# Group-Wise Learning for Weakly Supervised Semantic Segmentation

Tianfei Zhou, Liulei Li, Xueyi Li, Chun-Mei Feng, Jianwu Li, *Member, IEEE*, and Ling Shao, *Fellow, IEEE*

***Abstract*— Acquiring sufficient ground-truth supervision to train deep visual models has been a bottleneck over the years due to the data-hungry nature of deep learning. This is exacerbated in some structured prediction tasks, such as semantic segmentation, which require pixel-level annotations. This work addresses weakly supervised semantic segmentation (WSSS), with the goal of bridging the gap between image-level annotations and pixel-level segmentation. To achieve this, we propose, for the first time, a novel group-wise learning framework for WSSS. The framework explicitly encodes semantic dependencies in a group of images to discover rich semantic context for estimating more reliable pseudo ground-truths, which are subsequently employed to train more effective segmentation models. In particular, we solve the group-wise learning within a graph neural network (GNN), wherein input images are represented as graph nodes, and the underlying relations between a pair of images are characterized by graph edges. We then formulate semantic mining as an iterative reasoning process which propagates the common semantics shared by a group of images to enrich node representations. Moreover, in order to prevent the model from paying excessive attention to common semantics, we further propose a graph dropout layer to encourage the graph model to capture more accurate and complete object responses. With the above efforts, our model lays the foundation for more sophisticated and flexible group-wise semantic mining. We conduct comprehensive experiments on the popular PASCAL VOC 2012 and COCO benchmarks, and our model yields state-of-the-art performance. In addition, our model shows promising performance in weakly supervised object localization (WSOL) on the CUB-200-2011 dataset, demonstrating strong generalizability. Our code is available at: *https://github.com/Lixy1997/Group-WSSS*.**

***Index Terms*— Semantic segmentation, weakly supervised learning, group-wise learning, graph neural networks, object localization, neural attention.**

Manuscript received December 8, 2020; revised September 3, 2021; accepted November 22, 2021. Date of publication December 15, 2021; date of current version January 3, 2022. This work was supported in part by the Beijing Natural Science Foundation under Grant L191004, in part by the National Key Research and Development Program of China under Grant 2019YFB1310803, and in part by the China Computer Federation (CCF)-Baidu Open Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Senem Velipasalar. *(Corresponding author: Jianwu Li.)*

Tianfei Zhou is with the Computer Vision Laboratory, ETH Zürich, 8092 Zürich, Switzerland (e-mail: ztfei.debug@gmail.com).

Liulei Li, Xueyi Li, and Jianwu Li are with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100811, China (e-mail: ljw@bit.edu.cn).

Chun-Mei Feng is with the Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China.

Ling Shao is with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates (e-mail: ling.shao@ieee.org).

Digital Object Identifier 10.1109/TIP.2021.3132834

## I. Introduction

SEMANTIC segmentation is a fundamental task in computer vision, aiming to predict a semantic category for each pixel in an image. It can benefit a wide variety of applications including image editing [3], [4], visual analysis and understanding [5], [6] and medical diagnosis [7], [8]. With the recent renaissance of deep neural networks, semantic segmentation has achieved tremendous progress. However, most leading approaches fall into a fully supervised paradigm [9]–[11], requiring massive amounts of pixel-level annotated training data, which are extremely expensive and time-consuming to obtain. In contrast, the weak supervision alternatives, *e.g.,* image-level tags, scribbles or bounding-box annotations, are less costly. Thus, it is of interest to explore the potential of these weak supervision cues to provide a data-efficient solution for semantic segmentation. This work aims to address weakly supervised semantic segmentation (WSSS) under the supervision of image-level tags, which can be obtained effortlessly.

WSSS based on image tags is extremely challenging because fine-grained pixel-level annotations, which are typically required for semantic segmentation, are difficult to obtain from class labels. Starting from the pioneering work [12], most studies follow a two-stage pipeline for WSSS (see Fig. 1). The first stage aims at obtaining pseudo ground-truths by recognizing the discriminative regions based on class activation maps (CAMs), while the second stage employs these pseudo ground-truths to train a semantic segmentation network. However, CAMs are limited in only highlighting the most discriminative object parts rather than the whole object regions, causing unsatisfactory performance. Numerous approaches have been introduced to alleviate this problem. For example, some approaches [13]–[15] manipulate internal feature maps to guide the network to perceive easily ignored but essential parts, while others [16]–[19] adopt self-ensembling or self-supervision to improve object localization.

However, the mainstream methods mentioned above are merely based on ***single images*** (Fig. 2 (a)), ignoring the valuable semantic context existing in a group of images. Very recent studies [20], [21] utilize Siamese networks to model the relations between a pair of images, leading to a ***pair-wise*** solution (Fig. 2 (b)). These approaches have proven effective in locating more accurate object regions. However, seeking relations between two images at a time is still limited when it comes to capturing substantial semantic context. In this work, we introduce a more promising, and fundamentally different

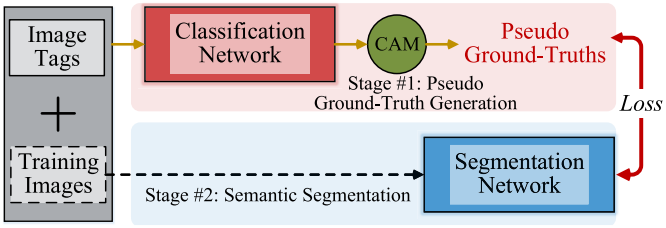


Fig. 1. **Mainstream two-stage pipeline for WSSS** with (1) pseudo ground-truth generation and (2) semantic segmentation. Our work follows this pipeline, but introduces a novel group-wise learning framework to achieve better pseudo ground-truths.

***group-wise*** solution (Fig. 2 (c)), which comprehensively mines richer semantics from a group of images. Our main motivation is that the availability of image groups containing instances of the *same semantic classes* can make up for the absence of detailed supervisory information. From this perspective, we hypothesize that it is desirable to take advantage of all available information for WSSS, including not only individual image properties, but also group-level synergetic relationships.

Based on the above analysis, we propose a novel deep learning model for WSSS. Unlike previous pair-wise approaches, our model aims for group-wise semantic mining to capture more comprehensive relations among input images. Specifically, we develop an efficient, end-to-end trainable graph neural network (GNN), and conduct recursive reasoning for group-wise semantic understanding. In our graph, the nodes represent a group of input images, and edges describe pair-wise relations between two connected images. We consider two images as connected only if they share common semantic objects with each other, and their relation is then characterized by an elaborately designed co-attention mechanism. Through iterative message passing, the information from individual elements can be efficiently integrated and broadcasted over the graph structure. In this way, our model is capable of leveraging explicit semantic dependencies among images to obtain better node representations. However, this graph reasoning strategy mainly focuses on co-occurring semantics in input images, ignoring isolated objects. To address this, we further introduce a graph dropout layer, which can be seamlessly integrated into the GNN for iterative inference. The graph dropout layer selectively suppresses the most salient objects, forcing the network to be biased toward other less salient counterparts.

The proposed method has two appealing characteristics over single-image and pair-wise methods. **(1)** It is capable of learning semantic relations from an arbitrary number of images using a flexible GNN framework. The GNN also empowers our model to inherit the complementary strengths of neural networks in learning capability and graphical models in structure representations. **(2)** Our model conducts multi-step, iterative inference to collect contextual knowledge for updating image representations. This is more favorable than directly producing image representations by one-step inference, as done in previous approaches.

Since all components in our model are differentiable, the whole network is fully end-to-end trainable. We perform extensive experiments on two popular WSSS benchmarks (*i.e.,* PASCAL VOC 2012 [22] and COCO [23]), and our model achieves consistent improvement in performance over the current state of the arts. Furthermore, we showcase the advantages and generalizability of our group-wise learning framework via the weakly supervised object localization (WSOL) task. Our model again shows superior performance on the widely used CUB-200-2011 dataset [24].

In summary, our main contributions are four-fold: **(1)** To the best of our knowledge, we are **the first** to propose group-wise semantic mining for WSSS, which significantly outperforms existing single-image and pair-wise based approaches. **(2)** We proffer a graph-aware solution to discover comprehensive semantic context from a group of images within an effective iterative reasoning process. **(3)** We introduce edge-aware supervision to enforce the learning of common semantics shared by a pair of images, thereby directly propagating group-wise errors backward to guide the network training. **(4)** Our model is evaluated on WSSS and WSOL tasks, and the results demonstrate its superiority and high generalizability.

An earlier conference version of this manuscript appeared in [25]. This paper makes several new contributions. **(1)** We demonstrate the generalizability of the proposed group-wise learning framework to the WSOL task (Sections V and VI-B). **(2)** We introduce a novel multi-granular supervision loss (Section IV-D) which enables explicit supervision of group-wise semantic mining, yielding consistent performance improvements (see Table IV)). **(3)** We also provide a more through overview of the recent work on weakly supervised semantic segmentation and object localization, graph neural networks as well as visual learning from multiple images (Section II). **(4)** We report much more experimental results and conduct failure case analysis (Section VI-E) for comprehensive and in-depth examinations of our model.

## II. Related Work

Our work is related to four lines of research, *i.e.,* weakly supervised semantic segmentation, weakly supervised object localization, graph neural networks, and visual learning from multiple images. We discuss each of them in the following.

### A. Weakly Supervised Semantic Segmentation

Recent years have seen a surge of interest in semantic segmentation under weak supervision (*e.g.,* image-level labels [14], [26]–[31], scribbles [32], [33], bounding boxes [34]–[36] or points [37]) to reduce the expensive up-front annotation costs. In particular, methods operating with image-level labels have attracted the most attention since they require minimal annotation efforts. Most of these methods follow the two-stage pipeline shown in Fig. 1. In the first stage, an image classification network is trained using only image-level tags, and CAM [12] is employed to highlight class-specific object regions, which serve as pseudo ground-truths. In the second stage, a semantic segmentation network is trained under the supervision of these pseudo ground-truths. However, CAMs tend to activate to small discriminative parts

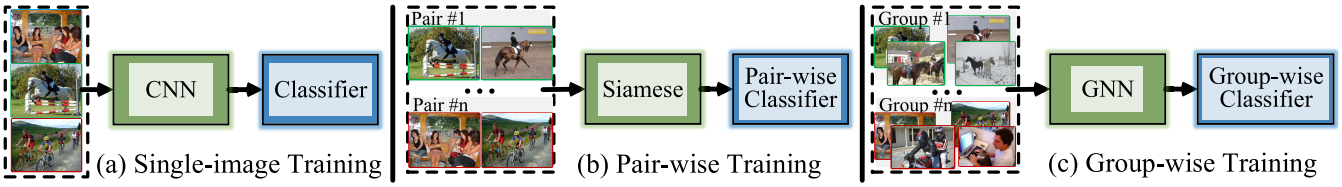


Fig. 2. **Conceptual architectures of existing approaches vs ours.** (a) Single-image models feed each image one by one into the network for training, which bears high similarity with standard classifiers (*e.g.,* VGG [1], ResNet [2]). (b) Pair-wise methods extract features from a pair of images using a Siamese network, and make predictions using a pair-wise classifier which has learned the correlation between the two images. (c) In contrast, we devise a group-wise method that accepts an arbitrary number of images as input. The group images are processed by a GNN within an iterative reasoning process, which enables information propagation to improve image representations. Finally, a group-wise classifier is employed for prediction.

of objects rather than complete object regions, resulting in inaccurate pseudo ground-truths for training semantic segmentation networks. To address this problem, many studies [13], [16], [38] hide or erase discriminative feature responses to force the model to focus on other relevant parts. A few methods [14], [39], [40] apply region growing to expand initial CAM-activated responses to cover the full extent of objects. Some recent efforts [17], [19], [41] investigate unsupervised signal (through self-supervised learning) to enrich feature representations and improve initial class activation maps. In addition, some works achieve more complete activation responses by learning pixel affinities [30], stochastic feature selection [42], joint learning with saliency detection [43] or casual inference [44].

All methods mentioned above consider each image individually. Recent efforts [20], [21] extend this into a pair-wise paradigm, employing Siamese networks to exploit correlations within image pairs to improve activation responses. Although impressive, we claim that these methods are still limited in capturing contextual information from only two images, missing higher-order relationship among image collection.

In this work, we take a further step towards discovering more comprehensive relations among groups of images. This is achieved by a graph neural network, which conducts iterative reasoning to refine image representations by propagating informative common semantics over the graph. Iterative learning [45] has also demonstrated its effectiveness in mining pseudo labels for semi-supervised learning. Moreover, to reduce the negative influences of unshared semantics, we devise a graph dropout layer which performs random feature selection during the graph inference to identify unshared object regions.

### B. Weakly Supervised Object Localization

WSOL aims to predict coarse bounding boxes for each object, instead of pixel-level semantic categories as done WSSS. Current WSOL approaches can be roughly divided into two categories: multiple instance learning based methods [46]–[48] and CAM based methods [15], [16], [49]–[52]. Here, we review CAM based methods since they are more similar to our approach. Most of these methods are also designed to discover complete object regions rather than small discriminative parts. For example, [49], [53] exploit adversarial erasing techniques to suppress discriminative regions and highlight other relevant parts. ADL [15] introduces an attentive dropout layer to facilitate network learning. SPG [50] discovers auxiliary supervision from high-confident response regions and gradually improves the activation maps to cover the whole objects. Regional dropout methods randomly remove regions from images [52], [54] or features [55] to improve the localization ability of CNNs. In this article, we transfer the proposed group-wise learning framework to the WSOL task to demonstrate its generalizability. With slight modification, our model shows compelling performance on WSOL.

### C. Graph Neural Networks

The concept of graph neural networks was first introduced in [56] as a generalization of recursive neural networks (RNNs) that can directly deal with a more general class of graph structure, *e.g.,* cyclic, directed and undirected graphs. Due to their convincing performance and high interpretability, GNNs have gained widespread attention for graph data analysis. Graph convolutional networks (GCNs) [57] directly define convolutions on the graph, and operate on groups of spatially close neighbors for message propagation. Graph attention networks (GATs) [58], [59] further incorporate a self-attention mechanism into the propagation step to compute the hidden states of each node by attending over its neighbors, leading to a leap in model capacity. Moreover, message passing graph networks (MPGNs) [60] abstract the commonalities among these popular GNN models. They model all graph elements (*i.e.,* nodes, edges) and iterative inference as learnable neural networks, gaining higher flexibility and learning capacity.

GNNs have achieved tremendous success in a variety of fields, including molecular biology [60], computer vision [61]–[65], medical image analysis [66] and machine learning [58], [67]. Motivated by this, we formulate group-wise learning within a GNN, in which each node represents an image, and each edge characterizes the semantic relations between image pairs. Through structured modeling and iterative reasoning, our model is able to mine comprehensive semantics from the graph, which alleviates the insufficient-label problem in WSSS.

### D. Visual Learning From Multiple Images

A typical paradigm in computer vision is to solve problems (*e.g.,* classification, detection, segmentation) for each individual image (or video) independently. However, they

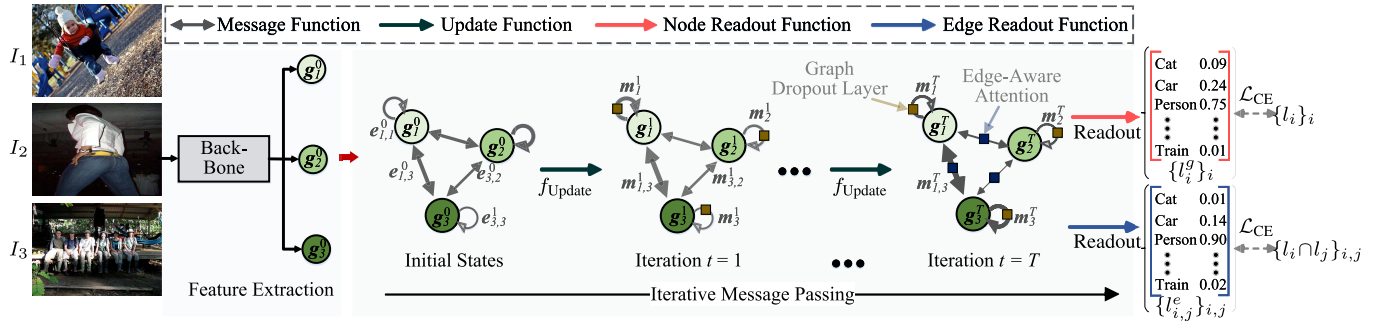


Fig. 3. **Overview of our group-wise learning framework** during the training phase. Given a group of images (*i.e.,* $\{I_i\}_{i=1}^3$), our model uses a backbone (*e.g.,* VGG16) to extract convolutional features (*i.e.,* $\{\boldsymbol{g}_i^0\}_{i=1}^3$), which are employed as the initial embeddings for graph construction. Next, our model conducts $T$-step message passing (Section IV-B) to iteratively refine the features by collecting shared semantic information across the graph. At each iteration, we introduce a graph dropout layer (Section IV-C) to force the network to focus more on unshared semantics. Finally, we apply two readout functions on graph nodes and edges, respectively, to predict the classification results. The whole network is supervised by multi-granularity losses (Section IV-D).

neglect valuable knowledge implicitly within visual data, which deserves careful studies to gain a more comprehensive understanding of scenes. Recently, [68] leverages memory-augmented networks to store and access relevant information of all available frames in an image sequence for segmentation. [11] learns intrinsic structures of labeled data by contrastive learning for fully supervised semantic segmentation task. Additionally, some works [69]–[71] model the interactions or matchings of multiple images in a group to segment visually similar objects at the same time. For example, [69] learns the relevance of group images through a recurrent neural network, under the supervision of pixel-level annotations. [70], [71] addresses weakly supervised object co-segmentation by seeking for explicit matching between common objects. However, one underlying assumption in these co-segmentation approaches is that there is ***only one*** common visual pattern (*e.g.,* `horse` or `human ride horse`) in each group. Thus, they are susceptible to noisy data and weak to generalize complex scenarios that multiple common object patterns exist. In contrast, our approach provides a more flexible and principled GNN solution, which is able to mine complex relations for an arbitrary number of common objects through pairwise co-attention.

## III. Preliminary

We start by revisiting the basic concept of GNNs. We define a graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$ by its node set $\mathcal{V}=\{v_1,\ldots,v_n\}$ and edge set $\mathcal{E}=\{e_{i,j}=(v_i,v_j)|v_i,v_j\in\mathcal{V}\}$. We assume that each node $v_i$ is associated with a feature embedding vector $\boldsymbol{g}_i$, and each edge $e_{i,j}$ has an edge representation $\boldsymbol{e}_{i,j}$. During inference, GNNs iteratively improve the feature representations at a node by aggregating its neighborhood features. Specifically, a GNN maps the graph $\mathcal{G}$ to the node outputs through two phases: a message passing phase and a readout phase. The message passing phase is defined in terms of a message function $f_{\text{Passing}}$, whose input is a node's features and output is a message, and an update function $f_{\text{Update}}$, whose input is a set of messages and output is the updated features. Suppose we conduct $T$ rounds of message passing; the $t$-th round for a node $v_i$ can be described as:

$$\boldsymbol{m}_i^t = \sum_{v_j\in\mathcal{N}_i} f_{\text{Passing}}^t(\boldsymbol{g}_j^{t-1},\boldsymbol{e}_{i,j}), \tag{1}$$

$$\boldsymbol{g}_i^t = f_{\text{Update}}(\boldsymbol{g}_i^{t-1},\boldsymbol{m}_i^t), \tag{2}$$

where for $v_i$, the message function firstly summarizes the information (*i.e.,* $\boldsymbol{m}_i^t$) from its neighbors $\mathcal{N}_i$, and then uses it to update the node state. Then, in the readout phase, a task-specific readout function $f_{\text{Readout}}$ operates on the final node representation $\boldsymbol{g}_i^T$ to produce a node output:

$$\hat{l}_i = f_{\text{Readout}}(\boldsymbol{g}_i^T). \tag{3}$$

## IV. Group-Wise Learning Framework

In this section, we elaborate on the proposed group-wise learning framework for WSSS. Given training images with only image-level labels, current efforts operate on two stages to achieve pixel-wise predictions. The first stage is *pseudo ground-truth generation*, which relies on an image classification network to localize object regions. The second stage is *semantic segmentation*, which conducts dense predictions using a fully convolutional network under the supervision of the pseudo labels. Our approach also adopts this two-stage pipeline. However, unlike previous approaches that treat each single image independently, our model aims to mine common semantic patterns from multiple images by iterative graph reasoning. In this way, our model can alleviate the incomplete-annotation problem in WSSS and produce more accurate pseudo labels.

### A. Problem Formulation

In sharp contrast to existing methods, our approach accepts a group of semantically relevant training images as input. Our goal is to exploit the semantic correlations of images to earn more comprehensive object pattern understanding, eventually yielding more accurate pseudo ground-truth for each group image. To achieve this, we formulate the problem as graph-based semantic co-mining from the input group. Formally, we denote $\mathcal{D}=\{(I_n,l_n)\}_{n=1}^N$ as the training data,

where $I_n \in \mathbb{R}^{w \times h \times 3}$ is an image and $l_n \in \{0,1\}^L$ is the corresponding image label with $L$ possible semantic categories. During training, we *selectively* sample $K$ images $\{I_i\}_{i=1}^K$ that at least share a common label as a mini-batch, and model their relations as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the image $I_i$ is denoted as node $v_i \in \mathcal{V}$, and the relation between $v_i$ and $v_j$ is represented by edge $e_{i,j} \in \mathcal{E}$. To better capture more comprehensive common semantics, we build a fully connected graph, *i.e.,* any two nodes $v_i$ and $v_j$ are linked by an edge and each node has a self-edge itself.

Given the above definitions, our network aims to conduct pseudo ground-truth generation in a graph learning scheme, under the full supervision of image-level labels as well as the implicit semantic relations among different images. In this manner, our model can capture richer semantic information and obtain more accurate pseudo labels. Next, we detail our group-wise learning framework based on GNN.

### B. GNN for Group-Wise Learning

*1) Initial Node Embedding:* As an initial step, we abstract a high-level feature representation for each input image. Formally, given $I_i$, we extract features $\boldsymbol{g}_i^0 \in \mathbb{R}^{W \times H \times C}$ from the convolutional stages of a standard classification network (*e.g.,* VGG-16 [1]). The initial state of node $v_i$ is then set as $\boldsymbol{g}_i^0$, which is a $(W, H, C)$-dimensional tensor preserving full spatial details for more effective pixel-level matching.

*2) Edge Embedding:* For each edge $e_{i,j}$ connecting $v_i$ to $v_j$, we aim to learn an edge embedding $\boldsymbol{e}_{i,j}^t$ at each iteration $t$ to find the correct semantic correspondence between the two nodes. This is achieved by dense matching over node embeddings using the following bilinear model:

$$\boldsymbol{e}_{i,j}^t = \boldsymbol{g}_i^t \boldsymbol{W} \boldsymbol{g}_j^{t\top} \in \mathbb{R}^{WH \times WH}, \tag{4}$$

where $\boldsymbol{g}_i^t \in \mathbb{R}^{WH \times C}$ and $\boldsymbol{g}_j^t \in \mathbb{R}^{WH \times C}$ are flattened into matrix representations for computational convenience. $\boldsymbol{W} \in \mathbb{R}^{C \times C}$ is a trainable weight matrix. In Eq. 4, $\boldsymbol{e}_{i,j}^t$ encodes the similarity between $\boldsymbol{g}_i^t$ and $\boldsymbol{g}_j^t$ for all pairs of spatial locations. For the edge $e_{j,i}$, its embedding at iteration $t$ is simply calculated as $\boldsymbol{e}_{j,i}^t = \boldsymbol{e}_{i,j}^{t\top}$.

It should be noted that Eq. 4 introduces a large number of parameters, increasing the computational cost. To alleviate this, $\boldsymbol{W}$ is approximately factorized into two low-rank matrices $\boldsymbol{P} \in \mathbb{R}^{C \times \frac{C}{d}}$ and $\boldsymbol{Q} \in \mathbb{R}^{C \times \frac{C}{d}}$, where $d$ $(d = 4)$ is a reduction ratio. Then, Eq. 4 can be rewritten as:

$$\boldsymbol{e}_{i,j}^t = \boldsymbol{g}_i^t \boldsymbol{P} \boldsymbol{Q}^\top \boldsymbol{g}_j^{t\top} \in \mathbb{R}^{WH \times WH}. \tag{5}$$

Eq. 5 has significant advantages over Eq. 4 in both model parameters and computational cost: 1) it reduces the number of parameters by $2/d$ times; 2) it only requires $(2WHC^2 + W^2H^2C)/d$ multiplication operations, instead of the $WHC^2 + W^2H^2C$ in Eq. 4.

In addition, for each self-edge $e_{i,i}$, its embedding $\boldsymbol{e}_{i,i}$ captures the self-relation over the node representation $\boldsymbol{g}_i$. We compute $\boldsymbol{e}_{i,i}^t$ at step $t$ by self-attention [72], [73], which can effectively capture non-local semantic dependencies. In particular, the self-attention calculates the response at a position by attending to all the positions within the same node embedding:

$$\boldsymbol{e}_{i,i}^t = \text{softmax}(\phi_f(\boldsymbol{g}_i^t)\phi_g^\top(\boldsymbol{g}_i^t))\phi_h(\boldsymbol{g}_i^t) + \boldsymbol{g}_i^t \in \mathbb{R}^{W \times H \times C}, \tag{6}$$

where $\phi_{\{f,g,h\}}$ are $1 \times 1$ convolutional operators. We also use a residual layer in Eq. 6, which can effectively preserve information in the original feature map.

*3) Iterative Message Passing:* Given the node and edge embeddings, our model iteratively updates the hidden states of graph nodes by applying message functions to collect contextual information from their neighboring nodes. More specifically, for a node $v_i$, it absorbs knowledge along two types of edges: 1) a self-edge $e_{i,i}$ that encodes rich context-aware knowledge in $v_i$; and 2) other edges $\{e_{j,i}\}_j$ that connect $v_j$ to $v_i$. For the former, our message function directly reads the message from $\boldsymbol{e}_{i,i}$, *i.e.,* $\boldsymbol{m}_{i,i}^t = \boldsymbol{e}_{i,i}^{t-1}$, while for the latter, the messages are summarized as:

$$\boldsymbol{m}_{j,i}^t = \text{softmax}(\boldsymbol{e}_{i,j}^{t-1})\boldsymbol{g}_j^{t-1} \in \mathbb{R}^{WH \times C}, \tag{7}$$

where softmax denotes the row-wise softmax normalization. In Eq. 7, we accumulate knowledge from $\boldsymbol{g}_j^{t-1}$, which is weighted based on the similarity between $\boldsymbol{g}_i^{t-1}$ and $\boldsymbol{g}_j^{t-1}$. $\boldsymbol{m}_{j,i}^t$ is then reshaped to a $(W, H, C)$-dimensional tensor. Then, we can easily summarize the message for $v_i$ at the $t$-th iteration:

$$\boldsymbol{m}_i^t = \sum_{v_j \in \mathcal{N}_i} \boldsymbol{m}_{j,i}^{t-1} + \boldsymbol{m}_{i,i}^{t-1}. \tag{8}$$

Next, the update function $f_{\text{Update}}$ updates the hidden states of nodes, as done in Eq. 2. In our method, $f_{\text{Update}}$ is instantiated by a ConvGRU network [74], which is an extension of the GRU update function used in [60]. We run the message passing algorithm for $T$ steps in total to iteratively collection messages and update node embeddings.

*4) Readout Phase:* Given the final node embedding $\boldsymbol{g}_i^T \in \mathbb{R}^{W \times H \times C}$ for $v_i$, we derive the final prediction for the corresponding image through a readout phase:

$$\boldsymbol{q}_i = f_{\text{CAM}}(\boldsymbol{g}_i^T) \in \mathbb{R}^{W \times H \times L}, \quad \hat{l}_i^g = f_{\text{GAP}}(\boldsymbol{q}_i) \in \mathbb{R}^L, \tag{9}$$

where $f_{\text{CAM}}$ is a class-aware convolutional layer with kernel size $1 \times 1$ that produces a class-aware attention map $\boldsymbol{q}_i$, and $f_{\text{GAP}}$ denotes a *global average pooling* layer to produce the final classification prediction $\hat{l}_i^g$.

*5) Deeply Supervised Learning:* In addition to the readout (Eq. 9) from the final embeddings, we apply an intermediate readout phase on the initial convolutional embedding $\boldsymbol{g}_i^0$:

$$\boldsymbol{d}_i = f_{\text{CAM}}(\boldsymbol{g}_i^T) \in \mathbb{R}^{W \times H \times L}, \quad \hat{l}_i^m = f_{\text{GAP}}(\boldsymbol{d}_i) \in \mathbb{R}^L, \tag{10}$$

where $\hat{l}_i^m$ is the intermediate classification prediction. Then, for each image, we ensemble the class-aware attention maps $\boldsymbol{q}_i$ and $\boldsymbol{d}_i$ to obtain the final object localization estimation. We find that the pseudo ground-truths from different outputs are complementary to each other, and self-ensembling them through averaging can further improve the performance (see Table IV).

### C. Graph Dropout Layer

The above graph reasoning scheme enables our model to discover common semantics present in different images (Eq. 5). The features of these semantics can be accordingly enriched by summarizing all the information from other images (Eq. 8). However, standalone categories, which may exist only in a single image, are almost entirely ignored in this procedure. To address this, we introduce a graph dropout layer to force the network to pay more attention to these categories. Formally, given the feature map $\boldsymbol{g}_i^t \in \mathbb{R}^{W\times H\times C}$ at the $t$-th iteration, we average it along the channel dimension to obtain $\boldsymbol{o}_i^t \in \mathbb{R}^{W\times H}$. Then, we generate a mask $\boldsymbol{s}^t \in \mathbb{R}^{W\times H}$ as follows:

$$\boldsymbol{s}_i^t = \begin{cases} \text{sigmoid}(\boldsymbol{o}_i^t), & \text{if } \ r < \delta_r; \\ \boldsymbol{o}_i^t \mathbb{1}(\boldsymbol{o}_i^t < \max(\boldsymbol{o}_i^t) * \delta_d), & \text{otherwise.} \end{cases} \tag{11}$$

Here, the parameter $\delta_r \in [0, 1]$ is a drop rate threshold, determining whether to carry out the dropout operation or not. The parameter $r$ is a scalar generated from a random generator. If $r < \delta_r$, $\boldsymbol{s}_i^t$ is an importance map which supports the activations in $\boldsymbol{g}_i^t$; otherwise, the layer drops the highly activated semantic regions to emphasize standalone semantics. $\mathbb{1}(x)$ is a matrix indicator function which returns 1 for the true elements in $x$, and 0 otherwise. The $\max(\cdot)$ operation calculates the maximum value for a 2D tensor. $\delta_d \in [0, 1]$ is a threshold controlling the dropout. Finally, we enhance the feature maps by:

$$\hat{\boldsymbol{g}}_i^t = \boldsymbol{g}_i^t \otimes \boldsymbol{s}_i^t, \tag{12}$$

where $\otimes$ denotes spatial-wise multiplication. Note that $\hat{\boldsymbol{g}}_i^t$ is then used to replace the original embedding $\boldsymbol{g}_i^t$ (Eqs. (5)-(7)) in the next iteration.

### D. Multi-Granularity Supervision

Our model is trained in an end-to-end manner with multi-granularity supervision:

$$\begin{aligned} \mathcal{L} = \lambda_1 \sum\nolimits_i \mathcal{L}_{\text{CE}}(\hat{l}_i^g, l_i) + \lambda_2 \sum\nolimits_i \mathcal{L}_{\text{CE}}(\hat{l}_i^m, l_i) \\ + \lambda_3 \sum\nolimits_i \sum\nolimits_j \mathcal{L}_{\text{CE}}(\hat{l}_{i,j}^e, l_i \cap l_j), \end{aligned} \tag{13}$$

where $\mathcal{L}_{\text{CE}}$ denotes the standard cross-entropy loss and $\{\lambda_i\}_{i=1}^3$ balances the three terms. Here, the first term evaluates the final outputs $\{\hat{l}_i^g\}_i$ (Eq. 9) to guide the learning of the GNN, while the second term evaluates the intermediate prediction $\{\hat{l}_i^m\}_i$ (Eq. 10). For the third term, we introduce an edge-aware supervision which encourages each edge in the graph to learn the shared semantics of the connected nodes. Here, $l_i \cap l_j \in \{0, 1\}^L$ represents a new label vector in which the items of shared semantic categories between nodes $v_i$ and $v_j$ are 1, and others are 0. $\hat{l}_{i,j}^e \in \mathbb{R}^L$ denotes the prediction of edge $e_{i,j}$.

*1) Edge-Aware Supervision:* For each edge $e_{i,j}$, we denote $\boldsymbol{g}_i^T \in \mathbb{R}^{W\times H\times C}$ and $\boldsymbol{g}_j^T \in \mathbb{R}^{W\times H\times C}$ as the learned final embeddings of nodes $v_i$ and $v_j$ at step $T$, and $\boldsymbol{g}_{i,j} = [\boldsymbol{g}_i^T, \boldsymbol{g}_j^T] \in \mathbb{R}^{W\times H\times 2C}$ as the concatenation of the two node features. Then, we learn a shared representation based on the dual-attention mechanism [75], which employs two parallel attention modules (*i.e.,* position-aware and channel-aware) to discover the shared semantics from $\boldsymbol{g}_{i,j}$.

For the position-aware attention module, we first compute the normalized cross-correlation between each pair of spatial feature vectors in $\boldsymbol{g}_{i,j}$ by the following pairwise dot product:

$$\boldsymbol{u}_{i,j} = \text{softmax}(\boldsymbol{g}_{i,j} \boldsymbol{g}_{i,j}^\top) \quad \in [0, 1]^{WH\times WH}. \tag{14}$$

Here, $\boldsymbol{g}_{i,j} \in \mathbb{R}^{WH\times C}$ is first flattened into a matrix representation for computational convenience. The affinity matrix $\boldsymbol{u}_{i,j}$ stores similarity scores corresponding to all pairs of features in $\boldsymbol{g}_{i,j}$. Next, attention summaries are computed based on $\boldsymbol{u}_{i,j}$, and used to generate a positive-attentive edge representation $\boldsymbol{p}_{i,j}$:

$$\boldsymbol{p}_{i,j} = \boldsymbol{u}_{i,j} \boldsymbol{g}_{i,j} + \boldsymbol{g}_{i,j} \quad \in \mathbb{R}^{WH\times 2C}. \tag{15}$$

For the channel-aware attention module, we first compute the cross-correlation between each pair of channel feature vectors in $\boldsymbol{g}_{i,j}$ as follows:

$$\boldsymbol{v}_{i,j} = \text{softmax}(\boldsymbol{g}_{i,j}^\top \boldsymbol{g}_{i,j}) \quad \in [0, 1]^{2C\times 2C}. \tag{16}$$

Here, $\boldsymbol{v}_{i,j}$ captures the channel-wise feature interdependencies in $\boldsymbol{g}_{i,j}$. Then, we obtain a channel-attentive edge feature map (similar to Eq. 15):

$$\boldsymbol{c}_{i,j} = \boldsymbol{g}_{i,j} \boldsymbol{v}_{i,j} + \boldsymbol{g}_{i,j} \quad \in \mathbb{R}^{WH\times 2C}. \tag{17}$$

The $\boldsymbol{p}_{i,j}$ in Eq. 15 and $\boldsymbol{c}_{i,j}$ in Eq. 17 capture long-range position- and channel-aware contextual information of the edge $e_{i,j}$, and we then combine them to obtain a more comprehensive representation. More specifically, we first reshape $\boldsymbol{p}_{i,j}$ and $\boldsymbol{c}_{i,j}$ into sizes of $W \times H \times 2C$, and apply two $1 \times 1$ convolutional layers separately to transform them into $(W, H, C)$-dimensional tensors. The transformed feature maps are fused by element-wise addition, and finally processed by a readout phase (similar with Eq. 9 and Eq. 10) to obtain the edge-aware prediction $\hat{l}_{i,j}^e \in \mathbb{R}^L$. We follow this procedure to achieve predictions for all the edges, and employ them in Eq. 13 to guide the learning of the GNN towards better discovering shared semantics in each group.

### E. Detailed Network Architecture

Our model consists of two sub-networks: a classification network for group-wise pseudo ground-truth generation and a semantic segmentation network for pixe-wise segmentation.

*1) Image Classification Network:* We use VGG16 [1] as our backbone network, which is initialized using weights pre-trained on ImageNet [76]. We replace the last convolutional layer in VGG16 by dilated convolutions with a rate of 2, and the feature maps from this layer are taken as the initial node representations for the GNN. After training, we obtain the CAMs for each training image from the two classification layers mentioned earlier, and combine them to obtain foreground object seeds. Moreover, we follow conventions [19], [77], [78] to estimate background seeds using an off-the-shelf salient object detection model [79]. The final pseudo labels are generated by combining the foreground and background seeds.

*2) Semantic Segmentation Network:* To make a fair comparison with existing methods [17], [19], [77], [78], we choose DeepLab-V2 [10] as the semantic segmentation network.

## V. Group-Wise Learning for WSOL

In addition to the WSSS task, our group-wise learning framework is also applicable to the WSOL task, which predicts the bounding box of each object using image-level labels only.

In particular, the training phase of the proposed framework is exactly the same for both the WSSS and WSOL tasks. However, during the testing phase, we need to feed a group of images to the trained classification network for prediction in WSOL, rather than a single image as the input of the segmentation network in WSSS. Therefore, we first employ another standalone VGG16 [1] to generate categorical predictions of each image in the test set, and accordingly sample a group of images as the input of our network for group-wise inference. For each test image, we can obtain its class activation map. Since the score statistics differ vastly across images, a common practice in WSOL is to normalize the activation maps per image. We choose the min-max normalization method in our experiment. After normalization, we utilize the *violent searching strategy* [15] to find an optimal threshold for bounding box generation.

## VI. Experiment

In this section, we first compare our method with state-of-the-art models on the WSSS task in Section VI-A. Then, we investigate the performance of our model on the WSOL task in Section VI-B. For each task, we introduce the corresponding datasets and their performance. To gain deeper insight into our model, we conduct detailed diagnostic experiments of the crucial components in Section VI-C. Lastly, we offer several failure cases for more comprehensive analysis in Section VI-E.

### A. Weakly Supervised Semantic Segmentation

*1) Datasets:* We conduct our experiments on two popular datasets: PASCAL VOC 2012 [22] and COCO [23]. 1) **PASCAL VOC 2012** [22] is currently the most popular benchmark for WSSS. The dataset contains 20 semantic categories (*e.g.,* person, bicycle, cow) and one background category. Following standard protocol [19], [40], [42], extra data from SBD [88] are also used for training, leading to a total of 10,582 training images. We evaluate our model on the standard validation and test sets, which have 1,449 and 1,456 images, respectively. 2) **COCO** [23] is a more challenging benchmark with 80 semantic classes. Since more complex contextual relations exist among these categories, it is interesting to examine the performance of our model in this dataset. Following [86], we use the default train/val splits (80k images for training and 40k for validation) in the experiment.

*2) Evaluation Metric:* We utilize the widely used *mean intersection-over-union (mIoU)* for evaluation. The scores on the test set of PASCAL VOC 2012 are obtained from the official evaluation server. We also consider CAM mIoU on the training set of PASCAL VOC 2012 to evaluate the qualities of generated CAM results.

TABLE I

QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON PASCAL VOC 2012 IN TERMS OF MIOU. *: VGG BACKBONE. †: RESNET BACKBONE

| Method | Publication | mIoU (%) | |
|---|---|---|---|
| | | val | test |
| *single-image based methods* | | | |
| *MEFF [80] | CVPR18 | - | 55.6 |
| *GAIN [81] | CVPR18 | 55.3 | 56.8 |
| *MDC [39] | CVPR18 | 60.4 | 60.8 |
| *RRM [82] | AAAI20 | 60.7 | 61.0 |
| †MCOF [83] | CVPR18 | 60.3 | 61.2 |
| †SeeNet [16] | NIPS18 | 63.1 | 62.8 |
| †DSRG [40] | CVPR18 | 61.4 | 63.2 |
| †AffinityNet [30] | CVPR18 | 61.7 | 63.7 |
| †SS-WSSS [84] | CVPR20 | 62.7 | 64.3 |
| †SSNet [43] | ICCV19 | 63.3 | 64.3 |
| †IRNet [85] | CVPR19 | 63.5 | 64.8 |
| †FickleNet [42] | CVPR19 | 64.9 | 65.3 |
| †IAL [86] | IJCV20 | 64.3 | 65.4 |
| †SSDD [87] | ICCV19 | 64.9 | 65.5 |
| †SEAM [19] | CVPR20 | 64.5 | 65.7 |
| †SubCat [17] | CVPR20 | 66.1 | 65.9 |
| †OAA+ [77] | ICCV19 | 65.2 | 66.4 |
| †RRM [82] | AAAI20 | 66.3 | 66.5 |
| †BES [78] | ECCV20 | 65.7 | 66.6 |
| †ICD [18] | CVPR20 | 67.8 | 68.0 |
| *pair-wise based methods* | | | |
| †CIAN [20] | AAAI20 | 64.3 | 65.3 |
| †MCIS [21] | ECCV20 | 66.2 | 66.9 |
| *group-wise based methods* | | | |
| ***Ours** | – | 63.7 | 64.1 |
| †**Ours** | – | **68.7** | **69.0** |

*3) Training Details:* 1) *Greedy Mini-Batch Sampling.* During training, we design a heuristic, greedy strategy to sample $K$ training images in each mini-batch. Starting from a randomly sampled image $I_i$, we further find another $K-1$ images, each of which shares as many common semantic objects with $I_i$ as possible. These $K$ images are then used to build a $K$-node GNN. This sampling strategy enables our model to better explore rich relationships among groups of images and improve the results. 2) *Training Settings.* For the classification network, the number of nodes $K$ and message passing steps $T$ in the GNN are separately set to 4 and 3 by default. The input image size is $224 \times 224$. The entire network is trained using the SGD optimizer with initial learning rates of 1e-3 for the backbone and 1e-2 for the GNN, which are reduced by 0.1 every five epochs. The total number of epochs, momentum and weight decay are set to 15, 0.9, and 5e-4, respectively. The $\lambda_1$, $\lambda_2$ and $\lambda_3$ in Eq. 13 are set to 0.7, 0.1, 0.2, respectively. For the segmentation network, we follow the training setting in [10], but use the pseudo ground-truths as the supervision.

*4) Performance on PASCAL VOC 2012:* We evaluate the proposed approach on PASCAL VOC 2012 against current top-performing WSSS methods that only operate with image-level labels. Following conventions, we evaluate the performance of our model using VGG16 [1] and ResNet101 [2] as the backbones, respectively. As reported in Table I, our model with ResNet101 achieves the best performance, scoring

Fig. 4. **Qualitative results** on PASCAL VOC 2012 `val`. From top to bottom: input images, ground-truths, and our results.

TABLE II

QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON COCO `val` IN TERMS OF MIOU. ALL METHODS USE VGG16 AS THE BACKBONE

| Method | Publication | mIoU (%) |
|---|---|---|
| BFBP [89] | ECCV16 | 20.4 |
| SEC [14] | ECCV16 | 22.4 |
| DSRG [40] | CVPR18 | 26.0 |
| IAL [86] | IJCV20 | 27.7 |
| **Ours** | – | **28.7** |

TABLE III

QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON CUB-200-2011 `test` IN TERMS OF TOP-1 CLS AND TOP-1 LOC. ALL METHODS USE VGG16 AS THE BACKBONE

| Method | Publication | Top-1 Cls (%) | Top-1 Loc (%) |
|---|---|---|---|
| CAM [12] | CVPR16 | 69.95 | 37.05 |
| AcoL [49] | CVPR18 | 71.90 | 45.92 |
| SPG [50] | ECCV18 | 75.50 | 48.93 |
| HaS-32 [38] | ICCV17 | 76.10 | 49.46 |
| DANet [51] | ICCV19 | 75.40 | 52.52 |
| CutMix [52] | ICCV19 | – | 52.53 |
| ADL [15] | CVPR19 | 65.27 | 52.36 |
| $I^2C$ [90] | ECCV20 | – | 55.99 |
| EIL [53] | CVPR20 | 72.26 | 56.21 |
| HaS-32 + RCAM [91] | ECCV20 | 70.12 | 57.37 |
| MEIL [53] | CVPR20 | 74.77 | 57.46 |
| GC-Net-Elli [92] | ECCV20 | **76.80** | 58.85 |
| ADL + RCAM [91] | ECCV20 | 75.01 | 58.96 |
| PSOL [93] | CVPR20 | – | 59.29 |
| **Ours** | – | 76.44 | **60.18** |

**68.7%** and **69.0%** on the *val* and *test* sets, respectively. It significantly outperforms the current leading approach, *i.e.,* ICD [18], by **+0.9%** and **+1.0%** on the two evaluation sets.

In addition, Table I also shows that the proposed approach outperforms both pair-wise models (*i.e.,* CIAN [20] and MCIS [21]), and all single-image based models (*e.g.,* RRM [82], OAA+ [77]), by a large margin. The reason lies in that existing methods exploit limited context in image collection, while our approach can learn more effective inter-image representations with GNNs.

In Fig. 4, we provide sample results for representative images in PASCAL VOC 2012 `val`. The images cover various challenging factors in WSSS, such as multiple objects, different semantic categories, small objects, and cluttered background. We see that our approach can deal with these difficulties well, resulting in appealing segmentation results.

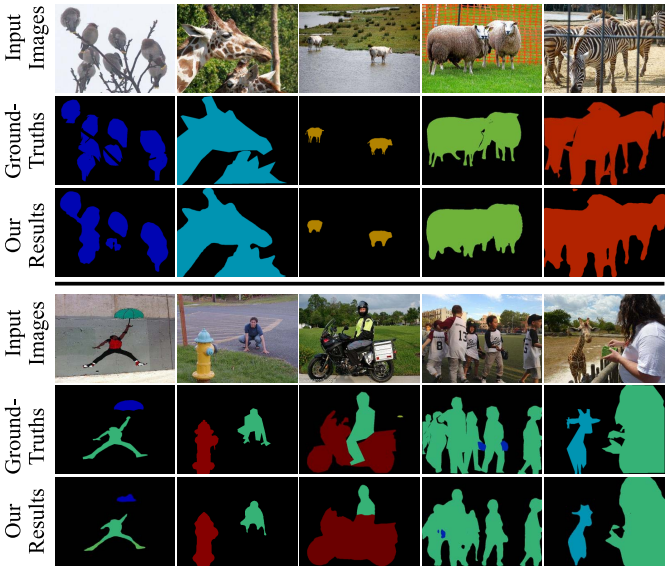


Fig. 5. **Qualitative segmentation results** on COCO `val`.

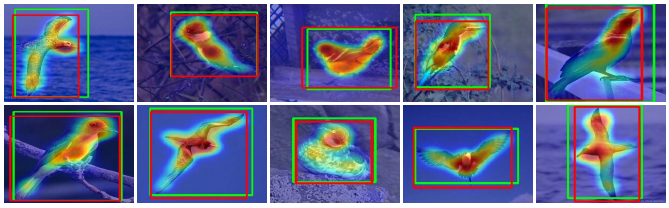

Fig. 6. **Qualitative results on CUB-200-2011 `test`**. For each image, we show its ground-truth box (in red), localization result (in green), as well as the overlayed class activation map.

*5) Performance on COCO:* We further examine the performance of our model on COCO. As reported in Table II, our model achieves the best mIoU score (*i.e.,* **28.7%**) on the validation set, outperforming the second-best model, *i.e.,* IAL [86], by **1.0%**. This further proves the powerful capability of our model in WSSS. Fig. 5 illustrates some representative segmentation results of our model on COCO `val`. We see that, even with only image-level labels, our approach can produce high-quality segmentation results in many challenging scenarios (*e.g.,* small objects, crowded scenes).

### *B. Weakly Supervised Object Localization*

*1) Datasets:* We use CUB-200-2011 [24] to evaluate the performance of our model on WSOL. The dataset consists of

TABLE IV
**DIAGNOSTIC EXPERIMENTS OF OUR MODEL** ON PASCAL VOC 2012 `val` IN TERMS OF MIOU. FOR ALL VARIANTS, WE USE RESNET101 AS THE BACKBONE

| Aspect | | Variant | mIoU (%) |
|---|---|---|---|
| **Full Model** | | $T = 3$, $K = 4$<br>$\delta_r = 0.8$, $\delta_d = 0.7$ | **68.7** |
| Graph Reasoning | Graph Node Number | $K = 2$ | 66.1 |
| | | $K = 3$ | 68.6 |
| | | $K = 5$ | 68.3 |
| | | $K = 6$ | 68.1 |
| | Message Passing Steps | $T = 2$ | 68.3 |
| | | $T = 4$ | 68.5 |
| | | $T = 5$ | 68.5 |
| | Graph Dropout Layer | $\delta_r = 0.8$, $\delta_d = 0.9$ | 68.5 |
| | | $\delta_r = 0.8$, $\delta_d = 0.5$ | 68.2 |
| | | $\delta_r = 0.6$, $\delta_d = 0.7$ | 67.3 |
| | | $\delta_r = 0.4$, $\delta_d = 0.7$ | 64.1 |
| | | *w/o* dropout | 68.2 |
| Multi-Granularity Supervision | | *w/o* intermediate supervision | 68.4 |
| | | *w/o* edge-aware supervision | 68.2 |
| Self-Ensembling | | intermediate output *only* | 64.6 |
| | | graph output *only* | 68.3 |
| Sampling Strategies | | random | 67.1 |
| | | one-common | 68.3 |

200 species of birds, with 11,788 images in total. Following [15], [28], [49], [93], we adopt the default splits in our experiments, including 5,994 for `train` and 5,794 for `test`.

*2) Metrics:* We use two standard evaluation metrics [15], [38] for evaluation: 1) Top-1 localization accuracy (*Top-1 Loc*) denotes the fraction of images for which the evaluated bounding box for the ground-truth class has more than 50% IoU with the ground-truth bounding box, as well as the predicted class with the highest probability is equal to the ground-truth class. 2) Top-1 classification accuracy (*Top-1 Cls*) indicates that the result is correct when the predicted class with the highest probability is equal to the ground-truth class.

*3) Performance on CUB-200-2011:* Table III reports the results of our approach in comparison with state-of-the-art methods on CUB-200-2011 `test`. As can be seen, our model achieves compelling performance on the WSOL task. In particular, it achieves the best Top-1 Loc score (*i.e.,* **60.18%**), leading to about 1% improvement compared with the second-best method (*i.e.,* PSOL [93]). For the Top-1 Cls, our model is slightly worse than GC-Net-Elli [92] (76.44% *vs.* 76.80%), but significantly outperforms all other methods. In summary, our group-wise learning framework shows strong robustness in the WSOL task.

Fig. 6 provides some representative localization results of our model on CUB-200-2011 `test`. Benefiting from group-wise learning, our model can obtain complete activation maps, leading to accurate bounding box localization.

### C. Diagnostic Experiment

We further conduct diagnostic analysis on PASCAL VOC 2012 `val` to verify the effectiveness of the essential modules in our approach. We use ResNet101 as the default backbone for all the studies. The performance of our full model with default parameters is given in the first row of Table IV.

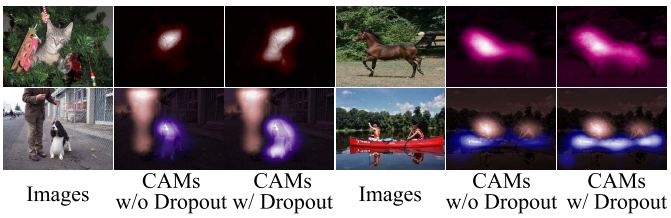


Fig. 7. **Visual comparisons of CAMs** generated with or without the graph dropout layer.

*1) Group-Wise Learning:* We first investigate the effect of the node number $K$ used in the GNN, which indicates the number of images in a group. As shown in Table IV, the model achieves comparably high performance with three or four nodes. However, when more nodes are added, the performance decreases significantly. This can be attributed to the trade-off between meaningful semantic relations and noise brought by input group images. When $K = 3$ or 4, the semantic relations can be fully exploited to improve the integral regions of objects. However, when more images are further considered, meaningful semantic cues reach a bottleneck and noise, introduced by imperfect localization of the classifier, dominates, thus leading to performance degradation.

We further evaluate the impact of the message passing steps by comparing the performance with different $T$ ranging from 2 to 5. From Table IV, we observe that the mIoU score is significantly improved when $T$ varies from 2 to 3. The performance decreases slightly when considering more steps. Therefore, we set $T = 3$ as default for message passing.

Moreover, we build a baseline model (*i.e., w/o group-wise learning* in Table V) which discards the group-wise learning mechanism. As can be seen, the CAM and segmentation results encounter significant drops, which demonstrates the importance of our group-wise learning mechanism.

*2) Graph Dropout Layer:* To examine the proposed graph dropout layer, we design multiple experiments to search an optimal configuration of parameters $\delta_r$ and $\delta_d$. We observe that both parameters have great influences on the performance. As observed in Table IV, our model reaches the best performance at $\delta_r = 0.8$ and $\delta_d = 0.7$. If the $\delta_d$ is higher (*e.g.,* 0.9), most discriminative regions will be kept, and thus inactivated regions will remain being ignored. In contrast, if $\delta_d$ is lower, the regions with high responses will be excessively dropped, leading to degraded classification accuracy.

In addition, the parameter $\delta_r$ controls whether or not the responses should be dropped during training. As shown in Table IV, a $\delta_r$ of 0.8 helps to achieve the best mIoU score. Such a setting not only maintains the classification ability of the network by keeping discriminative regions with a high probability, but also drives the network to mildly attend to other regions. We can also see that by setting $\delta_r$ to smaller values (*e.g.,* 0.6 or 0.4), the performance encounters a significant decrease.

Moreover, we examine the performance of our model without the graph dropout layer (*i.e., w/o graph dropout* in

TABLE V

ABLATION STUDY OF CAM RESULTS ON PASCAL VOC 2012 `train` AS WELL AS SEGMENTATION RESULTS ON `val`

| Variant | CAM mIoU (%) | | SEG mIoU (%) |
|---|---|---|---|
| | VGG16 | ResNet38 | |
| w/o Group-Wise Learning | 56.3 | 58.7 | 64.5 |
| w/o Graph Dropout | 61.9 | 64.9 | 68.2 |
| Full Model | **62.9** | **65.7** | **68.7** |

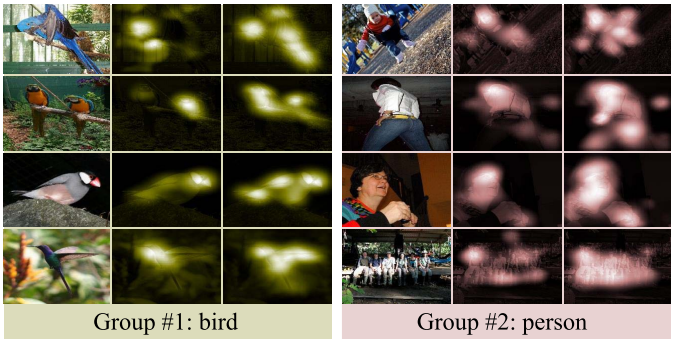


Fig. 8. **CAM visual comparisons**. We provide results for two groups of images. For each group, we show the input images, CAMs from the *intermediate readout layer* and CAMs from the *graph readout layer* (from left to right). Our model clearly yields more accurate CAMs after group-wise graph reasoning.

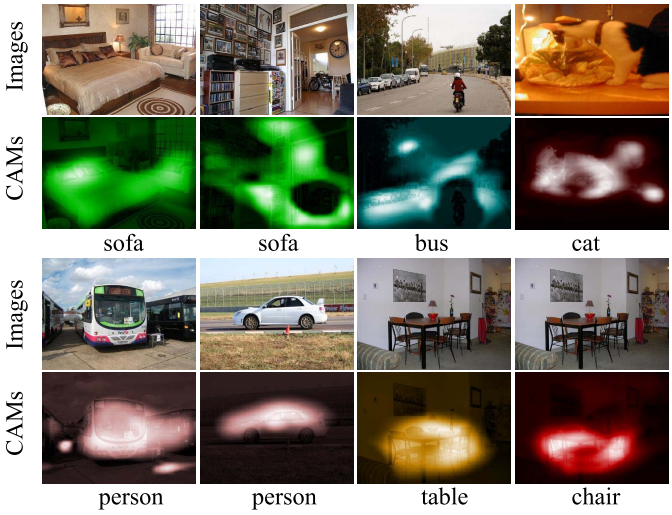


Fig. 9. **Failure cases of CAMs** on PASCAL VOC 2012 `val`.

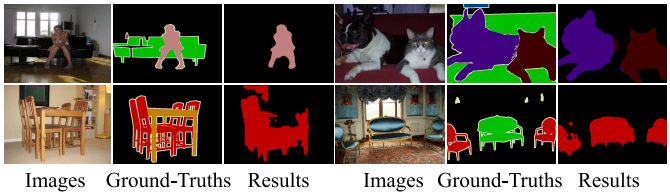


Fig. 10. **Failure cases of segmentation results** on PASCAL VOC 2012 `val`.

Table V). We can see that all metrics consistently decrease after discarding the graph dropout layer, revealing its necessity.

Finally, we illustrate some examples of the final CAMs generated *with* or *without* the graph dropout layer. As shown in Fig. 7, without the dropout layer, the network only focuses on the most discriminative parts (*e.g.,* heads of the cat and the horse). This is improved with our dropout layer, which helps to highlight non-discriminative object regions.

*3) Multi-Granularity Supervision:* We further study the multi-granularity supervision in Eq. 13. We design two baseline models, *i.e., w/o* intermediate supervision and *w/o* edge-aware supervision. As reported in Table IV, both baseline models show worse performance than our full model with multi-level supervision. We can also see that by dropping the edge-aware supervision, the baseline model encounters a great performance degradation (*i.e.,* 0.5%), revealing the importance of the edge-aware constraints.

*4) Self-Ensembling:* In addition to the supervision on the final outputs, we also introduce deep supervision signals on the intermediate features. Such multi-level supervision has proven effective for improving the performance of various vision tasks. Besides, this enables us to combine the multiple outputs with low cost to further boost the performance. Here, we examine the self-ensembling strategy by building three network variants, *i.e., intermediate output*, *graph output* and *self-ensembling*, in which the final CAMs are separately extracted from the intermediate readout layer, graph-aware readout layer, and their ensemble, respectively. As shown in Table IV, the *intermediate output* only obtains an mIoU score of 64.6%, greatly lagging behind the 68.3% obtained by the *graph output*. This demonstrates that through iterative graph reasoning, our model can improve the image representations by integrating information from image groups, leading to huge performance gains. Furthermore, the self-ensembling strategy boosts the performance to 68.7%.

In Fig. 8, we illustrate two groups of images with their CAMs from the *intermediate readout layer* and *graph readout layer*. As can be seen, in both groups, the CAMs are well-refined to cover more complete foreground regions after graph reasoning. Besides, in many cases, the CAMs from two output layers complement each other well, enabling better results to be obtained by self-ensembling.

*5) Greedy Mini-Batch Sampling:* We also conduct experiments to verify the effectiveness of our greedy mini-batch sampling strategy. For comparison, we design two baseline sampling methods, *i.e., random sampling* and *one-common sampling*. The *random sampling* method randomly samples $K$ training images as a mini-batch, while the *one-common sampling* method finds $K$ training images with at least one common semantic category. As reported in Table IV, the *random sampling* method only obtains an mIoU score of 67.1%, largely lagging behind the *one-common sampling* and *greedy sampling* methods. Moreover, our *greedy sampling* method encourages more comprehensive semantic discovery, yielding the best result (*i.e.,* 68.7%).

### *D. Model Efficiency Analysis*

In our approach, iterative graph reasoning is only used in the training stage for pseudo ground-truth generation, which takes

about eight hours for training on PASCAL VOC 2012. The training of the segmentation network is almost same to most existing WSSS approaches [17], [21], [78]. During inference, only the segmentation network will be used for segmentation prediction. Thus, the inference speed of our model is the same with previous approaches (about 12 fps under a NVIDIA RTX 2080Ti GPU).

### *E. Failure Case Analysis*

To gain deeper insight into our model, we provide several typical cases where our approach fails on PASCAL VOC 2012 `val`. In Fig. 9, we illustrate the failure cases of generated class activation maps for particular semantic categories. As can be seen, the failures are mainly caused by i) incorrect recognition due to similar appearance between some objects (*e.g.,* sofa and bed), ii) unrecognized small or occluded objects (*e.g.,* person in bus or car), iii) background distractors (*e.g.,* bus). The failure of CAM estimation will directly result in poor segmentation results (see Fig. 10).

## VII. Conclusion

In this work, we introduce a novel group-wise learning framework for weakly supervised semantic segmentation (WSSS). Unlike previous single-image or pair-wise based approaches, our framework is able to capture comprehensive semantic context to generate more accurate pseudo ground-truths. In particular, we formulate group-wise learning within a graph neural network, which operates on a group of images and conducts iterative graph reasoning to discover meaningful semantics. Moreover, we devise a graph dropout layer which randomly hides the most discriminative parts from the model to capture the integral extent of object regions. Extensive experiments are conducted on the WSSS and WSOL tasks, and our group-wise learning framework performs favorably against the state-of-the-art approaches.

## References


[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[3] S. Sengupta, V. Jayaram, B. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Background matting: The world is your green screen," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2291–2300.

[4] Y. Aksoy, T.-H. Oh, S. Paris, M. Pollefeys, and W. Matusik, "Semantic soft segmentation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–13, Aug. 2018.

[5] T. Zhou, J. Li, S. Wang, R. Tao, and J. Shen, "MATNet: Motion-attentive transition network for zero-shot video object segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 8326–8338, 2020.

[6] W. Wang, T. Zhou, F. Porikli, D. Crandall, and L. Van Gool, "A survey on deep learning technique for video segmentation," 2021, *arXiv:2107.01153*.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[8] T. Zhou, L. Li, G. Bredell, J. Li, and E. Konukoglu, "Quality-aware memory network for interactive volumetric image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 560–570.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[11] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. Van Gool, "Exploring cross-image pixel contrast for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 7303–7313.

[12] 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.

[13] 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.

[14] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 695–711.

[15] 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.

[16] 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.

[17] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, and M.-H. Yang, "Weakly-supervised semantic segmentation via sub-category exploration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8991–9000.

[18] J. Fan, Z. Zhang, C. Song, and T. Tan, "Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4283–4292.

[19] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12275–12284.

[20] J. Fan, Z. Zhang, T. Tan, C. Song, and J. Xiao, "CIAN: Cross-image affinity net for weakly supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10762–10769.

[21] G. Sun, W. Wang, J. Dai, and L. Van Gool, "Mining cross-image semantics for weakly supervised semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 347–365.

[22] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[23] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[24] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.

[25] X. Li, T. Zhou, J. Li, Y. Zhou, and Z. Zhang, "Group-wise semantic mining for weakly supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1984–1992.

[26] D. Pathak, P. Krahenbuhl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1796–1804.

[27] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia, "Augmented feedback in semantic segmentation under image level supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 90–105.

[28] Y. Wei *et al.*, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, Nov. 2017.

[29] A. Chaudhry, P. K. Dokania, and P. H. S. Torr, "Discovering class-specific pixels for weakly-supervised semantic segmentation," 2017, *arXiv:1707.05821*.

[30] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4981–4990.

[31] R. Fan, Q. Hou, M.-M. Cheng, G. Yu, R. R. Martin, and S.-M. Hu, "Associating inter-image salient instances for weakly supervised semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 367–383.

[32] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3159–3167.

[33] P. Vernaza and M. Chandraker, "Learning random-walk label propagation for weakly-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7158–7166.

[34] J. Dai, K. He, and J. Sun, "BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1635–1643.

[35] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 876–885.

[36] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3136–3145.

[37] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 549–565.

[38] 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.

[39] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7268–7277.

[40] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7014–7023.

[41] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised scale equivariant network for weakly supervised semantic segmentation," 2019, *arXiv:1909.03714*.

[42] J. Lee, E. Kim, S. Lee, J. Lee, and S. Yoon, "FickleNet: Weakly and semi-supervised semantic image segmentation using stochastic inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5267–5276.

[43] Z. Yu, Y. Zhuge, H. Lu, and L. Zhang, "Joint learning of saliency detection and weakly supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7223–7233.

[44] D. Zhang, H. Zhang, J. Tang, X. Hua, and Q. Sun, "Causal intervention for weakly-supervised semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020.

[45] R. Quan, Y. Wu, X. Yu, and Y. Yang, "Progressive transfer learning for face anti-spoofing," *IEEE Trans. Image Process.*, vol. 30, pp. 3946–3955, 2021.

[46] F. Wan, C. Liu, W. Ke, X. Ji, J. Jiao, and Q. Ye, "C-MIL: Continuation multiple instance learning for weakly supervised object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2199–2208.

[47] 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.

[48] A. Rahimi, A. Shaban, T. Ajanthan, R. Hartley, and B. Boots, "Pairwise similarity knowledge transfer for weakly supervised object localization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 395–412.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13001–13008.

[55] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "DropBlock: A regularization method for convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10727–10737.

[56] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[57] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[58] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[59] J. Zhang, X. Shi, J. Xie, H. Ma, I. King, and D.-Y. Yeung, "GaAN: Gated attention networks for learning on large and spatiotemporal graphs," 2018, *arXiv:1803.07294*.

[60] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1263–1272.

[61] T. Zhou, S. Qi, W. Wang, J. Shen, and S.-C. Zhu, "Cascaded parsing of human-object interaction recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 5, 2021, doi: 10.1109/TPAMI.2021.3049156.

[62] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 20–28.

[63] W. Wang, T. Zhou, S. Qi, J. Shen, and S.-C. Zhu, "Hierarchical human semantic parsing with comprehensive part-relation modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 29, 2021, doi: 10.1109/TPAMI.2021.3055780.

[64] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 401–417.

[65] W. Wang, X. Lu, J. Shen, D. Crandall, and L. Shao, "Zero-shot video object segmentation via attentive graph neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9236–9245.

[66] Y. Zhou, T. Zhou, T. Zhou, H. Fu, J. Liu, and L. Shao, "Contrast-attentive thoracic disease recognition with dual-weighting graph reasoning," *IEEE Trans. Med. Imag.*, vol. 40, no. 4, pp. 1196–1206, Apr. 2021.

[67] M. Qu, Y. Bengio, and J. Tang, "GMNN: Graph Markov neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 5241–5250.

[68] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9226–9235.

[69] B. Li, Z. Sun, Q. Li, Y. Wu, and H. Anqi, "Group-wise deep object co-segmentation with co-attention recurrent neural network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8519–8528.

[70] F. Meng, H. Li, Q. Wu, B. Luo, and K. N. Ngan, "Weakly supervised part proposal segmentation from multiple images," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 4019–4031, Aug. 2017.

[71] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3632–3647, Oct. 2021.

[72] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[73] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[74] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," in *Proc. Int. Conf. Learn. Represent.*, 2016.

[75] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[76] 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.

[77] P.-T. Jiang, Q. Hou, Y. Cao, M.-M. Cheng, Y. Wei, and H. Xiong, "Integral object mining via online attention accumulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2070–2079.

[78] L. Chen, W. Wu, C. Fu, X. Han, and Y. Zhang, "Weakly supervised semantic segmentation with boundary exploration," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 347–362.

[79] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3203–3212.
[80] W. Ge, S. Yang, and Y. Yu, "Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1277–1286.
[81] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9215–9223.
[82] B. Zhang, J. Xiao, Y. Wei, M. Sun, and K. Huang, "Reliability does matter: An end-to-end weakly supervised semantic segmentation approach," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12765–12772.
[83] X. Wang, S. You, X. Li, and H. Ma, "Weakly-supervised semantic segmentation by iteratively mining common object features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1354–1362.
[84] N. Araslanov and S. Roth, "Single-stage semantic segmentation from image labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4253–4262.
[85] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2209–2218.
[86] X. Wang, S. Liu, H. Ma, and M.-H. Yang, "Weakly-supervised semantic segmentation by iterative affinity learning," *Int. J. Comput. Vis.*, vol. 128, pp. 1736–1749, Jan. 2020.
[87] W. Shimoda and K. Yanai, "Self-supervised difference detection for weakly-supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5208–5217.
[88] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 991–998.
[89] F. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez, "Built-in foreground/background prior for weakly-supervised semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 413–432.
[90] X. Zhang, Y. Wei, and Y. Yang, "Inter-image communication for weakly supervised localization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 271–287.
[91] 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.
[92] 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.*, 2020, pp. 481–496.
[93] 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.

**Tianfei Zhou** received the Ph.D. degree from the Beijing Institute of Technology in 2017. From 2019 to 2020, he was a Researcher at the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. He is currently a Postdoctoral Researcher at ETH Zürich, Switzerland. His current research interests include computer vision, medical image analysis, and deep learning.

**Liulei Li** is currently pursuing the master's degree with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. His research interest is image and video segmentation.

**Xueyi Li** is currently pursuing the master's degree with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. Her research interests include weakly supervised learning and image segmentation.

**Chun-Mei Feng** is currently pursuing the Ph.D. degree with the School of Information Science and Technology, Harbin Institute of Technology, Shenzhen, China. Her research interests include medical imaging analysis, computer vision, and bioinformatics.

**Jianwu Li** (Member, IEEE) received the Ph.D. degree from Tianjin University, Tianjin, China, in 2003. He is currently an Associate Professor with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. His research interests cover computer vision, image processing, and machine learning.

**Ling Shao** (Fellow, IEEE) is currently the CEO and a Chief Scientist of the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. His research interests include computer vision, machine learning, and medical imaging. He is a fellow of IAPR, IET, and BCS. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and several other journals.