An Improved Algorithm for Semantic Segmentation of Remote Sensing Images Based on Deeplabv3+

Jiaqi Liu
Beijing University of Posts and Telecommunications
No.10 Xitucheng Road, Haidian District, Beijing, China
(+86)130-5186-6799
jiaqiliu@bupt.edu.cn

Zhili Wang
Beijing University of Posts and Telecommunications
No.10 Xitucheng Road, Haidian District, Beijing, China
(+86)138-1193-9753
zlwang@bupt.edu.cn

Kangxin Cheng
Beijing University of Posts and Telecommunications
No.10 Xitucheng Road, Haidian District, Beijing, China
(+86)166-1996-4107
ckx@bupt.edu.cn

ABSTRACT

Remote sensing image segmentation is a more and more popular topic currently, and obviously it depends on the improvement of semantic segmentation. Encoder-decoder structure is an effective architecture in semantic segmentation. Since the encoder network helps to obtain various scale feature from the deep convolution layers, and the decoder network often help to recover the spatial resolution and location information in detail. Drawing on this idea, Google proposed the DeepLabv3+ [9] after DeepLabv3[8], which reused DeepLabv3 as its encoder, and added the decoder part to help the network to recover accurate location information. However, the decoder part of DeepLabv3+ is simple and sometimes it's difficult to obtain enough details from the encoder, and achieves not so good results on remote sensing images. Therefore, we design our decoder by adding more skip connections and convolution layers, which improves the result of building detection in the dataset SpaceNet [1].

CCS Concepts

• **Computing methodologies → Image segmentation**

Keywords

semantic segmentation; remote sensing image; encoder-decoder structure

1. INTRODUCTION

Experts analyzed remote sensing image by the method of visual interpretation with their own eyes several years ago. With the development of computer vision, the algorithm based on feature point extraction has been a better choice than analysis by human judgement, but one of the disadvantages is that the feature of the remote sensing image must be extracted by the traditional method of computer vision, and the feature is not suitable for all situation, which limits the interpretation effect. Fortunately, semantic segmentation dealt with deep convolutional neural network made remote sensing image interpretation easier these years.

Semantic segmentation is a significant topic in computer vision nowadays. The main goal of semantic segmentation is to assign labels to each pixel, which is also can be used in remote sensing image segmentation. However, semantic segmentation faces two main challenges come from itself: location and semantics. Various deep network deal with these two problems using their own methods. In this paper, we propose a new model by modifying DeepLabv3+ and designing the decoder part.

DeepLabv3+ dealt with those two problems by using ASPP (Atrous Spatial Pyramid Pooling) and encoder-decoder architecture. ASPP module applies several parallel atrous convolution with different rates, which makes network obtain multi-scale feature and makes feature map eight or sixteen or thirty-six times smaller. The encoder networks extract multi-scale contextual information and the decoder networks help to capture more accurate boundaries.

However, the decoder of DeepLabv3+ upsamples only two times, and each for four or eight times' expanding on feature map, and it's difficult to recover accurate boundary information via two skip connections because of not making full use of the location information of the encoder side, so the decoder could be improved by adding more skip connections between encoder and encoder and applying more convolution layers between the skip connection to help decoder generate better results.

Therefore, motivated by DeepLabv3+, to obtain more accurate boundaries, we add more skip connections in the part of decoder networks, which maximums the advantages of the encoder by receiving more accurate spatial information at the pixel level.

2. RELATED WORK

Jonathan Long et al. applied the Convolutional Neural Network (CNN) structure to the field of image semantics and achieved outstanding results, which is the Full Convolutional Network [2]. FCN achieved a score of 62.2% on all types of average cross-references at PASCAL VOC 2012 dataset, and the processing time for a typical image was approximately one-fifth of a second.

Inspired by FCN, image segmentation networks based on the encoder-decoder architecture have evolved, such as U-net [3] proposed by Olaf Ronneberger et al. and SegNet [4] put forward by Vijay Badrinarayanan et al. U-net uses the skip connection of the encoder to connect to the decoder, and the decoder uses information from the encoder side to help restore the target's details and spatial dimensions. SegNet changes the fully-connected layer in VGG-16[5] to the convolutional layer, and uses the pooling in the encoder as a guide for upsampling at the

ICCIP 2019, November 15–17, 2019, Chongqing, China.
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7258-9/19/11…$15.00
https://doi.org/10.1145/3369985.3370027

decoder side, that is, after each max-pooling operation on the encoder side, recording the index of the maximum value. So that the index can be used to restore the value to a more accurate position, and other positions are filled with zero during upsamples.

Liang-Chieh Chen et al. also proposed a semantic segmentation model based on attention mechanism [10], which is used to assign different weights on different scale features on inputs of different scales to improve the accuracy. Hanchao Li et al. proposed the Pyramid Attention Network [11], which combines the attention mechanism and the spatial pyramid structure to extract precise and dense features.

Yi Li et al. proposed an end-to-end semantic segmentation model [12], which is also the first end-to-end solution to instance segmentation. Mengye Ren et al. proposed an instance segmentation method based on recurrent attention [13]. Vladimir I. Iglovikov [14] proposed an instance segmentation algorithm based on full convolutional neural network taken the skeleton of U-net. The result of the segmentation is also the result of the instance segmentation.

DeepLab [6,7,8,9] proposed by Liang-Chieh Chen et al. has also received extensive attention. To this day, four versions of DeepLabs have been proposed. DeepLab took atrous convolution to expand the receptive field to get more contextual information. Its first two versions [6, 7] used a fully connected Conditional Random Fields (CRF) to improve position accuracy, and the results proved to be able to produce more advanced semantic segmentation results at the time. The authors proposed and improved ASPP (Atrous Spatial Pyramid Pooling) in subsequent versions, that is, the spatial convolution of different sampling rates is used to acquire the context of images on multiple scales in parallel.

However, almost all segmentation methods face two challenges: semantic and location. DeepLabv3+ [9], uses Encoder-Decoder architecture, where encoder network encodes multi-scale contextual information, and decoder network captures more accurate object boundaries. Using encoder and decoder network, DeepLabv3+ achieved good results on VOC2012, Cityscapes and other datasets. Figure 1 shows the change of feature maps' size.

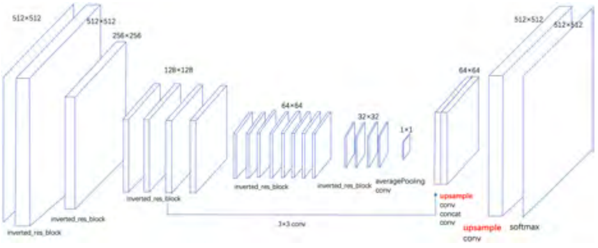

Figure 1. Size change of feature maps in DeepLabv3+

From DeepLabv3+, we know that there are twice upsample operations in decoder network, and each operation improves the size of feature map by four times. Not making full use of the location information of the encoder side, the decoder could be improved by adding more skip connections and convolution layers between encoder and encode to obtain more accurate boundaries, which maximums the advantages of the encoder by receiving more accurate spatial information at the pixel level.

3. MODEL

In this section, we introduce our proposed model based on DeepLabv3+ which reuses the encoder and decoder structure. We reuse the encoder part of DeepLabv3 and we change the decoder part, which could recover more accurate location information than the decoder of DeepLabv3+. Our new model achieves the mIOU (mean Intersection over Union) 0.7229 on SpaceNet and while DeepLabv3+ reaches only 0.7146. Next we introduce the two parts respectively and summarize our proposed model's new features.

3.1 Encoder

In the structure of encoder-decoder, the encoder extracts more and more abstract features from input images. With the help of ASPP (Atrous Spatial Pyramid Pooling), which is proposed in DeepLab, multi-scale contextual information is extracted by atrous convolution layers. ASPP uses several types of atrous convolution with various atrous rates, which can extract multi-scale context information, and help the whole network obtain more robust results. In our model, we apply MobileNetv2 [36] and Xception [37] as our encoder networks. After processing by encoder, the size of feature map become sixteen or thirty-six times smaller.

3.2 Decoder

Just like U-net, SegNet and DeepLabv3+ and other segmentation networks that follow encoder-decoder structure, decoder part of which is responsible for recovering spatial resolution and location information. Our decoder utilizes upsample to increase the size of the feature maps and recover the location information from the skip connections between encoder and decoder.

After encoder outputs, we get feature representation for the input image. In decoder part, the encoder features are first bilinearly upsampled by a factor of 2 and then concatenated with the corresponding low-level features from the encoder network that have the same spatial resolution. Notably, to prevent that the encoder features make training harder because of the corresponding low-level features' including a large number of channels, we apply 1×1 convolution layers after the low-level feature in encoder. After the concatenation, we apply a few 3×3 convolutions to refine the features. Then we apply the same operations with above repeatedly, which is upsampled bilinearly by a factor of 2 and concatenated with the low-level features after convolution layer from encoder and apply a few 3×3 convolutions to refine the features until the spatial resolution reaches the same shape with the input.

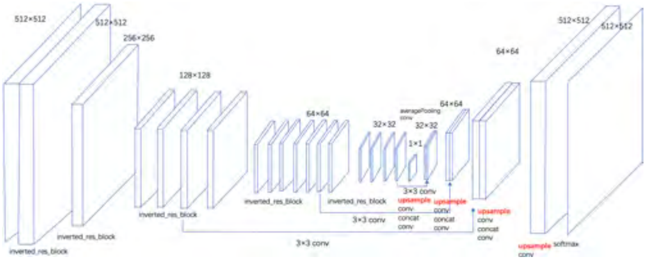

Figure 2. Size change of modified model by adding skip connections

3.3 New Features in our Network

The main idea of the encoder-decoder architecture is to extract abstract feature in encoder and recover accurate location information in decoder, from which we know the skip connections between encoder and decoder is to help the network obtain sharper segmentation boundaries. Motivated by this idea, we try to add the skip connection, which is shown in figure 2 and figure 3. Compared to DeepLabv3+, we add two upsample operations during decoder part in our proposed model.

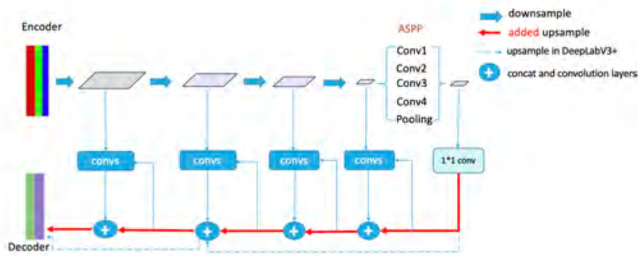

Figure 3. Modified model by adding skip connections

As is shown in figure 3, our new model has four upsample operations, and each upsample makes two times' expanding on feature maps. In decoder part, the feature maps perform upsample and concatenate with the feature maps after convolution layers from encoder, and More upsample connections in network make decoder obtain more details about location information.

A remarkable trick in our model is using batch normalization (BN) [16]. We add more connections and make the network more complex than DeepLabv3+, which increase the risk of overfitting, and maybe causes the network more difficult to converge. With the help of batch normalization, convergence becomes easier, and overfitting has been alleviated.

In conclusion, our proposed model draws on DeepLabv3+'s advanced ideas, and adds skip connections between encoder and decoder, which causes the network get sharper object boundaries.

4. EXPERIMENT

We perform our experiment on remote sensing image dataset SpaceNet. SpaceNet is a collection of remote sensing images provided by Digital Globe Commercial Satellite, which contains some tag information for machine learning research. The dataset uses satellite imagery with 30 cm resolution collected from DigitalGlobe's WorldView-3 satellite. Each image has 650×650 pixels and covers 195×195 m^2 of the earth surface. Moreover, each region consists of high-resolution RGB, panchromatic, and 8-channel low-resolution multi-spectral images. The satellite data comes from 4 different cities: Vegas, Paris, Shanghai, and Khartoum with different coverage, of (3831, 1148, 4582, 1012) images in the train and (1282, 381, 1528, 336) images in the test sets correspondingly [14].

We trained our proposed model on following four cities and tags for buildings: Las Vegas, Paris, Shanghai and Khartoum.

Our evaluation method is mIOU (mean Intersection over Union). The physical meaning of IOU is the ratio of the number of pixels of the intersection area and the union area of two regions. The value of IOU can be calculated by the formula below.

$$\mathrm{IOU} = \frac{\mathrm{Area}(\mathrm{A} \cap \mathrm{B})}{\mathrm{Area}(\mathrm{A} \cup \mathrm{B})}$$

We employ encoder as MobileNetv2 or Xception pretrained on ImageNet. Our implement is built on TensorFlow and works on two GPUs (NVIDA Titan X).

Our experiment results are shown below. The experiments represented by the two tables above use MobileNetv2 and Xception as encoder respectively.

Table 1. SpaceNet test set results (mIOU) with our proposed model using MobileNetv2 as encoder

Encoder	MobileNetv2				
city	Las Vegas	Paris	Shang hai	Khart oum	mean
DeepLabv3 + (%)	89.11	75.55	67.02	54.15	71.46
Our model (%)	89.87	75.34	68.71	55.23	72.29

Table 2. SpaceNet test set results (mIOU) with our proposed model using Xception as encoder

Encoder	Xception				
city	Las Vegas	Paris	Shang hai	Khartoum	mean
DeepLab v3+ (%)	89.14	76.90	70.03	56.19	73.07
Our model (%)	91.01	77.46	71.14	58.23	74.46

Regardless the network in encoder, whether it is MobileNetv2 or Xception, both table 1 and table 2 show that our proposed new decoder improves around 1% of mean IOU in all cities' buiding detection in the dataset of SpaceNet. Which can prove that our proposed new model could get a better fit on remote sensing image segmentation.

In order to show intuitively, we put some segmentation images below.

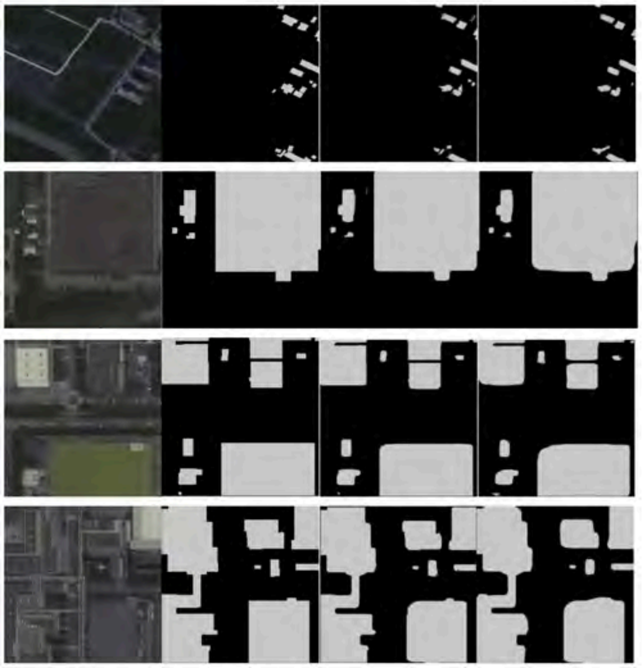

Figure 4. Segmentation results

As is shown in figure 4, each line has four images, and they are remote sensing image with buildings, the ground truth, segmentation results by our model, and segmentation results by DeepLabv3+ respectively. In the last three pictures of each line, gray pixel means the pixel label of the corresponding of the first image is building and the black pixel represents no building. From the pictures shown in figure 4, we found that our new model performs better in boundary processing and obtain more accurate location information because of our design for decoder.

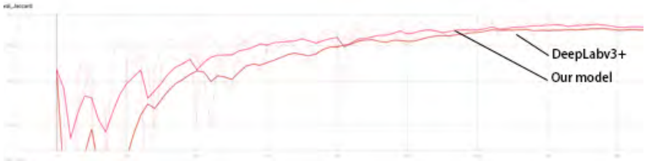

Figure 5. Convergence trend graph of our model and DeepLabv3+

The graph above shows the convergence of DeepLabv3+ and our model. The ordinate represents the Jaccard value of each model on validation set and abscissa represents the number of epochs of training. It's easy to find that our new model's convergence seems earlier and reaches a better result at last.

5. CONCLUSIONS

We developed a model for satellite imagery building detection. Our proposed model reuses the encoder-decoder structure, and use DeepLabv3 as our encoder, utilizing ASPP to extract features, which is proved to be useful and efficient again in our remote sensing dataset. We design more skip connections and convolution layers in decoder than DeepLabv3+, which helps model obtain more accurate location information and improves the remote sensing image segmentation results. Eventually, our experiment results show that our proposed model achieves the state-of-the-art result of building detection in SpaceNet dataset.

6. REFERENCES

[1] https://spacenetchallenge.github.io/

[2] Long, Jonathan, Evan Shelhamer and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015): 3431-3440.

[3] Ronneberger, Olaf, Philipp Fischer and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." *ArXiv* abs/1505.04597 (2015): n. pag.

[4] Badrinarayanan, Vijay, Alex Kendall and Roberto Cipolla. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence 39* (2015): 2481-2495.

[5] Simonyan, Karen and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *CoRR* abs/1409.1556 (2014): n. pag.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy and Alan L. Yuille. "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs." *CoRR* abs/1412.7062 (2014): n. pag.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy and Alan L. Yuille. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs." *IEEE Transactions on Pattern Analysis and Machine Intelligence 40* (2016): 834-848.

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff and Hartwig Adam. "Rethinking Atrous Convolution for Semantic Image Segmentation." *ArXiv* abs/1706.05587 (2017): n. pag.

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff and Hartwig Adam. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation." *ECCV* (2018).

[10] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu and Alan L. Yuille. "Attention to Scale: Scale-Aware Semantic Image Segmentation." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015): 3640-3649.

[11] Fan, Lei, Huifang Kong, Wei-Chien Wang and Jiapeng Yan. "Semantic Segmentation With Global Encoding and Dilated Decoder in Street Scenes." *IEEE Access 6* (2018): 50333-50343.

[12] Li, Yi, Haozhi Qi, Jifeng Dai, Xiangyang Ji and Yichen Wei. "Fully Convolutional Instance-Aware Semantic Segmentation." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016): 4438-4446.

[13] Ren, Mengye and Richard S. Zemel. "End-to-End Instance Segmentation with Recurrent Attention." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016): 293-301.

[14] Iglovikov, Vladimir I., Selim S. Seferbekov, Alexander V. Buslaev and Alexey Shvets. "TernausNetV2: Fully Convolutional Network for Instance Segmentation." *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2018): 228-2284.

[15] Visin, Francesco, Adriana Romero, Kyunghyun Cho, Matteo Matteucci, Marco Ciccone, Kyle Kastner, Yoshua Bengio and Aaron C. Courville. "ReSeg: A Recurrent Neural Network-Based Model for Semantic Segmentation." *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2015): 426-433.

[16] Ioffe, Sergey and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." *ArXiv* abs/1502.03167 (2015): n. pag.

[17] Wang, Guangrun, Ping Luo, Liang Lin and Xiaogang Wang. "Learning Object Interactions and Descriptions for Semantic Image Segmentation." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017): 5235-5243.

[18] Luo, Ping, Guangrun Wang, Liang Lin and Xiaogang Wang. "Deep Dual Learning for Semantic Image Segmentation." *2017 IEEE International Conference on Computer Vision (ICCV)* (2017): 2737-2745.

[19] Wang, Panqu, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou and Garrison W. Cottrell. "Understanding Convolution for Semantic Segmentation." *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2017): 1451-1460.

[20] Hariharan, Bharath, Pablo Andrés Arbeláez, Ross B. Girshick and Jitendra Malik. "Hypercolumns for object segmentation and fine-grained localization." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014): 447-456.

[21] Zhang, Zhenli, Xiangyu Zhang, Chao Peng, Xiangyang Xue and Jian Sun. "ExFuse: Enhancing Feature Fusion for Semantic Segmentation." *ECCV* (2018).

[22] Noh, Hyeonwoo, Seunghoon Hong and Bohyung Han. “Learning Deconvolution Network for Semantic Segmentation.” *2015 IEEE International Conference on Computer Vision (ICCV)* (2015): 1520-1528.

[23] Peng, Chao, Xiangyu Zhang, Gang Yu, Guiming Luo and Jian Sun. “Large Kernel Matters — Improve Semantic Segmentation by Global Convolutional Network.” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017): 1743-1751.

[24] Cote, Melissa and Parvaneh Saeedi. “Automatic Rooftop Extraction in Nadir Aerial Imagery of Suburban Regions Using Corners and Variational Level Set Evolution.” *IEEE Transactions on Geoscience and Remote Sensing 51* (2013): 313-328.

[25] Cohen, Joseph Paul, Wei Ding, Caitlin Kuhlman, Aijun Chen and Liping Di. “Rapid building detection using machine learning.” *Applied Intelligence 45* (2016): 443-457.

[26] Yuan, Jiangye. “Automatic Building Extraction in Aerial Scenes Using Convolutional Networks.” *ArXiv* abs/1602.06564 (2016): n. pag.

[27] Zhang, Amy, Xianming Liu, Andreas Gros and Tobias Tiecke. “Building Detection from Satellite Images on a Global Scale.” *ArXiv* abs/1707.08952 (2017): n. pag.

[28] Neuhold, Gerhard, Tobias Ollmann, Samuel Rota Bulò and Peter Kontschieder. “The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes.” *2017 IEEE International Conference on Computer Vision (ICCV)* (2017): 5000-5009.

[29] He, Kaiming, Xiangyu Zhang, Shaoqing Ren and Jian Sun. “Deep Residual Learning for Image Recognition.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015): 770-778.

[30] Demir, Ilke, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia and Ramesh Raskar. “DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images.” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2018): 172-17209.

[31] Cordts, Marius, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth and Bernt Schiele. “The Cityscapes Dataset for Semantic Urban Scene Understanding.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016): 3213-3223.

[32] Dai, Jifeng, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu and Yichen Wei. “Deformable Convolutional Networks.” *2017 IEEE International Conference on Computer Vision (ICCV)* (2017): 764-773.

[33] Islam, Md. Amirul, Mrigank Rochan, Neil D. B. Bruce and Yang Wang. “Gated Feedback Refinement Network for Dense Image Labeling.” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017): 4877-4885.

[34] Holschneider, Matthias, Richard Kronland-Martinet, Jean Morlet and Ph. Tchamitchian. “A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform.” (1989).

[35] Sandler, Mark, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov and Liang-Chieh Chen. “MobileNetV2: Inverted Residuals and Linear Bottlenecks.” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018): 4510-4520.

[36] Chollet, François. “Xception: Deep Learning with Depthwise Separable Convolutions.” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016): 1800-1807.