

Figure 8 shows two misclassfied examples along with the learned attributes. In the first row of Fig. 8, the object “tv” is misclassified to a similar category “laptop”. The importance of the screen as a feature for a laptop could be the reason for this. The second row shows a more complicated example. We can observe while the attention map produced by our model captures most of details in the image, it is unable to distinguish class-specific features required to make accurate predictions for the target class, i.e., “broccoli”. In future work, we must explore a more compatible feature extractor that can generate more robust local features.

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

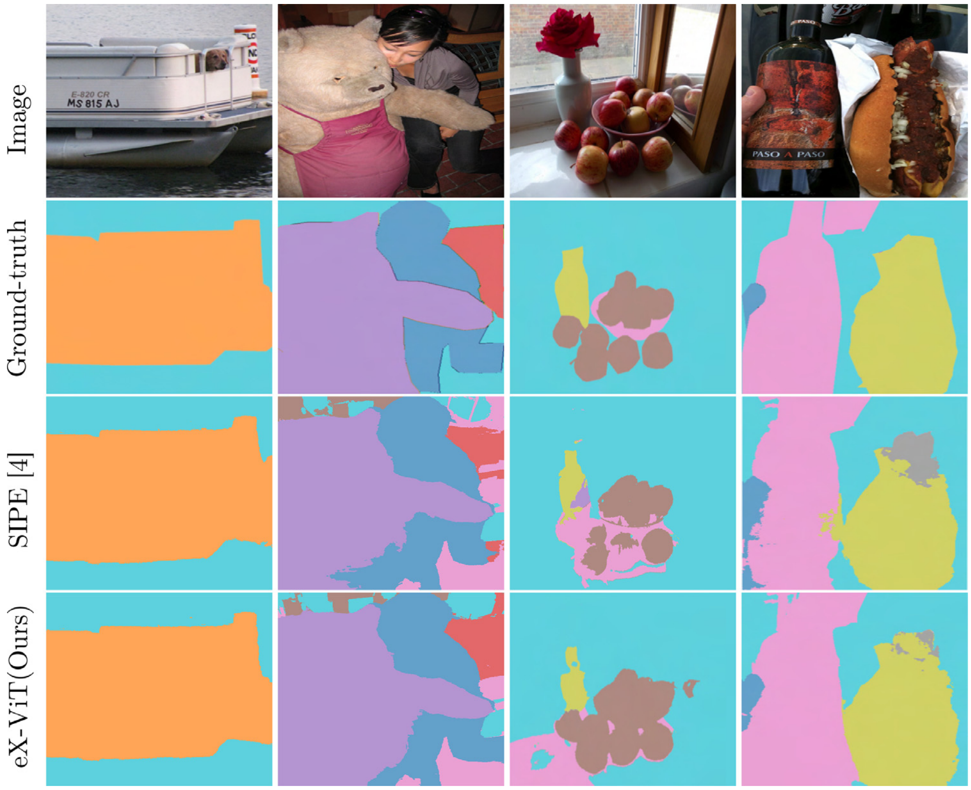
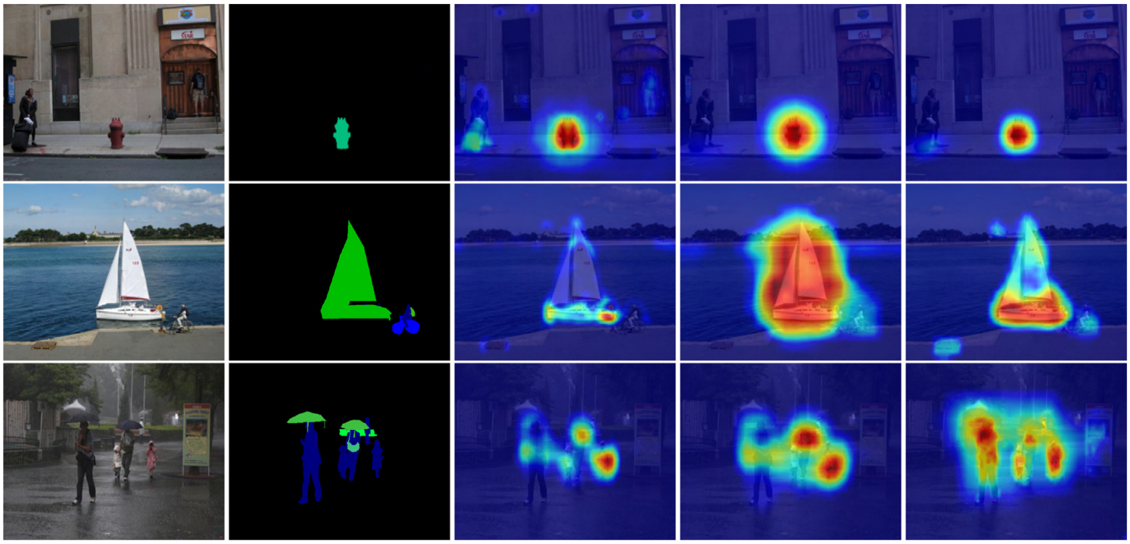


Fig. 6. Qualitative segmentation results on the validation set of MS COCO 2014. From top to down: original image, ground-truth, SIPE [4] and our eX-ViT.

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 [38]	2.3	2.3	4.7	5.5	5.9	10.7
(ACL'19) Partial LRP [39]	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 [6]	7.2	10.4	12.4	13.4	21.0	19.4
(ICCV'21) Chefer et al. [13]	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**



(a) Image (b) Ground-truth (c) E-MHA (d) w/o AttE (e) w/ AttE

Fig. 7. Visualization results on the MS COCO 2014 validation set.

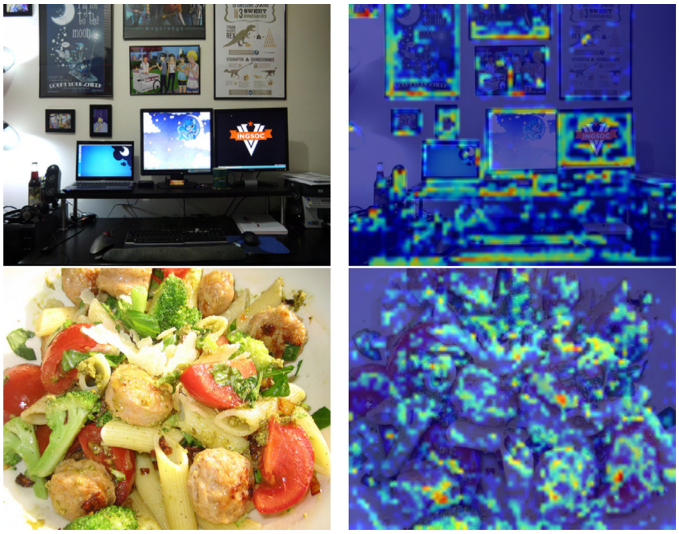


Fig. 8. Illustration of misclassified samples.

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

Method	mIoU(%)
(CVPR'19) ResNet50-CAM [40]	48.30
(CVPR'20) ResNet38-CAM [20]	47.43
(ICCV'21) Conformer-S-CAM [10]	51.70
(Ours) E-MHA	52.31

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

E^{θ}	$\mathcal{L}_{\text{global}}$	$\mathcal{L}_{\text{dis}}$	$\mathcal{L}_{\text{div}}$	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**

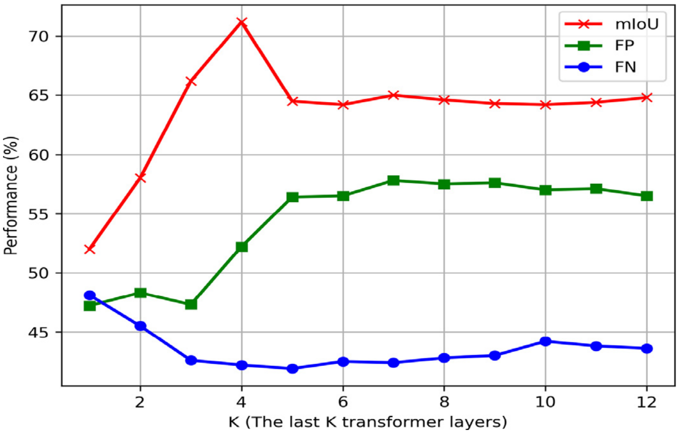


Fig. 9. Evaluation of object localization maps generated by fusing the class-specific attentions from the last K transformer layers in eX-ViT's encoder E^{θ} in terms of false positives (FP), false negatives (FN) and mIoU. The larger FP and FN values denote having more over-activated pixels, while the higher mIoU value indicates the generated localization maps have fewer over-activated pixels and more complete object coverage.

global-level attribute-guided loss function $\mathcal{L}_{\text{global}}$, the local-level attribute discriminability loss function $\mathcal{L}_{\text{dis}}$, and the attribute diversity loss function $\mathcal{L}_{\text{div}}$.

4.3.1. Effectiveness of E-MHA

It is an intuition that the improved transformer attention mechanism in E-MHA will improve the model's ability to generate pseudo segmentation labels. In order to verify this idea, we simply apply CAM to the last transformer encoder layer. Table 5 reports the mIoU results of the pseudo labels generated by CAM with the backbone of ResNet38, ResNet50, Conformer-S [10], and the encoder E^{θ} in our proposed eX-ViT. As can be seen, the backbone of 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% and +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 [10], we find that our proposed E-MHA still achieves a promising result. In details, compared to CrossFormer-S [10] 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 frequent and robust features by use of E-MHA is highly effective for WSSS tasks that 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 $\mathcal{L}_{\text{dis}}$ 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 $\mathcal{L}_{\text{dis}}$ further improves the mIoU to 54.63% on the validation set. By incorporating the attribute diversity loss function $\mathcal{L}_{\text{div}}$ to explicitly regularize 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 $\mathcal{L}_{\text{div}}$ 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 [20], 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. 9, 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 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 becomes 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

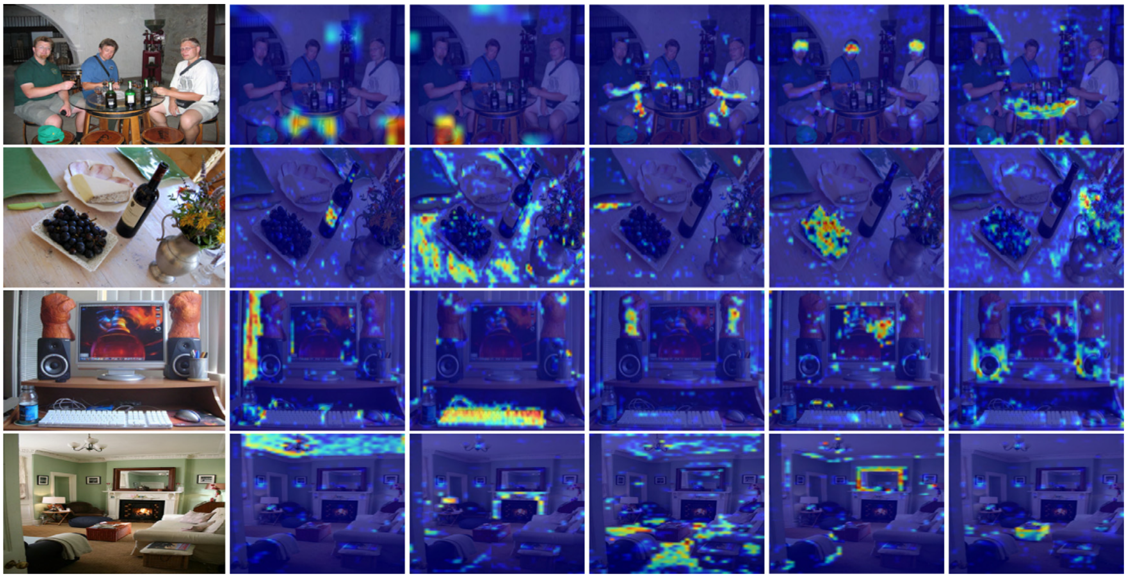


Fig. 10. Visualization of the learned attributes on the PASCAL VOC 2012 validation set, and MS COCO 2014 validation set, respectively. In each row, the left part is the input image, and the rest of images visualize the top-5 attributes, which shows that AttE attends to the discriminative attributes with a high degree of detail.

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

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

Table 8
The influence of hyperparameters in mIoU (%) on the PASCAL VOC validation and test sets.

Hyperparameter	value	PASCAL VOC 2012 *val*	PASCAL VOC 2012 *test*
α	0.1	69.8	69.4
	0.3	70.5	70.6
	0.5	**71.2**	**71.1**
	1.0	70.6	70.4
β	0.1	69.5	69.1
	0.3	70.1	70.3
	0.5	70.6	70.2
	1.0	**71.2**	**71.1**

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 512. However, as the number of attributes increases to 16, the model exhibits 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.

Additionally, we use the Grad-CAM as a tool to visualize the learned attributes and use Fig. 10 to present the visualization results. In each row of Fig. 10, the left column is the input image, and the images in the rest columns visualize the top 5 attributes. The first two rows are from the PASCAL validation set, whereas the last two rows are from the COCO validation set. By examining the visualization results presented in Fig. 10, several observations can be made regarding the effectiveness of the AttE in localizing object attributes. Firstly, the AttE is able to effectively focus on the compact regions of most objects, which is consistent with human observations. Secondly, for large-area attributes such as the table and ceiling, the learned attributes can accurately attend to the corresponding areas. Finally, the AttE is capable of attending to the regions of small but important attributes such as the fork and head. With these observations, we can ascertain how the AttE decomposes the feature map into different attributes.

4.3.5. Influence of hyperparameters

In this section, we explore how variations of hyperparameters can impact the performance of our model. For this purpose, we train models on PASCAL with each hyperparameter modification and report the accuracy in Table 8. It is observed that when α is small (<0.5), there is a slight performance drop. On the other hand, there is a significant accuracy drop when β is smaller than 1.0. This confirms that our model learns better features when our diversity loss enforces the attribute features to the maximally independent from each other, so as to capture broader visual clues of objects.

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.

Although our work has shown promising results, a limitation is that the proposed model does not incorporate attribute-level ground-truth labels. For future studies, the model should potentially be further improved if prior fine-grained knowledge of various attributes is integrated. Therefore, we plan to develop approaches to learn and integrate the knowledge in our future work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

[1] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaïd, D. Xu, Multi-class token transformer for weakly supervised semantic segmentation, in: IEEE 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: 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: IEEE 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: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1–11.

[5] 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.

[6] H. Chefer, S. Gur, L. Wolf, Transformer interpretability beyond attention visualization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 782–791.

[7] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: IEEE International Conference on Computer Vision (ICCV), 2021, pp. 9630–9640.

[8] R. Li, Z. Mai, Z. Zhang, J. Jang, S. Sanner, TransCAM: transformer attention-based cam refinement for weakly supervised semantic segmentation, J. Vis. Commun. Image Represent 92 (2023) 1–8.

[9] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: 38th International Conference on Machine Learning (ICML), 2021, pp. 10347–10357.

[10] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, Q. Ye, Conformer: local features coupling global representations for visual recognition, in: IEEE International Conference on Computer Vision (ICCV), 2021, pp. 357–366.

[11] 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: European Conference on Computer Vision (ECCV), 2014, pp. 740–755.

[12] S. Abnar, W.H. Zuidema, Quantifying attention flow in transformers, in: 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 4190–4197.

[13] H. Chefer, S. Gur, L. Wolf, Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers, in: IEEE International Conference on Computer Vision (ICCV), 2021, pp. 387–396.

[14] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2921–2929.

[15] Z. Chen, T. Wang, X. Wu, X. Hua, H. Zhang, Q. Sun, Class re-activation maps for weakly-supervised semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1–14.

[16] K. Yuan, G. Schaefer, Y.-K. Lai, Y. Wang, X. Liu, L. Guan, H. Fang, A multi-strategy contrastive learning framework for weakly supervised semantic segmentation, Pattern Recognit. (2023) 1–12.

[17] S. Lee, M. Lee, J. Lee, H. Shim, Railroad is not a train: saliency as pseudo-pixel supervision for weakly supervised semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 5495–5505.

[18] M. Everingham, S.M.A. Eslami, L.V. Gool, C.K.I. Williams, J.M. Winn, A. Zisserman, The pascal visual object classes challenge: aretrospective, Int. J. Comput. Vis. 111 (1) (2015) 98–136.

[19] Z. Wu, C. Shen, A. van den Hengel, Wider or deeper: revisiting the resnet model for visual recognition, Pattern Recognit. 90 (2019) 119–133.

[20] Y. Wang, J. Zhang, M. Kan, S. Shan, X. Chen, Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12272–12281.

[21] Y. Chang, Q. Wang, W. Hung, R. Piramuthu, Y. Tsai, M. Yang, Weakly-supervised semantic segmentation via sub-category exploration, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8988–8997.

[22] 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: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16765–16774.

[23] J. Lee, E. Kim, S. Yoon, Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4071–4080.

[24] K. Sun, H. Shi, Z. Zhang, Y. Huang, ECS-Net: improving weakly supervised semantic segmentation by using connections between class activation maps, in: IEEE International Conference on Computer Vision (ICCV), 2021, pp. 7263–7272.

[25] 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: IEEE International Conference on Computer Vision (ICCV), 2021, pp. 6974–6983.

[26] 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.

[27] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, in: Advances in Neural Information Processing Systems (NeurIPS), 2021, pp. 1–18.

[28] B. Zhang, J. Xiao, Y. Wei, M. Sun, K. Huang, Reliability does matter: an end-to-end weakly supervised semantic segmentation approach, in: AAAI Conference on Artificial Intelligence, 2020, pp. 12765–12772.

[29] D. Zhang, H. Zhang, J. Tang, X. Hua, Q. Sun, Causal intervention for weakly-supervised semantic segmentation, in: Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 1–12.

[30] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaïd, F. Sohel, D. Xu, Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation, in: IEEE International Conference on Computer Vision (ICCV), 2021, pp. 6964–6973.

[31] Y. Su, R. Sun, G. Lin, Q. Wu, Context decoupling augmentation for weakly supervised semantic segmentation, in: IEEE International Conference on Computer Vision (ICCV), 2021, pp. 6984–6994.

[32] J. Lee, J. Choi, J. Mok, S. Yoon, Reducing information bottleneck for weakly supervised semantic segmentation, in: Advances in Neural Information Processing Systems (NeurIPS), 2021, pp. 27408–27421.

[33] 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.

[34] 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.

[35] S. Kho, P. Lee, W. Lee, M. Ki, H. Byun, Exploiting shape cues for weakly supervised semantic segmentation, Pattern Recognit. 132 (2022) 1–13.

[36] B. Zhang, J. Xiao, Y. Wei, K. Huang, S. Luo, Y. Zhao, End-to-end weakly supervised semantic segmentation with reliable region mining, Pattern Recognit. 128 (2022) 1–13.

[37] Z. Huang, X. Wang, J. Wang, W. Liu, J. Wang, Weakly-supervised semantic segmentation network with deep seeded region growing, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7014–7023.

[38] 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: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.

[39] 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: 57th Conference of the Association for Computational Linguistics (ACL), 2019, pp. 5797–5808.

[40] J. Ahn, S. Cho, S. Kwak, Weakly supervised learning of instance segmentation with inter-pixel relations, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2209–2218.

Lu Yu received the Msc degree in college of computer science from Beijing University of Technology in 2015. She is currently working towards a PhD degree in computer science from James Cook University. Her research interests are computer vision and explainable AI.

Wei Xiang received the BEng and MEng degrees, both in electronic engineering, from the University of Electronic Science and Technology of China, Chengdu, China, in 1997 and 2000, respectively, and the PhD degree in telecommunications engineering from the University of South Australia, Adelaide, Australia, in 2004. Since January 2004, he has been with the School of Mechanical and Electrical Engineering, University of Southern Queensland, Toowoomba, Australia, where he currently holds a faculty post of Associate Professor. He has been awarded a number of prestigious fellowship titles, including the Queensland International Fellow (20102011) by the Queensland Government of Australia, the Endeavour Research Fellow (20122013) by the Commonwealth Government of Australia, the Smart Futures Fellow (20122015) by the Queensland Government of Australia, and JSPS Invitational Fellow (20142015) by the Japan Society for the Promotion of Science (JSPS). He received the Best Paper Award at 2011 IEEE WCNC, and the USQ Excellence in Research Award in 2013. He is an Editor of IEEE COMMUNICATIONS LETTERS. His research interests are in the broad area of communications and information theory, particularly coding and signal processing for multimedia communications systems.

Juan Fang received the PhD degree in college of computer science from Beijing University of Technology in 2005. She is a professor at Beijing University of Technology. Her research interests are in the area of computer architecture, edge computing and Artificial Intelligence.

Yi-Ping Phoebe Chen received the BInfTech (Hons.) and PhD degrees in computer science (bioinformatics) from the University of Queensland, Brisbane, QLD, Australia. She is currently a Professor and the Chair and Director of Research with the Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, VIC, Australia. She is the Chief Investigator with the ARC Centre of Excellence in Bioinformatics, Brisbane. She is involved in knowledge discovery technologies, and is especially interested in their application to genomics and biomedical science. She has been involved in the areas of bioinformatics, health informatics, multimedia databases, query systems, and systems biology. She has co-authored more than 200 research papers, with many published in top journals and conferences. Her current research interests include to find the best solutions for mining, integrating, and analyzing complex data structures and functions for scientific and biomedical applications. Prof. Chen is the Steering Committee Chair of the Asia-Pacific Bioinformatics Conference (Founder) and the International Conference on Multimedia Modelling. She has been on the program committees of more than 100 international conferences, including top ranking conferences, such as ICDE, ICPR, ISMB, and CIKM.

Lianhua Chi received the dual PhD degrees in computer science from the University of Technology Sydney, Australia, and the Huazhong University of Science and Technology, Wuhan, China, in 2015. She was a Post Doctoral Research Scientist in IBM Research Melbourne. Dr. Chi was a recipient of the Best Paper Award in PAKDD in 2013. Currently, she is a Lecturer with the Department of Computer Science and Information Technology at La Trobe University since 2018. Her current research interests include data mining and machine learning.