SEGMENTATION OF ISPRS VAIHINGEN DATASET INCLUDING DSM DATA

Yu-Chuan Cheng

yu-chuan.cheng@tu-braunschweig.de

ABSTRACT

Use this opportunity to compare DeepLabv3 with U-net in terms of segmentation and semantics, as well as to understand the differences between focal loss and cross entropy loss, between using just RGB channels and RGB plus the fourth channel, DSM, and between utilizing both.

***Index Terms*—**Vaihingen dataset, DSM, DeepLab3+, U-net

1. INTRODUCTION

We want to compare U-net and DeepLab3+ to other models we've studied throughout the course to see if they can still provide high-accuracy semantic segmentation with satellite images. We want to use DeepLab3+, U-net with Focal Loss, and Cross Entropy with Adam Optimizer to test which experimental strategy is best.

2. RELATED WORK

A. *Normalization (min-max)*

Min-Max Normalizing is the first normalization technique. It consistently performs worse than the second normalizing approach (about 40% mIOU). As a result, we only choose the second one to serve as our primary research basis.

B. *Normalization (standardization)*

Standardization, which has a 0 mean and 1 standard deviation, is the second normalizing technique we'll discuss. Since the results of this normalizing approach are significantly superior to those of the prior normalization method, we use it as the major component of the experiment. So, we'll concentrate on that.

C. *Preprocessing network*

We must perform some preprocessing in order to suit the original network design because the Deeplabv3+ can only be applied to three channels. Since we already performed standardization, we begin with batch normalization and then sigmoid. Conv2D assists us in blending the fourth channel into the other three, and the drop-out layer enables us to train our neural network more thoroughly while avoiding overfitting. However, the performance dropped by at least 10% nIOU when the drop out layer was inserted before the Conv2D layer, despite our efforts. Our final construction will thus resemble Figure 1.

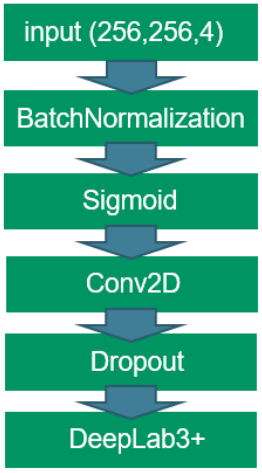

Fig. 1: preprocessing network of deeplabv3+

3. DATASET INTRODUCTION

This test data set was gathered over Vaihingen, Germany. The data set is an excerpt from the data used to evaluate digital aerial cameras by the German Association of Photogrammetry and Remote Sensing (DGPF). For each of the three test zones, reference data are given for a variety of object classes.

The data set consists of 33 patches (of varying sizes), each of which is composed of a true orthophoto (TOP) and a digital surface model (DSM) that were extracted from a larger TOP mosaic (see Figure 2). The ground sampling distance for the TOP and the DSM is 9 cm.

The data split can be formally described as image $x_1^T = \{x_t\}$, with T = {128, 38, 190}. The elements of T are respect to training set, validation set and testing set. An example image can be seen in Figure 2(a). For each image, there is a corresponding ground truth semantic segmentation mask $\bar{m}_t = (\bar{m}_{t,i}) \in S^{H \times W}$, with S being the set of semantic classes $S = \{1, \ldots, S\}$, and $S = 6$ for Vaihingen dataset.

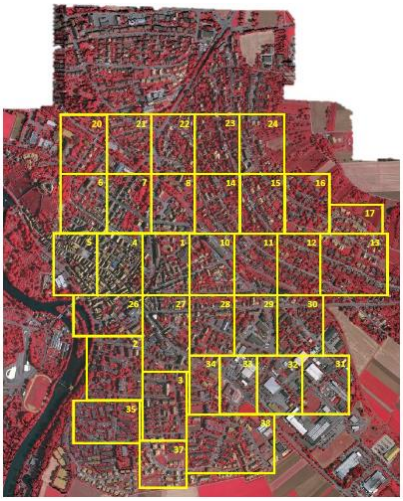

Fig. 2: The overview of the Vaihingen data. Numbers refer to the individual patch numbers as written at the filename endings. 2D Semantic Label. - Vaihingen (isprs.org)

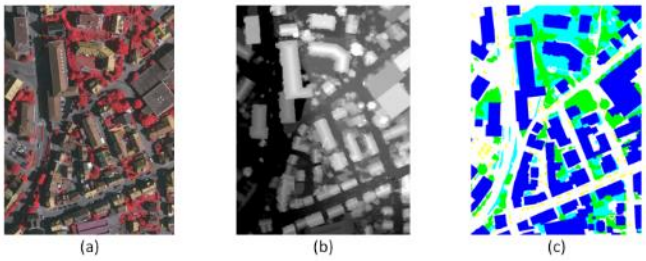

Fig. 3: Example patches of the semantic object classification contest with (a) true orthophoto, (b) DSM, and (c) ground truth. 2D Semantic Label. - Vaihingen (isprs.org)

4. EXPERIMENTAL EVALUATION AND DISCUSSION

A. *Metrics*

For the evaluation of Vaihingen semantic segmentation on the generated dataset, we propose the following setup with evaluation metrics based on the mean intersection over union (mIoU), which is typically used in semantic segmentation.

$$mIOU = \frac{1}{S}\sum_{s\in S}\frac{TP_s}{TP_s + FP_s + FN_s}$$

True positives TP_s per class $s \in S$ are defined by images with index t and pixel positions i, where $\bar{m}_{t,i} = m_{t,i} = s$ holds, wherein the latter denotes the network prediction. False positives FP_s are defined by $\bar{m}_{t,i} \neq s = m_{t,i}$, and false negatives FN_s by $\bar{m}_{t,i} = s \neq m_{t,i}$.

B. *Implementation and Training Details*

The network is trained on the Vaihingen data set. We monitor the training using the validation set. The exact split can be found on ISPRS. We use the Adam optimizer with an initial learning rate of 0.001 and exponential decay to train segmentation of ISPRS Vaihingen dataset for 200 epochs, but we also set early stopping patience as 40. We select the best performing model on the validation set for evaluation on the test set. The data augmentation scheme includes random horizontal flip, random vertical flip, random rotate 90˚ and shift scale rotate. Last but not least, our gamma value in focal loss is 2.

C. *Segmentation Results and discussion*

When the fourth channel is used, we can observe that the performance is not always improving. Deeplabv3+ performs practically better in every way, though. Given that Deeplabv3+ has 11.8 million parameters whereas U-net only has 8.5 million, we trade off computational power and time for greater mIOU.

We can observe that the model that uses a focal loss function misclassifies some impervious surfaces. The gamma value, in our opinion, is the cause. The focal loss, the primary component of the loss function, contains the hyper parameter known as gamma value. (see Figure 6)

Table 1: The experiment result. The Metrics is mIOU

	U-net		**Deeplabv3+**	
	Cross Entropy	Focal loss	Cross Entropy	Focal loss
With RGB	57.70%	53.20%	**64.31%**	63.56%
With RGB & DSM	55.03%	51.20%	52.78%	**58.65%**

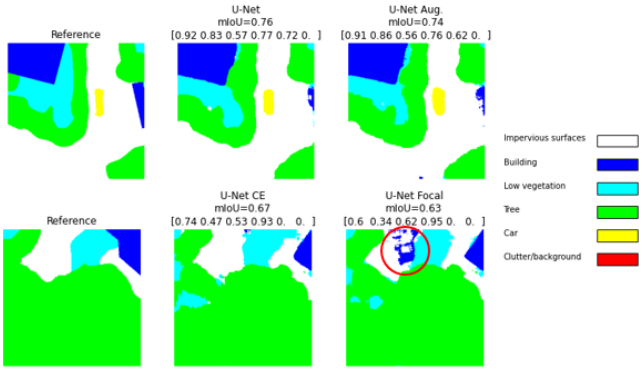

Fig. 4: The semantic segmentation result of U-Net

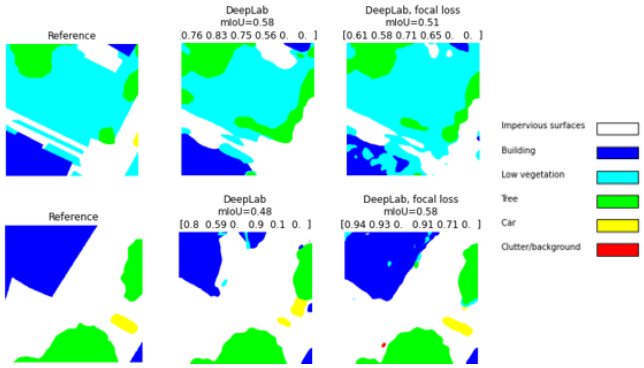

Fig. 5: The semantic segmentation result of DeepLabv3+

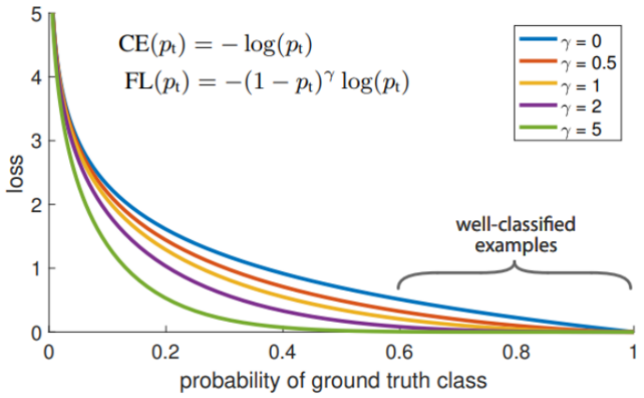

Fig. 6: Comparison between Focal loss and cross entropy loss. [5]

5. CONCLUSIONS

On impermeable surfaces, focal loss frequently produces incorrect semantic segmentations. The gamma value in the focus loss function is probably to blame for semantic segmentation's misclassification. Additionally, Deeplabv3+ has the highest performance when we train the model with 4 channels, with a focus loss of 58.65% mIOU, albeit it is still inferior to the model employing the cross-entropy loss function with just the RGB channel (64.31%).

The use of the many optimizers and hyper parameters is what the future holds. Given that the gamma value of the focal loss, which has a significant impact on the output, has been noted, we would want to set additional experimental variables under the same conditions. Additionally, the optimizer we employ is typically Adam, which outperforms other optimizers. But it uses SGD optimizer in the HRNet training environment. To do more analysis on this subject, we would attempt SGD.

6. REFERENCES

[1] Chen, L., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018a). *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation.*

[2] Chen, L., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018b). *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation.*

[3] L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, *DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs*, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834-848, 1 April 2018

[4] Wu, H., Zhang, J., Huang, K., Liang, K., & Yu, Y. (2019). *FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation.*

[5] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999-3007, doi: 10.1109/ICCV.2017.324.