# Generalized Weakly Supervised Object Localization

Dingwen Zhang, *Member, IEEE*, Guangyu Guo, Wenyuan Zeng, Lei Li, and Junwei Han, *Fellow, IEEE*

***Abstract*— With the goal of learning to localize specific object semantics using the low-cost image-level annotation, weakly supervised object localization (WSOL) has been receiving increasing attention in recent years. Although existing literatures have studied a number of major issues in this field, one important yet challenging scenario, where the test object semantics may appear in the training phase (seen categories) or never been observed before (unseen categories), is still beyond the exploration of the existing works. We define this scenario as the generalized WSOL (GWSOL) and make a pioneering effort to study it in this article. By leveraging attribute vectors to associate seen and unseen categories, we involve threefold modeling components, i.e., the class-sensitive modeling, semantic-agnostic modeling, and content-aware modeling, into a unified end-to-end learning framework. Such design enables our model to recognize and localize unconstrained object semantics, learn compact and discriminative features that could represent the potential unseen categories, and customize content-aware attribute weights to avoid localizing on misleading attribute elements. To advance this research direction, we contribute the bounding-box manual annotations to the widely used AwA2 dataset and benchmark the GWSOL methods. Comprehensive experiments demonstrate the effectiveness of our proposed learning framework and each of the considered modeling components.**

***Index Terms*— Object localization, unseen object category, weakly supervised learning.**

## I. Introduction

DEEP learning models have been playing critical roles in the modern computer vision community. By well-fitting the mature technology background in big data, internet, and high-quality computational devices, deep learning models have been widely used in various scenarios to solve real-world vision challenges. Although many promising results have been achieved, one important issue that has not been fully addressed is the label-hunger nature of deep learning models. In other words, the current mainstream deep learning models heavily rely on the manual annotations to accomplish the specific vision tasks. However, due to the inconvenience and inefficiency in acquiring the manual annotations in practice, deep models working under the coarse (i.e., weak) image-level annotation are highly desirable [1], [2], [3], [4], [5].

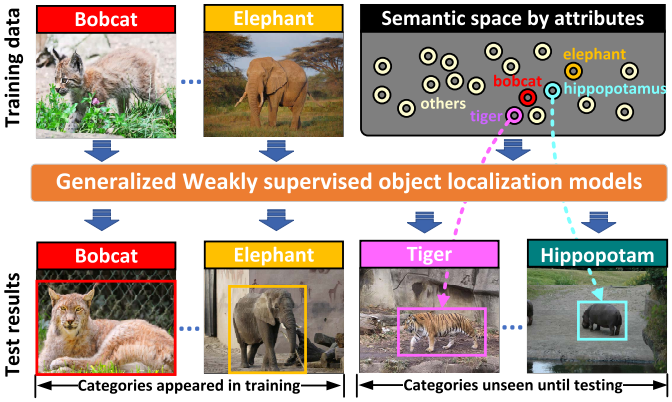



Fig. 1. Illustration of the newly explored GWSOL task. As can be seen, in GWSOL, we leverage the sematic space that is formed by attribute vectors to associate the unseen categories that would appear in testing with the seen categories appearing in the training phase, and finally localize both the seen objects and unseen ones in testing.

To accomplish the sematic-oriented localization task with low label cost, weakly supervised object localization (WSOL) methods are proposed [6], [7], [8], [9], where the models use the image-level manual annotation to automatically learn the potential locations for different object semantics. For achieving the promising WSOL performance, great efforts have been made in the past years and methods equipped with various region mining mechanisms, e.g., the complementary and discriminative visual pattern mining [10], erasing integrated learning [11], attention-based dropout [12], etc., appeared.

Among the existing literature, the core problem is how to localize the complete semantic-specific object regions that have been observed from the training data. While in this work, we study a new WSOL scenario, where the object semantics to be localized in testing may either appear in the training phase or never been observed before (see Fig. 1). We define this scenario as the generalized WSOL (GWSOL) and it turns out that learning under this scenario would be more suitable for real-world cases as one can hardly guarantee that all the object categories presenting in the test phase have been collected in the training data.

Compared with the common WSOL task, the GWSOL task works under the less constrained testing scenario, which makes

Manuscript received 19 August 2021; revised 1 May 2022; accepted 15 August 2022. This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2021B0101200001, in part by the National Science Foundation of China under Grant 61876140 and Grant U21B2048, and in part by the Open Research Projects of Zhejiang Laboratory under Grant 2019KD0AD01/010. *(Corresponding authors: Lei Li; Junwei Han.)*

Dingwen Zhang is with the Brain and Artificial Intelligence Laboratory, School of Automation, Northwestern Polytechnical University, Xi'an 710129, China, and also with the Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, Hefei 230026, China (e-mail: zhangdingwen2006yyy@gmail.com).

Guangyu Guo and Wenyuan Zeng are with the Brain and Artificial Intelligence Laboratory, School of Automation, Northwestern Polytechnical University, Xi'an 710129, China.

Lei Li is with the Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing 100074, China (e-mail: 381205977@qq.com).

Junwei Han is with the Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, Hefei 230026, China (e-mail: junweihan2010@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3204337.

Digital Object Identifier 10.1109/TNNLS.2022.3204337

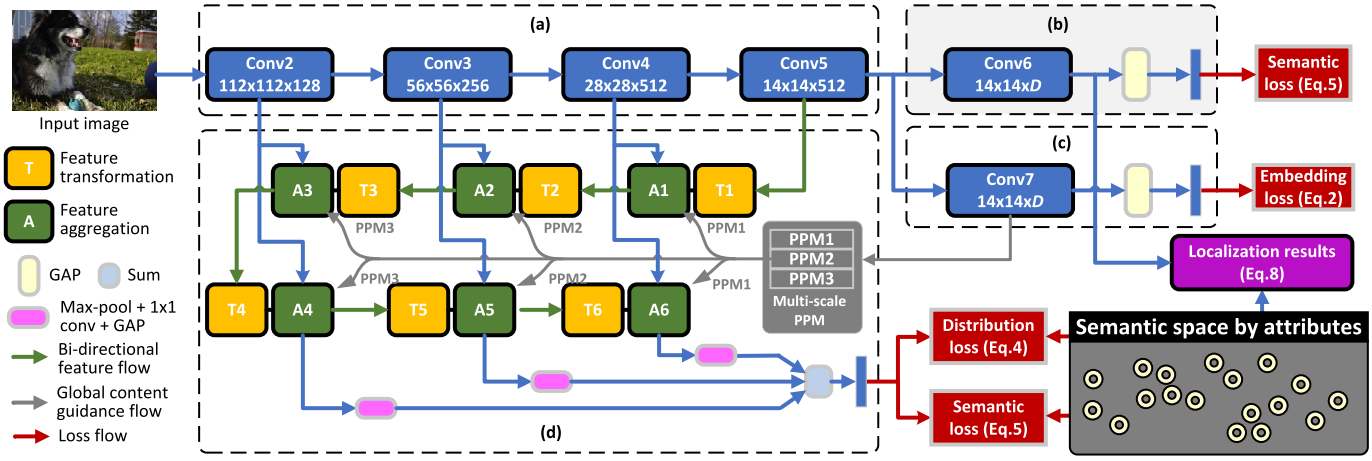



Fig. 2. Illustration of the proposed framework for GWSOL. It mainly contains four parts, including (a) feature extraction backbone, (b) classification-localization head, (c) semantic-agnostic embedding head, and (d) content-based discriminative attribute mining module. The losses are computed under the guidance of the attribute vectors. In testing, the localization maps are generated from Conv6. *D* indicates the dimension of the attribute vector.

the learning procedure much more challenging. In GWSOL, we can leverage category attributes to associate seen categories with unseen ones and address the following three key issues: 1) How to encode the visual features of the input images to the semantic space formed by the attribute vectors? 2) How to learn effective feature representations that can better generalize on the data with unseen categories? and 3) how to customize content-aware attribute weights to avoid localizing on misleading attribute elements.[1]

To deal with the aforementioned issues, we reveal threefold modeling for GWSOL. The first one is the class-sensitive modeling, which maps the input image to the semantic space built by either user-defined attribute annotations [13] or the unsupervised semantic attributes (word2vec [14], GloVe [15]).[2] Classification and localization processes performed in such a semantic space would associate the unseen object categories to the object categories explored in the training phase. The second one is the semantic-agnostic modeling, which aims at learning compact yet discriminative features for examples from each arbitrary semantic. As demonstrated in [16], such feature embedding would contribute to representing the potential unseen object categories. The third one is the content-aware modeling, which re-weights the attribute elements of each category attribute vector by mining patterns that correspond to the content of each specific input image.

Based on the above discussion, we build a novel deep learning framework to involve the threefold modeling components, i.e., the class-sensitive modeling, semantic-agnostic modeling, and content-aware modeling, into a unified end-to-end learning framework to solve the GWSOL problem. The concrete framework is shown in Fig. 2. As can be seen, the proposed learning framework mainly contains four parts. The first part is the network backbone, which is used to perform the basic feature extraction. The second part is the classification-localization head, which learns the attribute activation features to perform the class-sensitive classification and localization. The third part is the semantic-agnostic embedding head, which projects the input image to a high-dimensional, representative, yet nonsemantically oriented embedding space. The last part is the content-based attribute reweighting module, which uses the global image representation to guide a bidirectional feature flow to mine the discriminative attributes that correspond to the specific image content.

It is also worth mentioning that the class-activation mapping (CAM) [6] is a key technique in the conventional WSOL task and our method is also built upon it. However, as witnessed by recent WSOL methods (e.g., ADL [12] and Cutmix [17]), another nonnegligible factor to determine the final localization performance is the constructed **feature space**, while the threefold modeling proposed in this work is precisely aimed at constructing the effective feature space in GWSOL scenario. Perhaps these designs are a bit like those used in the classification tasks. However, this is due to that under the weakly supervised learning scenario, we can only use image labels as supervision to learn the feature space, making the whole framework **classification-driven**. The experimental comparison in Section IV verifies this by showing that our approach can achieve around 10% performance gains over the conventional WSOL methods, even though these methods are also equipped with CAM.

To sum up, this article mainly has the following threefold contributions.

1) We make one of the earliest efforts to study the GWSOL problem. GWSOL performs WSOL under a practical yet challenging scenario, where the test images contain both the object categories appearing in the training process and those never been observed until testing.
2) We reveal the key issues in GWSOL and integrate the class-sensitive modeling, semantic-agnostic modeling, and content-aware modeling into a unified and end-to-end deep learning framework.

[1]For example, for an image with a desert background, the attribute element of *color yellow* would mislead the localization of camels.

[2]It is worth mentioning that the attribute annotations have also been used in the conventional WSOL approaches, such as [10].

3) We provide the bounding-box manual annotation to the widely used AwA2 dataset and benchmark the GWSOL methods on AwA2 and CUB datasets. Comprehensive experiments demonstrate the effectiveness of our proposed learning framework and each of the considered modeling components.

## II. Related Works

### A. Weakly Supervised Object Localization

WSOL is a widely studied problem in recent years. Existing efforts mainly pursue high-quality localization results for the object categories collected in the training data. Early methods, e.g., [18], [19], [20], [21], address this problem by using either hand-crafted features or pretrained CNN features together with multiple computational stages. Later, in light of the discriminative localization capacity of the class activation mapping technique [6], recent works have developed a series of end-to-end deep learning models to automatically mine object regions that are associated with certain semantics under a classification framework [7], [22], [23]. The main challenge under this framework is how to deal with the learned compact classification-oriented features to avoid partial localization. To deal with this problem, Wei *et al.* [24] propose an adversarial erasing approach. By progressively erasing the discriminant parts, this method can sequentially discover complement object regions. Following this work, Hou *et al.* [25] introduce background prior to the learning process to prevent mined object regions from covering the undesired image background. Besides the adversarial erasing strategy, dropout attention-based strategy also obtains good localization performance but with even lower computational costs. Specifically, Choe and Shim [12] utilize the self-attention mechanism to process the feature maps and manually drops out the high attention regions from the attention map. Following this work, Mai *et al.* [11] jointly learn two parallel network branches, i.e., one with dropout attention and the other without, extract features for localization and classification separately. Recently, Xue *et al.* [10] propose to explore divergent activation for WSOL. Xie *et al.* [26] generate the localization maps by refining low-level feature maps of the deep network. Kim *et al.* [27] propose a normalization method for refining the class activation maps. Meng *et al.* [28] jointly optimize classification and object localization by foreground activation maps.

Instead of studying the conventional WSOL, this work promotes this community to a new research direction. Specifically, we focus on a more realistic and challenging scenario where objects need to be localized are not all be seen before. Performing object localization in such a scenario can not only reduce the annotation costs but also eliminate the constraint in testing.

### B. Zero-Shot Learning

Zero-Shot Learning aims to learn to generate a unified feature space that associates the visual representations extracted from the input images (or regions) and the semantic concepts characterized by attribute vectors. With such a feature space, we can predict the probability of any semantic concepts (including the seen categories and the unseen ones) with the given attribute vectors, thus achieving zero-shot image classification [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], zero-shot object detection [39], [40], and zero-shot semantic segmentation [41], [42], [43], [44]. Chen *et al.* [45] disentangle the visual features into the semantic-consistent and semantic-unrelated latent vectors. Chen *et al.* [46] utilize a feature refinement module to refine the visual features of seen and unseen class samples. Although great efforts have been made to promote the aforementioned zero-shot learning frameworks in recent years, the zero-shot image classification methods still do not have the capacity to localize concrete object regions from each input image while the zero-shot object detection or semantic segmentation methods always requires the labor-consuming manual annotation on the specific object regions of the training data.

From this perspective, this work also makes an attempt to fill the blank in the zero-shot learning community—the investigated task can achieve the similar purpose with the zero-shot object detection or semantic segmentation methods (i.e., discovering the region-level unseen semantics) but only requires the image-level annotation like in zero-shot image classification.

## III. Proposed Approach

### A. Problem Formulation

In GWSOL, the training set contains $N^{\text{tr}}$ images with image-level labels $y^{\text{tr}} \in \mathcal{Y}^{\mathcal{S}}$, where $\mathcal{Y}^{\mathcal{S}}$ indicates the categories of the seen objects. In addition, we are also given a testing set that contains $N^{te}$ images with unknown object categories and such objects belong to $y^{te} \in \mathcal{Y}^{\mathcal{S}} \cup \mathcal{Y}^{\mathcal{U}}$, where $\mathcal{Y}^{\mathcal{U}}$ indicates the unseen object categories. Note that $\mathcal{S}$ and $\mathcal{U}$ are disjoint, i.e., $\mathcal{S} \cap \mathcal{U} = \emptyset$. For each object category, we have a corresponding $D$-dimensional attribute vector $\mathbf{a}$, which is obtained either by user annotation [13] or unsupervised semantic representation [14], [15]. From Fig. 3, we can observe that in the semantic space generated by attribute vectors, object categories with similar semantics would be close to each other, e.g., *chihuahua* and *siamese cat*, while those categories with dissimilar semantics would be far from each other, e.g., *dolphin* and *tiger*. Consequently, using the attribute vectors, we can associate unseen categories with seen categories even though we actually have not observed any of the unseen categories before [34]. Finally, the goal of the GWSOL model is learning to map the input image to the desired feature space, where the features can finely align with the one formed by the attribute vectors and in the meantime discover useful hints to highlight the objects of interest. To archive this goal, we introduce the following threefold modeling components.

### B. Class-Sensitive Modeling

One basic component in GWSOL is how to map the input image to semantic spaces that can fit both seen and unseen object categories. To achieve this goal, we introduce a classification-localization head (CLH) [see Fig. 2(b)]. CLH basically has a $3 \times 3$ convolutional layer to map the

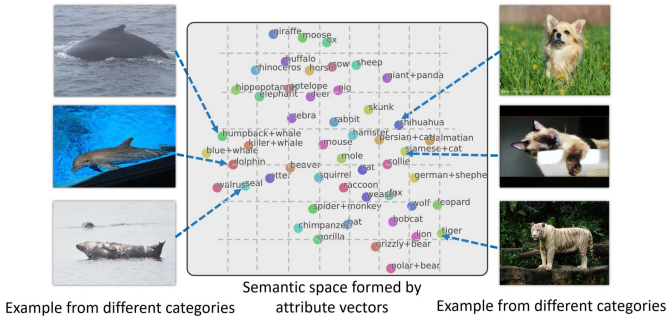



Fig. 3. Visualization of the attribute vectors for different object categories [47]. Examples are from AwA2 dataset.

deep features to have the same channel dimension as the attribute vector. Then, we use the same way as CAM [6] to build our classification-localization head, where a global average pooling (GAP) layer is used to transform the feature maps extracted from the final convolutional layer into a feature vector, thus obtaining the semantic representation $\mathbf{s}$. To align the obtained semantic feature space with the one formed by attribute vectors, we introduce the following semantic loss function $\mathcal{L}_S$ to guide the learning process:

$$\mathcal{L}_S = -\log \frac{\exp\left(\mathbf{s}^T \mathbf{a}_{y*}\right)}{\sum_{y \in \mathcal{Y}^S} \exp\left(\mathbf{s}^T \mathbf{a}_y\right)} \tag{1}$$

where $y*$ indicates the ground-truth category of the input image. We use the inner product operation to measure the alignment of the two spaces, where the obtained semantic representation is forced to better align to the attribute vector of its corresponding object category than those of other categories. Another interpretation is that the attribute vectors here can be considered as the classifiers to map the input features to a predefined semantic space.

### C. Semantic-Agnostic Modeling

Although the class-sensitive modeling plays a critical role in the GWSOL framework, only using it would make the learned features bias to the categories appearing in the training process. Consequently, we additionally introduce the semantic-agnostic modeling into the learning framework. The goal is to learn semantic-agnostic feature representations that can generally fit arbitrary object categories. We design a semantic-agnostic embedding head (SAEH) as shown in Fig. 2(c). Specifically, we introduce one additional network stream which is parallel to CLH. Same with CLH, SAEH also contains a $3 \times 3$ convolutional layer and a GAP layer, while the output is the feature embedding $\mathbf{e}$ that can be used to represent the semantic-agnostic content of the input image. To learn such a network head, we use the triplet embedding loss

$$\mathcal{L}_E = \max\left[||\mathbf{e}^o - \mathbf{e}^+||_2^2 - ||\mathbf{e}^o - \mathbf{e}^-||_2^2 + \alpha, 0\right] \tag{2}$$

where $(\mathbf{e}^o, \mathbf{e}^+, \mathbf{e}^-)$ is the sampled triplet, among which $\mathbf{e}^o$ indicates the anchor sample, $\mathbf{e}^+$ indicates the positive sample, and $\mathbf{e}^-$ indicates the negative sample. $\alpha$ is a parameter to control the margin [48], [49]. As can be seen, $\mathcal{L}_E$ will not associate the learned features to specific semantics but encourage the obtained embedding space to have high intra-class compactness and inter-class discriminativeness. Although the obtained feature embedding cannot totally overcome the bias brought by the seen categories, it can improve the representation capacity in dealing with the potential unseen categories.


Fig. 4. Illustration the situation that the attribute vector for one class is constant, while the specific object instance and its context can vary greatly.

### D. Content-Aware Modeling

It is still insufficient to introduce the class-sensitive modeling and the semantic-agnostic modeling to accomplish a promising GWSOL performance. The main issue is that an attribute vector is used to describe the general visual or semantic characteristics of an object category, which might not well fit the specific cases in each given image (see Fig. 4). To solve this problem, we design a content-based discriminative attribute mining module (CDAM).

The goal of CDAM is to generate the attribute weight vector according to the input image content. To comprehensively explore the image content, we introduce the top-down flow, bottom-up flow, and global information path to encode the high-level information, low-level information, and the global context, respectively. With the finely explored image content, we can then obtain more accurate attribute weight vectors. Suppose the input image contains a rabbit with blue color, which is rarely appeared. Then, the learner should assign a low weight on the attribute element corresponding to color so that the localization would not be influenced by this attribute element.

For implementing CDAM, we introduce a bi-directional feature flow (BDFF) to mutually transmit messages between adjacent levels of features from both bottom-up and top-down views with the guidance of the global image content captured by SAEH. In this way, CDAM explores threefold informative cues, including the top-down, bottom-up, and global information, at each feature level, thus making CDAM work effectively. Specifically, in the bi-directional feature flow, we use the features from different levels as the inputs and it outputs the content-based discriminative attribute weights. Specifically, the first part of the bi-directional feature flow is the top-down feature flow, where features from the deeper network layers are gradually transformed and aggregated to the shallower network layers. The second part of it is the bottom-up feature flow, where features from the shallower network layers are gradually transformed and aggregated to the deeper network layers. The concrete flow connection can be referred to in Fig. 2(c). The detailed design of the feature transformation block and feature aggregation block are displayed in Fig. 5(a) and (b). Afterward, we input each of the features extracted from the bottom-up feature flow to a GAP-based network layer to obtain the attribute weights.

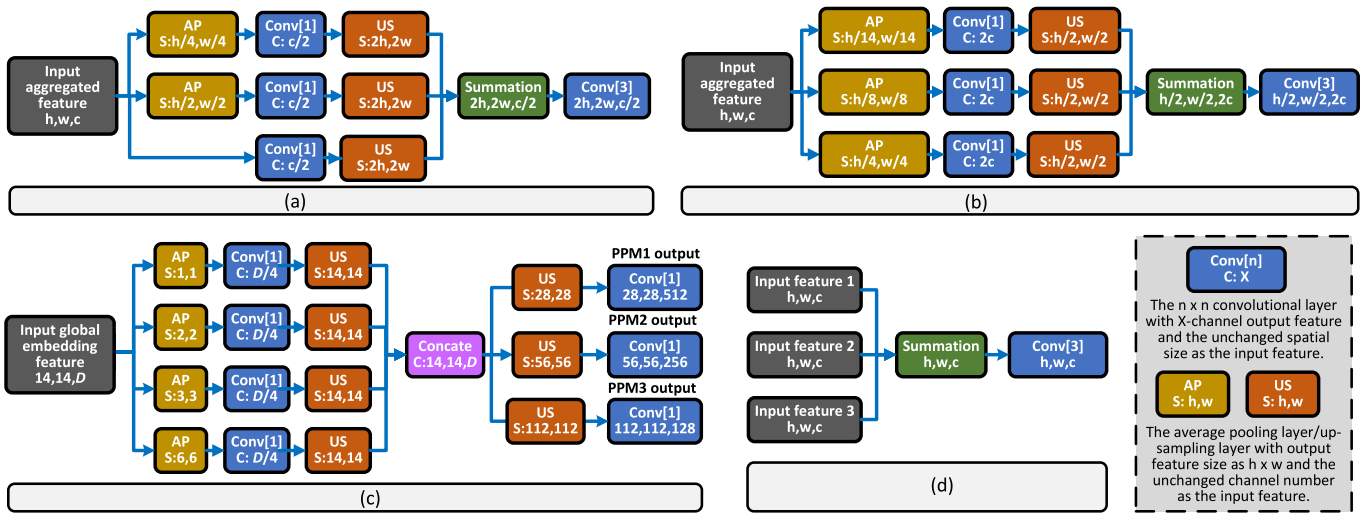



Fig. 5. Illustration of the architectures of the modules used to build our learning framework. Specifically, (a) and (b) show the architectures of the top-down feature transformation modules (T1-3 in Fig. 2) and the bottom-up feature transformation modules (T5 and T6 in Fig. 2), respectively. Note that T1 has a little different setting, where the convolutional layer would not change the channel number of the feature map. The absolute settings of the channel and spatial sizes of the feature maps in T4 are the same as those in T3, except for their different input feature sizes. (a) Detailed architecture of the feature transformation in the top-down flow. (b) Detailed architecture of the feature transformation in the bottom-up flow. (c) Detailed architecture of the PPMs. (d) Detailed architecture of the aggregation modules.

To introduce the global image content to guide the bi-directional feature flow, we feed the features learned from SAEH to a multiscale pyramid pooling block [see Fig. 5(c)], which contains three pyramid pooling modules (PPMs) to extract the multicontext features for each level. Inspired by [50], we design each PPM with hierarchical pooling layers to down-sample the input feature maps by four different rates and then perform the $1 \times 1$ convolution on the down-sampled features. Afterward, the obtained features are up-sampled to the same spatial size as the input feature. Finally, the output features are obtained by concatenating the four up-sampled feature maps and performing another $1 \times 1$ convolutional layer. Notice that three PPMs share the same pooling pyramid but with different settings on the output feature dimension.

To infer the content-based discriminative attribute weight vector $\bar{\mathbf{w}}$, we forward the feature maps obtained from the A4, A5, and A6 to three max pooling and GAP layers to obtain $\mathbf{w}_4$, $\mathbf{w}_5$, and $\mathbf{w}_6$, respectively, and then calculate $\bar{\mathbf{w}}$ as

$$\bar{\mathbf{w}} = \text{softmax}(\mathbf{w}_4 + \mathbf{w}_5 + \mathbf{w}_6). \tag{3}$$

In order to endow the obtained attribute weights the capacity to select the meaningful and discriminative attribute elements, we introduce the following loss function:

$$\mathcal{L}_W = \sum_{i=4}^{6} \eta ||\text{sigmoid}(\bar{\mathbf{w}}_i)||_1 + \tau D_{\text{KL}}(\bar{\mathbf{w}}||\mathbf{a}_{y*}) \tag{4}$$

where the first term is the $L_1$ norm of each inferred weight. It aims to make weight vectors select discriminative yet sparse elements. To prevent weight vectors from shifting to noise attribute elements that are less relative to the corresponding object category, we introduce the second term, i.e., the Kullback-Leibler (KL) divergence term. It encourages the distribution of the weights in $\bar{\mathbf{w}}$ to be consistent with the distribution of the elements in the attribute vector of the corresponding object category. $\tau$ is the weight assigned to the KL divergence term. By using these two terms, the obtained weight vector $\bar{\mathbf{w}}$ could sparsely select (with the soft weighting) the discriminative attribute elements and, ensure the selected elements are also under a reasonable distribution.

Finally, we introduce the inferred content-based discriminative attribute weights $\bar{\mathbf{w}}$ to help predict the object category label by converting (1) to

$$\mathcal{L}_S^* = -\log \frac{\exp\left(\mathbf{s}^T[\mathbf{a}_{y*} \odot \bar{\mathbf{w}}]\right)}{\sum_{y \in \mathcal{Y}^s} \exp\left(\mathbf{s}^T[\mathbf{a}_y \odot \bar{\mathbf{w}}]\right)} \tag{5}$$

where $\odot$ indicates the element-wise production. The whole learning framework is optimized by $\mathcal{L}_S^*$, $\mathcal{L}_E$ and $\mathcal{L}_W$.

### *E. GWSOL Prediction*

When predicting the object locations in the test process, we first identify the relationship between an unseen category $z$ with the seen categories by optimizing

$$\min_{\mathbf{r}^z} \left\| \boldsymbol{a}_z - \sum_{y \in \mathcal{Y}^s} r_y^z \boldsymbol{a}_y \right\|_2^2 + \lambda \left\| \mathbf{r}^z \right\|_2^2 \tag{6}$$

where $\mathbf{r}^z = [r_1^z, r_2^z, \ldots, r_{n_s}^z]$ is the relation vector, which encodes the relationship between the given unseen category $z$ with a seen category $y$ by the correspondence weight $r_y^z$, $n_s$ indicates the number of seen and unseen classes. $\lambda$ is a free parameter. Equation (6) is a standard ridge regression problem [51]. The solution is $\mathbf{r}^z = (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{a}_z$, where $\mathbf{A} \in \mathbb{R}^{D \times n_s}$ is formed by $\mathbf{a}_y$. Then, using the obtained relation vector, we infer the potential feature embedding of the given unseen category as $\mathbf{e}_z = \sum_{y \in \mathcal{Y}^s} r_y^z \mathbf{e}_y$, where $\mathbf{e}_y$ is the average feature embedding of the seen categories. Finally, we predict the classification result for the input image by measuring

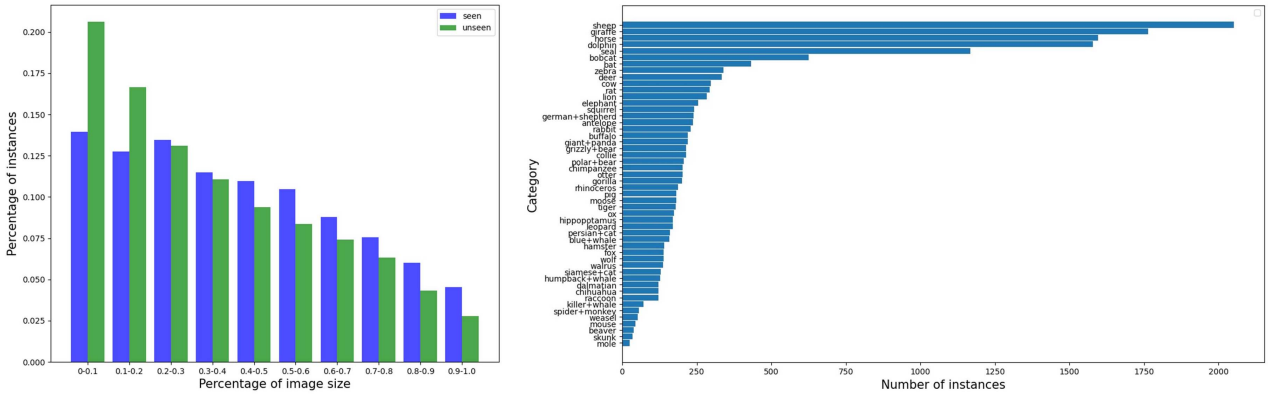



Fig. 6. Statistics on the instance sizes and instance numbers of the annotated AwA2 dataset.

the feature coherence both in the semantic space and the embedding space

$$\hat{y} = \underset{y \in \mathcal{Y}^{\mathcal{S}} \cup \mathcal{Y}^{\mathcal{U}}}{\arg\max} \quad \Phi\big(\mathbf{s}, [\bar{\mathbf{w}} \odot \boldsymbol{a}_y]\big) + \Phi\big(\mathbf{e}, \mathbf{e}_y\big) \tag{7}$$

where $\Phi(\cdot)$ indicates the cosine-similarity of two given vectors. The localization heatmap $\mathbf{M}$ is obtained by associating the Conv6 feature maps $\mathbf{F}_6$ with the attribute vector of the predicted object category $\mathbf{a}_{\hat{y}}$

$$\mathbf{M} = \sum_{i=1}^{D} \mathbf{F}_6^{(i)} a_{\hat{y}}^{(i)}. \tag{8}$$

Finally, we follow [6] to binarize the $\mathbf{M}$ by threshold 0.2 and generate a box that encloses the largest connected region for each image as the localization result.

## IV. Experiments

### A. Experimental Setup

*1) Datasets:* Our experiments are mainly conducted on the CUB (Caltech-UCSD Birds-200-2011) [52] and AwA2 (Animals with Attributes 2) [31] datasets. Specifically, CUB consists of 11 788 images from 200 different bird species with 312 user-defined attributes. It also provides the bounding-box annotation for each image. While AwA2 consists of 37 322 images from 50 animal classes with 85 user-defined attributes. Considering none bound-box annotation is provided for the AwA2 dataset, we collect them from human-annotators to facilitate the experimental comparisons. The statistics on the annotated box sizes (relative to the corresponding whole image size) and instance numbers are shown in Fig. 6. We follow the train-test split of [31], where the unseen classes do not contain the classes even used in backbone pretraining. Specifically, in CUB, we use 7057 images for training (containing 150 seen classes) and 4731 images for testing (1764 from the 150 seen classes and 2967 from the 50 unseen classes). In AwA2, we use 23 527 images for training (containing 40 seen classes) and 13 795 images for testing (5882 from the 40 seen classes and 7913 from the ten unseen classes).

*2) Test Settings:* There are two test settings in the experimental evaluation. The first setting is called the constrained setting (C-setting), where we only test on the images containing unseen classes. The second setting is called the generalized setting (G-setting), where the test images contain both the seen and unseen classes. The first setting evaluates how well the trained model can deal with the unseen classes while the latter setting further evaluates whether it can separate the unseen classes from the seen classes.

*3) Metrics:* We apply the top-$k$ localization accuracy (Top-$k$ Loc) for evaluating the performance. It considers the prediction as correct when both the estimated top-$k$ classes contain the ground truth class and the intersection over union (IoU) between the ground truth bounding box and the estimated box is 50% or more. Here we choose $k$ from 1 to 3 to see if the evaluated approaches can perform precise classification of the test categories. In G-setting, we separately evaluate the localization accuracy for seen and unseen classes and measure the overall performance by their harmonic mean (H-mean).

*4) Implementation Details:* We follow the existing works [48], [59] to use the user annotated attribute vector $\mathbf{a}$ provided by the datasets. Specifically, we first set the negative values in $\mathbf{a}$ to 0 s and then normalize each dimension to [0, 1]. We use the VGG19 [60] as the backbone, which is pretrained on the ImageNet while other network layers are initialized randomly. Input images are resized to $256 \times 256$ and then cropped to $224 \times 224$. We also use the color jitter for data augmentation on AwA2. When training the network, we select Adam as the optimizer. The learning rates are set to 1e-5 and 3e-5 on the AwA2 dataset and CUB dataset, respectively. The batch size is set as 24 in training. The loss weights of $\mathcal{L}_S^*$, $\mathcal{L}_E$ and $\mathcal{L}_W$ are set to 1, 5, 1, respectively. For all the experiments, both $\alpha$ and $\lambda$ are set to 1, $\eta$ is set to 0.0001. $\tau$ is set to 10 and 20 for AwA2 and CUB, respectively. Code is available at https://github.com/wmetaz/gwsol.

### B. GWSOL Benchmark

In this section, we benchmark the GWSOL performance on both the AwA2 dataset and the CUB dataset. Specifically, we compare the proposed approach with ten state-of-the-art WSOL approaches, including ADL [12], ACOL [58],

TABLE I

COMPARISON RESULTS ON THE AWA2 DATASET UNDER THE C-SETTING. THE RESULTS ARE MEASURED BY THE TOP-1, TOP-2, TOP-3 LOCALIZATION ACCURACY AS WELL AS THE AVERAGE ACCURACY

| Model | Top-1 | Top-2 | Top-3 | Ave |
|---|---|---|---|---|
| ACOL [12] | 15.32 | 24.66 | 35.92 | 25.30 |
| HAS [53] | 13.08 | 27.23 | 35.72 | 25.34 |
| DANet [10] | 17.75 | 25.88 | 36.6 | 26.74 |
| Cutmix [17] | 16.97 | 27.06 | 36.65 | 26.89 |
| Ccam [54] | 15.96 | 28.34 | 38.75 | 27.68 |
| Rcam [55] | 18.29 | 29.29 | 39.07 | 28.88 |
| SPG [56] | 17.48 | 32.18 | 39.17 | 29.61 |
| ADL [12] | 17.04 | 32.16 | 44.02 | 31.07 |
| PSOL [57] | 21.8 | 37.47 | 50.04 | 36.44 |
| GCNet [23] | 23.66 | 40.83 | 55.01 | 39.83 |
| Ours | **36.57** | **50.47** | **56.05** | **47.70** |

TABLE II

COMPARISON RESULTS ON THE CUB DATASET UNDER THE C-SETTING. THE RESULTS ARE MEASURED BY THE TOP-1, TOP-2, TOP-3 LOCALIZATION ACCURACY AS WELL AS THE AVERAGE ACCURACY

| Model | Top-1 | Top-2 | Top-3 | Ave |
|---|---|---|---|---|
| ADL [12] | 13.55 | 20.39 | 25.41 | 19.78 |
| SPG [56] | 14.70 | 22.25 | 24.95 | 20.63 |
| DANet [10] | 20.36 | 28.48 | 33.23 | 27.36 |
| HAS [53] | 22.72 | 30.06 | 34.24 | 29.01 |
| Ccam [54] | 21.94 | 31.75 | 37.07 | 30.25 |
| ACOL [58] | 23.86 | 34.38 | 38.29 | 32.18 |
| PSOL [57] | 25.88 | 35.39 | 40.07 | 33.78 |
| Cutmix [17] | 28.85 | 39.74 | 45.80 | 38.13 |
| Rcam [55] | 30.67 | 43.04 | 48.94 | 40.88 |
| GCNet [23] | 32.96 | 45.84 | 52.14 | 43.65 |
| Ours | **48.10** | **58.21** | **61.91** | **56.07** |

SPG [56], HAS [53], Cutmix [17], DANet [10], PSOL [57], Ccam [54], Rcam [55], and GCNet [23].[3] For facilitating such WSOL methods for working in the investigated generalized test scenario, we first adopt (6) to identify the relationship between an unseen category with the seen categories. Then, we obtain the classification score for an unseen category $z$ as $s_z = \sum_{y\in\mathcal{Y}^s} r_y^z s_y, z \in \mathcal{Y}^{\mathcal{U}}$. Similarly, we obtain the class activation map for an unseen category $z$ as $\mathbf{M}_z = \sum_{y\in\mathcal{Y}^s} r_y^z \mathbf{M}_y$. While the final localization map is obtained by

$$\mathbf{M} = \mathbf{M}_{\hat{y}}, \quad \hat{y} = \underset{y\in\mathcal{Y}^{\mathcal{S}}\cup\mathcal{Y}^{\mathcal{U}}}{\arg\max}\, s_y. \tag{9}$$

The comparison results under the C-setting are reported in Tables I and II, while the comparison results under the G-setting are reported in Tables III and IV. From the tables, we can observe the proposed approach consistently obtains the superior performance under different settings. The Top-1 localization accuracy under the G-setting is a very difficult evaluation scenario where almost all the existing WSOL approach obtains zero H-mean score due to that they could not get the right answer on the desired object semantics. Some approaches like GCnet and Ccam can obtain very promising localization capacity on the seen categories. However, these methods have limited generalization capacity for the unseen semantics. To our best knowledge, it is because the existing methods learn the classifier on seen classes, while we use attributes to generate the classification prediction. The attribute reveals the relationship between different classes but requires the model to respond to regions related to all attribute elements, thus hurting the discriminability of the model. Therefore, they cannot achieve as good performance as our proposed approach under the H-mean score. To sum up, the benchmark reveals the insufficient capacity of the existing WSOL methods as well as the effectiveness of our approach in dealing with different generalized test scenarios. We additionally visualize some localization results in Figs. 7 and 8.

[3]The backbone used in ACOL, ADL, HAS, DANet, Ccam, PSOL, and Rcam is VGG16 [60]. While the backbones used in Cutmix, SPG, and GCnet are Resnet50, InceptionNet, and GoogleNet, respectively.

### C. Ablation Study

We implement ablation studies on the AwA2 dataset to evaluate the effectiveness of each of the considered components. The base model is implemented by directly training a CAM-based classification network [6] on the seen categories and then use it to predict the unseen categories based on the attribute relationship formulated by (6). Upon this base model, we gradually add the proposed CLH, SAEH, and CDAM into the learning framework to see the obtained performance improvement. The experimental results are reported in Table V, from which we can observe that CLH can bring more performance gains for the C-setting than for the G-setting. On the contrary, SAEH and CDAM obtain more performance gains for the G-setting than for the C-setting. Finally, the combination of all the three considered components achieves the best performance. In addition, we also conduct the ablation studies on the design of the CDAM. The results reported in bottom rows of Table V reflect that the multiscale PPM module and the bi-directional flow path used in our framework provide more helpful information to the more challenging G-setting scenario. To further demonstrate that our inferred content-based discriminative attribute weights $\bar{\mathbf{w}}$ to help predict the object category label, we replace the learning objective function $\mathcal{L}_S^*$ by $\mathcal{L}_S$. Experimental results show that this obtains 9.82% and 24.40% performance degeneration under the C-setting and G-setting, respectively, which in turn verifies the importance of inferring content-based discriminative attribute weight in GWSOL.

### D. Zero-Shot Baseline

As shown in Table VI, we conduct experiments based on two zero-shot classification methods, i.e., ARE [37] and DAZLE [35], where the localization maps are generated by using Grad-CAM [61]. These results demonstrate that zero-shot classification methods can also be used in the proposed GWSOL task as they can establish relationship between unseen categories and seen categories. It can also be observed that our proposed method can outperform ARE [37] and DAZLE [35] by 5.92% and 24.93%, respectively.

### E. Failure Cases Analysis

As shown in Fig. 9, we provide several failure cases on the unseen categories of the CUB and AwA2 datasets.

TABLE III

COMPARISON RESULTS ON THE AWA2 DATASET UNDER THE G-SETTING. THE RESULTS ARE MEASURED BY THE TOP-1, TOP-2, TOP-3 LOCALIZATION ACCURACY AS WELL AS THE AVERAGE ACCURATE

| Localizer model | Top-1 Localization | | | Top-2 Localization | | | Top-3 Localization | | | Average Localization | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Seen | Unseen | H-mean | Seen | Unseen | H-mean | Seen | Unseen | H-mean | Seen | Unseen | H-mean |
| Cutmix [17] | 63.58 | 0.00 | 0.00 | 64.41 | 7.47 | 13.38 | 64.58 | 16.66 | 26.49 | 64.19 | 8.04 | 13.29 |
| PSOL [57] | 63.24 | 0.00 | 0.00 | 64.13 | 6.05 | 11.05 | 64.37 | 20.39 | 30.97 | 63.91 | 8.81 | 14.01 |
| ACOL [58] | 45.26 | 0.00 | 0.00 | 45.88 | 9.63 | 15.91 | 45.99 | 20.70 | 28.55 | 45.71 | 10.11 | 14.82 |
| SPG [56] | 38.17 | 0.02 | 0.03 | 39.26 | 9.53 | 15.34 | 39.70 | 24.20 | 30.07 | 39.04 | 11.25 | 15.15 |
| DANet [10] | 46.86 | 0.00 | 0.00 | 47.88 | 10.77 | 17.58 | 48.01 | 20.72 | 28.94 | 47.58 | 10.50 | 15.51 |
| Rcam [55] | 65.77 | 0.00 | 0.00 | 66.86 | 9.22 | 16.21 | 67.11 | 21.29 | 32.32 | 66.58 | 10.17 | 16.18 |
| GCNet [23] | **71.07** | 0.00 | 0.00 | **72.03** | 7.57 | 13.71 | **72.35** | 23.42 | 35.38 | **71.82** | 10.33 | 16.36 |
| ADL [12] | 56.87 | 0.00 | 0.00 | 57.76 | 10.52 | 17.80 | 57.98 | 22.62 | 36.49 | 57.54 | 11.05 | 18.10 |
| Ccam [54] | 67.39 | 0.00 | 0.00 | 68.60 | 11.01 | 18.98 | 68.95 | 26.96 | 38.77 | 68.31 | 12.66 | 19.25 |
| HAS [53] | 44.44 | 0.00 | 0.00 | 46.88 | 16.18 | 24.06 | 47.45 | 28.18 | 35.36 | 46.26 | 14.79 | 19.81 |
| Ours | 55.15 | **10.98** | **18.32** | 59.35 | **21.90** | **32.00** | 59.80 | **32.13** | **41.80** | 58.10 | **21.67** | **30.71** |

TABLE IV

COMPARISON RESULTS ON THE CUB DATASET UNDER THE G-SETTING. THE RESULTS ARE MEASURED BY THE TOP-1, TOP-2, TOP-3 LOCALIZATION ACCURACY AS WELL AS THE AVERAGE ACCURATE

| Localizer model | Top-1 Localization | | | Top-2 Localization | | | Top-3 Localization | | | Average Localization | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Seen | Unseen | H-mean | Seen | Unseen | H-mean | Seen | Unseen | H-mean | Seen | Unseen | H-mean |
| SPG [56] | 29.65 | 0.00 | 0.00 | 32.48 | 7.35 | 11.98 | 33.90 | 14.93 | 20.73 | 32.01 | 7.43 | 10.90 |
| PSOL [57] | 44.33 | 0.00 | 0.00 | 47.11 | 6.57 | 11.54 | 48.47 | 14.53 | 22.35 | 46.64 | 7.03 | 11.30 |
| GCNet [23] | **59.81** | 0.00 | 0.00 | **63.89** | 8.73 | 15.36 | **65.59** | 18.71 | 29.11 | **63.10** | 9.15 | 14.82 |
| Cutmix [17] | 51.53 | 0.00 | 0.00 | 54.42 | 10.04 | 16.96 | 55.05 | 19.31 | 28.59 | 53.67 | 9.78 | 15.18 |
| DANet [10] | 35.83 | 0.00 | 0.00 | 38.55 | 13.78 | 20.31 | 39.23 | 23.15 | 29.12 | 37.87 | 12.31 | 16.48 |
| HAS [53] | 35.26 | 0.00 | 0.00 | 39.46 | 14.73 | 21.45 | 40.25 | 23.83 | 29.94 | 38.32 | 12.85 | 17.13 |
| Ccam [54] | 50.28 | 0.00 | 0.00 | 54.42 | 12.13 | 19.84 | 55.39 | 22.38 | 31.88 | 53.36 | 11.50 | 17.24 |
| ACOL [58] | 44.44 | 0.00 | 0.00 | 46.88 | 16.18 | 24.06 | 47.45 | 28.18 | 35.36 | 46.26 | 14.79 | 19.81 |
| ADL [12] | 47.05 | 0.00 | 0.00 | 49.94 | 18.44 | 26.93 | 50.96 | 30.37 | 38.06 | 49.32 | 16.27 | 21.66 |
| Rcam [55] | 53.63 | 0.00 | 0.00 | 57.14 | 17.96 | 27.34 | 58.62 | 31.72 | 41.16 | 56.46 | 16.56 | 22.83 |
| Ours | 47.51 | **12.81** | **20.18** | 57.26 | **24.10** | **33.92** | 60.49 | **34.75** | **44.14** | 55.09 | **23.89** | **32.75** |

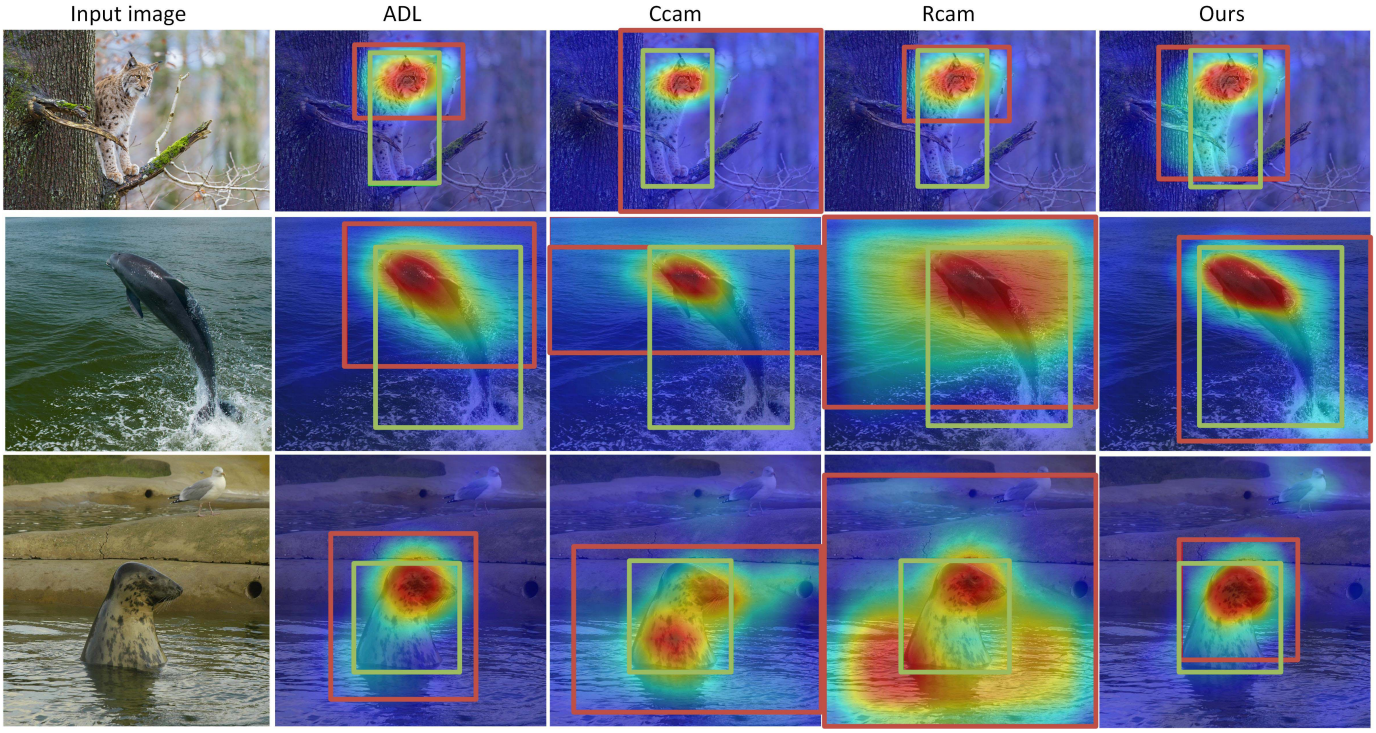



Fig. 7. Visualization comparison with the existing methods on the AwA2 dataset. The green and red boxes indicate the ground truth and the predicted object locations, respectively. The results are from unseen categories under the G-setting.

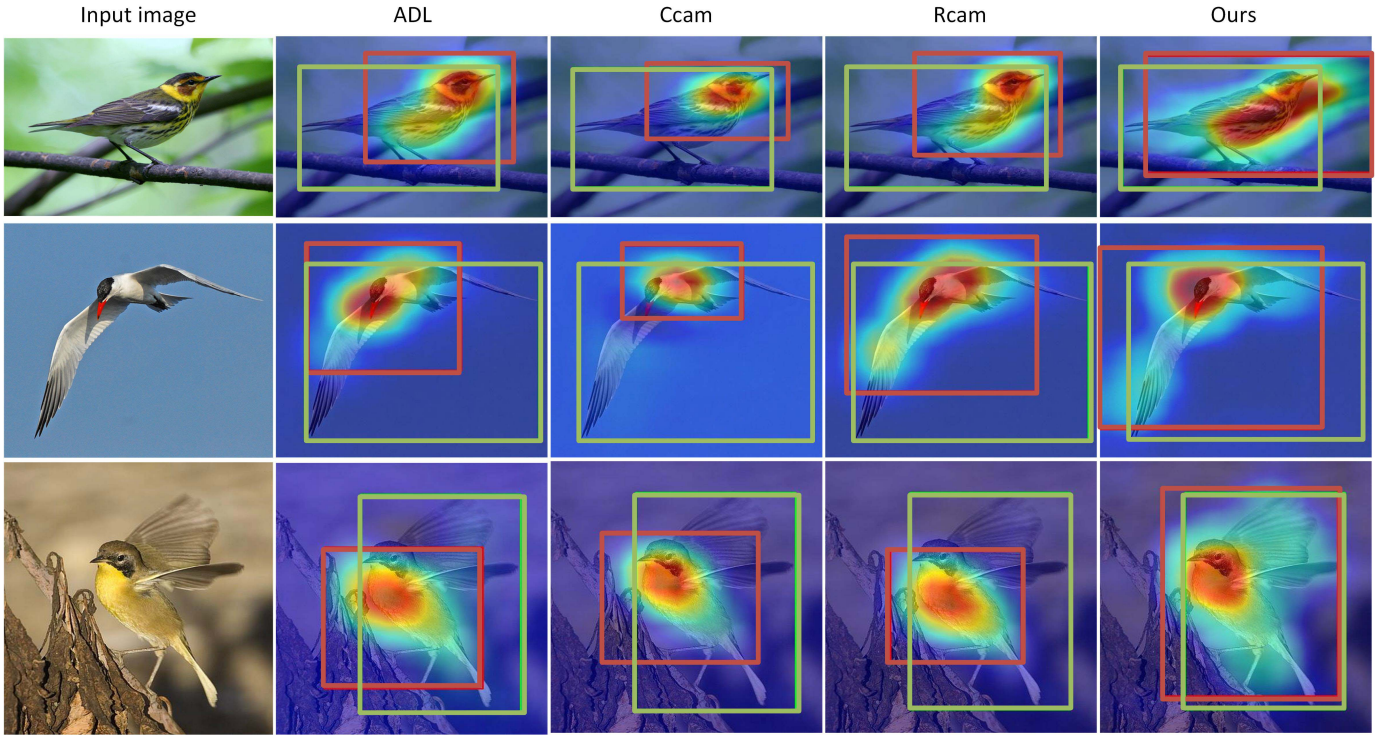



Fig. 8. Visualization comparison with the existing methods on the CUB dataset. The green and red boxes indicate the ground truth and the predicted object locations, respectively. The results are from unseen categories under the G-setting.

TABLE V

ABLATION STUDY ON THE AWA2 DATASET. UNDER G-SETTING, WE REPORT THE H-MEAN ACCURACY TO CONSIDER BOTH THE SEEN CATEGORIES AND UNSEEN CATEGORIES. WHILE UNDER C-SETTING, WE DIRECTLY REPORT THE LOCALIZATION ACCURACY ON UNSEEN CATEGORIES. ↓ INDICATES THE PERFORMANCE DEGENERATION WHEN COMPARED TO OUR FINAL MODEL

| **Module components** | | | | **Top-1** | **Top-2** | **Top-3** | **Ave** | **±Gains** | **Top-1** | **Top-2** | **Top-3** | **Ave** | **±Gains** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Base | CLH | SAEH | CDAM | Evaluation under C-setting | | | | | Evaluation under G-setting | | | | |
| ✓ | | | | 22.82 | 32.15 | 35.96 | 22.73 | - | 0.00 | 12.99 | 22.82 | 11.94 | - |
| ✓ | ✓ | | | 27.30 | 36.85 | 44.53 | 27.17 | 4.44 ↑ | 4.48 | 13.72 | 23.01 | 13.74 | 1.80 ↑ |
| ✓ | ✓ | ✓ | | 30.45 | 41.54 | 48.26 | 30.06 | 2.89 ↑ | 10.23 | 20.36 | 30.71 | 20.43 | 6.69 ↑ |
| ✓ | ✓ | ✓ | ✓ | 36.57 | **50.47** | **56.05** | **35.77** | 5.71 ↑ | **18.32** | **32.00** | 41.80 | **30.71** | 10.28 ↑ |
| Ours without mutil-scale PPM | | | | 37.41 | 48.97 | 52.51 | 34.72 | 1.05 ↓ | 16.49 | 29.96 | **42.10** | 29.52 | 1.19 ↓ |
| Ours without BDFF | | | | **38.24** | 49.84 | 54.34 | 35.61 | 0.17 ↓ | 13.59 | 27.82 | 39.34 | 26.92 | 3.79 ↓ |
| Ours using $\mathcal{L}_S$ instead of $\mathcal{L}_S^*$ | | | | 32.59 | 45.18 | 51.27 | 32.26 | 3.51 ↓ | 11.00 | 22.82 | 35.82 | 23.21 | 7.49 ↓ |

TABLE VI

COMPARISON WITH ZERO-SHOT BASELINES ON THE CUB DATASET UNDER THE C-SETTING. THE RESULTS ARE MEASURED BY THE TOP-1, TOP-2, TOP-3 LOCALIZATION ACCURACY AS WELL AS THE AVERAGE ACCURACY

| Model | Top-1 | Top-2 | Top-3 | Ave |
|---|---|---|---|---|
| ARE [37] | 44.15 | 52.04 | 55.31 | 50.51 |
| DAZLE [35] | 27.20 | 32.59 | 34.72 | 31.50 |
| Ours | 48.10 | 58.21 | 61.91 | 56.07 |

TABLE VII

ANALYSIS ON THE COMPUTATIONAL COST OF THE PROPOSED METHOD ON AWA2 DATASET

| **Base** | **CLH** | **SAEH** | **CDAM** | **ms/Image** | **MACs(G)** | **Param (M)** |
|---|---|---|---|---|---|---|
| √ | | | | 3.44 | 19.56 | 20.07 |
| √ | √ | | | 3.52 | 19.63 | 20.43 |
| √ | √ | √ | | 3.57 | 19.71 | 20.82 |
| √ | √ | √ | √ | 10.34 | 42.69 | 35.20 |

Most failure cases are caused by the part domination problem of CNNs, i.e., the localization maps are usually dominated by the most discriminative regions of the object. Besides, it is difficult for our model to establish the relationship between objects of unseen categories and seen categories when the background is too complex.

### *F. Efficiency Analysis*

In Table VII, we provide a detailed analysis of the computational cost for every single component in the training phase, we report the time cost (ms/image), multiply–accumulate operations (MACs) and the number of parameters (Param). We can observe that the proposed CLH and SAEH modules

TABLE VIII

RESULTS ON IMAGENET. UNDER G-SETTING, WE REPORT THE H-MEAN ACCURACY TO CONSIDER BOTH THE SEEN CATEGORIES AND UNSEEN CATEGORIES. WHILE UNDER C-SETTING, WE DIRECTLY REPORT THE ACCURACY ON UNSEEN CATEGORIES. BOTH THE LOCALIZATION AND CLASSIFICATION PERFORMANCE ARE REPORTED

| | Top-1 | Top-2 | Top-3 | Average | Top-1 | Top-2 | Top-3 | Average |
|---|---|---|---|---|---|---|---|---|
| | Evaluation under C-setting | | | | Evaluation under G-setting | | | |
| Classification | 8.11 | 13.67 | 18.02 | 13.27 | 0.38 | 0.95 | 1.64 | 0.99 |
| Localization | 3.41 | 5.71 | 7.67 | 5.60 | 0.10 | 0.28 | 0.48 | 0.29 |

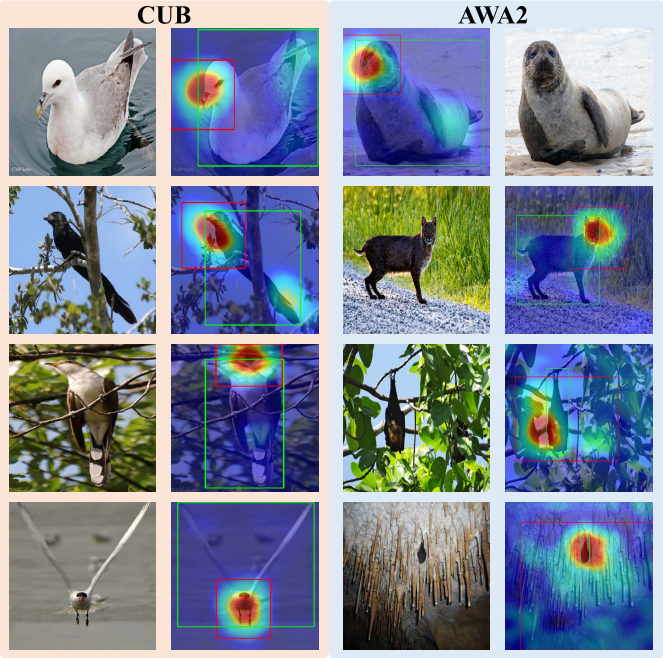



Fig. 9. Visualization of failure cases on the CUB and AwA2 dataset. The green and red boxes indicate the ground truth and the predicted object locations, respectively. The results are from unseen categories under the G-setting.

bring a little computational cost. The main computational cost of our method comes from the CDAM module. Using it makes the time cost increase from 3.57 to 10.34 ms/image, the MACs increases from 19.71G to 42.69G, and the number of parameters increase from 20.82M to 35.20M.

### G. Performance on ImageNet and Potential Limitation

We further test our method on ImageNet [62]. To implement the experiment, we follow [63] to split ImageNet-1K into a seen subset (800 classes) and a unseen subset (200 classes). We use Word2Vec [4] to extract attributes and the backbones are re-trained on the training subset rather than directly using the parameters pretrained on ImageNet-1K. The performance under the C-setting and G-setting are shown in Table VIII. As can observe, the model has difficulty predicting classification results correctly, which in turn affects the localization performance. The main issue might be that there are no manually annotated attributes on ImageNet—The attribute matrix is generated by unsupervised methods like Word2Vec [4]. Under this circumstance, there are two potential ways to facilitate the evaluation of GWSOL in the large-scaled test scenario. The first way is to collect human-defined attribute annotations on large-scale localization datasets. While the other way is to provide instance-level annotations for the test set of large-scale zero-shot classification datasets that have manually annotated attributes. LAD [64] is a zero-shot classification dataset that contains 78 017 images of 230 classes. It can be used for GWSOL in the future by providing instance-level annotations for the test images.

## V. CONCLUSION

This article has studied the GWSOL problem to fill the blank of the existing literature in localizing unseen object semantics under the weak supervision. To address this problem, we build a novel deep learning framework which involves the class-sensitive modeling, semantic-agnostic modeling, and content-aware modeling to localize objects corresponding to the specific semantics, learn compact and discriminative features that could represent the potential unseen categories, and mine content-aware patterns to facilitate attribute reweighting. Comprehensive experiments on the AwA2 and CUB datasets demonstrate the effectiveness of our proposed learning framework and each of the considered modeling components. Both the proposed learning framework and the annotated bounding-boxes will be released to promote the development of this research field. In the future, the proposed methods can also be used in the field of temporal action localization [65], [66], [67], [68], [69] when the action semantics to be localized in test videos may either appear in the training phase or never been observed before.

## REFERENCES

[1] L. Zhou, C. Gong, Z. Liu, and K. Fu, "SAL: Selection and attention losses for weakly supervised semantic segmentation," *IEEE Trans. Multimedia*, vol. 23, pp. 1035–1048, 2020.

[2] C. Gong, J. Yang, J. You, and M. Sugiyama, "Centroid estimation with guaranteed efficiency: A general framework for weakly supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2841–2855, Jun. 2022.

[3] D. Zhang, J. Han, G. Guo, and L. Zhao, "Learning object detectors with semi-annotated weak labels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3622–3635, Dec. 2019.

[4] D. Zhang, J. Han, L. Zhao, and T. Zhao, "From discriminant to complete: Reinforcement searching-agent learning for weakly supervised object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5549–5560, Dec. 2020.

[5] W. Zhao, J. Zhang, L. Li, N. Barnes, N. Liu, and J. Han, "Weakly supervised video salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16826–16835.

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[7] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Soft proposal networks for weakly supervised object localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1841–1850.

[8] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 20, 2021, doi: 10.1109/TPAMI.2021.3074313.

[9] G. Guo, J. Han, F. Wan, and D. Zhang, "Strengthen learning tolerance for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7403–7412.

[10] H. Xue, C. Liu, F. Wan, J. Jiao, X. Ji, and Q. Ye, "DANet: Divergent activation for weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6589–6598.

[11] J. Mai, M. Yang, and W. Luo, "Erasing integrated learning: A simple yet effective approach for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8766–8775.

[12] J. Choe and H. Shim, "Attention-based dropout layer for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2219–2228.

[13] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1778–1785.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[15] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[16] S. Rahman, S. H. Khan, and F. Porikli, "Zero-shot object detection: Joint recognition and localization of novel concepts," *Int. J. Comput. Vis.*, vol. 128, no. 12, pp. 2979–2999, Dec. 2020.

[17] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6023–6032.

[18] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1307–1314.

[19] Z. Shi, T. M. Hospedales, and T. Xiang, "Bayesian joint topic modelling for weakly supervised object localisation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2984–2991.

[20] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, Jan. 2017.

[21] C. Wang, W. Ren, K. Huang, and T. Tan, "Weakly supervised object localization with latent category learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 431–445.

[22] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim, "Evaluating weakly supervised object localization methods right," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3133–3142.

[23] W. Lu, X. Jia, W. Xie, L. Shen, Y. Zhou, and J. Duan, "Geometry constrained weakly supervised object localization," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 481–496.

[24] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1568–1576.

[25] Q. Hou, P. Jiang, Y. Wei, and M.-M. Cheng, "Self-erasing network for integral object attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 549–559.

[26] J. Xie, C. Luo, X. Zhu, Z. Jin, W. Lu, and L. Shen, "Online refinement of low-level feature based activation map for weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 132–141.

[27] J. Kim, J. Choe, S. Yun, and N. Kwak, "Normalization matters in weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3427–3436.

[28] M. Meng, T. Zhang, Q. Tian, Y. Zhang, and F. Wu, "Foreground activation maps for weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3385–3395.

[29] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3174–3183.

[30] S. Liu, M. Long, J. Wang, and M. I. Jordan, "Generalized zero-shot learning with deep calibration network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2005–2015.

[31] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, Sep. 2019.

[32] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5542–5551.

[33] Y. Yu, Z. Ji, J. Han, and Z. Zhang, "Episode-based prototype generating network for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14035–14044.

[34] Z. Han, Z. Fu, and J. Yang, "Learning the redundancy-free features for generalized zero-shot object recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12865–12874.

[35] D. Huynh and E. Elhamifar, "Fine-grained generalized zero-shot learning via dense attribute-based attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4483–4493.

[36] X. Chen, X. Lan, F. Sun, and N. Zheng, "A boundary based out-of-distribution classifier for generalized zero-shot learning," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 572–588.

[37] G.-S. Xie *et al.*, "Attentive region embedding network for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9384–9393.

[38] K. Li, M. R. Min, and Y. Fu, "Rethinking zero-shot learning: A conditional visual classification perspective," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3583–3592.

[39] S. Rahman, S. Khan, and N. Barnes, "Transductive learning for zero-shot object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6082–6091.

[40] Z. Li, L. Yao, X. Zhang, X. Wang, S. Kanhere, and H. Zhang, "Zero-shot object detection with textual descriptions," in *Proc. AAAI*, vol. 33, 2019, pp. 8690–8697.

[41] M. Bucher, V. Tuan-Hung, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 468–479.

[42] P. Hu, S. Sclaroff, and K. Saenko, "Uncertainty-aware learning for zero-shot semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1–12.

[43] Z. Gu, S. Zhou, L. Niu, Z. Zhao, and L. Zhang, "Context-aware feature generation for zero-shot semantic segmentation," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1921–1929.

[44] N. Kato, T. Yamasaki, and K. Aizawa, "Zero-shot semantic segmentation via variational mapping," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1363–1370.

[45] Z. Chen *et al.*, "Semantics disentangling for generalized zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8712–8720.

[46] S. Chen *et al.*, "FREE: Feature refinement for generalized zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 122–131.

[47] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

[48] Y. Liu, J. Guo, D. Cai, and X. He, "Attribute attention for semantic disambiguation in zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6698–6707.

[49] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[50] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[51] D. W. Marquardt and R. D. Snee, "Ridge regression in practice," *Amer. Statist.*, vol. 29, no. 1, pp. 3–20, 1975.

[52] P. Welinder *et al.*, "Caltech-UCSD birds 200," California Inst. Technol., Pasadena, CA, USA, Tech. Rep., 2010.

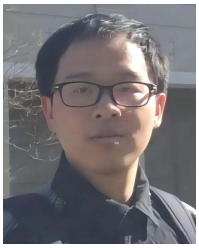
[53] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3544–3553.

[54] S. Yang, Y. Kim, Y. Kim, and C. Kim, "Combinational class activation maps for weakly supervised object localization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2941–2949.

[55] W. Bae, J. Noh, and G. Kim, "Rethinking class activation mapping for weakly supervised object localization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 618–634.

[56] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang, "Self-produced guidance for weakly-supervised object localization," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 597–613.

[57] C.-L. Zhang, Y.-H. Cao, and J. Wu, "Rethinking the route towards weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13460–13469.
[58] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1325–1334.
[59] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang, "Designing category-level attributes for discriminative visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 771–778.
[60] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.
[61] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[62] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[63] A. Frome *et al.*, "DeViSE: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–9.
[64] B. Zhao, Y. Fu, R. Liang, J. Wu, Y. Wang, and Y. Wang, "A large-scale attribute dataset for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 398–407.
[65] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, "UntrimmedNets for weakly supervised action recognition and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4325–4334.
[66] L. Yang, J. Han, T. Zhao, T. Lin, D. Zhang, and J. Chen, "Background-click supervision for temporal action localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 2, 2021, doi: 10.1109/TPAMI.2021.3132058.
[67] T. Zhao, J. Han, L. Yang, B. Wang, and D. Zhang, "SODA: Weakly supervised temporal action localization based on astute background response and self-distillation learning," *Int. J. Comput. Vis.*, vol. 129, no. 8, pp. 2474–2498, Aug. 2021.
[68] Y. Zhai, L. Wang, W. Tang, Q. Zhang, N. Zheng, and G. Hua, "Action coherence network for weakly-supervised temporal action localization," *IEEE Trans. Multimedia*, vol. 24, pp. 1857–1870, 2022.
[69] Z. Liu *et al.*, "Weakly supervised temporal action localization through contrast based evaluation networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3899–3908.



**Dingwen Zhang** (Member, IEEE) received the Ph.D. degree from Northwestern Polytechnical University (NPU), Xi'an, China, in 2018.

From 2015 to 2017, he was a Visiting Scholar with the Robotic Institute, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently a Professor with the School of Automation, NPU. His research interests include computer vision and multimedia processing, especially on saliency detection and weakly supervised learning.

**Guangyu Guo** received the B.E. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Automation, Northwestern Polytechnical University, Xi'an, China.

His research interests include computer vision and multimedia processing, especially on weakly supervised learning.

**Wenyuan Zeng** received the B.E. degree from Northwestern Polytechnical University, Xi'an, China, in 2019, where he is currently pursuing the M.S. degree with the School of Automation.

His research interests include computer vision and multimedia processing, especially on weakly supervised object detection and localization.

**Lei Li** is currently with the Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing, China. His current research interests include machine learning and computer vision, with particular interests in object detection and tracking.

**Junwei Han** (Fellow, IEEE) received the Ph.D. degree from Northwestern Polytechnical University (NPU), Xi'an, China, in 2003.

He was a Research Fellow with Nanyang Technological University, Singapore, The Chinese University of Hong Kong, Hong Kong, and the University of Dundee, Dundee, U.K. He is currently a Professor with NPU. He has published over 100 papers in IEEE TRANSACTIONS and top tier conferences. His research interests include computer vision and brain imaging analysis.

Prof. Han is currently an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON MULTIMEDIA.