MUNet: A Multi-scale U-Net Framework for Medical Image Segmentation

1st Wentao Zhang, 2rd Hao Cheng, 3rd Jun Gan
National Key Laboratory for Novel Software Technology
Department of Computer Science and Technology
Nanjing University
Nanjing, China
zhangwent963@gmail.com, chengh@smail.nju.edu.cn, junnor.gan@gmail.com

Abstract—Artificial intelligence has once again become the focus of attention in all fields, of which deep learning has brought a series of changes in the field of computer vision. In this paper, we propose the MUNet model, which is a more general convolutional neural network framework for medical image segmentation. The framework proposed in this paper is essentially a fully convolutional encoder-decoder network based on feature pyramids, in which the encoder and decoder are connected by skip connections. Not only is it suitable for image segmentation, but it can also identify categories of regions of interest. The full convolutional neural network architecture can implement multi-scale image input and prediction. We compared the MUNet and UNet models in cervical lymph node localization and benign and malignant diagnosis in ultrasound images. Experiments show that on our dataset, MUNet with multi-scale segmentation has achieved a Dice score improvement of 4.1% and an AUC score improvement of 8.1% compared to U-Net.

Index Terms—Artificial intelligence, Fully convolutional neural networks, Multi-scale, Cervical lymph nodes

I. Introduction

Convolutional neural networks (CNNs) have superior performance for high-level vision tasks, for example, object detection [1]–[4], and semantic segmentation [5]–[7]. CNNs have been widely used in the medical image diagnosis system. And performs well in segmentation and classification. The classical models for image segmentation are variants of the encoder-decoder architecture like U-Net [8] and fully convolutional network (FCN) [5]. They all have the same architectures: skip connections, which combine deep, semantic, coarse-grained feature maps from the decoder sub-network with shallow, low-level, fine-grained feature maps from the encoder sub-network. In FCN, upsampled feature maps are summed with feature maps skipped from the encoder, while U-Net concatenates them and add convolutions and non-linearities between each upsampling step. It can be said that image segmentation in natural images has reached a satisfactory level. However, the segmentation of lesions or abnormalities in medical images requires higher accuracy than in natural images. The marginal details and small objects in the medical image need to be segmented more accurately than in natural images.

The feature pyramid is the basic component of multi-scale object discovery systems, which is used to find marginal details and small objects. A top-down architecture with horizontal connections is proposed for constructing high-level semantic feature maps of various scales. This architecture is called feature pyramid networks (FPN) [9]. It can be trained end-to-end with arbitrary sizes and is used consistently at inference time. We present MUNet, a new framework for medical image segmentation based on the feature pyramid to address the need for more accurate segmentation in medical images. The hidden hypothesis behind our framework is that the model can more effectively capture (1) large object when the encoder sub-network layer deepens and receptive field increases and (2) fine-grained details and small objects when high-resolution feature maps from the encoder sub-network are gradually enriched before fusion with the corresponding semantically rich feature maps from the decoder sub-network. Our experiments demonstrate that the architecture is effective. Our main contributions lies in:

- We utilize a fully convolutional network framework that accepts inputs of arbitrary size and produces outputs of the corresponding size through effective learning and inference.
- The backbone of the framework can be replaced arbitrarily to meet the needs of different scenarios. For example, high efficiency, high accuracy or small network size.
- It can be applied to segment objects of different scales. The architecture can more effectively capture large objects, small objects and fine-grained details.

The rest of this paper is organized as follows. We first review the related work on deep classification networks and the latest methods for semantic segmentation using convolutional networks. The next section will explain the design of MUNet and introduce the architecture of classification subnets and multi-scale segmentation subnets. Experimental settings and results will be given, and the architecture is proven to be effective. At the same time, we have listed some useful techniques for classification and segmentation. Finally, we summarize this paper and

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

suggest possible future improvements.

II. Related Work

In recent years, deep neural convolutional networks have outperformed the state of the art in many visual recognition tasks such as image classification, semantic segmentation, object detection, posture recognition, inpainting, style transfer, and even image compression. Public datasets such as ImageNet, COCO, Pascal VOC made great contributions to the development of computer vision.

Image classification is a core problem in computer vision. LeNet [10] is the first successful application of CNN that used to read zip codes, digits, etc. The first work that popularized CNN in computer vision is AlexNet [11], which is the winner of ImageNet ILSVRC 2012. AlexNet is based on the LeNet, but AlexNet is deeper, larger scale, and convolution layers stacked on top of each other. The winners of ImageNet ILSVRC 2013 and ImageNet ILSVRC 2014 are ZF-Net [12], GoogLeNet [13], respectively. The former is a modified version on AlexNet by adjusting the hyperparameters, especially by extending the size of the convolution layers in the middle and the latter introduced an Inception Module that dramatically reduced the number of parameters in the network. After that the followup versions of GoogLeNet have been proposed, Inception-V2 [14], Inception-V3 [15], Inception-V4 [16] and InceptionResNet-V2 [16]. In the same challenge, the runner-up is the VGGNet [17]. It is found that blindly adding layers after the depth of the CNN networks reached a certain depth could not bring further improvement of classification performance but would lead to slower network convergence and worse classification accuracy of the test dataset. ResNet [18] designed by Kaiming He et al. solved the degradation problem and avoided the gradient explosion. Different from the characteristic of the InceptionNet family that extends the width of the network, and different from the ResNet family that increases the depth, DenseNet [19] emphasized the features. The special convolution block that called dense block is designed. It enhanced the transference of features and alleviated the vanishing-gradient.

Semantic segmentation is understanding an image at the pixel level, that is, we want to assign each pixel in the image an object category. Before deep learning took over computer vision, TextonForest [20] and Random Forest classifiers [21] for semantic segmentation are popular. A kind of initial deep learning approaches for semantic segmentation is patch classification [22], which predicts its classes separately through a patch of the image of each pixel. In 2014, the popularized CNN architecture for dense predictions without any fully connected layers that called Fully Convolutional Networks(FCN) [5] is proposed by Long et al. This allows segmentation maps to be generated for an image of any size and is also much faster compared to the patch classification approach. In addition to fully connected layers, the pooling layer is one of the main problems of using CNN for segmentation. Pooling layers can extract context information while partial location information is lost. There two kinds of architectures are proposed to discard the information. One is encoder-decoder architecture, the encoder reduces the spatial dimension and extracts features by pooling layers, the decoder gradually recovers the spatial dimension and object details. U-Net [8] is the most popular frame of this way. The second architecture is applying dilated convolutions or atrous convolutions instead of pooling layers. FCN [5] is the end-to-end convolutional network for semantic segmentation, upsamples feature maps with deconvolutional layers, and introduces skip connections to improve over the coarseness of upsampling. SegNet [6] does not copy encoder features like FCN, while the indices from max-pooling are copied, it makes more memory efficient and segmentation resolution better than FCN. Dilated Convolutions were used in the paper [23], and achieved the multi-scale aggregation. DeepLab-v1 [24] and DeepLab-v2 [25] applied atrous convolutions or dilated convolutions, proposed atrous spatial pyramid pooling (ASPP), and finally through fully connected CRF to predict structures. The improved version is DeepLab-v3, which improves the atrous spatial pyramid pooling and employs atrous convolutions in cascade. RefineNet [26] was constructed by an encoder and decoder, and both components were designed based on the resnet blocks.

The state-of-the-art models for image segmentation are variants of the encoder-decoder architecture like U-Net and fully convolutional networks. They all have the same architectures: skip connections, which combine deep, semantic, coarse-grained feature maps from the decoder sub-network with shallow, low-level, fine-grained feature maps from the encoder sub-network. Objects with multi-scale are segmented by the same decoder sub-network. The order of feature extraction in different depth of encoder sub-network is different. Shallow network layer tends to focus on lower-level features, which means small objects can be more effectively captured. The deep network layer more focuses on higher-level features, and large objects can be easier to be captured correspondingly. Therefore, it is inappropriate for a decoder layer to segment all sizes of objects. We propose a new framework MUNet, which predicts multi-scale objects through different network layers.

III. Proposed Architectures

Fig. 1 shows an overview of the suggested architecture. We input an image of arbitrary size. Next, the input image features are extracted by the backbone, which is the encoder sub-network. The Backbone consists of 5 convolution blocks (L_{1-5}), the output feature maps are defined as (C_{1-5}). We assumed that the size of input image is $256 \times 256 \times 3$, so the size of C_1 is $128 \times 128 \times 64$, C_2 is

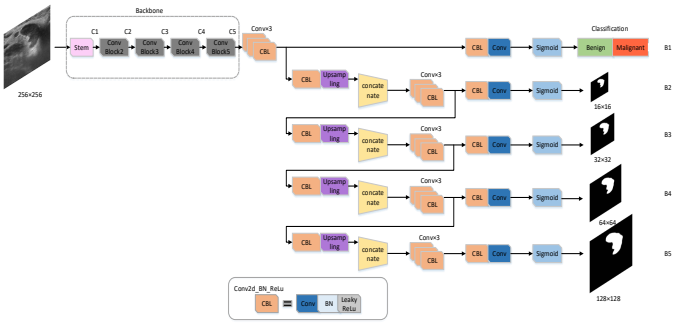

Fig. 1. MUNet for image segmentation. At first, We input an image of arbitrary size. Next, the input image features are extracted by the classification network as the backbone, which is the encoder sub-network. The Backbone consists of 5 convolution blocks (L_{1-5}), the output feature maps are defined as (C_{1-5}). Finally, a network branch is absorbed in predicting the categories of an input image, and four network branches are applied to generate segmentation masks with different sizes. Segmentation mask branches consist of feature maps, which are upsampled (C_{1-5}), and then concatenated with (L_{1-5}). For example, the input image size is 256×256.

$64 \times 64 \times 128$, C_3 is $32 \times 32 \times 256$, C_4 is $16 \times 16 \times 512$ and C_5 is $8 \times 8 \times 1024$. We replaced the max-pooling layer by the special convolution layer. These special convolution layers with settings that the size of the kernel is 3×3, padding is 1 and stride is 2, which makes sure that the width and height of feature maps halved and the number of channels doubled as the number of convolution blocks increasing. C_5 is entered into one network branch which is applied to predict the categories of the input image and enter into another network branch for generating segmentation masks. Every step in the decoder sub-network consists of an upsampling of the feature map, followed by a concatenation with the correspondingly C_* from the encoder sub-network, 3 DBL(a convolution layer, followed by a batch normalization layer and a leaky relu) blocks that halves the number of feature channels. For each upsampling stage, we add 1×1 convolution layer before sigmoid function, which is used to cross channel information, reduce dimension, introduce nonlinearity and accelerate the computation. At the last of each upsampling stage, we output the masks with different sizes to achieve the multi-scale segmentation.

A. Generation of The Best Mask

Since multiple masks are generated, we need to decide which mask should be retained as the optimal one. One way to do that is to compare the metrics with ground truth for each mask and select the mask with the highest metric. However, this goes against our original intention, multi-branch can not generate multi-scale segmentation masks. We used a weighted voting method to determine the best mask.

We assumed that the masks generated from B_{2-5} are two-dimensional matrix M_{2-5} with a different number of the number of rows and columns. Then we resized these matrices to the same size with input image through bilinear interpolation, which are defined as $M^{'}_{2-5}$. The areas of $M^{'}_{2-5}$ are A_{2-5}, correspondingly. W and H are the width and height of the input image. The best mask is defined as:

$$\begin{aligned} M &= \sum_{k=2}^{5} \alpha_k M^{'}_k \\ M^{'}_{kij} &\in [0,1], 1 \le i \le W, 1 \le j \le H \end{aligned} \tag{1}$$

The $\alpha_k, k = 2, 3, 4, 5$ is the weighting coefficient, which is defined as:

$$\alpha_k = 1 - \frac{A_k}{WH} \tag{2}$$

That means we pay more attention to small parts and details.

B. Loss Function

Each training input image is labeled with a ground truth class u and a ground truth segmentation mask target v. We use a multi-task loss of L on each input image to jointly train for classification and mask segmentation.

$$L = L_{cls} + \lambda L_{mas} \tag{3}$$

in which $L_{cls} = -\alpha(1 - p_u)^{\gamma} \log p_u$ is focal loss [27] for true class u. λ is the coefficient of balance.

The second task loss, L_{mas} is defined over the output of the four segmentation mask branches. We usually use the Dice coefficient to measure the quality of image segmentation, which is a similarity measure related to the Jaccard index. For segmentation output $v^{'}$ and target v, the coefficient is defined as:

$$D(v^{'}, v) = \frac{2|v \cap v^{'}|}{|v| + |v^{'}|} \tag{4}$$

, and the Dice coefficient loss is $1 - D$. Therefore L_{mas} is

$$L_{mas} = \sum_{k=2}^{5} 1 - D(v_k, v) \tag{5}$$

, in which v_k is the resized segmentation mask of the branch B_k.

IV. Experiments

A. Datasets

We evaluated MUNet in ultrasound images of cervical lymph nodes, which come from 3000 patients approximately. The dataset concludes about 4000 benign images and 1000 malignant images with the original size are 700×800. Data augmentation has been carried out to balance the categories of the data. The number of images is about 10 thousand after data augmentation. We split the data into the train set, verification set and test set according to $8 : 1 : 1$.

B. Data Augmentation

Data augmentation technology is used to increase the number of data and balance the categories. The ultrasound images of the cervical lymph node are usually class imbalance and expensive, this paper uses cost-sensitive learning and data augmentation to prevent it as far as possible. Several data augmentation techniques are applied to the original ultrasound image and ground truth correspondingly, for example, rotating an image by an arbitrary amount, fliping the image along its vertical axis or horizontal axis, performing a random elastic gaussian distortion on an image, zooming into an image at a random location within the image, cropping a random area of an image based on the percentage area, skewing an image by tilting by a random amount and shearing the image by a specified number of degrees.

C. Experimental Setup

All training and testing processes were performed on NVIDIA GeForce GTX 1080Ti 11G GPUs. We developed our models in the deep learning framework Keras. On the Ubuntu Linux system equipped with NVIDIA GPUs, training a single model took 4–6 hours depending on the architecture of the networks.

D. Implementation Details

We augment the set of training images by flipping, rotating each with $\pm 5°$, adding Gaussian white noises with variances of 0.001 and 0.01 and some others. The size of the images is about 700×800, in the preprocess of training we will resize them to different size. The optional W, H are selected from $\{256, 384, 512, 640\}$, which are integer multiples of 32. So we have multi-scale image with size $\{256 \times 256, 256 \times 384, 256 \times 512, 256 \times 640, 384 \times 384, 384 \times 512, ...\}$.

We monitored the Dice coefficient and AUC and applied an early-stop mechanism on the validation set. We then train our networks for 100 epochs with the following parameters: Adam optimizer with a learning rate of $1e-3$ and weight decay 0.01, batch size 64. We set focal loss for the classification branch. The total loss function is the weighted sum of the Dice coefficient loss and the focal loss. And the balanced coefficient λ is 0.25.

E. Results and Analysis

According to Table I, that compared the various metrics of MUNet with different backbones and the baseline U-Net on the test datasets. U-Net is the simplest model that we have slightly modified to classification and segmentation at the same time. MUNet-ResNet50 has not only a high Dice coefficient but also a high score of accuracy and AUC, which indicates it can not only extract features from the ultrasound images but also recognize the structure of the ultrasound image efficiently. Obviously, the depth of MUNet-ResNet50 is deeper than U-Net, and MUNet-ResNet50 can joint more resnet blocks

TABLE I
The Various Metrics of MUNet with Different Backbones on The Test Datasets

Models	Dice	Sen.	Spec.	Acc.	AUC
U-Net	0.877	0.910	0.746	0.863	0.860
MUNet-ResNet50	0.914	0.891	0.943	0.923	0.924
MUNet-ResNet152	0.916	0.918	0.942	0.941	0.941
MUNet-Inceptionv3	0.924	0.911	0.944	0.932	0.932
MUNet-Inceptionv4	0.878	0.824	0.897	0.861	0.863
MUNet-InceptionResNetv2	0.910	0.890	0.948	0.914	0.916
MUNet-DenseNet	0.906	0.858	0.905	0.927	0.913

[a]Dice is the Dice coefficient, Sen. is the sensivity, Spec. is the specificity, Acc. is the accuracy and AUC is the area under curve.
[b]Boldface represents the model that achieved the highest score for the indices we focused on in the experiments.

(a) original malignant image (b) predicted mask of the original malignant image (c) masked malignant image

(d) original benign image (e) predicted mask of the original benign image (f) masked benign image

Fig. 2. Predictions of MUNet-Inceptionv3 have been shown in the above figures.

such as MUNet-ResNet152 to improve the ability of extracting features from images. MUNet-DenseNet, MUNet-InceptionResNetv2 and MUNet-Inceptionv4 are average in performance. But dense blocks can be used to extend the depth of the network. And MUNet-InceptionResNetv2 has a better performance than that two models because of the effect of inception-resnet-blocks. MUNet-Inceptionv3 has the best Dice coefficient in these models, the wide Inception Module contributes to extracting location information from images. Different size convolution kernel means the different sizes of the receptive field, and finally splicing means the fusion of different scale location features. The feature map output of each Inception Block indicates that MUNet-Inceptionv3 is more advanced than others in extracting location information.

The backbone network structure is complex and diverse and how to select them in the application. Different application requirements need to select different backbone structures. ResNet is more suitable for applications where there is a strong need for accuracy. The ResNet with fewer layers extracts lower-level features that lead to a loss of precision, but the speed of calculation is faster. On the

contrary, the ResNet with more layers extracts higher-level features so that is easier to determine the lesion area is benign or malignant, but more calculated time and space resources are usually required. Inceptionv3 is more suitable for application where there is a strong need for locating lesion areas, but it is often accompanied by a loss of accuracy. If we want both high accuracy and high location ability, the network that ResNet and Inception can be combined and the features extracted by them, but it will be an extremely complex network structure.

F. Some Tricks That Lead to Increase of Dice Coefficient

As shown in Table II, we used some tricks to improve the Dice coefficient based on the MUNet-ResNet152.

TABLE II
Tricks That Lead to Increase of Dice Coefficient in MUNet

Tricks				
BN and 1×1 conv?	✓	✓	✓	✓
Hard example mining?		✓	✓	✓
Ensamble methods?			✓	✓
Multi-scale?				✓
Dice coefficient	0.914	0.924	0.929	0.938

Multi-scale training and inference: They are the common methods to improve model performance. We first load the network parameters that have been pre-trained in ImageNet. And then we fined tune the network with the different input images size $\{256 \times 256, 256 \times 384, 256 \times 512, 256 \times 640, 384 \times 384, 384 \times 512, ...\}$. When in the inference stage, we resized the test images to a different size, predicted the categories and masks.

Batch normalization and 1×1 convolution layer: Batch normalization is designed to solve the problem that called internal covariate shift [14] and accelerate deep Network training. During training time, a batch normalization layer does the following:

$$\begin{aligned} \mu_B &= \frac{1}{m}\sum_{i=1}^{m} x_i \\ \sigma_B^2 &= \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2 \\ \bar{x_i} &= \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \\ y_i &= \gamma \bar{x_i} + \beta \end{aligned} \tag{6}$$

μ_B is the mean of the mini-batch data and σ_B^2 is the variance. The third formula is used to normalize values. And the two variables γ, β are introduced, one for learning the mean and other for variance. During inference time, the mean and the variance are fixed. They are estimated using the previously calculated means and variances of each training batch. 1×1 convolution layer is used to cross channel information, reduce dimension, introduce nonlinearity and accelerate the computation.

Hard example mining: We labeled train samples as difficult or easy using prediction confidence. And then difficult samples can have more post or pre-process. Focal loss has been used to balance the difficult and easy examples.

Ensemble methods: We have trained multiple stronger models, but how to integrate the results of these models to increase metrics? We assumed that we have trained k different models, which is defined as $(p_{ui}, v_i) = F_i(x), 1 \leq i \leq k$, x is the input image, and p_{ui} is probability value that x belongs to the categories u of model i, v_i is the best mask that model i predicted. We use the linear summation method to combine the results.

$$\begin{aligned} F(x) &= \sum_{i=1}^{k} \lambda_i F_i(x) \\ p_u &= \sum_{i=1}^{k} \lambda_i p_{ui} \\ v &= \sum_{i=1}^{k} \lambda_i v_i \\ \lambda_i &\in (0,1), 1 \leq i \leq k \end{aligned} \tag{7}$$

In order to simplify the process, we did not use ensamble methods that like stacking or blending for classification prediction.

V. Conclusion

In this paper, we proposed a framework called MUNet for medical image segmentation. MUNet is essentially a fully convolutional encoder-decoder network based on the feature pyramids where the encoder and decoder are connected through skip-connection. It is not merely suitable for image segmentation, but also to identify the categories of the region of interest. To address the need for more accurate medical image segmentation, we designed a multiple-branching architecture for segmenting objects with different sizes. MUNet with multi-scale segmentation masks achieves a 4.1% Dice score improvements, and an 8.1% AUC score improvements compared with U-Net in our datasets. The Dice coefficient can be improved 1.4% when we bring in ensemble methods and multi-scale training and inference in our datasets. The backbone of the framework can be replaced arbitrarily to meet the needs of different scenarios. We are sure that the MUNet architecture can be applied easily to many more tasks.

VI. Acknowledgment

This paper is supported by the Collaborative Innovation Center of Novel Software Technology and Industrialization at Nanjing University.

References

[1] D. Miyashita, S. Kousai, T. Suzuki, and J. Deguchi, "A neuromorphic chip optimized for deep learning and CMOS technology with time-domain analog and digital mixed-signal processing," J. Solid-State Circuits, vol. 52, no. 10, pp. 2679–2689, 2017. [Online]. Available: https://doi.org/10.1109/JSSC.2017.2712626

[2] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, http://www.deeplearningbook.org.

[3] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, 2014, pp. 580–587. [Online]. Available: https://doi.org/10.1109/CVPR.2014.81

[4] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 6517–6525. [Online]. Available: https://doi.org/10.1109/CVPR.2017.690

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 3431–3440. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298965

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481–2495, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2644615

[7] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," CoRR, vol. abs/1706.05587, 2017. [Online]. Available: http://arxiv.org/abs/1706.05587

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28

[9] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 936–944. [Online]. Available: https://doi.org/10.1109/CVPR.2017.106

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017. [Online]. Available: http://doi.acm.org/10.1145/3065386

[12] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I, 2014, pp. 818–833. [Online]. Available: https://doi.org/10.1007/978-3-319-10590-1_53

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 1–9. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298594

[14] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, 2015, pp. 448–456. [Online]. Available: http://jmlr.org/proceedings/papers/v37/ioffe15.html

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 2818–2826. [Online]. Available: https://doi.org/10.1109/CVPR.2016.308

[16] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA., 2017, pp. 4278–4284. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14806

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[19] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 2261–2269. [Online]. Available: https://doi.org/10.1109/CVPR.2017.243

[20] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA, 2008. [Online]. Available: https://doi.org/10.1109/CVPR.2008.4587503

[21] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011, 2011, pp. 1297–1304. [Online]. Available: https://doi.org/10.1109/CVPR.2011.5995316

[22] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States., 2012, pp. 2852–2860. [Online]. Available: http://papers.nips.cc/paper/4741-deep-neural-networks-segment-neuronal-membranes-in-electron-microscopy-images

[23] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016. [Online]. Available: http://arxiv.org/abs/1511.07122

[24] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. [Online]. Available: http://arxiv.org/abs/1412.7062

[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, 2018. [Online]. Available: https://doi.org/10.1109/TPAMI.2017.2699184

[26] G. Lin, A. Milan, C. Shen, and I. D. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 5168–5177. [Online]. Available: https://doi.org/10.1109/CVPR.2017.549

[27] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, 2017, pp. 2999–3007. [Online]. Available: https://doi.org/10.1109/ICCV.2017.324