INVESTIGATIONS ON SKIP-CONNECTIONS WITH AN ADDITIONAL COSINE SIMILARITY LOSS FOR LAND COVER CLASSIFICATION

C. Yang *, F. Rottensteiner, C. Heipke

Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover - Germany
{yang, rottensteiner, heipke}@ipi.uni-hannover.de

Commission III. WG III/7

KEY WORDS: land cover classification, CNN, aerial imagery, skip-connections, cosine similarity loss

ABSTRACT:

Pixel-based *land cover* classification of aerial images is a standard task in remote sensing, whose goal is to identify the physical material of the earth's surface. Recently, most of the well-performing methods rely on encoder-decoder structure based convolutional neural networks (CNN). In the encoder part, many successive convolution and pooling operations are applied to obtain features at a lower spatial resolution, and in the decoder part these features are up-sampled gradually and layer by layer, in order to make predictions in the original spatial resolution. However, the loss of spatial resolution caused by pooling affects the final classification performance negatively, which is compensated by *skip-connections* between corresponding features in the encoder and the decoder. The most popular ways to combine features are element-wise addition of feature maps and 1x1 convolution. In this work, we investigate *skip-connections*. We argue that not every *skip-connections* are equally important. Therefore, we conducted experiments designed to find out which *skip-connections* are important. Moreover, we propose a new cosine similarity loss function to utilize the relationship of the features of the pixels belonging to the same category inside one mini-batch, i.e. these features should be close in feature space. Our experiments show that the new cosine similarity loss does help the classification. We evaluated our methods using the Vaihingen and Potsdam dataset of the ISPRS 2D semantic labelling challenge and achieved an overall accuracy of 91.1% for both test sites.

1. INTRODUCTION

The goal of land cover classification is to assign a class label for each image pixel so that the physical material of its surface (e.g. *grass, asphalt*) is identified. The pixel-based classification (*semantic segmentation* in computer vision) of images has been tackled by supervised methods. Recently, Convolutional Neural Networks (CNN) variants have mostly been applied for this task, in particular fully convolution networks (FCN, Long et al., 2015), sometimes using encoder-decoder architectures (e.g. Noh et al., 2015; Ronneberger et al., 2015; Badrinarayanan et al., 2017). Variants of these networks have also been applied for land cover classification while using aerial images as input, e.g. (Audebert et al., 2018; Marmanis et al., 2018; Maggiori et al., 2017). A remaining problem is the poor delineation of boundaries due to the loss of spatial resolution caused by the pooling layers. Many strategies have been designed to solve that problem. For instance, Sherrah (2016) used dilated convolutions to avoid pooling; Marmanis et al. (2018) extracted boundaries explicitly and considered this information in the CNN. Another promising strategy is to use skip-connections, i.e. upsampling low resolution feature maps and adding high resolution features from the encoder part of the CNN (Marmanis et al., 2018; Audebert et al., 2018). Element-wise addition of feature maps is the most popular method of combination. Yang et al. (2019) have shown successfully how the optimal combination of high-resolution features and upsampled ones can be learned in the form of 1x1 convolutions to combine the feature maps.

In this paper, we investigate the question whether all skip-connections between convolution blocks in the encoder part and corresponding blocks in decoder part are equally important. To do so, we compare different network variants with different sets of skip connections, removing one set of connections after the other one starting from the outermost ones (the ones relating the information at the highest spatial resolution). In this way we obtain a best-performing architecture for land cover classification. We also discuss the contribution of skip-connections to the classicisation near object boundaries by evaluating the classification performance for pixels inside boundary areas and outside boundary areas separately.

Moreover, the standard loss function for optimizing a CNN is cross-entropy, which tries to make the distribution of predictions approach the true distributions of categories. However, inside one mini-batch, it does not take into account the relationship between pixels belonging to the same category. It is obvious that the features of these pixels should be similar and, thus, close to each other in feature space. During training, we know the true labels of all pixels in one mini-batch, thus we can add an additional constraint on the features of the pixels belonging to the same category, so that their features are to be similar. To achieve this aim, we first calculate the mean feature vector of these pixels, and then calculate the cosine similarity between each feature and the mean feature vector, which is to be maximized. By designing the cosine similarity loss, we are inspired by the focal loss (Lin et al., 2017) to apply a penalty on the well-predicted pixels, so that their contributions to the loss are supressed. Therefore, our cosine similarity loss focuses on pixels that are hard to be classified.

In this paper, we use high-resolution aerial imagery and derived data such as a Digital Surface Model (DSM) and a Digital Terrain Model (DTM) as data source. We apply an encoder-decoder network for land cover classification, where the encoder consists of two branches. The first branch requires images of three bands (e.g. RGB) as input and the second branch requires a composite image (e.g. consisting of the normalised DSM (nDSM) and the red and infrared bands of an image) as input. The two branches

* Corresponding author.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-3-2020-339-2020 | © Authors 2020. CC BY 4.0 License.

are fused at the beginning of the decoder part. The scientific contributions of this paper can be summarized as follows:

- We investigate the skip-connections to differentiate important skip-connections and non-important ones, which results in an optimized CNN architecture.
- We investigate the contribution of skip-connections to the classification performance, showing that the strongest improvement is due to skip-connections between layers at coarse spatial resolutions.
- Beyond the cross-entropy loss, we propose a new loss, the cosine similarity loss, to exploit the inherent relationship of pixels belonging to the same category.

For all the tasks, we conduct experiments using the Vaihingen and Potsdam datasets of the ISPRS 2D semantic labelling challenge. In section 2, we give a review of related work. Our approaches for land cover classification are presented in section 3. Section 4 describes the experimental evaluation of our approach. Conclusions and an outlook are given in section 5.

2. RELATED WORK

The goal of land cover classification is to predict class labels at pixel-level for input images. Recently, this task has been solved by applying CNN variants which can directly deliver dense predictions, e.g. FCN (Long et al., 2015 or encoder-decoder based networks (Noh et al., 2015). In these networks, convolution and pooling operations are applied to the input image, resulting in lower spatial resolution signal maps, which are then up-sampled to the full resolution of input image for dense prediction. In (Long et al., 2015) the upsampling from the lowest spatial resolution to the full one is performed in one step, whereas encoder-decoder networks apply a decoder in a structure that is symmetric to the one of the encoder to upsample the low resolution feature map, e.g. *SegNet* (Badrinarayanan et al., 2017) and *U-Net* (Ronneberger et al., 2015), applying end-to-end learning of all parameters. In these networks, pooling is applied mainly to enlarge the receptive field to incorporate more context information in an implicit way. One main disadvantage caused by pooling is the loss of spatial resolution, leading to inaccurate object boundaries. Many authors apply skip-connections that directly connect feature maps from the encoder to their corresponding counterparts in the decoder to mitigate this problem e.g. (Long et al., 2015; Zhao et al., 2017). In land cover classification, variants of such networks have been used and achieved promising results. Marmanis et al. (2018) apply a Holistically-Nested Edge Detection (HED) framework (Xie et al., 2017) to extract edge maps from aerial images. Subsequently, the edge maps and the aerial images are combined, serving as input for *FCN* and *SegNet* for dense prediction. Although they achieve good results, they suffer from many training stages and a huge number of parameters. Audebert et al. (2018) investigate *SegNet* and *ResNet* (He et al., 2016) and the integration of multispectral and height information in one model, and achieve promising results. Both methods just cited use skip-connections by a simple elementwise addition of feature maps (Long et al., 2015). Thus, the combination of the features of different resolution cannot be learned. Maggiori et al. (2017) propose a method to learn feature combinations: first, they concatenate feature maps of different resolutions, and then they convolve the concatenated maps with 1 x 1 filters. All methods mentioned so far apply skip-connections solely before the classification layer. In a symmetric encoder-decoder structure, the feature maps of the encoder part can be utilized to enrich the representation in the decoder part, e.g. *U-Net* (Ronneberger et al., 2015), where the skip-connections are introduced between the last convolutional layers in corresponding encoder and decoder convolution blocks symmetrically. They only concatenate the feature maps for further processing. Yang et al. (2019) combined the ideas of Ronneberger et al. (2015) and Maggiori et al. (2017) by building a structure similar to *U-Net*, but concatenating the outputs of all convolutional layers at each resolution and using 1 x 1 convolutions to learn the combination of encoder and decoder features. A question that has not been investigated so far to the best of our knowledge is whether all skip connections are equally important and which skip connections have the highest impact on the classification results.

Up to now, a similarity loss has mainly been applied to explore the relationship between samples consisting of pairs. Hadsell et al. (2006) proposed the contrastive loss to minimize the Euclidian distances of similar pairs and maximize the Euclidian distances of dissimilar pairs, with the goal to yield a representation where the Euclidean distance can be used to measure the similarity of image pairs. Another example using Euclidian distance is Hoffer et al. (2015) where the authors proposed the triplet loss to learn representations that are useful for tasks such as image retrieval. Hoffer et al. build triplets consisting of positive and negative pairs and construct a loss that draws the feature vectors of positive pairs close to each other while pushing the feature vectors of negative pairs away from each other. However, the Euclidian distance of two feature vectors is unbounded. To obtain a result that is normalized between -1 and 1, the cosine similarity has been proposed. For instance, Yi et al. (2014) proposed the binomial deviance loss based on cosine similarity for person re-identification by employing a Siamese network, and achieved very promising results. This motivates our application of the cosine similarity to measure similarity of two feature vectors. Wang et al. (2019) proposed a framework to generalize the losses mentioned so far. All of them are applied in a pair-wise context to force the network to learn a good representation of the object. The similarity loss proposed in this paper is different from all losses mentioned so far because it leverages the available information about the class labels of the objects. Using this information, we want features of objects belonging to the same category to be similar. This is achieved by an additional cosine similarity loss for the objects belonging to the same category to make their features similar to the centroid of all feature vectors of that class.

3. NETWORK

3.1 Network architecture

3.1.1 Network architecture: The network architecture used in this paper, referred to as *FuseNet* (Fig. 1), requires two different input images, each of size 256 x 256 pixels with three bands. In the encoder phase, two separate branches are applied on the two input images to extract features, and then the features of the two branches are fused by 1x1 convolutions before decoding. In each encoder branch, there are four convolution blocks, each consisting of three convolutional layers followed by batch normalization (BN; Ioffe et al., 2015) and a rectified linear unit (ReLU) for non-linearity. At the end of the block, there is a max-pooling layer. Symmetrically, the decoder part consists of four blocks, each starting with an upsampling layer that applies bilinear interpolation, followed by three convolutional layers, batch normalization and a ReLU unit. The filter size of each convolution is 3 x 3. Optionally, at the end of each convolution block in the decoder part, there may be *skip-connections*; network variants differing by the type and number of skip-connections are described in section 3.2. Finally, to predict the class labels at the

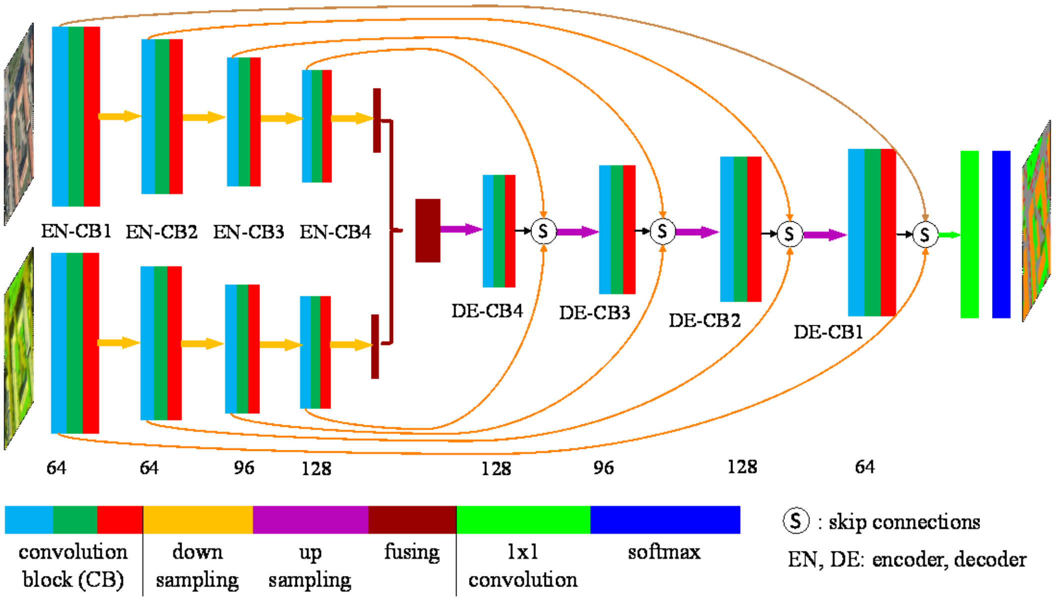

Figure 1: The network architecture (*FuseNet*).

resolution of the input image, there is a 1 x 1 convolutional layer converting the output of the previous layer to a vector of M class scores for each of the $H \times W$ pixels of the input image, where M denotes the number of classes to be differentiated. For each pixel i of the image to be classified, this results in a vector $\boldsymbol{z^i} = (z_1^i, \dots, z_M^i)^T$ of class scores, where $\mathbb{C} = \{C_1, \dots, C_M\}$ is the set of land cover classes and z_c^i is the class score for class C_c. These class scores are normalised by a softmax function delivering the posterior probability $P_i(C_c|x)$ for pixel i to take class label C_c given the image data x:

$$P_i(C_c|x) = softmax(\boldsymbol{z^i}, C_c) = \frac{exp(z_c^i)}{\sum_{l=1}^{M} exp(z_l^i)}, \qquad (1)$$

3.1.2 Skip-connections: The structure of the skip-connections used in this paper is shown in Fig. 2. It is an extension of the learnable skip-connection of in (Yang et al., 2019). When linking encoder and decoder blocks at level N, we take the feature maps delivered by all convolutional blocks of that level. We denote a single feature map (a 2D array of height H and width W) by $\boldsymbol{f^s}$, where s is an index of the feature map that runs over all feature maps delivered by all convolutional blocks. We apply a 3 x 3 depth-wise convolution to every feature map:

$$\boldsymbol{v^s} = ReLU(\omega_s * \boldsymbol{f^s} + b_s)\,. \qquad (2)$$

In equation 2, $\boldsymbol{v^s}$ is the output feature maps, ω_s and b_s are the parameters to be learnt, and the symbol * represents convolution. Note that these convolutions are only applied if a skip connection is established. Now we follow (Yang et al., 2019) and concatenate the feature maps $\boldsymbol{v^s}$ to form a 3D tensor **V** whose dimension is H x W x S, where S is the total number of feature maps concatenated in **V**. After concatenation, a set of D 1x1 convolutions is used to deliver D combined feature maps $\boldsymbol{g^d}$, where d is the index of the d^{th} feature map. Denoting the element at position (r,c) of feature maps $\boldsymbol{v^s}$ and $\boldsymbol{g^d}$ by $v_{r,c}^s$ and $g_{r,c}^d$, respectively, elements of the combined feature map are computed according to:

$$g_{r,c}^d = ReLU\left(\sum_{s=1}^{S} \theta_d^s \cdot v_{r,c}^s + b_d\right), \qquad (3)$$

where θ_d^s and b_d are parameters to be learnt. The feature maps $\boldsymbol{g^d}$ form the input to the first convolutional layer of the next decoder block in Fig. 1.

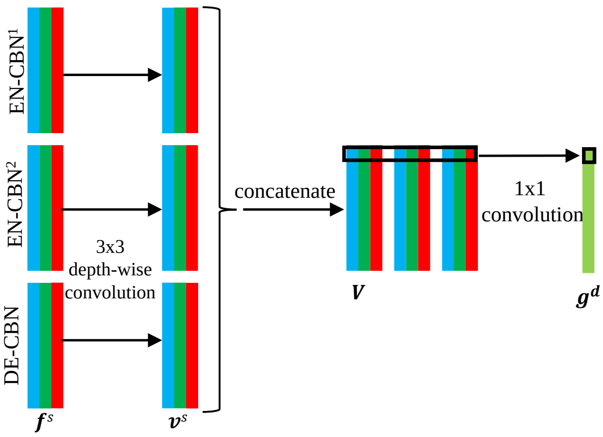

Figure 2: Skip-connections. EN-CBN^1, EN-CBN^2: feature maps delivered by all convolutional blocks of the encoder blocks N of the upper and lower branches in Fig. 1, respectively. DE-CBN: feature maps of all convolutional blocks of the decoder block N in Fig. 1. The colours indicate from which convolutional block in Fig. 1 a feature map $\boldsymbol{f^s}$ was delivered. $\boldsymbol{v^s}$: feature map derived from $\boldsymbol{f^s}$ by convolution; **V**: concatenation of all feature maps; $\boldsymbol{g^d}$: a feature map after 1 x 1 convolution.

3.1.3 Network variants: We developed some variants of the network described in Section 3.1.1 to investigate the effect of different definitions and configurations of skip-connections. The first network variant connects all convolution blocks of the encoder to their counterparts in the decoder (*FuseNet-All*). To investigate the importance of each group of skip-connections, we build different variants in which more and more skip-connections are removed. The names of these variants are shown in Table 1.

variant	**skip connections**
FuseNet-All	Connect all convolution blocks between encoder and decoder
FuseNet-None	no skip connections
FuseNet-234	connect convolution blocks 2, 3 and 4
FuseNet-34	connect convolution blocks 3 and 4
FuseNet-4	only connect convolution blocks 4
FuseNet-123	connect convolution blocks 1, 2 and 3
FuseNet-1	only connect convolution blocks 1

Table 1: Network variants with different skip-connections.

3.2 Training

3.2.1 Training using the extended focal loss: All parameters of the convolutional layers are learned during in the training process, which is based on stochastic mini-batch gradient descent using backpropagation for computing the gradients. The standard loss we use for training in the experiments related to the structure of the network is the *extended focal loss* (Yang et al., 2019):

$$L_{focal} = -\frac{1}{W\cdot H\cdot N}\sum_{c,i,k}\left[y_c^{ik}\cdot\left(1-P_i(C_c|X_k)\right)^{\gamma}\cdot log\left(P_i(C_c|X_k)\right)\right], \quad (4)$$

where k is the index of an image, X_k is the k^{th} image in the mini-batch and N is the number of images in a mini-batch. The indicator variable y_c^{ik} is 1 if the training label of pixel i in image k is identical to C_c and 0 otherwise, and γ is a hyperparameter. The sum in equation (4) is taken over all potential class labels for all pixels of all images of a mini-batch.

3.2.2 Cosine similarity loss: In addition to the extended focal loss (equation 4), we propose an extension based on feature similarity. In a mini-batch, the pixels belonging to the same category should be close to each other in feature space, i.e. their features of the last layer (the one before the softmax function (equation 1) is applied) should be similar. Thus, an additional constraint on these features may support the learning procedure to deliver a better classifier.

The implementation of this idea requires four steps, which are performed for all classes. First, for each pixel i of a class C_c in the current minibatch, the raw class scores $\boldsymbol{z^i}$ according to equation (1) are passed through the ReLU activation function, resulting in feature vectors $\boldsymbol{a^i}$, i.e. $\boldsymbol{a^i} = ReLU(\boldsymbol{z^i})$. After that, the mean feature vector $\boldsymbol{u^c}$ of all the N_c pixels of that class is determined:

$$\boldsymbol{u^c} = \frac{1}{N_c}\sum_i^{N_c}\boldsymbol{a^i}. \quad (5)$$

In the third step, the cosine similarity of each feature vector $\boldsymbol{a^i}$ and the mean feature vector $\boldsymbol{u^c}$ is computed:

$$cos\left(\boldsymbol{a^i},\boldsymbol{u^c}\right) = \frac{\boldsymbol{a^i}\cdot\boldsymbol{u^c}}{\|\boldsymbol{a^i}\|_2\|\boldsymbol{u^c}\|_2}. \quad (6)$$

Finally, we want to maximize the cosine similarities of all pixels over all classes inside the mini-batch. During maximization, we also apply a penalty term which is inspired by the focal loss (Lin et al., 2017). For pixels that are well predicted (i.e. with high probability for belonging to their correct class C_c), the losses are suppressed, so that the loss focuses on pixels which are hard to be classified. As in the original focal loss, we expect this penalty term to accelerate the training procedure and deliver better classification performance. Thus, the *cosine similarity loss* function is defined according to:

$$L_{cos} = \frac{1}{W\cdot H\cdot N}\sum_c\sum_i^{N_c}\left[\left(1-P_i(C_c)\right)^{\varsigma}\cdot max(1-cos(\boldsymbol{a^i},\boldsymbol{u^c})-m,0)\right], \quad (7)$$

where m is a margin to control the similarity and ς is a hyperparameter to control the influence of the penalty term. If this loss function is used in training, it is combined with the extended focal loss L_{focal}according to equation 4, so that the combined loss that is optimized in the experiments involving $L_{cos}$is

$$L_{combined} = L_{focal} + L_{cos} \quad (8)$$

3.2.3 Hyperparamete settings: In the training procedure, we apply weight decay with 0.0005, a step learning policy. The learning rate was set to 0.01 and decreased to 0.001 after 15 epochs in a total of 30 epochs training. The mini-batch size is 4. In the *extended focal loss*, the hyperparameter γ in equation 4 is set to 1 for all experiments. For all experiments in which the combined loss according to equation 8 is applied, the hyperparameters of the cosine similarity loss (equation 7) are set to ς = 1 and m = 0.2.

3.2.4 Implementation: All networks are implemented based on the tensorflow framework (Abadi et al., 2015). We use a GPU (Nvidia TitanX, 12GB) to accelerate training and inference.

4. EXPERIMENTS

4.1 Test Data und Test Setup

4.1.1 Test Data: Our approaches for classification of land cover are evaluated on the Vaihingen and Potsdam datasets of the ISPRS 2D semantic labelling challenge. The former one consists of 33 colour infrared (CIR) images with a Ground Sampling Distance (GSD) of 9 cm, whereas the latter one consists of 38 orthophotos (RGB-IR) with a GSD of 5 cm. In addition, nDSMs provided by Gerke (2015) were available. Following the benchmark protocol, in Vaihingen 16 images with known reference are used for training and the rest (17) for testing, and in Potsdam 24 images with known reference are used for training and the rest (14) for testing. There are six land cover classes: *impervious surface (imp. surf.), building (build.), low vegetation (low veg.), tree, car* and *clutter* (Wegner et al., 2017).

4.1.2 Test setup: We extract windows of 256 x 256 pixels with an overlap of 128 pixels in both spatial dimensions from the training images, which results in 4426 training patches in Vaihingen and 50784 training patches in Potsdam. In training, we applied data augmentation by rotations of 90°, 180°, 270°, horizontal and vertical flipping (i.e. 6 times more data). In Vaihingen, due to the lack of a blue band, we use CIR instead of RGB images as the first input and a composite of the red and near infrared bands and the nDSM (RID) as the second input. In Potsdam, RGB and the composite RID serve as the inputs. During inference, the class labels for a patch of 256 x 256 pixels are predicted six times for the original image and variants that are flipped and rotated as the training images, and the probabilistic scores are multiplied to obtain a combined score for classification.

We performed two sets of experiments. The first set was dedicated to the comparison of the different network variants defined in Table 1. Here we used the extended focal loss (eq. 4) for training in all cases. For Vaihingen, we trained and tested all variants of Table 1, while for Potsdam, we selected the two variants performing best in Vaihingen (*FuseNet-All* and *FuseNet-234*) as well as the variant without skip connections (*FuseNet-None*). In the second set of experiments, we evaluated the effectiveness of the cosine similarity loss. In this set, we used the loss function $L_{combined}$ (equation 8) for training. We only compared the network variants performing best on the basis of the extended focal loss (*FuseNet-234*) and the variant without skip-connections (*FuseNet-None*); the variants trained on the basis of that loss are identified by an asterisk, thus they are denoted by *FuseNet-None** and *FuseNet-234**, respectively.

For evaluation, there are two reference datasets: the full reference contains class labels for all pixels, while the eroded reference does not consider the pixels near object boundaries (erosion by a circular disc of 3-pixel radius). For a comparison of variants, we use the full reference to compute the Overall Accuracy *OA*, i.e., the percentage of pixels whose class label determined by the CNN is identical to the reference, and the class-specific *F1* scores, i.e. the harmonic mean of precision and recall determined on a per-pixel level. We also determine the average F1 score (*avg. F1*) as the mean of *F1* over all classes. For a comparison with the results of the ISPRS benchmark (Wegner et al., 2017), we report *OA* and *F1* also for the eroded reference, i.e. just considering pixels that are not near to an object boundary. To gain deeper insights into the behaviour of the networks near object boundaries, we additionally determine the OA just for pixels near object boundaries (i.e., for the pixels without class labels in the eroded reference). In Vaihingen, 9.1% of the pixels are inside a boundary area and 90.9% pixels are outside boundary area; In Potsdam, 7.2% of the pixels are inside a boundary area and 92.8% pixels are outside boundary area.

4.2 Evaluation: Comparison of Network Variants

In this section, we report the results of the first set of experiments, conducted to compare different network variants (Table 1). The results of the evaluation of all experiments are shown in Table 2. Fig. 3 shows some exemplary results for some network variants and both datasets. In general, the CNN works very well on the both datasets in all variants, with an OA of more than 87% and an average F1 score of more than 71%. However, there are also areas which all networks fail to classify correctly (red ellipses in Fig. 3).

4.2.1 Comparison of network variants: Comparing the results achieved by all network variants for Vaihingen (see Tab. 2), several observations can be made:

1) Comparing *FuseNet-234* and *FuseNet-All*, the former delivers slightly better results in terms of average F1 score while the OA is identical. The largest improvement in F1 occurs for class *clutter* (2.1%). The removal of the skip-connections between the blocks of the highest resolution has no impact on the OA in the boundary areas (OA_b). This would indicate that the quality of the classification in boundaries is not positively affected by this specific connection.

2) Additionally removing the skip-connection of convolution block 2 (*FuseNet-34*) still delivers comparable results to variants *FuseNet-234* and *FuseNet-All* in terms of *OA*, yet with a decrease of average F1 score of 0.8%, mostly due to the worse performance for the class *clutter*, which is very heterogeneous and for which there are only few samples.

3) Additionally removing the skip-connection of convolution block 3 (*FuseNet-4*) results in a negligible decrease of *OA* and a somewhat larger one for the average F1 score (1.6% compared to *FuseNet-234*), again mostly for the class *clutter*. Yet again, the classification accuracy in the boundary regions seems to be hardly affected, indicated by a slight decrease of 0.6% OA_b compared to *FuseNet-234*.

4) When not using any skip-connections at all (*FuseNet-None*), there is again only a small difference in *OA* compared to *FuseNet-234* (0.8%), and the difference in the boundary areas is of a similar size. However, there is a more obvious decrease in the F1 scores, particularly for the underrepresented classes *car* (3.8%) and *clutter* (21.3%). This leads to a significant drop in the average F1 score (6.2%).

These results show that skip-connections have hardly any impact on the *OA*, and the general assumption that they improve the quality of the classification near object boundaries is not confirmed, not even for the connections between high-resolution encoder and decoder blocks (blocks 1 in Fig. 1). Nevertheless, skip-connections have a positive impact on the performance of the classifier for underrepresented classes, as indicated by the F1 scores. In this context, the skip-connections of convolution block 4 are most important, which is somewhat counter-intuitive, because it is the block having the coarsest spatial resolution, and *cars*, corresponding to one of the classes that is most affected, have a small spatial resolution. This may be related to the observation made by He et al. (2016) that networks have difficulties in learning identity transformations, so that skip-connections (referred to as *bypass connections* in the reference) can support the training procedure. In general, the best-performing network is *FuseNet-234*, i.e. the network excluding the skip-connections between the outmost convolution blocks, though only by a very small margin.

Are the skip-connections really the most important one? If we compare the results for the variant *FuseNet-123*, which has all

Test Site	Network	*F1* [%]						*avg. F1*	*OA*	OA_{red}	OA_b
		imp. surf.	*build.*	*low. veg.*	*tree*	*car*	*clutter*	[%]	[%]	[%]	[%]
Vaihingen	*FuseNet-None*	89.0	93.5	81.2	86.1	77.7	15.9	73.9	87.3	90.2	58.3
	FuseNet-All	89.8	**94.0**	81.5	**87.2**	**81.1**	47.2	80.1	**88.1**	90.9	59.7
	FuseNet-234	**89.9**	**94.0**	81.6	86.9	80.8	49.3	**80.4**	**88.1**	**91.0**	**59.9**
	FuseNet-34	89.6	93.8	81.6	**87.2**	80.6	43.2	79.3	88.0	90.9	59.6
	FuseNet-4	89.6	93.8	81.2	87.0	80.8	40.4	78.8	87.9	90.7	59.3
	FuseNet-123	89.5	93.4	81.4	87.1	80.8	**50.2**	**80.4**	87.9	90.7	**59.9**
	FuseNet-1	89.5	93.5	80.9	86.2	75.9	0.3	71.5	87.5	90.3	59.1
Potsdam	*FuseNet-None*	90.1	95.1	84.8	85.3	88.8	53.1	82.9	87.8	90.1	58.4
	FuseNet-All	90.8	95.9	**85.5**	**85.7**	90.4	50.5	83.1	**88.6**	**90.7**	**61.2**
	FuseNet-234	**90.9**	**96.0**	85.4	**85.7**	**90.8**	**53.8**	**83.8**	**88.6**	**90.7**	**61.2**

Table 2. Results of land cover classification for different network variants defined in Table 1 using the extended focal loss (eq. 4) for training. *F1*: F1 score, *OA*: Overall Accuracy, both determined on the basis of the full reference; OA_{red}: Overall Accuracy based on the eroded reference; OA_b: Overall Accuracy for pixels in the boundary areas. Best scores per test site and metric are printed in bold font.

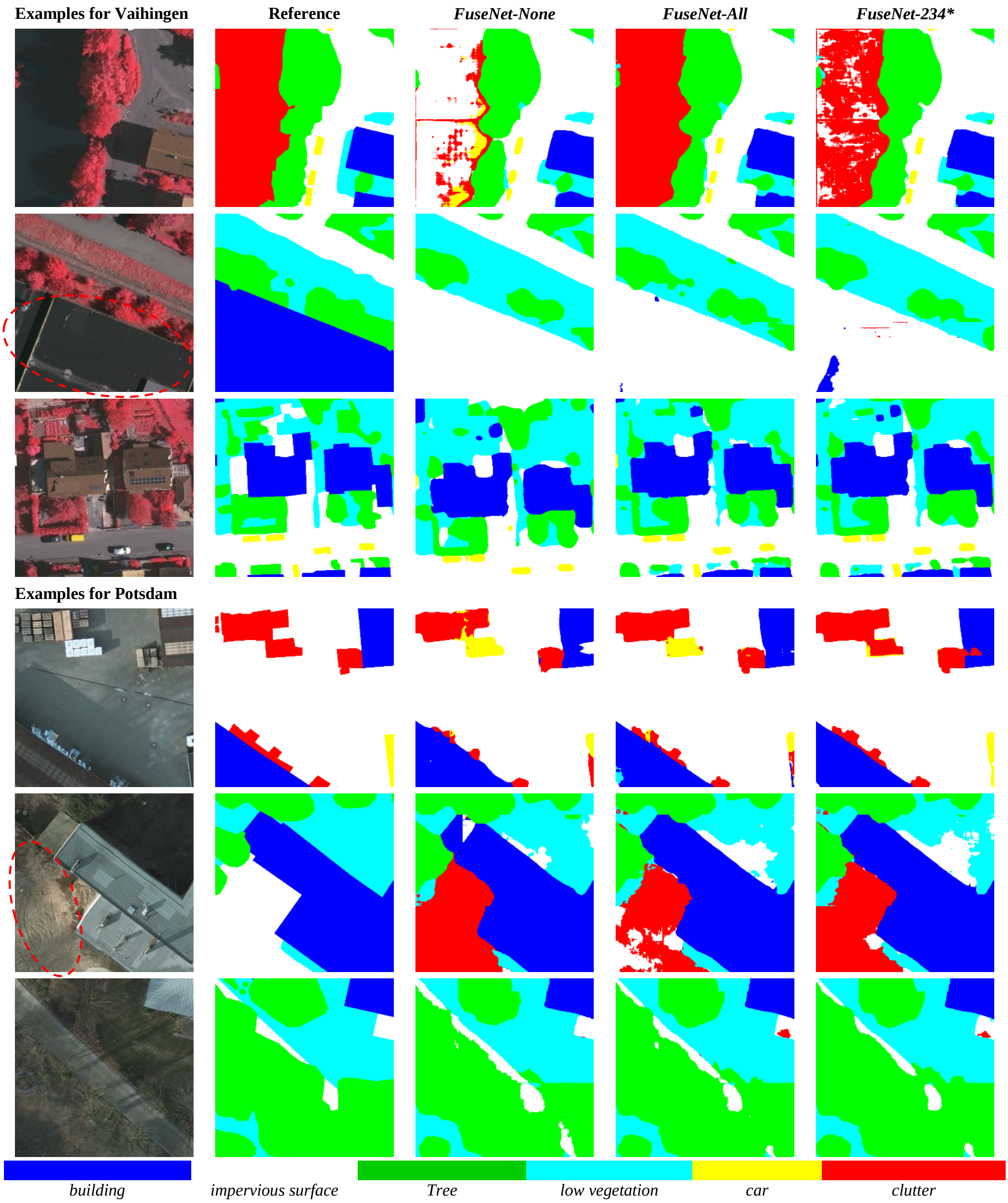

Figure 3: Data and exemplary classification results for both datasets. The first and second columns show the image and the reference, respectively; the other columns show the results for several network variants described in the main text. The colour code is given at the bottom of the figure. Red dashed ellipse: problematic areas for all network variants.

skip-connections except the one at convolution block 4 to the network which only uses this connection (*FuseNet-4*), the OA of both networks is identical, but there is an advantage of *FuseNet-123* in the average F1 score (1.6%), mainly due to a better performance for class *clutter*. However, the training time per epoch of *FuseNet-C-123* is about 2.5 times longer and it also requires more memory. We can say that the skip-connections of convolution block 4 improves the quality almost to the same level as the combination of all other skip-connections in our network while requiring much less computation capacity.

If we apply skip-connections only in convolution block 1 (*FuseNet-1*), the *OA* is nearly identical to the one achieved when not using any skip-connection at all (*FuseNet-None*), and it performs worse in the average F1 score (decrease of 2.4%), again due to problems with the classes *car* and *clutter*. Of course, *FuseNet-1* performs considerably worse than the one using only convolution block 4 (*FuseNet-4*). This indicates that the skip-connections of convolution block 1 does not play a significant role in the classification and underlines that the skip-connections at block 4 are more important.

The results for Potsdam shown in Tab. 2 confirm that using skip-connections improves the results by a margin in the order of about 1% in *OA* and mean F1 score. Here, the difference in the boundary areas is somewhat larger (2%), which indicates that in this case, the skip-connections do contribute to a better classification of object boundaries, though the effect is small. Again, *FuseNet-234* slightly outperforms *FuseNet-All* in terms of average F1 score.

4.2.2 Discussion of the contribution of skip-connections in boundary areas: The primary goal of applying skip-connections is to address the loss of spatial resolution caused by pooling operations. It is frequently expected that using skip-connections would deliver a better boundary delineation, because they add precise location information from high-resolution features. However, our experiments indicate that skip connections between low-resolution layers seem to be more important than those between high-resolution layers. An analysis of the classification accuracies in the boundary areas shows relatively small differences between the variants. In general, the improvements in OA in the boundary regions (OA_b) and in the areas outside the boundary areas (OA_{red}) are of a similar size; in Vaihingen, the improvement of the best variant (*FuseNet-234*) over the worst one (*FuseNet-None*) is 0.8% in OA_{red} and 1.6% in OA_b, indicating that there is a very small relative improvement of accuracy near boundary areas due to the skip-connection. A similar observation can be made when analysing the data for Potsdam (improvement of *FuseNet-234* over *FuseNet-None* by 0.6% outside the boundaries vs. 1.8% near boundaries). Somewhat counter-intuitively, this small improvement does not mainly depend on skip-connections between the high-resolution layers of the network. In summary, skip-connections do have a positive effect on the quality of the classification both in boundary areas and outside these areas. The improvement is larger in the boundary areas, but this effect is very small.

4.3 Evaluation: Comparison of loss functions

The intention of applying cosine similarity loss is to force the features of pixels belonging to the same category being similar, by making the features of the pixels of this category being close to their centroid. Tab. 3 shows the evaluation results of the network variants that were trained using the combined loss of equation 8 and, thus, considering the cosine similarity loss. Compared to *FuseNet-None* (Tab. 2), the results achieved when using the cosine similarity loss (*FuseNet-None**) are improved both in Vaihingen and Potsdam. In terms of OA, the increase is not very large (up to 0.5%). However, the average F1 scores are improved by 3.0% in Vaihingen and 0.7% in Potsdam, mainly due to a better performance for *car* and *clutter*. When comparing the variants using skip-connections (*FuseNet-234** vs. *FuseNet-234*), there are also slight improvement in terms of OA and average F1 score due to the new loss for both test sites, though they are smaller than 1% in all cases. The improvements inside and outside boundary areas due to the cosine similarity loss are of a similar size. In conclusion, the comparison shows that cosine similarity loss does help in the classification of land cover classes. While the improvement in overall accuracy is relatively small, there is a larger impact on the performance of classes that occur rarely in the data.

4.4 Comparison to the state of the art

A comparison with other methods based on the scoreboard of the ISPRS benchmark (Wegner et al., 2017) is shown in Tab. 4. Following the convention of the ISPRS benchmark, the *eroded reference* is used for evaluation; the comparison is based on the variant *FuseNet-234** (listed as "ours" in the table). For *Vaihingen*, the benchmark website only lists two (out of more than 100) contributions that deliver an OA that is better than the one of our method. The OA of *FuseNet-234** is only 0.5% worse than the best one (HUSTW5), yet our method outperforms their average F1 score (without considering the class *clutter*, following the benchmark protocol) by 2.2%, which is mainly due to our huge improvement of the identification of class *car* (more than 13% in term of F1 score). The other method better than ours in term of OA is NLPR3. However, their increase of OA is compensated by decrease of average F1 score (-0.6%). For *Potsdam*, the benchmark website lists three methods delivering better results than ours in term of OA. However, for the individual class *building*, *low vegetation* and *tree*, our method shows the ability of identification in first or second place. In conclusion, we take the results mentioned above as an indication that our method is on par with the current state of the art.

5. CONCLUSION

In this paper, we have proposed a variant of a CNN similar to *U-net* for land cover classification. First, we generated different variants of the network with a different number of skip-connections to investigate the relevance of these skip-connections for the classification performance. Our experiments indicate that skip-connections between the low-resolution layers of the encoder and the decoder might be more important than the ones between the high-resolution layers. In general, the impact on the OA is low, but skip-connections lead to a noticeable improvement in the classification of classes having few samples. We also analysed the contributions of skip-connections in boundary areas, and found that they do have a very small positive effect. Second, we proposed a new cosine similarity loss to push pixels belonging to the same category inside one mini-batch to have similar feature vectors. The land cover classification profits from this loss slightly, but again, the effect is relatively small.

Test Site	Network	F1 [%]						avg. F1	*OA*	OA_{red}	OA_b
		imp. surf.	*build.*	*low. veg.*	*tree*	*car*	*clutter*	[%]	[%]	[%]	[%]
Vaihingen	*FuseNet-None**	89.3	93.6	81.5	86.6	78.8	31.6	76.9	87.6	90.6	58.5
	*FuseNet-234**	**90.1**	**94.0**	**81.8**	**87.1**	**82.2**	**49.7**	**80.8**	**88.3**	**91.1**	**60.2**
Potsdam	*FuseNet-None**	90.6	95.8	85.0	85.6	90.5	54.1	83.6	88.3	90.5	59.4
	*FuseNet-234**	**91.2**	**96.3**	**85.6**	**86.2**	**91.1**	**54.9**	**84.2**	**88.9**	**91.1**	**61.4**

Table 3. Results of land cover classification for different network variants defined in Table 1 using the extended focal loss (eq. 4) for training. *F1*: *F1* score, *OA*: Overall Accuracy, both determined on the basis of the full reference; $\underline{OA_{red}}$: Overall Accuracy based on the eroded reference; $\underline{OA_b}$: Overall Accuracy for pixels in the boundary areas. Best scores per test site and metric are printed in bold font.

Finally, we compare our best-performing method with the state-of-art from the ISPRS benchmark and found its performance to be on par with the best methods reported on the benchmark website.

In future work, we want to verify the current findings regarding skip-connections for other methods of combining the signals from different layers (e.g. element-wise addition). Second, we are going to investigate the cosine similarity loss in more detail by tuning the hyper-parameters and comparing it to loss functions based on other measures of similarity of feature vectors, e.g. on the Euclidian distance.

Network	F1 [%]					$\overline{F1}$ [%]	OA [%]
	imp. surf.	*build.*	*low. veg.*	*tree*	*car*		
Vaihingen							
HUSTW5	93.3	96.1	86.4	90.8	74.6	88.2	91.6
NLPR3	93.0	95.6	85.6	90.3	84.5	89.8	91.2
ours	92.7	95.8	85.2	90.1	88.0	90.4	91.1
Potsdam							
SWJ_2	94.4	97.4	87.8	87.6	94.7	92.4	91.7
HUSTW3	93.8	96.7	88.0	89.0	96.0	92.7	91.6
AMA_1	93.4	96.8	87.7	88.8	96.0	92.5	91.2
ours	93.1	97.3	88.1	88.8	95.2	92.5	91.1

Table 4. Comparison to the state-of-art (with eroded reference). $\overline{F1}$: average F1 score, OA: Overall Accuracy.

ACKNOWLEDGEMENT

We thank the Landesamt für Geoinformation und Landesvermessung Niedersachsen (LGLN), the Landesamt für Vermessung und Geoinformation Schleswig Holstein (LVermGeo) and Landesamt für innere Verwaltung Mecklenburg-Vorpommern (LaiV-MV) for their support of this project.

REFERENCES

Abadi, M. et al., 2015. Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org (accessed 20/01/2020).

Albert, L., Rottensteiner, F., Heipke, C., 2017. A higher order conditional random field model for simultaneous classification of land cover and land use. ISPRS Journal of Photogrammetry and Remote Sensing 130: 63-80.

Audebert, N., Saux, B. L., Lefevre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS Journal of Photogrammetry and Remote Sensing 140: 20-32

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(12): 2481-2495.

Gerke, M., 2015. Use of the stair vision library within the ISPRS 2d semantic labelling benchmark (Vaihingen). Tech. rep., International Institute for Geo-Information Science and Earth Observation.

Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping. IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 1735-1742

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778.

Hoffer E., Ailon N. 2015: Deep metric learning using triplet network. Lecture Notes in Computer Sicence, Vol. 9370, Srpinger, Cham, pp. 84-92

Ioffe, S., Szegedy, C., 2015. Batch Normalization: accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning, pp. 448-456.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431-3440.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P., 2017. Focal loss for dense object detection. IEEE International Conference on Computer Vision (ICCV), pp. 2999-3007

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. High-resolution semantic labelling with convolutional neural networks. IEEE Transactions on Geosciences and Remote Sensing, Vol. 55 (12), pp. 7092-7103

Marmanis, D., Schindler, K., Wegner, J. D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS Journal of Photogrammetry and Remote Sensing 135: 158–172.

Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. IEEE International Conference on Computer Vision, pp. 1520-1528.

Ronneberger O., Fischer P., Brox T., 2015. U-Net: Convolutional neworks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234-241

Wang, X., Han, X., Huang, W., Dong D., Scott, M.R., 2019. Multi-similarity loss with general pair weighting for deep metric learning. IEEE Conference on Computer Vision and Pattern Recognition, pp. 5017-5025.

Wegner, J.D., Rottensteiner, F., Gerke, M., Sohn, Gunho, 2017. The ISPRS labelling challenge. Available in the WWW: http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html (accessed 20/01/2020).

Xie, S.N., Tu, Z.W., 2017. Holistically-Nested Edge Detection. International Journal of Computer Vision, Vol. 125 (3), pp. 3-18

Yang, C., Rottensteiner, F., Heipke, C., 2018: Classification of land cover and land use based on convolutional nerual networks. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. IV-3, pp. 251-258

Yang, C., Rottensteiner, F., Heipke, C., 2019: Towards better classification of land cover and land use based on convolutional nerual networks. ISPRS Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XLII-2/W13, pp. 139-146

Yi, D., Lei, Z., Li, S.Z., 2014. Deep metric learning for practical person re-identification. 22nd International Conference on Pattern Recognition, Stockholm, pp. 34-39

Zhao, H.S., Shi, J.P., Qi, X.J., Wang, X.G., Jia, J.Y., 2017. Pyramid scene parsing network. IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881-2890