LocLoc: Low-level Cues and Local-area Guides for Weakly Supervised Object Localization

Xinzi Cao
Sun Yat-Sen University
Guangzhou, China
Peng Cheng Laboratory
Shenzhen, China
caoxz@mail2.sysu.edu.cn

Xiawu Zheng
Peng Cheng Laboratory
Shenzhen, China
zhengxw01@pcl.ac.cn

Yunhang Shen
Youtu Lab
Tencent
Shanghai, China
shenyunhang01@gmail.com

Ke Li
Youtu Lab
Tencent
Shanghai, China
tristanli@tencent.com

Jie Chen
Peng Cheng Laboratory
Shenzhen, China
chenj@pcl.ac.cn

Yutong Lu*
Sun Yat-Sen University
Guangzhou, China
The National Supercomputer Center
in Guangzhou
Guangzhou, China
luyutong@mail.sysu.edu.cn

Yonghong Tian*
Peking University
Beijing, China
Peng Cheng Laboratory
Shenzhen, China
yhtian@pku.edu.cn

ABSTRACT

Weakly Supervised Object Localization (WSOL) aims to localize objects using only image-level labels while ensuring competitive classification performance. However, previous efforts have prioritized localization over classification accuracy in discriminative features, in which low-level information is neglected. We argue that low-level image representations, such as edges, color, texture, and motions are crucial for accurate detection. That is, using such information further achieves more refined localization, which can be used to promote classification accuracy. In this paper, we propose a unified framework that simultaneously improves localization and classification accuracy, termed as LocLoc (Low-level Cues and Local-area Guides). It leverages low-level image cues to explore global and local representations for accurate localization and classification. Specifically, we introduce a GrabCut-Enhanced Generator (GEG) to learn global semantic representations for localization based on graph cuts to enhance low-level information based on long-range dependencies captured by the transformer. We further design a Local Feature Digging Module (LFDM) that utilizes low-level cues to guide the learning route of local feature representations for accurate classification. Extensive experiments demonstrate the effectiveness of LocLoc with 84.4%(↑ 5.2%) Top-1 *Loc.*, 85.8% Top-1 *Cls.* on CUB-200-2011 and 57.6% (↑ 1.5%) Top-1 *Loc.*, 78.6% Top-1 *Cls.* on ILSVRC 2012, indicating that our method achieves competitive performance with a large margin compared to previous approaches. Code and Appendix are available at https://github.com/Cliffia123/LocLoc.

CCS CONCEPTS

• **Computing methodologies → Object detection**.

KEYWORDS

weakly supervised object localization, low-level cues, transformer

ACM Reference Format:
Xinzi Cao, Xiawu Zheng, Yunhang Shen, Ke Li, Jie Chen, Yutong Lu, and Yonghong Tian. 2023. LocLoc: Low-level Cues and Local-area Guides for Weakly Supervised Object Localization. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612165

1 INTRODUCTION

Weakly supervised object localization (WSOL) aims to achieve efficient localization and classification simultaneously [39] by using only image-level supervision without bounding box annotations. It has attracted increasing attention in the research community for its low annotation costs [7, 9, 33, 34, 39]. Existing works mainly

[*] Corresponding authors.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612165

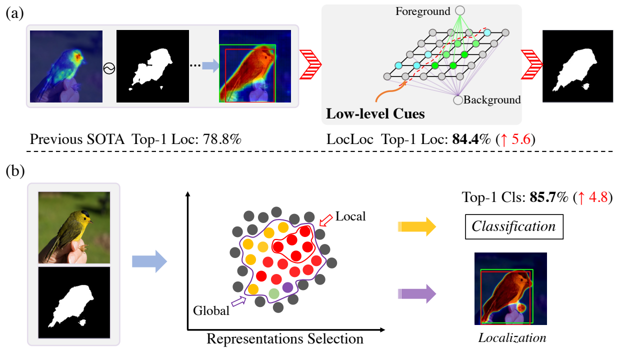


Figure 1: Comparison between previous SOTA method [29] and our LocLoc. (a) On the left, [29] produces a pseudo mask using a hard threshold, whereas we propose incorporates low-level cues (Details see Section 3.2 and Appendix in code link) to produce a GrabCut-Enhanced mask. (b) LocLoc explores feature representations for both classification and localization based on the enhanced mask (Colorful circles in Representation Selection indicates response values for classification). The red color highlights the significant improvement achieved by our LocLoc approach.

prioritize localization performance over classification [1, 13, 29], leading to unbalanced development of the two tasks and negatively affecting WSOL performance, particularly for Top-1 localization (Top-1 *Loc.*). The Top-1 *Loc.* metric is considered accurate only if both the Top-1 classification (Top-1 *Cls.*) and pure localization are precise. Any decrease in classification accuracy reduces the Top-1 *Loc.*. For example, Bai et al. [1] introduce a spatial calibration module for accurate localization, sacrificing classification accuracy, resulting in subpar Top-1 *Loc.*. In contrast, Chen et al. [6] improve both classification and pure localization, leading to higher Top-1 *Loc.* accuracy despite similar localization performance with [1]. Meanwhile, Xie et al. [28] achieve a considerable margin above SOTA on Top-1 *Loc.* by emphasizing improving both classification and localization accuracy. However, their approach relies on two separate models for WSOL, resulting in increased computational overhead. Therefore, the challenge in WSOL is to unify classification and localization into a single framework and improve both tasks simultaneously.

CAM [21, 31] is the representative WSOL method, but it only highlights the most discriminative regions, ignoring the full context of the object. To address this issue, several CAM-based techniques have been proposed, including adversarial erasing [7, 13, 34, 35], spatial relation activation [1, 11, 30, 37], and low-level features activation like SPOL [26]. Compared with the first two methods, SPOL shows a more competitive performance. SPOL emphasizes the importance of low-level features, which have richer global information that can help with global localization. However, it also activates background noise in shallow layers, impeding global localization. Therefore, it is essential to find a way to leverage low-level cues while reducing background noise activation for effective global localization in WSOL.

Recently, transformer architecture is now used in computer vision to extract features for image recognition. Specifically, the vision transformer (ViT) uses self-attention mechanisms to capture long-range dependencies and maintain global spatial information [8, 18]. Caron et al. [3] proposed a self-supervised ViT named DINO that leverages the advantage of the ViT features to produce self-attention maps covering more object context. Xu et al. [29] further proposed a Proxy Probing Decoder (PPD) that uses DINO self-attention maps as pseudo-supervision to promote localization, which has significantly alleviated the partial activation in WSOL. Inspired by this, we utilized the long-range capabilities of transformers to capture more global context.

In this paper, we propose a novel approach for weakly supervised object localization (WSOL) with Low-level Cues and Local-area Guides (LocLoc). The key idea of LocLoc is to develop a unified framework that integrates global and local representations for accurate localization and classification, respectively. We propose a GrabCut-Enhanced Generator (GEG) that enhances the low-level image cues in the long-range dependencies captured by DINO [3] to extract global semantic representations for localization. Second, we design a Local Feature Digging Module (LFDM) that employs low-level information to guide the learning of local feature representations for accurate classification. The LFDM consists of transformer encoders and two convolutional layers that mine global and local representations to activate class-specific regions for higher classification. We further adopt a novel data augmentation technique to enhance recognition accuracy. Only the GEG and LFDM are updated during training, while pre-trained model parameters are fixed. We conduct extensive experiments to demonstrate the effectiveness of the proposed LocLoc. The results show that our method achieves competitive performance in WSOL for both localization and classification tasks.

Collectively, our main contributions are summarized as follows:

- We propose a novel approach LocLoc for WSOL, which is the first attempt to leverage low-level cues to facilitate global representations and local feature representations learning, respectively.
- We introduce a GrabCut-Enhanced Generator (GEG), which utilizes the GrabCut-Enhanced activation as pseudo-labels rich in low-level information to learn global representations for localization.
- We design a Local Feature Digging Module (LFDM) to activate local class-specific representations based on long-range feature dependencies from the transformer, leading to improved classification accuracy.
- Our LocLoc method achieves state-of-the-art performance on both CUB-200-2011 and ILSVRC 2012 datasets with 85.4% (↑ 5.2%) and 57.6% (↑ 1.5%) Top-1 localization accuracy compared to the best transformer-based methods, respectively.

2 RELATED WORK

2.1 Low-level Information

Low-level information in images refers to fundamental visual features such as color, texture, and shape crucial for higher-level image understanding [4, 23]. These features are commonly extracted using techniques such as edge detection, gradient analysis, and texture analysis and are used in various computer vision applications, including image segmentation, object recognition, and image retrieval [17]. Edge detection identifies object boundaries in an image and has been used for object proposal generation in object recognition systems, as demonstrated by Zitnick et al. [42] with their Edge Boxes method. Histogram of Oriented Gradients (HOG) is another feature extraction method that analyzes image pixel gradients. Chen et al. [5] proposed a spectral imaging-based approach using HOG that utilizes unique spectral signatures to distinguish objects of interest more discriminatively. GrabCut [19] is an interactive image segmentation technique based on graph cuts optimization that iteratively refines segmentation by incorporating color and texture information. He et al. [12] proposed an improved GrabCut method for image segmentation using multi-scale features. Utilizing low-level information is crucial as it can detect features ignored by high-level semantics, such as edges and colors. Therefore, we focus on effectively motivating low-level cues for image-processing tasks.

2.2 CNN-based methods for WSOL

CNNs are powerful feature extractors that capture relevant information in the input image. There are many works based on CNNs to generate feature maps as final localization maps. One of the most representative works is class activation maps (CAMs) from the global average pooling (GAP) layer concerning the predicted class by Zhou et al. [40]. Unfortunately, limited by the category instruction, CAMs tend to identify the most discriminative regions within an image responsible for the specific class. Therefore, many works have focused on alleviating this issue. Zhang et al. [35] introduced ACoL with an adversarial erasing training strategy to discover complementary regions. Mai et al. [14] further proposed MEIL that shares the classification parameters of the branches of anti-erasure learning and pushes the erasure anti-complementary learning to the peak. More and more methods divide WSOL tasks into class-agnostic localization task and classification task as two independent subtasks, such as SPOL [26], BAS [27] and C^{2}AM [28], and achieves remarkably improvement in WSOL.

These methods relieve the issue of focusing on discriminative regions. Unfortunately, CNNs only capture partial semantic features with a limited receptive field, making it challenging to model object localization. Meanwhile, the CNN-based method for WSOL typically relies on high-level semantic cues highlighting part of the class-specific regions, further resulting in part activation. However, these class-specific regions are helpful for local feature mining in classification. Therefore, we propose to leverage the vision transformer to activate features for localization and combine CNNs to explore classification.

2.3 Transformer-based methods for WSOL

Vaswani et al. [24] were the pioneers of the transformer, which was designed to overcome the limitations of RNNs in handling long data sequences by using a self-attention mechanism to capture long-range dependencies. Recent studies have shown that transformers perform well in computer vision tasks [2, 8, 38]. Dosovitskiy et al. [8] designed the ViT model, which captures long-range dependencies and global information in image classification tasks. Carion et al. [2] introduced BETR, a transformer-based model that preserves spatial information effectively and achieves promising results in object detection. Gao et al. [10] proposed TS-CAM, which leverages the semantic-agnostic attention map of Deit [22] for accurate and semantically-aware localization compare CAM in CNNs, inspiring researchers to focus on the semantic relevance of transformers. Moreover, PPD [29] utilizes self-attention maps as pseudo-labels for training a localization generator by exploiting the global context regions through self-attention maps.

However, PPD [29] must use DINO [3] as the backbone, leading to reduced model generalization and lower classification performance compared to current state-of-the-art methods. In contrast, our method maintains high localization accuracy while achieving superior generalization across different models and maintaining the highest level of WSOL in classification.

3 MTEHOD

In this section, we elaborate on our proposed LocLoc, organized as follows. Firstly, we review the GrabCut algorithm and produce the GrabCut-Enhanced Mask (Sec. 3.2). Secondly, we present a detailed description of GEG, which is learning global representations for object localization (Sec. 3.3). Thirdly, we describe LFDM, which explores local feature representations for object classification (Sec. 3.4). Finally, we demonstrate how these components are incorporated into the inference process (Sec. 3.5).

3.1 Overview

As shown in the figure 2, LocLoc consists of three parts: low-level information enhanced by GrabCut, GrabCut-Enhanced Generator (GEG), and Local Feature Digging Module (LFDM). As discussed in [3], the self-attention map generated by the self-supervised ViT can segment an object's foreground. However, unlike PPD [29], which simply thresholds the self-attention map as pseudo supervision to train a generator, we first propose to utilize GrabCut to explore low-level information upon the self-attention map from the corresponding image. Secondly, we have devised a new generator GEG that combines the UNet structure and deconvolution layers, which enables the utilization of features from the transformer with long-range dependencies and discovering more global representations to alleviate the part object activation. Finally, we introduce LFDM as a classifier to search the local class-specific feature in long-range dependencies. Besides, we also leverage the GrabCut-Enhanced mask for image augmentation in LFDM, which leads to the highest classification performance in weakly supervised object localization.

The input images as $I = \{(I_i, y_i)\}_{i=0}^{M-1}$, where $y_i \in \{0, 1, \cdots, C-1\}$ represents the class of image I_i, M and C refer to the total number of images and classes, respectively. As shown in Figure 2, an input image is split into $N = w \times h$ patches, each patch represented as a token with a resolution of (w, h), where $w = W/P$, $h = H/P$, P is the width/height of a patch, and W and H represent the width and height of the input image, respectively. After grouping the aforementioned patch tokens and the CLS token into a sequence, the input is fed into a DINO [3] model pre-trained on the ImageNet dataset and generates a self-attention map. In order to generate the self-attention map $\boldsymbol{F}^{att} \in \mathbb{R}^{w \times h}$, DINO initially calculates the average of self-attention heads in each layer, resulting in the attention weight $\boldsymbol{W}_i \in \mathbb{R}^{(N+1)\times(N+1)}$, which includes the class token

$\boldsymbol{W}_i^{cls} \in \mathbb{R}^{(w)\times(h)}$ at the i^{th} layer. Subsequently, it concatenates the attention weights across all layers and averages them to generate the final attention map $\boldsymbol{F}^{att}$.

3.2 GrabCut-Enhanced Mask

In this section, we describe how to use the GrabCut algorithm to generate GrabCut-Enhanced Mask. GrabCut is an interactive algorithm involving the user initializing foreground and background pixels as seed pixels. The algorithm then iteratively refines the segmentation by incorporating color and texture information using a Gaussian Mixture Model (GMM) [4]. Therefore, once we have a rough feature map of the foreground and background, we can utilize the GrabCut algorithm to generate a binarized mask with a clearer foreground.

Moreover, the self-attention maps generated by DINO [3] have been observed to cover more foreground areas. We leverage this feature by using the GrabCut algorithm to generate a GrabCut-Enhanced mask with better foreground and background separation instead of binarizing the rough self-attention map into a mask as a pseudo-label in PPD [29], Details see appendix in code link.

Specifically, we divide the pixels into three distinct categories: foreground regions, background regions, and uncertain regions. The partitioning of the self-attention map $\boldsymbol{F}^{att}$ into three parts is achieved using double thresholds δ_0 and δ_1, as the green, blue and grey circles illustrated in Figure 1 (a). Values with high responses correspond to the foreground (green circles), while those with low responses represent the background (grey circles). The uncertain regions are identified with low confidence (blue circles) between the foreground and background. We then utilize a Gaussian Mixture Model (GMM) within GrabCut, which learns representations of both foreground and background pixels. This learned information is then used to classify the pixels in uncertain regions, and each pixel is assigned to either foreground or background based on the highest probability. This produces the GrabCut-Enhanced mask M^G, which is rich in low-level cues, and significantly improves the accuracy of image segmentation, especially for images with complex backgrounds.

3.3 GrabCut-Enhanced Generator

In this section, we utilize GrabCut-Enhanced mask M^G to supervise the GrabCut-Enhanced Generator for exploring global representations while disregarding any specific category information and generate activation maps that are bounded in the range of [0-1], which can be used to identify foreground for object localization.

First, we propose to use a UNet architecture as a semantic coder to extract global semantic information. UNet is a widely-used network architecture that consists of an encoder and decoder, which can effectively capture global semantic details of the long-range dependencies from the transformer.

Let $\boldsymbol{F}^i \in \mathbb{R}^{N\times D}$ be the output from the i-th transformer block in a frozen model, where D represents the dimension of each patch. We reshape $\boldsymbol{F}^i \in \mathbb{R}^{D\times w\times h}$ and apply the proposed GEG for further activating semantic features.

$$\boldsymbol{F}^u = f^u(\boldsymbol{F}^i; \mathbf{W}^{\mathrm{u}}_{\mathrm{G}}), \tag{1}$$

where $\mathbf{W}^{\mathrm{u}}_{\mathrm{G}}$ represents the parameters of the UNet f^u, $\boldsymbol{F}^u$ denotes the output of the segmentation module, with a dimension of $\boldsymbol{F}^u \in \mathbb{R}^{D'\times w\times h}$. To generate the activation map, we feed $\boldsymbol{F}^u$ into a block that comprises three deconvolution layers, a 2D convolution layer (Conv), and a Batch Normalization layer (BN).

$$\boldsymbol{F}^d = f^d(\boldsymbol{F}^u; \mathbf{W}^{\mathrm{d}}_{\mathrm{G}}), \boldsymbol{F}^a = \mathrm{Sigmoid}\left(\mathrm{BN}\left(\mathrm{Conv}\left(\boldsymbol{F}^d; \mathbf{W}^{a}_{G}\right)\right)\right), \tag{2}$$

where $\mathbf{W}^{\mathrm{d}}_{\mathrm{G}}$ and $\mathbf{W}^{\mathrm{a}}_{\mathrm{G}}$ denote the learning parameters of the decoder f^d and the Conv-BN layer, respectively. The output $\boldsymbol{F}^u$ of the UNet module is fed into this feature representation block, resulting activation map for localization, represented by $\boldsymbol{F}^a \in \mathbb{R}^{1\times(2^3\times w)\times(2^3\times h)}$.

The GEG is supervised using a mean square error loss, as expressed in Eq (3). The loss measures the discrepancy between the GrabCut-Enhanced mask M^G and generated activation map $\boldsymbol{F}^a$,

$$\mathcal{L}_{mse} = \frac{\sum_{i=0}^{h-1}\sum_{j=0}^{w-1}\left(M^G(i,j) - \boldsymbol{F}^a(i,j)\right)^2}{h\times w}. \tag{3}$$

By minimizing the loss, the GrabCut-Enhanced Generator can effectively leverage the learned global semantic representations to activate the object regions with the aim of low-level information through a straightforward regression approach.

3.4 Local Feature Digging Module

In the classification module, we leverage a combination of transformer encoder and convolutional neural layers to learn local class-specific representations, as illustrated in Figure 2. Convolutional neural networks are proficient in capturing local spatial patterns, while attention mechanisms excel in modeling global dependencies and long-range relationships between features. By fusing these two types of layers, the model can effectively exploit global information and further mine local representations. Additionally, we introduce a novel data augmentation technique wherein the background in the original image is masked, and a background patch from the same image is randomly filled, which encourages the model to learn to recognize and classify objects in the context of their surroundings, resulting in improved generalization performance.

Specifically, I and M^G denote the original image and its corresponding GrabCut-Enhanced mask. Firstly, a random bounding box coordinates $A = (r_x, r_y, r_w, r_h)$ is sampled uniformly, and the corresponding regions in M^G are filled with a patch cropped from A with value 1 to produce the augmented image $\tilde{M}^G$. Similar to CutMix[32], we sample the box coordinates according to a uniform distribution concerning the image width and height, where r_w and r_h are scaled by the parameter λ to control the size of the bounding box :

$$\begin{aligned} r_x &\sim \mathrm{Unif}(0, W), \quad r_w = W\sqrt{1-\lambda}, \\ r_y &\sim \mathrm{Unif}(0, H), \quad r_h = H\sqrt{1-\lambda}. \end{aligned} \tag{4}$$

Let $r_x \sim \mathrm{Unif}(0, W)$ and $r_w = W\sqrt{1-\lambda}$ denote the uniformly sampled x-coordinate and width of the bounding box, respectively. Similarly, let $r_y \sim \mathrm{Unif}(0, H)$ and $r_h = H\sqrt{1-\lambda}$ denote the uniformly sampled y-coordinate and height of the bounding box, respectively. Where λ is sampled from the uniform distribution (0, 1), next, the original image I is multiplied by the mask $\tilde{M}^G$ to generate

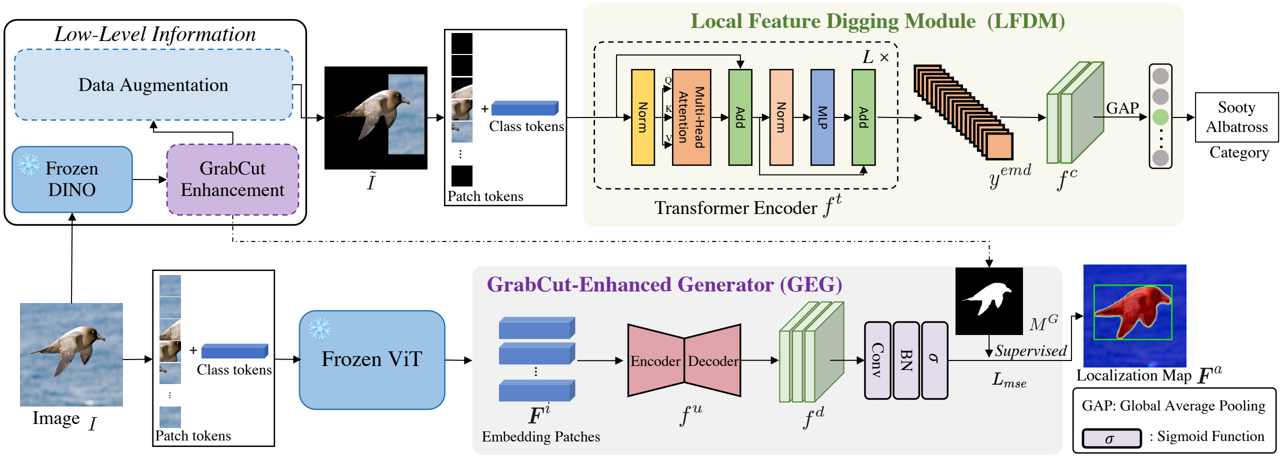


Figure 2: Overview of our proposed LocLoc framework, which consists of a low-level cues enhancement plugin, GrabCut-Enhanced Generator (GEG), and Local Feature Digging Module (LFDM). Firstly, the GrabCut algorithm enhances low-level cues and generates the GrabCut-Enhanced Mask M^G. Secondly, the ViT is frozen, and the GEG is trained to learn global feature representations for localization. Thirdly, the input image $\tilde{I}$, augmented with low-level cues, is used to extract local class-specific representations for classification. During inference, the image I is only passed through the frozen ViT, GEG, and LFDM to achieve localization and classification, respectively.

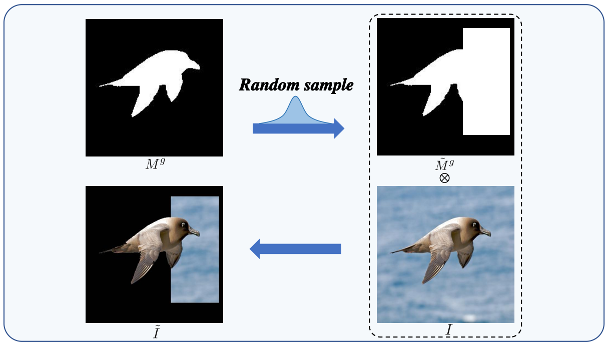


Figure 3: Data augmentation. Given the original image I, and its corresponding GrabCut-Enhanced mask M^G, we utilize the random sample strategy to produce a box whose area pixel will be set to 1. Finally, multiply M^G and image I to get $\tilde{I}$.

a new image $\tilde{I}$ that retains the whole foreground regions and adds some background information:

$$\tilde{I} = I \otimes \tilde{M}^G, \tag{5}$$

the $\tilde{I}$ is the same size as the original images, after that, the images $\tilde{I}$ are fed into transformer encoder f^t to get feature y_{emd}. The feature tensor y_{emd} is then reshaped to $\mathbb{R}^{D\times w\times h}$, where D represents the dimension of each patch, and then applied into CNNs f^c and global average pooling (GAP) [41] to generate the class probability distributions $\hat{y}$ which can be expressed as follows :

$$y_{emd} = f^t\left(\tilde{I};\mathbf{W}^{\mathbf{t}}\right), \hat{y} = GAP\left(f^c\left(y_{emd};\mathbf{W}^{\mathbf{c}}\right)\right). \tag{6}$$

Besides, we obtain the classification loss to $\mathcal{L}_{cls}$ by applying a cross-entropy loss to $\hat{y}$, which is used for classification learning of the entire image:

$$\mathcal{L}_{cls} = -\sum_{i=0}^{C-1} y_i \log \frac{e^{\hat{y}_i}}{\sum_j^C e^{\hat{y}_j}}. \tag{7}$$

Finally, the classifier is realized for exploring local representations with respect to class-specific regions with simple classification loss $\mathcal{L}_{cls}$.

3.5 Inference

During inference, we remove the low-level informance enhanced module. We input an original image and feed it into the frozen ViT to generate the patch embedding for both localization and classification. Next, the patch embedding is obtained after the attention blocks are fed to the GEG for localization. On the other hand, the patch embedding before the attention blocks are utilized for classification and are fed to LFDM to predict the corresponding class. The final result of the WSOL model is achieved by combining the predicted class and the localized object regions obtained from the activation map. This approach significantly improves the performance of WSOL tasks in terms of classification and localization.

4 EXPERIMENT

4.1 Experiment Setup

Datasets. We evaluate the effectiveness of our proposed LocLoc on two widely used datasets, including CUB-200-2011 [25] and ILSVRC 2012 [20]. CUB-200-2011 is a small dataset containing 200 species of birds with 5,994 training images and 5,794 testing images. In contrast, ILSVRC 2012 is a large-scale dataset with 1K classes and 1,281,197 training images, and 50,000 testing images. We use only the image-level labels for training, while for evaluation, we incorporate both the image-level annotations and bounding box annotations.

Evaluation Metrics. Following prior studies [20, 39], we adopt Top-1/Top-5 classification accuracy (Top-1/Top-5 *Cls.*), Top-1/Top-5 localization accuracy (Top-1/Top-5 *Loc.*), and GT-known localization (GT-known *Loc.*) accuracy as our evaluation metrics. Specifically, Top-1/Top-5 *Cls.* is considered correct if the Top-1/Top-5 predicted category contains the correct label. GT-known *Loc.* is deemed correct if the intersection over union (IOU) between the ground-truth and the predicted bounding box is greater than 0.5. Top-1/Top-5 *Loc.* is regarded correct only as both Top-1/Top-5 *Cls.* and GT-known *Loc.* are correct.

Implementation details. We empirically evaluate the proposed LocLoc on the widely used transformer-based backbone, namely Deit [22], which is pre-trained on the ILSVRC [20] and use self-supervised model DINO [3] to produce self-attention maps for GrabCut-Enhancing. We replace the MLP head with our Local Feature Digging Module (LFDM) while training the classifier. We obtain the self-attention maps as the input of our GrabCut-Enhanced method from the self-supervised DINO model, and resize the input images and corresponding GrabCut-Enhanced mask to 256 × 256 and randomly crop them to 224 × 224. For training the GrabCut-Enhanced Generator (GEG), we employ the Adam optimizer with a learning rate of 0.001 for 20 epochs. For training the classifier, we use the SGD optimizer with a learning rate of 0.00015. On the CUB-200-2011 dataset, we use a batch size of 128 and train the classifier for 80 epochs, while on ILSVRC 2012, we use a batch size of 256 and train the classifier for 40 epochs. All experiments are conducted using PyTorch and trained on Nvidia Tesla V100 GPUs.

4.2 Comparison with the State-Of-The-Arts

Localization. We first compare our results with the baseline and competitive methods on the CUB-200-2011 test set, as illustrated in Table 1. Our method surpasses all CNNs-based methods by a large margin in terms of Top-1/Top-5/GT-known *Loc.*, particularly with GoogLeNet, InceptionV3, and VGG16 backbones. Moreover, LocLoc achieves competitive performance compared to the transformer-based method, outperforming baseline PPD [29] by **5.6%**, LCTR [6] by **5.2%** in Top-1 *Loc.*. Table 3 reports the results in ILSVRC 2012 validation set. Besides, Table 3 illustrates the localization accuracy on the ILSVRC validation set. It shows that LocLoc surpasses the baseline PPD and over the SOTA LCTR by 2.6% and 1.5%. It is worth noting that although our method is lower than BAS [27], but our method is more capable than BAS when BAS use VGG16 backbone.

Classification. Table 2 and Table 4 show the classification accuracy on the CUB-200-2011 test set and ILSVRC 2012 validation set, respectively. Our proposed LocLoc method achieves 85.8%/97.0% in Top-1/Top-5 *Cls.* on the CUB-200-2011 dataset, outperforming the state-of-the-art LCTR [6] by 0.8% in terms of Top-1 *Cls.* Furthermore, on the ILSVRC 2012 validation set, LocLoc achieves 78.6%/94.0% in Top-1/Top-5 *Cls.*, surpassing all other methods. In addition, LocLoc achieves 1.5% improvement comparable with LCTR [6] on Top-1 *Cls.*. Overall, the proposed LocLoc method achieves state-of-the-art (SOTA) performance on both localization and classification tasks.

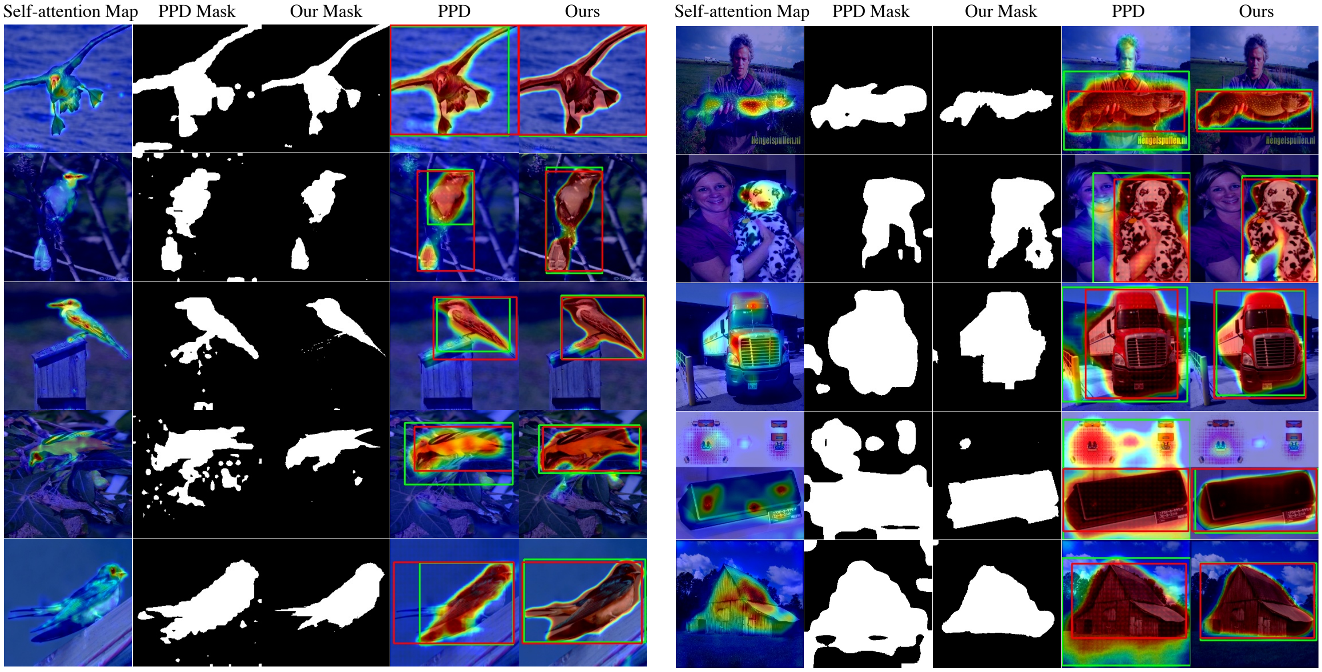
Table 1: Localization accuracy on dataset CUB-200-2011 test set. Best results are highlighted in bold, second-best are underlined.

Methods	Backbone	Loc. Acc		
		Top-1	Top-5	GT-known
CAM [39]	GoogLeNet	41.1	50.7	55.1
ACoL [34]	GoogLeNet	46.7	57.4	-
ADL [7]	InceptionV3	48.7	-	-
DANet [30]	InceptionV3	49.5	60.6	67.0
SPA [16]	InceptionV3	53.6	66.5	62.1
PSOL [33]	InceptionV3	65.5	83.4	-
SLT [11]	InceptionV3	66.1	-	86.5
I^{2}C [37]	InceptionV3	56.0	-	72.6
FAM [15]	InceptionV3	70.6	-	87.3
BAS [27]	InceptionV3	73.3	86.3	92.2
CAM [39]	VGG16	36.1	-	56.0
ACoL [34]	VGG16	45.9	56.5	63.0
ADL [7]	VGG16	52.4	-	74.0
DANet [30]	VGG16	52.5	-	-
SPG [36]	VGG16	49.0	57.2	58.9
MEIL [13]	VGG16	57.5	-	-
SPA [16]	VGG16	60.3	72.5	77.3
PSOL [33]	VGG16	66.3	84.1	-
SLT [11]	VGG16	67.8	-	87.6
FAM [15]	VGG16	69.3	-	89.3
BAS [27]	InceptionV3	71.3	85.3	91.1
PPD [29]	ViT-S	78.8	-	97.0
TS-CAM [9]	Deit-S	71.3	83.8	87.7
LCTR [6]	Deit-S	79.2	89.9	92.4
SCM [1]	Deit-S	76.4	91.6	96.6
Ours	Deit-S	**84.4**	**93.3**	**98.1**

Visualization. Figure 4 presents a visual comparison between the baseline PPD [29] and our proposed LocLoc on CUB-200-2011 and ILSVRC 2012 datasets. The first column displays the self-attention maps generated by DINO [3], which cover more object foreground, demonstrating the transformer's ability to capture long-range dependencies. The second and third columns represent PPD Mask and our GrabCut-Enhanced Mask, respectively, which are used as pseudo-labels for digging global representations. Our GrabCut-Enhanced mask evidently generates much sharper boundaries and explores more of the object regions with low-level information, compared to PPD, which applies a hard threshold on self-attention maps. The last two columns show the activation maps of PPD and LocLoc, with predicted bounding boxes (green color) and the ground-truth one (red color). The activation maps produced by LocLoc have sharper boundaries, benefiting from the GrabCut-Enhanced masks, and precisely localize object regions with predicted bounding boxes, outperforming PPD and showing the superiority of our proposed LocLoc.

4.3 Ablation Study

In this section, we conduct a series of ablation studies in our proposed LocLoc model on the CUB-200-2011 dataset.



(a) CUB-200-2011 dataset

(b) ILSVRC 2012 dataset

Figure 4: Visualization of comparison of baseline PPD [29] and LocLoc. The 1st column shows the Self-attention Map from DINO [3]. The 2nd column shows the PPD Mask by applying a hard threshold on Self-attention Map. The 3rd column displays our mask by GrabCut-Enhanced. The last two columns illustrate the localization map, ground truth bounding box (red color), and predicted bounding box (green color) of the baseline PPD and our LocLoc, respectively.

Table 2: Classification accuracy on dataset CUB-200-2011 test set. Best results are highlighted in bold, second-best are underlined.

Methods	Backbone	Cls. Acc	
		Top-1	Top-5
CAM [39]	GoogLeNet	73.8	91.5
DANet [30]	InceptionV3	74.6	90.6
ADL [7]	InceptionV3	74.6	-
SLT [11]	InceptionV3	76.4	-
CAM [39]	VGG16	76.6	92.5
ACoL [34]	VGG16	71.9	-
DANet [30]	VGG16	75.4	-
SPG [36]	VGG16	75.5	92.1
MEIL [13]	VGG16	74.8	-
SLT [11]	VGG16	76.6	-
PPD [29]	ViT-S	80.9	-
TS-CAM [9]	Deit-S	80.3	94.8
LCTR [6]	Deit-S	85.0	97.1
SCM [1]	Deit-S	78.5	94.5
Ours	Deit-S	**85.7**	**98.4**

Double thresholds on GrabCut-Enhancing. We explore the impact of δ_0 and δ_1, which partition the self-attention maps into three regions on GrabCut algorithm. Notably, we set δ_0 to 0.02, which effectively identifies almost all potential foreground regions. Then, we change δ_1 as illustrated in Figure 5, the performance first increases and then decreases as the δ_1 increases, because increasing $delta_1$, leads to a gradual reduction in the size of the GrabCut-enhanced mask, causing it to shrink from the edge of the object. Our classifier module LFDM mines local representations that rely on the class-specific area of the object center, thus minimizing the impact of the reduced non-discriminative area on classification performance. However, this reduction in mask size primarily affects object localization, which requires more global information. when δ_1 near to 0.75, we achieve best performance in Top-1 *Cls*, Top-1 and GT-known *Loc*.

Ablation studies of LFDM components. The experimental results of different components in the classifier are presented in Table 5. We are observing a significant improvement of 4.6% and 2.7% in Top-1 *Cls.* and Top-5 *Cls.*, respectively, when we replace the Linear Head with CNNs layers from LFDM, which demonstrates the ability of CNNs to capture more local features for classification based on the features extracted by the transformer, which is designed to capture long-range dependencies. Additionally, we find that incorporating the new augmentation (Aug.) based on GrabCut-Enhanced Mask in both Linear Head and CNNs classifier leads to increased classification accuracy. This suggests that low-level information is also beneficial for classification. Finally, we are employing CNN layers and the new augmentation technique as the Local Feature Digging Module (LFDM) for the classifier.

Table 3: Localization accuracy on dataset ILSVRC 2012 validation set. Best results are highlighted in bold, second-best are underlined.

Methods	Backbone	Loc. Acc		
		Top-1	Top-5	GT-known
ACoL [34]	GoogLeNet	46.7	57.4	-
DANet [30]	GoogLeNet	49.5	60.5	67.0
CAM [39]	InceptionV3	46.3	58.2	68.5
ADL [7]	InceptionV3	48.7	-	-
SPG [36]	InceptionV3	48.6	60.0	64.7
MEIL [13]	InceptionV3	48.8	-	-
SPA [16]	InceptionV3	52.7	64.3	68.3
I^{2}C [37]	InceptionV3	53.1	-	68.5
FAM [15]	InceptionV3	55.2	-	68.6
SLT [11]	InceptionV3	55.7	65.4	67.6
BAS [27]	InceptionV3	**58.5**	**69.0**	**71.9**
CAM [39]	VGG16	42.8	54.9	59.0
ACoL [34]	VGG16	45.8	59.4	63.0
ADL [7]	VGG16	44.9	-	-
MEIL [13]	VGG16	46.3	-	-
SPA [16]	VGG16	49.6	61.3	65.1
PSOL [33]	VGG16	50.9	60.9	64.0
SLT [11]	VGG16	51.2	52.4	67.2
FAM [15]	VGG16	52.0	-	71.7
BAS [27]	VGG16	53.0	65.4	69.6
PPD [29]	ViT-S	55.0	-	67.5
TS-CAM [9]	Deit-S	53.4	64.3	67.6
LCTR [6]	Deit-S	56.1	65.8	68.7
SCM [1]	Deit-S	56.1	66.4	68.8
Ours	Deit-S	57.6	67.2	70.0

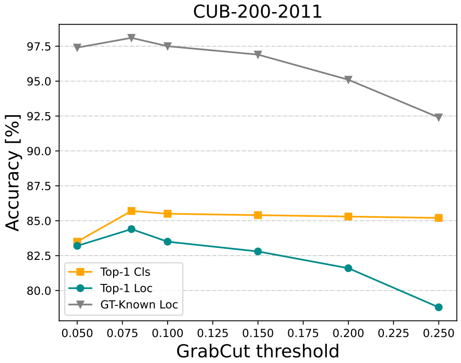


Figure 5: The impact of the threshold δ_1 in GrabCut-Enhanced Mask on CUB-200-2011 test set.

Table 4: Classification accuracy on dataset ILSVRC 2012 validation set. Best results are highlighted in bold, second-best are underlined.

Methods	Backbone	Cls. Acc	
		Top-1	Top-5
ACoL [34]	GoogLeNet	71.0	-
DANet [30]	GoogLeNet	63.5	91.4
CAM [39]	InceptionV3	73.3	91.8
ADL [7]	InceptionV3	72.8	-
I^{2}C [37]	InceptionV3	72.3	-
MEIL [13]	InceptionV3	73.3	-
FAM [15]	InceptionV3	77.6	-
SLT [11]	InceptionV3	78.1	-
CAM [39]	VGG16	68.8	88.6
ACoL [34]	VGG16	67.6	88.0
I^{2}C [37]	VGG16	69.4	89.3
MEIL [13]	VGG16	70.3	-
FAM [15]	VGG16	70.9	-
SLT [11]	VGG16	72.4	-
PPD [29]	ViT-S	76.9	-
TS-CAM [9]	Deit-S	74.3	92.1
LCTR [6]	Deit-S	77.1	93.4
SCM [1]	Deit-S	76.7	93.0
Ours	Deit-S	**78.6**	**94.0**

Table 5: The impact of classifier components on CUB-200-2011 test set.

	Baseline	Linear Head	CNNs	Aug.	Top-1 *Cls.*	Top-5 *Cls.*
(a)	✓	✓			80.9	95.1
(b)	✓	✓		✓	81.3	95.4
(c)	✓		✓		85.4	97.8
(d)	✓		✓	✓	**85.7**	**98.4**

5 CONCLUSION

In this paper, we propose a unified framework termed LocLoc for weakly supervised object localization, which leverages low-level cues present in images and feature representations with self-supervised ViT for object localization and classification. We first analyze the importance of low-level information and the superiority of transformers in capturing long-range dependencies. To exploit the low-level cues in images, we introduce the GrabCut Enhanced strategy based on self-attention maps from self-supervised ViT and design a GrabCut-Enhanced Generator (GEG) that uses the GrabCut Enhanced mask as pseudo labels to explore global localization representations. Moreover, we introduce a local feature digging module (LFDM) that combines low-level augmentation and mines local feature representations for classification. Extensive experiments on CUB-200-2011 and ILSVRC 2012 datasets demonstrate that our proposed LocLoc significantly outperforms baseline PPD, achieving state-of-the-art performance in both localization and classification.

6 ACKNOWLEDGMENTS

This work is supported by Key-Area Research and Development Program of Guangdong Province (No. 2021B0101400002), The Major Program of Guangdong Basic and Applied Research (No.2019B030302002) and The Major Key Project of PCL (PCL2021A13).

REFERENCES

[1] Haotian Bai, Ruimao Zhang, Jiong Wang, and Xiang Wan. 2022. Weakly Supervised Object Localization via Transformer with Implicit Spatial Calibration. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX (Lecture Notes in Computer Science, Vol. 13669)*. Springer, 612–628. https://doi.org/10.1007/978-3-031-20077-9_36

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 12346)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 213–229. https://doi.org/10.1007/978-3-030-58452-8_13

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 9630–9640. https://doi.org/10.1109/ICCV48922.2021.00951

[4] Tony F. Chan and Jianhong Shen. 2005. *Image processing and analysis - variational, PDE, wavelet, and stochastic methods*. SIAM. https://doi.org/10.1137/1.9780898717877

[5] Lulu Chen, Yongqiang Zhao, Jonathan Cheung-Wai Chan, and Seong G. Kong. 2022. Histograms of oriented mosaic gradients for snapshot spectral image description. *ISPRS Journal of Photogrammetry and Remote Sensing* 183 (2022), 79–93. https://doi.org/10.1016/j.isprsjprs.2021.10.018

[6] Zhiwei Chen, Changan Wang, Yabiao Wang, Guannan Jiang, Yunhang Shen, Ying Tai, Chengjie Wang, Wei Zhang, and Liujuan Cao. 2022. LCTR: On Awakening the Local Continuity of Transformer for Weakly Supervised Object Localization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 410–418. https://ojs.aaai.org/index.php/AAAI/article/view/19918

[7] Junsuk Choe, Seungho Lee, and Hyunjung Shim. 2021. Attention-Based Dropout Layer for Weakly Supervised Single Object Localization and Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 12 (2021), 4256–4271. https://doi.org/10.1109/TPAMI.2020.2999099

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

[9] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. 2021. TS-CAM: Token Semantic Coupled Attention Map for Weakly Supervised Object Localization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2866–2875. https://doi.org/10.1109/ICCV48922.2021.00288

[10] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. 2021. TS-CAM: Token Semantic Coupled Attention Map for Weakly Supervised Object Localization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2866–2875. https://doi.org/10.1109/ICCV48922.2021.00288

[11] Guangyu Guo, Junwei Han, Fang Wan, and Dingwen Zhang. 2021. Strengthen Learning Tolerance for Weakly Supervised Object Localization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 7403–7412. https://doi.org/10.1109/CVPR46437.2021.00732

[12] Kun He, Dan Wang, Miao Tong, and Zhijuan Zhu. 2020. An improved GrabCut on multiscale features. *Pattern Recognition* 103 (2020), 107292. https://doi.org/10.1016/j.patcog.2020.107292

[13] Jinjie Mai, Meng Yang, and Wenfeng Luo. 2020. Erasing Integrated Learning: A Simple Yet Effective Approach for Weakly Supervised Object Localization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 8763–8772. https://doi.org/10.1109/CVPR42600.2020.00879

[14] Jinjie Mai, Meng Yang, and Wenfeng Luo. 2020. Erasing Integrated Learning: A Simple Yet Effective Approach for Weakly Supervised Object Localization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 8763–8772. https://doi.org/10.1109/CVPR42600.2020.00879

[15] Meng Meng, Tianzhu Zhang, Qi Tian, Yongdong Zhang, and Feng Wu. 2021. Foreground Activation Maps for Weakly Supervised Object Localization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 3365–3375. https://doi.org/10.1109/ICCV48922.2021.00337

[16] Xingjia Pan, Yingguo Gao, Zhiwen Lin, Fan Tang, Weiming Dong, Haolei Yuan, Feiyue Huang, and Changsheng Xu. 2021. Unveiling the Potential of Structure Preserving for Weakly Supervised Object Localization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 11642–11651. https://doi.org/10.1109/CVPR46437.2021.01147

[17] Nikos Paragios, Yunmei Chen, and Olivier D. Faugeras (Eds.). 2006. *Handbook of Mathematical Models in Computer Vision*. Springer. https://doi.org/10.1007/0-387-28831-7

[18] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. Do Vision Transformers See Like Convolutional Neural Networks?. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 12116–12128. https://proceedings.neurips.cc/paper/2021/hash/652cf38361a209088302ba2b8b7f51e0-Abstract.html

[19] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.* 23, 3 (2004), 309–314. https://doi.org/10.1145/1015706.1015720

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[21] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* 128, 2 (2020), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[22] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 10347–10357. http://proceedings.mlr.press/v139/touvron21a.html

[23] Scott E. Umbaugh, Jeffrey Snyder, and Elena A. Fedorovskaya. 2011. Digital Image Processing and Analysis: Human and Computer Vision Applications with CVIPtools, Second Edition. *J. Electronic Imaging* 20, 3 (2011), 039901. https://doi.org/10.1117/1.3628179

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[25] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).

[26] Jun Wei, Qin Wang, Zhen Li, Sheng Wang, S. Kevin Zhou, and Shuguang Cui. 2021. Shallow Feature Matters for Weakly Supervised Object Localization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 5993–6001. https://doi.org/10.1109/CVPR46437.2021.00593

[27] Pingyu Wu, Wei Zhai, and Yang Cao. 2022. Background Activation Suppression for Weakly Supervised Object Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 14228–14237. https://doi.org/10.1109/CVPR52688.2022.01385

[28] Jinheng Xie, Jianfeng Xiang, Junliang Chen, Xianxu Hou, Xiaodong Zhao, and Linlin Shen. 2022. C^{2} AM: Contrastive learning of Class-agnostic Activation Map for Weakly Supervised Object Localization and Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 979–988. https://doi.org/10.1109/CVPR52688.2022.00106

[29] Jingyuan Xu, Hongtao Xie, Chuanbin Liu, and Yongdong Zhang. 2022. Proxy Probing Decoder for Weakly Supervised Object Localization: A Baseline Investigation. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 4185–4193. https://doi.org/10.1145/3503161.3547945

[30] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. 2019. DANet: Divergent Activation for Weakly Supervised Object Localization. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 6588–6597. https://doi.org/10.1109/ICCV.2019.00669

[31] Ke Yang, Dongsheng Li, and Yong Dou. 2019. Towards Precise End-to-End Weakly Supervised Object Detection Network. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 8371–8380. https://doi.org/10.1109/ICCV.2019.00846

[32] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 6022–6031. https://doi.org/10.1109/ICCV.2019.00612

[33] Chen-Lin Zhang, Yun-Hao Cao, and Jianxin Wu. 2020. Rethinking the Route Towards Weakly Supervised Object Localization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 13457–13466. https://doi.org/10.1109/CVPR42600.2020.01347

[34] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S. Huang. 2018. Adversarial Complementary Learning for Weakly Supervised Object Localization. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 1325–1334. https://doi.org/10.1109/CVPR.2018.00144

[35] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S. Huang. 2018. Adversarial Complementary Learning for Weakly Supervised Object Localization. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 1325–1334. https://doi.org/10.1109/CVPR.2018.00144

[36] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas S. Huang. 2018. Self-produced Guidance for Weakly-Supervised Object Localization. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII (Lecture Notes in Computer Science, Vol. 11216)*. Springer, 610–625. https://doi.org/10.1007/978-3-030-01258-8_37

[37] Xiaolin Zhang, Yunchao Wei, and Yi Yang. 2020. Inter-Image Communication for Weakly Supervised Localization. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIX (Lecture Notes in Computer Science, Vol. 12364)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 271–287. https://doi.org/10.1007/978-3-030-58529-7_17

[38] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. 2021. Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 6881–6890. https://doi.org/10.1109/CVPR46437.2021.00681

[39] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2921–2929. https://doi.org/10.1109/CVPR.2016.319

[40] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2921–2929. https://doi.org/10.1109/CVPR.2016.319

[41] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2921–2929. https://doi.org/10.1109/CVPR.2016.319

[42] C. Lawrence Zitnick and Piotr Dollár. 2014. Edge Boxes: Locating Object Proposals from Edges. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 8693)*, David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer, 391–405. https://doi.org/10.1007/978-3-319-10602-1_26