# Balanced and Hierarchical Relation Learning for One-shot Object Detection

Hanqing Yang[1,2] Sijia Cai[2] Hualian Sheng[2,3] Bing Deng[2]
Jianqiang Huang[2] Xian-Sheng Hua[2] Yong Tang[4] Yu Zhang[1*]
[1] State Key Laboratory of Industrial Control Technology,
College of Control Science and Engineering, Zhejiang University
[2]DAMO Academy, Alibaba Group
[3] College of Information Science and Electronic Engineering, Zhejiang University
[4]Shudao Investment Group Co., Ltd
{hanqing.yang, hlsheng, zhangyu80}@zju.edu.cn, jianqiang.jqh@gmail.com, ty640106@163.com
{stephen.csj, dengbing.db, xiansheng.hxs}@alibaba-inc.com

## Abstract

*Instance-level feature matching is significantly important to the success of modern one-shot object detectors. Recently, the methods based on the metric-learning paradigm have achieved an impressive process. Most of these works only measure the relations between query and target objects on a single level, resulting in suboptimal performance overall. In this paper, we introduce the balanced and hierarchical learning for our detector. The contributions are two-fold: firstly, a novel Instance-level Hierarchical Relation (IHR) module is proposed to encode the contrastive-level, salient-level, and attention-level relations simultaneously to enhance the query-relevant similarity representation. Secondly, we notice that the batch training of the IHR module is substantially hindered by the positive-negative sample imbalance in the one-shot scenario. We then introduce a simple but effective Ratio-Preserving Loss (RPL) to protect the learning of rare positive samples and suppress the effects of negative samples. Our loss can adjust the weight for each sample adaptively, ensuring the desired positive-negative ratio consistency and boosting query-related IHR learning. Extensive experiments show that our method outperforms the state-of-the-art method by 1.6% and 1.3% on PASCAL VOC and MS COCO datasets for unseen classes, respectively. The code will be available at https://github.com/hero-y/BHRL.*

## 1. Introduction

Development of modern deep learning architectures gives rise to great advances in general object detection [10, 19, 24, 33]. However, deep detectors necessitate massive, high-quality yet expensive annotated data to reach performance saturation, which is extremely difficult for practical applications. Inspired by the human ability of learning new concepts with very little supervision, some recent works [23, 25, 29, 30, 34] attempt to apply few-shot learning techniques to detect novel-class objects from extremely few novel-class data. Yet, the majority focus on finetune-based strategies by adapting pre-trained detectors to limited unseen-class samples. Such deep detectors usually suffer from generalization limitations, such as the catastrophic forgetting of base classes, the underutilization of novel data, and the severe shift in data distribution.

The One-Shot Object Detection (OSOD) [2,12,22] without finetuning aims to detect all interesting objects in the target image with the same novel class of single query image patch. A noteworthy effort has been made to exploit the instance matching techniques and build semantic relations for the query-target region proposals. Some early developments [12, 22] integrate effective metric-learning solutions with the general object detectors (e.g., Faster R-CNN [24]) to learn a similarity metric. For example, SiamMask [22] uses the matching module to correlate the query image and target image. CoAE [12] proposes to apply non-local scheme [28] for exploring the relation feature. Although such methods are capable of achieving fast and effective adaption to novel classes through a single query image patch, they lack accurate modeling of multi-level semantic relations for the query-target pairs. Intuitively, the single relation measure may lead to a certain similarity deviation, thus reducing the generalization ability of the learned detector. Some modern approaches [2, 8, 30] propose to capture seen-unseen semantic structure via more fine-grained relation learning. For example, [8] introduces three relation

[*] Corresponding author.

heads to exploit different matching relationships. However, it ignores the importance of the fusion of different relation features and adopts the global relation features that leads to spatial information loss of relation features. AIT [2] deploys the transformer architecture to explore the visual characteristics in each proposal-query pair, but it does not explicitly tie visual semantics for query and target representations in a compositional way.

To this end, we first introduce a novel Instance-level Hierarchical Relation (IHR) module that can infer multi-level semantic relations for generating query-target similarity features. Specifically, we initially use region proposal network to extract instance-level feature maps. Then, the IHR module decomposes query-target feature matching into three hierarchical semantic levels, which are responsible to capture the global difference, local salient region, and local discriminative part, respectively. The global difference reveals that the target object should be described by using its contrastive characteristics when being compared with the query object. The local salient region is extracted by depthwise convolution architecture to better contain different activation patterns and infer instance-level saliency semantics. The local discriminative part is learned by the attention mechanism to capture the distinct features that affect the matching. Different levels of semantic coverage can guide the learning of diverse and hierarchical features for query-target matching to aggregate both global and local details. For each relation branch, we maintain the resolution consistency of the output relation feature and the input region feature to avoid the loss of context information. Subsequently, these three kinds of relation features are integrated to promote the discriminative and localization ability.

Additionally, we claim that the prevailing wisdom such as random sampling scheme [24] in general two-stage detectors for dealing with sample imbalance is not efficient due to a small positive-negative ratio in the OSOD task. This leads to the unbalanced training and suboptimal performance of the above IHR module. Accordingly, we propose a simple but effective Ratio-Preserving Loss (RPL) to reweight the samples in the training process for achieving balanced IHR learning. We adaptively adjust the sample weights to maintain a suitable and stable positive-negative sampling ratio. With such a one-shot sample reweighting scheme, the rare positive-pair relations for seen classes would be identified and contribute more to the final discriminative detection. Thus, our learned detector offers great potential to detect novel classes with complex semantic similarities and differences.

Our key contributions are summarized as follows:

- We design a powerful multi-level relation module named IHR for the OSOD task. It exploits the semantic similarities on the contrastive-level, salient-level, and attention-level simultaneously, aiming to find more comprehensive relations between query image patch and target image.
- We propose a simple but effective RPL to solve the imbalance issue of positive-negative samples for achieving balanced and effective learning of the IHR module.
- Extensive experiments show that our detector outperforms the state-of-the-art method by 1.6% and 1.3% on PASCAL VOC and MS COCO datasets, respectively, which validates its effectiveness.

## 2. Related Work

**Few-shot Object Detection.** In few-shot object detection, most approaches first train the model with abundant base-class data and then fine-tune the model on both novel-class and base-class data. They treat this task as a multi-classification and localization problem, which aims to achieve incremental detection through a small amount of data. There are two main streams: meta-learning-based methods [13, 30, 31] and transfer-learning-based methods [25, 29]. Meta-learning-based methods extract meta-level knowledge to help the model adapt to novel categories. Meta R-CNN [31] predicts per-class channel-wise attention vector to reweight the corresponding feature map. FS-DetView [30] proposes a joint-feature embedding module. FSOD [8] introduces attention-RPN and multi-relation detector to detect novel classes. For the transfer-learning-based methods, TFA [29] proposes only fine-tuning the last layers of the detector and freezing the other parameters. FSCE [25] introduces a contrastive proposal encoding loss to facilitate the classification of detected objects. In the case of extremely few novel-class data, this fine-tuning method can easily overfit novel classes and degrade the performance of base classes. In addition, the reliance on the fine-tuning process limits the practicality to a certain extent.

**One-shot Object Detection.** One-shot object detection is a particular case of few-shot object detection. Unlike most few-shot object detectors, one-shot object detectors only use the base-class data for training and directly detect novel-class objects without fine-tuning. They treat the OSOD task as a two-classification and localization problem, which aims to directly detect novel-class objects under a template matching scheme. SiamMask [22] inserts a matching module into Mask R-CNN [10] to generate the similarity feature map between the query image patch and the entire target image. CoAE [12] uses the non-local scheme [28] and squeeze-excitation scheme [14] to correlate the query image patch and target image. FOC OSOD [32] introduces the classification feature deformation-and-attention module and split iterative head to improve the classification power. AIT [2] develops an attention-based encoder-decoder architecture with transformer [27] to evaluate the relation among

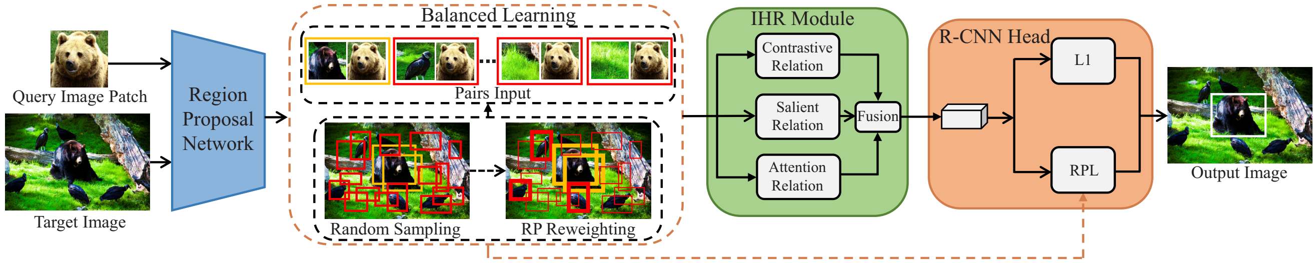

Figure 1. The overall architecture of the proposed BHRL for one-shot object detection.

query-target pairs. By contrast, our method comprehensively extracts the instance-level relation feature in a hierarchical structure. Moreover, we propose an effective ratio-preserving loss induced by soft sampling to ensure the balanced and effective learning of our relation module.

## 3. Our Method

### 3.1. Problem Definition

In OSOD task, the object classes are split into seen classes $S$ and unseen classes $U$, where $S \cap U = \varnothing$. Given an arbitrary query image patch $q$, the one-shot object detector aims to detect all instances in the target image $I$ that are consistent with this query patch category. The one-shot object detector is trained with the data of seen classes $S$. Once the detector is trained, it can be generalized to directly detect unseen classes $U$ with only one query patch.

### 3.2. Overall Architecture

Figure 1 shows the overall architecture of our BHRL (**B**alanced and **H**ierarchical **R**elation **L**earning). It mainly consists of three parts: the process of generating region proposals, the IHR module for multi-level relation modeling, and the balanced design of detection head that strengthens the IHR learning. (1) In the proposal generation, the shared-weight siamese ResNet-50 [11] with feature pyramid network (FPN) [18] is adopted to extract the visual features of the query image patch and target image. We then follow SiamMask [22] to use the matching module to compute the similarity feature between the query vector and each position of the whole target feature. The similarity feature enables the standard RPN [24] to generate a set of potential region proposals more relevant to the query patch. Based on these region proposals, we retrieve the proposal features in the whole target feature using the RoI pooling operator. The query feature is pooled to the same size as the target proposal features. (2) The proposed IHR module then learns relation representation to highlight the complex interdependencies for query and target pairs in a hierarchical manner. (3) At last, the R-CNN head [24] is used to detect the query-related instances, and the proposed RPL is introduced to rebalance the involved samples, thus achieving a more balanced training of the IHR module with accurate discriminative and localization ability.

### 3.3. Instance-level Hierarchical Relation Module

Most of the existing studies [2, 12, 22] measure the semantic relation between the query image patch and target image utilizing a single relation module (e.g, relation network [26]). The relation network [26] extracts the similarity feature between the instance-level query feature $\mathbf{F}_q \in \mathbb{R}^{C\times K\times K}$ and instance-level target feature $\mathbf{F}_t \in \mathbb{R}^{C\times K\times K}$ by concatenation operation. Here, $C$ is the number of channels and $K$ is the height or width of the feature map. The output relation feature is as follows:

$$\mathbf{R}_r = \text{Conv}_{\frac{3C}{2}}([\mathbf{F}_q, \mathbf{F}_t]), \tag{1}$$

where $[\cdot,\cdot]$ denotes concatenation and $\text{Conv}_{\frac{3C}{2}}(\cdot)$ is a $1\times1$ convolution layer with an output channel of $\frac{3C}{2}$.

However, it may be ineffective when faced with indistinguishable distractors. [8] proposes a multi-relation detector to model different relations. However, this multi-relation detector has two drawbacks. 1) For each relation head, it obtains the global representation of the relation feature, which results in spatial information loss of the relation feature. 2) It simply adds classification scores generated by different relation heads instead of fusing relation features. Therefore, this method is useful for classification task but not suitable for localization task. Unlike it, the proposed IHR module eliminates the above shortcomings and adopts a hierarchical manner to comprehensively describe the semantic relations. Figure 2 sketches the architecture of the IHR module. It encodes the contrastive-level, salient-level, and attention-level relations simultaneously. For each relation branch, we generate the relation feature with the same size as the input feature to preserve the global contextual coherence and spatial consistency relationship. These semantic relation clues are then integrated to enhance the query-relevant similarity representation, thus facilitating the subsequent classification and localization tasks.

**Contrastive-level Relation.** We introduce a contrastive-level relation branch to compute the relation between the

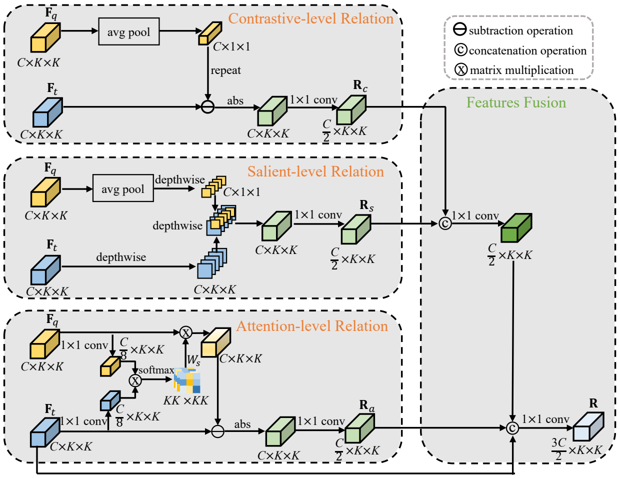

Figure 2. Illustration of the proposed instance-level hierarchical relation (IHR) module.

global query feature and the local target feature. We compare the query vector to each location of the target feature using a subtraction operation. Unlike a previous study [22] that calculates the relation between the query patch and the whole target image, we describe the relation between the instance-level objects in a more straightforward manner. The output relation feature is given by the following:

$$\mathbf{R}_c = \mathrm{Conv}_{\frac{C}{2}}(|\mathcal{R}(\mathcal{P}(\mathbf{F}_q)) - \mathbf{F}_t|), \tag{2}$$

where $|\cdot|$ denotes the absolute value operator. $\mathcal{R}(\cdot)$ is a repeat operator that makes $\mathbb{R}^{C\times 1\times 1} \rightarrow \mathbb{R}^{C\times K\times K}$. $\mathcal{P}(\cdot)$ denotes the average pooling operation.

**Salient-level Relation.** We build a salient-level relation branch to learn the instance-level saliency relation [21]. The query vector is treated as the convolution kernel to extract the relation feature with the local target feature in a depthwise manner [16]. In contrast to the multi-relation detector [8], we use the global query feature instead of the local query feature to ensure that the generated relation feature has the same resolution as the input feature. This way can preserve rich semantic information, thus improving the ability of modeling relation. The output relation feature can be expressed as follows:

$$\mathbf{R}_s = \mathrm{Conv}_{\frac{C}{2}}(\varphi(\mathcal{P}(\mathbf{F}_q), \mathbf{F}_t)), \tag{3}$$

where $\varphi(\cdot,\cdot)$ denotes the depthwise convolution operation.

**Attention-level Relation.** To learn the more detailed local relation, we apply an attention-level relation branch. The local comparison between the query feature and target feature may meet the spatial misalignment issue [3]. To alleviate this issue, we adopt cross attention to generate a spatial-aware query feature. First, two embedding features for the query and target features are generated by the convolution layer $\mathrm{Conv}_{\frac{C}{8}}$ [13]. The query-target similarity at the spatial level can be computed based on the matrix multiplication between the two embeddings. The spatial attention matrix $W_s$ is obtained by applying softmax to the query-target similarity. The above process can be summarized as:

$$W_s = \mathrm{softmax}((\mathrm{Conv}_{\frac{C}{8}}(\mathbf{F}_t))^T \mathrm{Conv}_{\frac{C}{8}}(\mathbf{F}_q)). \tag{4}$$

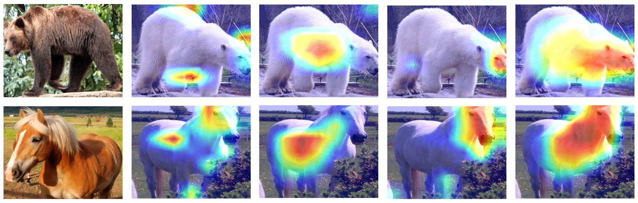

Figure 3. Visualization of the proposal relation feature for each level in the IHR module for unseen classes.

Next, the attention matrix $W_s$ is treated as a soft weight to generate the spatial-aware query feature. Different from [3, 13] that concatenate the weighted query feature and target feature as output, we extract the local semantic relation between the spatial-aware query feature and target feature through a subtraction operation, which is more effective. The output relation feature is as follows:

$$\mathbf{R}_a = \mathrm{Conv}_{\frac{C}{2}}(|W_s\mathbf{F}_q - \mathbf{F}_t|). \tag{5}$$

**Fusion-level Relation.** After obtaining these three relation features, we first integrate $\mathbf{R}_c$ and $\mathbf{R}_s$ produced by the global query feature (using average pooling operator), then concatenate $\mathbf{R}_a$ produced by the local query feature to get relation feature with dimension $C$. Finally, it is fused with the target feature $\mathbf{F}_t$ to obtain the final relation feature.

$$\mathbf{R} = \mathrm{Conv}_{\frac{3C}{2}}([\mathrm{Conv}_{\frac{C}{2}}([\mathbf{R}_s, \mathbf{R}_c]), \mathbf{R}_a, \mathbf{F}_t]), \tag{6}$$

where we follow [22] to use $\frac{3C}{2}$ dimension for the final relation feature, and use $\frac{C}{2}$ dimension for $\mathbf{R}_c$, $\mathbf{R}_s$ and $\mathbf{R}_a$ for reducing computational cost.

**Discussions.** In Figure 3, we visualize the heatmaps of contrastive-level relation feature, salient-level relation feature, attention-level relation feature, and fusion-level relation feature in the IHR module, respectively. It can be seen that contrastive-level relation module concentrates more on global characteristics of objects, such as the object contour boundary information. Salient-level relation module places a high priority on the middle salient region of the object that contrastive-level relation module misses. Attention-level relation module can capture rich and subtle regions such as nose, mouth, and eyes due to the spatial attention mechanism. Fusion relation feature can take full advantage of the complementary benefits thus is able to provide discriminative semantic cues generated by three different kinds of relation modules comprehensively.

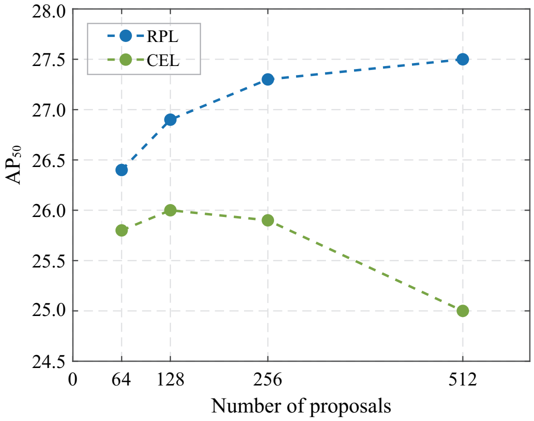

Figure 4. Performance for CEL and the proposed RPL under the different number of proposals.

### 3.4. Balanced Learning of the IHR Module

We observe that the effectiveness of the IHR module will be overwhelmed by a large number of negative query-target pairs under the common sampling scheme. Our goal is to reduce the level of imbalance that exists in the training process using only lightweight loss design, thus exploiting the potential of the IHR-driven detector as much as possible.

**Imbalance Issue in the OSOD Task.** Existing OSOD methods [2, 12, 22] follow the popular two-stage detector Faster R-CNN [24] to apply a random sampling scheme, which aims to keep a reasonable sample ratio for positive and negative samples during the second stage. The cross-entropy loss (CEL) is suitable for this setup. Its formulation for all $N$ proposals can be formulated as:

$$L_{CE} = \frac{1}{N}(\sum_{i\in R_p} L^i_{SCE} + \sum_{i\in R_n} L^i_{SCE}), \tag{7}$$

where $R_p$ and $R_n$ denotes the set of positive samples and negative samples, respectively. $L^i_{SCE}$ denotes the softmax cross-entropy loss value for $i$-th proposal.

However, in the OSOD task, this sampling scheme is unable to achieve the desired effect under the default number of proposals. This reason is that the OSOD models only sample the rare proposals with the same class as the query patch as positive samples. A large number of proposals that are inconsistent with the query category are treated as negative samples. The too small positive-negative ratio makes the model difficult to learn from the positive samples. Alternatively, we can sample a small number of proposals to alleviate the positive-negative imbalance. As shown in Figure 4 (green polyline), the performance with a small number of proposals is superior to the performance with the default number of proposals (i.e., 512). This is due to the fact that the desired positive-negative ratio is more likely to be maintained when the number of proposals is small. This shows the importance of ensuring the desired positive-negative ratio consistency. However, the significant sample reduction will also sacrifice the crucial proposals that are beneficial for learning semantic relations.

**Ratio-Preserving Loss.** A common method for addressing positive-negative imbalance is to introduce a weighting factor $\alpha \in$ [0, 1] for positive samples and 1-$\alpha$ for negative samples. The balanced CEL can be formulated as:

$$L_{BCE} = \frac{1}{N}(\alpha\sum_{i\in R_p} L^i_{SCE} + (1-\alpha)\sum_{i\in R_n} L^i_{SCE}), \tag{8}$$

The above method is unsuitable for direct use in the OSOD task. The reason is that it adopts the static balancing parameter $\alpha$ for all images. However, this is suboptimal for the challenging OSOD task.

To address the above issues, we propose an effective ratio-preserving (RP) reweighting strategy and its induced RPL to ensure a reasonable and stable positive-negative ratio without filtering important proposals. That is, we dynamically increase the weight of positive samples and decrease the weight of negative samples to retain a suitable and specific number-weighted ratio (neither too big nor too small). The too-small ratio makes the model difficult to learn from positive samples, while the too-large ratio causes overfitting. Moreover, to enhance the learning of false positives, we separate false positives from negative samples. Then, we take false positives and positive samples as a whole to increase their weights and decrease the weights of true negatives, which can be seen as a special hard negative mining strategy for the OSOD task. The above process can be summarized as:

$$\begin{gathered} L_{RP} = \frac{1}{N}(u\sum_{i\in R_p\cup R_{fp}} L^i_{SCE} + v\sum_{i\in R_{tn}} L^i_{SCE}), \\ u = \frac{N\cdot\alpha}{N_p+N_{fp}}, v = \frac{N\cdot(1-\alpha)}{N_{tn}}, \end{gathered} \tag{9}$$

where $R_{fp}$ denotes the set of false positives (decided by comparing prediction and ground truth) and $R_{tn}$ denotes the set of true negatives. $N_p$ denotes the number of positive samples, $N_{fp}$ denotes the number of false positives, and $N_{tn}$ denotes the number of true negatives.

As shown in Figure 4, our RPL is consistently better than CEL under the different number of sampled proposals, which verifies the effectiveness of ratio-preserving mechanism in boosting the relation learning.

## 4. Experiments

### 4.1. Datasets and Settings

**Datasets and Evaluation Metrics.** For a fair comparison, we follow the same OSOD settings as the previous works [2, 12, 22]. For the PASCAL VOC dataset [7], we divide 20

| Methods | Seen class | | | | | | | | | | | | | | | | Unseen class | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | plant | sofa | tv | car | bottle | boat | chair | person | bus | train | horse | bike | dog | bird | mbike | table | Average | cow | sheep | cat | aero | Average |
| SiamFC (ECCV2016) [1] | 3.2 | 22.8 | 5.0 | 16.7 | 0.5 | 8.1 | 1.2 | 4.2 | 22.2 | 22.6 | 35.4 | 14.2 | 25.8 | 11.7 | 19.7 | 27.8 | 15.1 | 6.8 | 2.28 | 31.6 | 12.4 | 13.3 |
| SiamRPN (CVPR2018) [15] | 1.9 | 15.7 | 4.5 | 12.8 | 1.0 | 1.1 | 6.1 | 8.7 | 7.9 | 6.9 | 17.4 | 17.8 | 20.5 | 7.2 | 18.5 | 5.1 | 9.6 | 15.9 | 15.7 | 21.7 | 3.5 | 14.2 |
| OSCD (Neurocomputing2020) [9] | 28.4 | 41.5 | 65.0 | 66.4 | 37.1 | 49.8 | 16.2 | 31.7 | 69.7 | 73.1 | 75.6 | 71.6 | 61.4 | 52.3 | 63.4 | 39.8 | 52.7 | 75.3 | 60.0 | 47.9 | 25.3 | 52.1 |
| CoAE (NIPS2019) [12] | 24.9 | 50.1 | 58.8 | 64.3 | 32.9 | 48.9 | 14.2 | 53.2 | 71.5 | 74.7 | 74.0 | 66.3 | 75.7 | 61.5 | 68.5 | 42.7 | 55.1 | 78.0 | 61.9 | 72.0 | 43.5 | 63.8 |
| AIT (CVPR2021) [2] | 46.4 | **60.5** | 68.0 | 73.6 | 49.0 | **65.1** | 26.6 | 68.2 | **82.6** | 85.4 | **82.9** | 77.1 | 82.7 | 71.8 | 75.1 | **60.0** | 67.2 | **85.5** | **72.8** | 80.4 | 50.2 | 72.2 |
| BHRL (Ours) | **57.5** | 49.4 | **76.8** | **80.4** | **61.2** | 58.4 | **48.1** | **83.3** | 74.3 | **87.3** | 80.1 | **81.0** | **87.2** | **73.0** | **78.8** | 38.8 | **69.7 (+2.5)** | 81.0 | 67.9 | **86.9** | **59.3** | **73.8 (+1.6)** |

Table 1. Performance comparisons with state-of-the-art methods on the PASCAL VOC dataset in terms of $\text{AP}_{50}$.

| Methods | Seen | | | | | Unseen | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | split-1 | split-2 | split-3 | split-4 | Average | split-1 | split-2 | split-3 | split-4 | Average |
| SiamMask (Arxiv2018) [22] | 38.9 | 37.1 | 37.8 | 36.6 | 37.6 | 15.3 | 17.6 | 17.4 | 17.0 | 16.8 |
| CoAE (NIPS2019) [12] | 42.2 | 40.2 | 39.9 | 41.3 | 40.9 | 23.4 | 23.6 | 20.5 | 20.4 | 22.0 |
| AIT (CVPR2021) [2] | 50.1 | 47.2 | 45.8 | 46.9 | 47.5 | 26.0 | 26.4 | 22.3 | 22.6 | 24.3 |
| BHRL (Ours) | **56.0** | **52.1** | **52.6** | **53.4** | **53.5 (+6.0)** | **26.1** | **29.0** | **22.7** | **24.5** | **25.6 (+1.3)** |

Table 2. Performance comparisons with state-of-the-art methods on the COCO dataset in terms of $\text{AP}_{50}$.

classes into 16 seen classes and 4 unseen classes. For the COCO dataset [20], 80 classes are equally split into 4 parts ($P_1$, $P_2$, $P_3$, $P_4$), alternately taking 3 parts (60 classes) as seen classes and 1 part (20 classes) as unseen classes. For evaluation metrics, we follow [2, 12] to report $\text{AP}_{50}$ for both the PASCAL VOC dataset and COCO dataset.

**Implementation Details.** We train our model with a batch size of 16 on 8 GPUs using the SGD optimizer for 9 epochs. The learning rate starts at 0.02 and decays by a factor of 10 at the 7-th epoch. We use ResNet-50 as our backbone network, which is is pre-trained on the reduced ImageNet [6] where the COCO-related ImageNet classes are removed [12]. This ensures that the model does not foresee the unseen-class objects. We use the deformable RoI pooling [5] to generate the target proposals on the PASCAL VOC dataset, and use RoIAlign [10] to generate the target proposals on the COCO dataset.

**Target-query Pairs.** We follow [2, 12] to generate the target-query image pairs. In training stage, for a given target image containing the seen-class object, we randomly select a query patch with the same seen class. In the test stage, for each class in the target image, the query patches of the same class are shuffled with a random seed of the target image ID, then the first five query patches are chosen to test five times and average the metrics scores as the reported results.

### 4.2. Comparison with State-of-the-art Methods

**Evaluation on the PASCAL VOC Dataset.** In Table 1, we compare our BHRL with state-of-the-art methods on the PASCAL VOC dataset for both seen classes and unseen classes. It can be seen that our BHRL achieves the best performance in most cases for both seen classes and unseen classes. For seen classes, BHRL outperforms the state-of-the-art AIT [2] by 2.5% $\text{AP}_{50}$. For unseen classes, our method outperforms the previous popular CoAE [12] significantly with a 10.0 % $\text{AP}_{50}$ gain and outperforms previous SOTA AIT with a 1.6 % $\text{AP}_{50}$ gain. This significant improvement mainly comes from the fact that BHRL can comprehensively explore the relations between query patch and target image and achieve balanced relation learning.

**Evaluation on the COCO Dataset.** To further validate the effectiveness of our proposed BHRL, we evaluate the performance of BHRL on the challenging COCO dataset for all four splits. Table 2 shows the results. It can be seen that, although COCO is much more challenging than PASCAL VOC with higher complexity like occlusions and more classes, BHRL still achieves superior performance compared with other methods in all splits. As shown in the "Average" columns of Table 2, BHRL achieves 53.5% $\text{AP}_{50}$ on seen classes and 25.6% $\text{AP}_{50}$ on unseen classes. It surpasses the second-best AIT [2] by 6.0% $\text{AP}_{50}$ and 1.3% $\text{AP}_{50}$ on seen classes and unseen classes, respectively.

### 4.3. Ablation Studies

In this section, we conduct extensive ablation experiments to analyze each component of our proposed BHRL. Following the previous works [2, 12], we use $\text{AP}_{50}$ as the main performance indicator.

**Component-wise Analysis.** We conduct the experiments to verify the effectiveness of the proposed IHR module and RPL, and summarize the results for unseen classes on the COCO split-2 dataset and PASCAL VOC dataset in Table 3. The method in the $1^{st}$ row adopts the widely-used relation network [26] to extract the relation feature, and softmax CEL to supervise the classification. As shown in the $1^{st}$ and $2^{nd}$ rows, the IHR module contributes to a 2.9% $\text{AP}_{50}$ improvement and 4.2% $\text{AP}_{50}$ improvement on the COCO dataset and PASCAL VOC dataset, respectively. This benefits from the fact that the IHR module can generate a comprehensive and discriminating relation feature. As shown in the $2^{nd}$ and $4^{th}$ rows, the RPL improves the IHR module by 1.1% $\text{AP}_{50}$ and 2.1% $\text{AP}_{50}$ on the COCO dataset and PASCAL VOC dataset, respectively. This demonstrates that the RPL can boost the effective learning of the IHR module by solving the positive-negative imbalance.

**Impact of Relation Levels in the IHR Module.** In Table

4, we investigate the importance of each relation level in the IHR module. For a fair comparison, we fuse $\mathbf{F}_t$ and use the same channel dimension. The first four rows show that the performance of using each relation module alone is effective but limited since the single relation module may lead to certain similarity deviation. As shown in $5^{th}$ to $7^{th}$ rows, fusing any two relation features on different levels can bring performance improvement. And the best performance is achieved by fusing all three relation features, as shown in the last row. This indicates that measuring the relation feature comprehensively is beneficial.

| IHR | RPL | coco | | voc |
|---|---|---|---|---|
| | | AP | $AP_{50}$ | $AP_{50}$ |
| | | 15.2 | 25.0 | 67.5 |
| ✓ | | 16.9 | 27.9 | 71.7 |
| | ✓ | 16.5 | 27.5 | 70.9 |
| ✓ | ✓ | **17.4** | **29.0** | **73.8** |

Table 3. Effects of each component in our design on the COCO split-2 dataset and PASCAL VOC dataset for unseen classes.

| C.R. | S.R. | A.R. | $F_t$ | Level Fusion | $AP_{50}$ |
|---|---|---|---|---|---|
| | | | | - | 25.0 |
| ✓ | | | ✓ | - | 26.4 |
| | ✓ | | ✓ | - | 26.6 |
| | | ✓ | ✓ | - | 26.7 |
| ✓ | ✓ | | ✓ | features fusion | 27.1 |
| | ✓ | ✓ | ✓ | features fusion | 27.0 |
| ✓ | | ✓ | ✓ | features fusion | 27.0 |
| ✓ | ✓ | ✓ | | scores addition | 27.2 |
| ✓ | ✓ | ✓ | | features fusion | 27.6 |
| ✓ | ✓ | ✓ | ✓ | features fusion | **27.9** |

Table 4. Ablation studies for the IHR module on the COCO split-2 dataset for unseen classes. "C.R.", "S.R.", and "A.R." represent the contrastive-level relation, salient-level relation, and attention-level relation, respectively.

| Relation Extraction Methods | $AP_{50}$ |
|---|---|
| multi-relation detector [8] | 20.4 |
| feature aggregation module [30] | 21.0 |
| non-local attention [28] | 23.7 |
| dense relation distillation [13] | 25.4 |
| IHR (Ours) | **27.9** |

Table 5. Performance comparison with other popular relation extraction methods on the COCO split-2 dataset for unseen classes.

**Performance Comparison with Other Relation Extraction Methods.** We further validate the effectiveness of the IHR module by comparing it with other popular relation extraction methods in Table 5. We re-implement relation modules in [8, 13, 28, 30] into our network for fair comparisons.

- multi-relation detector [8]: It contains three relation heads to produce three classification scores, which are added as the output score.
- feature aggregation module [30]: It uses channel-wise multiplication and subtraction operation to process the query vector and target vector.
- non-local attention [28]: It utilizes attention mechanism to generate attention-wise query feature, which is integrated with target feature in summation.
- dense relation distillation [13]: It extracts pixel-wise similarity in a non-local manner, which is concatenated with the target value map as relation feature.

*Discussions.* The multi-relation detector [8] and feature aggregation module [30] encode the relation in vector format, which results in damaging spatial information of relation feature. Moreover, the relations extracted by these two methods cannot be presented in an explicit manner. The non-local attention [28] and dense relation distillation [13] adopt similar way to generate the weighted query feature, followed by integrating with target feature using summation or concatenation. But this way cannot directly and comprehensively conduct the differentiated information. Unlike these methods, our IHR module adopts a hierarchical and explicit manner to comprehensively describe the semantic relations. As shown in Table 5, the IHR module outperforms other methods by a significant margin, which demonstrates the effectiveness of the IHR module.

**Impact of Hyperparameter $\alpha$ in the RPL.** As shown in Table 6, the RPL with a reasonable hyperparameter $\alpha$ can outperform CEL significantly, which indicates a suitable and stable positive-negative ratio can bring better performance. We choose $\alpha$ to be equal to 1/4 in all experiments.

**Performance Comparison with Other Balanced Losses.** In Table 7, we compare our RPL with some popular balanced losses. We focus on discussing Focal Loss [19]. When the weight of the classification task is the default value 1, Focal Loss reduces the model's performance (27.9% $AP_{50}$ $vs.$ 25.8% $AP_{50}$). This reason is that it leads to an imbalance between the classification task and localization task. This is detrimental to the OSOD task that is more challenging for the classification task. With an increase in the weight of classification task, its performance grows steadily and reaches saturation at last. GHM Loss [17] has similar issues. For the proposed RPL, the weight of the classification task does not need to be changed because it does not destroy the balance between the classification task and localization task. Table 7 shows that our RPL in the default weight value can obtain better performance than the careful weight-adjusted Focal Loss. This benefits from the fact that our RPL can adjust positive-negative weights adaptively, ensuring a suitable and stable positive-negative ratio.

**Performance Comparison for Seen Classes.** We take Faster R-CNN as baseline to verify the effectiveness of our model for seen classes. We train Faster R-CNN with the same setting as ours (e.g., training epochs and backbone's weights). As shown in Table 8, our BHRL can achieve com-

| Methods | CEL | $\alpha$ in RPL | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 2/3 | 1/2 | 1/3 | 1/4 | 1/5 | 1/6 | 1/7 |
| $\text{AP}_{50}$ | 27.9 | 28.3 | 28.7 | 28.9 | **29.0** | 28.9 | 28.4 | 28.0 |

Table 6. Experimental results for different $\alpha$ in the RPL on the COCO split-2 dataset for unseen classes.

| Methods | W.C.T | $\text{AP}_{50}$ |
|---|---|---|
| CEL | 1 | 27.9 |
| Class-Balanced Loss [4] | 1 | 22.3 |
| GHM Loss [17] | 1 | 26.5 |
| Focal Loss [19] | 1 | 25.8 |
| | 2 | 27.0 |
| | 5 | 28.2 |
| | 10 | 27.8 |
| RPL (ours) | 1 | **29.0** |

Table 7. Performance comparison with other popular balanced losses on the COCO split-2 dataset for unseen classes. "W.C.T" indicates the weight of the classification task.

| Methods | $\text{AP}_{50}$ |
|---|---|
| Faster R-CNN | 56.6 |
| AIT [2] (one-shot) | 50.1 |
| BHRL (one-shot) | 56.0 |
| BHRL (two-shot) | 57.8 |

Table 8. Comparison for seen classes on COCO split-1 dataset.

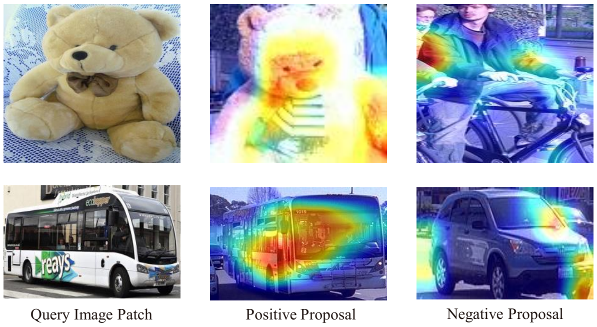

Figure 5. Visualization heatmaps of the positive proposal relation feature and negative proposal relation feature.

petitive performance with Faster R-CNN and outperform AIT on COCO split-1 dataset. Moreover, when using more query patches (e.g., 2) to aggregate their features after neck during inference, it will bring further improvement.

### 4.4. Qualitative Results

Figure 5 visualizes the heatmaps of the positive proposal relation feature and negative proposal relation feature extracted by the IHR module. The positive proposal has more obvious activation regions compared with the negative proposal, which demonstrates that the IHR module can construct different intensity similarity features for positive sample and negative sample. In Figure 6, we visualize the detection results of CoAE (re-implemented by us) and our BHRL for unseen classes. White boxes indicate correct detections, and red boxes indicate false detections. It can be observed that our BHRL can effectively detect unseen-class objects. Compared to CoAE, our proposed BHRL generates fewer false detections.

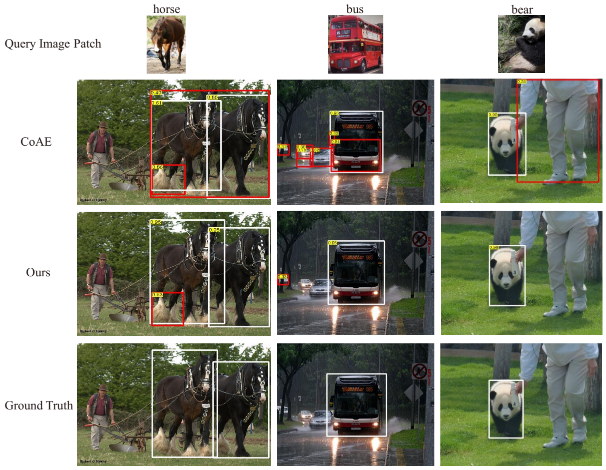

Figure 6. Visualization comparison between CoAE and our BHRL for unseen classes. White boxes indicate correct detections, and red boxes indicate false detections.

## 5. Conclusion

In this paper, we present a novel BHRL to tackle the OSOD task by improving instance-level semantic relation learning. Firstly, we propose the IHR module to comprehensively explore the semantic relation between instance-level target-query pairs in a hierarchical manner. Secondly, we propose the RPL to effectively solve the positive-negative imbalance, thus boosting the IHR learning process. Our BHRL achieves new state-of-the-art on two benchmark datasets. We hope that our work can offer good insights and inspire more research regarding the OSOD task.

**Limitations.** Although the proposed model can achieve superior performance compared with previous models, it still generates some false detections in complex scenes.

## Acknowledgement

This work was supported by the National Key Research and Development Program of China under Grant 2021ZD0201400, the National Natural Science Foundation of China (Grant No.62088101, 61673341), the Project of State Key Laboratory of Industrial Control Technology, Zhejiang University, China (No.ICT2021A10), the Fundamental Research Funds for the Central Universities (No.2021FZZX003-01-06), Double First Class University Plan (CN).

## References

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. *ECCV*, 2016. 6

[2] Ding-Jie Chen, He-Yen Hsieh, and Tyng-Luh Liu. Adaptive image transformer for one-shot object detection. *CVPR*, 2021. 1, 2, 3, 5, 6, 8

[3] Tung-I Chen, Yueh-Cheng Liu, Hung-Ting Su, Yu-Cheng Chang, Yu-Hsiang Lin, Jia-Fong Yeh, and Winston H Hsu. Should i look at the head or the tail? dual-awareness attention for few-shot object detection. *ArXiv*, 2021. 4

[4] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. *CVPR*, 2019. 8

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. *ICCV*, 2017. 6

[6] Jia Deng, W. Dong, R. Socher, L. Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009. 6

[7] M. Everingham, L. Gool, C. K. Williams, J. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2009. 5

[8] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with attention-rpn and multi-relation detector. *CVPR*, 2020. 1, 2, 3, 4, 7

[9] Kun Fu, T. Zhang, Yue Zhang, and Xian Sun. Oscd: A one-shot conditional object detection framework. *Neurocomputing*, 2020. 6

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *ICCV*, 2017. 1, 2, 6

[11] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016. 3

[12] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. *NeurIPS*, 2019. 1, 2, 3, 5, 6

[13] Hanzhe Hu, Shuai Bai, Aoxue Li, Jinshi Cui, and Liwei Wang. Dense relation distillation with context-aware aggregation for few-shot object detection. *CVPR*, 2021. 2, 4, 7

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *CVPR*, 2018. 2

[15] B. Li, J. Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. *CVPR*, 2018. 6

[16] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. *CVPR*, 2019. 4

[17] Buyu Li, Yu Liu, and Xiaogang Wang. Gradient harmonized single-stage detector. *AAAI*, 2019. 7, 8

[18] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *CVPR*, 2017. 3

[19] Tsung-Yi Lin, Priyal Goyal, Ross B. Girshick, Kaiming He, and P. Dollár. Focal loss for dense object detection. *TPAMI*, 2020. 1, 7, 8

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *ECCV*, 2014. 6

[21] Yi Liu, Qiang Zhang, Dingwen Zhang, and Jungong Han. Employing deep part-object relationships for salient object detection. *ICCV*, 2019. 4

[22] Claudio Michaelis, Ivan Ustyuzhaninov, Matthias Bethge, and Alexander S. Ecker. One-shot instance segmentation. *ArXiv*, 2018. 1, 2, 3, 4, 5, 6

[23] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. *ICCV*, 2021. 1

[24] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 2016. 1, 2, 3, 5

[25] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. *CVPR*, 2021. 1, 2

[26] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. *CVPR*, 2018. 3, 6

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2

[28] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CVPR*, 2018. 1, 2, 7

[29] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *ICML*, 2020. 1, 2

[30] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. *ECCV*, 2020. 1, 2, 7

[31] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. *ICCV*, 2019. 2

[32] Hanqing Yang, Yongliang Lin, Hong Zhang, Yu Zhang, and Bin Xu. Towards improving classification power for one-shot object detection. *Neurocomputing*, 2021. 2

[33] Hanqing Yang, Liyang Zheng, Saba Ghorbani Barzegar, Yu Zhang, and Bin Xu. Borderpointsmask: One-stage instance segmentation with boundary points representation. *Neurocomputing*, 2022. 1

[34] Ze Yang, Yali Wang, Xianyu Chen, Jianzhuang Liu, and Yu Qiao. Context-transformer: tackling object confusion for few-shot detection. *AAAI*, 2020. 1