A COMPARATIVE STUDY FOR BUILDING SEGMENTATION IN REMOTE SENSING IMAGES USING DEEP NETWORKS: CSCRS Istanbul Building Dataset and Results

B. Amirgan[a, d], B. Awad[b,d], I. Erer[b,d], N. Musaoğlu[a,c,d]

[a]ITU, Center for Satellite Communications and Remote Sensing, - (burcu.amirgan, musaoglune)@cscrs.itu.edu.tr
[b]ITU, Electrical and Electronic Engineering Faculty, Electronics and Communication Department, - (awad, ierer)@itu.edu.tr
[c]ITU, Civil Engineering Faculty, Geomatic Engineering Department, - musaoglune@itu.edu.tr
[d]34469 Maslak, Istanbul, TURKEY

KEY WORDS: Semantic segmentation, very high resolution satellite imagery, building detection, deep learning, fully convolutional neural networks, CSCRS Istanbul Building Dataset.

ABSTRACT:

Building semantic segmentation is an exceedingly important issue in the field of remote sensing. A new building dataset as created consisting of very high-resolution optical satellite images provided by the Center for Satellite Communications and Remote Sensing (CSCRS). The imagery is obtained by Pleiades satellite and have a resolution of 0.5 meters. Segmentation results have been obtained using post-FCN architectures. Architectures examined in this work fall under one of few categories. The first category is Encoder-Decoder Network: an encoder that reduces the spatial resolution of the data and a decoder that recreates the lower resolution result of the encoder and upsamples it. The second category is Feature Pyramid Network, in this type of network scene information is aggregated across pyramid structures which produce more comprehensive results. The third category is Dilated Network, due to its atrous structure, which can calculate any layer at any desired resolution, with the presence of holes in the filter. The final category is Attention-Based Network, in these networks, certain aspects of the data are emphasized while other aspects are ignored. After this work, it can be seen that according to several metrics Dilated and Attention-Based Networks perform better than their counterparts. As a result of the training of 100 epochs with the data set in architectures belonging to Dilated and Attention-Based Networks, IoU values above 0.90 were obtained.

1. INTRODUCTION

Determining building boundaries as accurately as possible is a tremendously significant challenge faced in the field of remote sensing (Huang et al., 2016). It forms the basis for a myriad of important applications such as land use and land cover classification, urban sprawl monitoring, and risk assessment.

In recent years, deep neural networks have been the corner stone of every improvement made in the field of image semantic segmentation. However, deep models have been developed and tested for natural images and have only recently been used in the remote sensing field. One of the reasons for this is the lack of labeled data. Therefore, in this work a dataset for building semantic segmentation is introduced for public use. Geographic conditions and cultural influences present differences in building structures. For this reason, many building datasets published as open-source may not offer the desired performance when tested in dissimilar geographical regions. For this reason, the dataset which consists of very high resolution (VHR) satellite imagery is meant to represent Istanbul city and its natural diversity. Moreover, performance of the most popular deep networks is measured for this dataset to set a baseline for the current state of building semantic segmentation in this region.

Fully convolutional neural networks (FCNs) brought on an area of change by replacing the final fully connected layer of typical convolutional deep networks (CNN) with a convolutional layer (Long et al., 2015). Compared to standard CNNs, FCNs provide a significant improvement in speed, accuracy, and efficiency. Furthermore, FCNs allow input images of any arbitrary size which eliminates the need for uniform size across all the images. The main purpose of FCNs is to fine-tune classification networks and transfer learned weights of previously networks.

FCNs lack the ability to utilize local information present and emphasize global information (Ulku and Akagündüz, 2022). This does not bode well for building semantic segmentation since it has an abundance of locally dense information. Therefore, this work focuses on examining network architectures published in the post-FCN era.

UNet architecture takes its name from its structure that resembles a “U” shape in the way it narrows and then expands symmetrically (Ronneberger et al., 2015). This architecture can adapt to different problems easily. The defining feature of this architecture is that it replaces pooling operators with upsampling operators which increasing the outputs resolution in the decoder layers. LinkNet architecture was proposed as a real-time application by (Chaurasia and Culurciello, 2017). The main difference from other architectures is the method used in connecting the encoder to the decoder. Each encoder level is connected to its corresponding decoder, which causes the information that would be lost at the encoder to be preserved. This both reduces processing time and increases accuracy. SegNet architecture was proposed by (Badrinarayanan et al., 2015). This architecture uses and Encoder-Decoder Network followed by a pixel-wise classification layer. Furthermore, an important factor that distinguishes SegNet from other architectures is its use of indices to connect corresponding pooling layers across the decoder and encoder.

FPN architecture was initially created as multi-class image segmentation method based on FCN architecture (Seferbekov et

al., 2018). FPN consists of bottom-up and top-down paths and lateral connections to connect them. There is a pyramid level for each stage in the bottom-up path. Each stage is added to the corresponding top-down path level with a lateral connection and the bottom-up path. PspNet architecture was proposed for the FCN based pixel prediction framework by (Zhao et al., 2016). Different region-based semantic segmentation is done with the Pyramid pooling module. A semantic segmentation model with local and global clusters is suggested for state-of-the-art scene parsing. To reduce information loss between different sub-regions, it is recommended to combine information at different scales hierarchically.

DeepLabV3 architecture was proposed to effectively expand the field of view to capture multi-scale context. It uses atrous convolution gradually and in parallel with the ASPP structure (Chen et al., 2017). On the other hand, DeepLabV3+ architecture makes sharper segmentation at the borders with the combination of FPN Network and Encoder-Decoder Network features (Chen et al., 2018). DeepLabV3+ is created by combining ASPP used in DeepLabV3 with a simple encoder-decoder.

PAN architecture consists of a combination of Feature Pyramid Attention (FPA) and Global Attention Upsample (GAU) methods as well as the encoder-decoder structure (Li et al., 2018). FPA provides context information at different scales while GAU is a decoder method that effectively distributes features at different scales taking in consideration both local and global information. MA-Net architecture identifies focal features with their global dependencies to extract context information using multi-scale feature fusion (Fan et al., 2020). This is a novel architecture based on improving the existing UNet architecture. Ma-Net consists of two different blocks: Position-wise attention Block (PAB), which used to capture spatial dependencies of global feature maps and finds spatial dependencies between pixels. Multi-scale Fusion Attention Block (MFAB) which combines high-level and low-level feature maps used to locate exchange dependencies between any feature maps.

In applications made for building segmentation, when looking at the results obtained in (Xu et al., 2022), SpaceNet Las Vegas dataset the accuracy is %77.0, %78.5, %78.1 for UNet, Deeplabv3+, PspNet architectures respectively. Meanwhile, the Massachusetts building dataset IoU metric, the MA-Net architecture is given as %72.2. According to the experimental results obtained in (Wu et al., 2021) WHU Building dataset accuracy is %86.2, %84.9, %85.6, %87.3 for UNet, LinkNet, SegNet, Deeplabv3+ architectures.

Dividing exiting neural networks into separate categories facilitates a simpler method for testing as many neural networks as possible. For this reason choosing these categories is of extreme importance. (Minaee et al., 2020), groups some of the most used deep learning architectures into separate groups based on their technical contribution. According (Li et al., 2018) deep networks can be grouped into these categories: Encoder-decoder, Global Context Attention and Spatial Pyramid structures based on their architecture type. (Jiang et al., 2022), looks at segmentation networks used in the remote sensing and suggests a grouping based on their merit in this field. Based on this, 4 categories are chosen as base for this work: Encoder-Decoder Network, Feature Pyramid Network, Dilated Network, and Attention-Based Network. This work will present a comprehensive comparison of the previously mention networks for building semantic segmentation. Moreover, this comparison will be conducted using the dataset presented in this work.

The flow of the paper is organized as follows: the dataset will be introduced in the second section. The third section covers the deep neural networks belonging to each of these categories: Encoder-Decoder Network, Feature Pyramid Network, Dilated Network, and Attention-Based Network, which will be used to conduct the comparison of building semantic segmentation based on the introduced dataset. Furthermore, the post-processing accomplished using Conditional Random Fields (CRF) is explained in this section. In the fourth section, the training process of the architectures belonging to the deep networks is explored, and comparisons are made between different networks in accordance to widely used metrics such as Intersection over Union (IoU), overall accuracy (OA), and F1-score. Chapter five offers some conclusions and insight regarding past and future research.

2. CSCRS ISTANBUL BUILDING DATASET

Istanbul is one of the most populated cities in the world and the largest city in Turkey and Europe. Due to its unique geographical location and diverse history, Istanbul's buildings have a great structural and visual variety. Furthermore, the density of the building changes dramatically across the city giving rise to both densely and sparsely distributed buildings. For these reasons it is important to have an accurate data set that represents the city alongside the best model that are able to take advantage of it.

In this work a novel dataset containing images obtained from Pleiades satellite is created by the help of Center for Satellite Communications and Remote Sensing (CSCRS) and is shown in Figure 1.

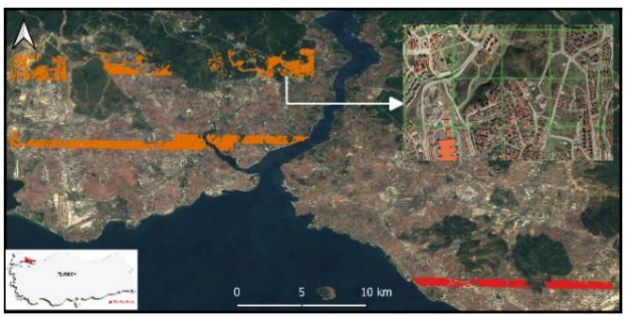

Figure 1. CSCRS Istanbul Building Dataset.

CSCRS Istanbul Building Dataset covers certain regions of the Anatolian and European sides of Istanbul. The dataset is comprised of very high resolution (VHR), pansharp images, with three channels Red, Green, Blue (RGB), quantized to 8 bits and spatial resolution of 0.5 m. One image in this dataset has the size of 1500x1500 pixels and is further divided into 9 tiles of 512x512 pixels. The size of the dataset is approximately 1.0 GB. Building roof boundaries were delineated by on screen digitizing using a GIS environment. Each individual mask represents a building area while non- delineated regions represent the background making this a binary dataset. The data set consists of two parts: the first part is 9764 building masks that were delineated over 21 images (red area in Figure 1) representing the Anatolian Side of the dataset. The second part consists of 30047 building masks that were delineated over 129 images (orange area in Figure 1) representing the European side of the data set. The masks of the remaining images (550 images) are to be added to the data set at later time. After the dataset is completed, it will be publicly accessible from the ("ITU - Satellite Communication and Remote Sensing Center," n.d.) website. The 150 Pleiades satellite images of the dataset were divided into %70 train, %20 validation, and

%10 test data. In total the dataset contains approximately 40,000 building masks.

Finally, due diligence is taken to ensure that the dataset represents Istanbul as much as possible. For this reason, the dataset contains a great variety of building structure types, for example it contains small buildings and large ones, complex structures and simple ones, densely populated areas and sparsely populated areas. Furthermore, the regions contained in the dataset also are of different types, for example: industrial areas, residential areas, forest, and many more as seen in Figure 2.

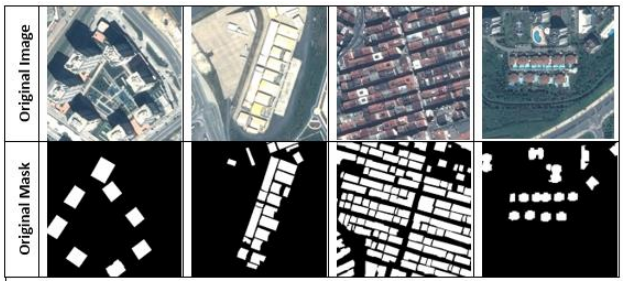

Figure 2. CSCRS Istanbul Building Dataset content includes different types.

3. BUILDING SEGMENTATION WITH DEEP NETWORKS

In this section, deep networks used in in this work are explored based on their previously determined categorization. Following that, a brief explanation of the post-process method is conducted.

3.1 Encoder-Decoder Network

The first category is the Encoder-Decoder Network, which typically consists of two parts an encoder and decoder as seen in Figure 3. At the encoder stage, pooling and strides are used to drop the resolution of the image, and low-resolution feature maps are created. This leads to the preservation of context information, but caused the degradation of the loss of spatial information. At the decoder stage upsampling is accomplished using pooling index and full convolutions which leads to an equal increment of resolution. Hence, feature extraction is achieved and spatial information loss in the encoder is recovered. Skip connections can be used to transport the information from feature maps located at the same level in the encoder and decoder. This allows networks to capture low-level features without focusing on global context information. The most commonly used architectures belonging to this network type are UNet, LinkNet, SegNet, etc.

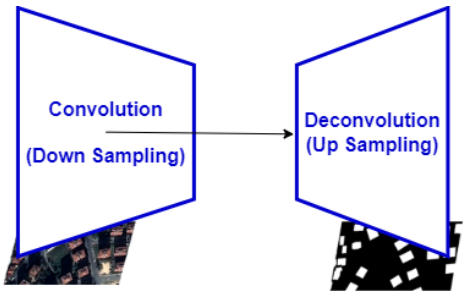

Figure 3. Encoder-Decoder Network.

3.2 Feature Pyramid Network

The second category, Feature Pyramid Network seen in Figure 4, provides great improvement in the identification of objects at different scales (Lin et al., 2016). It was created to capture multi-scale features with pyramidal hierarchy. It creates multi-scale feature maps in the using full convolutions regardless of the input image's size. This is very useful for capturing objects of different sizes, which proves to be very useful in the case of building detection. Feature Pyramid Network is used for both object detection and object segmentation (Lin et al., 2016; Seferbekov et al., 2018). The most commonly used architectures belonging to this network type are Feature Pyramid Network (FPN), Pyramid Scene Parsing Network (PspNet), etc.

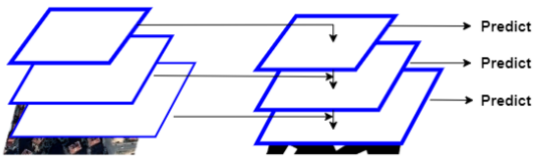

Figure 4. Feature Pyramid Network.

3.3 Dilated Network

Dilated Network defining feature is the fact that its convolutions contain holes with an atrous structure. This atrous structure allows any layer can be calculated at any desired resolution (Chen et al., 2016a). At the same time, the field of view of the filters can be expanded without increasing the number of parameters and the amount of calculations needed. With the Atrous Spatial Pyramid Pooling (ASPP) structure, objects can be captured at multiple scales with multiple parallel filters at different rates as seen in Figure 5. The most commonly used architectures belonging to this network are DeepLabV3, DeepLabV3+, etc.

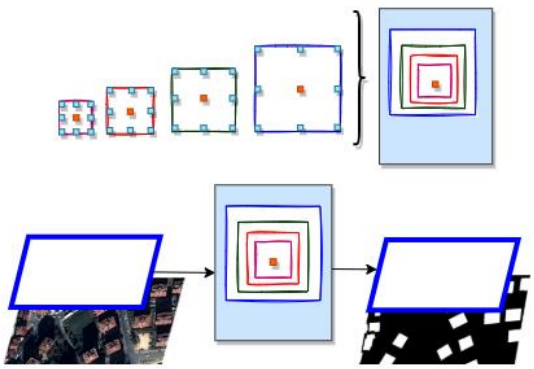

Figure 5. Dilated Network.

3.4 Attention-Based Network

Attention-Based Network have become very popular in recent years and they tend to produce fairly accurate results. In the Attention-Based Network as seen in Figure 6, each pixel is assigned a weight value, highlighting certain areas of the data while other areas are ignored(Chen et al., 2016b; Oktay et al., 2018). Multi-scale features at each pixel location also have an attention mechanism that assigns soft weights to them which provide an improvement in extracting objects of different sizes. The most commonly used architectures belonging to this network type are Pyramid Attention Network (PAN), Multi-Scale Attention Network (MA-Net), etc.

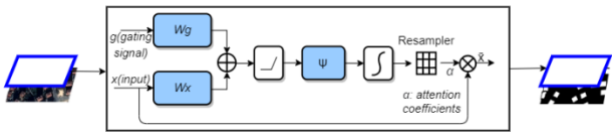

Figure 6. Attention-Based Network.

3.5 Post processing with CRF

Many segmentation architectures lack an emphasis on intersection areas. Post-process is preferred in segmentation applications to avoid noise on the borderlines and to obtain better clarity at the edges. Post-process also provides an improvement in metric values. CRF algorithms are one of the most preferred methods in segmentation applications. There are various CRF models and these models are preferred for the application (Dhawan et al., 2019). Linear CRFs are inherently applied to linear problems. Grid CRFs are two-dimensional and by nature, 1 node is connected to 4 nodes around it (i.e., in a grid structure). Grid CRFs are used in pattern recognition or simple image segmentation applications. Dense CRFs are used in structures containing complex relationships. This method gives the best results among CRF models for image segmentation. The fully connected CRFs version of this model is preferred for operation complexity and time-saving. Fully connected CRFs are defined by a linear combination of Pairwise edge potentials and Gaussian kernels (Krähenbühl and Koltun, 2012). CRFs maximize accurate labelling between similar pixels by modelling relationships between object classes. Post process was applied to all results using Fully Connected CRF as seen in (Dhawan, 2019; Lucas, n.d.). Post processing examples are as shown in Figure 7.

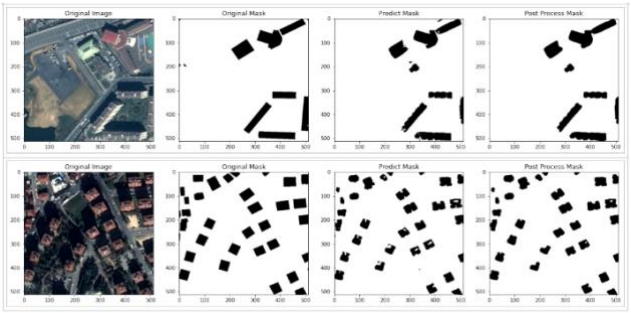

Figure 7. Post processing results on sample images from the CSCRS Istanbul Building dataset.

In the Figure 7 the columns show the original RGB images, the original mask images, the prediction images, and the post-processing results.

4. RESULTS AND DISCUSSIONS

Testing for each of the previous architectures was conducted using the codes provided by the authors of the original papers. UNet, LinkNet, FPN, PspNet architecture use Keras Segmentation Model library (Lakubovskii, 2019). While DeepLabV3, DeepLabV3+, PAN, MA-Net architectures use PyTorch Segmentation Model library (Lakubovskii, 2020). and SegNet architecture also uses the only Keras (Divam, 2019).

The hyperparameters used for architectures trained with Keras and PyTorch Segmentation Models are as follows: The backbones used are VGG16 and EfficientNet. The batch size is selected to be 4. The number for epochs is determined to 50 while the input image size is set to 512×512 px. Learning rate is chosen as 0.0001. As optimizers ADAM and SGD were used. Several loss functions were used depending on the model such as: Total/Dice and binary cross-entropy losses. Kaggle provides free online access to NVIDIA TESLA P100 GPU for training. Hence, the training in Kaggle environment. Overall Accuracy (OA), Intersection over Union (IoU), and F1-score were used as metrics to quantify the quality of each model. Metric values calculated following 50 epochs of training can be seen in Table 1.

Deep Network	Model	OA	IoU	F1-Score
Encoder-Decoder	UNet	0.90060	0.86848	0.92406
	LinkNet	0.85863	0.82579	0.89636
	SegNet	0.90339	0.85165	0.70445
Feature Pyramid	FPN	0.88373	0.83781	0.90465
	PspNet	0.82489	0.74738	0.82719
Dilated	DeepLabV3	**0.93736**	0.91592	0.95258
	DeepLabV3+	0.93693	**0.91737**	**0.95324**
Attention-Based	PAN	0.92500	0.90390	0.94532
	MA-Net	0.93630	0.91614	0.95239

Table 1. Comparison of different segmentation architectures on the test datasets.

According to Table 1, the best segmentation results were obtained using Attention-Based and dilated Network architectures. DeepLabV3 achieves the highest overall accuracy, while DeeplabV3+ achieves the highest IoU and F1-score. Other architectures are both less modern and have lower complexity structures therefore they give worse results. Architectures that preserve both local and global information are expected to give higher accuracy values, this was indeed the case. With segmentation networks, building areas can be detected as unique-mask as well as multi-mask detection. This situation is also because the ground truth masks of complex structured closely located building areas are selected collectively in the data set. Visual representation of the results can be seen in Figure 7.

Training metric values were observed to be higher than test metric values. This is due to the nature of artificial intelligence, and in general, we cannot obtain metric values as high in test data as in training data. Prediction mask images were post-processed using the Fully Connected CRF method. After the post-processing, the noises at the segmentation borders softened and the borderlines became sharper.

The best performing models: DeepLabV3, DeepLabV3+, PAN, and MA-Net were trained for another 50 epochs. This was done in order to identify the best-performing model. The results can be seen in Table 2.

Deep Network	Model	OA	IoU	F1-Score
Dilated	DeepLabV3	0.93502	0.91496	0.95190
	DeepLabV3+	0.94318	0.92388	0.95722
Attention-Based	PAN	0.93066	0.90844	0.94816
	MA-Net	**0.94430**	**0.92620**	**0.95840**

Table 2. Comparison of DeepLabV3, DeepLabV3+, PAN and MA-Net architectures on the test datasets.

As a result of 100 epoch training, a slight improvement was achieved in the metric values and the best values were observed to shift from dilated Networks such as DeeplapV3 and

DeeplapV3+ to the attention-based MA-Net architecture. However, it is important to note that all the dilated and Attention-Based architectures achieve high results in all metrics.

The results shown in Figure 8, represent the deep network categories belonging to the best performing two categories seen in Table 2. Furthermore, depict the variety of building types and distribution present in the dataset. For example, both sparsely and densely populated areas are shown. Moreover, the differences in results of the top preforming models are hard discern, this is also can be seen in Table 2, where all the top models have accuracy metrics exceeding %90.

5. CONCLUSION

Despite the fact that building semantic segmentation is a significant area in the remote sensing field, the lack of data makes the problem much more difficult to approach. In this work a novel dataset is presented to mitigate this issue. This dataset is meant to represent the city of Istanbul. For this reason, great care was taken to ensure that the dataset contains as many various examples of buildings in Istanbul as possible. Furthermore, the diversity of building types, structures and distribution was emphasized upon. Region diversity was also taken in consideration when constructing this dataset. Lastly, this work presents a comparative study for the performance of deep neural networks using this dataset. This is meant to be a baseline for future works wishing to use this dataset or conduct building semantics segmentation in Istanbul or Turkey. The networks compared where divided into four categories: Encoder-Decoder Networks, Feature Pyramid Networks, Dilated Networks, and Attention-Based Networks. It is concluded that Attention-Based and Dilated Networks achieve similarly good results. Whereas MA-Net achieves the highest score across all metrics.

Figure 8. Comparison of DeepLabV3, DeepLabV3+, PAN and MA-Net architectures prediction on the test datasets.

REFERENCES

Badrinarayanan, V., Kendall, A., Cipolla, R., 2015. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

Chaurasia, A., Culurciello, E., 2017. LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation. 2017 IEEE Vis. Commun. Image Process. VCIP 2017 2018-January, 1–4. https://doi.org/10.1109/VCIP.2017.8305148

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2016a. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L., 2016b. Attention to Scale: Scale-Aware Semantic Image Segmentation. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-December, 3640–3649. https://doi.org/10.1109/CVPR.2016.396

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 11211 LNCS, 833–851. https://doi.org/10.1007/978-3-030-01234-2_49

Dhawan, A., 2019. Post Processing of Image Segmentation using CRF [WWW Document]. URL https://github.com/dhawan98/Post-Processing-of-Image-Segmentation-using-CRF (accessed 1.27.22).

Dhawan, A., Bodani, P., Garg, V., 2019. Post Processing of Image Segmentation using Conditional Random Fields. 2019 6th Int. Conf. Comput. Sustain. Glob. Dev.

Divam, G., 2019. Implementation of Segnet, FCN, UNet , PSPNet and other models in Keras. [WWW Document]. URL https://github.com/divamgupta/image-segmentation-keras (accessed 1.28.22).

Fan, T., Wang, G., Li, Y., Wang, H., 2020. Ma-net: A multi-scale attention network for liver and tumor segmentation. IEEE Access 8, 179656–179665. https://doi.org/10.1109/ACCESS.2020.3025372

Huang, Z., Cheng, G., Wang, H., Li, H., Shi, L., Pan, C., 2016. Building extraction from multi-source remote sensing images via deep deconvolution neural networks. Int. Geosci. Remote Sens. Symp. 2016-November, 1835–1838. https://doi.org/10.1109/IGARSS.2016.7729471

ITU - Satellite Communication and Remote Sensing Center [WWW Document], n.d. URL https://web.cscrs.itu.edu.tr/homepage/ (accessed 4.15.22).

Jiang, B., An, X., Xu, S., Chen, Z., 2022. Intelligent Image Semantic Segmentation: A Review Through Deep Learning Techniques for Remote Sensing Image Analysis. J. Indian Soc. Remote Sens. 1–14. https://doi.org/10.1007/S12524-022-01496-W/TABLES/1

Krähenbühl, P., Koltun, V., 2012. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. Adv. Neural Inf. Process. Syst. 24 25th Annu. Conf. Neural Inf. Process. Syst. 2011, NIPS 2011.

Lakubovskii, P., 2020. Segmentation models with pretrained backbones: PyTorch [WWW Document]. URL https://github.com/qubvel/segmentation_models.pytorch (accessed 1.27.22).

Lakubovskii, P., 2019. Segmentation models with pretrained backbones: Keras and TensorFlow Keras [WWW Document]. URL https://github.com/qubvel/segmentation_models (accessed 1.27.22).

Li, H., Xiong, P., An, J., Wang, L., 2018. Pyramid Attention Network for Semantic Segmentation. Br. Mach. Vis. Conf. 2018, BMVC 2018.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2016. Feature Pyramid Networks for Object Detection.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 07-12-June-2015, 431–440. https://doi.org/10.1109/CVPR.2015.7298965

Lucas, B., n.d. Python wrapper to Philipp Krähenbühl's dense (fully connected) CRFs with gaussian edge potentials. [WWW Document]. URL https://github.com/lucasb-eyer/pydensecrf (accessed 1.27.22).

Minaee, S., Boykov, Y.Y., Porikli, F., Plaza, A.J., Kehtarnavaz, N., Terzopoulos, D., 2020. Image Segmentation Using Deep Learning: A Survey. IEEE Trans. Pattern Anal. Mach. Intell. https://doi.org/10.48550/arxiv.2001.05566

Oktay, O., Schlemper, J., Folgoc, L. Le, Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: Learning Where to Look for the Pancreas. https://doi.org/10.48550/arxiv.1804.03999

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 9351, 234–241.

Seferbekov, S.S., Iglovikov, V.I., Buslaev, A. V, Shvets, A.A., 2018. Feature Pyramid Network for Multi-Class Land Segmentation. Comput. Vis. Pattern Recognit.

Ulku, I., Akagündüz, E., 2022. A Survey on Deep Learning-based Architectures for Semantic Segmentation on 2D Images. https://doi.org/10.1080/08839514.2022.2032924

Wu, M., Shu, Z., Zhang, J., Hu, X., 2021. Hrlinknet: Linknet with High-Resolution Representation for High-Resolution Satellite Imagery 2504–2507. https://doi.org/10.1109/IGARSS47720.2021.9554601

Xu, H., Zhu, P., Luo, X., Xie, T., Zhang, L., 2022. Extracting Buildings from Remote Sensing Images Using a Multitask Encoder-Decoder Network with Boundary Refinement. Remote Sens. 2022, Vol. 14, Page 564 14, 564. https://doi.org/10.3390/RS14030564

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2016. Pyramid Scene Parsing Network. Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017 2017-January, 6230–6239. https://doi.org/10.1109/CVPR.2017.660