EXPLORING CATEGORY CONSISTENCY FOR WEAKLY SUPERVISED SEMANTIC SEGMENTATION

*Zhaozhi Xie, Hongtao Lu**

Department of computer science and engineering, Shanghai Jiao Tong University, China

ABSTRACT

Self-supervised framework has been widely used in weakly supervised semantic segmentation. Generating a reliable and detailed pseudo mask label is the main challenge for improving the quality of predicted mask. In this paper, we propose Category Consistency Mask Refinement (CCMR) to explore the category consistency cued with the input image, and inject such information to mask refinement, guaranteeing the completeness of the refined mask. Moreover, we exploit Selective Weighted Pooling (SWP) to restrict the backward propagation of background, limiting the update of the background. Experimental results demonstrate that our methods can boost the performance on the PASCAL VOC 2012 segmentation benchmark, outperforming the state-of-the-art weakly supervised semantic segmentation methods.

Index Terms— Self-supervised, pseudo label, category consistency, mask refinement

1. INTRODUCTION

Weakly supervised semantic segmentation (WSSS) is defined as utilizing weak annotation such as image label, bounding box, saliency maps, etc, to achieve pixel-label classification, which broadens the usage restrictions and application scenarios of classic semantic segmentation. In recent years, WSSS has developed rapidly and some work [1] has achieved remarkable performance on par or outperforming the fully supervised methods. Since the image label contains a lack of structural and position information, many approaches prefer to leverage stronger annotation especially saliency maps for WSSS. Nevertheless, there are also some works focusing on image-label WSSS, whereas the results are a far cry from fully supervised method.

Class Activation Map (CAM) [2] successfully excavates the pixel-label semantic information from the trained classification network, which establishes a strong baseline for image-label WSSS. Nevertheless, CAM is either diffused or incomplete compared with pixel-label annotation. Obviously, this is because of the absence of pixel-label annotation in the training process. Moreover, with deep consideration of this issue, the training way of classification is contradictory with yielding a perfect segmentation mask. When training an image, since the label is shared for both foreground and background, the backward propagation will raise the value of the whole image without ignoring the background area. Though the value of background may change slowly, the trend is undesirable for a segmentation mask generation. Consequently, attaching pseudo mask to classification network can prevent such feedback of background, and hence improve the quality of CAM.

In the past few years, many researchers focus on generating the fine pseudo label as segmentation supervision. Araslanov et al. [3] propose pixel-adaptive mask refinement (PAMR) to refine the predicted mask cued with the RGB value of the input image. Lee et al. [4] propose a bounding box attribution map (BBAM), which can draw on the rich semantics learned by an object detector to instruct generating pseudo label. Xu et al. [5] perform affinity learning across saliency detection and semantic segmentation, obtaining the pseudo label according to the affinity map.

In this paper, we propose an effective way called Category Consistency Mask Refinement (CCMR) to generate pseudo label as segmentation supervision. To mine the misclassified area of predicted mask, we explore the category consistency to group together pixels of the same class. Inspired by [6], we adopt EM algorithm cued with reference feature to refine the predicted mask. To avoid that if the values of the same class are of large variation, leading to a fuzzy class prototype, the propagation of neighbors [3] is also taken into consideration. Specifically, we first generate the class prototypes by multiplying the predicted mask and reference feature. In practice, we utilize the RGB space of the input image as the reference feature. Then we will calculate the similarity between each position and its neighbors, together with the class prototypes. With the expanded affinity map, we refine the predicted mask iteratively to obtain the final pseudo label. Additionally, we improve the design of nGWP [3] by exploiting a background-shielding pooling way called Selective Weighted Pooling (SWP). As analyzed above, the backward propagation of background in the classification network is harmful for CAM generation. To this end, we leverage the pseudo la-

∗Corresponding author: Hongtao Lu, also with MOE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China, htlu@sjtu.edu.cn. This paper is supported by NSFC (62176155), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), NSFC (61876109, 61772330) and the Interdisciplinary Program of Shanghai Jiao Tong University (no. YG2019QNA09)

bel to select foreground area and do the Weighted Pooling as the score of class.

2. MODEL

In this section, we will present the overall architecture and principal components of our model. We first provide an overview of our WSSS framework in Sec. 2.1, exploring the self-supervised model for semantic segmentation. Next, we explain Category Consistency Mask Refinement (CCMR) in Sec. 2.2. In Sec. 2.3, we develop Selective Weighted Pooling (SWP) for class score calculation, limiting the response of background area.

2.1. Network Architecture

As illustrated in Fig. 1, we design a simple self-supervised framework for WSSS. The backbone is VGG16 with output stride of 8. A segmentation head will be attached to the last layer of VGG16 and the predicted head comprises three convolutions. The first two convolutions are designed for dimensionally reduction and information exchanging, each followed by batch normalization and ReLU activation. The last convolution, with the kernel size of 1×1, will output the score maps S. Next we attach a background map of a constant value to the score maps and do the softmax operation to obtain the predicted mask M. Then we utilize SWP to calculate the score values of each class, which will participate in the classification loss calculation. Additionally, CCMR is applied to M for generating pseudo label M', which will guide the network for a better mask generation.

2.2. Category Consistency Mask Refinement

Category consistency is an important character of feature representation. Many previous works [6, 7] had proved that keeping category consistency can yield a compact representation. In this paper, we explore the category consistency information from the input image, and transfer the signal to the predicted mask. To this end, we design Category Consistency Mask Refinement (CCMR) to generate pseudo mask label, which is illustrated in Fig. 2. The calculation can be decomposed into two parts: prototypes generation and pixel refinement.

2.2.1. *Prototypes generation*

Given the predicted mask $M \in \mathcal{R}^{K*H*W}$, we firstly reshape M to $\mathcal{R}^{K*N}$. Note that K is the number of class and N is the number of pixels in M. To obtain the class prototypes, we need to make sure that the sum of the mask across each class is equal to 1. Hence we do the normalization as follows:

$$\bar{m}_{k,n} = \frac{m_{k,n}}{\epsilon + \sum_{i=1}^{N} m_{k,i}} \tag{1}$$

When obtaining $\bar{M}$, we can aggregate the features of each pixel to form the class prototypes. For pixel i, $\bar{m}_{k,i}$ indicates

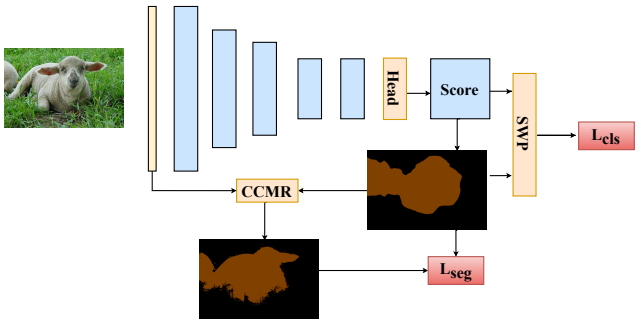

Fig. 1. Overall architecture of our network. We use VGG16 backbone to extract deep feature, followed by a predicted head. CCMR provides the pseudo segmentation label and SWP calculates the score values of each class.

how possible i belongs to class k among all the pixels. So the prototype of class k is given by:

$$p_k = \frac{\sum_{i=1}^{N} \bar{m}_{k,i} f_i}{\epsilon + \sum_{i=1}^{N} \bar{m}_{k,i}} \tag{2}$$

where $k \in \mathcal{P}$, $\mathcal{P}$ is the set of class that exist in the corresponding image label set. f_i is the feature of pixel i. In practice, we use RGB space vector of input image as the feature f_i.

2.2.2. *Pixel refinement*

In this part, we will display how we use the prototype to refine the predicted mask. Since we use the color values as the only cue for prototype generation and similarity calculation, if the color values of the same class are of large variation, the corresponding prototype is useless for mask refinement. To this end, we follow the idea of [3], taking the neighbors of each pixel into consideration. Then we can obtain two affinity maps A^p and A^l as follows:

$$a^p_{n,k} = \frac{\mathcal{K}(f_n, p_k)}{\sum_{i \in \mathcal{P}} \mathcal{K}(f_n, p_i) + \sum_{j \in \mathcal{N}_n} \mathcal{K}(f_n, f_j)} \tag{3}$$

$$a^l_{n,l} = \frac{\mathcal{K}(f_n, f_l)}{\sum_{i \in \mathcal{P}} \mathcal{K}(f_n, p_i) + \sum_{j \in \mathcal{N}_n} \mathcal{K}(f_n, f_j)} \tag{4}$$

where $a^p_{n,k}$ indicates the affinity value between the feature of pixel n f_n and the prototype of class k p_k. $a^p_{n,k}$ indicates the affinity value between the feature of pixel n f_n and the feature of adjacent pixel l f_l. $l \in \mathcal{N}_n$ and $\mathcal{N}_n$ is the neighbours of n. The kernel $\mathcal{K}$ has many choices and we take the exponential negative Euclidean distance in our paper, which is defined as:

$$\mathcal{K}(a, b) = \exp(-\|a - b\|) \tag{5}$$

Finally, we can obtain the refined mask M' according to the original mask and two affinity maps. Formally, M' is defined as:

$$m'_{k,n} = a^p_{n,k} + \sum_{j \in \mathcal{N}_n} m_{k,j} a^l_{n,j} \tag{6}$$

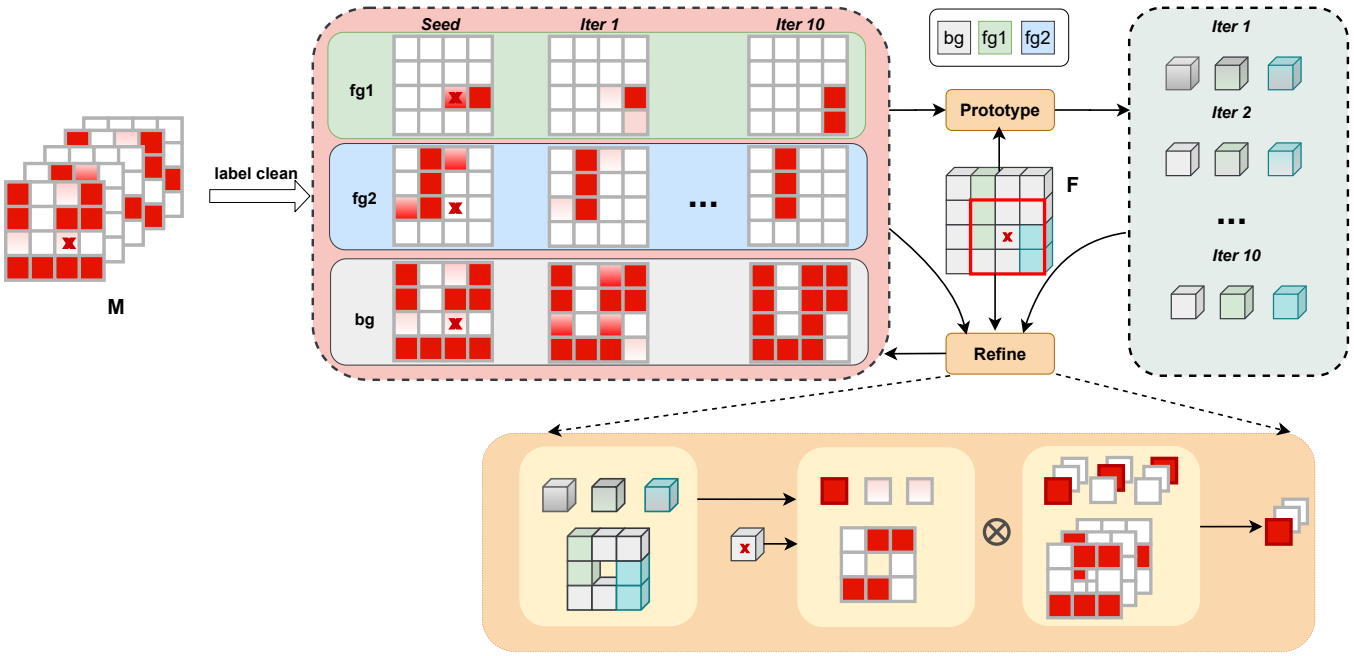

Fig. 2. The structure of CCMR. Given the predicted mask M, CCMR will firstly generate the class prototypes, then do the pixel refinement to obtain pseudo mask label. The box below illustrates the refined process of a certain position.

Note that in Eq. 6, only one prototype is used for reconstruction. This is because each prototype is the representation of each class, and hence each prototype only belongs to the corresponding class, with no possibility of belonging to other class. That is to say, the mask of prototype is equal to a one-hot vector, so we only need to calculate one corresponding prototype for each class.

Now we finish one update of predicted mask, whereas we know that EM algorithm executes expectation step and maximization step alternately. Here expectation step is pixel refinement and maximization step is prototypes generation. Thus we will run these two steps iteratively 10 times to obtain the final refined mask as the pseudo mask label.

2.3. Selective Weighted Pooling

In Section 1, we analyze that the score map values of the target class will rise with backward propagation. However, we don't want the score map values of background to become too big, leading to a bad predicted mask. To this end, we propose Selective Weighted Pooling (SWP) to restrict the propagation of background. We firstly multiply the refined mask to the predicted mask in the following way:

$$m_{k,n} = m_{k,n} \cdot \mathbb{1}\left[m'_{k,n} > \tau\right], z_k = 1 \tag{7}$$

where τ is the threshold measuring the degree of background-neglect and is fixed by 0.6 in the paper. Additionally, to avoid the missing of small objects, we only consider the objects of large scale. Then we follow [3] to calculate the score values of each class, which is given by:

$$y'_k = \frac{\sum_{i=1}^{N} m_{k,i} s_{k,i}}{\epsilon + \sum_{j=1}^{N} m_{k,j}} \tag{8}$$

where S are the score maps and Y are score values of each class.

3. EXPERIMENTS

3.1. Experimental Setup

3.1.1. Dataset

We train and evaluate our models using PASCAL VOC 2012 segmentation benchmark [15]. One background category and 20 object categories are annotated in this dataset. Following the standard practice [11, 16, 10], we used the augmented PASCAL VOC training data provided by Hariharan et al. [17]. In total, we use 10 582 images with image-level annotation for training and 1449 images for validation.

3.1.2. Implementation Details

We adopt VGG16 backbone network for feature extraction, initialized with ImageNet [18] pre-training. Our network is trained with SGD using a learning rate of 0.01 (0.0001 for VGG16 backbone) and a weight decay of 0.0005. Training parameters in the learning phase are: batch size = 16, max

Model	background	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	Mean IoU
MIL+seg [8]	79.6	50.2	21.6	40.9	34.9	40.5	45.9	51.5	60.6	12.6	51.2	11.6	56.8	52.9	44.8	42.7	31.2	55.4	21.5	38.8	36.9	42.0
SEC [9]	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5	28.3	65.8	57.8	62.3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
AdvErasing [10]	83.4	71.1	30.5	72.9	41.6	55.9	63.1	60.2	74.0	18.0	66.5	32.4	71.7	56.3	64.8	52.4	37.4	69.1	31.4	58.9	43.9	55.0
PSA [11]	88.2	68.2	30.6	81.1	49.6	61.0	77.8	66.1	75.1	29.0	66.0	40.2	80.4	62.0	70.4	73.7	42.5	70.7	42.6	68.1	51.6	61.7
Single [3]	88.7	70.4	35.1	75.7	51.9	65.8	71.9	64.2	81.1	30.8	73.3	28.1	81.6	69.1	62.6	74.8	48.6	71.0	40.1	68.5	**64.3**	62.7
SEAM [12]	88.8	68.5	33.3	**85.7**	40.4	67.3	78.9	76.3	81.9	29.1	75.5	**48.1**	79.9	73.8	71.4	**75.2**	48.9	79.8	40.9	58.2	53.0	64.5
SSDD [13]	89.0	62.5	28.9	83.7	52.9	59.5	77.6	73.7	**87.0**	34.0	**83.7**	47.6	84.1	**77.0**	73.9	69.6	29.8	**84.0**	**43.2**	68.0	53.4	64.9
BES[14]	88.9	74.1	29.8	81.3	53.3	**69.9**	**89.4**	**79.8**	84.2	27.9	76.9	46.6	78.8	75.9	72.2	70.4	**50.8**	79.4	39.9	65.3	44.8	65.7
Ours+crf	89.2	75.5	34.4	77.3	53.7	69.1	70.4	68.4	86.0	31.2	76.5	25.9	83.3	71.1	68.2	69.1	49.8	75.1	38.4	69.4	56.4	63.7
Ours+DeepLabv3+	**89.9**	**79.1**	**38.7**	78.3	**59.0**	59.3	78.4	74.6	85.9	**34.6**	72.2	27.6	**84.5**	74.3	**74.2**	72.7	47.5	82.1	37.4	**76.3**	61.3	**66.1**

Table 1. Semantic segmentation performance on PASCAL VOC 2012 *val set*, with image-level supervision.

Table 2. Ablation study on CCMR and SWP

Method		mIoU
CCMR	SWP	**58.6**
	nGWP [3]	56.2
	Average Pooling	53.7
SWP	CCMR	**58.6**
	PAMR [3]	57.3
	CCMR w/o $\mathcal{N}$	46.2

epoch = 30. In the first 5 epochs, the network is trained with classification loss, and we add the segmentation loss in the final 25 epochs. We also use some standard data augmentation techniques to train our network following the setup of [11], including random rescaling, horizontal flipping, color jittering and random crops of size 448×448.

3.2. Ablation Study

This part will demonstrate the impact of CCMR and SWP based on the corresponding experiments. Tab. 2 shows the ablation study of these two components, which are the results of PASCAL VOC 2021 val set by using the image-level labels for mask cleaning and dense CRF for postprocessing. We compare SWP with other pooling methods, including nGWP [11] and average pooling. Obviously, SWP performs the best among all the pooling methods. This illustrates that the restriction on background is useful for segmentation training. Additionally, we compare CCMR with other mask refinement methods. Note that PAMR [11] only considers the neighbors while CCMR w/o $\mathcal{N}$ only considers the class prototypes. We find that CCMR achieves the highest mIoU among these methods. Remarkably, if we only use class prototypes, the result will drop sharply. This is because the RGB space is not powerful enough to describe the character of a class. Only if we use the low-level feature as the reference for similarity calculation, CCMR w/o $\mathcal{N}$ can achieve a good result.

3.3. Comparing with the State-of-the-Art

The quantitative results are summarized in Tab 1. For fair comparison, we use the framework of Single [3] and change the network with the proposed CCMR and SWP. Additionally, since we generate the prediction with a single-stage architecture, we further feed the predicted mask processed with CRF to DeepLab v3+ (ResNet-101) [19] as the pseudo label and obtain final output from DeepLab v3+. The final result achieves the mIoU of 66.1%, outperforming the state-of-the-art weakly supervised semantic segmentation methods with only image label.

4. CONCLUSION

In this work, we propose Category Consistency Mask Refinement (CCMR) which explores the category consistency for weakly supervised semantic segmentation. Instead of generating a compact feature representation by strengthening the category consistency, we attempt to discover the category consistency from the input image. Then this signal will be used for mask refinement. Apart from class prototypes, we also use the neighbors to refine the predicted mask and generate the pseudo label. Moreover, we research the training process of classification network and find that the background area of a target class map has a trend to rise, leading to a diffused CAM. To this end, we propose Selective Weighted Pooling (SWP) to restrict the backward propagation of background, and hence avoid the update of background area of the target class map. Experimental results demonstrate the effectiveness of both CCMR and SWP, and when applying these two components to a strong baseline, we can achieve the state-of-the-art result among all the weakly supervised semantic segmentation methods with only image label.

5. REFERENCES

[1] Youngmin Oh, Beomjun Kim, and Bumsub Ham, "Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6913–6922.

[2] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[3] Nikita Araslanov and Stefan Roth, "Single-stage semantic segmentation from image labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4253–4262.

[4] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon, "Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2643–2652.

[5] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, Ferdous Sohel, and Dan Xu, "Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6984–6993.

[6] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9167–9176.

[7] Fan Zhang, Yanqin Chen, Zhihang Li, Zhibin Hong, Jingtuo Liu, Feifei Ma, Junyu Han, and Errui Ding, "Acfnet: Attentional class feature network for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6798–6807.

[8] Pedro O Pinheiro and Ronan Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1713–1721.

[9] Alexander Kolesnikov and Christoph H Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *European conference on computer vision*. Springer, 2016, pp. 695–711.

[10] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1568–1576.

[11] Jiwoon Ahn and Suha Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4981–4990.

[12] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12275–12284.

[13] Wataru Shimoda and Keiji Yanai, "Self-supervised difference detection for weakly-supervised semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5208–5217.

[14] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang, "Weakly supervised semantic segmentation with boundary exploration," in *European Conference on Computer Vision*. Springer, 2020, pp. 347–362.

[15] M. Everingham, L. Gool, C. K. Williams, J. Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2009.

[16] Yunchao Wei, H. Xiao, Humphrey Shi, Zequn Jie, Jiashi Feng, and T. Huang, "Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7268–7277, 2018.

[17] Bharath Hariharan, Pablo Arbeláez, Lubomir D. Bourdev, Subhransu Maji, and Jitendra Malik, "Semantic contours from inverse detectors," *2011 International Conference on Computer Vision*, pp. 991–998, 2011.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[19] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.