DeepMAO: Deep Multi-scale Aware Overcomplete Network for Building Segmentation in Satellite Imagery

Aniruddh Sikdar[†1], Sumanth Udupa[†2], Prajwal Gurunath[†2], Suresh Sundaram[2]
[1]Robert Bosch Centre for Cyber Physical Systems, Indian Institute of Science, Bengaluru, India
[2]Department of Aerospace Engineering, Indian Institute of Science, Bengaluru, India
{aniruddhss, sumanthudupa, prajwalg, vssuresh} @iisc.ac.in

Abstract

Building segmentation in large-scale aerial images is challenging, especially for small buildings in dense and cluttered urban environments. Complex building structures with highly varied geometric footprints pose an additional challenge for the building segmentation task in satellite imagery. In this work, we propose to tackle the issue of detecting and segmenting small and complex-shaped buildings in Electro-Optical (EO) and SAR satellite imagery. A novel architecture Deep Multi-scale Aware Overcomplete Network (DeepMAO), is proposed that comprises an overcomplete branch that focuses on fine structural features and an undercomplete (U-Net) branch tasked to focus on coarse, semantic-rich features. Additionally, a novel self-regulating augmentation strategy, "Loss-Mix," is proposed to increase pixel representation of misclassified pixels. DeepMAO is simple and efficient in accurately identifying small and geometrically complex buildings. Experimental results on SpaceNet 6 dataset, on both EO and SAR modalities, and the INRIA dataset show that DeepMAO achieves state-of-the-art building segmentation performance, including small and complex-shaped buildings with a negligible increase in the parameter count. In addition, the presence of the overcomplete branch in DeepMAO helps in handling the speckle noise present in the SAR image modality.

1. Introduction

The availability of high-resolution satellite images from different modalities, such as Electro-Optical (EO) and Synthetic Aperture Radar (SAR), has made it easy to monitor urban environments on a large scale. Building segmentation is important in monitoring changes in the urban landscape, as buildings are critical components in these regions. It aims to classify the area occupied by the building in the image by pixel-level classification. Building information is used in many applications like change monitoring, map updating, disaster response [1], population density estimation [23], humanitarian aid, and 3-D modeling [4]. In these applications, high-resolution SAR imaging is highly beneficial, as it provides consistent information over EO imaging due to all-weather operational capabilities. However, SAR sensors have certain drawbacks such as speckle noise and less semantic information, which makes interpretation challenging for computer vision systems as well as human interpreters [20]. Automated building detection still faces significant challenges due to the diversity of buildings in terms of shapes and sizes, the complex background environment, and the complexities introduced due to SAR sensors.

Advances in deep learning-based segmentation models have largely improved building segmentation performance in high-resolution remote-sensing images. Semantic segmentation architectures such as U-Net [19], DeepLabv3+ [5], FPN [12], PSPNet [31] have achieved competitive results on datasets like MS-COCO [13], Cityscapes [6]. However, the small objects defined in these benchmarking datasets differ from those defined for remote sensing datasets. The small buildings defined in SpaceNet 6 dataset [21] are smaller than pixel area of 225 in the full image size of 900 x 900 pixels. Fig. 1 shows the predictions of buildings of various sizes made by U-Net and DeepMAO. U-Net falls short in recognizing buildings of different shapes and sizes, especially in dense and cluttered environments. The aforementioned standard segmentation models exhibit poor building segmentation results for buildings with small-area footprints due to low pixel representation, even in a supervised setting. Small buildings in close proximity to larger buildings are not detected accurately, leading to a huge variation in performance between buildings with smaller and larger footprints.

Convolutional encoder-decoder architectures [3], [5], [19] are widely used to extract semantic information from remote-sensing images [28]. The input images undergo compression in the encoder and are subsequently decompressed in the decoder. The focus of these networks is lim-

[†]Equal contribution of authors.

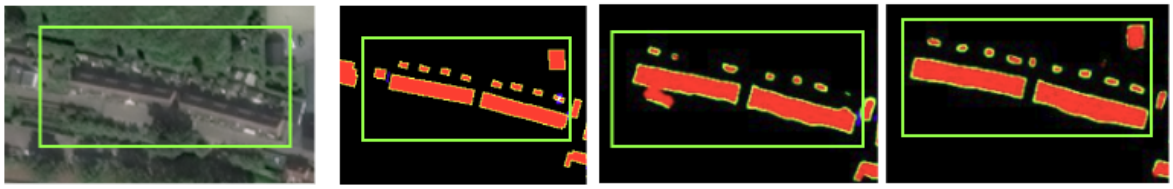

Figure 1. Building predictions by U-Net and the proposed model DeepMAO. The predictions made by U-Net have coarse boundaries, and inaccurate small buildings. Proposed model identifies small buildings and poses stricter boundaries. In (b), red denotes building footprints, yellow denotes the boundary pixels.

ited to relatively large objects, as only the shallow layers of the encoders are responsible for the low-level features. As the receptive field increases in size over the depth of the encoder blocks, the network focuses more on high-level features. However, smaller receptive fields are required for making fine-grained predictions in a dense and cluttered environment and identifying small buildings accurately.

To address the structural limitation of under-complete networks, over-complete architectures were proposed to detect small anatomical structures in the medical imaging community [26]. In these networks, the input is projected to higher dimensions (in a spatial sense) in the intermediate layers to restrict the size of the receptive field deeper into the network. This constriction forces the network to focus on smaller, intricate features of the input images. The use of these architectures in remote sensing applications has been limited due to their inherent drawbacks, such as the need for greater computing power, GPU memory, and longer training times, which pose significant challenges. As a result, they have been under-explored in this domain.

In this paper, we propose Deep Multi-scale Aware Overcomplete network (DeepMAO) for accurately detecting small and complex shaped buildings in dense and cluttered urban environments in the EO and SAR image modality. It contains two branches, where one is an overcomplete branch to extract finer details, and the other is an undercomplete branch, i.e., U-Net. The overall architecture is made to be very simple to adjust to the computational overload of the upsampling layers. The effectiveness of the proposed method is shown in Fig. 1, where the building predictions are superior. Additionally, to increase the representation of small and complex geometric shaped buildings, a new self-regulative training strategy *Loss-Mix* is adopted to further aid DeepMAO's low-level feature extraction capabilities. The main contributions of this paper are as follows:

1. A Deep Multi-scale Aware Overcomplete Network (DeepMAO) is proposed, which comprises of an overcomplete branch and an undercomplete branch (U-Net). Its simple structure focuses on efficiently learning fine-grained information along with coarse information for building segmentation.

2. A new training strategy called Loss-Mix is proposed that uses self-regulative cut-mix augmentation to increase the pixel representation of harder patches. This strategy aids the training process of standard building segmentation models and DeepMAO, for optical images.

3. DeepMAO outperforms standard building segmentation models across two public building detection datasets - SpaceNet 6 [21] and INRIA [17]. It achieves state-of-the-art results for building detection including small buildings on EO and SAR modalities on SpaceNet 6 dataset.

2. Related work

Building segmentation networks: FCN [15], U-Net [19] and their variants have been used for building extraction for both EO and SAR sensors [29], [33]. A fully convolutional network (FCN) extracts building footprints based on a skip connection-based architecture to fuse low-level and high-level semantic information [2]. U-Net [19] has feature fusion on an equal level of encoder and decoder to learn the high-level features of the buildings. DeepLabv3+ [5] contains an atrous spatial pyramid pooling block to capture the contextual information and is used to extract buildings of different scales. In [8], a modified auto-encoder structure using a selective spatial pyramid network (SSPD) was proposed for multiscale context fusion to extract buildings in SAR images. CVCMFF-Net [4] was proposed for building segmentation of high-resolution complex-valued SAR images. Furthermore, a fully complex-valued segmentation model FC^{2}MFN [22] was proposed, which showed empirically that complex-valued models perform better than their real-valued counterparts for complex-valued SAR images. In [32], the authors first showed the performance gaps of deep learning models between EO and SAR modalities, and a drop in performance for SAR images was reported. A dynamic network framework was proposed to learn the

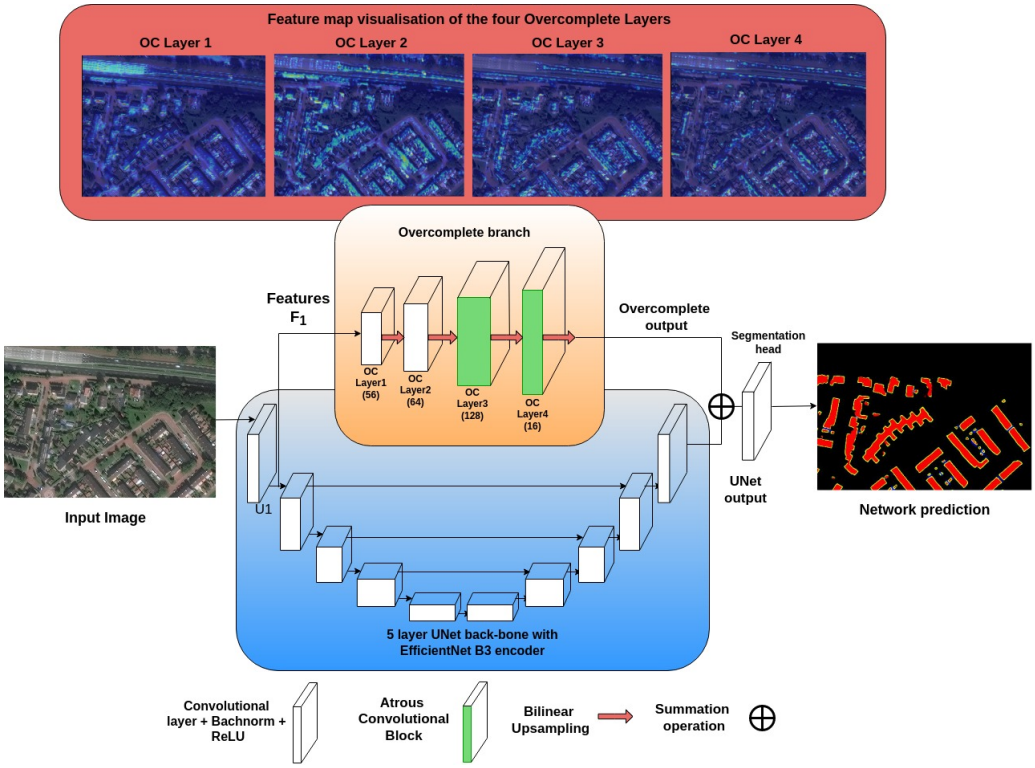

Figure 2. The proposed architecture for Deep Multi-scale Aware Overcomplete Network (DeepMAO). The input image is fed to the U-Net model with EfficientNet B3 as the encoder. The features of the first U-Net encoder block is passed to the overcomplete branch. The final segmentation mask is attained by adding the outputs of overcomplete and under-complete branches and passed through a segmentation head. Feature map visualisations of the four overcomplete layers reveal that the each successive layer focuses on finer features.

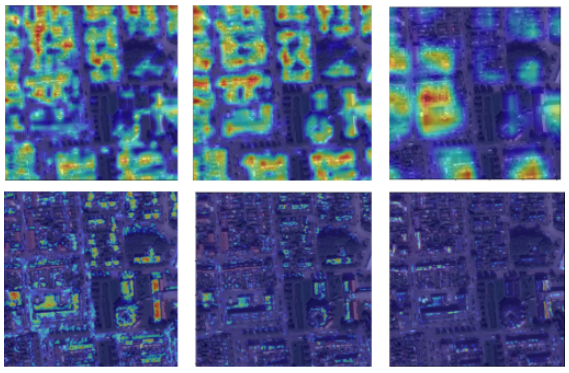

Figure 3. Feature maps collected from DeepMAO trained on Spacenet6 dataset. The top row corresponds to feature maps collected from U-Net, while the bottom row corresponds to the feature maps collected from the layers of the overcomplete branch.

metasensory representation from both the sensors and test on SAR images. In a two-step training strategy, a knowledge distillation framework known as DisOptNet [9] was proposed to distill the high-level semantic knowledge from optical images to SAR images. The trained model was tested only on SAR images for building detection.

Overcomplete representations: These representations were explored in signal processing to represent the input signal samples with more number basis functions. It was shown that not only are the overcomplete bases better approximators of the underlying distribution of data but are also more robust to noise [26]. Models like denoising autoencoder with overcomplete layers performed better as feature detectors [27]. Overcomplete convolutional networks have been used for SAR despeckling [18]. The finer details captured by the overcomplete branch are useful for removing the fine speckles and a Multi-Scale Feature Fusion block is proposed to transfer the low-level features of the overcomplete branch to the under-complete branch. A Fine Context-aware Shadow Detection Network (FCSD-Net) [25] was proposed to detect unclear and blurry shadow regions by using upsampling layers to reduce the receptive field as the network propagates deeper.

3. Methodology

In this section, the proposed DeepMAO is explained in detail along with the self-regulative learning scheme, Loss-Mix.

3.1. Deep Multi-scale Aware Overcomplete network

To accurately identify buildings, especially small and complex shaped buildings, a Multi-scale Aware Overcomplete network (DeepMAO) is proposed, which is an overcomplete branch augmented to U-Net. DeepMAO architecture is illustrated in Fig. 2. In DeepMAO, a parallel overcomplete branch is added to the under-complete branch, i.e., U-Net, to exploit the low-level features while the U-Net focuses on the higher-level features, as shown in the feature maps in Fig. 3. The input image I is fed to the U-Net, with EfficientNet B3 [24] its encoder. For the task of semantic segmentation, in an encoder-decoder structure, the receptive field enlarges in an under-complete network as the network propagates deeper [26]. This is mainly due to the pooling layers. This increase in the receptive field causes the network to focus more on the global features of the input image, as shown in the feature visualisations in Fig. 3. The penultimate layers of the U-Net encoder focus on larger features to gain high-level semantic context.

From the output of the first encoder block of U-Net, features F_1 are passed to the overcomplete branch, as shown in Fig. 2. The proposed overcomplete branch has four bilinear interpolate upsampling layers. The scaling factor of the first three layers is 1.2, and the last upsampling layer is set to 1.15 to constrain the receptive field after each layer, allowing the overcomplete branch to not shift focus toward semantic information, but instead focus on low-level features. The ability of DeepMAO to focus on large-sized buildings and relatively smaller buildings simultaneously without a drop in performance in either makes it multi-scale aware. The inputs to the last two upsampling layers are passed through an atrous convolution layer, with a dilation rate and padding of 2. Contrary to the pooling layers, the upsampling layer projects feature maps to higher-dimension space to constrict the receptive field and focus more on the low-level features and finer details. The kernel size and upsampling coefficient are selected considering the computational overload due to upsampling layers. The output feature maps of both branches are fused together by a summation block. These fused features are passed through a segmentation head of 1×1 convolution layers to generate the final predictions.

3.2. Self-Regulative Learning scheme: Loss-Mix

CutMix [30] augmentation has been widely used to generate new images by cutting out random image patches and pasting them onto one another, leading to an increase in performance for supervised segmentation methods. To increase the number of samples in training data corresponding to small and complicated building segments, we propose a self-regulative learning scheme [10] based on CutMix augmentation termed *Loss-Mix* on the labeled data D_L, which is inspired from SP-CutMix [11].

Samples x_i from D_L are first fed to the segmentation model f_θ to obtain the predictions, $\tilde{l_i} = f_\theta\,(x_i)$. These predictions $\tilde{l_i}$ and their corresponding labels l_i are divided into $H \times W$ patches, and patch-wise loss scores L_S are calculated. These scores are used for evaluating the easy and difficult patches for the model to predict. The loss score for the p^{th} patch of the i^{th} sample is defined as $L_S^p(l_i, \tilde{l_i})$. It is calculated by taking the L_1 loss between the model predictions and their corresponding ground-truth labels, given by the following equation,

$$L_S^p(l_i, \tilde{l_i}) = \sum |ReLU(\tilde{l_i}(p)) - l_i(p)|_1 \tag{1}$$

A binary mask $M \in \{0,1\}^{H \times W}$ is generated, keeping $M = 1$ based on the patch with the maximum loss $L_S^p(l_i, \tilde{l_i})$. The augmented training samples x'_k and its label l'_k are generated as follows:

$$x'_k = M \odot x_i + (1 - M) \odot x_j \tag{2}$$

$$l'_k = M \odot l_i + (1 - M) \odot l_j \tag{3}$$

where x_j, l_j are image and ground truth label patches corresponding to the maximum patch-wise loss L_S^p, and $\odot$ is element-wise multiplication operation. The sampling probability of misclassified pixels is increased due to the aforementioned augmentation which provides loss-based supervision for small and complex shaped buildings.

Training Strategy: The model is trained using L_{seg}, which is the summation of dice loss and focal loss and is defined as :

$$L_{seg}(y, p) = L_{Dice}(y, p) + L_{FocalLoss}(y, p), \tag{4}$$

$$L_{seg}(y, p) = 1 - \frac{2\sum_i p_i y_i}{\sum_i y_i + \sum_i p_i} - \sum_i \alpha_i (1 - p_i)^\gamma \log(p_i) \tag{5}$$

where y_i, p_i denote the pixel-wise ground truth and predicted probability values respectively, α_i and γ_i are hyper-parameters, where α_i is generally in the range of [0,1] and $\gamma_i > 1$. Unlike [11], the models in our training strategy are trained for sufficient epochs before the proposed augmentation is used. We call this strategy Loss-Mix as it is a sequential update in the augmentation by first training on simple augmentations like scaling and rotation, followed by the proposed augmentation. This is done in interest of training stability and so that once the model is sufficiently confident of its predictions of medium and big-sized buildings, the proposed augmentations focus only on the harder patches. Also, SP-CutMix [11] uses a threshold for the confidence score of the patches, while in Loss-Mix, the patch with the maximum loss is considered. This makes Loss-Mix self-regulating in nature. The supervision for harder patches is not set manually but governed by the loss function. The effectiveness of the training strategy is shown in the ablation studies for optical images.

4. Experimental results

In this section, the qualitative and quantitative results of DeepMAO are presented. Before proceeding with the results, the details about the SpaceNet 6 [21] and INRIA [17], [7] datasets are first described, followed by the evaluation metrics and the implementation details. The building segmentation performance of DeepMAO is evaluated with state-of-the-art models like U-Net [19], DeepLab v3+ [5], and FPN [12]. The backbones for all the networks are the same, i.e., EfficientNet B3 [24] pretrained on ImageNet.

4.1. Dataset

All the experiments are conducted on three building detection datasets, two from SpaceNet 6 and a third from the INRIA dataset. In SpaceNet 6 dataset, EO and SAR imaging modalities are used separately to evaluate the performance of various models.

SpaceNet 6: SpaceNet 6 [21] (Multi-Sensor All-Weather Mapping) is a large-scale dataset, covering 120 km^2 area of Rotterdam, The Netherlands, with over 48,000 buildings footprints. The dataset includes high-resolution optical and SAR images with a resolution of 0.5m with each tile of size 900 $\times$ 900 pixels. The dataset is split into training and testing sets based on the officially released repository in order to avoid data leakage*. There are about 2654 images in the training set, and about 747 images in the validation set, for both EO and SAR images. All four channels of SAR images are used.

INRIA: The INRIA Aerial Image Labeling Dataset [17], [7] covers various cities like Austin, Chicago, Kitsap, Vienna, and San Francisco. The spatial resolution of each image is 5000 $\times$ 5000 pixels with a surface coverage of 1500 $\times$ 1500 m^2. We split each image into 25 smaller images of 1000 $\times$ 1000 spatial resolution. Following previous investigations [14], the first five images of each city are selected for validation, and the rest are selected for training. There are 3875 training samples and 625 in the validation set. The supervision is provided by labels which are pixel-wise binary masks.

4.2. Evaluation metrics

Four evaluation metrics are chosen to evaluate the models, namely precision, recall, F1 score, and Intersection over Union (IoU), which are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$\textit{F1}\text{-score} = \frac{2TP}{2TP + FP + FN} \tag{8}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{9}$$

where *TP, TN, FP*, and *FN* are true positive, true negative, false positive, and false negative respectively. With pixel-wise classification, failing to detect smaller buildings does not necessarily cause a significant decrement in the overall *F1* score or the other metrics. Hence, a threshold is introduced for the evaluation metrics to identify the small buildings correctly. If the buildings are correctly identified with a threshold greater than 0.5, they are counted as true positives, else they are counted as false positives. The threshold is applied so that potential bias in the overall dice score resulting from the presence of larger buildings is not prevalent. The threshold is set for SpaceNet 6 dataset but not for the INRIA dataset due to the nature of the ground truth labels of the same.

4.3. Implementation details

We choose U-Net [19] with EfficientNet-B3 [24] backbone pre-trained on ImageNet as the under-complete model architecture. All the models are trained using AdamW optimizer [16] with weight decay of 10^{-2}. The initial learning rate is 2.0 x 10^{-4} with a stepwise decrement of 0.5 at epochs 80, 100, and 120. All models are trained for 150 epochs with a batch size of 8. In the training phase, the optical images are randomly cropped into the spatial size of 512 x 512 pixels, while random scaling and 10^o random rotation are used as the augmentations. The same process is followed for SAR images, with the only addition being random flips being added as an augmentation. Loss-Mix is added after half number of epochs. Loss-Mix is only applied to optical images and not SAR images. All the models are trained using L_{seg} loss, i.e., summation of dice and focal loss. All experiments are implemented using Pytorch deep learning framework on Nvidia RTX Quadro 5000 GPU with 16GB memory.

4.4. Results

Quantitative results: Table 1 shows the building segmentation performance of DeepMAO with other methods on the SpaceNet 6 optical images. Buildings with a pixel area range of 40 to 225 are classified as small buildings, as the performance of segmentation models for buildings above the pixel area of 225 is more confident, as seen from the *F1* scores in Table 1. As the resolution of images is 0.5

*SpaceNet 6 winner's code: https://github.com/SpaceNetChallenge/SpaceNet_SAR_Buildings_Solutions/tree/master/1-zbigniewwojna

Method	# Params	F1 Score			Precision	Recall
		F1 Score ($10\ m^2$ <Area <$56.25\ m^2$)	F1 Score ($56.25\ m^2$ <Area)	Overall F1 Score ($10\ m^2$ <Area)		
DeepLab (v3+)	11.6 M	15.78	62.17	43.80	52.41	37.67
FPN	12.47 M	30.45	74.93	56.35	64.61	49.96
U-Net	13.15 M	34.07	74.60	57.35	64.03	51.93
DeepMAO (+Loss-Mix)	13.30 M	**38.42**	**76.29**	**59.92**	**66.2**	**54.73**

Table 1. Evaluation metrics (%) of building segmentation networks and DeepMAO on SpaceNet 6 dataset EO Images.

Method	# Params	F1 Score			Precision	Recall
		F1 Score ($10\ m^2$ <Area <$56.25\ m^2$)	F1 Score ($56.25\ m^2$ <Area)	Overall F1 Score ($10\ m^2$ <Area)		
DeepLab (v3+)	11.6 M	10.33	41.5	29.12	45.28	21.46
FPN	12.47 M	**11.29**	42.05	29.75	45.18	22.18
U-Net	13.15 M	8.61	44.64	30.91	46.62	23.12
DeepMAO	13.30 M	9.99	**45.46**	**31.86**	**47.53**	**23.96**

Table 2. Evaluation metrics(%) of state of the art building segmentation models and DeepMAO on SpaceNet 6 SAR images.

Method	Intersection over Union
DeepLab (v3+)	77.42
FPN	78.09
U-Net	78.86
DeepMAO (+LossMix)	**80.01**

Table 3. Evaluation metrics (%) of building segmentation networks and DeepMAO on INRIA dataset (EO Images).

$m \times 0.5\ m$ per pixel in the dataset, a building pixel area range of 40 to 225 translates to an area range of 10 m^2 to 56.25 m^2. The F1 score is measured for buildings with three different area settings, (1) buildings of a small area, i.e., 10 m^2 to 56.25 m^2, (2) buildings with more than 56.25 m^2 area, and (3) all the buildings with an area greater than 10 m^2. As seen in the table, the *F1* score of DeepMAO has an increase of 4.35% compared to U-Net and 7.97% compared to FPN on small buildings.

This indicates the effectiveness of the overcomplete branch in preserving the semantic knowledge of the small buildings. The overall *F1* score of DeepMAO exceeds U-Net by 2.57% and FPN by 3.57%. This performance improvement comes from an increase of just 1.14% of the parameters of U-Net.

In Table 2, all the models have been evaluated on SAR images from SpaceNet 6 without using the Loss-Mix training strategy. Although FPN performs better than DeepMAO on small buildings, DeepMAO achieves the best overall *F1* score among all the compared methods proving that it makes better building predictions for buildings of all sizes. It achieves a better overall *F1* score compared to FPN and U-Net by 2.11% and 0.95%, respectively. This observation is in concordance with [26], [18], who observed that overcomplete models can capture finer details and are more robust to noise. Table 3 shows the INRIA dataset's evaluation of DeepMAO and other networks. DeepMAO trained with Loss-Mix achieves a better performance of 1.24 % compared to U-Net and 1.92% compared to FPN.

As seen in Fig. 5, DeepMAO performs consistently better than other models on SAR images, and converges early. Fig. 4 shows that with an increase in the building footprint size, the F1 score of all the models increases. Although DeepMAO performs well on buildings with extremely small area, its performance is greatly aided by the Loss-Mix training scheme. The regulative nature of identification of harder patches from Loss-Mix helps DeepMAO generalize better.

Qualitative Results: The predictions of DeepMAO and the best-performing models are visualized in Fig. 6 and Fig. 7. The predictions made by DeepMAO are closer to the ground truth. As seen in the figures, the building predictions by other models are inaccurate in dense and cluttered urban environments. The building boundaries are blurry, especially for smaller buildings. This is not the case with DeepMAO, as the predictions for small and large complex-shaped buildings are relatively more accurate and precise.

4.4.1 Ablation study

We perform several ablation studies to characterize the proposed DeepMAO architecture and Loss-Mix on both EO and SAR image modalities. The false positive rate when employing overcomplete layers was observed as Deep-

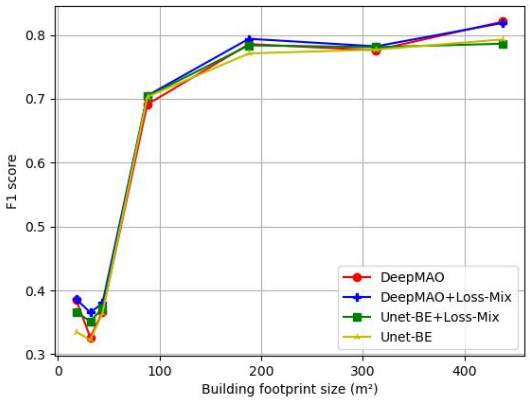

Figure 4. Performance comparison of overall F1 score vs building footprint size on SpaceNet 6 optical images.

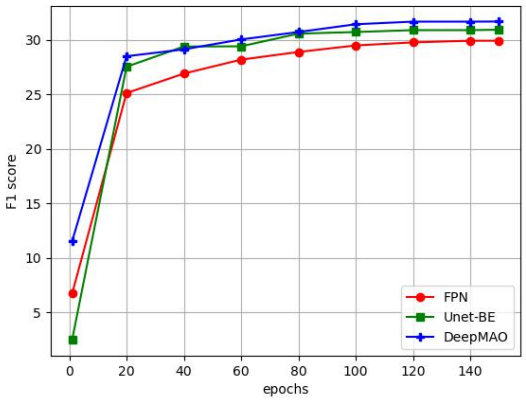

Figure 5. Performance curves of DeepMAO, FPN and U-Net models on SpaceNet 6 SAR images. The F1 score is plotted for all buildings with area more than 10 m^2.

Method	#Params	F1 Score			Precision	Recall
		F1 Score ($10\ m^2$ <Area <$56.25\ m^2$)	F1 score ($56.25\ m^2$<Area)	Overall F1 Score (Area >$10m^2$)		
DeepLab (v3+)	11.6M	15.78	62.17	43.80	52.41	37.67
FPN	12.47M	30.45	74.93	56.35	64.61	49.96
FPN + Loss-Mix	12.47M	32.34	74.66	56.76	63.25	51.48
U-Net	13.15M	34.07	74.60	57.35	64.03	51.93
U-Net+Loss-Mix	13.15M	36.68	75.08	58.43	63.61	54.03
DeepMAO	13.30M	37.66	75.33	58.82	63.59	54.72
DeepMAO +Loss-Mix	13.30M	**38.42**	**76.29**	**59.92**	**66.20**	**54.73**

Table 4. Ablation study of Loss-Mix strategy on DeepMAO and state-of-the-art models on SpaceNet 6 dataset - EO Images.

MAO's focus shifted towards very fine features. Thereby, rich semantic contextual information was lost due to the constriction in the receptive field. Hence, atrous convolution blocks were introduced in the penultimate layers to ease said constriction on the receptive field, enabling the network to regain some vital semantic context. Before passing the feature maps to the segmentation head, the outputs of the two branches are added. Better fusion strategies of the two branches may be employed to further enhance the results.

From Table 4, it can be seen that when the models are trained with Loss-Mix on SpaceNet 6 EO images, there is a clear increase in the segmentation performance across all the metrics, particularly for buildings of smaller sizes. The F1 score of DeepMAO trained with Loss-Mix is 4.35% better than the simple U-Net and 1.74% better compared to U-Net trained with Loss-Mix for small building segmentation. In the overall F1-score, it has an increase of 2.57% compared to the simple U-Net and 1.49% better compared to U-Net trained with Loss-Mix. The superiority of the proposed vanilla DeepMAO can be seen as it performs better than U-Net trained with Loss-Mix on small buildings by approximately 1% and 0.39% on the overall F1 score. Adding Loss-Mix to DeepMAO further improves its performance. A similar observation is also made on the INRIA dataset, as shown in Table 5, where all the models are trained using Loss-Mix. Loss-Mix is not employed for SAR images as a

Method	Intersection over Union
DeepLab (v3+)	77.42
FPN	78.09
FPN + Loss-Mix	78.97
U-Net - BE	78.86
U-Net - BE + Loss-Mix	79.60
DeepMAO	79.31
DeepMAO + Loss-Mix	**80.01**

Table 5. Ablation study of the effectiveness of Loss-Mix strategy on building segmentation networks on INRIA dataset (EO Images).

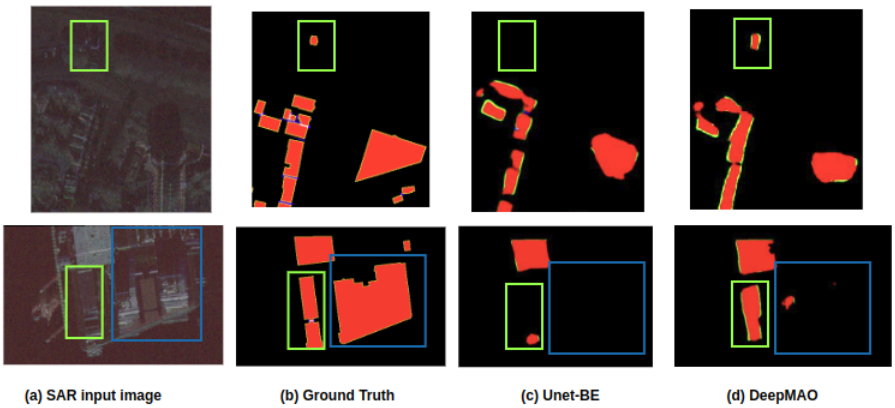

Figure 6. Selective crops of SAR input image has been used for visualization purposes. (a) SAR input image. (b) Respective ground truths. (c) U-Net-BE prediction. (d) DeepMAO predictions. Green box indicates a case where DeepMAO performance is better than Unet-BE. Blue box indicates a case wherein both models fail to detect a building.

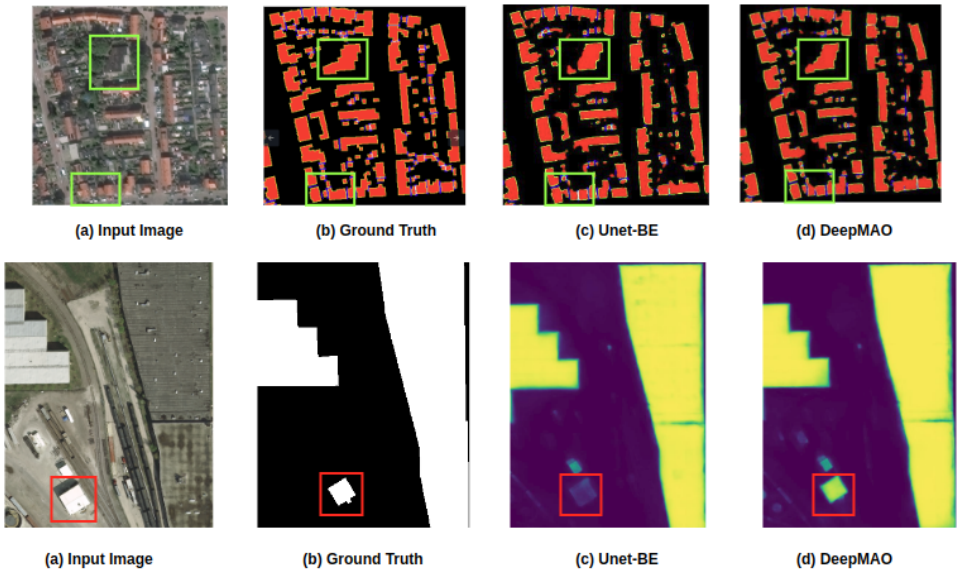

Figure 7. Selective crops of EO predictions. First row consists of Spacenet 6 Challenge dataset. Second row consists of INRIA dataset. (a) Input EO images. (b) Respective ground truth labels. (c) U-Net-BE predictions. (d) DeepMAO predictions.

minor drop in performance across all models was observed.

5. Conclusion

This work proposes a novel Deep Multi-scale Aware Overcomplete network (DeepMAO) for building segmentation tasks in satellite imagery for SAR and EO modalities. Current state-of-the-art models showcase good segmentation performance for either small or big buildings, whereas DeepMAO delivers a considerable bump across the board. The model contains two parallel branches, an overcomplete branch and the other under-complete branch. The proposed network focuses on high-level semantic information as well as the smaller and finer structural features by constricting the receptive field deeper into the network. A self-regulative learning scheme termed Loss-Mix is used, where harder patches are used to augment the image and further enhance the segmentation performance. Experimental results indicate that DeepMAO outperforms other state-of-the-art building segmentation models on EO and SAR images. It achieves a gain in overall F1 score of 1-2.5% on SpaceNet 6 and INRIA datasets, while more accurately detecting buildings of smaller area.

Acknowledgements: This work was financially supported by Centre for Airborne Systems.

References

[1] Bruno Adriano, Naoto Yokoya, Junshi Xia, Hiroyuki Miura, Wen Liu, Masashi Matsuoka, and Shunichi Koshimura. Learning from multimodal and multitemporal earth observation data for building damage mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:132–143, 2021. 1

[2] Rasha Alshehhi, Prashanth Reddy Marpu, Wei Lee Woon, and Mauro Dalla Mura. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130:139–149, 2017. 2

[3] 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. 1

[4] Jiankun Chen, Xiaolan Qiu, Chibiao Ding, and Yirong Wu. Cvcmff net: Complex-valued convolutional and multifeature fusion network for building semantic segmentation of insar images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021. 1, 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. 1, 2, 5

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[7] Bohao Huang, Kangkang Lu, Nicolas Audeberr, Andrew Khalel, Yuliya Tarabalka, Jordan Malof, Alexandre Boulch, Bertr Le Saux, Leslie Collins, Kyle Bradbury, et al. Large-scale semantic classification: outcome of the first year of inria aerial image labeling benchmark. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 6947–6950. IEEE, 2018. 5

[8] Hao Jing, Xian Sun, Zhirui Wang, Kaiqiang Chen, Wenhui Diao, and Kun Fu. Fine building segmentation in high-resolution sar images via selective pyramid dilated network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:6608–6623, 2021. 2

[9] Jian Kang, Zhirui Wang, Ruoxin Zhu, Junshi Xia, Xian Sun, Ruben Fernandez-Beltran, and Antonio Plaza. Disoptnet: Distilling semantic knowledge from optical images for weather-independent building segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022. 3

[10] Sannidhi P Kumar, Chandan Gautam, and Suresh Sundaram. Meta-cognition-based simple and effective approach to object detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3795–3799. IEEE, 2021. 4

[11] Eungbean Lee, Somi Jeong, Junhee Kim, and Kwanghoon Sohn. Semantic equalization learning for semi-supervised sar building segmentation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 4

[12] 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, 5

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[14] Yaohui Liu, Jie Zhou, Wenhua Qi, Xiaoli Li, Lutz Gross, Qi Shao, Zhengguang Zhao, Li Ni, Xiwei Fan, and Zhiqiang Li. Arc-net: An efficient network for building extraction from high-resolution aerial images. *IEEE Access*, 8:154997–155010, 2020. 5

[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[17] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3226–3229. IEEE, 2017. 2, 5

[18] Malsha V Perera, Wele Gedara Chaminda Bandara, Jeya Maria Jose Valanarasu, and Vishal M Patel. Sar despeckling using overcomplete convolutional networks. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 401–404. IEEE, 2022. 3, 6

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1, 2, 5

[20] Mohammad Rostami, Soheil Kolouri, Eric Eaton, and Kyungnam Kim. Sar image classification using few-shot cross-domain transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[21] Jacob Shermeyer, Daniel Hogan, Jason Brown, Adam Van Etten, Nicholas Weir, Fabio Pacifici, Ronny Hansch, Alexei Bastidas, Scott Soenen, Todd Bacastow, et al. Spacenet 6: Multi-sensor all weather mapping dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 196–197, 2020. 1, 2, 5

[22] Aniruddh Sikdar, Sumanth Udupa, Suresh Sundaram, and Narasimhan Sundararajan. Fully complex-valued fully convolutional multi-feature fusion network(FC2MFN) for building segmentation of insar images. In *2022 IEEE Symposium*

Series on Computational Intelligence (SSCI), pages 581–587, 2022. 2

[23] K Steinnocher, Andréa De Bono, Bruno Chatenoux, Dirk Tiede, and L Wendt. Estimating urban population patterns from stereo-satellite imagery. *European Journal of Remote Sensing*, 52(sup2):12–25, 2019. 1

[24] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 4, 5

[25] Jeya Maria Jose Valanarasu and Vishal M Patel. Fine-context shadow detection using shadow removal. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1705–1714, 2023. 3

[26] Jeya Maria Jose Valanarasu, Vishwanath A Sindagi, Ilker Hacihaliloglu, and Vishal M Patel. Kiu-net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation. *IEEE Transactions on Medical Imaging*, 41(4):965–976, 2021. 2, 3, 4, 6

[27] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 3

[28] Yongzhi Wang, Hua Lv, Rui Deng, and Shengbing Zhuang. A comprehensive survey of optical remote sensing image segmentation methods. *Canadian Journal of Remote Sensing*, 46(5):501–531, 2020. 1

[29] Wei Yao, Dimitrios Marmanis, and Mihai Datcu. Semantic segmentation using deep neural networks for sar and optical image pairs. 2017. 2

[30] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 4

[31] 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

[32] Zhuo Zheng, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Deep multisensor learning for missing-modality all-weather mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 174:254–264, 2021. 2

[33] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4):8–36, 2017. 2