Lifei He, Ryutarou Ohbuchi, Ming Jiang, Takahiko Furuya, Min Zhang, Cascaded Multi-Channel Feature Fusion for Object Detection, in Proc. ICCCV'20: 2020 the 3rd International Conference on Control and Computer Vision, August 2020, Pages 11–16

Cascaded Multi-Channel Feature Fusion for Object Detection

1st Author
1st author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
1st author's E-mail address

2nd Author
2nd author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
2nd E-mail

3rd Author
3rd author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
3rd E-mail

ABSTRACT

In this paper, we propose and evaluate a novel object detection architecture called Cascaded Multi-Channel Feature Pyramid Network, or CM-FPN. The proposed architecture, based on the idea of feature pyramid network [10], employs cascaded feature pyramids to obtain salient, a highly semantic feature pyramid for object location proposal/regression and classification. The architecture uses deeper channels for the more semantic stages of the cascade so that the salient features are extracted and preserved. Experimental evaluation of the proposed approach has shown that the proposed combination of cascaded top-down feature pyramid and deeper channel contribute to higher object detection accuracy.

CCS Concepts

• Computing methodologies→Artificial intelligence→Computer vision→Computer vision problems→Object detection

Keywords

Feature pyramid network, .

1. INTRODUCTION

Object detection aims to detect instances of objects in images and associate these objects with one of a set of class labels. Current generation of object detection algorithms, arguably originating at R-CNN [1], are based on deep covolutional neural network (CNN), achieving significantly higher performance than the algorithms in pre-CNN era. Current CNN-based object detection architectures can be classified into two; two-stage network and one-stage network. The two-stage network consists of three distinct steps, the feature map extraction, region proposal generation, and classification. The two-stage network include R-CNN [1], SPP-net [2], Fast R-CNN [3], and Faster R-CNN [4]. One stage-network, on the other hand, merges region proposal generation and classification into one. One stage-networks include SSD [5], several incarnations of YOLO (YOLO v1 [6], etc.) and RetinaNet [7].

In both classes of architecture, object detection performance depends on the performance of the feature extraction part, or CNN backbone network. A backbone is a CNN that accepts an input image and produces a feature map or set of feature maps having higher semantic content. Accurate detection/regression of object regions and accurate classification of objects in the detected regions depends on quality of the feature map. For accurate classification of objects, the feature map should have strong semantic content. For accurate localization of objects and for detecting small objects, the feature map should have high resolution while maintaining high semantic content.

Earlier object detection algorithms, such as Faster R-CNN [4] used a “standard” image recognition CNNs, e.g., VGG[8] or ResNet [9] as their backbone. These CNN has a multi-layer architecture in which feature maps in earlier (closer to input) layers have higher resolution yet lower semantic content, while feature maps in the later layers have lower resolution yet higher semantic content. Using feature maps in later layers having higher semantic content yet lower resolution lead to lower detection/classification accuracy of smaller objects in the images.

To alleviate this issue, Lin et al. in [10] introduced Feature Pyramid Networks (FPN). The FPN is a combination of convolutional, or bottom-up network and deconvolutional, or top-down, network (Figure 2a). The bottom-up network is used in a classical image recognition pipeline, whose successive layers increases semantic content while decreasing image resolution. The top-down network takes as its input the output of the convolution network having low-resolution yet high semantic content. The top-down network, with its up-sampling layers, successively increases resolution while trying to maintain high semantic content. Lateral connection from the bottom-up network to the top-down network supply high resolution image information, which is fused, via dimension reduction using 1×1 convolution, with the feature maps of the top-down network. Resulting feature pyramid in the top-down network (P5~P2 in Figure 2a) provides high-resolution and highly semantic multi-resolution feature maps for effective object localization and/or classification. The Inside-Outside Net (ION) [11] employs similar idea.

To further strengthen semantic content of the multiresolution feature maps, cascading of backbone network is proposed in Cascade R-CNN [12] and Cascade RPN [13]. By repeating the FPN-like structure multiple times, it is hoped that the semantic saliency would improve. Cascading pyramid produced improvement in object detection accuracy. However, further increases in number of pyramid stages brought reduction in accuracy, possibly due to loss of information, especially in terms of image resolution, in later stages caused by repeated convolution and deconvolution.

In this paper, we try to combine the idea of cascade with deeper, multi-channel feature maps appropriate for the semantic content of

Conference'10, Month 1–2, 2010, City, State, Country.

DOI: http://dx.doi.org/10.1145/12345.67890

Lifei He, Ryutarou Ohbuchi, Ming Jiang, Takahiko Furuya, Min Zhang, Cascaded Multi-Channel Feature Fusion for Object Detection, in Proc. ICCCV'20: 2020 the 3rd International Conference on Control and Computer Vision, August 2020, Pages 11–16

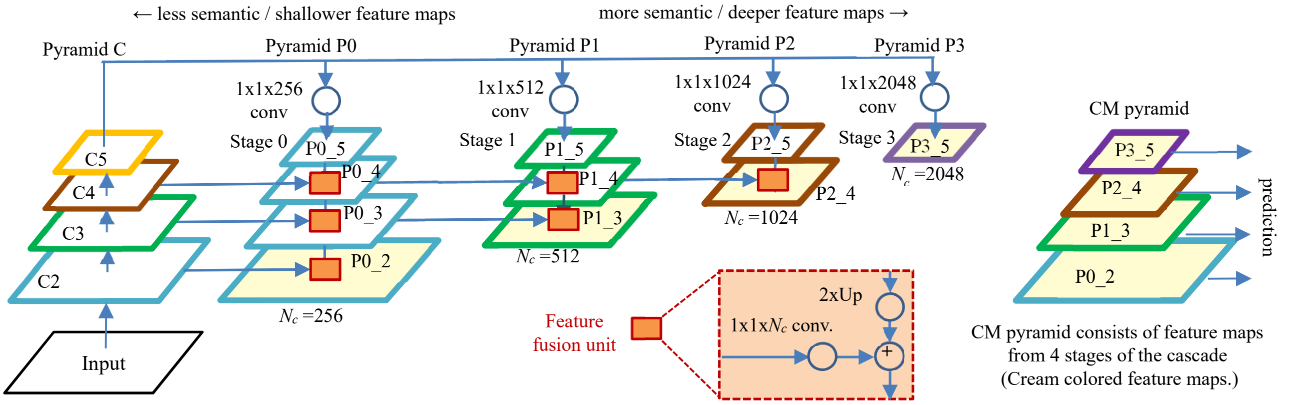


Figure 1. CM-FPN having number of cascade stages T = 4. Pyramid at each stage uses different channel depth; the more semantic the stage, the deeper the channel. The 1x1 convolution for feature fusion at each stage uses corresponding channel depth.

each stage of the cascade. That is, in the earlier stages of the cascade, where semantic content is less, feature maps are shallower. In the later stages of the cascade, where semantic content is higher, feature maps are deeper. The depth of 1 × 1 convolution used in fusing feature maps also vary according to the stages of the cascades. We call the proposed method *Cascaded Multi--Channel Feature Pyramid Network*, or CM-FPN. Figure 1 illustrates our proposed method. We experimentally evaluated the proposed CM-FPN by applying it to both 2-stage as well as 1-stage object detection networks, namely, Faster R-CNN and RetinaNet. The experiment showed that the proposed CM-FPN performs better than the original FPN as well as its cascade only and multi-channel fusion only variants.

Contribution of this paper can be summarized as below;

- Proposal of CM-FPN, which takes better advantage higher semantic content produced by cascaded pyramid approach while reducing the detrimental effect of feature blurring due to the cascade.
- Experimental evaluation of the proposed CM-FPN applied to both Faster R-CNN and RetinaNet using Pascal VOC dataset. Evaluation results show that the proposed CM-FPN performs better than the original FPN. It also performs better than the cascade-only approach and multi-fusion only approach.

2. METHOD

2.1 Cascaded Multi-Channel FPN

The Cascaded Multi-Channel FPN, or CM-FPN uses cascaded multiple auxially pyramids (P0~P3 in Figure 1) to extract highly semantic feature maps at different resolution levels. It also adopts deeper feature maps at cascade stages having higher semantic content.

The overall architecture of the CM-FPN is illustrated in Figure 1. The bottom-up network produces multi-resolution feature pyramid C. The network we use in the experiment is ResNet 101, but other network can be used as the backbone. An input image is processed by the network to produce feature maps C_2~C_5. The C_5 has the highest semantic content yet lowest resolution. The feature map C_5 is processed by multiple deconvolution, or top-down networks with upsampling, to produce pyramidal feature maps P_0~P_3 having higher resolution and high semantic content. Pyramids P_0~P_3 are formed by the most semantic yet lowest-resolution feature map C_5, as well as other feature maps C_2~C_5 of the bottom-up pyramid C. Information from feature maps C_2~C_5 are brought to pyramids P_0~P_3 via direct lateral connections.

In Figure 1, the number of cascade stages (or auxially top-down networks) is set at 4. In the experiments, we vary feature map depths to see their effect without changing the overall network structure, i.e., number of auxiliary pyramids and their respective number of levels.

In the pyramids P_0~P_3, lower resolution yet semantic feature maps trickle down from the top of the top-down network. As it trickles down, up-sampling followed by fusion with feature maps C_2~C_5 bottom-up network produces higher resolution yet semantic feature maps. Note that the pyramids in the later stages of the cascade, e.g., P_3 in Fig. 1, is designed to have higher semantic content by the use of deeper feature maps and correspondingly deeper 1 × 1 convolution used in fusing features. For example, while pyramid P_0 uses the depth 256, pyramid P_3 uses the depth 2048. In the experiments, we used channel depths of 256, 512, 1024, 2048 are used for the pyramids P_0, P_1, P_2, P_3, respectively.

At each pyramid, feature map C_5 of the bottom-up network is processed by $1 \times 1 \times Nc$ convolution to form lowest resolution feature map P_i^5 at the top of the top-down pyramid. Here, the channel depth Nc vary according to the stages of the cascade. They are then processed by the top down pyramid to produce successively higher-resolution feature maps. Feature maps P_i^j of pyramid i at level j, $j = 2$~4 (that, except for the Pi_5 at the top) is computed by using the following equation:

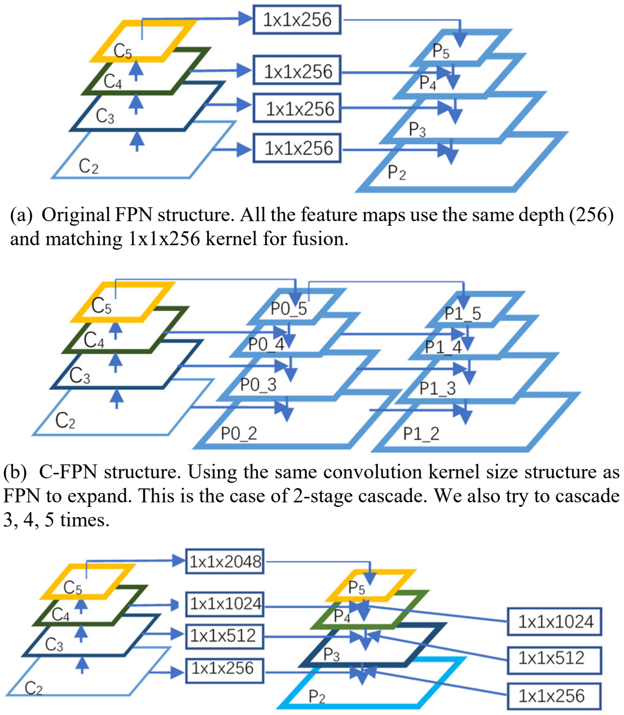
$$P_i^{j-1} = \check{C}_i^{j-1} + P_i^j \qquad (1)$$

where $\check{C}_i^{j-1}$ denotes a feature C_i^{j-1} from Pyramid C convolved with a $1 \times 1 \times Nc$ filter.

3. EXPERIMENTS AND RESULTS

3.1 Experimental Setup

In this section, the proposed CM-FPN is experimentally evaluated by using dataset Pascal VOC 2007 and Pascal VOC 2012 datasets. The Pascal VOC 2007 consists of 5k trainval images + 12k annotated objects, and the Pascal VOC 2012 consists of 11k trainval images + 27k annotated objects. These two datasets annotate 20 types of objects that are common in life.



(a) Original FPN structure. All the feature maps use the same depth (256) and matching 1x1x256 kernel for fusion.

(b) C-FPN structure. Using the same convolution kernel size structure as FPN to expand. This is the case of 2-stage cascade. We also try to cascade 3, 4, 5 times.

(c) M-FPN structure. High-level feature map having higher semantic content are fused using deeper feature map and convolution kernels, and vice versa.

Figure 2. Feature pyramid structure used for object detector. Original FPN(a), proposed C-FPN(b) and M-FPN(c).

We compare the proposed CM-FPN with the original FPN [10] as well as the "ablated" versions of the CM-FPN called Cascaded FPN (C-FPN) and Multi-channel FPN (M-FPN) described below. To evaluate, we add original FPN [10], C-FPN, M-FPN and the proposed CM-FPN to two-stage network Faster R-CNN. We also add the latter three to the RetinaNet. (The RetinaNet contains FPN since its birth so we replaced it with C-FPN, etc.) We use Resnet-101 as the backbone network for all the cases.

FPN (original): In the original FPN, 1×1 convolution is applied to the feature maps of the bottom up network prior to their fusion with the feature maps of the top-down network. The 1×1 convolution is applied to reduce dimensionality of the feature maps so they coincides with those of the top-down network.

Cascaded FPN (C-FPN): Cascaded FPN cascades K top-down pyramids with the hope of increasein semantic content of the feature maps. Figure 3b illustrates its structure for the case K=2. We vary K in the range 2~5 to observe the influence of *K* on accuracy. For the experiment, all the auxiliary pyramids have 4 levels and feature map depth of 256 at all their levels. Note that C-FPN with K=1 corresponds to the orignal FPN.

Multi-channel FPN (M-FPN): Multi-channel FPN has only one top-down pyramid, but employs feature maps having depth "appropriate" for their semantic content. That is, highly semantic, low-resolution feature maps would have deeper feature maps. Similar to the original FPN or FPNs in the Cascaded-FPN, 1×1 convolutions are applied to the feature maps of the bottom-up network. It is illustrated in Figure 2c for the case of 4 level feature map. Let $\mathrm{d}\left(P_i^j\right)$ be the depth of level *j* feature map P_i^j of the pyramid *i* of the M-FPN. We tested four sets of channel depths $\left(\mathrm{d}\left(P_i^2\right), \mathrm{d}\left(P_i^3\right), \mathrm{d}\left(P_i^4\right), \mathrm{d}\left(P_i^5\right)\right)$ in the experiments. Those are;

Case 1: (256, 256, 256, 256), i.e., the original FPN.
Case 2: (256, 512, 512, 512).
Case 3: (256, 512, 1024, 1024).
Case 4: (256, 512, 1024, 2048).

Cascaded Multi-Channel FPN (CM-FPN): Details of CM-FPN is as described in the previous section. We tried the sets of channel depths for the CM-FPN. We tried 4 sets of parameters 1~4 as listed in Table 1.

Table 1. The set of CM-FPN channel depths experimented.

	P0	P1	P2	P3
Case 1	256, 256, 256, 256	256, 256, 256	256, 256	256
Case 2	256, 256, 256, 256	512, 512, 512	512, 512	512
Case 3	256, 256, 256, 256	512, 512, 512	1024, 1024	1024
Case 4	256, 256, 256, 256	512, 512, 512	1024, 1024	2048

Three networks are trained end-to-end, using the training sets of the respective databases, on Nvidia 1080ti GPU. We choose Resnet-101 networks as the backbone and use Adam optimizer with momentum of 0.9 and mini-batch size of 1. We trained 100k mini-batches for Pascal VOC 2007 dataset and 150k mini-batches for Pascal VOC 2007+2012 joint dataset. The learning rate is set at 1×10^{-3} for the first 50k mini-batches, change to 1×10^{-4} for next 20k, and change to 1×10^{-5} for the rest. Other implementation details are the same as the implementations of Faster R-CNN found at [14] and RetinaNet found at [15]. As the numerical index of object detection accuracy mean Average Precision (mAP) is used.

3.2 RESULTS

Tables 2 and 3 shows the accuracy in mAP of the C-FPN and M-FPN, which are ablated versions of the CM-FPN. Similarly, Table 4 shows the accuracy in mAP of the proposed CM-FPN. Results shows in Table 2, 3, and 4 are the results using the Pascal VOC 2007 database. Note that the best accuracies for each case are indicated by bold letters. From the tables, CM-FPN produced best accuracy for both Faster R-CNN and RetinaNet.

Table 2 shows that C-FPN having more than one auxiliary pyramid brings better accuracy than the original FPN having only one auxiliary pyramid. However, when K is varied, there is a peak in accuracy. For Faster R-CNN, the peak is observed at K=3 or K=4, and the accuracy is reduced for a larger value of K=5. We observe that this is caused by attenuation or "blurring" of information due to excessive convolution. Table 3 shows that M-FPN yields higher accuracy when deeper channels are used for feature maps near the top of the pyramid, as in the Case 4, in which depth at the top of the top-down pyramid, $\mathrm{d}\left(P_i^5\right)$ is 2048. Closer look at CM-FPN results in Table 4 also shows that the set of parameters case 3, not case 4, in Table 1 brought the best accuracy. Case 3 has less depth (at 1024) than the case 4 (at 2048) for the feature map at the top of the pyramid. There appears to be an optimal channel depth for the CM-FPN as well.

Table 5 shows the accuracy of CM-FPN using Pascal VOC 2007 + 2012 trainval data set, which is significantly larger than Pascal VOC 2007 only. Accuracy figures of CM-FPN in Table 5 using the extended dataset is higher, for both Faster R-CNN and for RetinaNet, than Table 4 that uses smaller Pascal VOC 2007.

Table 6 shows the size of parameters, in Mbytes, for the FPN and CM-FPN combined with either Faster R-CNN or RetinaNet. While the increase in size over FPN due to CM-FPN is significant, according increase in accuracy would justify it in certain applications.

Table 2. C-FPN detection accuracy in mAP [%] (Pascal VOC 2007 trainval dataset.)

	K=1	K=2	K=3	K=4	K=5
Faster R-CNN	76.1	77.2	**78.0**	**78.0**	77.7
RetinaNet	73.3	**74.2**	74.0	73.2	72.7

Table 3. M-FPN detection accuracy in mAP [%] (Pascal VOC 2007 trainval dataset)

	Case 1	Case 2	Case 3	Case 4
Faster R-CNN	76.1	78.3	**78.6**	**78.6**
RetinaNet	73.3	73.9	**74.2**	74.1

Table 4. CM-FPN detection accuracy in mAP [%] (Pascal VOC 2007 trainval dataset)

	Case 1	Case 2	Case 3	Case 4
Faster R-CNN	76.1	78.0	**79.2**	78.9
RetinaNet	73.3	74.5	**74.8**	74.3

Table 5. CM-FPN detection accuracy in mAP [%] (VOC 2007+2012 trainval dataset)

	T=1(FPN)	T=2	T=3	T=4
Faster R-CNN	78.4	80.5	**81.9**	81.7
RetinaNet	76.3	79.2	**80.5**	80.3

Table 6. Parameter size in MByte.

	Faster R-CNN	RetinaNet
+FPN	459MB	393MB
+CM-FPN (case3)	528MB	497MB

Figure 3 shows examples of object detection. FPN misses the bicycle and its rider or the child by the car. CM-FPN detects these smaller objects with higher confidence than the other methods.

4. CONCLUSION

This paper proposed and evaluated a novel object detection architecture called Cascaded Multi-Channel Feature Pyramid Network, or CM-FPN. The architecture employs cascades of multiple top-down feature pyramid having channel depth adapted to their respective semantic levels to extract a highly semantic and high resolution feature pyramid. That is, the pyramid assigned for more semantic and lower resolution feature map uses the deeper channel. Evaluation using the Pascal VOC 2007 and Pascal VOC 2012 datasets has shown that the proposed CM-FPN performs better than FPN. It also performs better than cascade only C-FPN and deeper multi-channel fusion only M-FPN.

5. REFERENCES

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", in proc. CVPR 2014, pp. 580-587.

[2] K. He, X. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” in IEEE Trans. PAMI, 2015, pp. 1904–1916.

[3] R. Girshick, “Fast R-CNN,” in Proc. IEEE ICCV, 2015, pp. 1440-1448.

[4] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. NIPS, 2015, pp. 91–99.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A.-C. Berg, “SSD: Single shot multibox detector,” in Proc. ECCV 2016, pp. 21–37.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, . “You only look once: Unified, real-time object detection,” in Proc. CVPR 2016, pp. 779-788.

[7] T.-Y. Lin, P. Goyal, R. Girshick, and K. He, “Focal loss for dense object detection,” in Proc. ICCV, 2017, pp. 2980-2988.

[8] K. Simonyan, and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. ICLR 2014.

[9] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. CVPR 2016, pp. 770-778.

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. CVPR 2017, pp. 2117–2125.

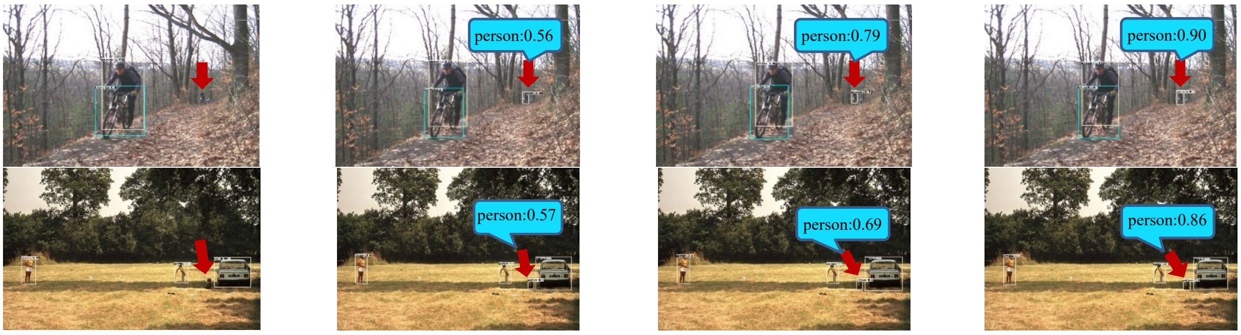
[11] S. Bell, C. Lawrence Zitnick, K. Bala, & R. Girshick, “Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks,” in Proc. CVPR 2016, pp. 2874-2883.

[12] Z. Cai, & N. Vasconcelos, “Cascade R-CNN: High Quality Object Detection and Instance Segmentation,” in Proc. CVPR 2018, pp. 6154-6162.

[13] T. Vu, H. Jang, T.-X. Pham, & C. Yoo, “Cascade RPN: Delving into High-Quality Region Proposal Network with Adaptive Convolution,” in proc. NIPS 2019, pp. 1430-1440.

[14] Y. Yang, “Faster-RCNN_Tensorflow,” https://github.com/DetectionTeamUCAS/Faster-RCNN_Tensorflow (accessed May 24, 2020) .

[15] Y. Yang, "Focal Loss for Dense Object Detection," https://github.com/DetectionTeamUCAS/RetinaNet_Tensorflow (accessed May 24, 2020)



(a) FPN **(b) C-FPN** **(c) M-FPN** **(d) CM-FPN**

Lifei He, Ryutarou Ohbuchi, Ming Jiang, Takahiko Furuya, Min Zhang, Cascaded Multi-Channel Feature Fusion for Object Detection, in Proc. ICCCV'20: 2020 the 3rd International Conference on Control and Computer Vision, August 2020, Pages 11–16

Figure 3. Object detection examples. CM-FPN detects smaller objects with higher confidence values then other methods.

Lifei He, Ryutarou Ohbuchi, Ming Jiang, Takahiko Furuya, Min Zhang, Cascaded Multi-Channel Feature Fusion for Object Detection, in Proc. ICCCV'20: 2020 the 3rd International Conference on Control and Computer Vision, August 2020, Pages 11–16

In order to have a better understanding of the paper, please fill in the following information of all the authors.

Name	Email	Title	Research Field
Lifei He	hlf@hdu.edu.cn	Master student	computer vision
Ryutarou Ohbuchi	ohbuchi@yamanashi.ac.jp	Full professor	3D shape retrieval
Ming Jiang	13588161992@163.com	Full professor	computer vision
Takahiko Furuya	takahikof@yamanashi.ac.jp	Assistant professor	3D shape analysis
Min Zhang	hz_andy@163.com	Lecturer	Computer vision

*Title could be chosen from: master student, PhD candidate, researcher, lecturer, senior lecturer, assistant professor, associate professor, full professor, etc.

Note: this page will not be published. It should be deleted in the camera-ready paper.