Sci-Net: a Scale Invariant Model for Buildings Segmentation from Aerial Images

Hasan Nasrallah [*] Mustafa Shukor [†] Ali J. Ghandour [‡]

Abstract

Buildings' segmentation is a fundamental task in the field of earth observation and aerial imagery analysis. Most existing deep learning-based methods in the literature can be applied to fixed or narrow-ranged spatial resolution imagery. In practical scenarios, users deal with a broad spectrum of image resolutions. Thus, a given aerial image often needs to be re-sampled to match the spatial resolution of the dataset used to train the deep learning model, which results in a degradation in segmentation performance. To overcome this, we propose a Scale-invariant Neural Network (Sci-Net) that can segment buildings present in aerial images at different spatial resolutions. Specifically, our approach leverages UNet hierarchical representations and dilated convolutions to extract fine-grained multi-scale representations. Our method significantly outperforms other state of the art models on the Open Cities AI dataset with a steady improvements margin across different resolutions.

1. Introduction

Semantic segmentation is one of the most investigated computer vision topics, where the aim is to provide a pixel-wise classification over several classes in a particular image. With the current deep learning breakthrough, several fully connected neural network models are proposed for semantic segmentation [3, 24, 30, 33, 38, 44], and employed for several applications such as autonomous driving [20], buildings footprint extraction [8, 19] and medical applications [23].

Buildings' footprint segmentation from aerial imagery [14, 17, 26, 31, 34] is important for several applications such as urban planning, disaster assessment, and change analysis. In addition, several online challenges have addressed deep-leaning-based buildings' segmentation topics such as different nadir-angles and non-optical and noisy data. DIUx's xView2 [13], SpaceNet challenges (1, 2, 4, 5 and 7) [37], and Open Cities AI [22] are examples of recent well-known competitions focusing on this research topic.

Existing buildings' segmentation models are trained on a fixed spatial resolution. Training a robust and accurate deep learning model capable of segmenting buildings from a wide range of input spatial resolution images remains little investigated in the literature. Sate-of-the art buildings' segmentation models perform well on test images of the exact spatial resolution as the training dataset used to generate the model. However, in practical scenarios, test images might be of various resolutions, resulting in non-optimal performance. This is due to several issues, first, the fragmentation of building segments in high-resolution images, as the model fails to acquire a large enough receptive field, and it becomes harder to classify pixels closer to the center of large buildings accurately. Second, in low-resolution test images the model seamlessly merges the pixels of small building instances with the background, this is known as under-segmentation of buildings' instances, where the model suffers from an over-segmentation of the background, leaving false negative holes in the mask. These problems are commonly addressed by resampling the test images to match the resolution of the training dataset. However, as the gap between inference and training image resolution increases in both directions, the quality of the resultant segmentation masks decreases.

In this context, we propose Scale-invariant neural network (Sci-Net) that is able to extract a multi-scale representation with wider receptive field of an aerial image to cope with varying spatial resolutions during test. The contribution of this paper is two-fold: *(i)* show that existing SoA buildings' segmentation models often suffer from fragmentation, under-segmentation, or over-segmentation, *(ii)* propose Sci-Net, a new model that significantly outperforms other approaches on different test image resolutions.

The rest of the paper is organized as follows: Section 2 reviews current research in the literature related to buildings' segmentation from aerial images. Practical problems in the process of buildings' segmentation from aerial images are discussed in Section 3. Section 4 introduces the proposed Sci-Net model and relevant background details. Section 5 describes the Open Cities AI dataset and the experimental results. Finally, the manuscript is concluded in Section 6.

[*] Lebanese University, Beirut, Lebanon

[†] Sorbonne University/ISIR (MLIA), Paris, France

[‡] National Center for Remote Sensing - CNRS, Lebanon. Corresponding author: aghandour@cnrs.edu.lb

2. Related Work

In this section, we details some related work relevant to the problem in hand.

2.1. Multi-Scale Context for Semantic Segmentation:

Capturing multi-scale context has gained a lot of attention due to its importance for semantic segmentation.

To this end, several methods have been proposed. Image Pyramid methods [4, 11, 28, 32] are one of the first approaches, where the feature extractor is applied on the same input with varying resolutions, then different aggregation mechanisms are applied to gather the features from all resolutions.

Encoder-Decoder approaches [1, 12, 18, 27, 33] have proven to be successful, where the input is processed by a feature extractor that reduces the spatial size and increases the number of channels progressively, then the decoder tries to decode the features and produce the output map. These approaches exploit the multi-scale feature in the encoder (*e.g.* using skip connections [33] or transferred pool indices [1]).

Spatial Pyramid Pooling are another way to capture global context, methods like DeepLab [3] rely on dilated convolutions [43] to process the feature map using different rates in parallel. PSPNet [44] proposes to process different pooled feature maps with different resolution. In [45], authors introduce a modified version of squeeze and excitation blocks denoted as Squeeze and Attention (SA) module. SA module re-weights spatial locations in the features according to the local and global context, thus, improving semantic segmentation. Dilated or Atrous convolutions are widely used in this context [7, 39]

2.2. Multi-Scale Context for Semantic Segmentation in Remote Sensing:

Being able to segment objects with multiple resolutions is of high interest for the remote sensing community. While most of the work proposed in the computer vision community can be adapted to satellite images, several work have been also proposed in this context. [15] leverage multi-task learning and distillation to produce several output maps depending on the buildings size. [25] propose to reuse previous feature maps by the help of the connections from each layer to the same sized subsequent layers. [29] propose an efficient model based on separable factorized residual block in addition to dilated convolution.

Authors in [31] introduce channel relation module that applies global average pooling over the features and spatial relation module to obtain global spatial relation features capable of capturing global contextual dependencies for identifying various objects. However, the validation of their results is based on the Postdam dataset with a fixed high resolution.

Furthermore, authors in [14] introduce the Local Feature Extractor (LFE) module, which is composed of a series of dilated convolutions of decreasing rates, after aggressively increasing the rates of dilated convolutions used in the front-end module to attain a high receptive field throughout the feature extraction process. They show that LFE module helps with tiny objects by recovering the spatial inconsistency and extracting local structure at higher layers.

3. Problem Description

3.1. Challenges

Designing a model that is capable of segmenting buildings footprint from aerial images at different spatial resolutions faces the following challenges:

(i) **Features Resolution**: Most Feature extractors [6, 16, 35, 36, 40] are a series of five down-sampling stages, where each stage outputs denser and more meaningful representations than the previous one. However, features spatial resolution is reduced to half at each stage using a 2x2 pooling operation or a convolution with a stride $= 2$. This reduction leads to a loss in spatial information the deeper we go in the network, as the features extracted by the last stage have a resolution $32\times$ smaller than the input size (output stride $= 32$).

(ii) **Field of View**: In feature extraction networks, convolutions are applied with a 3×3 receptive field (kernel size). Although this works very well in segmenting small to medium-sized objects, it often fails when dealing with larger objects, (*i.e.*, building footprints at very high resolutions such as 2cm/pixel). In the latter case, predicted segments often suffer from fragmentation, under-segmentation, and noise because the field of view is too small for the network to decide if the pixel belongs to a larger object or the background. On the contrary, increasing receptive field by applying convolutions with an increasing kernel size leads to losing local spatial information and exponential growth in both time and computational complexity.

3.2. Motivation

Motivated by the observations above, we propose the following solution:

(i) To avoid losing spatial features details and keep a reasonable computational complexity, we adopt the skip connections from the encoder to the decoder.

(ii) To increase the field of view or the receptive field: *(a)* we use an encoder/decoder framework where the feature map spatial size decreases/increases with depth for the encoder/decoder, *(a)* we propose to include dilated convolutions, at the bottleneck of the encoder.

To this end, we propose Sci-Net; a UNet like architecture augmented with DenseASPP modules. We provide more details in the following sections.

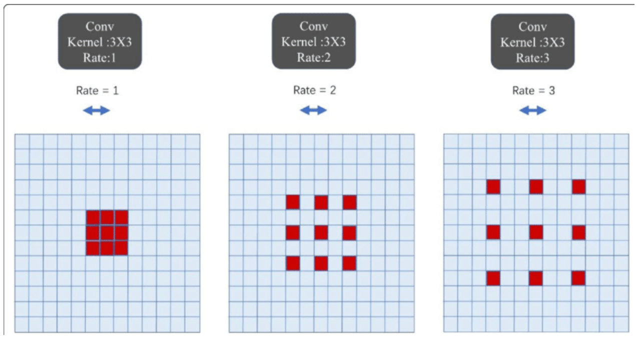


Figure 1. Dilated 3×3 convolutions with rates = 1, 2 and 3 acquiring 3×3, 5×5 and 7×7 receptive fields, respectively.

4. Sci-Net

4.1. Dilated Convolutions

Dilated convolutions work just like regular convolutions; however, they manage to increase the size of the receptive field by the insertion of holes between the weights of the kernel, according to a selected rate denoted by r. For a 2-dimensional input feature map x, the output feature map y obtained as a result of a dilated convolution at every spatial location i with a rate r is defined according to the following function:

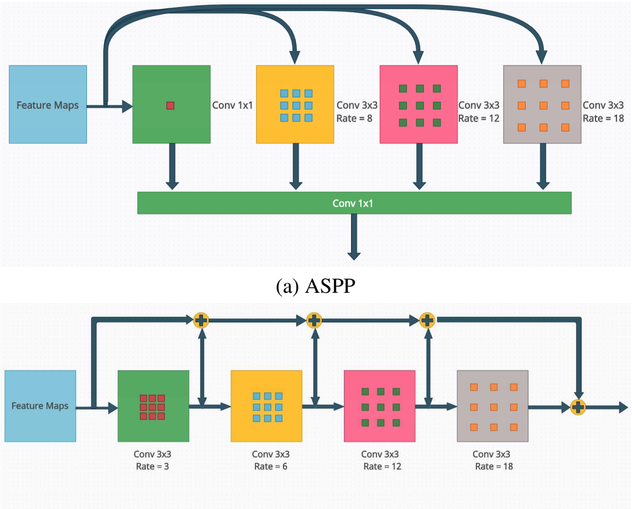
$$y[i] = \sum (x[i + r * k] * w[k]). \tag{1}$$

The rate r corresponds to the distance between the kernel weights. In this manner, a 3×3 convolution with rates $= 1, 2 and 3$ acquires the same receptive field size as 3×3, 5×5, and 7×7 regular convolutions respectively with the same number of parameters as a 3×3 regular convolution as shown in Figure 1. Dilated convolution increases the receptive field size without incurring additional time or computational complexity.

4.2. Atrous Spatial Pyramid Pooling (ASPP)

ASPP is a module that applies multiple dilated convolutions with different rates on the feature maps to capture multi-scale representations. The concept has been first introduced in [2] and then further developed in [3, 5].

Typically one 1×1 convolution, three 3×3 dilated convolutions of atrous rates equal to (8, 12 and 18), and a global average pooling layer are applied in parallel. The resulting representations are then concatenated together and pooled with a 1×1 convolution as shown in Figure 2 *(a)*. Applying three different and separate dilated convolutions allows the model to extract spatial information at three different scales, with a maximum receptive field size equal to 37 pixels.



(a) ASPP

(b) Dense ASPP

Figure 2. ASPP vs. Dense ASPP architectures.

4.3. Dense ASPP

Dense ASPP [42] applies dilated convolutions with increasing atrous rates in a cascade manner. The input to each dilated convolution block is the initially extracted feature maps concatenated with all the representation from previous dilated convolutions of lower rates as shown in Figure 2 *(b)*. Typically, four atrous convolutions are applied with rates equal to (3, 6, 12 and 18). When two convolutions of different receptive fields are stacked together, the resulting receptive field increases in a linear manner, as follows:

$$R_{new} = R_1 + R_2 - 1. \tag{2}$$

where R_1 and R_2 denote receptive fields size in pixels of the 1^{st} and 2^{nd} convolutional layers, and R_{new} is the size of the new receptive field after stacking the two convolutional layers together.

Usage of Dense ASPP would lead to 16 receptive field scales and a maximum value equal to 79 pixels, which means that more pixels are involved in the convolution resulting in a denser feature pyramid than ASPP. Thus, based on the above analysis, Dense ASPP is integrated into the proposed Sci-Net model.

4.4. Sci-Net Architecture

In this subsection, we provide a detailed description of the proposed Sci-Net model architecture and illustrate the

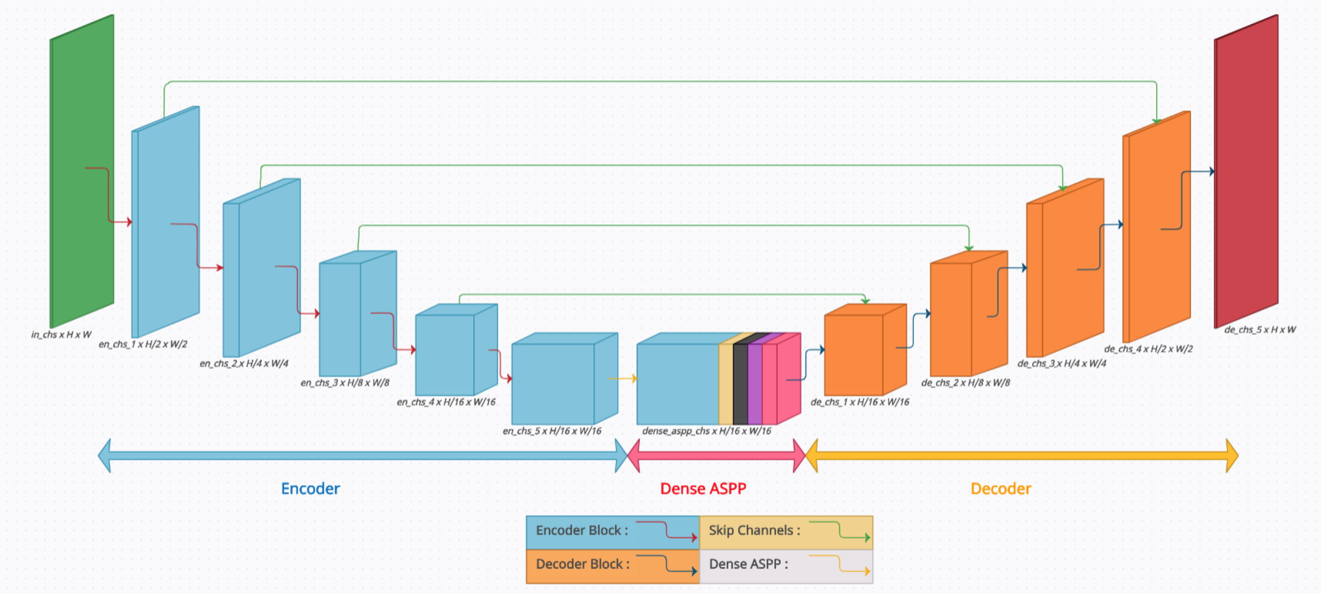


Figure 3. Proposed Sci-Net Architecture.

role of the modifications that we apply to deliver a better performance.

The proposed Sci-Net model shown in Figure 3 adapts conventional UNet encoder-decoder architecture [33] with the following modifications:

(a) Replacing the encoder with a more powerful yet light-weight feature extractor from the RegNet Family ($RegNetY - 1.6GF$). RegNets have similar performance to their Efficient-Net counterparts while being 3x to 5x times faster.

(b) Integration of a Dense ASPP block to extract multi-scale representation from the features of the last encoder stage. The output features of Dense ASPP are the input to the first decoder block. We used the following rates $(3, 6, 12, 18)$ for the dilated blocks, and we set the output channels to 256 for each block. When concatenated, the multi-scale representations alone are a total of $256 \times 4 = 1024$ channels, and the initial feature maps contain 888 channels.

(c) Substitution of the last 3×3 convolution with a kernel stride $= 2$ in the 5-th encoder stage, with a dilated convolution of a low atrous rate $= 2$ and kernel stride $= 1$ to prevent downsampling of features produced by stage 5 and thus the output stride remains equal to 16 instead of 32. This modification preserves a sufficiently good spatial resolution at the Dense ASPP input.

(d) No upsampling is applied at the first decoder block as both stages 4 and 5 feature maps have the exact spatial resolution.

Each decoder block comprises two 3x3 convolutions with a stride equal to 1 followed by a 2x upsampling bi-linear interpolation.

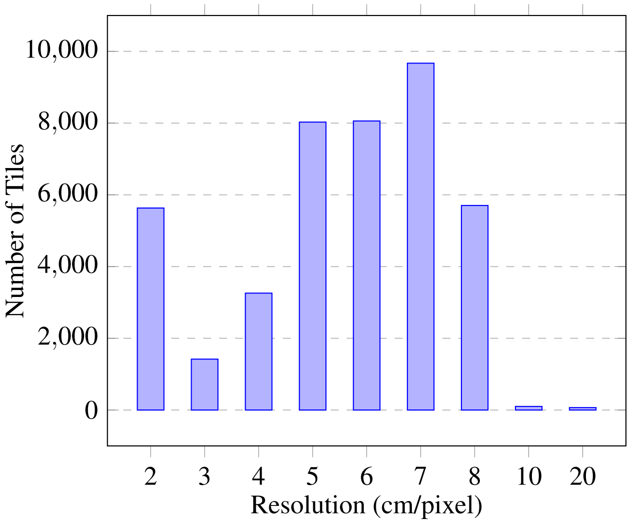


Figure 4. Tiles distribution per resolution in cm/pixel in the Open Cities AI dataset.

5. Experiments

In this section, we present the dataset, the training and implementation details and finally, the experimental results.

5.1. Dataset

As far as we know, Open Cities AI Challenge dataset (OCAIC) is the only aerial images dataset suitable for the underlying task:

OCAIC [22]: To assess the performance evaluation of the proposed Sci-Net model over different images' spatial resolution, we used the Open Cities AI Challenge dataset (also known as Segmenting buildings for disaster resilience). The majority of the data are collected across different African cities, where the images and labels quality varies from one region to another. Open Cities AI dataset is split into two tiers.For the scope of this work, tier 1 images are used, where imagery and labels are distributed under the **CC-BY-4** and **ODbL-1.0** licenses, respectively. Tier 1 data is made of 31 GeoTiff images of different spatial resolution and size. Resolution varies from very high (2cm/pixel) up to medium (20cm/pixel) resolutions. Moreover, Figure 4 shows images distribution across different resolutions.

The resulting dataset contains 40,000 tiles with their corresponding buildings' masks. The dataset is split as 90 % for training and 10 % for testing.

5.2. Implementation details

Here we explain the implementation details for other approaches (For Sci-Net please refer to section 4 for more details).

For fair comparison with Sci-Net, we ensured, as far as possible, that we have the same parameter choices for all compared methods. Specifically, the encoder is replaced by $RegNetY-1.6GF$ for all methods, except for HRNet which uses HRNetV2-W32 as backbone. The decoder channels are fixed to (256, 128, 64, 32, 16) except for Deeplabv3+ where the number of channels is fixed to 256, and PSPNet where the number of output channels is 512 (number of filters in Spatial Pyramid). The number of channels for the PAB module in MANet is 64.

All encoders are initialized with ImageNet [9] weights.

5.3. Training details

In all our experiments, the models are trained until convergence (50 epochs) using Adam optimizer [21] and polynomial learning rate policy [44] where the learning rate is decayed from the initial one of 0.0001 till zero at the last epoch as follows:

$$lr_{t+1} = lr_t * (1 - \frac{epoch_{t+1}}{epoch_{max}})^{0.9} \tag{3}$$

A weighted combination of Dice loss and Binary Cross-Entropy (BCE) loss is used as defined in the following:

$$\text{Loss} = \Gamma_1 \cdot BCE + \Gamma_2 \cdot Dice \tag{4}$$

where $\Gamma_1 = \Gamma_2 = 0.5$ is considered for simplicity.

During training, we use a batch-size of 12, random 512 × 512 chips of the original 1024 × 1024 tiles and apply only positional augmentations like horizontal-flipping, vertical-flipping, and 180° rotation with an 80% probability. These augmentations help to introduce some randomness at every training iteration and prevent the model from over-fitting.

Moreover, the training is performed over a Single Titan-XP GPU card with 12 GB of VRAM. In addition to the proposed Sci-Net model, we trained some well-known SoA models for comparison. Training these models took between 20 and 120 hours, depending on the model complexity.

Training framework is done in PyTorch using mixed precision functionalities [**?**] and $pytorch-segmentation-models$ implementation [41].

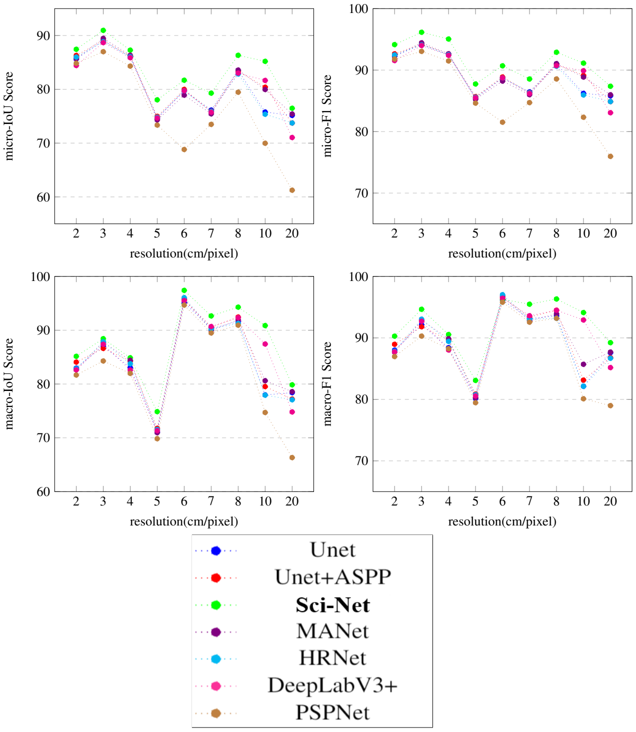


Figure 5. Performance comparison of Sci-Net (in green) versus SoA models in terms of micro- and macro- IoU and F1 scores.

5.4. Metrics

Performance evaluation of the proposed Sci-Net and SoA models is measured using macro, and micro Intersection over Union (IoU) and F1-score defined below:

$$IoU = \frac{TP + \varepsilon}{TP + FP + FN + \varepsilon} \tag{5}$$

Model	micro-IoU	micro-F1	macro-IoU	macro-F1	Params. (M)	GFLOPs
PSPNet [44]	78.47	87.93	85.04	89.55	12.0	8.7
DeepLabV3+ [5]	79.83	88.78	86.28	90.50	12.0	13.5
MANet [10]	80.08	88.93	86.20	90.38	35.5	25.0
HRNet [38]	80.30	89.07	82.85	90.46	29.5	45.0
Sci-Net	**82.25**	**91.04**	**88.42**	**92.62**	24.2	33.1

Table 1. Performance metrics of Sci-Net architecture compared to existing SoA models on the testset.

$$F1-score = \frac{(1+\beta^2)\cdot TP+\varepsilon}{(1+\beta^2)\cdot TP+\beta^2\cdot FN+FP+\varepsilon} \tag{6}$$

where $\beta = 1$ and $\varepsilon = 0.0001$. TP, FP, and FN are True Positive, False Positive, and False Negative, respectively.

The employed metrics are used according to the following two types of averaging across the whole test dataset:

1. **Macro** scores are calculated per prediction according to TP, FP, and FN for each prediction mask and then averaged afterward.

2. **Micro** scores are calculated using the total number of TP, FP, and FN across all prediction masks, and then the final score is computed accordingly.

Macro scoring helps in assessing the average performance per image, while micro scoring assesses the overall performance. Correctly classified images with blank ground truth masks (True Negative) provide a boost in macro scoring. However, such images do not affect micro scoring, as it treats the whole dataset as having one large ground truth mask.

It is worth noting that simulation results are validated every epoch, and the model's weights that maximize micro-IoU score over the validation set are saved.

5.5. Comparison to other methods

Experimental results reveal the performance superiority of the proposed Sci-Net over several SoA models in terms of IoU and F1-score across varying resolutions. At inference time, the images are fed to the neural network at full scale (1024 × 1024). Table 1 shows that the proposed Sci-Net model provides a significant improvement of at least 2% score over benchmarked SoA models. Sci-Net attained a micro-IoU score of 82.25%. Table 1 also shows that Sci-Net scored the highest macro-IoU value of 88.42%, which indicates that it leverages the best per-image performance. Furthermore, the 2% minimum score improvement margin is observed in micro and macro F1-scores (91.04% and 92.62% respectively). Thus, the proposed model attains better precision and recall against competitor models. Models like PSPNet and DeepLabV3+ failed to provide near good results at some resolutions leading to a degradation in their scores. Complexity analysis...

Furthermore, we plot micro and macro IoU and F1-scores for each resolution (cm/pixel) for every model as shown in Figure 5. The green curve corresponding to the Sci-Net model always performs better than all other models across different metrics and resolutions, which shows that it can effectively extract better multi-scale representations than the existing models. Sci-Net curve is consistently above all other curves for the four presented score graphs, which indicates that it is less prone to performance degradation when the scale changes. It is unclear which model holds the ”runner up” spot, as these models alternate places on varying resolutions. For instance, UNet+ASPP (in red) achieves competitive scores for resolutions in range (2cm/pixel up-to 8cm/pixel), however its performance deteriorates for larger resolutions. PSPNet (in brown) benchmarks the worst performance for all resolutions.

To better visualize our results, a sample of predicted masks by each model are shown in Figure 6. For instance, problems like fragmentation and under-fitting are solved using Sci-Net by acquiring sufficiently large receptive fields capable of relating far pixels that belong to large building instances at high resolutions (Masks in rows 1, 2, 3, and 4). In the first four rows, it is clear that Sci-Net succeeds in segmenting large building instances. For example, models like DeepLabV3+, UNet, and HRNet showed a critical level of mask fragmentation for that large building instance in the first row. Also, at lower resolutions (rows 6 and 7), Sci-Net can avoid over-segmentation, unlike other architectures such as HRNet that misclassified a significant amount of background pixels, as shown in row 6.

While models that use pyramid pooling like PSPNet performed well in segmenting large building instances (row 1 and 2), they often fails in capturing small to medium-sized buildings at lower resolutions. Finally, Unet+ASPP could not capture enough multi-scale information to segment large structures properly (rows 1, 2, and 3). Sci-Net proved to be the most efficient at all the presented resolutions, as shown in Figure 6.

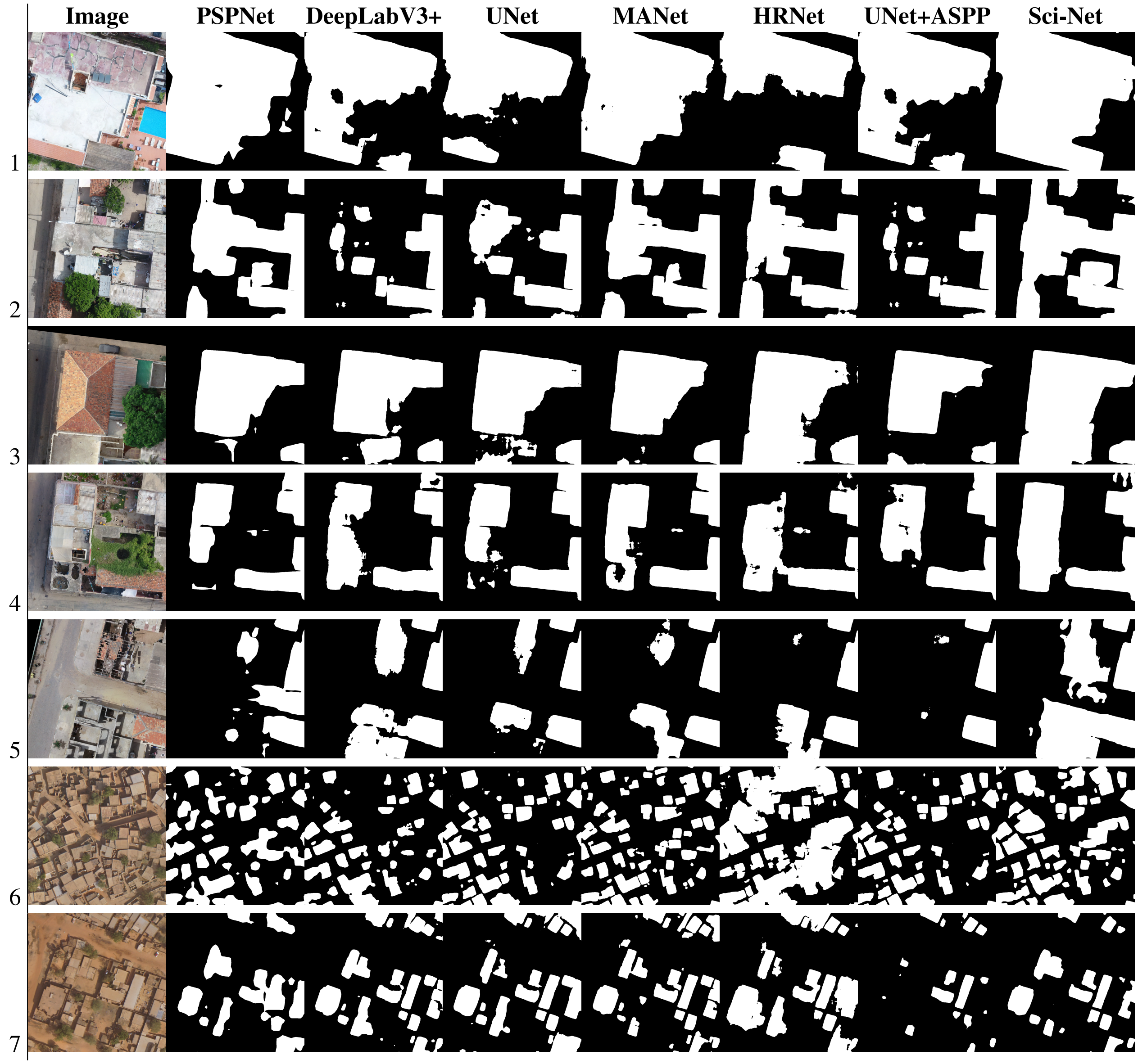


Figure 6. Predicted masks for Sci-Net and benchmarked SoA models for various images at different resolutions revealing fragmentation, under- and over-segmentation issues.

Model	micro-IoU	micro-F1	macro-IoU	macro-F1	Params. (M)	GFLOPs
UNet [33]	80.18	89.00	86.01	90.16	14.5	23.2
UNet+ASPP	80.47	89.18	86.65	90.79	21.0	28.5
UNet+DenseASPP (Sci-Net)	**82.25**	**91.04**	**88.42**	**92.62**	24.1	33.1

Table 2. Ablation Study; Dens ASPP brings a significant improvement to UNet on the test set, compared to ASPP.

In terms of model complexity, Table 1 shows that our model is smaller than MANet and HRNet, which indicates that the improvements is not due only to adding more modules/parameters.

5.6. Ablation Study

In this subsection, we investigate the importance of our design choices. From Table 2, we can notice that the dilated convolutions at the bottleneck, brings slight improvements to the UNet model. On the other hand, the Dense ASPP module leads to significant improvements compared to ASPP for the UNet model. This observation also holds for the evaluation on several resolutions (Figure 5) and in the qualitative comparison in Figure 6.

6. Conclusion

This paper proposes Sci-Net, a new model capable of accurately segmenting buildings' footprint at multi-scale spatial resolutions. We compare the performance of Sci-Net with other well-known SoA models. We show that the proposed Sci-Net architecture does not suffer from fragmentation, over-segmentation, and under-segmentation problems.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 2

[2] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 40(04):834–848, apr 2018. 3

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1, 2, 3

[4] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016. 2

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 3, 6

[6] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng. Dual path networks. *Advances in neural information processing systems*, 30, 2017. 2

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2

[8] Remi Delassus and Romain Giot. Cnns fusion for building detection in aerial images for the building detection challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 242–246, 2018. 1

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[10] Tongle Fan, Guanglei Wang, Yan Li, and Hongrui Wang. Ma-net: A multi-scale attention network for liver and tumor segmentation. *IEEE Access*, 8:179656–179665, 2020. 6

[11] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013. 2

[12] Golnaz Ghiasi and Charless C Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European conference on computer vision*, pages 519–534. Springer, 2016. 2

[13] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 10–17, 2019. 1

[14] Ryuhei Hamaguchi, Aito Fujita, Keisuke Nemoto, Tomoyuki Imaizumi, and Shuhei Hikosaka. Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1442–1450. IEEE, 2018. 1, 2

[15] Ryuhei Hamaguchi and Shuhei Hikosaka. Building detection from satellite imagery using ensemble of size-specific detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[17] Vladimir Iglovikov, Selim Seferbekov, Alexander Buslaev, and Alexey Shvets. Ternausnetv2: Fully convolutional network for instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 1

[18] Md Amirul Islam, Mrigank Rochan, Neil D. B. Bruce, and Yang Wang. Gated feedback refinement network for dense image labeling. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4877–4885, 2017. 2

[19] Aatif Jiwani, Shubhrakanti Ganguly, Chao Ding, Nan Zhou, and David M Chan. A semantic segmentation network for urban-scale building footprint extraction using rgb satellite imagery. *arXiv preprint arXiv:2104.01263*, 2021. 1

[20] Çağrı Kaymak and Ayşegül Uçar. A brief survey and an application of semantic image segmentation for autonomous driving. In *Handbook of Deep Learning Applications*, pages 161–200. Springer, 2019. 1

[21] Diederik P Kingma and J Adam Ba. A method for stochastic optimization. arxiv 2014. *arXiv preprint arXiv:1412.6980*, 434, 2019. 5

[22] GFDRR Labs. Open cities ai challenge dataset. *Version 1.0, Radiant MLHub. [Date Accessed] https://doi.org/10.34911/rdnt.f94cxb*, 2020. 1, 5

[23] Tao Lei, Risheng Wang, Yong Wan, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. Medical image segmentation using deep learning: a survey. *arXiv preprint arXiv:2009.13120*, 2020. 1

[24] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018. 1

[25] Lin Li, Jian Liang, Mingrui Weng, and Haihong Zhu. A multiple-feature reuse network to extract buildings from remote sensing imagery. *Remote. Sens.*, 10:1350, 2018. 2

[26] Qingyu Li, Yilei Shi, Stefan Auer, Robert Roschlaub, Karin Möst, Michael Schmitt, Clemens Glock, and Xiaoxiang Zhu. Detection of undocumented building constructions from official geodata using a convolutional neural network. *Remote Sensing*, 12(21), 2020. 1

[27] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 2

[28] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3194–3203, 2016. 2

[29] Jingbo Lin, Weipeng Jing, Houbing Song, and Guangsheng Chen. Esfnet: Efficient network for building extraction from high-resolution aerial images. *IEEE Access*, 7:54285–54294, 2019. 2

[30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[31] Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12416–12425, 2019. 1, 2

[32] Pedro Pinheiro and Ronan Collobert. Recurrent convolutional neural networks for scene labeling. In *International conference on machine learning*, pages 82–90. PMLR, 2014. 2

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 2, 4, 7

[34] Yilei Shi, Qingyu Li, and Xiao Xiang Zhu. Building segmentation through a gated graph convolutional neural network with deep structured feature embedding. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:184–197, 2020. 1

[35] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 2

[36] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 2

[37] Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The multi-temporal urban development spacenet dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2021. 1

[38] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 1, 6

[39] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1451–1460. Ieee, 2018. 2

[40] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 2

[41] Pavel Yakubovskiy. Segmentation models pytorch, 2020. 5

[42] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[43] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 2

[44] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1, 2, 5, 6

[45] Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei-Shi Zheng, Jonathan Li, and Alexander Wong. Squeeze-and-attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13065–13074, 2020. 2