YOLOv5 with Dual Attention Network for Object Detection on Drone

Jinsu An
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
Jinsu5023@gmail.com

Muhamad Dwisnanto Putro
Department of Electrical Engineering
Universitas Sam Ratulangi
Manado, Indonesia
dwisnantoputro@unsrat.ac.id

Adri Priadana
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
priadana3202@mail.ulsan.ac.kr

Youlkyeong Lee
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
yklee00815@gmail.com

Junmyeong Kim
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
kjm7029@islab.ulsan.ac.kr

Kanghyun Jo
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
acejo@ulsan.ac.kr

2023 International Workshop on Intelligent Systems (IWIS) | 979-8-3503-0504-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/IWIS58789.2023.10284592

***Abstract*—Detecting the objects in images is a fundamental and crucial research field in computer vision, aiming to locate and classify objects in images. Applying object detection to drones offers numerous advantages. However, to achieve this, the lightweight algorithm is mandatory to operate in real-time on low-cost devices such as Lattepanda and Jetson Nano. In this paper, we propose the YOLOv5 Network with the incorporation of Efficient Residual Bottleneck and DANet. The Efficient Residual Bottleneck can decrease the parameters, making the model lightweight and reducing computational complexity. DANet emphasizes important image features while suppressing noise and unnecessary information, thereby improving object detection performance. We trained our model on the VisDrone dataset, and compared to YOLOv5, the proposed model achieved approximately a 12% improvement in mAP (mean Average Precision) with an mAP value of 22.8, while reducing the number of parameters by approximately 55% to 3,897,433.**

***Index Terms*—Computer Vision, Object Detection, Attention Module, Low-cost Device, Drone Vision**

I. Introduction

A drone is an Unmanned Aerial Vehicle (UAV) that is operated through computers or remote control. Drone technology is continuously advancing and being utilized in various fields. Drone technology refers to the capability of unmanned aircraft to perform flights without onboard human pilots. Drones are used for a wide range of purposes and offer high mobility, detection, collection, and communication capabilities. Due to their small size and aerial maneuverability, drones find applications in diverse fields. In recent times, drones have become extensively used for aerial cinematography. Equipped with high-resolution cameras and gimbal stabilization systems, drones are employed in movie and advertising productions. Additionally, drones have opened up possibilities for capturing footage in previously inaccessible locations with limited human presence, making them crucial in industrial applications. They are commonly employed for video and photo capturing and are widely utilized in industries such as news, film, and advertising. Furthermore, drones are utilized for exploration in hazardous areas, disaster site investigations, agriculture, and environmental monitoring. Drone technology continues to evolve with advancements in autonomous flight, long-range capabilities, and battery life improvement. These developments expand the range of applications for drones, and they are expected to be increasingly utilized in various fields. For example, in industrial settings, drones are used for precise inspections and monitoring tasks, supplementing hazardous tasks and addressing labor shortages. In agriculture, drones are utilized to assess crop conditions and analyze the growth environment, making it easier to manage crops.

Drone Vision refers to the technology that analyzes and interprets the images collected by drones during flight. It is closely related to Computer Vision technology. Computer Vision is a field that enables computers to understand and interpret images using machine learning, pattern recognition, and image processing techniques. Computer Vision technology can perform various tasks such as object classification, detection, tracking, and segmentation. Drone Vision applies these Computer Vision techniques to drones, enabling real-time analysis of images during drone flights. Drones acquire visual information about the surrounding environment through onboard cameras or other sensors and process the images using Computer Vision technology. Drone Vision is utilized in various applications. For example, when performing terrain monitoring using drones, Drone Vision technology can be used to detect and analyze changes or damage in the terrain. Additionally, Drone Vision technology is useful in monitoring crop growth in agricultural fields or inspecting progress at road and construction sites using drones. Drone Vision is a field

within Computer Vision that utilizes image processing, pattern recognition, deep learning, and other techniques to enable drones to effectively perceive the environment and understand the situation. This allows drones to perform tasks such as autonomous flight, object detection and tracking, and terrain monitoring effectively.

II. Related Work

A. Convolutional Neural Network

CNN(Convolutional Neural Network) is one of the types of ANN(artificial neural networks) widely used for image processing and computer vision tasks. CNN extracts features from input images through convolutional and pooling operations and performs tasks such as object detection, classification, and segmentation based on these features. YOLOv5 [1] is an object detection algorithm based on CNN, which stands for "You Only Look Once". YOLOv5 utilizes a one-stage neural network architecture to detect the position and class of objects in real-time. YOLOv5 consists of Backbone, Neck, and Head, where the backbone network is in charge of extracting features from input images.

YOLOv5, based on CNN, performs object detection tasks with the backbone network using CSPDarknet53. This network can handle inputs of various sizes and resolutions. The neck combines and adjusts feature maps of different sizes to improve object detection performance. In the head part, predictions for object classes and bounding boxes are made for each grid cell.

YOLOv5 is a lightweight CNN architecture suitable for real-time object detection. It can effectively detect objects of various sizes and offers flexibility to balance performance and accuracy. YOLOv5 is based on CNN technology and is commonly used when high accuracy and real-time processing are required for object detection tasks.

B. C3 Layer

The C3 layer in YOLOv5 is a component of the network's backbone. C3 is used in the CSPDarknet53 (backbone network) and is responsible for combining feature maps of different sizes based on the depth of the network. The C3 layer is an effective method for reducing or increasing the size of feature maps in the network. To achieve this, the C3 layer alternates between 1x1 convolutions and 3x3 convolutions to generate feature maps of various sizes. This configuration helps balance the efficiency and performance of the network. The operation of the C3 layer is as follows:

The input feature map is dimensionally reduced using a 1x1 convolution, generating feature maps of different sizes. The feature map reduced by the 1x1 convolution is then processed by a 3x3 convolution. This step captures features of various scales while preserving spatial information. The 1x1 and 3x3 convolutions are applied alternately multiple times. This allows the combination of feature maps of different sizes, effectively capturing the varied sizes and features of objects. The C3 layer is used in the CSPDarknet53 backbone network and is effectively applied to object detection tasks in YOLOv5. The C3 layer flexibly handles feature maps of different sizes and improves the accuracy of object detection. This enables YOLOv5 to provide performance in effectively detecting objects of various sizes.

C. Path Aggregation Network

PANet (Path Aggregation Network) is a neck structure used in YOLOv5. The neck combines and adjusts feature maps of various sizes to improve the object detection performance. PANet consists of top-down and bottom-up pathway. The bottom-up pathway is responsible for generating feature maps of different sizes. This pathway extracts feature maps of various scales from a given input image, where smaller feature maps capture contextual information from a wider area and larger feature maps capture more detailed object information. The top-down pathway starts with smaller feature maps and gradually moves to larger-sized feature maps. This pathway utilizes higher-level features generated from smaller feature maps to enhance the fine-grained details of objects. PANet flexibly handles feature maps of different sizes, improves the accuracy of object detection, and is designed to effectively capture the size and features of objects. It contributes to achieving high performance and accuracy in YOLOv5.

III. Proposed Method

It is necessary to reduce the parameter count or computational load of the deep learning object detection network to operate object detection algorithms in real-time. In this paper, two improvements are made to the architecture of YOLOv5, as shown in the figure. First, the Efficient Residual Bottleneck (ERB) [2] is proposed to enhance the model's lightweight design and reduce computational load, with a focus on improving the C3 layer. Second, the emphasis is placed on enhancing the object detection performance by applying the Dual Attention Network (DANet) [3] instead of Spatial Pyramid Pooling Fast (SPPF).

- An object detection algorithm designed to swiftly locate objects in real-time, suitable for use on low-cost devices.
- More efficient and effective feature fusion provides a method for improving the speed and detection performance of the model.

A. The Backbone

YOLOv5 architecture is comprised of three main components: Backbone, Neck, and Head. The Backbone captures image features and transfers them to the Neck, which subsequently relays these features to the Head. Extracting good image features is a crucial aspect of object detection, making it vital to extract high-quality image features in the Backbone. Firstly, good image features capture the essential visual characteristics of objects. They extract various visual features, including shape, color, texture, and more, which contribute to object detection. Secondly, good image features accurately extract the position and size of object bounding boxes, aiding in the precise localization of objects.

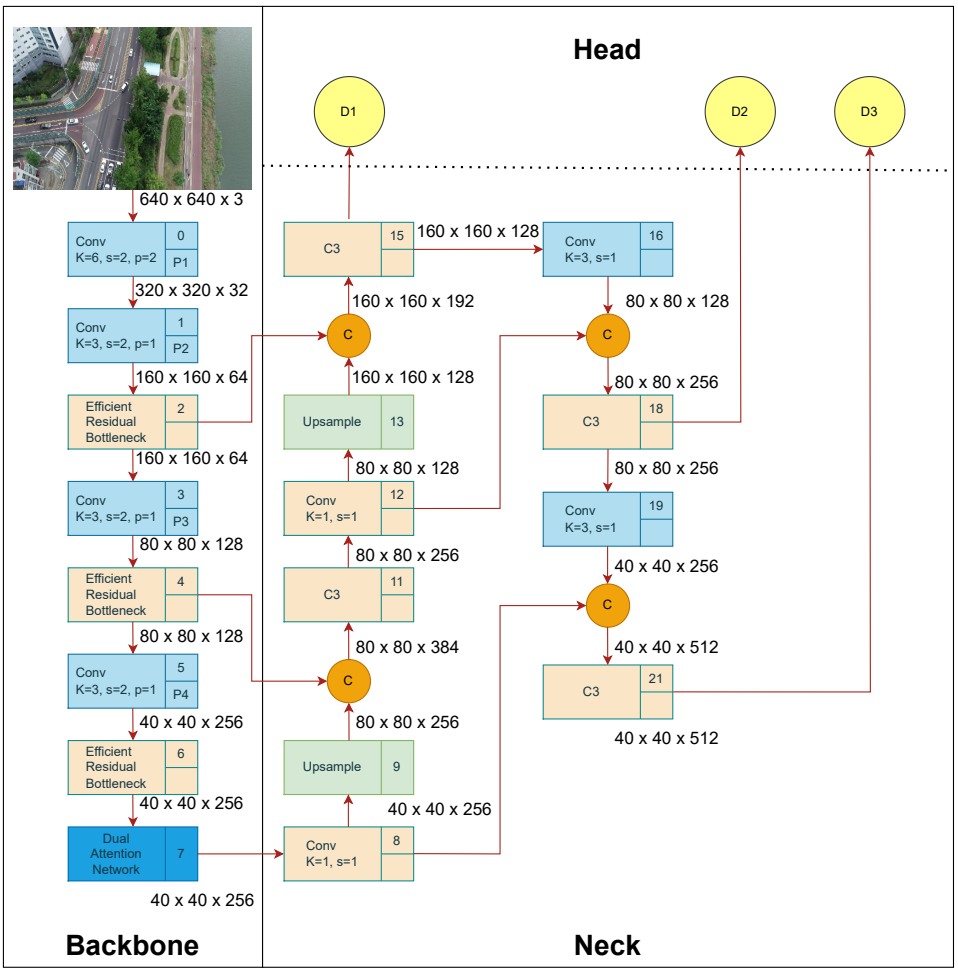

Fig. 1. Proposed Architecture. YOLOv5 network applied with Efficient Residual Bottleneck and Dual Attention Network.

B. Efficient Residual Bottleneck

Efficient Residual Bottleneck is a layer designed to reduce the computational load of the C3 layer in the existing YOLOv5. Decreasing the number of convolutional layers helps to reduce the parameter count and computational load, which in turn improves the speed of the model. Additionally, changing the order of feature fusion allows for adjusting the interaction between the extracted image features.

C. Dual Attention Network

Dual Attention Network is one of the neural network architectures used for image analysis tasks. This architecture focuses on emphasizing important features by leveraging both spatial and channel information of the image while suppressing noise or irrelevant information to improve performance. The Dual Attention Network is comprised of two main components. The first is the Spatial Attention mechanism, which utilizes the spatial information of the input image to highlight important locations and features. It processes the input image in a specific manner to calculate spatial weights and applies them to the original image. These spatial weights help to focus on significant objects or regions while suppressing unnecessary background or noise. The second is the Channel Attention mechanism, which utilizes the channel information of the input image to emphasize important features. It considers the interrelationships between channels in the input image to calculate channel weights and applies them on a per-channel basis. This highlights important features while suppressing noise or irrelevant channel information, leading to performance enhancement. By combining Spatial Attention and Channel Attention in parallel, the Dual Attention Network emphasizes important features of the input image, resulting in performance improvements in various scenarios of computer vision. This architecture is applied instead of the SPPF layer of YOLOv5 for performance improvement, and it is particularly useful when focusing on important parts of an image.

D. Loss Function

The loss function in the proposed method comprises three main components: Localization, Confidence, and Class Loss. Localization loss measures the disparity between the predicted bounding box and the actual ground truth bounding box. This component quantifies the error in terms of the

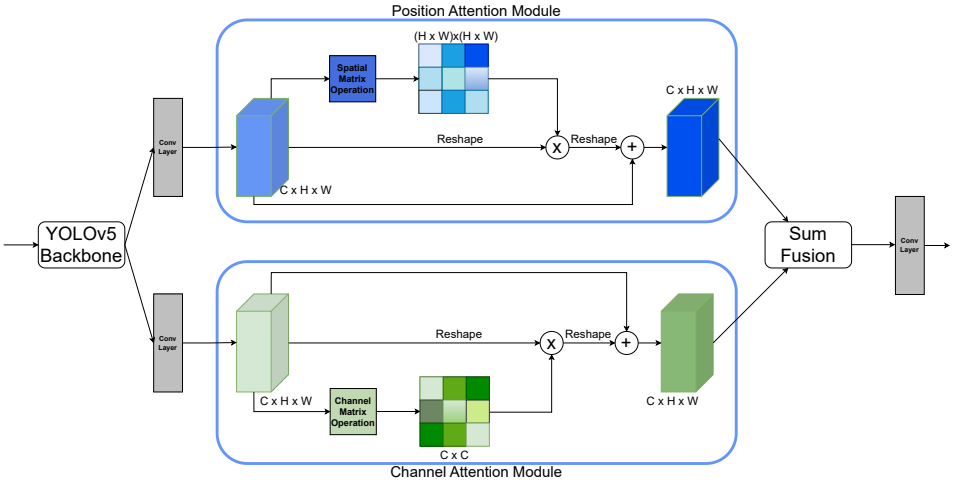

Fig. 2. Overview of Dual Attention Network.

center coordinates and dimensions of the predicted bounding box, employing Mean Square Error (MSE). Confidence loss evaluates the dissimilarity in Intersection over Union (IoU) values between the predicted and actual bounding boxes. This is determined as a binary cross-entropy loss between the predicted confidence in the bounding box and the actual confidence. Class loss assesses the differentiation between the predicted object category and the true category, using Multi-class Cross Entropy loss. Each loss function measures the divergence between ground truth and predicted values, and the model is trained to minimize these disparities. In YOLOv5, These three primary components are merged to get the total loss. The weights of each loss function are adjusted considering the presence of objects, the position of bounding boxes, and the object's class. The loss function compares the model's predictions to the ground truth during the training process of YOLOv5, optimizing the model to improve the accuracy of object detection.

$$
\begin{aligned}
L_{MB} = \lambda_{coord} \sum_{g=1}^{G^2} \sum_{a=1}^{A} 1_{ga}^{obj} L_{coord} + \\
\lambda_{obj} \sum_{g=1}^{G^2} \sum_{a=1}^{A} 1_{ga}^{obj} L_{obj} + \\
\lambda_{cls} \sum_{g=1}^{G^2} \sum_{a=1}^{A} 1_{ga}^{obj} L_{cls}.
\end{aligned} \tag{1}
$$

IV. Implementation Setup

This session explains the experimentation with the YOLOv5 network, which incorporates Efficient Residual Bottleneck and Dual Attention Network, using the VisDrone [4] dataset. The experiments are conducted in a Linux environment, and the model is trained and tested using PyTorch. The equipment used for the experiments includes an Intel Xeon Gold CPU and Nvidia Tesla A100 40GB GPU.

V. Experimental Results

A. Evaluation on VisDrone Dataset

VisDrone dataset is a large-scale dataset for computer vision research based on aerial drone-captured images. It is used for various tasks such as object detection, object tracking, and visual tracking. The VisDrone dataset consists of images captured in various environments, including urban areas, agricultural fields, and highways.

The dataset contains diverse types of objects, including people, vehicles, bicycles, buildings, and more. These objects appear in various sizes and positions in the images, providing an effective platform for research on object detection and tracking in different scenarios.

VisDrone dataset serves as a standardized benchmark for computer vision and machine learning researchers to develop and evaluate various algorithms. It offers high-resolution images and a wide range of objects in different environments, posing realistic and challenging problems for the advancement and improvement of technologies. As a result, the VisDrone dataset contributes significantly to computer vision research and is recognized as a valuable resource with applications in various fields.

Our proposed method experiments object detection with drone images and videos. Since the VisDrone dataset contains many small-sized objects, high-resolution images or methods capable of extracting object features effectively are required for their detection. In this paper, we propose a method using DANet to enhance the accuracy of object features. Through this approach, we observed an improvement in object detection performance, achieving 22.8mAP, which is approximately 12% higher than the original YOLOv5. The parameter count also reduced significantly by approximately 55%, showing 3,897,433.

VI. Conclusion

This paper proposes an improved model by applying Efficient Residual Bottleneck and Dual Attention Network to

Fig. 3. Object Detection Result on VisDrone 2019 dataset.

TABLE I
DETECTION PERFORMANCE COMPARISONS ON VISDRONE DATASET

Model	AP	AP50	Backbone
HTC-drone [5]	22.6	45.2	ResNet50
TridentNet [6]	22.5	43.3	ResNet101
CenterNet-Hourglass [7]	22.4	41.8	Hourglass-104
Retinaplus [8]	20.6	40.6	ResNeXt-101
ERCNNs [9]	20.5	41.2	ResNeXt-101
SAMFR-Cas RCNN [10]	20.2	40.0	SERexNeXt-50
Cascade R-CNN++ [10]	18.3	33.5	SERexNeXt-50
EnDet	17.8	37.3	ResNet101-fpn
DCRCNN [11]	17.8	42.0	ResNeXt-101
ODAC	17.4	40.6	VGG
DA-RetianNet [12]	17.1	35.9	ResNet101
MOD-RETINANET [8]	17	33.8	ResNet50
DBCL [13]	16.8	31.1	Hourglass-104
ConstraintNet [14]	16.1	30.7	Hourglass-104
CornetNet* [15]	17.4	34.1	Hourglass-104
Light-RCNN* [16]	16.5	32.8	ResNet101
FPN* [17]	16.5	32.2	ResNet50
Cascade R-CNN* [18]	16.1	31.9	ResNeXt-101
DetNet59* [19]	15.3	29.2	ResNet50
RefineDet* [20]	14.9	28.8	ResNet101
RetinaNet* [8]	11.8	21.4	ResNet101
Proposed Method	**22.8**	**39.0**	**Improved CSPDarknet53**

TABLE II
DETECTION PERFORMANCE COMPARISONS ON OTHER METHODS

Model	# parameters	GFLOPs	AP
YOLOv5 small	7,046,599	15.9	20.1
YOLOv5s w DANet 123	7,951,977	16.8	19.7
YOLOv5s w ERB DANet-123	7,776,937	16.4	19.9
YOLOv5s w ERB 4det DANet-123	8,134,614	33.5	21.2
YOLOv5s w DANet-111	7,582,057	15.2	19.7
YOLOv5s w ERB DANet-111	7,407,017	14.8	19.2
YOLOv5s w ERB 4det DANet-111	7,764,694	31.9	20.9
YOLOv5s w ERB DANet wo C5	3,897,433	28.5	22.8

YOLOv5. The Efficient Residual Bottleneck is utilized to decrease the complexity of computation, and the Dual Attention Network is employed instead of the SPPF layer to enhance object detection rates through various image feature fusion methods. we train our proposed model on the VisDrone dataset, achieving an mAP of 22.8mAP, approximately 12% higher than YOLOv5 small, and reducing the number of parameters to 3,897,433, about 55% lower than the original YOLOv5.

In future plans, we aim to reference the detector of YOLOv8 to further improve the detector of YOLOv5 in order to enhance object detection performance. Additionally, we plan to conduct research on more effective ERB layers to make the network enable object detection on low-cost devices in real-time.

ACKNOWLEDGMENT

This result is supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003)

REFERENCES

[1] G. Jocher, A. Stoken, and J. Borovec, "ultralytics/yolov5: v3.0." [Online]. Available: https://doi.org/10.5281/zenodo.3983579

[2] J. An, M. D. Putro, and K.-H. Jo, "Efficient residual bottleneck for object detection on cpu," in *2022 International Workshop on Intelligent Systems (IWIS)*. IEEE, 2022, pp. 1–4.
[3] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3141–3149.
[4] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.
[5] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "Hybrid task cascade for instance segmentation," *CoRR*, vol. abs/1901.07518, 2019. [Online]. Available: http://arxiv.org/abs/1901.07518
[6] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," *CoRR*, vol. abs/1901.01892, 2019. [Online]. Available: http://arxiv.org/abs/1901.01892
[7] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 6568–6577.
[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2018.
[9] N. Xie, S. Li, and J. Zhao, "Ercnn: Enhanced recurrent convolutional neural networks for learning sentence similarity," in *Chinese Computational Linguistics*, M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu, Eds. Cham: Springer International Publishing, 2019, pp. 119–130.
[10] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 6154–6162.
[11] S. Chakraborty, S. Aich, A. Kumar, S. Sarkar, J.-S. Sim, and H.-C. Kim, "Detection of cancerous tissue in histopathological images using dual-channel residual convolutional neural networks (dcrcnn)," in *2020 22nd International Conference on Advanced Communication Technology (ICACT)*, 2020, pp. 197–202.
[12] G. Pasqualino, A. Furnari, G. Signorello, and G. M. Farinella, "An unsupervised domain adaptation scheme for single-stage artwork recognition in cultural sites," *Image and Vision Computing*, p. 104098, 2021.
[13] Y. Wu, Z. Cheng, Z. Xu, and W. Wang, "Segmentation is all you need," *CoRR*, vol. abs/1904.13300, 2019. [Online]. Available: http://arxiv.org/abs/1904.13300
[14] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *CoRR*, vol. abs/1904.07850, 2019. [Online]. Available: http://arxiv.org/abs/1904.07850
[15] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," *International Journal of Computer Vision*, vol. 128, pp. 642–656, 2020.
[16] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Light-head R-CNN: in defense of two-stage object detector," *CoRR*, vol. abs/1711.07264, 2017. [Online]. Available: http://arxiv.org/abs/1711.07264
[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 936–944.
[18] Z. Cai and N. Vasconcelos, "Cascade R-CNN: delving into high quality object detection," *CoRR*, vol. abs/1712.00726, 2017. [Online]. Available: http://arxiv.org/abs/1712.00726
[19] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Detnet: A backbone network for object detection," *CoRR*, vol. abs/1804.06215, 2018. [Online]. Available: http://arxiv.org/abs/1804.06215
[20] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 4203–4212.