Deep Neural Networks Capabilities for Semantic Segmentation of Noisy Aerial Images

Aleksandr Markelov, Ivan Krivorotov and Vadim Gorbachev

FSUE «GosNIIAS» (SSC RF), Moscow, Russia
{markelov.ao,krivorotov.ia,vadim.gorbachev}@gosniias.ru

Abstract. Semantic segmentation is one of the important ways of extracting information about objects in images. State of the art neural network algorithms allow to perform highly accurate semantic segmentation of images, including aerial photos. However, in most of the works authors use high-quality low-noise images. In this work, we study the ability of neural networks to correctly segment images with intensive uncorrelated Gaussian noise. The study brings us three main conclusions. Firstly, it demonstrates that neural network algorithms are capable of working with extreme image distortions without using additional filtration or image recovery techniques. Secondly, the experiments quantitatively show that distortion intensity can be negated with increased training set size. Such process is similar to model's quality improvement and generalization due to training dataset enlargement. Finally, we quantitatively demonstrate how image aggregation techniques affect training with noised data.

Keywords: Convolutional Neural Networks, Semantic Segmentation, Image Distortion, Aerial Images, Image Aggregation.

1 Introduction

Nowadays, there is an increased interest in the field of computer vision. This is due to significant progress in the field of deep neural networks (DNN) design, increase in available computational resources, as well as availability of huge databases of labeled data. The combination of these factors allows us to solve a wide variety of tasks that were previously inaccessible to classical computer vision algorithms.

Along with the range of tasks expansion, we naturally encounter questions about limit of the applicability of given methods. Such limitations can be determined by the problem formulation, available computational power, DNN building and training techniques, data quality, etc. In this paper, we study limit of applicability of DNN in case of noisy data. We also suggest ways of negative effects reduction with image aggregation methods.

* Publication is supported by RFBR grant №19-07-00844.

One of the practically important tasks in high-level image analysis is the semantic segmentation of images, in particular aerial images. A similar problem arises in the planning and administration of territories, environmental monitoring, etc. One of the most effective ways of solving such problems today are DNNs. With the development of unmanned aircraft, aerial data becomes more accessible. At the same time, it is known that the accuracy of the method strongly depends on the quality of the input data. Good results can be achieved mainly in the case of high-quality aerial input images with perfect weather conditions. In practice, collecting high-quality data is a complex and financially costly procedure. It is much easier to obtain data that has a significantly lower level of quality and a relatively high level of noise, but abundance of such data causes a great interest in their use. Noises and distortions can have a different nature: camera matrix noise, compression artifacts, distortions arising in the processing and transmission of information, atmospheric artifacts, etc.

In this work we tried to quantitatively study the behavior of neural network segmentation algorithms in the case of highly noised data, answering two main questions:

1. how solution accuracy depends on the noise level of input data.
2. is it possible to compensate lack of data quality with training dataset volume.

Neural network development progress inspires great optimism among community of researchers and suggest positive answer to the second question. However, it is extremely difficult to find exact quantitative studies of the issue on public data collections. The result of study may expand possibilities of using data mining and neural algorithms in wide range of industrial tasks. It can also show ways of reducing requirements for computer vision systems. In particular it may reduce data compression accuracy requirements.

1.1 Problem formulation

The paper investigates the problem of multiclass segmentation of aerial images. Dataset of such tagged images is the ISPRS Semantic Labeling Contest. It consists of images of Potsdam city [1]. The goal is to determine if each pixel of four-channel (RGB, IR) aerial image belongs to one of the classes. This results in semantic map of aerial image. Table 1 shows the classes and their corresponding color on the segmentation maps.

Table 1. Image segmentation mapping

Class	Color	HEX-code
Buildings	Blue	0000ff
Vegetation	Green	00ff00
Concrete, asphalt	White	ffffff
Cars	Yellow	ffff00
Clutter	Red	ff0000
Pedestrian space	Turquoise	00ffff

Contestants have access to 38 images with a resolution of 6000x6000 pixels. It is worth noting, that only 24 images have segmentation maps and are suitable for supervised learning. In addition to standard RGB images, there are also images with an infrared channel. To exploit all of the available information for segmentation we used four-channel images with IR channel. An example of such image and corresponding is shown in Fig. 1.

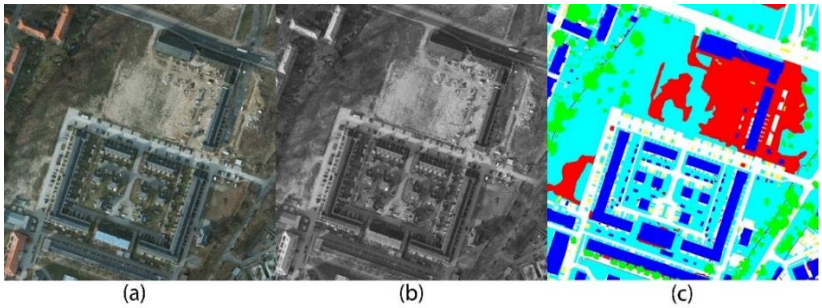

Fig. 1. – Optical image (a), infrared image (b), ground truth segmentation map (c)

1.2 Segmentation methods overview

Historically there are a large number of methods for semantic segmentation. The most successful models have the encoder-decoder architecture. Encoder transforms image into a vector of features. Then this feature vector is transformed into an image matrix using a decoder network. One of the first architectures for neural network segmentation is FCN-8s [2], released in 2014. Pre-trained convolutional networks, such as ResNet [3] and VGG [4], are often used as an encryption network. In turn, decoder is chosen from diverse implementation possibilities. For example, the SegNet architecture [5] uses the unpooling operation. During the max-pooling operation, at the convolution stage in the encoder, the maximum value indices are stored and later used to increase the discretization of the corresponding feature maps in the decryption network by performing the unpooling operation using stored indexes. The U-net model [6] uses the idea of skip-connections to preserve spatial information. Feature maps from the encryption network are directly transmitted and concatenated with feature maps on the corresponding layers of the decoder network, in parallel with the usual convolutional layers. LinkNet [7] uses the addition of feature maps instead of concatenation. The DeepLab[8] architecture introduced three innovations. Firstly they implemented convolution filters with increased receptive field (atrous convolution, dilated convolution). Secondly, the authors were the first to propose a spatial pyramidal union (ASPP) of such filters for segmenting objects at different scales. Thirdly, the localization of object boundaries was improved by combining methods from deep convolutional neural networks and probabilistic graphical models (CRF) to take into account contextual information.

2 Network architecture

In this work we use a DeepLabV3+ architecture with ResNet-101 backbone. The choice of this architecture is due to its highest segmentation performance according to the IoU metric on validation dataset. Comparison results are present in Table 2.

Table 2. Model comparison

Architecture	Backbone	IoU metric on validation set, %
DeepLabV3+	ResNet-101	**81,65**
DeepLabV3+	ResNet-34	78,83
Unet	ResNet-101	79,21
PSPNet	ResNet-101	73,09
LinkNet	ResNet-101	78,99

We also compare model performance with other existing methods. Such methods are presented in ISPRS Potsdam Semantic Labeling Contest. One of the officially evaluated reports uses graph-based segmentation method, chessboard segmentation and conditional random field pixel classification. Methods are compared via F1-score metric in Table 3. Results for SVL_1 are taken from official competition leaderboard [9].

Table 3. Segmentation methods comparison

Method	Average F1-score on validation set, %
DeepLabV3+ with ResNet-101 backbone	**84,3**
Conditional Random Field classification (SVL_1)	77,8
Chessboard segmentation with 5x5 raster size (SVL_3)	77,2

It is clear, that DNN approach, provided in this paper, outperforms classical computer vision algorithms in terms of F1-score. This is mainly due to graph-based models' incorrect labeling of small objects such as cars and clutter. Classical computer vision algorithms tend to merge such objects with background. DNNs on the contrary tend to correctly classify pixels of small objects. This can be further improved by applying class weighting for small object classes while training DNN.

Model used in this paper belongs to DeepLab family. They make extensive use of convolutions with large receptive field to improve context extraction. DeepLabV3+ incorporates several approaches of neural network construction. It uses Pyramid Pooling with expanded convolutions as in DeepLabV3 (Fig. 2(a)). This allows for efficient information extraction from the entire image. It is also combined with another widely

used method of encoder-decoder feature transfer (Fig. 2(b)), which allows for more accurate restoration of original image resolution. This results in hybrid architecture, shown in Fig. 2(c) and Fig. 3.

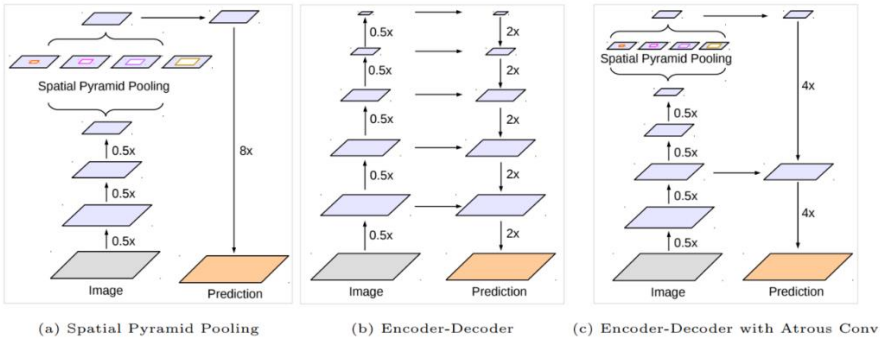

Fig. 2. Segmentation architecture comparison

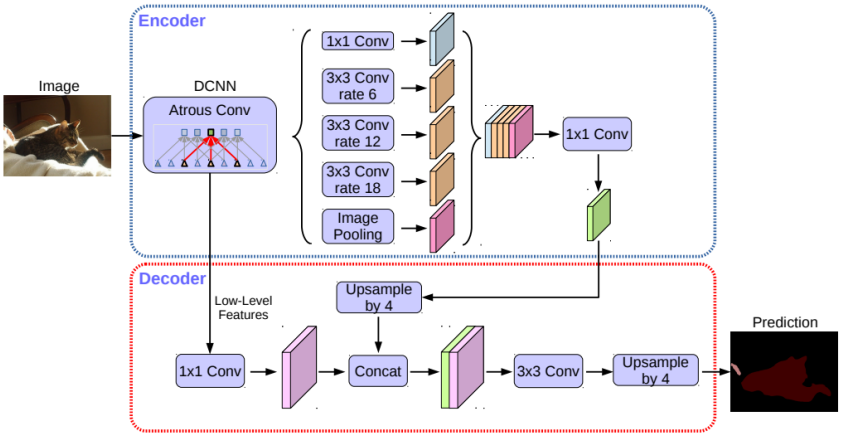

Fig. 3. DeepLabV3+ model

In 2015 Microsoft introduced new deep convolutional network architecture – ResNet (Residual Network). ResNet-34 model is shown in Fig. 4.

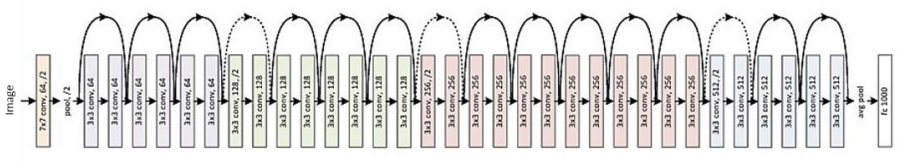

Fig. 4. ResNet-34 model

When training deep neural networks, most encounter a significant problem: with increasing depth of the network, accuracy first increases and then deteriorates rapidly. This is due to the vanishing gradients of the loss function during back propagation. To solve this problem, authors propose to use blocks with the skip-connection operation. In Fig. 5 2 types of commonly used blocks are shown. The second type of blocks is used in deeper architectures, for example, ResNet-101, to reduce the number of network parameters. Such blocks prevent vanishing gradients and allow building deeper networks. Thus, in this work we used the DeepLabV3 + architecture with a network-decoder ResNet-101 from the ResNet family.

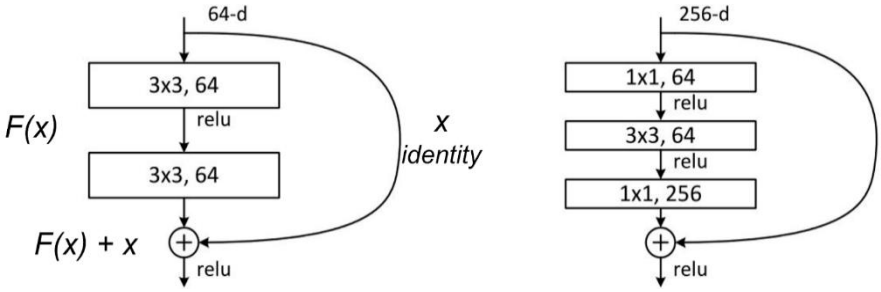

Fig. 5. Residual blocks

3 Experiments

3.1 Data preparation

As mentioned earlier, the resolution of the original images is 6000x6000 pixels. Such large images are unsuitable for direct processing on GPU. Therefore, some data preparation is needed. Training and validation samples are cut into segments with a resolution of 512x512 pixels. This compromise solution allows you to use multiple images for the gradient step while retaining most of the context. After slicing, 2904 images were obtained. Of these, 2604 are used for training and 300 for validation metrics. Examples of cropped segments and a digital mask are presented in Fig. 6.

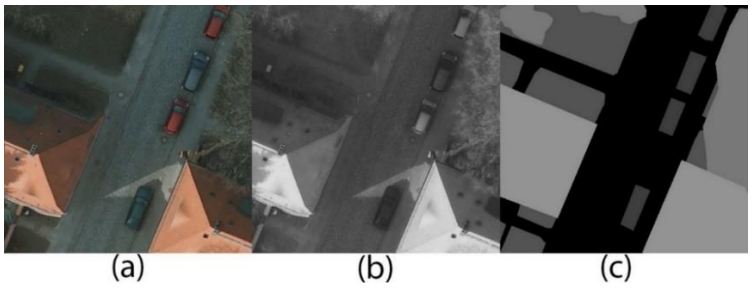

Fig. 6. RGB image segment (a), IR channel image segment (b), digital segmentation map (c)

To conduct experiments with noisy images, several duplicates of 2904 images with varying degrees of noise were created. An ordinary Gaussian noise with an average of 0 was used as a noise model. The standard deviation ranges from 0 to 0,3 with a step of 0,05. Examples of noisy images are shown in Fig. 7.

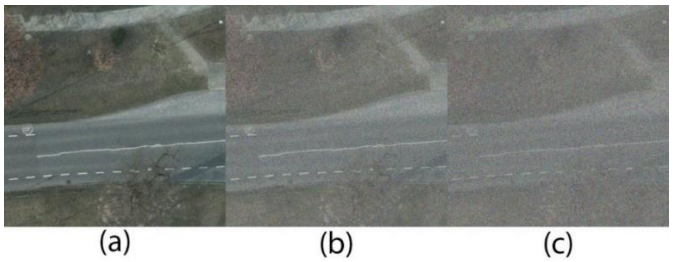

Fig. 7. Noisy segments with standard deviation of 0,05(a), 0,2(b) and 0,3(c)

Due to the small amount of training data, augmentation techniques have also been applied. Different images may have different photometric features and orientation of objects. To increase the generalizing ability of the network, it is reasonable to simulate various conditions by changing the brightness, contrast and orientation of the image. Thus, the following augmentations are applied:

- Random 90 degree turns.
- Random multiplicative brightness changes.
- Random contrast changes.

Augmentation examples are presented in Fig. 8.

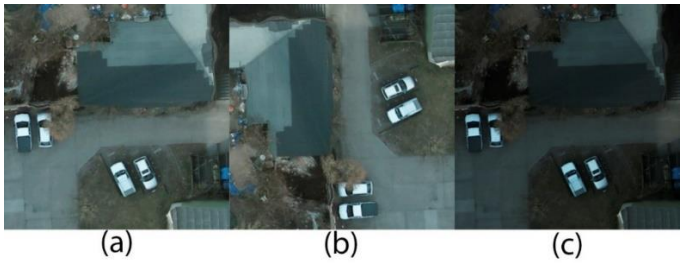

Fig. 8. Source crop (a), random 270º turn (b), random brightness and contrast change(c)

3.2 Model training setup

The model was trained by minimizing the cross-entropy loss function. The cross-entropy (CE) loss function is often used in semantic segmentation problems. Its output signal is a probability value ranging from 0 to 1. The magnitude of the cross-entropy loss function increases when the predicted probability deviates from the target label. In

a binary classification, where the number of classes is two, cross-entropy can be calculated as follows:

$$CE(p, y) = -(y \ln) + (1 - y)\ln(-p), \quad (1)$$

where $y = 0$ for an object of first class and $y = 1$ for the second class, p - probability that the object belongs to the second class. If there are more than two classes, values are calculated for each class and then summed up:

$$CE(p, y) = -\sum_i y_i \ln(p_i), \quad (2)$$

$y_i = 1$ when object belongs to class i, and $y_i = 0$ otherwise, p_i– predicted probability that an object belongs to a class i.

The loss function was minimized using the Adadelta optimizer [10]. It allows for automatic gradient descent parameter optimization in the learning process, and is resistant to noisy gradients.

The quality metric of the model is the Intersection over Union (IoU) metric. It ranges from 0 to 1 and shows same internal volume between two non-empty sets. Formally, for two nonempty sets A and B, the function IoU is defined as:

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

where set A and B are ground truth and predicted segmentation maps. IoU is calculated for each class of the segmentation map, and then averaged over classes. In the training process, the value of the metric is maximized.

Models were trained on batches of eight images with resolution of 512x512 pixels due to limited GPU memory. Each model was trained for 200 epochs on Nvidia GeForce RTX 2080 GPU.

3.3 Noise level and dataset size impact on model quality

To study the influence of training dataset size on the effectiveness of training on noised images, several training sets were created. First of all, the initial training dataset with 2604 images was prepared. After that, 1000 and 1500 images were randomly sampled from it. Thus, three training sets with 1000, 1500 and 2604 images were obtained. This allows for simulation of having different amount of data for training.

By training models on datasets of various sizes, it is possible to obtain dependence of the quality of the trained model versus the amount of data for training. At the same time, it is possible to carry out training on datasets with different noise intensities. As a result, a pair dependence between the amount of data and the noise intensity can be obtained. Studying it, we can draw conclusions about whether it is possible to overcome data noise by increasing the training dataset. The results are presented in Table 4.

Table 4. Segmentation results on clear and noisy data.

Noise intensity	Number of crops in training dataset		
	1000	1500	2604
0	79,32	80,54	81,65
0,05	78,74	79,22	80,60
0,1	77,62	78,54	79,06
0,15	75,32	76,86	77,97
0,2	74,12	75,05	76,44
0,25	73,32	74,12	75,36
0,3	72,03	72,33	73,35

For clarity the dependence of model IoU metric versus noise intensity and number of crops in training dataset were plotted. Resulting plots are presented in Fig. 9-10.

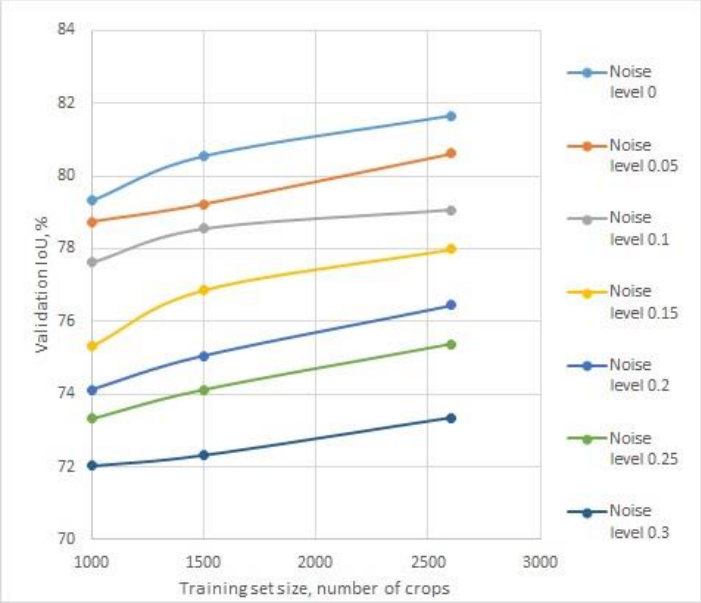

Fig. 9. Validation IoU metric plot versus number of crops in training dataset

The obtained dependence coincides with the expected one. An increase in number of training crops allows model to overcome data noise. The repeatability of the experiment is also worth noting, as model IoU reduction manifested itself on all noise levels.

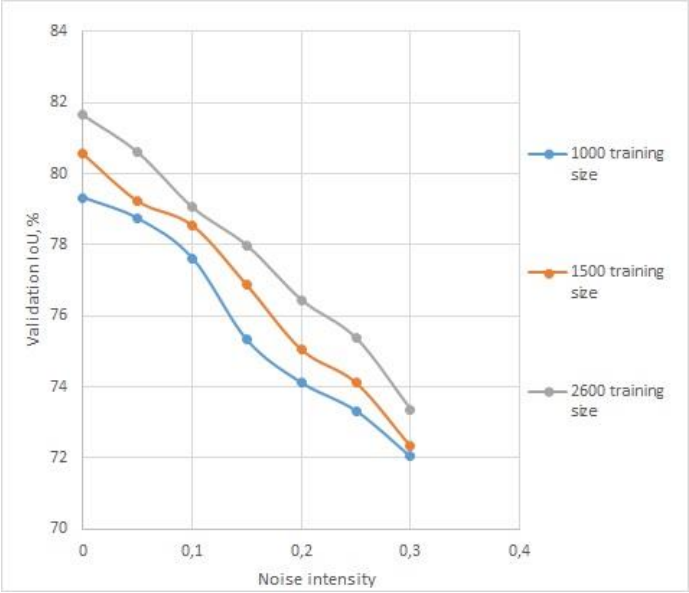

Fig. 10. Validation IoU metric plot versus noise intensity

In Fig. 10 the expected dependence can also be observed. Increase in noise intensity provokes decrease in IoU metric of the model. However, expanding training dataset can reduce the negative impact of noisy data on training process. Examples of model predictions trained on noise with intensity of 0,3 are presented in Fig. 11.

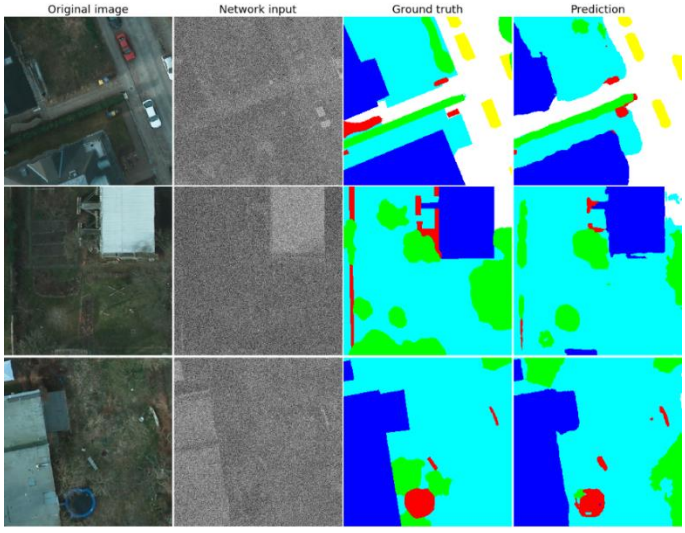

Fig. 11. Original image, noised image feeded into network ground truth and predicted segmentation maps

3.4 Image aggregation

One of possible ways of noise reduction (following CLT) is obtaining a set of noisy variable observations and averaging the results. One can perform similar process for noisy images. Suggested we may have a set of images with similar viewpoint, we tried to imitate such a noise reduction for semantic segmentation. For simulation each training image was duplicated. After that random noise was applied to each instance. While training, each image was loaded along with its duplicates. Images were pixel-wise averaged and resulting averaged image was fed to network input. We refer to such action as image aggregation. Both network training and validation pipeline had image aggregation embedded. The results of training with aggregation pipeline are presented in Table 5.

Table 5. Mean aggregation results

Noise intensity	Number of crops in training dataset	Number of aggregated images	Validation IoU, %
0,3	2604	1	73,35
		3	74,66
		5	75,90

The obtained results suggest that image aggregation techniques can improve model performance on noisy data. This is due to noise lessening capabilities of mean aggregation. Quantitatively we can compare 5-image aggregation with noise reduction of about 0,1.

4 Discussion

Further development of the method of reducing data noise influence can be based on the following approaches:

1. Ensemble models. If computing resources are available, several models can participate in the final prediction. For this, the final predictions of all models are averaged pixel by pixel. Each model can be trained with data with different noise levels. Ensembling such models will increase the generalizing ability of predictions regardless of the noise intensity in the image.
2. Knowledge distillation. One of the ways of increasing the generalization ability of models is the knowledge transfer. Instead of explicitly transferring knowledge by training the model with images with a given noise intensity, one can train teacher models at different noise intensities. After that, when teaching the student's model on data with various noise intensity, the distillation loss function is added to the main loss function, which is responsible for the deviation of the student's predictions from the teacher's predictions. Thus, knowledge about the correct recognition of images of a given noise can be implicitly transferred to the student model.

5 Conclusion

This paper demonstrates the ability of neural network-based segmentation algorithms to operate under extreme distortion conditions. Experimentally acquired dependence of the model validation metric on available training data and data noise level was studied. The experiments showed that additional training data allows to compensate the higher noise level in images and achieve same values of accuracy as on cleaner data. We can draw an analogy with how increase in available data can allow network to learn more classes or generalize better. Mean image aggregation technique have also proven useful in noisy image segmentation labeling. The results of the study shows the possibility of neural networks usage in complex industrial problems where collecting high-quality data is difficult, or when noise levels in data make recognition a difficult task even for human operator.

References

1. 2D Semantic Labeling Contest - Potsdam, http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html.
2. Long, J. et al.: Fully Convolutional Networks for Semantic Segmentation. (2015).
3. He, K. et al.: Deep Residual Learning for Image Recognition. (2015).
4. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. (2015).
5. Badrinarayanan, V. et al.: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. (2016).
6. Ronneberger, O. et al.: U-Net: Convolutional Networks for Biomedical Image Segmentation. (2015).
7. Chaurasia, A., Culurciello, E.: LinkNet: Exploiting encoder representations for efficient semantic segmentation. (2017).
8. Chen, L. et al.: DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. (2017).
9. ISPRS Semantic Labeling Contest (2D): Results, http://www2.isprs.org/commissions/comm2/wg4/potsdam-2d-semantic-labeling.html.
10. Zeiler, M.: ADADELTA: An Adaptive Learning Rate Method. (2012).