

Complementary Features Learning from RGB and Depth Information for Semantic Image Labelling

LIN CHEN¹, DAIXIN ZHAO¹, CHRISTIAN HEIPKE¹

Abstract: In this paper, we present a complementarity constraint for features computed from different sources of input data before fusion in semantic labelling. A two-branch encoder-decoder architecture with ResNet-50 is proposed and used as classification network. Our proposed complementarity constraint is added to the standard softmax cross-entropy classification loss. The impact of different weights for this constraint in multi-modal data fusion is investigated. The result of the two branch network is also compared to the one obtained with only the spectral information. The constraint is shown to improve the results consistently in our experiments. Different amounts of improvement are achieved when different weights for the complementarity constraint are used.

1 Introduction

Semantic image labelling, called classification in remote sensing, is the process of assigning an object class label to each pixel in an input image. This labelling process for aerial and satellite images is one of the fundamental tasks in photogrammetry and remote sensing. Applications, e.g. land cover classification, land-use classification and change detection, rely heavily on semantic labelling. Therefore, semantic image labelling has been a focus of research in photogrammetry and remote sensing community for a long time.

Spectral information of images, such as RGB or IRRG, is normally the first data source for classification. However, multi-modal data such a combination of spectral and depth information is frequently available, and the height data can add additional information. For instance, the appearance of objects may change due to shadows and weather conditions (e.g. cloudy, snowy), but a digital surface model is not influenced by these effects.

In recent years, the research focus of semantic labelling has shifted from probabilistic graphical models, e.g., conditional random field (CRF), to deep neural networks. Deep neural networks are designed to extract useful features from the input automatically for the underlying task. Among many of the deep neural networks, convolutional neural networks (CNN) (LECUN et al. 1998) attracts most attention in image classification and semantic labelling. Fully Convolutional Networks (FCN), a variant of CNN without a fully connected layer, directly generates a classification map for the entire input images and is now a stand tool for semantic labelling. The encoder-decoder architecture based on FCN (NOH et al. 2015; BADRINARAYANAN et al. 2017; CHEN et al. 2018) shows state-of-the-art performance on many semantic labelling benchmarks.

One of the major concerns in using multi-source data for semantic labelling is how to properly fuse the data in a multi input branch encoder-decoder FCN. Features extracted from different

¹ Leibniz Universität Hannover, Institute of Photogrammetry and Geoinformation, Nienburger Str. 1, D-30167 Hannover, E-Mail: chen@ipi.uni-hannover.de, daixin.zhao@gmail.com, heipke@ipi.uni-hannover.de

sources of input is normally fused in the early or middle stages (i.e., before decoding) of the network. In this paper, we build a complementarity constraint to motivate the features from different source to be perpendicular to each other, thus “different” distinctive features are learned by the network. A comparison of adding this constraint with different strength for an encoder-decoder form of network proposed by us is also presented and analyzed.

2 Related Work

Semantic labelling is one of the major tasks in remote sensing image interpretation. Fully Convolutional Networks (FCN) (LONG et al. 2015) take an arbitrary size input image and then learn the features through a gradually down-sampled convolution with trainable kernels; a dense pixel-wise classification map is then generated for the input image through an up-sampling stage. The final classification map can be 1/4 or 1/8 of the input image size. Several studies based on this architecture, e.g. SegNet (BADRINARAYANAN et al. 2017) and DeepLab (CHEN et al. 2015), are now standard baselines for semantic labelling. The following networks take the idea from FCN and achieve better accuracy in semantic labelling tasks. DeconvNet (NOH et al. 2015) builds a convolution-deconvolution network based on the VGG 16-layer Net (SIMONYAN et al. 2014) to upsample the convolutional feature maps to the original input image size. With the same backbone of VGGNet, SegNet presents an encoder-decoder architecture (bottleneck architecture) and stores the encoder max-pooling indices for producing sparse feature maps in the decoder stage. DeepLab v3+ (CHEN et al. 2018) takes advantages of the encoder-decoder architecture and Atrous (HOLSCHNEIDER et al. 1989) Spatial Pyramid Pooling (ASPP) from the first version of DeepLab. They utilize the dilated convolution (a.k.a. atrous convolution) to enlarge the receptive field of the convolution filter. Specifically, zero values are inserted between filter values to form a larger field-of-view without increasing the number of parameters. The rate parameter determines the number of zeros inserted between two adjacent filter matrix values, and thus controls the size of the receptive field. By using different rates of same size filters and stacking the results, a multi-scale response is imitated. These architectures have achieved notable performance improvements in semantic labelling tasks, e.g. using the PASCAL VOC 2012 (EVERINGHAM et al. 2014) and Cityscapes (CORDTS et al. 2015) datasets.

On the other hand, multi-modal data is frequently available in practical applications and is also provided in recent semantic segmentation contests, such as the ISPRS benchmark for 2D Semantic Labeling Contest (ROTTENSTEINER et al. 2014). Using multi-modal data can potentially improve the semantic segmentation performance because of the complementary information from different source. FuseNet (HAZIRBAS et al. 2016) makes use of RGB and depth (2.5D) with a two-branch-encoder and fuses the output from each branch at the end of encoder stage, and the fused information is then fed into a RGB-D decoder for generating a pixel-wise classification map. CHEN et al. (2018) investigate the impact of combinations of hand-crafted radiometric, e.g., NDVI (normalized difference vegetation index), and geometric features (e.g., nDSM, change of curvature) derived from the true orthophotos and DSM. The combination of RGB and normalized DSM (nDSM) delivers the best results in their investigation. AUDEBERT et al. (2018) explore the influence of where to fuse different source of information, e.g., IRRG, nDSM, NDVI and DSM data, in an encoder-decoder structure based variants of FuseNet. Their results show

that early fusion improves the semantic segmentation performance by learning multi-modal features jointly. Nevertheless, late fusion is helpful for hard pixels as the authors observed in experiments. In those mentioned works, two frequent fusion manners, summation and concatenation of input feature vectors for different modal, are used.

DualNet (HOU et al. 2017) focuses on learning complementary features from different sub-networks for same-source input data in object classification. Two parallel sub-networks without shared parameters extract features from the same source and then their extracted features are fused by summation (main-branch). A cross-entropy classifier based on the fused features acts as the main loss and two other cross-entropy classifiers built based on the features computed from each sub-network act as auxiliary losses. To motivate sub-networks to learn complementary features, only the loss from one sub-network and the main branch are calculated and the parameters of the other sub-network keep fixed in a single iteration. The two sub-networks are trained alternately, which means the sub-networks exchange the role of being optimized and fixed after each iteration. This alternative learning strategy can prevent one sub-network from moving towards the same weights as the other, and hence allows both sub-networks to capture discriminative, but complementary features from the input data. Once the networks are well trained, a joint fine tuning of all three branches is conducted, which yielded a minor performance improvement. As far as we know, DualNet is the most closely related work to our research. However, their work concentrates on descriptor learning for object classification. In this paper, we hypothesize that if the mid-level features (those extracted at the end of the feature extraction stage in the network) computed from RGB and depth information are complementary to each other, then the information computed from RGB and depth will reinforce each other and thus provide more distinctive and complementary features for classification. Based on this hypothesis, we formulate a complementarity constraint for the mid-level features of RGB and depth information. The complementarity constraint leads to features which are perpendicular to each other in high dimensional feature space, thus the features span different dimensions in feature space, which means they are aimed at extracting “different” useful features for semantic segmentation.

3 Methodology

In this section, we first report the network we propose for semantic labelling. Then, the complementarity constraint is introduced, followed by details on the online augmentation we apply for training.

3.1 Network Architecture

DeepLab v3+ (CHEN et al. 2018) is selected as the basic architecture of our baseline network. Several modifications for DeepLab v3+ are introduced in our classification network. It contains an encoder and decoder stage. In the encoder stage the input image is fed into convolution blocks and downsized to a series of lower spatial resolution, e.g. 1/2, 1/4 of the input image, high dimensional mid-level feature maps are thus obtained. Then, dilated convolution is applied to the mid-level feature maps to enlarge the field-of-view. This procedure is called Atrous Spatial Pyramid Pooling (ASPP). In the decoder stage, different scales of down-sized features maps of the encoder stage are fused by concatenation to the corresponding scale feature maps in the decoder

stage. Through this fusion the spatial context and boundary details of objects can be preserved for generating finer segmentation results. After obtaining the full size up-sampled feature map, a pixel-wise classification digits map is built through a few convolution blocks. If the underlying task is pixel-wise classification, then the cross-entropy is calculated as the classification loss by using the digits map. Thus, the network output is a classification map with a same size of the input image. However, the output can also be at a lower resolution, like 1/4 of the input image, once the output quality is able to meet the requirements of application.

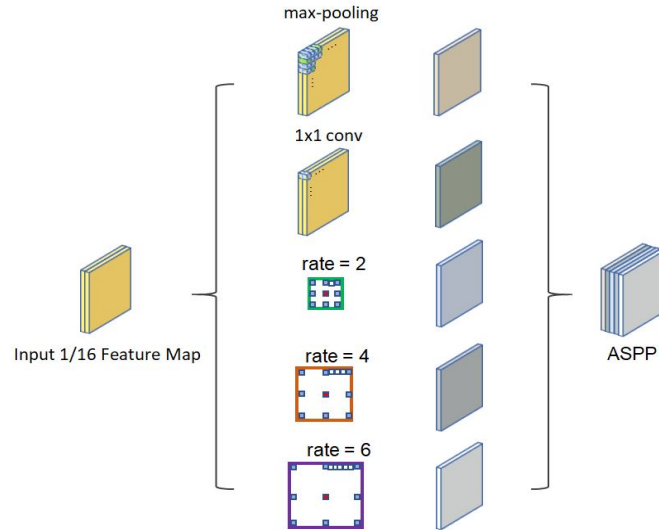


Fig. 1: Illustration of Atrous Spatial Pyramid Pooling (ASPP). Max-pooling, a 1×1 convolution and three atrous convolutions with rate parameters 2, 4, 6, are used to generate the ASPP feature map.

To increase the learning ability of the network, our variant of DeepLab v3+ takes advantage of the ResNet-50 network (HE et al. 2016), we modify it for our encoder stage. The ResNet-50 network has 5 stages, more specifically, 50 convolutional layers. Each stage consists of one convolutional block and 3 to 5 identity blocks. In our network, the 7×7 convolution with stride 2 in the first block is replaced by a 5×5 convolution (with unchanged stride), and the following 3×3 max-pooling is dropped. By using a smaller convolution kernel and dropping max-pooling in the first block, less smoothing is conducted and thus more explicit boundary information is preserved to differentiate objects, like buildings, that normally have sharp boundary information. Later in the encoder stage we choose dilated convolution with rate parameters 2, 4, 6 in ASPP to preserve more continuous spatial information at differing scales, as shown in Fig. 1.

To classify the aerial images with different modality, two branches for different sources of data are used in the encoder stage. IRRG and nDSM/NDVI data are fed into the two-branch-encoder separately to extract features. The extracted mid-level features are fused by summation. The feature decoder up-samples the ASPP feature map with bilinear interpolation. During up-sampling, the 1/4 and 1/8 feature maps from both branches are concatenated with corresponding scale of feature maps in the decoder. In the end, a classification map with the same size of the input image is calculated through a softmax classifier. The network architecture is described in Fig. 2.

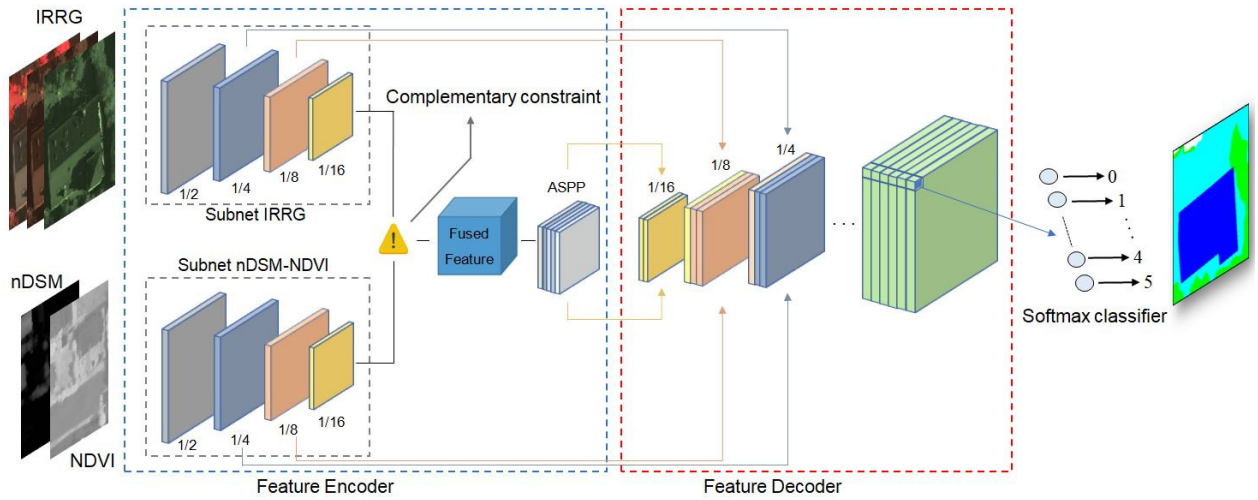


Fig. 2: Demonstration of network architecture. 1/4, 1/8 and 1/16 feature maps from encoder stages are fused in the decoder stage with the same size up-sampled feature maps. An ASPP feature map is generated using the fused feature map up-sampled for the decoder stage. The network is trained with a softmax cross-entropy loss and a complementarity constraint.

Input of the second branch are the nDSM and the NDVI. The nDSM captures representations of height differences, while the NDVI contains the discriminative information for vegetations and other classes. Therefore, the combination of the two input data can distinguish high vegetation from low vegetation, or different classes of objects with similar height, for instance, trees and buildings. The NDVI for each individual image pixel is derived by:

$$NDVI = \frac{IR - R}{IR + R} \quad (1)$$

in which IR stands for the near infrared band, R represents red band.

3.2 Complementary loss

Following our hypothesis that mid-level features computed from the two branches should be complementary to each other, we formulate a complementarity constraint for learning discriminative, yet complementary features for the classification. To enforce dissimilarity of the two feature vectors computed from different source of data, they should be perpendicular to each other. The complementarity constraint is illustrated in Fig. 3.

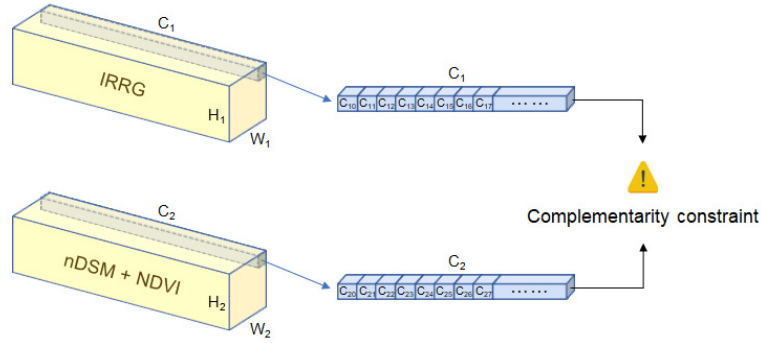


Fig. 3: Complementarity constraint. $H_1, H_2, W_1, W_2, C_1, C_2$, stand for the height, width and channel numbers for feature maps from the IRRG and the nDSM + NDVI branch, respectively.

Our complementary loss is described by the cosine similarity between two non-zero vectors in high dimensional feature space. The cosine similarity lies the range $[-1, 1]$, where a similarity of -1 is computed from two antiparallel vectors and a similarity of 1 is computed from two parallel vectors. Perpendicular vectors have a similarity of 0 , which means the two features provide complementary information. The cosine similarity is defined as:

$$\text{cosine similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (2)$$

where \mathbf{A} and \mathbf{B} stand for two non-zero vectors with n entries. Since the complementary loss should penalizes vector combinations providing similar information, the absolute value of cosine similarity that neatly bound the loss value between 0 and 1 is used for simplicity. When training the network with the complementary loss, we treat it as a regularization term. The final loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{classification}} + \lambda \cdot \mathcal{L}_{\text{complementary}} \quad (3)$$

The parameter λ represents the regularization strength of the complementary loss. We investigate different values of λ in our experiments, the results are discussed in Section 4.

3.3 Online augmentation

To overcome the relatively small amount of available training data and alleviate potential overfitting, online augmentation is applied in our training process. The augmentation contains random flipping (vertically or horizontally) and random rotation ($0^\circ, 90^\circ, 180^\circ, 270^\circ$) of all input data and the corresponding ground truth label. A random tag for each training sample in each mini-batch is set to switch the augmentation during training on or off. Overall, 50% of the samples used for forward propagation are augmented, and the other 50% samples are original data without augmentation.

4 Experimental results

4.1 Dataset

We use the Vaihingen image dataset from the ISPRS 2D Semantic Labelling Contest⁽¹⁾ for our experiments. The goal of this contest is to label images using multiple object categories, namely tree, building, low vegetation, impervious surfaces, car and clutter/background. This dataset contains very high resolution true orthophotos (TOP) with three bands (near infrared, red and green), manually labelled ground truth and a corresponding DSM derived from dense image matching. Instead of directly using the provided DSM, we use an nDSM released by GERKE (2014) to be independent of absolute heights. As a side effect, the noise inherent in the DSM is decreased by the filtering operation during the generation of the nDSM.

The Vaihingen dataset contains 33 TOP images with a ground sampling distance (GSD) of 9 cm, for 16 of them ground truth is provided. The test results are based on the test set containing 17 images with so called uneroded and eroded ground truth released recently. Uneroded ground truth contains the complete reference of the test data, whereas eroded ground truth ignores the object boundaries in a buffer of 3 pixels width to reduce the effects of uncertainty in boundary definition.

4.2 Parameters setup

We randomly choose 13 images for training and the other 3 images for validation. The TOP images, nDSM, NDVI and ground truth images are cropped into 256×256 pixel patches by sliding a 256×256 window with step of 64 pixels. That means adjacent patches overlap by 75%. After cropping, 4037 image patches are obtained for training.

We also utilize dropout introduced by SRIVASTAVA et al. (2014) after the 1/8 feature map concatenation in the decoder stage with the *keep_prob* parameter as 0.7, i.e. randomly dropping 30% connections during the training process to prevent overfitting.

Our model is implemented using the Tensorflow framework (ABADI et al. 2016) from scratch. We train our model for 200,000 iterations using Nesterov’s accelerated gradient with momentum 0.9 and the learning rate 0.0003. Additionally, the validation is based on a sliding window with a step of 50 pixels and the classification results from each 256×256 pixel patch are assembled with equal weights to generate final classification maps. The model with the best validation overall accuracy is retained.

4.3 Evaluation Criteria

The evaluation is based on the pixel-wise confusion matrix. The correctness (precision), completeness (recall), F1 score and overall accuracy are derived from the confusion matrix:

$$\text{Precision} = \frac{t_p}{t_p + f_p} \quad (4)$$

⁽¹⁾ <http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html> (accessed on Nov 21, 2018)

$$\text{Recall} = \frac{t_p}{t_p + f_n} \quad (5)$$

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (6)$$

$$\text{Overall Accuracy} = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \quad (7)$$

Where t_p , t_n , f_p , f_n represent the number of true positives, true negatives, false positives and false negatives, respectively.

4.4 Results

First, we compare the result of only using IRRG and using IRRG and nDSM/NDVI. Then different complementary strength λ , i.e. 0.1, 0.01, 0.001, for features computed from IRRG and nDSM/NDVI are compared. Table 1 shows the overall accuracy of different experiment setups on uneroded test dataset and eroded test dataset.

Tab. 1: Overall Accuracy (OA) on different experiment setups

Experiment setups	OA on uneroded test dataset	OA on eroded test dataset
Only IRRG	84.4%	87.2%
$\lambda = 0$	85.3%	88.2%
$\lambda = 0.1$	86.0%	89.0%
$\lambda = 0.01$	86.2%	89.1%
$\lambda = 0.001$	86.1%	89.0%

As illustrated in Table 1, the model trained only by a single IRRG branch encoder is outperformed by the two-branch-encoder network with IRRG and nDSM/NDVI as inputs, which is expected as the multi-modal data provide richer information for semantic segmentation. Notably, setting the complementary strength λ to 0.01 provides the best overall accuracy, as also shown in Table 2 and Table 3.

Table 2 and Table 3 show the experimental results with the precision, recall and F1 score calculated from the eroded and the uneroded data, respectively. Interestingly, we find that the highest F1 scores for each class are actually not given by the experiment setup with λ equals 0.01. For instance, cars are better classified by the network trained only with optical images. Our first interpretation is that the NDVI does not contain enough distinguishable features to separate cars from impervious surfaces and buildings. Also, the nDSM height difference might contain ambiguities on these relatively small objects due to problems in dense matching caused by lack of texture on cars. Therefore, the nDSM can contain some noisy information when classifying cars compared to other categories. Moreover, when the complementary strength is set to 0.1 or 0.001, we observe a relatively higher F1 score on building or trees and low vegetation.

Tab. 2: Experimental results based on the uneroded dataset

Experiment setups	Evaluation criteria	Impervious surface	Building	Low vegetation	Tree	Car	Clutter/Background
Only IRRG	Precision	0.845	0.907	0.807	0.809	0.781	0.764
	Recall	0.888	0.890	0.719	0.894	0.597	0.184
	F1 score	0.866	0.898	0.760	0.850	0.677	0.296
$\lambda = 0$	Precision	0.836	0.925	0.836	0.815	0.771	0.805
	Recall	0.910	0.912	0.723	0.895	0.429	0.013
	F1 score	0.871	0.918	0.775	0.853	0.551	0.025
$\lambda = 0.1$	Precision	0.844	0.925	0.841	0.836	0.648	0.687
	Recall	0.914	0.931	0.729	0.883	0.588	0.097
	F1 score	0.878	0.928	0.781	0.859	0.617	0.170
$\lambda = 0.01$	Precision	0.850	0.933	0.845	0.820	0.778	0.844
	Recall	0.910	0.916	0.741	0.904	0.562	0.097
	F1 score	0.879	0.924	0.790	0.860	0.652	0.174
$\lambda = 0.001$	Precision	0.839	0.929	0.831	0.845	0.792	0.797
	Recall	0.910	0.900	0.773	0.888	0.514	0.136
	F1 score	0.873	0.914	0.801	0.866	0.623	0.232

Tab. 3: Experimental results based on the eroded dataset

Experiment setups	Evaluation criteria	Impervious surface	Building	Low vegetation	Tree	Car	Clutter/Background
Only IRRG	Precision	0.872	0.929	0.841	0.837	0.809	0.797
	Recall	0.916	0.905	0.752	0.922	0.699	0.193
	F1 score	0.893	0.917	0.794	0.877	0.750	0.310
$\lambda = 0$	Precision	0.867	0.947	0.867	0.844	0.791	0.827
	Recall	0.935	0.929	0.762	0.922	0.528	0.014
	F1 score	0.899	0.938	0.811	0.881	0.633	0.027
$\lambda = 0.1$	Precision	0.877	0.947	0.874	0.863	0.670	0.726
	Recall	0.939	0.948	0.768	0.914	0.676	0.104
	F1 score	0.907	0.947	0.818	0.888	0.673	0.182
$\lambda = 0.01$	Precision	0.881	0.952	0.877	0.852	0.806	0.867
	Recall	0.934	0.934	0.781	0.930	0.669	0.104
	F1 score	0.907	0.943	0.826	0.889	0.731	0.186
$\lambda = 0.001$	Precision	0.868	0.951	0.864	0.876	0.816	0.825
	Recall	0.936	0.915	0.811	0.917	0.619	0.144
	F1 score	0.900	0.933	0.837	0.896	0.704	0.246

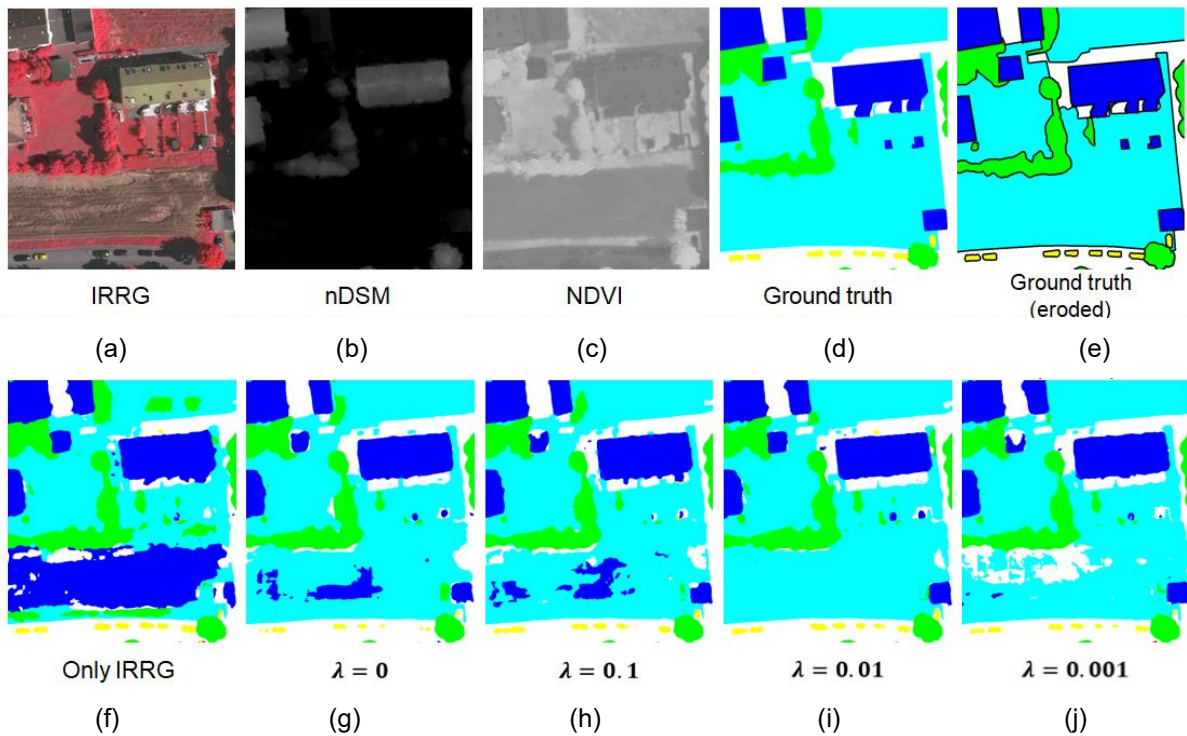


Fig. 4: Top: the top-left part of the test image 38, from left to the right, are shown: IRRG image (a), nDSM (b), NDVI (c), ground truth (d) and eroded ground truth (e). Bottom row from left to right: Classification results for network trained only using IRRG (f), multi-modal inputs network trained with complementary strength λ of 0 (g), 0.1 (h), 0.01 (i) and 0.001 (j).

Fig. 4 emphasizes that by using the complementary strength λ of 0.01, the network generates better classification results. The top row shows the input and ground-truth labels and the bottom row shows the classification result. In this area, low vegetation is classified as building without information from other sources. The bottom part of the input image is poorly classified except for the case of $\lambda = 0.01$. Our interpretation is that the lower vegetation is dry and shows more ambiguity to spectral features and the NDVI of buildings, and thus it cannot be differentiated well from buildings as shown in Fig. 4 (f). Involving the nDSM introduces depth information and thus contributes a lot towards a better differentiation in this case, as shown in the Fig. 4 (g, h, i, j). However, the strength of the complementary constraint does matter, as $\lambda = 0.01$ (Fig. 4(i)) delivers the best features to differentiate the lower drier vegetation and buildings.

Our experiments show that even though the complementary strength of 0.01 does not give the best results for each class, it strikes a better balance for all frequent categories and thus proves its effect.

5 Conclusion

In this work, we investigate a two-branch-encoder and decoder architecture with complementarity constraint for semantic segmentation using high resolution aerial images. Our proposed complementary constraint shows a notable performance improvement for semantic labelling with

different source of input data. In future work we will test more choices for the complementarity constraint and test our approach with different datasets.

Acknowledgement

Part of this work is done by DAIXIN ZHAO during an internship at Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover. The authors LIN CHEN and CHRISTIAN HEIPKE are grateful to NVIDIA Corp. for the GPU donation. Also, the author LIN CHEN and DAIXIN ZHAO are grateful to CHUN YANG and JUNHUA KANG for valuable discussions.

References

- ABADI, M., BARHAM, P., CHEN, J., CHEN, Z., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., IRVING, G., ISARD, M., KUDLUR, M., LEVENBERG, J., MONGA, R., MOORE, S., MURRAY, D.G., STEINER, B., TUCKER, P., VASUDEVAN, V., WARDEN, P., WICKE, M., YU, Y. & ZHENG, X., 2016: Tensorflow: a system for large-scale machine learning. In OSDI, **16**, 265-283.
- AUDEBERT, N., LE SAUX, B. & LEFÈVRE, 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**, 2481-2495.
- CHEN, L.C., PAPANDREOU, G., KOKKINOS, I., MURPHY, K. & YUILLE, A.L., 2018: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, **40**(4), 834-848.
- CHEN, L.C., ZHU, Y., PAPANDREOU, G., SCHROFF, F. & ADAM, H., 2018: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *European Conference on Computer Vision (ECCV)*. 8-14 September 2018, Munich, Germany
- CHEN, K., WEINMANN, M., SUN, X., YAN, M., HINZ, S., JUTZI, B. & WEINMANN, M., 2018: Semantic Segmentation of Aerial Imagery via Multi-scale Shuffling Convolutional Neural Networks with Deep Supervision. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, **4**(1). 29-36
- CORDTS, M., OMRAN, M., RAMOS, S., SCHARWÄCHTER, T., ENZWEILER, M., BENENSON, R., FRANKE, U., ROTH, S. & SCHIELE, B., 2015: The cityscapes dataset. *CVPR Workshop on the Future of Datasets in Vision*, **1**(2).
- EVERINGHAM, M., ESLAMI, S.M.A., GOOL, L.V., WILLIAMS, C.K.I., WINN, J. & ZISSERMAN, A., 2014: The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, **111**(1), 98-136.
- GERKE, M., 2014: Use of the stair vision library within the ISPRS 2D semantic labeling benchmark (Vaihingen). Technical Report. ITC, University of Twente.

- HAZIRBAS, C., MA, L., DOMOKOS, C. & CREMERS, D., 2016: Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. Asian Conference on Computer Vision, Taipei Taiwan, China, 213-228.
- HE, K., ZHANG, X., REN, S. & SUN, J., 2016: Deep residual learning for image recognition. IEEE conference on computer vision and pattern recognition, Caesars Palace in Las Vegas, Nevada, USA, 770-778.
- HOU, S., LIU, X. & WANG, Z., 2017: Dualnet: Learn complementary features for image recognition. International Conference on Computer Vision (ICCV), Oct. 22-29 2017, Venice, Italy, 502-510.
- LECUN, Y. BOTTOU, L. BENGIO, Y. & HAFFNER, P., 1998: Gradient-based learning applied to document recognition. Proceedings of the IEEE, **86**(11), 2278-2324.
- LONG, J., SELHAMER, E. & DARRELL, T., 2015: Fully convolutional networks for semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 3431-3440.
- HOLSCHNEIDER, M., KRONLAND-MARTINET, R., MORLET, J. & TCHAMITCHIAN, P., 1989: A REAL-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets: Time-Frequency Methods and Phase Space*, 289-297.
- NOH, H., HONG, S. & HAN, B., 2015: Learning deconvolution network for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 1520-1528.
- ROTTENSTEINER, F., SOHN, G., GERKE, M. & WEGNER, J.D., 2014: ISPRS semantic labeling contest. <http://www2.isprs.org/semantic-labeling.html>.
- SIMONYAN, K. & ZISSERMAN, A., 2014: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv: 1409.1556.
- SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I. & SALAKHUTDINOV, R., 2014: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, **15**(1), 1929-1958.