Segmentation of retinal detachment and retinoschisis in OCT images based on improved U-shaped network with cross-fusion global feature module

Yang, Changqing, Chen, Xinjian, Su, Jinzhu, Zhu, Weifang, Chen, Qiuying, et al.

Changqing Yang, Xinjian Chen, Jinzhu Su, Weifang Zhu, Qiuying Chen, Jiayi Yu, Ying Fan, Fei Shi, "Segmentation of retinal detachment and retinoschisis in OCT images based on improved U-shaped network with cross-fusion global feature module," Proc. SPIE 11596, Medical Imaging 2021: Image Processing, 1159621 (15 February 2021); doi: 10.1117/12.2580665

SPIE.

Event: SPIE Medical Imaging, 2021, Online Only

Segmentation of retinal detachment and retinoschisis in OCT images based on improved U-shaped network with cross-fusion global feature module

Changqing Yang[1] , Xinjian Chen[1,2] , Jinzhu Su[1] , Weifang Zhu[1], Qiuying Chen[3], Jiayi Yu[3], Ying Fan[3], Fei Shi[1,*]

[1]School of Electronics and Information Engineering, Soochow University, Suzhou, 215006, China
[2]State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou, 215123, China
[3]Shanghai General Hospital, Shanghai 200080, China

ABSTRACT

Retinal detachment (RD) refers to the separation of the retinal neuroepithelium layer (RNE) and retinal pigment epithelium (RPE), and retinoschisis (RS) is characterized by the RNE splitting into multiple layers. Retinal detachment and retinoschisis are the main complications leading to vision loss in high myopia. Optical coherence tomography (OCT) is the main imaging method for observing retinal detachment and retinoschisis. This paper proposes a U-shaped convolutional neural network with a cross-fusion global feature module (CFCNN) to achieve automatic segmentation of retinal detachment and retinoschisis. Main contributions include: (1) A new cross-fusion global feature module (CFGF) is proposed. (2) The residual block is integrated into the encoder of the U-Net network to enhance the extraction of semantic information. The method was tested on a dataset consisting of 540 OCT B-scans. With the proposed CFCNN method, the mean Dice similarity coefficient of retinal detachment and retinoschisis segmentation reached 94.33% and 90.29% and were better than some existing advanced segmentation networks.

Keywords: Optical coherence tomography, cross-fusion global feature module, retinal detachment, retinoschisis

1. INTRODUCTION

For patients with high myopia, retinal detachment and retinoschisis in the macular area of retina seriously impair visual function [1]. With optical coherence tomography (OCT), retinal detachment and retinoschisis can be observed clearly and non-invasively. At present, there are very few technologies that realize the automatic segmentation of retinal detachment and retinoschisis, which is of great significance for tracking the progress of high myopia, for treatment planning, and for prognosis analysis.

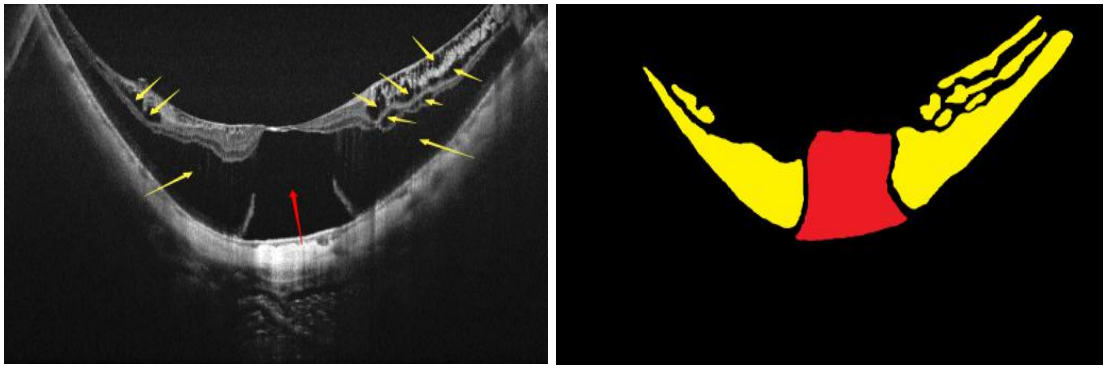
Convolutional neural networks (CNNs) have been widely applied for medical image segmentation. the U-Net [2] integrates the low-level semantic information into the high-level semantic information by adding skip-connections on the basis of the fully convolutional network [3] . This makes up for the loss of detail information due to continued downsampling, and has achieved good segmentation performance in medical image segmentation. Various modifications on the U-net have been proposed to further enhance its segmentation ability. CE-net [4] improved U-Net using dense atrous convolution (DAC) block and residual multi-kernel pooling (RMP) block to capture more high-level features and preserve more spatial information. CPFNet [5] was also based on the U-shape structure, which combines two pyramidal modules to fuse global/multi-scale context information.

*Corresponding author: E-mail: shifei@suda.edu.cn,

Medical Imaging 2021: Image Processing, edited by Ivana Išgum, Bennett A. Landman, Proc. of SPIE Vol. 11596, 1159621 · © 2021 SPIE · CCC code: 1605-7422/21/$21 · doi: 10.1117/12.2580665

As shown in Figure 1, retinal detachment is characterized by a cavity area between the retinal neuroepithelium layer (RNE) and retinal pigment epithelium (RPE). retinoschisis is characterized by the RNE splitting into multiple layers, usually with multiple columnar structures between the layers. The main challenges of retinal detachment and retinoschisis are the different sizes, large space span and the subtle differences between the two types of pathological regions.

To solve the above problems, we propose an end-to-end CFCNN deep neural network with a cross-fusion global feature module integrated into a U-shaped network. To deal with small target areas and to preserve detailed structures, we replace some downsampling operations of the U-net with dilated convolutions to retain sufficient spatial resolution. To deal with target areas with a large spatial span and to extract more contexts, we propose a cross-fusion global feature module, which can integrate global feature information into all locations of the feature map and increase the receptive field of the network.



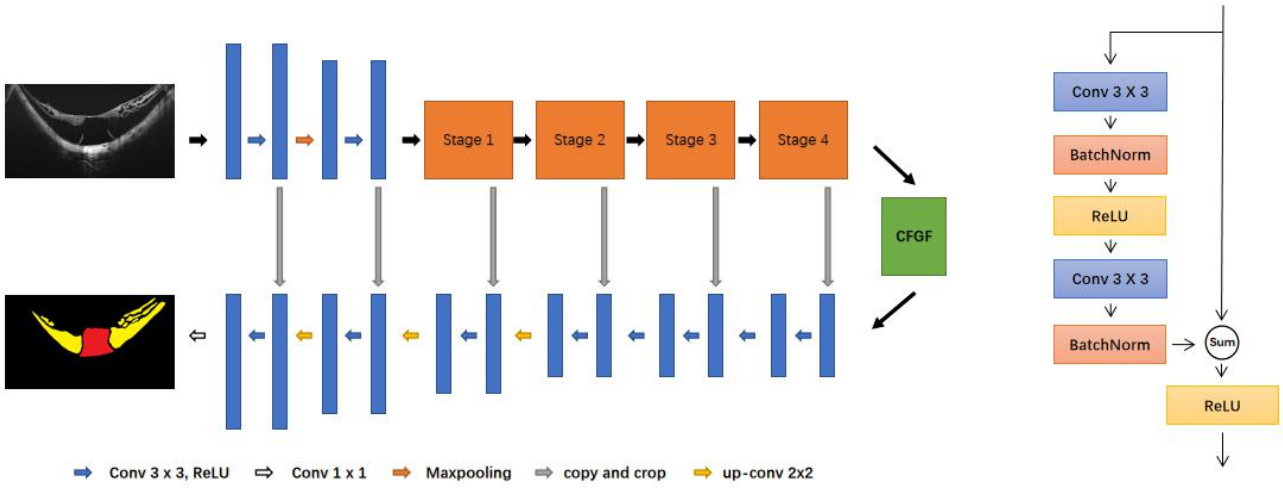
(a) (b)

Fig.1 (a) Retinal detachment and retinoschisis on OCT images (the red arrow indicate RD, and the yellow arrows indicate RS). (b) The ground truth of segmentation (the red region represents RD, and the yellow regions represent RS).

2. METHODS

We introduce the proposed method in the following three parts: the structure of the proposed deep network, cross-fusion global feature module (CFGF) and the loss function.

2.1 Network structure of CFCNN



(a) Whole Network (b) Residual Block

Fig.2 An overview of the CFCNN. (a) Network Architecture. (b) Residual Block.

One of the most classic image segmentation deep neural networks is the U-Net proposed by Olaf Ronneberger et al. in 2015 [2] . The network is composed of an encoder and a decoder. As shown in Figure 2, to extract richer semantic information, we use the middle part of Resnet34 [6] with 4 stages consisting of residual blocks to replace the lower half of the U-net encoder. The four stages are composed of 3, 4, 6 and 3 residual blocks respectively, and the first residual block in each stage adopts the downsampling operation with convolution strides of 2 and the number of channels is doubled. Since RD and RS regions are sometimes small, and fine structures such as the columnar structures inside RS can help distinguish the two types, we use dilated convolution instead of the downsampling in the last two stages to reduce the loss of detailed features. The CFGF module is added to the bottom of the encoder.

2.2 Cross-fusion global feature module (CFGF)

For large segmentation targets, the larger the network receptive field, the more information will be obtained, and the better the segmentation results will be achieved. And for a segmentation target with a large spatial span, it is crucial that the network can effectively capture long-range contextual information. Based on the above considerations, we creatively proposed the cross-fusion global feature module (CFGF).

As shown in Figure 3, first, let $X \in R^{C\times H\times W}$ be a three-dimensional input tensor, where C, H and W are the channel number, height and width, respectively. After fusing the features in the horizontal and vertical spatial dimensions respectively, $y_1 \in R^{C\times H\times 1}$ and $y_2 \in R^{C\times 1\times W}$ are obtained. It is worth noting that the weighted fusion coefficients are obtained by network learning. Since this fusion is along a spatial dimension, it can capture the long-range relations of isolated regions, which helps to capture global context information and prevent extraneous regions from hindering label prediction. Then, multiply y_1 and y_2 and pass the Sigmoid activation function, and multiply the result with X and add X to obtain $Y \in R^{C\times H\times W}$, so that a certain position of *Y* will contain merged feature information from the horizontal and vertical dimensions of the position. Formally, the cross-fusion operation can be written as:

$$y_{ij} = x_{ij} \times \mathrm{Sigmoid}\left(\sum_{m}^{W} a_m x_{im} \sum_{n}^{H} b_n x_{nj} \right) + x_{ij} \tag{1}$$

where a_m and b_n are the weighted fusion coefficients in the horizontal and vertical dimensions, respectively. Finally, after more than one operation, each position of the feature matrix will contain the global feature information relative to that position.

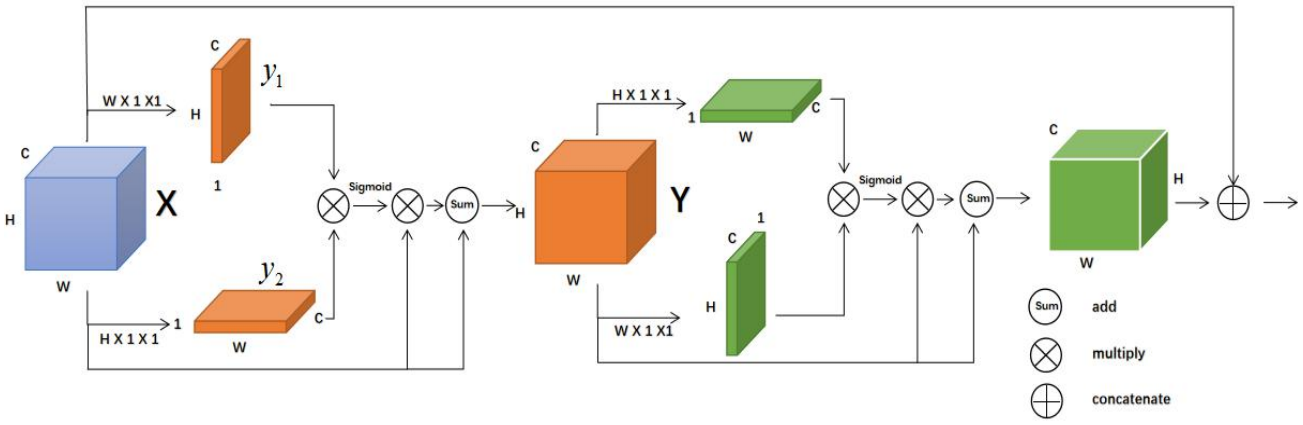


Fig 3. Cross Fusion Global Feature Module (CFGF)

The cross-fusion global feature module we proposed can effectively capture long-range contextual information while increasing the receptive field of the network, which is crucial for the goal of segmenting retinal detachment and retinoschisis with a large space span.

2.3 Loss function

For the segmentation of retinal detachment and retinoschisis, there will be problems of class imbalance and small segmentation targets. Based on the above problems, we use the Dice loss function [7], which has good performances in such scenario. The Dice loss function is obtained by calculating the overlapping area of the ground truth and the prediction map divided by the sum of their areas. Let x represent the pixel value in the ground truth, y represents the value predicted by the network, n represents the total number of pixels in the image, and i represents the *i-th* pixel. The loss function can be defined as follows:

$$loss = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{2x_i y_i}{x_i^2 + y_i^2} \tag{2}$$

3. RESULTS

3.1 Datasets

The datasets used in this paper are two-dimensional OCT images acquired by Topcon swept source OCT with 12-line scanning mode at Shanghai General Hospital, from patients with high myopia. The experimental dataset was drawn from a total of 540 OCT B-scans from 45 eyes, with 12 OCT B-scans per eye. The original image size was 1024×992 corresponding to 6×2.6 mm^2 (width $\times$ height). The ground truth is achieved by manual annotation under the supervision of a senior physician. A total of 420 labeled images with retinal detachment and retinoschisis from 35 eyes were used as the training data set, while the remaining 120 images from 10 eyes were used as the test data set. To be fair, both eyes of the same patient were present only in the training data set or the test data set.

3.2 Parameter settings

Our proposed network was trained in an end-to-end way. Considering the GPU memory cost and training time cost, we resized the image and ground truth to 512×256 before input. We used the Adam algorithm with an initial learning rate of 0.0001 to optimize the weight of the network in the training process, and our model was trained for 150 epochs. When the total number of epochs reached 40 and 100, the learning rate was multiplied by 0.1. In our experiment, the batch size was set to 4. The models were implemented based on the Pytorch framework on NVIDIA GPU 1660TI with 6GB memory.

3.3 Evaluation metrics

To evaluate the segmentation results of the network for RD and RS, we selected 4 evaluation indicators: Dice, Intersection over Union (IoU), Sensitivity (Sen) and Specificity (Spe). Their definitions are as follows:

$$Dice = \frac{2 \times TP}{(TP + FP) + (TP + FN)} \tag{3}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{4}$$

$$Sen = \frac{TP}{TP + FN} \tag{5}$$

$$Spe = \frac{TN}{FP + TN} \tag{6}$$

Where TP is the number of pixels that are true positive (the target pixels are correctly divided into target pixels), TN is the number of pixels that are true negative (the background pixels are correctly divided into background pixels), and FP is the false positive (the background pixels are incorrectly divided into target pixels), FN is the number of false negatives (the target pixels are incorrectly divided into background pixels).

3.4 Experimental results and analysis

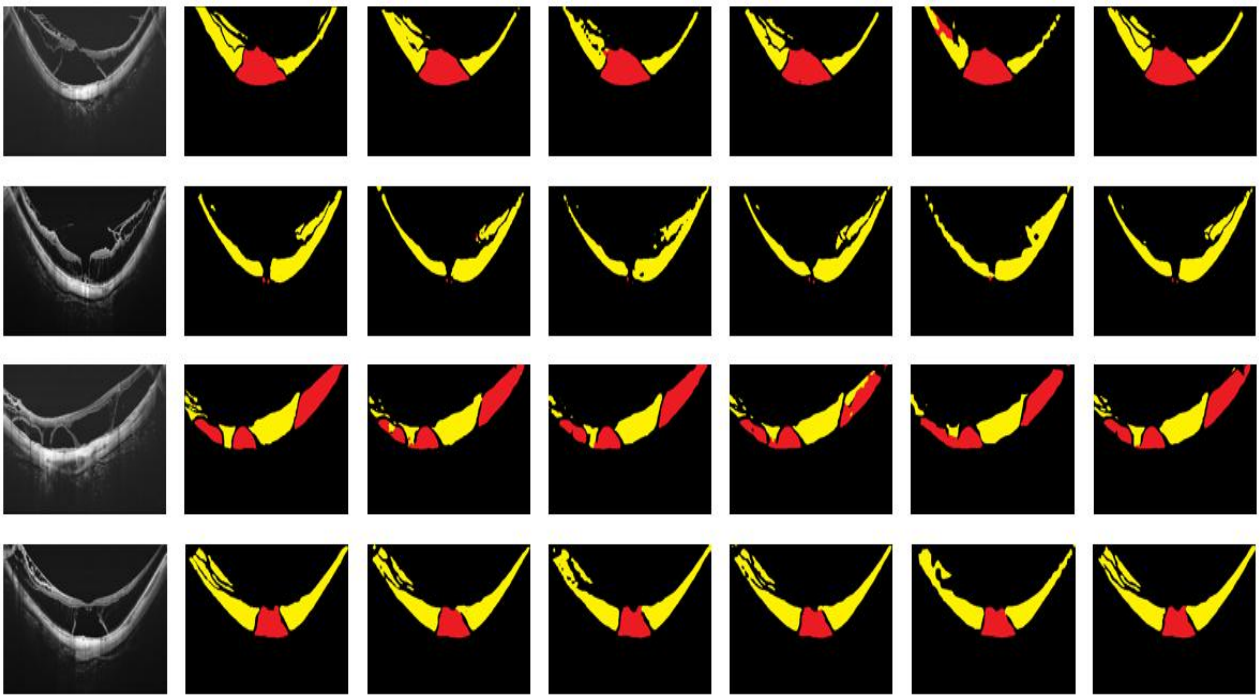
We compared the original U-net network with our proposed baseline without the cross-fusion global feature module. From Table 1, we can see that our baseline achieved better performance compared to the original U-net because it can obtain richer semantic information while reducing the loss of detailed features. After adding the cross-fusion global feature module to the baseline, we can see that the results have been further improved. The mean Dice similarity coefficient of retinal detachment and retinoschisis reached 94.33% and 90.29%, respectively. Finally, we compare the proposed CFCNN with some existing advanced segmentation networks CE-Net, PSP-Net [8] , DeeplabV3 [9] and CPF-Net. It can be seen from Table 1 that our proposed segmentation network is better than these advanced segmentation networks in most indicators.

Table 1 the mean Dice, IoU, Sen and Spe compared with different methods

		Dice	IoU	Sen	Spe
Retinoschisis	U-net	88.59	80.69	**91.23**	99.28
	Baseline	89.53	82.60	89.98	99.47
	CE-Net	88.84	81.03	88.15	**99.50**
	PSP-Net	86.40	77.44	86.56	99.33
	DeeplabV3	79.05	67.62	75.73	99.27
	CPF-Net	87.15	78.79	87.90	99.38
	CFCNN	**90.29**	**83.24**	90.73	99.47
Retinal detachment	U-net	88.03	84.79	93.62	99.85
	Baseline	93.27	90.33	93.31	**99.93**
	CE-Net	90.30	87.21	93.29	99.89
	PSP-Net	92.95	89.37	93.02	99.88
	DeeplabV3	85.96	81.20	92.82	99.74
	CPF-Net	89.38	85.80	92.42	99.90
	CFCNN	**94.33**	**91.41**	**93.94**	99.90

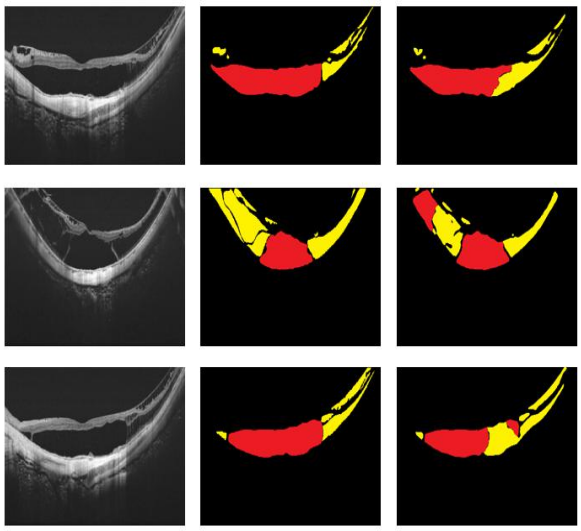
Figure 4 visualizes the segmentation results of various methods. The segmentation method we proposed can obtain good results for both small and large targets. It can be seen from the figure that the proposed network is more accurate than other networks in segmenting the target boundary region, as the cross-fusion global feature module improves the receptive field of the whole network. Compared with other networks, the segmentation of small targets is more accurate, because we use dilated convolution instead of downsampling to reduce the number of downsampling and retain more detailed information.

Figure 5 visualizes some cases of segmentation failure. It can be seen from the figure that the segmentation results are not good for cases where the boundary between retinal detachment and retinoschisis is blurred. Particularly, as shown in the second row, when some retinal layers become very fuzzy and even break, it is difficult for the network to obtain the information between the retinal layers, leading to segmentation errors. In the following work, we will try to enhance the network's attention to the target edge through edge loss, and add retinal layer information to alleviate the incorrect segmentation problem caused by boundary ambiguity.



(a) (b) (c) (d) (e) (f) (g)

Fig 4. Segmentation results for some B-scans (the red region represents RD, and the yellow regions represent RS). (a) Raw image. (b) Ground truth. (c) CE-Net. (d) PSP-Net. (e) CPF-Net. (f) DeeplabV3. (g) CFCNN.



(a) (b) (c)

Fig 5. Segmentation failure cases for some B-scans (the red region represents RD, and the yellow regions represent RS). (a) Raw image. (b) Ground truth. (c) CFCNN.

4. CONCLUSIONS

In this paper, we propose a new cross-fusion global feature module (CFGF) to capture long-range contextual information and increase the receptive field of the network. To adapt to the segmentation of retinal detachment and retinoschisis on OCT images, we design a new CFCNN network which is a U-shaped network that incorporates the cross-fusion global

feature module (CFGF). The network also integrates residual blocks into the encoder to enhance the extraction of semantic information, and the number of downsampling is reduced to obtain higher feature resolution. Experiments have proved that the CFCNN achieved good performance on the segmentation of retinal detachment and retinoschisis. The proposed model and network can be also useful for other medical imaging segmentation tasks with varied target sizes.

5. ACKNOWLEDGEMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0701700, in part by the National Nature Science Foundation of China under Grant 61622114, 61971298, and 61771326.

REFERENCES

[1] M. Fujimoto, M. Hangai et al., “Features Associated With Foveal Retinal Detachment in Myopic Macular Retinoschisis[J],” American Journal of Ophthalmology, 150(6), 2010.
[2] O. Ronneberger, P. Fischer et al., “U-net: Convolutional networks for biomedical image segmentation,” MICCAI.016, 234-241,2015.
[3] J. Long, E. Shelhamer et al., "Fully Convolutional Networks for Semantic Segmentation," IEEE Trans Pattern Anal Mach Intell, doi: 10.1109/TPAMI.2016.2572683. 2017
[4] Z. Gu, J. Cheng et al., “CE-Net: Context Encoder Network for 2D Medical Image Segmentation,” in IEEE Transactions on Medical Imaging, vol. 38, no. 10, pp. 2281-2292, Oct. 2019, doi:10.11 09/TMI.2903562.2019.
[5] S. Feng, H. Zhao et al., “CPFNet: Context Pyramid Fusion Network for Medical Image Segmentation,” in IEEE Transactions on Medical Imaging, doi: 10.1109/TMI.2983721.2020.
[6] K. He, X. Zhang et al., “Deep Residual Learning for Image Recognition,” arXiv:1512.03385v1 [cs.CV] 2015.
[7] F. Milletari, N. Navab et al, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *3D Vision (3DV), 2016 Fourth International Conference on.* IEEE, 2016, pp. 565–571.
[8] H. Zhao, J. Shi et al., “Pyramid Scene Parsing Network,” IEEE Conference on Computer Vision and Pattern Recongnition 2017, pp. 2881-2890.
[9] L. Chen, G. Papandreou et al., “Rethinking Atrous Convolution for Semantic Image Segmentation,” arXiv: 1706.05587v3 [cs.CV] 2017.