CRPN: DISTINGUISH NOVEL CATEGORIES VIA CLASS-RELEVANT REGION PROPOSAL NETWORK FOR FEW-SHOT OBJECT DETECTION

Han Wang, Yali Li, Shengjin Wang

Beijing National Research Center for Information Science and Technology (BNRist),
Department of Electronic Engineering, Tsinghua University

ABSTRACT

Few-shot object detection (FSOD) has attracted more attention in computer vision, where only very few training examples are presented during model learning process. A commonly-overlooked issue in FSOD is that novel classes are usually classified as background clutters in the pre-training process. Another difficulty of FSOD is that the detection performance degrades especially under higher IoU thresholds since previous deep metric learning (DML) requires frozen region proposals without class-relevant box regression. In this work, we propose a Class-relevant Region Proposal Network (CRPN). The CRPN can derive network parameters for novel classes from pre-trained convolution kernels according to their feature similarity, which is used to eliminate the above mentioned adverse effects and improve the performance of few-shot object detection. The proposed CPRN is able to kill two birds with one stone and has two main contributions: (1) transfer a region proposal network pre-trained on base classes to novel classes; (2) perform class-dependent bounding-box regression which previous DML classifier lacks. For experimental testing, we achieve 12.7% AP75 in MS COCO dataset and 28.6% AP75 in ImageNet2015 dataset under the few-shot setting introduced by previous works, which exceeds the state-of-the-art by a certain margin.

Index Terms— Object Detection, Novel Class, Few-Shot Learning, Deep Metric Learning, RPN

1. INTRODUCTION

Deep convolutional neural networks (CNNs) can suffer from severe over-fitting and fail to generalize when lack of labeled training data. The ability of learning from few-shot examples has motivated the research on few-shot learning (FSL). While significant progress has been made, most of these methods focus on image classification [1, 2, 3, 4, 5] and only a few generalize [6, 7, 8, 9, 10] to object detection.

Distinguishing novel class objects from cluttered background is the major difficulty in few-shot object detection task. Potential region proposals can be easily missed or false alarms can be produced in the background. We believe this is caused by a long-existing issue that novel classes are classified as background in base class pre-training (Fig. 1) and a simple fine-tuning in [6] is insufficient to overcome the prior bias. Since region proposal is class-relevant when transferring from base classes to novel classes, [10] suggests an Attention Region Proposal Network (Attention RPN), introducing a cross-correlation between support and query sample feature. This Attention RPN may alleviate the issue if support feature and query feature are highly correlated, which does not always hold due to the variance of object appearance.

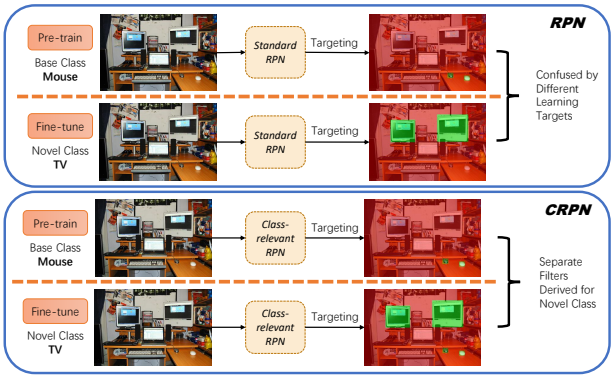

Fig. 1. **Top**: A standard RPN classifies anchors corresponding to `TV` as background in pre-train but as foreground in fine-tune, and vice versa for `mouse`. The foreground and background filters in RPN will be confused. **Bottom**: CRPN eliminates such confusion by deriving new filter parameters for novel classes based on feature similarity. CRPN will recognize both `mouse` and `TV` as foreground and perform class-dependent bounding box regression simultaneously.

In this paper, we propose two main contributions: First, we propose a Class-relevant Region Proposal Network (CRPN) to condition objectiveness filters and anchor delta filters on object class based on a simplified assumption that a linear transform maps class feature to filter weight corresponding to a category and anchor type. Second, we adopt prototypical

This work was supported by the State Key Development Program in 14th Five-Year under Grant No. 2021YFF0602103, 2021YFF0602102, 2021QY1702. We also thank for the research fund under Grant No. 2019GQG0001 from the Institute for Guo Qiang, Tsinghua University.

metric learning to replace the fully-connected classification head. We construct several prototype vectors for each class and compare the distance of an embedded object feature to prototypes to determine the posterior. Our proposed few-shot object detection approach achieved state-of-the-art performance in MS COCO [11] and ImageNet2015 [12] few-shot detection experiments comparing to [6, 9, 8, 7, 10].

2. RELATED WORK

Few-Shot Learning. Few-shot learning (FSL) means learning a model from just a few training samples per category. Previous works are mainly focused on image classification scenario. Existing FSL approaches are categorized into two types: generative models and discriminative models. Most of early FSL methods before the emergence of deep learning are generative models. With the great success of CNNs, many FSL researchers shift their work to discriminative models. Their approaches can be summarized into augmentation [13, 14], metric learning [15, 1, 16] and meta learning [17, 2].

Few-Shot Object Detection. Few-shot object detection is a new subject in FSL. [18] exploits unlabeled data and multiple models are alternatively trained and tested on images without annotations, which may be caught into a dilemma if an unlabeled object is incorrectly detected. LSTD [6] proposes a transfer loss between source domain and target domain and its performance is highly dependent on their similarity. RepMet [7] constructs representatives for each class with proposals generated by Selective Search[19] and fails to work with online RPN. Recently, some other works [9, 10] have been proposed to exploit the correlation between support and query samples. [20] addresses the same issue as we raised in our paper that the RPN may ignore or regard novel samples as background.

Our work is motivated by the research line leading by Prototypical Networks [1] and we generalize it into object detection with CRPN. Filter weights for novel classes in CRPN is derived based on their feature similarity base classes and prototype vectors are maintained for each category for distance calculation.

3. METHODOLOGY

3.1. Class-Relevant Region Proposal

Convolution feature of a backbone network is forwarded through a 3×3 convolution layer in CRPN and then split into two branches as shown in Figure 2, one responsible for predicting objectiveness for each anchor type (*objectiveness layer*) and the other for predicting anchor deltas (*anchor delta layer*).

The objectness layer of CRPN can be represented by $f_{obj}(\cdot; \mathbf{W}_o)$, where each row of $\mathbf{W}_o \in \mathbb{R}^{n \times p}$ represents the flattened and concatenated N_{anchor} filters $\mathbf{w}_o \in \mathbb{R}^p$ corresponding to a category. We follow the idea in [21] and assume the objectness filter of a class-specific $\mathbf{w}_o$ can be obtained by

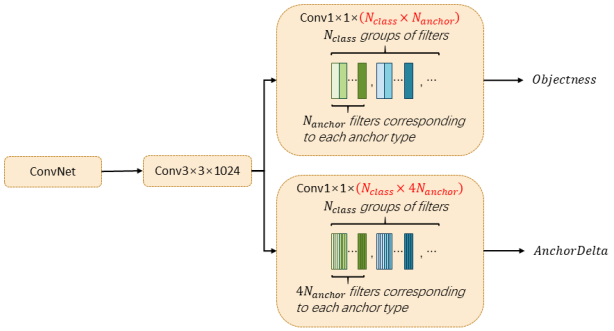

Fig. 2. Class-relevant Region Proposal Network (CRPN)

applying a linear transform matrix $T_o \in \mathbb{R}^{q \times p}$ to its hidden variable $\mathbf{v}_o \in \mathbb{R}^q$.

We sample a batch of object images cropped from whole picture for each class and resize them to a fixed size with zero padding. To measure the similarity between categories i, j, the sampled image patches X_i, X_j are forwarded through backbone feature extractor to get their mean average convolution feature vector $\overline{\mathbf{z}_i}, \overline{\mathbf{z}_j}$ and the cosine distance of A_{ij} is calculated according to:

$$A_{ij} = \frac{\overline{\mathbf{z}_i} \cdot \overline{\mathbf{z}_j}}{\|\overline{\mathbf{z}_i}\| \cdot \|\overline{\mathbf{z}_j}\|} \tag{1}$$

Base class pre-train determines base class objectness filter weight matrix $\mathbf{W}_o$. In order to find the optimal transform T_o, we need to minimize the error between $V_o T_o$ and W_o, where each row of $\mathbf{V}_o \in \mathbb{R}^{n \times q}$ is a hidden variable $\mathbf{v}_o \in \mathbb{R}^q$ attached to each category. Typically, the hidden variables $\mathbf{v}_o$ should keep the similarity defined by similarity matrix A, which gives a regularizer in Eq.2.

$$L(V_o, T_o) = \|V_o T_o - W_o\|_F^2 + \beta \|A - V_o V_o^T\|_F^2 \tag{2}$$

The optimal solution of V_o, T_o can be achieved by performing alternative gradient descent on base classes.

To construct objectness filters for a novel class, we first calculate its mean average convolution feature vector $\overline{\mathbf{z}_{new}}$ with support images and its similarity $\mathbf{a}^{new}$ to base classes. Suppose the transform T_o holds for novel classes, we can omit the first term in Eq.2 and get

$$L(\mathbf{v}_o^{new}) = \left\| \begin{bmatrix} A & \mathbf{a}^{newT} \\ \mathbf{a}^{new} & 1 \end{bmatrix} - \begin{bmatrix} V_o \\ \mathbf{v}_o^{new} \end{bmatrix} \begin{bmatrix} V_o^T & \mathbf{v}_o^{newT} \end{bmatrix} \right\|_F^2 \tag{3}$$

Removing all terms independent to $\mathbf{v}_o^{new}$ and the regularizer we have the transformed objectness filters in CRPN corresponding to this new class:

$$\begin{cases} \mathbf{v}_o^{new} & = \mathbf{a}^{new} (V_o^T)^+ \\ \mathbf{w}_o^{new} & = \mathbf{v}_o^{new} T_o \end{cases} \tag{4}$$

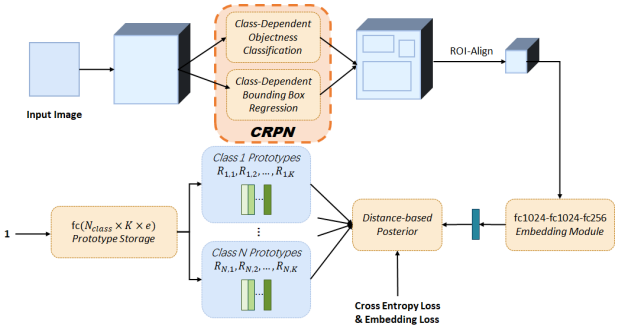

Fig. 3. Our proposed few-shot object detection framework

Fig. 4. **Example 10-shot detection results on MS COCO dataset.** Yellow frames indicate groundtruth. A threshold of 0.5 on the detection score is used throughout

where $(\cdot)^+$ is the pseudo inverse of a matrix.

The derived filter weights corresponding to a category act as a regularizer combined with sigmoid cross entropy loss in Eq. 5 during novel class fine-tuning.

$$L_{obj}(X,Y) = L_{ce}(f_{obj}(X;W_o),Y) + \lambda \sum_{i=1}^{N_{class}} \|\mathbf{w}_o^{(i)} - \mathbf{w}_{o,new}^{(i)}\|^2 \quad (5)$$

where X represents an input image and Y represents the corresponding category label.

The anchor delta regression layer $f_{reg}(\cdot;W_r)$ is computed in a similar approach except each row of $\mathbf{W}_r \in \mathbb{R}^{n\times 4p}$ represents the flattened and concatenated $4N_{anchor}$ filters $\mathbf{w}_r \in \mathbb{R}^{4p}$ corresponding to a category. The derived filter weights $\mathbf{w}_{r,new}^{(i)}$ act as a regularizer combined with smooth L1 loss in Eq 6 during novel class fine-tuning.

$$L_{reg}(X,B) = L_{SmoothL1}(f_{reg}(X;W_r),B) + \lambda \sum_{i=1}^{N_{class}} \|\mathbf{w}_r^{(i)} - \mathbf{w}_{r,new}^{(i)}\|^2 \quad (6)$$

where X represents an input image and B represents the corresponding bounding box annotation.

3.2. Deep Metric Learning Classifier

With our proposed CRPN, we can generalize DML into few-shot object detection task and generate prototypical vectors for each detection category (Figure 3). The bounding box regression module is unnecessary in our approach because its redundancy with CRPN.

Object features of the same category may vary due to visual angle, illumination, occlusion and individual differences. We keep K different prototypes R_{ij} for each class. Object features after ROI-Align[22] are forwarded through a feature embedding module and compared with prototypes of each class to compute distance-based posterior (Eq. 7).

$$P(Y=i|X) = P(Y=i|E) = \max_{j=1,\dots,K} p_{ij}(E) \quad where \quad p_{ij}(E) \propto \exp\left(-\frac{d(E,R_{ij})^2}{2\sigma^2}\right) \quad (7)$$

The posterior for background class is estimated by its lower bound:

$$P(Y=0|X) = P(Y=0|E) = 1 - \max_{i,j} p_{ij}(E) \quad (8)$$

During test phase, the prototypes of novel classes are replaced with embedded feature of support samples.

The deep metric learning module can be optimized with the sum of two losses: standard cross-entropy loss and embedding loss (Eq. 9).

$$L_{DML}(X,Y) = L_{ce}(\max_j p_{ij}(E),Y) + \left|\delta - \min_{j,i\neq Y} d(E,R_{ij}) + \min_j d(E,R_{Yj})\right|_+ \quad (9)$$

where Y is the groundtruth class and $|\cdot|_+$ is ReLU function. The embedding loss is similar to Triplet loss [23] which leads to a classification margin larger than δ.

4. EXPERIMENTS

4.1. Implementation Details

We use ResNet-50[24] pre-trained on ImageNet[25] as convolution backbone and our CRPN is attached to C4. $N_{anchor} = 15$ corresponding to anchor size $\{32, 64, 128, 256, 512\}$ and aspect ratio $\{0.5, 1.0, 2.0\}$. The feature embedding module is consist of two fully-connected layers of width 1024 with BN and ReLU, and a final layer with width $e = 256$. We set $\beta = 1.0, \lambda = 1.0, \sigma = 0.5, \delta = 0.2$. The shorter side of a query image is resized to 600px and the longer side is capped at 1000. The support sample is cropped around the bounding box with 16px padding and resized to 320px×320px with zero padding. The model is pre-trained on base classes and fine-tuned on few-shot novel classes.

4.2. Comparison with State-of-the-art Algorithms

We compare our proposed approach with recent state-of-the-art algorithms [9, 8, 10] on MS COCO [11] dataset. The 20 categories included in PASCAL VOC[26] are used as novel

Method	AP	AP_{50}	AP_{75}	AP_{80}	AP_{90}
FRCNN[27]	3.1	7.9	1.7	-	-
ReWeight[9]	5.6	12.3	4.6	-	-
MetaRCNN[8]	8.7	19.1	6.6	-	-
FSOD[10]	11.1	20.4	10.6	-	-
Ours w/ RPN	12.0	22.4	11.8	7.3	3.8
Ours w/ CRPN	12.9	22.9	12.7	8.3	5.1

Table 1. Results on MS COCO dataset.

Method	AP_{50}	AP_{75}
Faster R-CNN[27]	12.2	-
LSTD[6]	37.4	-
RepMet[7]	39.6	-
FSOD[10]	41.3	21.9
Ours	41.7	28.6

Table 2. Results on ImageNet-based 50-way detection.

classes and the rest 60 as base classes. $K = 10$ support samples are used for fine-tuning on novel classes. Table 1 shows our methods with RPN and CRPN both outperform the previous leading method FSOD[10]. Figure 4 includes examples of 10-shot detection test results.

Another comparison with [6, 7, 10] is performed on ImageNet2015 [12] dataset in a 50-way 5-shot setting (Table 2). We use MS COCO as large-scale pre-train dataset and choose 50 non-overlapping classes in ImageNet2015 as few-shot train/test dataset. The performance gain in classification accuracy is not as much as MS COCO because most objects are salient in ImageNet2015 but a significant gain in AP_{75} reflects that the class-dependent bounding box regression in CRPN leads to more accurate object localization.

4.3. Ablation Study

Backbone network. We adopt different backbone networks in the MS COCO experiment and the results are listed in Table 3. The improvement brought by deeper backbones are not significant because of the limited support samples. The grouped convolution in ResNeXt[28] achieves a larger performance gain comparing with ResNet with similar number of parameters.

Distance metrics. The squared Euclidean distance is a particular type of Bregman divergence and the cluster representative achieving minimal distance to its assigned points is the cluster mean in terms of Bregman divergence. Our experiments (Table 4) show that squared Euclidean achieves better performance over cosine distance in low-shot settings as suggested by [1]. More support shots in our method yields more representative vectors with various appearance features. Cosine distance is commonly adopted for vector comparison as it captures the orientation of vectors and not the magnitude.

Backbone	AP	AP_{50}	AP_{75}
ResNet-50	12.9	22.9	12.7
ResNet-101	13.2	23.3	13.1
ResNeXt-101-32×4d[28]	13.6	23.8	13.5
ResNeXt-101-64×4d[28]	13.9	24.2	13.9

Table 3. Different backbone networks on MS COCO dataset.

Distance Metric	#Shots	AP	AP_{50}	AP_{75}
Squared Euclidean	5	7.9	15.4	6.3
Squared Euclidean	10	12.5	22.2	12.3
Squared Euclidean	30	15.8	28.4	15.0
Cosine	5	7.5	14.7	6.0
Cosine	10	12.9	22.9	12.7
Cosine	30	16.4	29.5	15.7

Table 4. Different distance metrics on MS COCO dataset.

CRPN module. We first evaluate the recall on top 100 proposals over 0.5 IoU threshold and CRPN achieves better than standard RPN by 4.4%. We then evaluate the Average Best Overlap ratio (ABO [19]) and CRPN achieves 0.7294 while standard RPN gets 0.7127. The effectiveness of CRPN can also be validated by comparing AP's of RPN/CRPN in Table 1. CRPN outperforms RPN by $1.0\%/1.3\%$ under tighter IoU thresholds such as AP_{80} and AP_{90}. These results indicate that our proposed CRPN can generate proposals with higher quality.

DML module. The major difficulty of few-shot detection is over-fitting and the fully-connected classification head will easily fall into a local minimum when trained with few-shot examples. DML is an effective way to tackle this problem because it transforms support examples to prototypical vectors and calculated the distance of a query samples to each prototype. As shown in Table 1, a standard Faster R-CNN fine-tuned on MS COCO novel classes only achieves 3.1% AP and our DML-based approach with standard RPN achieves 12.0%.

5. CONCLUSION

We contribute a novel few-shot object detection approach with Class-relevant Region Proposal (CRPN) and Deep Metric Learning. CRPN addresses a long-existing problem that novel classes appear as background in base class pre-training and a simple fine-tuning is insufficient to overcome the prior bias. We condition objectiveness filters and anchor delta filters on object class and derive filter weights for novel classes based on feature similarity to base categories. Deep Metric Learning alleviates the necessity of optimizing fully-connected classification head with few-shot samples. Our model has been validated on MS COCO and ImageNet2015 datasets and achieves state-of-the-art performance.

6. REFERENCES

[1] Jake Snell, Kevin Swersky, and Richard Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*. 2017, Curran Associates, Inc.

[2] Sachin Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *ICLR*, 2017.

[3] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*. PMLR, 2016, pp. 1842–1850.

[4] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei, "Memory matching networks for one-shot image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4080–4088.

[5] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang, "Cross-domain few-shot classification via learned feature-wise transformation," *arXiv preprint arXiv:2001.08735*, 2020.

[6] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao, "Lstd: A low-shot transfer detector for object detection," in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32.

[7] 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 few-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[8] 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," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9577–9586.

[9] Bingyi Kang, Z. Liu, Xin Wang, F. Yu, Jiashi Feng, and Trevor Darrell, "Few-shot object detection via feature reweighting," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8419–8428, 2019.

[10] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai, "Few-shot object detection with attention-rpn and multi-relation detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4013–4022.

[11] 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," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, 2015.

[13] Roland Kwitt, Sebastian Hegenbart, and Marc Niethammer, "One-shot learning of scene locations via feature trajectory transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] Y. Wang, R. Girshick, M. Hebert, and B. Hariharan, "Low-shot learning from imaginary data," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7278–7286.

[15] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al., "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*. Lille, 2015, vol. 2.

[16] Meng Ye and Yuhong Guo, "Deep triplet ranking networks for one-shot recognition," *arXiv preprint arXiv:1804.07275*, 2018.

[17] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[18] X. Dong, L. Zheng, F. Ma, Y. Yang, and D. Meng, "Few-example object detection with model communication," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[19] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[20] Junying Huang, Fan Chen, Liang Lin, and Dongyu Zhang, "Plug-and-play few-shot object detection with meta strategy and explicit localization inference," *arXiv preprint arXiv:2110.13377*, 2021.

[21] Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, and Qi Tian, "Learning to learn image classifiers with visual analogy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11497–11506.

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[23] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012.

[26] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, 2010.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015.

[28] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.