

# Cascaded Feature Fusion Network for Small-size Pedestrian Detection

Shilong Yu<sup>1</sup> and Chenhui Yang<sup>2</sup>

<sup>1</sup>Department of Computer Science and Technology, Xiamen University, Xiamen 361000, Fujian, China

yushilong@stu.xmu.edu.cn

<sup>2</sup>Intelligent Laboratory, Xiamen University, Xiamen 361000, Fujian, China

chyang@xmu.edu.cn

**Abstract.** Deep neural network-based target detectors cannot sufficiently extract effective features for detecting small-size pedestrians. In this letter, we propose a deep cascaded network framework for small-size pedestrian detection, which contains an Iterative Feature Augmentation module and a Residual Attention Fusion module. Specifically, the Iterative Feature Augmentation module adopts bilinear interpolation sampling and channel reshaping in the deep backbone network to achieve feature fusion at different scales. Moreover, we also introduce a feature fusion coefficient to select small-size features. The Residual Attention Fusion module is constructed by stacking attention modules, and the attention modules at different depths produce adaptive changes in perceptual features. Each attention module is a bottom-up feedforward structure and features are re-constructed by residual connection between attention modules. Experiments on Tiny Citypersons, Caltech, and Tiny Person challenging datasets show that our proposed modules achieve significant gains, with an almost 10% improvement in pedestrian average miss rate and precision compared to baseline networks.

**Keywords:** Cascaded convolutional neural network (CNN), Pedestrian Detection, Residual Attention, Image Processing.

## 1 Introduction

Pedestrian detection is critical in real-world applications in advanced driver assistance systems, disaster rescue, and video surveillance[1]. Although the current advanced single-stage object detection algorithms such as SSD, YOLO series, and two-stage object detection algorithm Fast/Faster-RCNN, etc., have achieved the best performance on specific scene datasets (such as MS-COCO, ILSVRC).However, small-size pedestrians and low-resolution images(As Fig.1 shows) pose great challenges for these applications in real-life environments.

Early pedestrian detection algorithms are based on hand-designed features and classifiers. Navneet Dalal *et al.* [2] propose a pedestrian detection algorithm based on HOG + SVM, which has achieved good performance in the field of machine learning. Since HOG features are sensitive to noise, P Dollár, *et al.* [3] propose Integral Channel

Features (ICF) and train classifiers with multiple scales by using the cascade strategy of AdaBoost classifiers, achieving good accuracy. However, when the target size changes on a large scale, the performance of the hand-designed algorithms will be significantly declined.

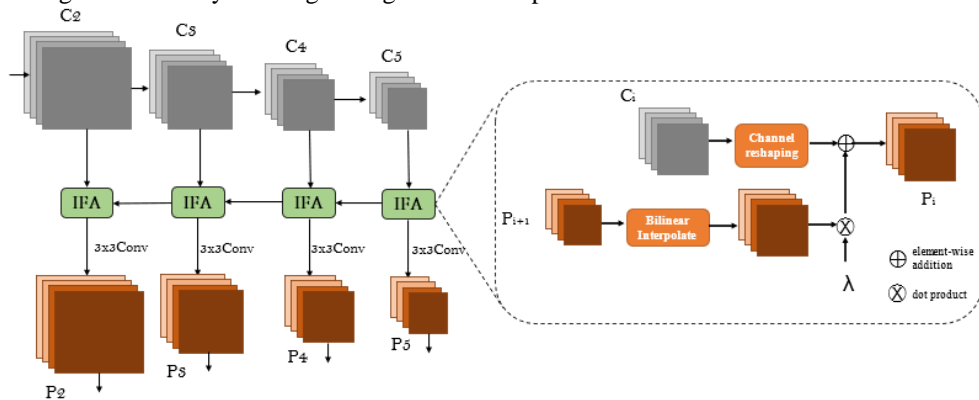


**Fig. 1.** Some small-size pedestrians in Tiny Citypersons, Caltech, and Tiny Person. The mean and standard deviation of the average pedestrian size are 18.0 and 17.4 pixels respectively.

Since deep learning technology has been applied to image[4] and video classification[5], researchers have found that features obtained based on deep learning have strong hierarchical representation and good robustness. Pedestrian detection methods [6]-[12] based on deep learning are emerging. Angelova, A., *et al.* [13] propose a scheme for pedestrian detection using a cascaded convolutional network, drawing on the idea of cascading the AdaBoost classifiers, which significantly improves the detection speed while ensuring detection accuracy. However, the algorithm has not been studied for small pedestrian detection. Li, J., *et al.* [14] analyze the distribution of objects in the Caltech pedestrian detection database, and propose a method called SA-Fast RCNN. Since the features extracted by large and small pedestrians show significant differences, the author designs two sub-networks for detection separately, but multiple networks add a lot of parameters and inference time to the model.

In addition to the above methods, feature fusion is widely used in the field of small-size object detection as an effective method to improve network performance. Y. Liu *et al.* [15] design an encoder architecture combined with convolutional networks to extract multi-scale features and global context information, and use nested join structures

to reconstruct fused features, but the fusion of multi-scale and context features brings a huge amount of computation. Woo, S., *et al.* [16] use deconvolution instead of proximity interpolation in the upsampling module and add a normalization layer and activation function after the FPN[17] convolution layer to form an efficient convolution, which effectively unifies the multi-scale representation and semantic distribution, but introducing additional layers brings a large number of parameters.



**Fig. 2.** An illustration of the first stage and the proposed Iterative Feature Augmentation (IFA) module. The whole process can be divided into two parts: deep-to-shallow iteration and feature fusion inside IFA.

Attention mechanism [18]-[20] act as essential parts in detecting small-size targets. SENet[21] and DFN[22] enhance special channels by explicit feature selection. DANet[23] provides correlation between features by implementing channel attention modules and spatial attention modules.

Based on the analysis of existing methods[24], we propose a new deep cascaded network. The proposed network consists of two stages: *shallow fusion* and *reselection*. In the first stage, the Iterative Feature Augmentation (IFA) module iteratively fuses multi-scale features through a feature weight coefficient; in the second stage, the Residual Attention Fusion (RAF) module learns and fuses features through multiple random HEAD to select valid features. The major contributions of this letter are summarized as follows:

1. For small pedestrian targets, we design an Iterative Feature Augmentation (IFA) module with a weight coefficient to iteratively enhance the initial pedestrian features and reduce the weight of uncorrelated noise.
2. We propose a residual attention fusion (RFA) module that superimposes multiple attention layers and adds residual connections[25] between layers to focus, screen, and fuse pedestrian features. The two modules are cascaded together to obtain valid features for target classification and bounding box prediction.
3. We conduct experiments on different small-size pedestrian datasets, and the results show that the method we design could effectively reduce the miss rate of pedestrian detection, and all the results are better than the baseline network.

## 2 Method

In this section, we will describe our method for small-size pedestrian detection in detail.

### 2.1 Iterative Feature Augmentation

As illustrated in Fig. 2, we propose the IFA module as a feature extractor to learn the initial features of the pedestrian in the first stage, and the whole process can be divided into two parts: deep-to-shallow iteration and feature fusion inside IFA.

For iteration, we select four feature layers of different scales from a fully convolutional neural network. The tensors are output in two forms after being processed by the IFA module, one is iteratively fused with the upper tensors as the input of the next IFA, and the other is input to the next stage after a 3x3 convolution layer. The process can be described mathematically as:

$$P = f(C, f(C', f(C'', P''))) \quad (1)$$

where  $P''$  and  $C'$  represent top-down features and bottom-up features, respectively, which must be consistent in the number of feature channels,  $C''$  represents shallower features, and  $f$  represents the IFA module.

For feature fusion inside IFA, shallow network has smaller receptive fields and could capture more details, which is crucial for detecting small objects; while the deep network has large receptive fields, which will contain too much background noise if detecting small objects. However, the advanced semantic information from deep layers is necessary for object detection.

Therefore, we use the idea of feature fusion of skip connection[25] for reference, firstly, the input tensor  $P''$  ( $C \times H' \times W'$ ) from deep is sampled by bilinear interpolation to obtain a tensor of size ( $C \times H \times W$ ), and then a weight coefficient  $\lambda$  is introduced to multiply with the element at each position of  $P''$ . The tensor  $C'$  from the middle layer ( $C' \times H \times W$ ) is reshaped by the channel to obtain a tensor of size ( $C \times H \times W$ ), and finally these two tensors are added by elements to obtain the output tensor. The process can be described mathematically as:

$$P' = C' \oplus (f(P'') \otimes \lambda) \quad (2)$$

where  $P''$  represents the upper layer features,  $P'$  represents the next layer output features,  $C'$  represents the middle layer features to be processed,  $\lambda$  represents the feature weight coefficient,  $f$  represents the bilinear interpolation function,  $\oplus$  represents element-wise addition, and  $\otimes$  represents dot product.

Full fusion ( $\lambda=1$ ) will introduce additional noise, so  $\lambda$  ranges from 0 to 1 and this process is called *shallow fusion*.

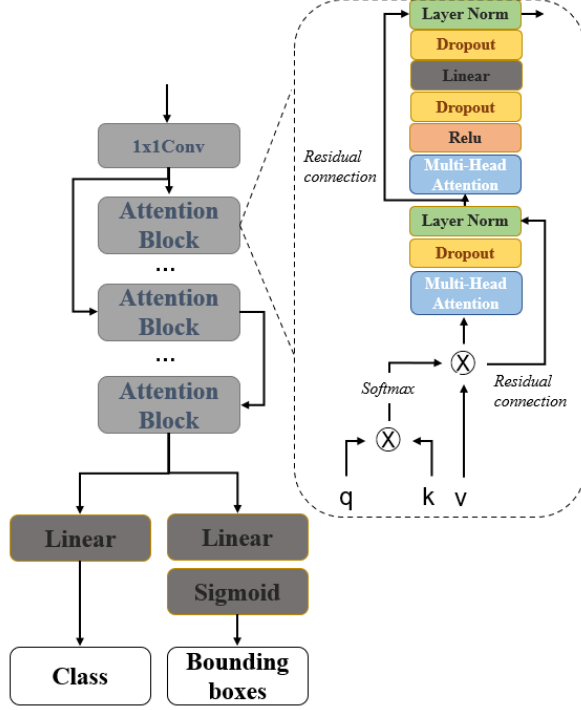


Fig. 3. The overall structure of the Residual Attention Fusion module.

## 2.2 Residual Attention Fusion

As shown in Fig. 3, we propose the Residual Attention Fusion module. By dividing the model into multiple heads and forming multiple subspaces, the model can focus on different aspects of information. The Multi-Head Attention in the figure above is Scaled Dot-Product Attention done multiple times and combined. We extend the ability of the model to focus on different locations by using multiple heads of attention, thus giving the residual attention model multiple seeds of expression. Specifically, the features from the previous stage are fed into the RAF after channel reshaping, and the first Attention Block has three inputs:  $q$ ,  $k$ , and  $v$ . The values of  $q$  and  $k$  are the values of  $v$  plus the values of positional embedding to calculate the correlation. The process can be described mathematically as:

$$\alpha = \frac{q \cdot k}{\sqrt{d_k}} \quad (3)$$

where  $\alpha$  represents the correlation coefficient and  $d_k$  represents the length of  $k$ . Then calculate the input of the first Attention Block, express mathematically as:

$$Input(q, k, v) = softmax\left(\frac{q \cdot k}{\sqrt{d_k}}\right) \cdot v \quad (4)$$

For each Attention Block, we use two residual connections, Layer Norm [26] is used after each residual connection to ensure the stability of the feature distribution, and Dropout [27] is used after Multi-Head Attention to prevent overfitting. Not only

that, but we also add residual connections between Attention Blocks to ensure feature validity. Iterative fusion and residual fusion are cascaded together, and the features are subjected to class discrimination and bounding box coordinate prediction after passing through the linear layer and sigmoid function.

### 3 Experiments

In this section, we first demonstrate the effectiveness of the proposed method by showing the ablation studies on different datasets (Tiny Citypersons[28], Caltech, and Tiny Person[28]).

#### 3.1 Implementation Details

Yu *et al.* [28] propose a scale-matching algorithm to unify the target scale among different datasets. Tiny Person is a small-size pedestrian dataset collected in various ways. The mean and standard deviation of the target size in the dataset are 18.0 and 17.4. Tiny Citypersons is a small pedestrian dataset obtained by applying the scale matching algorithm to the Citypersons dataset. The pedestrian size in Tiny Citypersons dataset is similar to that of Tiny Person. The Caltech dataset is also scaled.

We use the Pytorch framework to implement the proposed method, where the initial learning rate is set to 0.0001 and decays 0.1 per 30 epochs, batch size is 2, cross entropy loss is used as optimization. There is no special processing for the original image in data augmentation, only horizontal flip and vertical flip are used. We train 100 epochs per ablation experiment.

#### 3.2 Ablation Studies

Generally speaking, object detection focuses on AP values. More specifically, it focuses on aps with different IOU values, such as those with IOU values of 0.75 and 0.5. Pedestrian detection will also focus on AP, but generally focus on AP when the IOU is 0.5. In addition, Miss Rate will be paid attention to separately. MR means that the average miss rate of FPPI under 9 points uniformly taken in log space within the range of [0.01, 1]. That is, the average miss rate when FPPI is [0.0100, 0.0178, 0.03160, 0.0562, 0.1000, 0.1778, 0.3162, 0.5623, 1.000] respectively. In general, the lower the MR, the better, and the higher the AP, the better.

##### 3.2.1 Ablation for weight coefficient $\lambda$

We introduce the feature weight coefficient to integrate more small target features into the fusion features and reduce the proportion of irrelevant features. To verify the validity of  $\lambda$ , we perform ablation experiments at intervals in the IFA module. As can be seen from Table 1, when  $\lambda=0.3$ , the average miss rate on Tiny Citypersons can reduce to 29.25%, and the average precision can increase to 62.71%. On Caltech, the

average miss rate decrease to 45.64%, and the average precision rise to 46.21%. These results show that complete feature fusion will cause interference in small target detection, and the appropriate fusion ratio can improve the detection performance.

### 3.2.2 Ablation for Attention Block

As summarized in Table 2, we try stacking different numbers of Attention Blocks to build our network. In our experiment, there is an optimal value(layers=6) for the number of blocks, and when the number decreases or increases, it does not improve the performance of the network.

**Table 1.** Studies of Weight Coefficient  $\lambda$ .  $\lambda$  has a value range of 0-1, and  $\lambda = 1$  indicates complete fusion

$\lambda$	Tiny Citypersons		Caltech	
	$MR_{50}$	$AP_{50}$	$MR_{50}$	$AP_{50}$
0.2	32.17%	59.04%	51.73%	42.93%
0.3	<b>29.25%</b>	<b>62.71%</b>	<b>45.64%</b>	<b>46.21%</b>
0.4	32.05%	60.24%	48.51%	44.62%
0.6	33.55%	59.20%	50.25%	43.68%
0.8	34.10%	58.41%	50.41%	42.22%
1.0	35.15%	57.91%	52.18%	41.84%

### 3.2.3 Ablation for head of the Attention

The multi-head mechanism of the attention module can learn features from different aspects. Abundant features are considered helpful for detecting targets, but it is not absolute. We vary the number of heads used for feature learning in the ablation experiment and the results are shown in Table 3. The average miss rate of 8 heads is 4.75% lower than that of 4 heads and 4.4% lower than that of 16 heads. The average precision of 8 heads is improved by 8.16% and 4.84% compared to 4 heads and 16 heads, respectively.

**Table 2.** Studies of Layers of Attention Block. When Layers=6, the model achieves the best performance.

Tiny Citypersons
------------------

Layers of Attention Block	$MR_{50}$	$AP_{50}$
4	35.30%	56.25%
6	<b>29.25%</b>	<b>62.71%</b>
8	42.39%	55.12%

### 3.3 Comparing With Other Methods

We compare our method with other object detection methods on the Tiny Citypersons, Caltech, and Tiny Person datasets, respectively. Our experiments are carried out by integrating the proposed modules into the Detection Transformer framework, so we first conducted experiments with the DETR original model, respectively on Tiny Citypersons and Caltech datasets, and obtained corresponding results.

Subsequently, in order to prove the effectiveness of our design method, we conducted experiments under unified experimental conditions, and the results obtained were summarized in Table 4. It can be seen that on the basis of the original method, on the Tiny Citypersons dataset, our designed method reduces the average false detection rate by nearly 10 percentage points and increases the average accuracy by 7 percentage points under the condition of 50 IOU. In the Caltech dataset, we respectively reduced the average false detection rate by 15 percentage points and increased the average accuracy by 1 percentage point.

**Table 3.** Studies of Number of Head in the Multi-head Attention Module. When Number=8, our method gains 29.25% miss rate and 62.71% average precision.

Number of head	Tiny Citypersons	
	$MR_{50}$	$AP_{50}$
4	34.00%	54.55%
8	<b>29.25%</b>	<b>62.71%</b>
16	33.65%	57.87%

Table 5 shows the comparison between our method and Faster-RCNN method. Specifically, Faster-RCNN is pre-trained on five different data sets and then tested on Tiny Person data set. It can be seen that, although our average accuracy is slightly lower than that of the pre-trained Faster-RCNN on the MSM COCO, we can still reduce the average miss rate by two percentage points compared with the Faster-RCNN.

At the end of the paper, we show some experimental results of our model on the test set, as shown in Fig.4. It can be seen that our model can still detect pedestrians at a high detection rate in different road scenes, even in areas where it is difficult for human eyes to recognize whether there are pedestrians, which indicates the effectiveness of our method.



**Table 4.** Comparing With Baseline Network. Our experiments are carried out by integrating the proposed module into the detection transformer framework.

Method	Tiny Citypersons		Caltech	
	$MR_{50}$	$AP_{50}$	$MR_{50}$	$AP_{50}$
DETR[29]	39.40%	55.08%	50.81%	36.39%
Ours	<b>29.25%</b>	<b>62.71%</b>	<b>45.64%</b>	<b>46.21%</b>

**Table 5.** Comparing With Other Methods. Other methods are pre-trained on different datasets based on Faster-RCNN.

Method	Tiny Person	
	$MR_{50}$	$AP_{50}$
ImageNet[4]	87.78%	43.55%
COCO	86.58%	43.38%
COCO100	87.67%	43.03%
SM COCO	86.30%	46.77%
MSM COCO	85.71%	<b>47.29%</b>
Ours	<b>83.51%</b>	46.68%

## 4 Conclusion

In this letter, we propose a deep cascaded network for small-size pedestrian detection, which contains an Iterative Feature Augmentation module and a Residual Attention Fusion module. *Shallow fusion* proves that fusion of multi-scale features with appropriate proportions is helpful to detect small pedestrian targets. *Reselection* starts from the overall feature map and focuses on the areas of interest and the residual connection can help the network converge quickly. Experiment results show that our method has better performance on several challenging datasets, and the proposed modules obtain significant gains in the integration framework without introducing additional parameters and computation.

However, due to the large size of DETR model, its inference speed is greatly limited, and multiple attention modules make our model look very heavy. Therefore, in the future, we hope to move towards a lightweight model so that it can be used in other areas.



**Fig. 4.** Some detection results for small-size pedestrians in Tiny Citypersons, Caltech, and Tiny Person.

## References

1. Li, Ying et al. "Decoupled Pose and Similarity Based Graph Neural Network for Video Person Re-Identification." *IEEE Signal Processing Letters* 29 (2022): 264-268.
2. Dalal, N. , and B. Triggs . "Histograms of Oriented Gradients for Human Detection." *IEEE Computer Society Conference on Computer Vision & Pattern Recognition* IEEE, 2005.
3. P Dollár, et al. "Integral Channel Features." British Machine Vision Conference, BMVC 2009, London, UK, September 7-10, 2009. Proceedings DBLP, 2009.

4. Krizhevsky, A. , I. Sutskever , and G. Hinton . "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in neural information processing systems* 25.2(2012).
5. W. Lu, J. Lin, P. Jing and Y. Su, "A Multimodal Aggregation Network With Serial Self-Attention Mechanism for Micro-Video Multi-Label Classification," in *IEEE Signal Processing Letters*, vol. 30, pp. 60-64, 2023, doi: 10.1109/LSP.2023.3240889.
6. S. Kim, S. Kang, H. Choi, S. S. Kim and K. Seo, "Keypoint Aware Robust Representation for Transformer-Based Re-Identification of Occluded Person," in *IEEE Signal Processing Letters*, vol. 30, pp. 65-69, 2023, doi: 10.1109/LSP.2023.3240596.
7. B. Wu, Y. Feng, Y. Sun and Y. Ji, "Feature Aggregation via Attention Mechanism for Visible-Thermal Person Re-Identification," in *IEEE Signal Processing Letters*, vol. 30, pp. 140-144, 2023, doi: 10.1109/LSP.2023.3244747.
8. Rajaram, R. N. , E. Ohn-Bar , and M. M. Trivedi . "An Exploration of Why and When Pedestrian Detection Fails." *2015 IEEE 18th International Conference on Intelligent Transportation Systems - (ITSC 2015)* IEEE, 2015.
9. Luo, Q. , J. H. Gai , and H. Y. Zheng . "Small-scale Pedestrian Detection Based on Multi-scale Feature Fusion." *Computer Engineering & Software* (2019).
10. Hong, M. , et al. "SSPNet: Scale Selection Pyramid Network for Tiny Person Detection from UAV Images." (2021).
11. Zhang, X. , et al. "Too Far to See? Not Really! --- Pedestrian Detection with Scale-aware Localization Policy." *IEEE Transactions on Image Processing* (2018):1-1.
12. Bunel, R. , F. Davoine , and P. Xu . "Detection of Pedestrians at Far distance. " *2016 IEEE International Conference on Robotics and Automation (ICRA)* IEEE, 2016.
13. Angelova, A. , et al. "Real-Time Pedestrian Detection With Deep Network Cascades." *British Machine Vision Conference* 2015.
14. Li, J. , et al. "Scale-aware Fast R-CNN for Pedestrian Detection." IEEE, 10.1109/TMM.2017.2759508. 2015.
15. Y. Liu, Z. Yang, J. Cheng and X. Chen, "Multi-Exposure Image Fusion via Multi-Scale and Context-Aware Feature Learning," in *IEEE Signal Processing Letters*, vol. 30, pp. 100-104, 2023, doi: 10.1109/LSP.2023.3243767.
16. Woo, S. , S. Hwang , and I. S. Kweon . "StairNet: Top-Down Semantic Aggregation for Accurate One Shot Detection." IEEE Computer Society, 10.48550/arXiv.1709.05788. 2017.
17. Lin, T. Y. , et al. "Feature Pyramid Networks for Object Detection." *IEEE Computer Society* (2017).
18. G. Li, L. Li and J. Zhang, "BiAttnNet: Bilateral Attention for Improving Real-Time Semantic Segmentation," in *IEEE Signal Processing Letters*, vol. 29, pp. 46-50, 2022, doi: 10.1109/LSP.2021.3124186.
19. M. Garg, D. Ghosh and P. M. Pradhan, "Multiscaled Multi-Head Attention-Based Video Transformer Network for Hand Gesture Recognition," in *IEEE Signal Processing Letters*, vol. 30, pp. 80-84, 2023, doi: 10.1109/LSP.2023.3241857.
20. Min, K. et al. "Attentional feature pyramid network for small object detection." *Neural networks : the official journal of the International Neural Network Society* 155 (2022): 439-450 .
21. Jie, H. , et al. "Squeeze-and-Excitation Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP.99(2017).
22. Yu, C. , et al. "Learning a Discriminative Feature Network for Semantic Segmentation." *IEEE* (2018).
23. Fu, J. , et al. "Dual Attention Network for Scene Segmentation." *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* IEEE, 2020.

24. K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multi-task Cascaded Convolutional Networks," in *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.
25. Orhan, A. E. , and X. Pitkow . "Skip Connections Eliminate Singularities." (2017).
26. Ba, J. L. , J. R. Kiros , and G. E. Hinton . "Layer Normalization." (2016).
27. Srivastava, N. , et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research* 15.1(2014):1929-1958.
28. Yu, X. , et al. "Scale Match for Tiny Person Detection." *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)* IEEE, 2020.
29. Carion, Nicolas , et al. "End-to-End Object Detection with Transformers." (2020).
30. Ren, Shaoqing , et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 39.6(2017):1137-1149.