Fusing Adaptive Meta Feature Weighting for Few-shot Object Detection

Peng Zhou* 2021107254@ZUT.EDU.CN
Zhongyuan University of Technology, Zhengzhou, 450007, China

Anzhan Liu
Zhongyuan University of Technology, Zhengzhou, 450007, China

Editors: Nianyin Zeng and Ram Bilas Pachori

Abstract

Few-shot object detection models often lack the perceptual ability to detect the target objects and fine-tuning the model on base class images to quickly adapt to new tasks can lead to feature shift issues. We propose an Adaptive Meta-Feature Weighting (AMFW-YOLO) object detection model for solving these problems. This model introduces an attention mechanism based on spatial and channel-wise squeeze-and-excitation (scSE) blocks, which helps the model focus on the regions of interest in the target samples and suppresses interference from background regions. Additionally, to compensate for the feature shift caused during the fine-tuning stage, we design an Adaptive Meta-Feature Weighting module (AMFW), This module embeds positional information into spatial features, captures long-range dependencies along two directions, and adaptively compensates for the weights of deep global features, effectively improving the accuracy of the model.

Keywords: Few-shot Learning, Few-shot Object Detection, Feature Reweighting, Adaptive Meta-feature weighting

1. Introduction

With the leap in the development of deep learning, significant breakthroughs have been achieved in object detection technology. Many new object detection models have emerged, but these models rely heavily on a large number of sample images and require extensive annotated data. In certain special application scenarios, such as endangered animals, medical image diagnosis, industrial quality inspection, and environmental monitoring, there is a severe shortage of available data. Therefore, how to perform object detection with a small amount of annotated data has become a research hotspot in recent years (Liu, 2023).

Few-shot learning aims to detect new classes by using a small amount of annotated data, effectively reducing sample dependency and improving model generalization. Currently, few-shot learning has wide applications in computer vision fields such as image classification, image segmentation, and object detection. Few-shot object detection combines few-shot learning with object detection algorithms to achieve accurate recognition and localization of objects with a small number of annotated samples. Existing few-shot object detection methods mainly include data augmentation-based, metric learning-based, fine-tuning-based, transfer learning-based, and meta-learning-based approaches (Antonelli et al., 2022).

Data Augmentation-based Methods: Data augmentation-based methods aim to enhance the model's generalization to new classes by augmenting the base-class data. For example, Zhou et al. (2023) proposed a multi-scale positive sample optimization method that adjusts the scale of positive samples using a target feature pyramid and refines feature maps to increase sample diversity. Kim et al. (2022) increase the number of interested area samples by adjusting the scale of target regions multiple times using spatial features. Xu et al. (2021) introduced a Positive Sample Augmentation (PSA) module to balance the scale distribution of positive samples, suppress the negative sample ratio, and improve detection accuracy.

Metric Learning-based Methods: Metric learning, also known as similarity learning, is a method that first induces class vectors by summarizing support images of the same category. Based on the principle that closer distances represent higher similarity, it calculates the distance between the test sample and each category of support images using distance functions such as Euclidean distance and Minkowski distance. This way, it determines the category of the test sample. Karlinsky et al. (2018) introduced a novel distance metric learning approach that can simultaneously learn the parameters of the backbone network, the embedding space, and the category representation vectors. This method can also learn a joint representation space, effectively capturing the underlying relationships between data from different modalities, and using this space for metric learning tasks. Zhang et al. (2020) proposed a contrastive network object detection framework. This framework uses siamese networks to extract features from query images and target images, serving as the edge probabilities in the feature space for metric learning. By comparing the similarity of these features in the feature space using a learnable metric, it accomplishes object detection. Li et al. (2019) introduced a deep nearest neighbour neural network. This network uses local descriptors to replace image-level features.

Fine-tuning-based Methods: Fine-tuning-based methods adopt transfer learning principles, transferring features learned on a large-scale dataset to a few-shot dataset. Models pretrained on a large dataset can extract more robust feature representations. Sun et al. (2021) introduced a contrastive proposal encoding method for few-shot object detection, reducing variance in the embedding of proposals of the same class. Wei et al. (2019) proposed an adaptive adversarial sample generator to enhance detection performance in few-shot object detection.

Transfer Learning-based Methods: Transfer learning methods transfer prior knowledge learned from a source domain to a target domain, analogous to how humans apply previous experiences to tackle new problems. Cao et al. (2022) divided the fine-tuning phase into two stages, correlation, and discrimination, to ensure both intraclass feature space coherence and inter-class separability. Guirguis et al. (2022) used replay techniques in continual learning to transfer knowledge from base-class samples to new-class samples.

Meta-Learning-based Methods: Meta-learning methods train models on a per-task basis, learning commonalities among different tasks, optimizing model parameters quickly to improve generalization. Santoro et al. (2016) introduced a memory-augmented network inspired by neural Turing machines. Kang et al. (2019) proposed a meta-feature reweighting model for few-shot object detection, although it suffers from feature shift in the fine-tuning stage. Runchao LIN (2022) designed a meta-feature secondary reweighting module to address feature shift issues, but further improvement is needed.

This paper makes three main contributions: To mitigate the interference caused by negative samples in feature extraction networks, the paper introduces a Spatial and Channel Parallel Squeeze and Excitation (scSE) block-based attention mechanism, reducing the negative impact on model recognition accuracy. Adaptive Meta-Feature Weighting (AMFW) Module: this module incorporates position and spatial information, adaptively adjusts meta-feature weights, and compensates for offset feature coefficients, further improving model recognition accuracy. The AMFW-YOLO model achieves significant improvements in average accuracy for different shot tasks on the PASCAL VOC dataset, with the highest average accuracy improvement of nine point seven percent. This approach effectively addresses feature shift issues.

2. Adaptive Meta-Feature Weighting Model

2.1. Adaptive Meta-Feature Weighting Network Structure

This paper addresses the issues of the Re-weighting Model's low attention to important information, lack of perception, and feature shift, proposing the Adaptive Meta-Feature Weighting (AMFW-YOLO) network model. The primary network architecture adopted by the Adaptive Meta-Feature Weighting Model is YOLOv2. Firstly, the model introduces an attention mechanism based on Spatial and Channel Squeeze & Excitation blocks, which effectively filters out negative sample information, thereby reducing the interference of negative samples in feature extraction and generating fine-grained masks. Secondly, an Adaptive Meta-Feature Weighting (AMFW) module is designed, which embeds positional information into spatial information, adaptively adjusts the weights of meta-features, compensates for feature shift caused by fine-tuning, and further enhances the model's detection capabilities. The overall network structure of AMFW-YOLO is depicted in Figure 1.

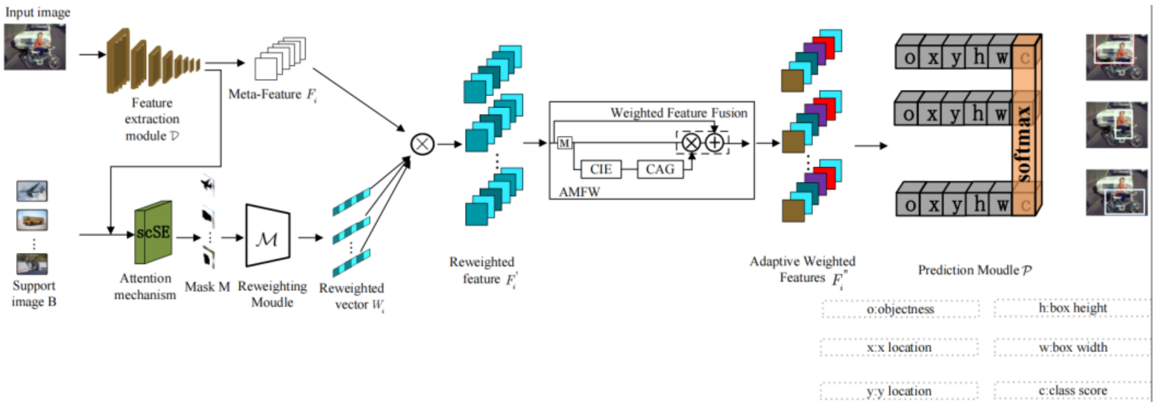

Figure 1: Overall network structure of AMFW-YOLO.

In this paper, the defined support image is denoted as B. Both the support image and query image undergo feature extraction. The support image yields meta-features with m channels, represented as $F \in \mathbb{R}^{w\times h\times m}$, and denoted as $F = D(I)$.Through the attention mechanism module based on scSE, fine-grained segmentation masks M are generated. Subsequently, the support image is concatenated with the mask, denoted as (B_i, M_i), which serves as the input to the re-weighting module, resulting in the re-weighting vector W_i, where $W_i \in \mathbb{R}^m$ represents a specific operation. Query the image to obtain the corresponding elemental feature F_i, and obtain the reweighted elemental feature n through Equation 1, completing the transfer of prior knowledge from base classes to new classes.

$$F_i^{'} = F_i \otimes W_i,i\text{= } 1,\dots,\text{N} \tag{1}$$

Here, $\otimes$ denotes channel convolution, implemented using a 1×1 depth-wise convolution. $F_i^{'}$ serves as the input to the AMFW module, and through the embedding of spatial position information

(CIE) and the generation of spatial position attention (CAG), weighting coefficients are adjusted to obtain adaptive weighted features $F_i^{''}$.

These features are then input into the detection module P, yielding confidence scores for target categories o_i, predicted bounding box positions and sizes (x_i, y_i, h_i, w_i), and category classification scores c_i. Based on the output information from the detection module P, target predictions are made.

2.2. Adaptive Meta-Feature Weighting Module

During the process of encoding information between channels, the importance of spatial information is often overlooked, making it difficult to capture deep global features. This paper introduces an Adaptive Meta-Feature Weighting (AMFW) module, which embeds positional information into channel relationships to explore deep feature information and adaptively compensate for its weight coefficients. The specific structure of the Adaptive Meta-Feature Weighting module is illustrated in Figure 2. The workflow of this module primarily consists of three main parts:

Spatial Position Information Embedding (CIE): Capturing global spatial information is typically done using global average pooling, but it can be challenging to retain important spatial details. Therefore, global average pooling is split into two 1-dimensional vector encoding operations. Given input X, encoding operations are performed using pooling kernels $(H, 1)$ and $(1, W)$ to capture horizontal and vertical features, respectively. The output for the c-th channel with a height of h can be represented as given in Equation 2.

$$z_c^h(h) = \frac{1}{W} \sum_{0 \leq i < W} x_c(h, i) \tag{2}$$

Similarly, the output for the c-th channel with a width of w can be represented as given in Equation 3.

$$z_c^w(w) = \frac{1}{H} \sum_{0 \leq i < H} x_c(j, w) \tag{3}$$

The advantage of aggregating features from both directions is that while capturing spatial relationships along one spatial axis, it preserves spatial information along the other axis.

Spatial Position Attention Generation (CAG): Concatenate $z_c^h(h)$ and $z_c^w(w)$, then perform feature transformation through a 1x1 convolution, BatchNorm, and non-linear activation to obtain intermediate feature f using Equation 4.

$$f = \delta(F_1([z^h, z^w])) \tag{4}$$

where $[z^h, z^w]$ represents the concatenation of vertical and horizontal features, $F_1(\cdot)$ denotes a 1x1 convolution, $\delta(\cdot)$ is a non-linear activation function, resulting in intermediate feature $f \in \mathbb{R}^{C/r \times W}$. r represents a scaling parameter used to reduce the number of channels in the intermediate feature, thereby reducing model complexity. Subsequently, f is split into two independently distributed features, $f^h \in \mathbb{R}^{C/r \times H}$ and $f^w \in \mathbb{R}^{C/r \times W}$, based on the spatial dimension. These features are then separately input into 1x1 convolutions and sigmoid activation functions for feature transformation and adjustment of their dimensions to match the input feature. This is done using Equations 5 and 6 to obtain g^h and g^w.

$$g^h = \sigma(F_h(f^h)) \tag{5}$$

$$g^w = \sigma(F_w(f^w)) \tag{6}$$

Weighted Feature Fusion: The combination of g^h and g^w is merged into a weight matrix, and the final calculation formula for the output $y_c(i,j)$ is represented in Equation 7.

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{7}$$

$y_c(i,j)$ is weighted and fused with the input features, completing the process of Adaptive Meta-Feature Reweighting. The Adaptive Meta-Feature Weighting module approaches feature importance from a global perspective, combining positional and spatial information to analyze the significance of a feature. It adaptively compensates for weight coefficients that offset feature shifts.

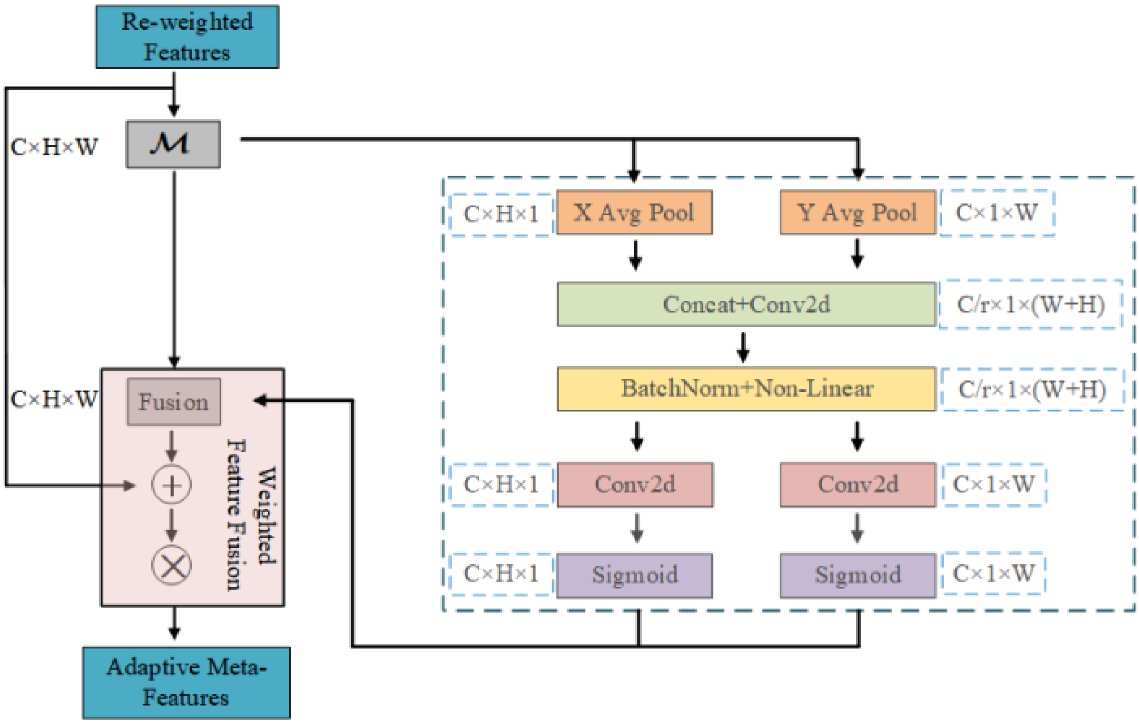

Figure 2: scSE Attention Mechanism Module.

3. Experimental Results and Analysis

3.1. Comparative Experiment

This article compares the performance of the model after integrating the AMFW module with other object detection models, including FSRW, SE-SMFR, FRCN+ft-full (Zhang et al., 2021), MetaDet (Huang et al., 2020), LSTD (Runchao LIN, 2022), YOLOv2-fit (Hu et al., 2018), and FRCN+ft (Zhang et al., 2021).

Table 1 provides a comparison of the average precision results between this paper's method and the comparison models in three classification groups on the VOC dataset. Table 2 shows the average accuracy of this paper's method and other methods in PASCAL VOC dataset for new and base class categories.

Table 1: Comparison of average accuracy between this method and other methods under different groups of PASCAL VOC.

Algorithm	*Novel Set 1*					*Novel Set 2*					*Novel Set 3*				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
YOLOv2-ft	6.6	10.7	12.5	24.8	38.6	12.5	4.5	11.6	16.1	33.9	13.0	15.9	15.0	32.2	38.4
LSTD	8.2	11.0	12.4	29.1	38.5	11.4	3.8	5.0	15.7	31.0	12.6	8.5	15.0	27.3	36.3
FRCN+ft-full	13.8	19.6	32.8	41.5	45.6	7.9	15.3	26.2	31.6	39.1	9.8	11.3	19.1	35.0	45.1
MetaDet	**17.1**	19.1	28.9	35.0	48.8	**18.2**	20.6	25.9	30.6	41.5	20.1	22.3	27.9	41.9	42.9
FRCN+ft	11.9	16.4	29.0	36.9	36.9	5.9	8.5	23.4	29.1	28.8	5.0	9.6	18.1	30.8	43.4
FSRW	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	**21.3**	25.6	28.4	42.8	45.9
SE-SMFR	15.5	21.2	28.1	36.2	40.3	13.9	**22.6**	28.8	35.6	41.9	15.1	27.8	29.6	38.4	41.7
AMFW	16.4	**22.3**	30.7	**43.6**	46.5	14.8	**23.7**	**29.4**	**38.6**	**45.2**	17.9	**28.4**	**30.1**	39.8	44.6

From the results in Table 1, it can be seen that compared to the seven other object detection models, t-he model proposed in this paper demonstrates good performance in all three grouping scenarios. The AMFW module's design effectively mitigates feature shift issues caused by the two-stage fine-tuning.

Compared to the FSRW model, the mAP is improved by up to 9.7%. In Novel Set 1, the overall aver-age precision for different shot tasks improved by 4.28%, which is a 3.64% improvement compared to the improved FSRW model, SE-SMFR. The average precision for different shot tasks in the three subset groups increased by 3.25%. AMFW-YOLO achieved an average improvement of 6% for 2-shot, 4.13% for 3-shot, and 5.06% for 5-shot compared to FSRW. In comparison to SE-SMFR, it achieved the highest mAP improvement of 7.4% and an overall performance improvement of 2.36%. The average accuracy for different shot tasks in Novel Set 1, Novel Set 2, and Novel Set 3 improved by 3.64%, 1.78%, and 1.64% respectively. AMFW-YOLO improved by 1.53%, 0.94%, 1.23%, 3.93%, and 4.13% for 1-shot, 2-shot, 3-shot, 4-shot, and 5-shot tasks. The experimental data shows that the improvement in AMFW-YOLO's performance becomes more pronounced with an increase in labeled data.

While the model in this paper significantly improved average accuracy in the three different shot tasks, it didn't achieve ideal results in the 1-shot task. The main reason for this is that in the 1-shot task, there is insufficient transferable information from a single instance image, resulting in subpar recognition performance. In comparison to MetaDet, which achieved the best performance in 1-shot recognition, it introduced a weight prediction meta-model that can predict specific class parameters from a small number of samples, allowing it to maintain high accuracy even with low samples. In the future, the approach in this paper can consider incorporating ideas from the MetaDet model to improve AMFW-YOLO's accuracy in low-sample scenarios.

3.2. Ablation Experiment

In order to validate the impact of the scSE and AMFW modules on the performance of small-sample object detection, this paper conducted an ablation experiment on the 10-shot task in Novel Set 1. The specific experimental data is presented in Table 2.

The data from the first row and the second row of the experimental results indicate that with the introduction of the scSE attention mechanism, the model's mAP for new class images improved by 0.78%, and the mAP for base class images increased by 1.31%. New class images inherently

Table 2: The results of ablating scSE and AMFW.

Algorithm	*scSE*	*AMFW*	*New class*	*Base class*
FSRW	×	×	47.42	63.59
AMFW-YOLO	√	×	48.20	64.90
AMFW-YOLO	×	√	49.72	65.48
AMFW-YOLO	√	√	50.88	67.04

lack sufficient feature information. By introducing the attention mechanism and combining it with the weighting module to transfer base class image meta-feature information to new class images, the model's recognition accuracy for new class images is effectively enhanced. Additionally, scSE filters out negative sample information, increasing the weight of features in the target regions of objects under detection, thereby enhancing the model's focus on important information.Comparing the data from the first row with the third row, it is evident that the designed AMFW module increased the mAP for new class images by 2.3 percentage points and improved the accuracy for base class images by 1.89 percentage points. This validates the good performance of AMFW in compensating for feature shift issues.

4. Conclusion

To address the issue of feature shift leading to a decrease in accuracy, this paper introduces the AMFW-YOLO model. This model incorporates the scSE attention mechanism to generate fine-grained masks for base class images, reducing the impact of negative samples on detection accuracy. By designing the Adaptive Meta-Feature Weighting (AMFW) module to compensate for feature bias in transferring features from new classes to base classes, AMFW embeds positional information in spatial information in parallel, capturing deep global features and adaptively compensating for their weight coefficients, alleviating the impact of shifted features on recognition accuracy. To demonstrate the improvement in the model's detection capabilities, validation was performed on the PASCAL VOC dataset. The model in this paper showed improvements in average accuracy for various detection tasks, with the highest improvement being 9.7% compared to the FSRM model. For 2-shot, 3-shot, and 5-shot tasks, there were average improvements of 6%, 4.13%, and 5.06%, respectively. Experimental data suggests that the proposed method in this paper exhibits strong performance, greatly improving the issue of feature shift and further enhancing the accuracy of model detection.

References

Simone Antonelli, Danilo Avola, Luigi Cinque, Donato Crisostomi, Gian Luca Foresti, Fabio Galasso, Marco Raoul Marini, Alessio Mecca, and Daniele Pannone. Few-shot object detection: A survey. *ACM Comput. Surv.*, 54(11s), sep 2022. ISSN 0360-0300. doi: 10.1145/3519022.

Yuhang Cao, Jiaqi Wang, Ying Jin, Tong Wu, Kai Chen, Ziwei Liu, and Dahua Lin. Few-shot object detection via association and discrimination, 2022.

Karim Guirguis, Ahmed Hendawy, George Eskandar, Mohamed Abdelsamad, Matthias Kayser, and Juergen Beyerer. Cfa: Constraint-based finetuning approach for generalized few-shot object detection, 2022.

Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection, 2018.

Zilong Huang, Xinggang Wang, Yunchao Wei, Lichao Huang, Humphrey Shi, Wenyu Liu, and Thomas S. Huang. Ccnet: Criss-cross attention for semantic segmentation, 2020.

Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8419–8428, 2019. doi: 10.1109/ICCV.2019.00851.

Leonid Karlinsky, Joseph Shtok, Sivan Harary, Eli Schwartz, Amit Aides, Rogerio Feris, Raja Giryes, and Alex M. Bronstein. Repmet: Representative-based metric learning for classification and one-shot object detection, 2018.

Geonuk Kim, Hong-Gyu Jung, and Seong-Whan Lee. Spatial reasoning for few-shot object detection, 2022.

Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7253–7260, 2019. doi: 10.1109/CVPR.2019.00743.

Chen T. Wang C. Jiang S. Chen D Liu, C. Review of small sample object detection research. *Journal of Frontiers o-f Computer Science Technology*, 17(1):53, 2023. doi: 10.2197/ipsjjip.24.49.

Aihua DONG Runchao LIN, Rong HUANG. Few-shot object detection based on attention mechanism and secondary reweighting of meta-features. 42(10):3025, 2022. doi: 10.11772/j.issn.1001-9081.2021091571.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1842–1850. JMLR.org, 2016.

Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding, 2021.

Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection, 2019.

Honghui Xu, Xinqing Wang, Faming Shao, Baoguo Duan, and Peng Zhang. Few-shot object detection via sample processing. *IEEE Access*, 9:29207–29221, 2021. doi: 10.1109/ACCESS.2021.3059446.

Hu Zhang, Keke Zu, Jian Lu, Yuru Zou, and Deyu Meng. Epsanet: An efficient pyramid squeeze attention block on convolutional neural network, 2021.

Tengfei Zhang, Yue Zhang, Xian Sun, Hao Sun, Menglong Yan, Xue Yang, and Kun Fu. Comparison network for one-shot conditional object detection, 2020.

Xiao Zhou, Yujie Zhong, Zhen Cheng, Fan Liang, and Lin Ma. Adaptive sparse pairwise loss for object re-identification, 2023.