SALIENCY AWARE: WEAKLY SUPERVISED OBJECT LOCALIZATION

Yun-Chun Chen *Winston H. Hsu*

National Taiwan University, Taipei, Taiwan

ABSTRACT

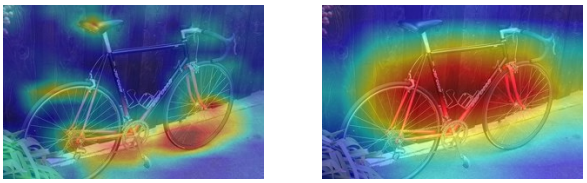
Object localization aims at localizing the object in a given image. Due to the recent success of convolutional neural networks (CNNs), existing methods have shown promising results in weakly-supervised learning fashion. By training a classifier, these methods learn to localize the objects by visualizing the class discriminative localization maps based on the classification prediction. However, correct classification results would not guarantee sufficient localization performance since the model may only focus on the most discriminative parts rather than the entire object. To address the aforementioned issue, we propose a novel and end-to-end trainable network for weakly-supervised object localization. The key insights to our algorithm are two-fold. First, to encourage our model to focus on detecting foreground objects, we develop a *salient object detection* module. Second, we propose a *perceptual triplet loss* that further enhances the foreground object detection capability. As such, our model learns to predict objectness, resulting in more accurate localization results. We conduct experiments on the challenging ILSVRC dataset. Extensive experimental results demonstrate that the proposed approach performs favorably against the state-of-the-arts.

***Index Terms*—** Weakly-supervised object localization

1. INTRODUCTION

Object localization is an active research topic in computer vision that aims at predicting the boundaries of the object in a given image. With a wide range of applications ranging from, object segmentation [1, 2], saliency detection [3, 4], object detection [5], semantic matching [6], to person re-identification [7], object localization has received substantial attention from academia. Nevertheless, due to the presence of background clutter, occlusion, and large intra-class appearance variations, object localization remains a challenging task.

With the advances of convolutional neural networks (CNNs), numerous CNN-based methods are proposed to tackle the aforementioned issues. While directly training a bounding box estimator could result in superior localization performance, the dependency of manual supervision would restrict the method's scalability. In addition, collecting large-scale datasets with manually annotated bounding boxes, however, is expensive and labor-intensive. To address this issue, several weakly-supervised methods [8, 9, 10] are proposed. Using only weak image-level supervision (i.e., class labels) to train a classifier, these methods are capable of localizing objects in the given images by utilizing the class activation maps (CAMs). While promising results have been shown, these methods still suffer from the following limitation. Training a classifier that can correctly classify objects does not guarantee sufficient localization performance since the model may only localize the most discriminative parts of an object to achieve correct classification results.



(a) Existing methods. (b) Ours.

Fig. 1: **Visual comparisons.** (a) Existing methods may only localize the most discriminative part (e.g., the wheels of the bicycle) when performing image classification. (b) Our proposed approach takes into account the salient object detection, resulting in more accurate localization result.

In this paper, we propose an end-to-end trainable network for weakly-supervised object localization. To encourage our model to focus on detecting foreground objects, we develop a salient object detection module. As shown in Fig. 1, with the integration of the salient object detection module, our model learns to detect foreground regions while performing classification, resulting in more accurate localization results. To enhance the foreground detection ability, we propose a *perceptual triplet loss*. We evaluate the effectiveness of our approach on the ILSVRC dataset. Extensive experimental results show that the proposed approach performs favorably against the state-of-the-art methods.

The contributions of this paper are highlighted as follows.

- We propose an end-to-end trainable network for weakly-supervised object localization.
- We propose a perceptual triplet loss which further improves the localization capability.
- Extensive comparisons with existing algorithms demonstrate that the proposed approach achieves the state-of-the-art performance.

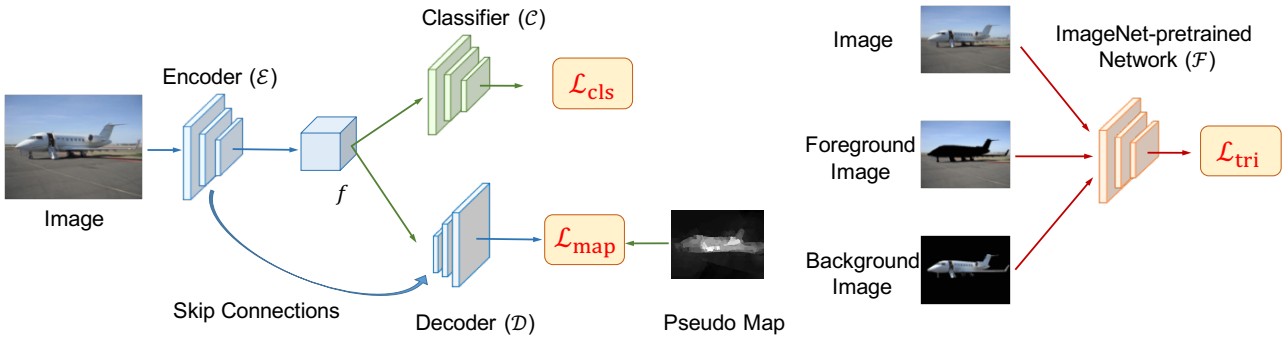


Fig. 2: **Overview of the proposed model**. Our model is composed of four sub-networks, including an encoder $\mathcal{E}$ (for feature extraction), a classifier $\mathcal{C}$ (for performing image classification), a decoder $\mathcal{D}$ (for detecting salient objects), and an ImageNet-pretrained network $\mathcal{F}$ (for enhancing the capability of salient object detection). The model training is driven by the classification loss $\mathcal{L}_{\text{cls}}$, the map loss $\mathcal{L}_{\text{map}}$, and the perceptual triplet loss $\mathcal{L}_{\text{tri}}$.

2. THE PROPOSED APPROACH

In this section, we first present an overview of the proposed approach. We then describe the details of the proposed framework for weakly-supervised object localization.

2.1. Algorithmic overview

Let $X = \{x_i\}_{i=1}^N$ be a set of images and denote its corresponding label set as $Y = \{y_i\}_{i=1}^N$, where $x_i \in \mathbb{R}^{H\times W\times 3}$ and $y_i \in \mathbb{R}$ represent the i^{th} image and its corresponding class label, respectively. Our goal is to learn a CNN-based model that can produce accurate object localization results by performing a classification task using only weak image-level supervision (i.e., class label). This task is referred to as weakly-supervised object localization since no ground truth bounding box annotations are used during training.

Network. As shown in Fig. 2, our model comprises four sub-networks: the encoder $\mathcal{E}$, the decoder $\mathcal{D}$, the classifier $\mathcal{C}$, and the ImageNet-pretrained network $\mathcal{F}$. To achieve weakly-supervised object localization, the encoder $\mathcal{E}$ learns to extract semantic feature maps. The classifier $\mathcal{C}$ takes the extracted feature maps as inputs to perform image classification. To encourage our model to focus on detecting foreground object, the decoder $\mathcal{D}$ learns to predict objectness based on the extracted feature maps. To further improve the capability of detecting salient objects, the proposed perceptual triplet loss $\mathcal{L}_{\text{tri}}$ is introduced to (1) minimize the appearance discrepancy between the input image and its foreground image while (2) maximizing the dissimilarity between the foreground image and the background image.

As for testing, the decoder $\mathcal{D}$ and the ImageNet-pretrained network $\mathcal{F}$ are discarded (i.e., only the encoder $\mathcal{E}$ and the classifier $\mathcal{C}$ are left). Thus, our proposed framework does not have additional computational overhead during inference. Given an input image, we forward it to the encoder $\mathcal{E}$ to extract its feature map. We then pass the feature map to the classifier $\mathcal{C}$. Based on the classification prediction, we obtain the corresponding class activation map for the input image. In the following, we elaborate each component in details.

2.2. CAM for weakly-supervised object localization

Encoder $\mathcal{E}$. Given an input image $x \in X$[1], we pass x to the encoder $\mathcal{E}$ to obtain its feature map $f = \mathcal{E}(x) \in \mathbb{R}^{h\times w\times d}$, where d denotes the number of channels.

Classifier $\mathcal{C}$. To perform classification using labeled information from training data, we introduce a classifier $\mathcal{C}$. The input of the classifier $\mathcal{C}$ is the feature vector v from the global average pooling (GAP) layer on the extracted feature map f, i.e., $v = \text{GAP}(f) \in \mathbb{R}^d$. Specifically, we define the classification loss $\mathcal{L}_{\text{cls}}$ using the labeled data to train the classifier $\mathcal{C}$ and the encoder $\mathcal{E}$ as

$$\mathcal{L}_{\text{cls}}(X, Y; \mathcal{E}, \mathcal{C}) = -\mathbb{E}_{(x,y)\sim(X,Y)} \sum_{k=1}^{K} \hat{y}_k \log(\tilde{y}_k), \quad (1)$$

where $\tilde{y} = \mathcal{C}(f) \in \mathbb{R}^K$ is the classification prediction, $\hat{y} \in \mathbb{R}^K$ is the ground truth one-hot vector, and K is the number of classes.

We note that weighted classification loss proposed by Chen *et al.* [11] can also be adopted to improve the image classification performance.

2.3. The proposed salient object detection module

To encourage our model to focus on localizing foreground object, we develop a salient object detection module that consists of two network components, including a decoder $\mathcal{D}$ and an ImageNet-pretrained network $\mathcal{F}$.

Decoder $\mathcal{D}$. To allow our model to focus on detecting foreground object, we introduce a decoder $\mathcal{D}$ that takes in the extracted feature map f and learns to predict objectness. To achieve this, we introduce a map loss $\mathcal{L}_{\text{map}}$. Our key insight is that encouraging our model to predict objectness would result

[1]We often omit the subscript i, denote input image as x, and represent its corresponding class label as y for simplicity.

in better localization capability. To achieve this, we generate a pseudo saliency map for each input image by averaging a set of object proposals generated by an off-the-shelf object proposal algorithm, e.g., *unsupervised* geodesic object proposals [12] used in this work. Using the pseudo saliency map M^p for input image x, we define the map loss $\mathcal{L}_{\text{map}}$ as

$$\mathcal{L}_{\text{map}}(X;\mathcal{E},\mathcal{D}) = \|M^p - \mathcal{D}(\mathcal{E}(x))\|_2, \tag{2}$$

where the map loss $\mathcal{L}_{\text{map}}$ is defined by using the L2 norm.

ImageNet-pretrained network $\mathcal{F}$. To further enhance the foreground object detecting capability, we propose the perceptual triplet loss $\mathcal{L}_{\text{tri}}$ by introducing an auxiliary ImageNet-pretrained network $\mathcal{F}$. The perceptual triplet loss $\mathcal{L}_{\text{tri}}$ enhances the quality of the saliency prediction produced by the decoder $\mathcal{D}$ based on two criteria: (1) low input image and foreground image distinctness and (2) high foreground image and background image discrepancy. As shown in Fig. 2, the decoder $\mathcal{D}$ produces saliency map $\mathcal{D}(x)$ for each input image x. With the saliency map $\mathcal{D}(x)$, we can generate the *foreground image* x^f and the *background image* x^b for image x. Namely,

$$x^f = \mathcal{D}(x) \otimes x, \text{and} \tag{3}$$

$$x^b = (1 - \mathcal{D}(x)) \otimes x, \tag{4}$$

where $\otimes$ represents the pixel-wise multiplication between the two operands.

We then apply an ImageNet-pretrained network $\mathcal{F}$ to x, x^f and x^b and extracts their semantic feature vectors $\mathcal{F}(x)$, $\mathcal{F}(x^f)$, and $\mathcal{F}(x^b)$, respectively. The perceptual triplet loss $\mathcal{L}_{\text{tri}}$ is then defined as

$$\mathcal{L}_{\text{tri}}(X;\mathcal{E},\mathcal{D},\mathcal{F}) = d_{\text{neg}} - d_{\text{pos}}, \tag{5}$$

$$d_{\text{pos}} = \frac{1}{n}\|\mathcal{F}(x) - \mathcal{F}(x^o)\|_2^2, \text{ and} \tag{6}$$

$$d_{\text{neg}} = \frac{1}{n}\|\mathcal{F}(x^o) - \mathcal{F}(x^b)\|_2^2, \tag{7}$$

where constant $n = 2,048$ is the dimension of the semantic features produced by the ImageNet-pretrained network $\mathcal{F}$.

2.4. Full objective

Overall, the full objective for training the proposed model can be expressed as

$$\begin{aligned}\mathcal{L}(X,Y;\mathcal{E},\mathcal{C},\mathcal{D},\mathcal{F}) = \mathcal{L}_{\text{cls}}(X,Y;\mathcal{E},\mathcal{C}) \\ + \mathcal{L}_{\text{map}}(X;\mathcal{E},\mathcal{D}) \\ + \mathcal{L}_{\text{tri}}(X;\mathcal{E},\mathcal{D},\mathcal{F}).\end{aligned} \tag{8}$$

2.5. Implementation details

We implement our model using PyTorch. For the encoder $\mathcal{E}$ and the classifier $\mathcal{C}$, we use the same modified AlexNet (i.e., AlexNet-GAP) and GoogLeNet (i.e., GoogLeNet-GAP) as introduced in [8]. For the decoder $\mathcal{D}$, it contains five blocks, each of which is composed of one deconvolutional layer and two convolutional layers. We add skip connections between each block of the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$ to facilitate saliency prediction and encourage more efficient gradient propagation. The ImageNet-pretrained network $\mathcal{F}$ is the ResNet-50 and is fixed during the course of training. The encoder $\mathcal{E}$, the classifier $\mathcal{C}$, and the decoder $\mathcal{D}$ are all randomly initialized. We apply data augmentation techniques by random cropping and horizontal flipping. The batch size is set to 32. We train our model for 100 epochs using the Adam optimizer [13] with learning rate 1×10^{-2}.

3. EXPERIMENTS

We describe the experimental settings for evaluating weakly-supervised object localization in this section.

3.1. Settings

We evaluate our approach on a standard benchmark: the ILSVRC dataset [14]. We evaluate the effectiveness of the proposed approach using two standard CNNs: the AlexNet [15] and the GoogLeNet [16]. Following Zhou *et al.* [8], we remove the fully-connected layers before the output and replace them with a global average pooling layer and a fully-connected layer with softmax serves as the activation function.

3.2. Evaluation protocol

We evaluate the proposed approach on two tasks: object localization and image classification.

Localization. For weakly-supervised object localization, we adopt the `GT-known Loc` and the `Top-1 Loc` as the evaluation metrics. For `GT-known Loc`, we obtain the corresponding class activation map with respect to the classification ground truth. For the `Top-1 Loc`, we obtain the corresponding class activation map with respect to the top-1 classification result. With the class activation map, we generate the bounding box by using the method proposed in [8].

Classification. For image classification, we compute the top-1 classification accuracy denoted as `Top-1 Class`.

3.3. Experimental results using AlexNet.

We compare our proposed approach with two existing weakly-supervised object localization methods: AlexNet-GAP [8] (adopts the CAM [8] method) and AlexNet-HaS [10] (adopts the Grad-CAM [8] method). As shown in Table 1, our method achieves 67.23% in `GT-known Loc`, 44.16% in `Top-1 Loc`, and 62.01% in `top-1 Class`. The proposed method

Table 1: Experimental results on the ILSVRC dataset [14] using the AlexNet [15] architecture. The bold numbers indicate the best results.

Method	GT-known Loc (%)	Top-1 Loc (%)	Top-1 Class (%)
AlexNet-GAP [8]	54.90	36.25	60.23
AlexNet-HaS [10]	58.68	37.65	58.68
Ours w/o $\mathcal{L}_{\text{map}}$	59.30	38.78	60.25
Ours w/o $\mathcal{L}_{\text{tri}}$	60.23	39.98	60.44
Ours	**67.23**	**44.16**	**62.01**

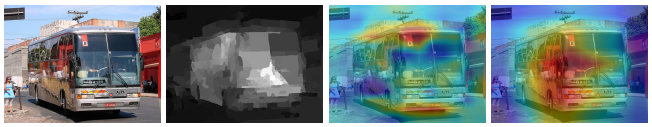


Image Pseudo Map AlexNet-GAP Ours

Fig. 3: **Visual comparisons.**. We present the qualitative results using the AlexNet [15] architecture. We observe that our method produces more accurate localization result.

performs favorably against the state-of-the-art methods, outperforming the previous best competitor [10] by 8.55% in GT-known Loc, 6.51% in Top-1 Loc, and 3.33% in top-1 Class. Our performance gain can be ascribed to the following factor. Unlike most existing methods that only learns a classifier, our method further takes into account salient object detection. With the proposed salient object detection module, our model learns to focus on detecting foreground objects, resulting in better localization performance.

Ablation study. To analyze the importance of the proposed components (i.e., the map loss $\mathcal{L}_{map}$ and the perceptual triplet loss $\mathcal{L}_{tri}$), we conduct an ablation study using AlexNet [15]. As shown in Table 1, without the map loss $\mathcal{L}_{map}$, our model suffers 7.93% and 5.38% performance drops in GT-known Loc and Top-1 Loc, respectively. On the other hand, if our model is trained without the perceptual triplet loss $\mathcal{L}_{tri}$, our model suffers 7.00% and 4.18% performance drops in GT-known Loc and Top-1 Loc, respectively. These results indicate that while our model is still able to reliably perform image classification, without either of the proposed loss functions, the proposed model is not able to accurately localize the foreground object. We note that if both of the proposed losses are turned off, our results will be reduced to those of AlexNet-GAP [8] since our encoder and classifier are the same as their model. The ablation experiments show that all the proposed loss terms play crucial roles in achieving the state-of-the-art performance.

3.4. Experimental results using the GoogLeNet.

We also evaluate the effectiveness of the proposed approach using the GoogLeNet [16]. Table 2 reports the experimental results. Our method achieves 66.23% in GT-known Loc, 51.30% in Top-1 Loc, and 74.56% in top-1 Class. The proposed method performs favorably against the state-of-the-art methods and outperforms the previous best competitor [10] by 6.3% in GT-known Loc, 6.52% in Top-1 Loc, and 4.19% in top-1 Class. We also observe that without either the map loss $\mathcal{L}_{map}$ or the perceptual triplet loss $\mathcal{L}_{tri}$, our model suffers significant performance drops in terms of the localization performance.

Table 2: Experimental results on the ILSVRC dataset [14] using the GoogLeNet [16] architecture. The bold numbers indicate the best results.

Method	GT-known Loc (%)	Top-1 Loc (%)	Top-1 Class (%)
GoogLeNet-GAP [8]	58.41	43.60	71.95
GoogLeNet-HaS [10]	59.93	44.78	70.37
Ours w/o $\mathcal{L}_{\text{map}}$	61.13	46.01	71.90
Ours w/o $\mathcal{L}_{\text{tri}}$	62.74	46.47	72.03
Ours	**66.23**	**51.30**	**74.56**

Fig. 4: **Qualitative results.**. We present more visual results. We observe that the proposed salient object detection module effectively encourages our model to accurately localize objects with only weak image-level supervision.

With these experiments, we confirm that the unique design of the proposed salient object detection module as well as the proposed loss terms allow our model to simultaneously focus on detecting the foreground objects while performs image classification. With only weak image-level supervision (i.e., class label), our model achieves the state-of-the-art performances.

4. CONCLUSIONS

We have presented a weakly-supervised and end-to-end trainable network for object localization. The core technical novelty of our approach lies in the integration of a *salient object detection* module to encourage our model to predict objectness while performing image classification. To further enhance the foreground object capability, we propose the perceptually triplet loss which minimizes the xxx while maximizes the figure-ground distinctness. We further investigate the multi-scale architecture which effectively improves the localization capability. Our network training requires only weak image-level supervision and thus significantly alleviates the cost of constructing and labeling large-scale training datasets. Experimental results demonstrate that our approach performs favorably against existing weakly-supervised object localization algorithms on one standard benchmark. We hope that our method could facilitate other vision tasks such as object segmentation, saliency detection, and object detection.

Acknowledgement. This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2634-F-002-004. We also benefit from the NVIDIA grants and the DGX-1 AI Supercomputer.

5. REFERENCES

[1] Thomas Brox and Jitendra Malik, "Object segmentation by long term analysis of point trajectories," in *ECCV*, 2010.

[2] Jhih-Yuan Lin, Min-Sheng Wu, Yu-Cheng Chang, Yun-Chun Chen, Chao-Te Chou, Chun-Ting Wu, and Winston H. Hsu, "Learning volumetric segmentation for lung tumor," *IEEE ICIP VIP Cup Tech. report*, 2018.

[3] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li, "Salient object detection: A benchmark," *TIP*, 2015.

[4] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li, "Salient object detection: A discriminative regional feature integration approach," in *CVPR*, 2013.

[5] Ross Girshick, "Fast r-cnn," in *ICCV*, 2015.

[6] Yun-Chun Chen, Po-Hsiang Huang, Li-Yu Yu, Jia-Bin Huang, Ming-Hsuan Yang, and Yen-Yu Lin, "Deep semantic matching with foreground detection and cycle-consistency," in *ACCV*, 2018.

[7] Yun-Chun Chen, Yu-Jhe Li, Xiaofei Du, and Yu-Chiang Frank Wang, "Learning resolution-invariant deep representations for person re-identification," in *AAAI*, 2019.

[8] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.

[9] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization.," in *ICCV*, 2017.

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

[11] Yun-Chun Chen, Yu-Jhe Li, Aragorn Tseng, and Tsungnan Lin, "Deep learning for malicious flow detection," in *IEEE PIMRC*, 2017.

[12] Philipp Krähenbühl and Vladlen Koltun, "Geodesic object proposals," in *ECCV*, 2014.

[13] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv*, 2014.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *IJCV*, 2015.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[16] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.