Semantic Segmentation of High-Resolution Aerial Imagery with W-Net Models

Maria Dias[1,2](✉), João Monteiro[1,2], Jacinto Estima[2,3], Joel Silva[2],
and Bruno Martins[1,2]

[1] Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
{maria.l.dias,joao.miguel.monteiro}@tecnico.ulisboa.pt,
bruno.g.martins@ist.utl.pt
[2] INESC-ID, Lisbon, Portugal
jacinto.estima@gmail.com, dinis.joel@gmail.com
[3] Instituto Politécnico de Setúbal, Setúbal, Portugal

Abstract. The semantic segmentation of high-resolution aerial images concerns the task of determining, for each pixel, the most likely class label from a finite set of possible labels (e.g., discriminating pixels referring to roads, buildings, or vegetation, in high-resolution images depicting urban areas). Following recent work in the area related to the use of fully-convolutional neural networks for semantic segmentation, we evaluated the performance of an adapted version of the W-Net architecture, which has achieved very good results on other types of image segmentation tasks. Through experiments with two distinct datasets frequently used in previous studies in the area, we show that the proposed W-Net architecture is quite effective in this task, outperforming a baseline corresponding to the U-Net model, and also some of the other recently proposed approaches.

Keywords: Semantic segmentation of satellite imagery · Fully-convolutional neural networks · W-Net architecture

1 Introduction

Large amounts of high-resolution remote sensing images are acquired daily through satellites and aerial vehicles, and used as base data for mapping and Earth observation activities. An intermediate step for converting these raw images into map layers in vector format is semantic image segmentation, which aims at determining, for each pixel, the most likely class label from a finite set of possible labels, corresponding to the desired object categories to map (i.e., discriminating pixels referring to roads, buildings, or vegetation). In the particular case of urban areas, semantic segmentation is quite challenging, given that objects in cities can be small, composed of many different materials, and

P. Moura Oliveira et al. (Eds.): EPIA 2019, LNAI 11805, pp. 486–498, 2019.
https://doi.org/10.1007/978-3-030-30244-3_40

have interactions with each other through occlusions, cast shadows, and interreflections. One can easily formulate the segmentation task as a pixel classification problem, to be addressed through supervised learning (i.e., given some labeled training data, infer the parameters of a model that estimates the conditional probabilities of the different classes, with basis on pixel intensities in the different color channels), although high within-class variability of the image intensities, and low inter-class differences, can be simultaneously expected.

In this paper, following on recent work in the area, we explore the potential of deep neural networks for semantic segmentation of high-resolution aerial images depicting urban regions. We propose to perform the segmentation using an adapted version of recent neural network architectures that have achieved very good results on other image segmentation tasks, namely W-Net architectures [1,2]. Through experiments with distinct datasets used in previous studies and competitions, we compared the performance of our particular W-Net architecture against simpler baselines, including the standard U-Net architecture. The obtained results attest to the effectiveness of the proposed method, which outperforms the baseline on the considered datasets, and also other recently proposed methods.

The rest of this document is organized as follows: Sect. 2 presents previous research in the area. Section 3 presents the deep learning method that was considered for aerial image segmentation, specifically detailing the adaptations implemented over previous W-Net approaches, as well as the hyper-parameters and the model training strategy. Section 4 details the evaluation methodology, including the selected datasets and evaluation metrics, and discusses the obtained results. Finally, Sect. 5 summarizes our conclusions and discusses possible directions for future work.

2 Related Work

Several previous studies have addressed the semantic segmentation of high-resolution aerial images through deep neural networks. For instance, Audebert et al. adapted previous deep learning models to use with multi-modal remote sensing data, i.e., multispectral imagery and digital surface model data [3]. In particular, the authors combined SegNet [4], an encoder-decoder architecture, and ResNet [5], a general convolutional architecture for image classification, that uses residual blocks as base models. Audebert et al. also proposed early and late fusion strategies for the multi-modal data. Early fusion is a strategy that learns fused feature maps of multi-modal features during the encoding process, whereas late fusion considers a separate network for each modality and fuses predictions after the decoding process. The proposed architectures were validated on multilabel datasets from previous challenges, particularly the ISPRS Vaihingen[1] and the ISPRS Potsdam[2] datasets, also used in our study.

[1] http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html.

[2] http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html.

Mou and Zhu advanced a model named RiFCN for semantic segmentation of high-resolution remote sensing imagery [6], based on what can be captured by the feature maps outputted at different depths of a deep neural network, as suggested in previous studies [7,8] that argued that shallower layers can capture low-level features such as object boundaries, while deeper layers are capable of learning high-level features. The RiFCN architecture is based on a bidirectional network that has a forward stream to extract multi-level feature maps and a backward stream that uses recurrent connections to fuse these multi-level feature maps from both streams. The forward stream is inspired by VGG-16 [9], a model with convolution blocks composed of multiple convolutional and max-pooling layers. In turn, the backward stream has autoregressive recurrent connections to embed the multi-level feature maps hierarchically. The authors demonstrated the effectiveness of the RiFCN architecture on the multi-label ISPRS Potsdam dataset, and on a binary dataset named INRIA Aerial Image Labeling[3].

In turn, Chen et al. proposed two similar encoder-decoder architectures with shortcut blocks [10], based on deep fully convolutional networks [11]. The shortcut block employed by the authors has two branches, namely a main branch with three convolutional layers, and a shortcut branch with one convolutional layer to help the direct gradient propagation. Both branches are merged in the complete network, and a ReLU activation function is applied to the result. The encoder is also composed of max-pooling layers, whereas the decoder has transpose convolutional layers. Leveraging the previous ideas, the authors propose two encoder-decoder architectures with convolutional and shortcut blocks, namely a symmetrical normal-shortcut fully convolutional network (SNFCN) and a symmetrical dense-shortcut fully convolutional network (SDFCN). The latter differs because it has three additional identity mappings between symmetrical pairs of the encoder and the decoder parts of the network. For validation, tests were conducted with the ISPRS Vaihingen and ISPRS Potsdam datasets.

Liu et al. described a self-cascaded convolutional neural network (ScasNet) [12], which corresponds to an architecture that has an encoder with multiple convolutional layers to extract features based on previous works, such as a 16-layer VGG-Net [9] and a 101-layer ResNet [5]. Dilated convolutions with different rates are applied in the last layer of the encoder to capture multi-level contexts. These contexts are sequentially aggregated from global to local in a self-cascaded approach to maintain the information about hierarchical dependencies. Then, to obtain an output with the same dimensionality of the input, corresponding to the segmentation mask, upsampling is performed, while shallow layers of the encoder are reused to identify low-level details of objects. The authors conducted experiments in two multi-label datasets, i.e. ISPRS Vaihingen and ISPRS Potsdam, and in a binary dataset, namely the Massachusetts Buildings Dataset [13], verifying the advantages of ScasNet.

[3] http://project.inria.fr/aerialimagelabeling.

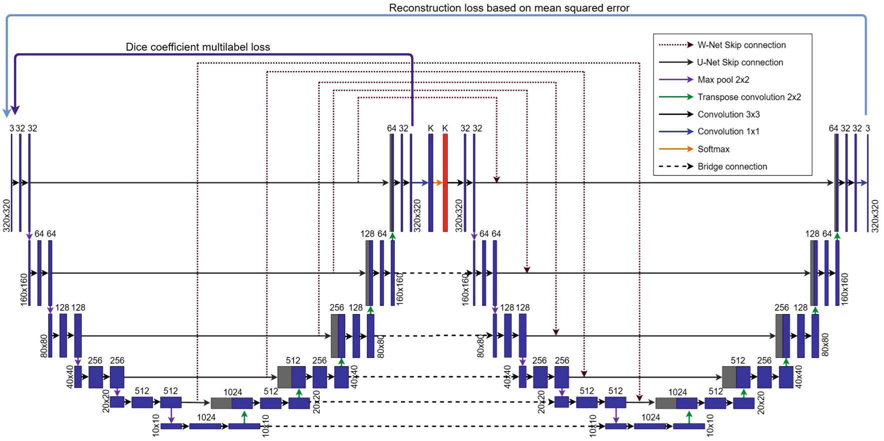


Fig. 1. Graphical representation for the considered W-Net network architecture.

3 The Proposed Segmentation Method

This section details the segmentation method used in our experiments. First, Sect. 3.1 presents the fully-convolutional neural network architecture adapted from previously proposed W-Net models [1,2]. Then, Sect. 3.2 presents the training and hyper-parameter tuning strategies that were considered.

3.1 The W-Net Neural Architecture

The neural network architecture used in our experiments borrows on ideas from previously proposed W-Net models, used for instance in the context of unsupervised image segmentation [1], or in the context of supervised medical image segmentation [2]. These models extend the typical U-shaped architecture (also commonly referred to as an hour-glass architecture) of the U-Net model [14]. Our network architecture is illustrated in Fig. 1, stacking/bridging together encoder (on the left-side) and decoder (on the right) U-Nets that, together, form a W-shaped architecture. The model is trained to simultaneously reconstruct the original input images from intermediate representations, and to predict the segmentation maps for the input aerial images. The input images are provided to the model in the LAB colour space, due to the fact that this representation is better at capturing human perceptual differences, important for segmentation and image reconstruction problems [15].

In total, our W-Net model has 44 convolutional layers which are structured into 22 modules. Each module consists of two 3×3 convolutional layers, each followed by a Leaky ReLU non-linearity [16] and batch normalization [17] operations. The first eleven modules form the dense prediction base of the network, and the second eleven modules correspond to a reconstruction decoder.

As in the regular U-Net architecture, the encoder consists of a contracting path (i.e., the first half of the first U-Net) to capture context, and a corresponding expansive path (the second half) that ends with a 1×1 convolutional layer, followed by a fully-connected layer with a softmax activation. The contracting path starts with an initial module which performs a convolution on the input images. In the figure, the output sizes are reported for an example input image resolution of 320×320 (i.e., the size of the image patches used in our experiments). Modules are connected via 2×2 max-pooling layers, and we double the number of feature channels at each down-sampling step. In the expansive path, modules are connected via transposed 2D convolution layers. The number of feature channels is halved at each up-sampling step. The input of each module in the contracting path is also bypassed to the output of its corresponding module in the expansive path, to recover lost spatial information due to down-sampling. The combination of the representations from the contractive and the expansive paths is made through a concatenation operation, following the ideas from Chen et al. [2]. The final 1×1 convolution maps each 32-component feature vector to the desired number of segmentation classes K, and then the softmax activation re-scales the values so that the elements of the K-dimensional output lie in the range [0, 1] and sum to 1, thus forming a probability distribution over the segmentation classes.

The architecture of the decoder U-Net is similar to that of the encoder, in this case reading the output of the encoder with a dimensionality of $320 \times 320 \times K$. The final 1×1 convolutional layer of the decoder, considering a sigmoid activation function, maps a 32-component feature vector back to a reconstruction of the original input, with the 3 channels of the LAB colour model.

The W-Net proposal from Xia et al. [1] considered the task of unsupervised image segmentation, using a soft normalized cut loss as a global criterion for the segmentation [18], which measures both the total dissimilarity between different classes and the total similarity within the same classes. We instead consider a supervised setting, in which the output semantic segmentation classes are pre-established, and having the ground-truth segmentation labels informing the training of the encoder part of the W-Net. We specifically used the Dice coefficient loss function [19]. When considering multi-class segmentation problems, the Dice coefficient loss is computed for each class separately, and then summed (i.e., all the classes contribute equally to the final loss, thus addressing issues of class imbalance in the training images).

In turn, the decoder part of the W-Net is trained to minimize a reconstruction loss (i.e., the mean squared error between the predicted and the true LAB values for the images), forcing the encoded representations to contain as much information of the original inputs as possible. In our specific context, the model promotes the idea that semantic categories relate to how one can perceive and distinguish colors and pixel intensities on the images, jointly optimizing a semantic segmentation loss and a reconstruction loss for the input coloured images.

3.2 Hyper-parameters and Model Training Strategy

The selection of hyper-parameters and model training strategies relied on the general guidelines discussed by Xie et al. [20]. The last layer of the first U-Net features a number of nodes compatible with the number of classes in the segmentation task. If considering a binary segmentation mask, this last layer would consist of a single channel with a sigmoid activation function, and the model training would involve minimizing a binary cross-entropy loss or the standard Dice coefficient. The tests reported on this paper always considered multiple classes, and the last layer of the first U-Net consists of multiple channels. In this case, training involves a softmax activation function together with the multi-class Dice coefficient loss. An initial set of tests verified that this loss function always outperformed the traditional categorical cross-entropy loss. Moreover, also through initial tests, we verified that the LAB colour space lead to slightly better results than those obtained with the standard RGB colour space, in terms of both the reconstruction and segmentation quality. Besides converting the input images to the LAB colour space, we also applied a contrast enhancement procedure, analyzing the distribution of pixel intensities and re-scaling the representation to include all intensities that fall within the 2^{nd} and 98^{th} percentiles. This simple pre-processing procedure was also found to improve the results.

Training relied on the Adam [21] optimization algorithm together with a Cyclical Learning Rate (CLR) update procedure, as described by Smith [22]. In more detail, the learning rate varied between 10^{-5} and 10^{-4}, according to a triangular policy that decreases the cycle amplitude by half after each period (i.e., annealing the learning rate), while keeping the base learning rate constant. We used mini-batches of 5 image patches with dimensionality $320 \times 320 \times 3$, created through a generator that considered simple real-time data augmentation procedures (i.e., randomly flipping the input images horizontally, vertically or diagonally when providing them as input to the training algorithm). The loss function regulating the training of the complete network corresponds to a weighted combination of the segmentation (i.e., with a weight of 0.95) and reconstruction (i.e., with a weight of 0.05) losses.

Model training proceeded for up to a maximum of 50 epochs. However, a small validation set was used to define an early stopping criterion, and the training stopped if the validation loss (i.e., the weighted sum of the segmentation and reconstruction losses over the validation data) did not decrease for 5 consecutive epochs. The final model weights were taken from the training epoch with the smallest value for the validation loss.

4 Experimental Evaluation

This section presents the experimental evaluation of the proposed method, comparing it against a standard U-Net model, and also against other proposals in the literature. Section 4.1 starts by presenting the datasets and the experimental methodology, and then Sect. 4.2 discusses the obtained results. The model was

Table 1. Characterization of the aerial imagery datasets.

Dataset	Size (pixels)	Number (train/test)	Resolution (cm/pixel)
ISPRS Vaihingen	2100 × 2100	16/17	9
ISPRS Postdam	6000 × 6000	24/14	5

implemented through a Python deep learning library named Keras[4], and the corresponding source code is available online[5].

4.1 Datasets and Evaluation Metrics

We conducted experiments on two datasets from a previous challenge focused on semantic segmentation of high resolution aerial imagery, namely from the ISPRS 2D Semantic Labeling Contest, with images of the cities of Potsdam and Vaihingen. These datasets have been extensively used within previous studies in the area, as discussed in Sect. 2. The task proposed in the context of both datasets is to classify each pixel in the image with a given class from a fixed set of six classes: impervious surface, car, building, background, low vegetation, and tree. Overall, the Vaihingen dataset has 33 images with different sizes, approximately with 2100 × 2100 pixels each and a Ground Sample Distance (GSD) of 9 cm. We used 12 images for training, 4 for validation, and 17 for testing. The Potsdam dataset has 38 images of 6000 × 6000 pixels each and a GSD of 5 cm. For this dataset, we used a split of 18 images for training, 6 for validation, and 14 for testing. Table 1 features a characterization of both datasets. In both cases, the data splits that we considered are common to those used in most previous studies. All images were initially converted from the RGB to the LAB colour space [15], and we also pre-processed them with a contrast enhancement procedure based on the distribution of pixel intensities, as described in Sect. 3.2.

To assess the quantitative performance of the segmentation methods, we computed the precision, recall, and F1 scores over each segmentation class, macro-averages and class-weighted averages for precision, recall, and F1 scores, and the overall accuracy (OA). Following the practice of other studies leveraging the datasets used in our tests, all the evaluation scores were computed over just five classes, ignoring the background pixels. Additionally, as suggested by the challenge organizers that made available the datasets, we evaluated the results on an alternative ground truth with the borders of the objects eroded by a radius of 3 pixels, to reduce the impact of ambiguous boundaries in the evaluation results.

4.2 Experimental Results

Table 2 features a comparison between the results of state-of-the-art methods in the selected datasets, against our U-Net and W-Net models, specifically considering the overall accuracy. Additionally, Table 3 presents the per-class metrics

[4] http://keras.io.

[5] http://github.com/martamaria96/deep-wnet.

that were used to assess the quantitative performance, together with the percentage of pixels, in each dataset, associated to each of the classes (not considering the background class). Summarizing the results, Table 3 also presents macro-averaged results and a weighted average of the metrics that takes into account the percentage of pixels belonging to each class. Notice that the results presented for previous models are merely indicative, given that we are directly reporting the values given in previous publications, obtained from the author's own tests, and not all those studies used exactly the same experimental protocol.

The results show that the classes corresponding to impervious surfaces, cars, and buildings achieved better results than the other classes. The results also show that the proposed W-Net model outperformed a standard U-Net (i.e., a model using just the first U-Net from Fig. 1, relying just on the loss function corresponding to the Dice coefficient) on both datasets, at the same time achieving comparable results to some of the previous state-of-the-art approaches. A more detailed analysis of previous results is available online, in the leader-boards

Table 2. Comparison in terms of the overall accuracy for different methods over the ISPRS Potsdam and ISPRS Vaihingen test datasets.

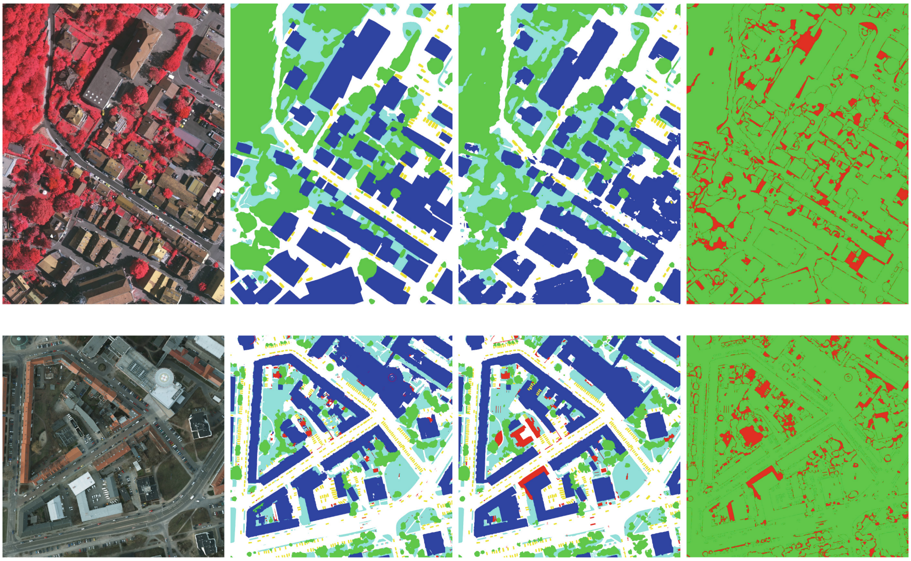
Model	Overall accuracy	
	Vaihingen	Potsdam
RiFCN [6]	-	86.6
CONC_2 [23]	86.5	-
ScasNet [12]	91.1	91.1
UOA [24]	87.6	-
SegNet-RC [3]	89.8	89.0
SDFCN [10]	88.3	89.0
UFMG_4 [25]	89.4	87.9
U-Net (ours)	87.05	89.12
W-Net (ours)	88.08	89.14

Table 3. Per-class precision, recall, and F1 metrics over the ISPRS Potsdam and ISPRS Vaihingen test datasets, with the W-Net model.

Class	Potsdam				Vaihingen			
	Percentage	Precision	Recall	F1-score	Percentage	Precision	Recall	F1-score
Impervious	31.07	92.88	90.59	91.69	30.87	89.82	90.04	89.79
Car	1.88	98.26	90.56	94.22	1.84	73.19	83.66	76.37
Building	26.99	95.29	94.73	94.98	26.94	93.77	91.95	92.75
Low vegetation	23.75	77.13	90.97	83.29	23.63	81.33	73.00	76.66
Tree	16.30	89.91	78.52	83.48	16.72	81.32	92.42	86.08
Macro-average	-	90.69	89.07	89.53	-	83.89	86.22	84.33
Weighted average	-	89.41	89.83	89.29	-	87.15	86.81	86.62

associated to the Potsdam[6] and Vaihingen[7] datasets. Many of the ideas advanced in state-of-the-art models for the task can also be combined with our W-Net approach, perhaps further improving the results. For instance, the state-of-the-art model named UFMG_4 [25] uses convolutional blocks inspired by the DenseNet architecture, i.e. an idea that can be easily combined with our W-Net [26,27].

Figure 2 illustrates the segmentation results for an example image taken from each dataset, where the third column presents the results of the W-Net model, and the fourth column compares our results with the ground truth, highlighting in red the pixels corresponding to incorrect predictions.



(a) Image (b) Ground Truth (c) W-Net (d) Comparison

Fig. 2. Qualitative results on the ISPRS Vaihingen (first row) and Potsdam (second row) datasets. The mask has six categories: impervious surface (white), building (blue), low vegetation (cyan), tree (green), car (yellow) and background/clutter (red). The columns correspond to (i) the original image, (ii) the W-Net results, (iii) the ground truth mask, and (iv) an indication of pixels corresponding to wrong predictions. (A color version of the figure is given in the online version)

5 Conclusions and Future Work

This paper reported on a set of experiments for evaluating the performance of an adapted version of the recently proposed W-Net neural network architecture,

[6] http://www2.isprs.org/commissions/comm2/wg4/potsdam-2d-semantic-labeling.html.

[7] http://www2.isprs.org/commissions/comm2/wg4/vaihingen-2d-semantic-labeling-contest.html.

originally proposed for the segmentation of medical images, on the semantic segmentation of high-resolution satellite images. Through experiments with several distinct datasets used in previous studies and in competitions, we showed that the W-Net architecture is quite effective in this task, outperforming the standard U-Net model and other approaches on common datasets in the area.

Despite the interesting results, there are also many ideas for future work. Several previous studies have proposed alternative connectivity patterns for U-Nets, which we could also explore. Besides the two W-Net architectures that inspired the model used in our tests, other studies have also proposed to stack/bridge multiple U-Nets together [28–30], with features going sequentially from the first U-Net to the last in an attempt to better capture high-order spatial relationships. Several ideas from these other approaches can also be borrowed, e.g. related to the use of residuals [31] or other intermediate supervisions. Other authors still have proposed to combine dense connectivity with the original U-Net architecture, following the general design of DenseNets for image classification [26,27]. In particular, each top-down or bottom-up module can feature a dense block with densely connected convolutional layers. Besides dense connections in the encoder and decoder blocks, one can also consider dense connections across both these parts [32]. Tang et al. proposed the coupled U-Net model, i.e. a hybrid of the dense U-Net and the stacked U-Net that integrates the merits of both dense connectivity and multi-stage top-down and bottom-up refinement [33]. Given several stacked U-Nets, these authors proposed to add shortcut connections for each U-Net pair, thus generating the coupled U-Net. Similar ideas can easily be considered as extensions to our model.

Besides taking ideas from previous developments in fully-convolutional networks for image segmentation, other recent developments in CNNs for image classification can also be integrated into the proposed approach. Examples include octave convolutions [34] or attention augmented CNNs [35].

Currently ongoing experiments relate to the use of W-Net model variants for the semantic segmentation of historical aerial photos available as gray-scale images. The segmentation of historical aerial photos can have many interesting practical applications, and we are interested in exploring the segmentation of building footprints for dasymetric disaggregation [36] of historical census data. In particular, we are studying the application of W-Net models for jointly coloring and segmenting the aerial photos, in which the first fully-convolutional part performs semantic segmentation from gray-scale inputs, and the second part returns a colored version of the original image. The entire model can be trained with modern datasets of colored aerial photos, such as the ones used in the experiments reported on this paper, ignoring the color information from the input images in the encoder part, and instead using the color information for computing the loss function from the second part of the model.

Acknowledgements. This research was supported through Fundação para a Ciência e Tecnologia (FCT), through the project grants with references PTDC/EEI-SCR/1743/2014 (Saturn), PTDC/CTA-OHR/29360/2017 (RiverCure), and PTDC/CCI-CIF/32607/2017 (MIMU), as well as through the INESC-ID multi-annual funding from the PIDDAC programme (UID/CEC/50021/2019). We also gratefully acknowledge the support of NVIDIA Corporation, with the donation of two Titan Xp GPUs used in the experiments reported on the paper.

References

1. Xia, X., Kulis, B.: W-Net: a deep model for fully unsupervised image segmentation. arXiv preprint arXiv:1711.08506 (2017)
2. Chen, W., et al.: W-Net: Bridged U-Net for 2D medical image segmentation. arXiv preprint arXiv:1807.04459 (2018)
3. Audebert, N., Le Saux, B., Lefèvre, S.: Beyond RGB: very high resolution urban remote sensing with multimodal deep networks. ISPRS J. Photogramm. Remote Sens. **140**, 20–32 (2018)
4. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. Mou, L., Zhu, X.X.: RiFCN: recurrent network in fully convolutional network for semantic segmentation of high resolution remote sensing images. arXiv preprint arXiv:1805.02091 (2018)
7. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
8. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5188–5196 (2015)
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations (2015)
10. Chen, G., Zhang, X., Wang, Q., Dai, F., Gong, Y., Zhu, K.: Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **11**(5), 1633–1644 (2018)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
12. Liu, Y., Fan, B., Wang, L., Bai, J., Xiang, S., Pan, C.: Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. ISPRS J. Photogramm. Remote Sens. **145**, 78–95 (2018)
13. Mnih, V.: Machine learning for aerial image labeling. Ph.D. thesis, University of Toronto (2013)
14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
15. Jain, A.K.: Fundamentals of Digital Image Processing. Prentice Hall, Upper Saddle River (1989)

16. Xu, B., Huang, R., Li, M.: Revise saturated activation functions. arXiv preprint arXiv:1602.05980 (2016)
17. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the International Conference on Machine Learning, pp. 448–456 (2015)
18. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 888–905 (2000)
19. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 240–248. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_28
20. Xie, J., He, T., Zhang, Z., Zhang, H., Zhang, Z., Li, M.: Bag of tricks for image classification with convolutional neural networks. arXiv preprint arXiv:1812.01187 (2018)
21. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (2014)
22. Smith, L.N.: Cyclical learning rates for training neural networks. In: Proceeedings of the IEEE Winter Conference on Applications of Computer Vision, pp. 464–472 (20170
23. Forbes, T., He, Y., Mudur, S., Poullis, C.: Aggregated residual convolutional neural network for multi-label pixel wise classification of geospatial features. In: Online Abstracts of the ISPRS Benchmark on Urban Object Classification and 3D Building Reconstruction (2018)
24. Lin, G., Shen, C., van den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3194–3203 (2016)
25. Nogueira, K., Mura, M.D., Chanussot, J., Schwartz, W.R., Santos, J.A.: Dynamic multi-scale segmentation of remote sensing images based on convolutional networks. arXiv preprint arXiv:1804.04020 (2018)
26. Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y.: The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: Proceedings of the Workshops at the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1175–1183 (2017)
27. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., Heng, P.-A.: H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imaging **37**(12), 2663–2674 (2018)
28. Newell, A., Yang, K., Deng, J.: Stacked Hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
29. Sun, T., Chen, Z., Yang, W., Wang, Y.: Stacked U-Nets with multi-output for road extraction. In: Proceedings of the Workshops at the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 187–192 (2018)
30. Khalel, A., El-Saban, M.: Automatic pixelwise object labeling for aerial imagery using stacked U-Nets. arXiv preprint arXiv:1803.04953 (2018)
31. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual U-Net. IEEE Geosci. Remote Sens. Lett. **15**(5), 749–753 (2018)
32. Zhang, J., Jin, Y., Xu, J., Xu, X., Zhang, Y.: MDU-Net: multi-scale densely connected U-Net for biomedical image segmentation. arXiv preprint arXiv:1812.00352 (2018)

33. Tang, Z., Peng, X., Geng, S., Zhu, Y., Metaxas, D.: CU-Net: coupled U-Nets. In: Proceedings of the British Machine Vision Conference, pp. 305–316 (2018)
34. Chen, Y., et al.: Drop an octave: reducing spatial redundancy in convolutional neural networks with octave convolution. arXiv preprint arXiv:1904.05049 (2019)
35. Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q.V.: Attention augmented convolutional networks. arXiv preprint arXiv:904.09925 (2019)
36. Monteiro, J., Martins, B., Pires, J.M.: A hybrid approach for the spatial disaggregation of socio-economic indicators. Int. J. Data Sci. Anal. **5**(2–3), 189–211 (2018)