SEGMENTATION OF IMBALANCED CLASSES IN SATELLITE IMAGERY USING ADAPTIVE UNCERTAINTY WEIGHTED CLASS LOSS

Benjamin Bischke [1, 2] *Patrick Helber* [1, 2] *Damian Borth* [2] *Andreas Dengel* [1, 2]

[1] TU Kaiserslautern, Germany
[2] German Research Center for Artificial Intelligence (DFKI), Germany

ABSTRACT

We propose a novel loss function for the training of deep Convolutional Neural Networks (CNNs) focusing on land use and land cover classification in remote sensed data. In satellite imagery, object classes are often highly imbalanced leading to poor pixel-wise classification results when using standard training methods only. In this work, we introduce a loss function which leverages the per class uncertainty of the model during training together with median frequency balancing of the class pixels. We evaluate our result on aerial images of the state-of-the-art dataset Vaihingen. We obtain a significant improvement of the F1-Score and pixel accuracy against the standard cross entropy loss on the small car class. The overall F1-Score using a single CNN achieves 89.35% resulting in an error reduction of 21.22% against the baseline.

Index Terms— Deep Learning, Semantic Segmentation, Class Uncertainty Weighting, Satellite Imagery, High-Resolution Imagery

1. INTRODUCTION

The increasing number of satellites constantly sensing our planet and the corresponding availability of remote sensed imagery raises a set of new challenges for the automated analysis of this data source at a large scale. Up-to date information of our earth's surface with detailed land use and land cover classes has potential to change how environmental monitoring, agriculture, forestry, emergency response and urban planning will be done in the future.

In order to support the research in the extraction of features in remote sensed imagery, dedicated datasets have been developed addressing specific use cases. For example, the Multimedia Satellite Task 2017 [2] aims to help relief agencies during flooding events by identifying areas that are affected hardest by such an event. Other large-scale datasets such as the EuroSAT [3] focus on a global land use and land cover classification by using publicly available satellite data from Sentinel 2. The ISPRS Vaihingen 2D semantic labeling contest dataset [1], focuses on the semantic segmentation of urban land use classes on very high resolution airborne data to support urban development.

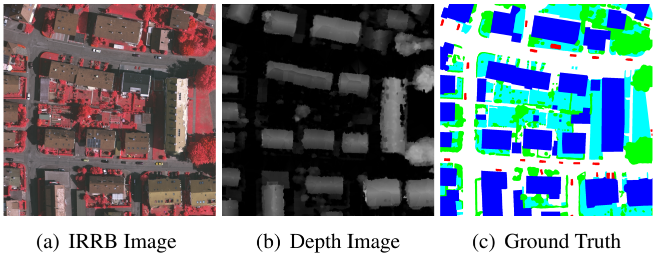

(a) IRRB Image (b) Depth Image (c) Ground Truth

Fig. 1. One sample patch of the same location in different representations: (a) an IRRG-image, (b) a normalized Depth image and (c) the Ground Truth. It can be seen that the object classes in this scene are highly imbalanced. The amount of pixels corresponding to the car class (shown in red color in (c)) is significantly lower compared to pixels related to building or vegetation classes (blue and green color in (c)).

In this paper, we rely on deep convolutional neural networks (CNNs) and the ISPRS Vaihingen 2D semantic dataset. CNNs are currently the major approach for semantic segmentation on RGB images [4, 5] and on satellite/aerial imagery of the remote sensing domain [6, 7, 8]. Their supremacy against traditional approaches [9] is also reflected in the ISPRS Vaihingen challenge where CNN achieve the best performing results. However, one major challenge when training CNNs for semantic segmentation on remote sensed imagery is the presence of a high class imbalance. This can be seen in Fig 1, where only a few pixels are assigned to the *car* class and a high number of pixels are assigned to vegetation class. Training with a standard cross entropy loss yields to poor classification results of small classes, since the gradients of the minority class are computed only for a few pixels. A major approach to overcome this problem of small classes, is the usage of a weighted cross-entropy loss. The idea is to assign small classes a higher cost value in case of miss-classifications compared to bigger classes. The work by

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the DGX-1 used for this research and the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) for providing the Vaihingen data set [1]. This work was partially funded by the BMBF Project DeFuseNN (01IW17002).

Kampffmeyer [7] showed that by weighting the cross entropy loss with median frequency balancing (MFB) [10], the segmentation accuracy for small classes in urban remote sensing can be effectively improved. While such a weighting yields to more balanced classification result by boosting small classes, larger classes tend to not improve after a certain threshold.

In this paper, we analyze the potential of using a novel loss function to leverage on the one hand side a median frequency balanced weighted loss and on the other hand the uncertainty of the network classes with respect to particular classes. The contributions of this paper can be summarized as follows:

- We propose a new loss for high resolution satellite images with imbalanced classes. The loss takes the frequency of class pixels and class uncertainty of the model into account.
- Results on the Vaihingen dataset [1] show an improvement of about 4 percentage points on the F1-Score against the standard cross entropy loss, which corresponds to error reduction of 21.22% percent.

2. APPROACH

In this section, we describe the details of the network architecture and the proposed loss function.

2.1. Fully Convolutional Network Architecture

Our method for semantic segmentation of land use and land cover classes in satellite imagery, relies on a Fully Convolutional Network (FCN). We adopt the originally proposed FCN [11], which is based on the VGG16 network [12], with the recently introduced state-of-the-art architecture ResNet50 [13]. This modification is motivated by two major reasons: (1) ResNet50 introduces residual connections into the architecture which allows the training of more layers compared to VGG16. Thereby, more complex features can be learned, resulting in a more accurate model with respect to image classification performance. (2) At the same time, less parameters are used compared to VGG16, yielding to a faster model during training and inference time.

We obtain pixel-wise class predictions from the network by applying a 1x1 convolution to the feature maps after the last bottleneck layer (res4b22) and up-sampling the predictions via bilinear interpolation to the output size. We additionally augment the resulting FCN with a pyramid pooling module [4]. Pyramid pooling modules are an attempt to extract global contextual information from feature maps and have shown to be an important component for improving the accuracy in semantic segmentation tasks[4]. Motivated by recent success of these multi-scale features [14, 4], we attach a pyramid pooling module to the last bottleneck layer of the network. The final network is trained end-to-end using backpropagation.

2.2. Adaptive Uncertainty Weighted Class Loss

The aim of our approach is to incorporate (1) the imbalanced frequency of pixels per class as well as (2) the uncertainty of the model with respect to the classes into a single loss function. Both objectives can be expressed as multi task problem, in which the target loss is defined as follows:

$$L_{total}(x;\theta) = L_c(x;\theta) + L_u(x;\theta) \qquad (1)$$

where L_i the corresponds to the task loss functions to be minimized with respect to the network parameters θ.

The first term $L_c(x;\theta)$ is used to cope with the imbalanced classes in the dataset. We use a weighted cross-entropy loss with a medium frequency balancing on the classes in the set of classes C as follows:

$$L_c(x,\theta) = \sum_{c=1}^{C} -C_c log P(C_c = 1|x,\theta,\sigma_t) * w_c \qquad (2)$$

Each class c is weighted by w_c the ratio of the median class frequency and the class frequency f_c (computed over the training dataset) as follows:

$$w_c = \frac{median(f_c|c \in C)}{f_c} \qquad (3)$$

The weighting gives classes with a few pixels higher weights and classes with many pixels smaller weights.

The second term $L_u(x;\theta)$ is used to weight classes according to the per-class uncertainty of the model. The loss has the same form as the weighted cross entropy loss defined in Eq. 2. Instead of using the median frequency balancing as weighting for w_c, the class uncertainty of the model is used. We compute the per class uncertainties of the model, from uncertainty maps using Monte Carlo dropout [15] for a given pixel classification. We therefore sample ten Monte Carlo samples from the network, calculate the variance over the softmax output and aggregated over all pixels per class. Each resulting class uncertainty is normalized by the uncertainty of the remaining classes. The loss adaptively assigns classes with an relative high uncertainty an higher weight.

3. EXPERIMENTS AND RESULTS

3.1. Dataset and Metrics

Dataset. We evaluate our proposed approach on the ISPRS Vaihingen 2D semantic labeling contest dataset [1]. The dataset consists of airborne images of Vaihingen, a town in Germany, containing high resolution true ortho photo (TOP) tiles in a ground sampling distance of 9 cm. Additionally, the dataset comes with Digital Surface Models (DSMs) of Vaihingen in the same spatial resolution. Normalized DSMs are provided by Gerke [16] to handle effects of varying ground

Table 1. Performance of the different models on the validation set. The F1 scores and accuracies are shown as percentages.

	Imp. Surface		Building		Low Veg.		Tree		Car		Overall	
Method	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.
FCN	90.69	90.16	95.15	95.46	76.77	74.53	87.64	90.31	82.15	73.98	86.48	84.89
FCN + MFB	90.96	89.11	95.25	95.66	77.58	76.32	88.24	**90.89**	89.15	**88.65**	88.24	88.13
FCN + MFB + UWC	**91.95**	**90.60**	**95.59**	**96.55**	**79.58**	**78.59**	**88.77**	90.26	**90.88**	87.79	**89.35**	**88.76**

heights. The pixels in the dataset are labeled with the following six land use and land cover classes: *Impervious Surfaces, Building, Low Vegetation, Tree, Car* and *Background*. Ground truth labels are publicly available for 16 of the 33 tiles in the dataset, the remaining ground truth is used for the private test set. Following recent methods [7], we split the 16 tiles into a training set with the 11 images (areas: 1, 3, 5, 7, 13, 17, 21, 23, 26, 32, 37) and the validation set having the 5 images (areas: 11, 15, 28, 30, 34).

Metrics. We follow the evaluation procedure defined by the ISPRS [17] to report our results on the contest. The segmentation accuracy is measured by the F1-Score $F1 = (2 * Precision * Recall/(Precision + Recall))$ and the overall accuracy, defined by the percentage of correctly labeled pixels. In order to cope with labeling errors and noise in the labels, class boundaries were eroded with a disk of radius three and annotated with the background class.

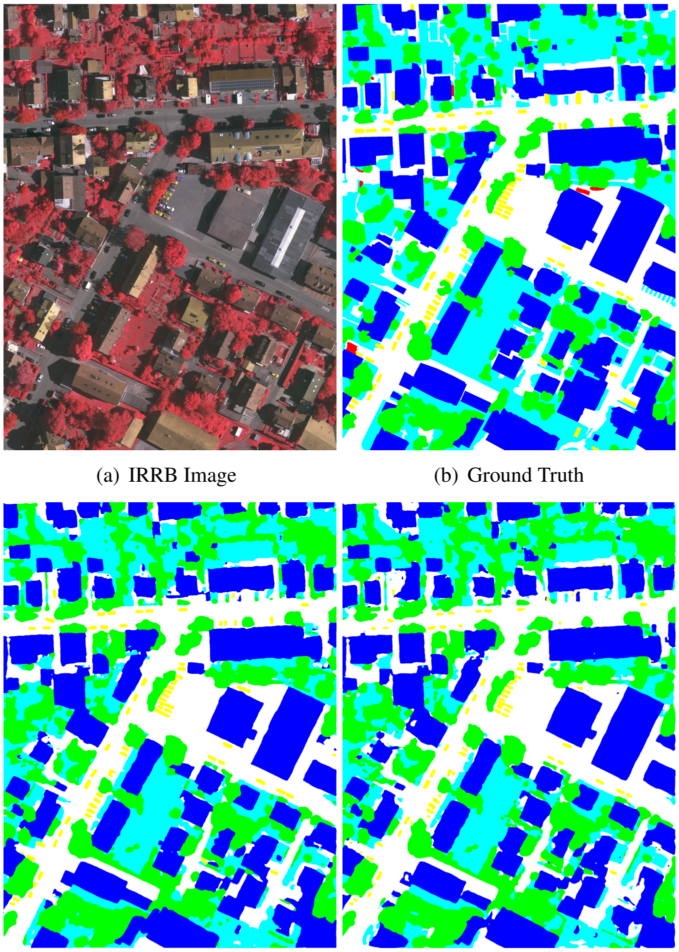
3.2. Network Training

Early Fusion. Since the Vaihingen dataset includes multimodal information of the optical and depth domain, we perform an early fusion approach. Thereby, we stack the three IR-R-G channels of the optical domain together with the two grayscale channels of the DSM and the normalized DSM images. The resulting five dimensional tensor is then passed to the network with the architecture described in Section 2.1.

Training Setup. We initialize all networks in this paper with a ResNet50 model pretrained on ImageNet [18]. We average the weights of the first convolution over the channels and extend the weights in the channel dimension to five to cope with two additional dimensions for the depth information not being present in the pre-trained RGB model. The non RGB channels are initialized with the average across the RGB channels. All networks in this paper are trained for 100 epochs with a batch size of 8. We extract overlapping image patches (with a surface overlap of 50%) of size 256x256 pixels and apply random flipping in horizontal and vertical dimension as data augmentation. We use Stochastic Gradient Descent with a momentum of 0.9, weight decay of 0.0005 and learning rate of 0.01 to optimize the network parameters. Inspired by [4], we use the poly learning rate policy, in which the learning rate determined by $(1 - \frac{iter}{max\_iter})^{power}$. We set the parameter *power* to 0.9. For a fair comparison, all networks are trained with the same experimental setup.

3.3. Experimental Results

The overall performance of the segmentation network trained with the proposed loss can be seen in Table 1 (row three). This network achieves an overall F1-Score of 89.36% and class accuracy of 88.76% on the Vaihingen validation set. We compare the results against the baseline, which uses the same network and experimental setup but was trained with the standard cross entropy loss only. Compared to this baseline, training with the novel loss improves the overall accuracy by 3.87 percentage points and improves the F1-Score 2.88 percent-



(a) IRRB Image

(b) Ground Truth

(c) Prediction FCN (MFB+UWC)

(d) Prediction FCN

Fig. 2. Performance of the different models on single image tile of the validation set of the ISPRS Vaihingen dataset.

age points. For the F1-Score, this results in a reduction of the classification error by 21.22%. An major improvement against the baseline can be recognized in the car class which gets improved from 82.15% to 90.88% on the F1-Score (row one against row three).

As second baseline, we trained the fully connected network using median frequency balancing weighted cross entropy loss. The performance against the first baseline gets improved to an overall F1-Score of 88.24% and to 88.13% for the overall pixel accuracy. However, compared to our approach, the second baseline achieves less accurate results as shown in Table 1 (row two against row three). While the F1-Score for the car class gets improved by about one percent, also the big classes with higher uncertainty such as *Low Vegetation* and *Impervious Surface* can still be improved.

4. CONCLUSION

In this paper, we introduced a novel loss function for segmentation networks relying on high resolution satellite imagery. The proposed loss takes imbalanced pixels of small classes into account as well as the per class uncertainty of the model. Compared to standard cross entropy loss the error of the F1-Score is reduced by 21.22% and compared to median frequency balancing by 9.43%. In the future, we additionally plan to weight the two losses by an scalar λ_i to model the importance of each task on the combined loss L_{total} and let the network learn an adaptive task weight as done in the multi-task scenario [19].

5. REFERENCES

[1] Michael Cramer, "The dgpf-test on digital airborne camera evaluation–overview and test design," *Photogrammetrie-Fernerkundung-Geoinformation*, vol. 2010, no. 2, pp. 73–82, 2010.

[2] Benjamin Bischke, Patrick Helber, Christian Schulze, Srinivasan Venkat, Andreas Dengel, and Damian Borth, "The multimedia satellite task at mediaeval 2017: Emergence response for flooding events," in *Proc. of the MediaEval 2017 Workshop (Sept. 13-15, 2017). Dublin, Ireland*, 2017.

[3] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *arXiv preprint arXiv:1709.00029*, 2017.

[4] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," *arXiv preprint arXiv:1612.01105*, 2016.

[5] Evan Shelhamer, Jonathan Long, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[6] Keiller Nogueira, Samuel G Fadel, Ícaro C Dourado, Rafael de O Werneck, Javier AV Muñoz, Otávio AB Penatti, Rodrigo T Calumby, Lin Tzy Li, Jefersson A dos Santos, and Ricardo da S Torres, "Exploiting convnet diversity for flooding identification," *arXiv preprint arXiv:1711.03564*, 2017.

[7] Michael Kampffmeyer, Arnt-Borre Salberg, and Robert Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–9.

[8] Maggiori, Emmanuel and Tarabalka, Yuliya and Charpiat, Guillaume and Alliez, Pierre, "High-Resolution Semantic Labeling with Convolutional Neural Networks," *arXiv preprint arXiv:1409.1556*, 2016.

[9] Joachim Niemeyer, Franz Rottensteiner, and Uwe Soergel, "Classification of urban lidar data using conditional random field and random forests," in *Urban Remote Sensing Event (JURSE), 2013 Joint*. IEEE, 2013, pp. 139–142.

[10] David Eigen and Rob Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2650–2658.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[12] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[15] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.

[16] Markus Gerke, "Use of the stair vision library within the isprs 2d semantic labeling benchmark (vaihingen)," 2014.

[17] //www2.isprs.org/commissions/comm3/wg4/ semantic labeling.html, "Isprs 2d semantic labeling contest," 2015.

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[19] Benjamin Bischke, Patrick Helber, Joachim Folz, Damian Borth, and Andreas Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," *arXiv preprint arXiv:1709.05932*, 2017.