Using Neural Networks in Segmentation of Aerial Images

Pavel Kohout
Brno University of Technology
Brno, Czech Republic
xkohou15@stud.fit.vutbr.cz

Alena Tesařová
Brno University of Technology
Brno, Czech Republic
xtesar36@stud.fit.vutbr.cz

Petr Kohout
Brno University of Technology
Brno, Czech Republic
xkohou14@stud.fit.vutbr.cz

ABSTRACT

Semantic segmentation on high-resolution aerial images plays a significant role in many remote sensing applications. Despite the rapid development of methods for image processing and huge success in the area of deep learning, reliable high-resolution aerial images segmentation are still a challenging task. To tackle this problem, the difficult datasets as *iSAID* and *DOTA* were created. For our project, we decided to implement a Dual-Path neural network that combines spatial and global context over the spatial and edge path. It helps to solve the problem with intra-class heterogeneity and inter-class homogeneity. We managed to compare our implementation with existing methods that were tested on *iSAID* dataset that contains 15 classes, such as plane, cars, harbor, large vehicles etc. Due to the lack of computation resources we could not get results as good as the results of the state-of-the-art methods. However, we were able to get the comparable results within 3 classes.

Keywords: semantic segmentation, high-resolutional aerial images, neural networks, Dual-path, deep learning, iSAID dataset

1 INTRODUCTION

With the rapid development of remote sensing technologies, we get more and more high resolution images from various domain such us urban planning, landscape classification, relief, etc. As a result, real-time semantic segmentation is becoming more and more important and receive more attention. Generally, there are some traditional image segmentation, such as watershed algorithm [10], graph cuts [2] and random forest [11], but they usually need to set some settings manually and can not provide accurate semantic segmentation. In the past few years, deep learning methods have significantly promoted the development of semantic segmentation, especially the *Fully Convolution Network* (FCN). However, they are still not perfect and struggle with 2 big problems, which are intra-class heterogeneity and inter-class homogeneity. To tackle these two challenges, we need to consider each category of pixels as a whole, instead of assigning semantic label to each single pixel independently. It's important to combine multi-level and global context feature so we can be able to categorize various objects belonged to the same semantic label.

The purpose of this work was to compare the accuracy of existing network implementations specialized on semantic segmentation with our implementation based on architecture proposed in article *A Dual-Path and Lightweight Convolutional Neural Network for High-Resolution Aerial Image Segmentation* [19]. The proposed network is composed of many parts. The first is the backbone. In the original article this backbone is based on *MobileNetV2* architecture [13]. Intermediate results from a few last layers of the backbone are distributed into two separated paths. In these paths, they process the image for semantic and edge analysis. According to this knowledge we experimented with the backbone of this network and with the settings of dual path to achieve the best results.

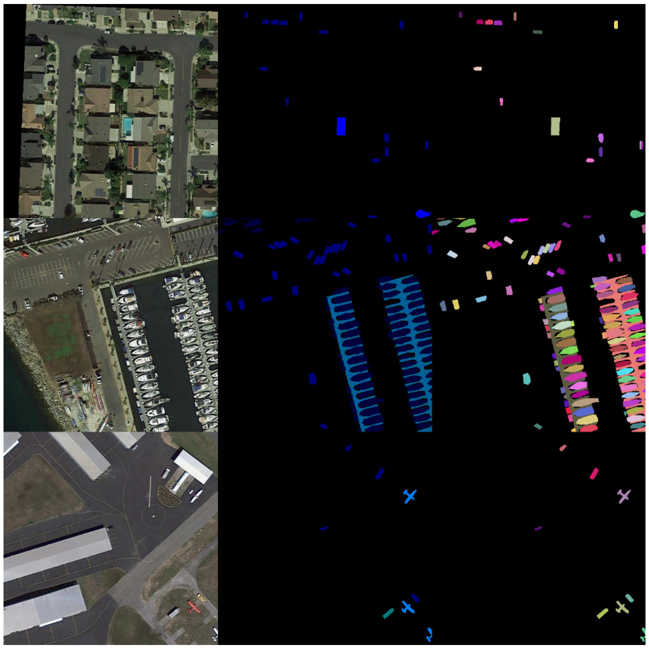

Figure 1: Example of segmentation tasks for our network. The area contains 15 categories to segment. iSAID dataset [16] selects only a few objects to segment on the images – e.x. *boat, harbor, car, plane* . . .

2 RELATED WORK

While convolution network have been used for a long time, their success was limited by the amount of available training images and computing resources. Since *AlexNet* [9] won the *ImageNet* Competition in 2012, Deep Convolution Neural Network (DCNN) has become the mainstream method in the field of computer vision and achieved also great results in semantic segmentation.

The goal of semantic segmentation is to assign each pixel with a semantic label that can represent people, animals, objects, etc. FCN is considered as a milestone in deep learning techniques for semantic segmentation. *SegNet* [1] introduces an encoder–decoder architecture for semantic segmentation. The encoder extracts features via convolution, max pooling and activation layers. The decoder is similar to the encoder. It upsamples the input, using indices stored from the encoding stage. U-net [12] is a U-shaped architecture, which is symmetric DCNN and uses skip connections between the sampling path and the downsampling path. Deeplab [3] introduces atrous convolution in DCNN to enlarge the receptive fields without increasing the number of network parameters. In DenseNet [7], each layer receives feature maps from all preceding layers and passes on its output feature maps to all subsequent layers. Therefore, the loss is propagated directly and this way the problem with vanishing gradient is solved. The state-of-the-art models employ the following technologies that are widely used in semantic segmentation algorithm:

(1) **skip connections** between lower convolutions layers and higher convolutional layers to use features from different levels,
(2) **use atrous convolution** to enlarge receptive fields without increasing computational parameters,
(3) **global pooling** convolutional layer to guide the location of objects.

These technologies are used to tackle the challenge of inter-class heterogeneity problem. In this work we used Holistically-nested Edge Detection-structured sub network to extract semantic boundaries of our network to help with the problem of the inter-class homogeneity, as described in article [19].

All these methods are trained on conventional datasets. However these method does not have to work on aerial image dataset. They often fail to provide satisfactory results due to large domain shift, high density objects with large variations in orientation and scale. In addition, instance segmentation is a challenging problem that goes one step ahead than regular object detection as it aims to achieve precise per-pixel localization for each object instance.

3 DATASET

For the purpose of our task we chose the *iSAID* [16] dataset. This dataset contains a huge amount of manually labeled aerial images. It is based and it extends *DOTA-v1.0* dataset [17].

The data for this dataset is possible to download from corresponding websites [5, 15] or you can run *dataset_preparator* script that downloads both datasets – from *iSAID* and *DOTA-v1.0* websites.

This dataset contains 665, 451 annotated instances divided into 15 categories – *Plane*, *Ship*, *Storage tank*, *Baseball diamond*, *Tennis court*, *Basketball court*, *Ground track field*, *Harbor*, *Bridge*, *Large vehicle*, *Small vehicle*, *Helicopter*, *Roundabout*, *Swimming pool*, *Swimming pool*. Occurrences of classes in the individual figures are uneven (as shown in Figure 2).

3.1 Preprocessing

The dataset is very huge and it contains pictures of different sizes and scales. We created subset from existing dataset to shorten training times of our models. In creation of the subdataset we focused on keeping similar distribution as in the whole dataset. The pictures have to be cut into batches of crops with the same size. We decided to crop images to sizes of *128x128* and *512x512* pixels. In these crops we can find many pictures that don't contain any instance or only the small parts of them. We tried to identify and filter these pictures during the runtime but it prolonged iteration time for long period. For this purpose we created cache records that describe which crops contain whole instances. We used *OpenCV* for cropping the images. In the *Dual-Path* 4.2 solution, authors use the edge path classifier to find edges in the pictures but the template (reference for the edge network) can not be found in the dataset. For this purpose we used *OpenCV* for the edge detection.

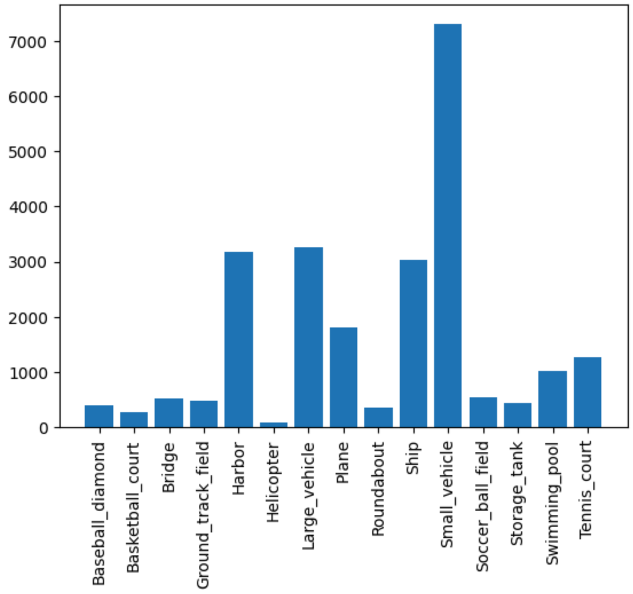

Figure 2: Illustration of the class distribution over the dataset.

4 OUR SOLUTION

In this chapter, we are describing our work done on this project. The aim of the work was to compare existing implementations of segmentation networks with ours based on architecture proposed in [19]. Therefore, this section is divided into three parts:

(1) implemented baseline solution,
(2) Dual-path implementation of segmentation network,
(3) comparing results.

4.1 Baseline solution

As a baseline solution we firstly picked U-net but after few rounds of training on this model we found out that the dataset is too hard for U-net and results were not sufficient. That is why we decided to try *FPN* (Feature Pyramid Network) because it is already implemented in *pytorch* so we did not have to modify anything and could easily test it. This network has been created by Facebook for semantic segmentation. The main trick is in building feature pyramids inside the convolutional neural network. It composes of bottom-up and top-down pathway. The bottom-up uses *ResNet* and is used for feature extraction and as we go up, the spatial resolution decreases and we get more semantic value for each layer (as shown in Figure 3). [8, 14]

Implementation details. We trained pretrained model on Imagenet with batch-size 16, *BCEDiceLoss* loss function, base learning rate 0.001. As a optimizer we picked *RAdam*. We trained for 100 epochs. As the images of iSAID dataset are very high-resolution, we could

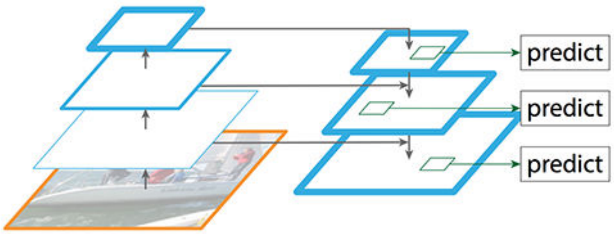

Figure 3: FPN composes of a bottom-up and a top-down pathway.

not feed them directly into the DCNN, We cropped all the images into 128x128 as described in section 3.1. Our baseline solution was implemented under pytorch [6] framework. All experiments run in Google Colab Notebook which executes the code on Linux PC with 64 bit Ubuntu 18.04, CPU with 16GB memory and Tesla P4.

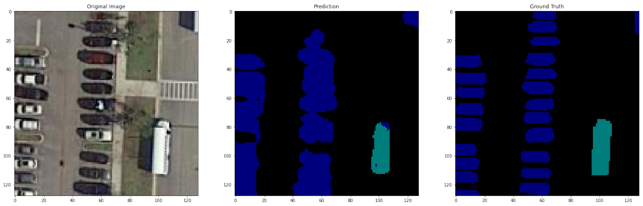

Figure 4: Result of pretrained FPN with iou over 50% in cars and large vehicles category.

4.2 Our implementation of Dual-Path architecture

In this section, we introduce our solution for the high-resolution aerial image segmentation problem. The high-resolution aerial images obtain a lot of information, which makes the precised segmentation task really challenging. To tackle this problem, we are formulating the high-resolution segmentation task as a multi-task learning framework by exploring the complementary information, which can predict the results of semantic labels and boundaries simultaneously. Therefore, we decided to apply the dual-path network architecture proposed in [19]. It combines spatial information extracted from multiple levels of the encoder and its global features dealing with intra-class heterogeneity problems. The boundary path helps differentiate objects with similar appearance but different semantic labels. As the feature extraction network, we adopt a high-performance lightweight network architecture *MobileNetV2.*

Network architecture. The overall structure of dual-path architecture is shown in Figure 5. The framework is based on the *MobileNetV2* network with several improvements. The first is the application of atrous convolution in the last four Bottleneck Residual Blocks (BRBs) instead of classical convolution. Atrous convolution enlarges receptive fields without reducing the resolution [4]. The second is removing the convolution stride of 2 in the first layer of BRBs 4 and 6.

The decoder is composed of spatial and edge paths. The spatial path combines multilevel features with a global context. Multilevel features are obtained from BRBs 3–8 of the *MobileNetV2* network. Each feature is propagated through the channel attention model (CAM) and features of adjacent blocks are summed stage-by-stage. The CAM outputs are concatenated and combined with global average features from the last encoder block. Therefore, the spatial path is able to combine the spatial information from lower blocks with fine semantic information from higher levels thanks to large receptive fields.

The edge path is a *HED-like* [18] network, which employs deep supervision at each side-output layer. The aim of the edge path is to improve network discriminative ability in the cases where two objects from different classes share a similar appearance and they are spatially adjacent.

According to experiments, we decided to modify the structure of *MobileNetV2* described in the original paper [19] where we set the stride to 1 in the first bottleneck residual block (BRB).

Loss function. In this work, we use the **cross-entropy** loss function for both spatial and edge paths. The training of the network is formulated as a per-pixel classification problem regarding the ground-truth segmentation masks. Mask is composed of the semantic object and its boundaries. Therefore, the loss function is the combination of both losses from spatial and edge paths respectively. The balance between them is realized by applying α and β coefficients.

$$L_{total} = \alpha \times L_{spatial} + \beta \times L_{edge} \tag{1}$$

Implementation details. Due to constrained computation capacity, we used images with *128x128* resolution. According to the experimental results, we selected Adam optimizer and tested several setups with different weight decay parameters (0.0005, 0.0001, 0.00005) and learning rate values. We decided to exclude the weight decay factor from the training phase, therefore it was set to 0. We trained the dual-path framework for 50 epochs and divided the learning rate by 10 after 10, 25, and 45 epochs. We started with a learning rate set to 0.001. The batch size was set to 10.

Our Dual-Path solution was implemented under pytorch [6] framework. All experiments were run in Google Colab Notebook which executes the code on Linux PC with 64 bit Ubuntu 18.04, CPU with 16GB memory and Tesla P4.

We wanted to test the effect of additional edge path to a solution. For that reason, two models were examined, denoted as *SpatialPath* and *SpatialEdgePath*. *SpatialPath* model ignores the prediction of edge path completely and uses only prediction from spatial path, while *SpatialEdgePath* uses both predictions with coefficients set as $\alpha = 1$ and $\beta = 0.0005$.

5 RESULTS

In this work, we tried to run many segmentation networks and compare them on our dataset. We run classical segmentation network *FPN* as our baseline [12] but mainly focused on the implementation of the *Dual-Path* network. The aim was to compare the accuracy of these existing network implementations on *iSAID* dataset with our implementation. As a reference solution we looked up to the

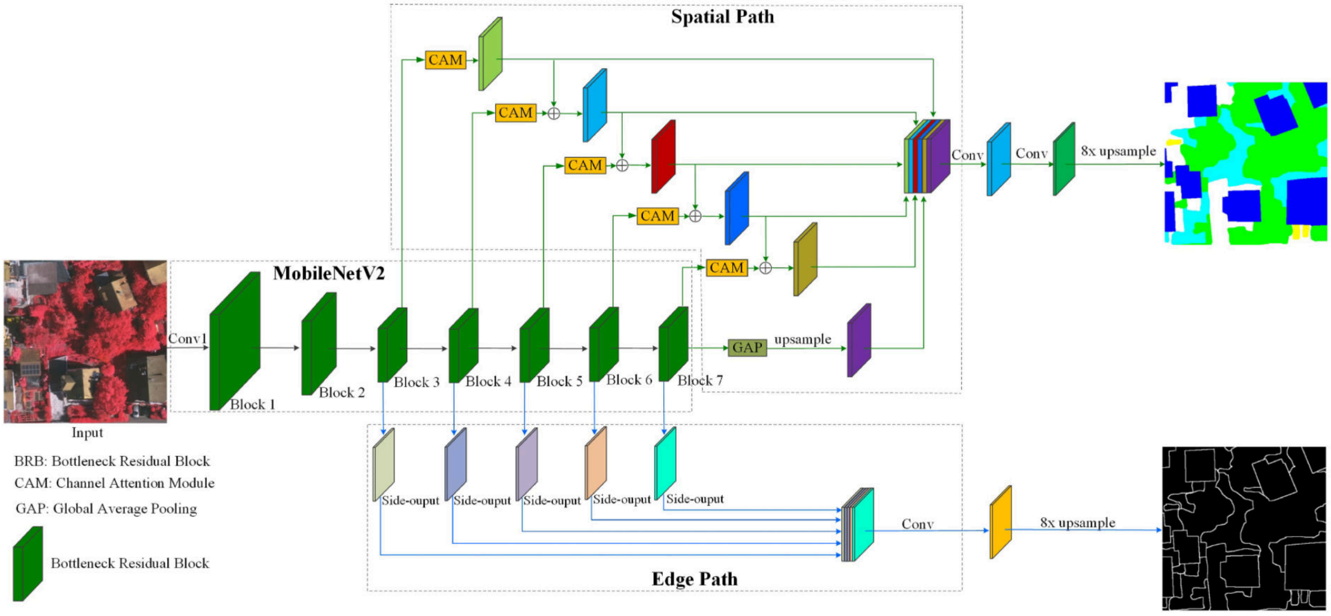

Figure 5: Structure of Dual-path architecture as proposed in [19]. It is composed from 3 main parts: encoder, spatial path and edge path. The encoder extracts several features while spatial and edge paths are propagating multilevel information to the output feature map.

Method	AP	Plane	BD	Bridge	GTF	SV	LV	Ship	TC	BC	ST	SBF	RA	Harbor	SP	HC
Mask R-CNN [16]	25.7	37.7	42.5	13.0	23.6	6.9	7.4	26.6	54.9	34.6	28.8	20.8	35.9	22.5	25.1	5.3
Mask R-CNN+ [16]	33.4	41.7	39.6	15.2	25.9	16.9	30.4	48.8	72.9	43.1	32.0	26.7	36.0	29.6	36.7	5.6
PANet [16]	34.2	39.2	45.5	15.1	29.3	15.0	28.8	45.9	64.1	47.4	29.6	33.9	36.9	26.3	36.1	9.5
PANet++ [16]	40.0	48.7	50.3	18.9	32.5	20.4	34.4	56.5	78.4	52.3	35.4	38.8	40.2	35.8	42.5	13.7
FPN	6.4	0.0	0.0	0.0	0.0	41.1	28.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SpatialPath	4.7	0	0.0	0.0	0.0	12.5	5.6	0.03	0.0	0.0	5.2	0.0	0.0	28.3	0.0	0.0
SpatialEdgePath	3.6	7.4	0.0	0.0	0.0	14.0	15.8	1.88	0.0	0.0	0.01	0.0	0.0	0.28	0.0	0.0

Table 1: Class-wise instance segmentation results on iSAID test set. Note that short names are used to define categories: BD-Baseball diamond, GTF-Ground field track, SV-Small vehicle, LV-Large vehicle TC-Tennis court, BC-Basketball court, ST-Storage tank, SBF-Soccer-ball field, RA-Roundabout, SP-Swimming pool, and HC-Helicopter.

results mentioned in article [16] about *iSAID* dataset where they use *PAN* and *R-CNN* as their baseline networks. All the results are compared in the final table 1 using metrics *Intersection over Union* (IoU) computed as:

$$IoU = \frac{Area\_of\_Intersection\_of\_two\_objects}{Area\_of\_Union\_of\_two\_objects} \tag{2}$$

We ran multiple experiments with our implementation and found some important factors to correct network function. The most influencing factor for our model was the batch size. At the beginning, we used batch size of two and cropped the images to size 512x512. After obtaining very poor results we decided to go for smaller crops of original images (128x128). It enabled us to increase the batch size up to 10 images per batch which had a good effect on learning. As loss functions, **cross-entropy**, **dice** and **mean square error** were used. The cross-entropy has shown the best results on *iSAID* dataset. The example of predicted segmentation mask by our network can be seen in Figure 6. It also shows that our solution is able to highlight cars that occur in the clear area. Furthermore, the proposed segmentation by our model seems to fit real shapes of cars quite well even if the target segmentation mask is marking cars as plain rectangles.

The *iSaid* dataset of aerial images is a challenging task for segmentation. The results obtained from the literature show that even the state-of-the-art methods for segmentation are not able to reach high prediction accuracy on this dataset. They trained the net on 8 GPU for 180k iterations with mini-batch size of 16 on the whole dataset (in comparison with our small dataset subset) with image size of 800x800. Even with this settings and available computational resources they could not reach neither 50% average accuracy. The maximum average precision they got was 40%. It shows the complexity of the chosen dataset.

Regarding our implemented solutions (*FPN*, *SpatialPath* and *SpatialEdgePath*), these models were not able to learn the classification for each class. Instead of detecting every class, the models focus on a few of the most frequently represented classes. Within these individual classes (small and large vehicles, harbors) some of our networks reach comparable results with the other state-of-the art methods. Especially, *SpatialPath* network performs well in the case of harbor segmentation reaching 28% accuracy and *FPN* which gets the best results for small vehicles of all nets reaching accuracy of 41%. It also got great results for large vehicles comparable with PANet results (both reaching 28.8%).

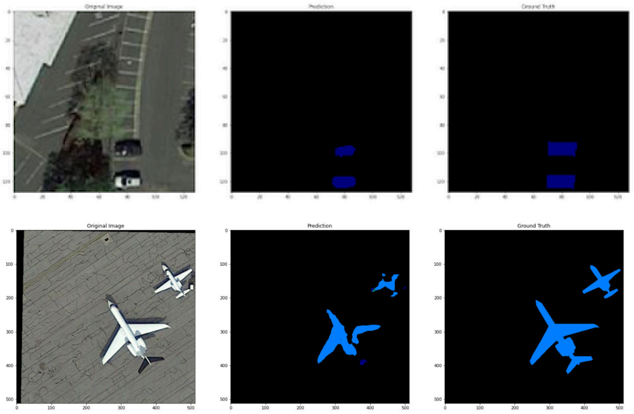

Figure 6: Results of our Dual-Path network implementation (denoted as *SpatialEdgePath*) with spatial and edge path having IoU over 14% for small car segmentation and 7.4% accuracy on planes.

6 CONCLUSION

The purpose of this work was to compare the accuracy of existing network implementations specialized on semantic segmentation with our implementation based on architecture proposed in article *A Dual-Path and Lightweight Convolutional Neural Network for High-Resolution Aerial Image Segmentation* [19].

We successfully implemented this network but we did not reach the same results as reference methods because the whole dataset was not used for training and we did not have enough computational resources to train the net properly. The proposed network was compared with other DCNNs, such as FPN (our baseline), PANet and Mask R-CNN and achieved good results in classification of Harbor class. The best segmentation results got our baseline model *FPN* in small vehicle class detection reaching precision over 40%. During this project, we found out that the increase of batch size was the most beneficial techniques for segmentation improvement.

Based on our opinion, we would reach better results by balancing distribution of all classes. The network would be forced to learn detection of other classes more frequently which may cause more successful results in every class segmentation. In the case of *FPN*, we observed the benefits of transfer learning. Using already trained model increase accuracy significantly. Also, an increase in training time is expected to make models more accurate as the usage of whole training *iSAID* dataset.

CODE AVAILABILITY

The code is available on:
https://gitlab.com/xkohou15/KNN_segmentation/-/tree/master.

REFERENCES

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39, 12 (December 2017), 2481—2495. https://doi.org/10.1109/tpami.2016.2644615

[2] Y.Y. Boykov and M.-P. Jolly. 2001. Interactive graph cuts for optimal boundary region segmentation of objects in N-D images. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, Vol. 1. 105–112 vol.1. https://doi.org/10.1109/ICCV.2001.937505

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2017. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. arXiv:1606.00915 [cs.CV]

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

[5] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Micheal Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. [n.d.]. DOTA A Large-Scale Benchmark and Challenges for Object Detection in Aerial Images. https://captain-whu.github.io/DOTA/dataset.html. Accessed: 2021-05-08.

[6] Facebook. [n.d.]. Pytorch. online. http://pytorch.org

[7] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[8] Jonathan Hui. 2018. Understanding Feature Pyramid Networks for object detection (FPN). https://jonathan-hui.medium.com/understanding-feature-pyramid-networks-for-object-detection-fpn-45b227b9106c

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems* 25 (01 2012). https://doi.org/10.1145/3065386

[10] F. Meyer and S. Beucher. 1990. Morphological segmentation. *Journal of Visual Communication and Image Representation* 1, 1 (1990), 21–46. https://doi.org/10.1016/1047-3203(90)90014-M

[11] M. Pal. 2005. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing* 26, 1 (2005), 217–222. https://doi.org/10.1080/01431160412331269698 arXiv:https://doi.org/10.1080/01431160412331269698

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *CoRR* abs/1505.04597 (2015). arXiv:1505.04597 http://arxiv.org/abs/1505.04597

[13] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[14] Sik-Ho Tsang. 2019. Review: FPN — Feature Pyramid Network (Object Detection). https://towardsdatascience.com/review-fpn-feature-pyramid-network-object-detection-262fc7482610

[15] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. [n.d.]. iSAID A Large-scale Dataset for Instance Segmentation in Aerial Images. https://captain-whu.github.io/iSAID/dataset.html. Accessed: 2021-05-08.

[16] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. 2019. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 28–37.

[17] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. 2018. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Saining Xie and Zhuowen Tu. 2015. Holistically-nested edge detection (2015). *arXiv preprint arXiv:1504.06375* (2015).

[19] Gang Zhang, Tao Lei, Yi Cui, and Ping Jiang. 2019. A dual-path and lightweight convolutional neural network for high-resolution aerial image segmentation. *ISPRS International Journal of Geo-Information* 8, 12 (2019), 582.