IMPROVING CNN-BASED BUILDING SEMANTIC SEGMENTATION USING OBJECT BOUNDARIES

E. Bousias Alexakis*, C. Armenakis

Geomatics Engineering, GeoICT Lab, Department of Earth and Space Science and Engineering, Lassonde School of Engineering, York University, Toronto, Canada – {bousiasa, armenc}@yorku.ca

Commission III, WG III/1

KEY WORDS: Building Extraction, CNN, Building Boundaries, Semantic Segmentation, Decoupled Body and Edge Segmentation

ABSTRACT:

Semantic segmentation is an active area of research with a wide range of applications including autonomous driving, digital mapping, urban monitoring, land use analysis and disaster management. For the past few years approaches based on Convolutional Neural Networks, especially end-to-end approaches based on architectures like the Fully Convolutional Networks (FCN) and UNet, have made great progress and are considered the current state-of-the-art. Nevertheless, there is still room for improvement as CNN-based supervised-learning models require a very large amount of labelled data in order to generalize effectively to new data and the segmentation results often lack detail, mostly in areas near the boundaries between objects. In this work we leverage the semantic information provided by the objects' boundaries to improve the quality and detail of an encoder-decoder model's semantic segmentation output. We use a UNet-based model with ResNet as an encoder for our backbone architecture in which we incorporate a decoupling module that separates the boundaries from the main body of the objects and thus learns explicit representations for both body and edges of each object. We evaluate our proposed approach on the Inria Aerial Image Labelling dataset and compare the results to a more traditional Unet-based architecture. We show that the proposed approach marginally outperforms the baseline on the mean precision, F1-score and IoU metrics by 1.1 to 1.6%. Finally, we examine certain cases of misclassification in the ground truth data and discuss how the trained models perform in such cases.

1. INTRODUCTION

Today, there is an unprecedented abundance of Earth observation image data available for geospatial applications, many of which are also provided free-of-charge. This data abundance, however, leads to new challenges regarding the development of methods for the automatic processing and organization of the data in order to facilitate their effective use in various applications.

Semantic segmentation is a fundamental computer vision task whose goal is to assign an object class label to each pixel of an image. For the past few years approaches based on Deep Convolution Neural Network (DCNN) architectures that perform end-to-end semantic segmentation have been very successful and are considered the current state-of-the-art in remote sensing supervised pixel-wise classification. CNN-based semantic segmentation algorithms have been successfully applied for the extraction of building footprints from satellite and aerial imagery. The automatic extraction of building footprints can be a very useful tool for many applications including but not limited to digital mapping, urban monitoring, land use analysis and disaster management.

The first successful DCNN architecture to perform end-to-end semantic segmentation, meaning that the network produced a final prediction map for each object class without the need for any post processing steps, was the Fully Convolutional Networks (Long et al., 2014). Although FCN excelled at many semantic segmentation benchmarks, they suffered from blurry outputs due to the repeated application of down sampling operators (either through max-pooling or striding) performed at consecutive layers of the encoder network which leads to features of reduced spatial resolution. Another challenge related to the existence of objects at various scales and the fact that the receptive field of FCN grew slowly with respect to the depth of the network limiting the ability of the network to fully model complex long-range relationships between image regions (Li et al., 2020). Many approaches have been proposed since then that aim to address these issues such as UNet (Ronneberger et al., 2015) and UNet++ (Zhou et al., 2018) architectures that introduce skip connections to improve the spatial accuracy of the model representations. Also, the multiple versions of the DeepLab architectures (Chen et al., 2017, 2018) improve the spatial accuracy and increase the receptive fields of convolutional layers by introducing dilated, also known as atrous (à trous), convolutions and the Atrous Spatial Pyramid Pooling (ASPP) layer to capture objects and image context at multiple scales.

Despite the significant improvements, there are still challenges that need to be addressed when it comes to end-to-end semantic segmentation. CNN-based supervised learning models still require a very large amount of labelled data in order to generalize effectively to new images and the segmentation outputs often lack in localization accuracy and detail, mostly in areas between objects boundaries. Furthermore, the presence of other objects near the buildings may introduce some uncertainty and/or some discontinuities in the extraction of the building footprint.

Various recent research works focus on leveraging the semantic information of object boundaries in order to improve the localization accuracy of the model's semantic segmentation predictions (Girard et al., 2021; Li et al., 2020; Marmanis et al., 2018; Zhao et al., 2018, 2020). Li et al., (2020) propose decoupling the body from the boundary of each segmented object within the model in order to learn separate representations for the body and edge parts, thus improving the final segmentation performance.

* Corresponding author

Inspired by the work of Li et al. (2020) we have developed a new model that learns explicit representations of the main body and the boundaries of buildings by incorporating a decoupling module into a UNet-based CNN architecture. We investigate the potential benefits of explicitly using the objects' boundary information in a CNN architecture to help the model learn better feature representations and improve the accuracy of the predicted outputs.

We compare the proposed model to an equivalent 'plain' UNet-based architecture and to a UNet-based network which besides the building segmented area also learns to predict the building edges as a separate object class. We also investigate whether the simpler approach, compared to the decoupled body and edge segmentation, of introducing a separate edge class helps the model learn better representations and improve the predicted footprints compared to the plain UNet-based architecture.

In section 2 we review recent research works pertaining to building extraction and the utilization of edges in CNN-based architectures for semantic segmentation. In section 3 we present the proposed network architecture, discuss the training details, and briefly introduce the Inria Aerial Image Labelling Dataset (Maggiori et al., 2017) that was used for the training and evaluation of all our models. In section 4 we present the results analysis and discussion and finally in section 5 we summarize our conclusions.

2. RELATED WORK

Building extraction has been widely studied by the remote sensing community with many CNN-based approaches being proposed (Chatterjee & Poullis, 2019; Ji et al., 2019; Shao et al., 2020; Xu et al., 2018; Yuan, 2018) and certain among them focusing on the refinement of the buildings' boundaries (Girard et al., 2021; Zhao et al., 2018, 2020). Zhao et al. (2020) propose a two-step method for a refined extraction of building boundaries that first uses a variation of mask R-CNN for building instance segmentation and then refine the noisy building information using a Graph Convolutional Network (GCN) that learns the geometric shapes of building polygons. Girard et al. (2021) introduced an additional frame field output, besides the building interiors and building boundary outputs to an encoder-decoder network for building extraction. The additional frame field information improves the segmentation quality and is also useful for the building boundaries polygonization via a newly introduced algorithm that extends the concept of the Active Contour Model (ACM).

In a more general context, there have been a few recent research works in the remote sensing and computer vision communities that studied the semantic segmentation enhancement via the use of boundary information.

Chen et al. (2016) proposed a method that combines a CNN architecture based on the DeepLab model and Domain Transform (DT) filtering, an edge preserving filtering method which smooths images based on an edge reference map. Marmanis et al. (2018) present a combination of the SEGNET encoder-decoder architecture and the Holistically-Nested Edge Detection (HED) network that has been trained using deep supervision on the ISPRS Vaihingen and the ISPRS Potsdam datasets. Their results suggest that including boundary information into the models improves the semantic segmentation performance of the networks. Similarly, Liu et al. (2018) developed ERN, an Edge loss Reinforced semantic segmentation Network that leverages the boundary context information that improves the semantic segmentation performance on UAV images. The network has an encoder-decoder architecture similar to UNet, where the convolutional blocks have been replaced by inception blocks. It incorporates edge outputs for certain intermediate blocks of the network that are then aggregated to the segmentation loss function of the model. Similarly, Jung et al. (2022) adopts HED, which extracts edge features at an encoder of a given architecture and in the proposed boundary enhancement module, an extracted edge and segmentation mask are combined, sharing mutual information.

Lyu et al. (2019) combine the outputs of an edge detection network based on MobileNet V2 and of a semantic segmentation network based on ESPNetV2 through a multilayer fusion module in order to perform real-time semantic segmentation on the Cityscapes dataset. Similarly, He et al. (2020) experimented with fusing the outputs of a Fully Convolutional Network (FCN) with edge information derived by a HED model and showed improved semantic segmentation performance compared to simply using an FCN network on the ESAR GID remote sensing datasets (Tong et al., 2020). He et al. (2021) propose an enhanced boundary learning approach in the very challenging scenario of glass-like object segmentation. They introduce a Refined Differential Module, which works with multiple resolution input features in a coarse-to-fine manner and learns to predict the edge, body and the complete glass object, and a Point-based Graph convolution network Module (PGM) which is used to improve the final prediction.

Li et al. (2020) developed a framework that models explicitly the body and edges of objects during semantic segmentation. They integrate their framework into a DeepLabv3+ architecture and they use the principles of label relaxation (Zhu et al., 2019) to train their models. Their approach, named decoupled body and edge supervision, achieved state-of-the-art results on road scene semantic segmentation benchmarks.

In a previous work we experimented with the potential benefits of combining an encoder-decoder architecture like UNet with boundary information produced by a deep architecture for change detection applications (Bousias Alexakis & Armenakis, 2021). We introduced the semantically rich boundaries, which were determined using DexiNeD (Soria et al., 2020), at multiple parts of the encoder-decoder architecture and concluded that the highest benefits in the segmentation quality were reported when the features were introduced both at the first and the last convolutional block of the architecture.

3. METHODOLOGY

In this work we try to leverage the semantic information of building boundaries to improve the localization accuracy of the semantic segmentation predictions. The proposed architecture was designed to combine the benefits of a UNet based encoder-decoder architecture and the body and edge decoupling module introduced by Li et al. (2020).

In section 3.1 we describe our backbone UNet-based architecture, which will also be used as our baseline in our experiments. In section 3.2 we introduce the body and edge decoupling module and present the proposed architecture. In sections 3.3 to 3.5 we present the loss functions and the dataset we used for the training of the models as well as some training details.

3.1 Baseline Model – Backbone architecture

UNet consists of a down-sampling part (encoder) that comprises multiple blocks of convolutional, ReLU and max-pooling layers and a symmetric up-sampling part (decoder) in which the max-

pooling layers have been replaced by up-sampling layers in order to expand their input to its original dimensions. The decoder is symmetric to the encoder in the sense that the up-sampling blocks output the same number of feature channels as their corresponding down-sampling blocks. Thanks to the larger number of feature channels in the decoding blocks compared to FCN and the skip connections that pass information from the contracting to the corresponding expanding blocks the model can retain both the rich contextual information of the features coming from deeper convolutional blocks as well as the spatial detail and localization accuracy of the features coming from the shallower blocks of the network (Figure 1).

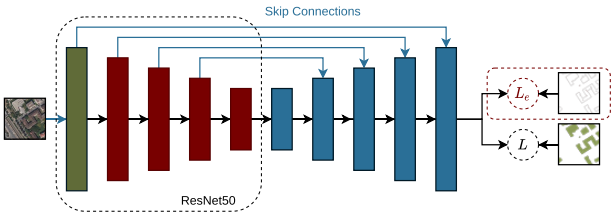

Figure 1. Baseline model.

Our baseline model has certain improvements compared to the traditional UNet architecture. First, we have replaced the original encoder with the much deeper and more expressive ResNet50 model (He et al., 2016). Residual Networks (ResNets) introduce shortcut connections between layers to address the degradation of training accuracy that is observed in very deep networks. Second, we have introduced batch normalization layers (Ioffe & Szegedy, 2015) in all up-sampling blocks as they have been shown to facilitate deep models training and help address the vanishing/ exploding gradients problem. We have also experimented with introducing the building edges as an additional output of the network besides the building footprints to investigate whether adding an extra supervised task would lead to more descriptive features and to even small improvement in the accuracy of the building footprints predictions.

3.2 Body and Edge Segmentation

As we mentioned earlier, we aim to improve the performance of our backbone architecture by incorporating a decoupled attention module into our architecture that decomposes the incoming features/signals into a low frequency (the object's body) and a high frequency component (the object's edges). By separating the body and edge of each object of interest the module helps the network learn features that explicitly relate to the edges of objects and thus produce more accurate segmentation predictions especially near the objects' boundaries.

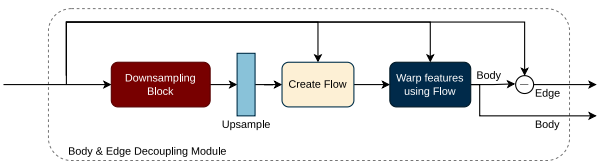

Figure 2. Body and edge decoupling module.

The decoupled attention module we used is based on the work of (Li et al., 2020). The model first learns to detect the body of objects by comparing a certain set of intermediate features that are used as input to the module to a low frequency/simplified version of them, which is produced by first down-sampling the feature set by a factor of four and then up-sampling the output to its original resolution through linear interpolation (Figure 2). The next step is to use a sequence of convolutional layers to learn a flow field that maps the low frequency output of the up-sampling layer to the sharper and more detailed original features that were used as input to the module creating a field with a flow pointing towards the body's main solid area of each feature map object. Next, the learned flow field is used to warp the original features by linear interpolation and thus estimate a new body feature map for each feature of the original feature map. The feature edge maps can then be computed by simply subtracting the body from the original feature map.

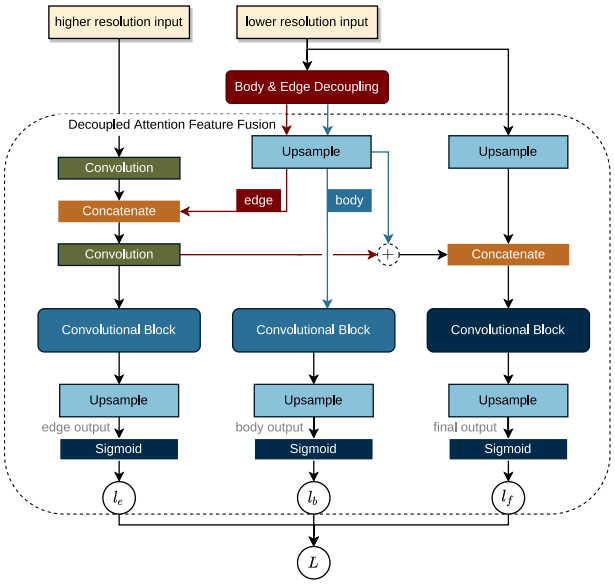

Figure 3. Body and Edge feature fusion. Model final outputs.

After computing the initial body and edge feature maps via the decoupling module we combine the module's outputs with the high-resolution features produced by the first convolutional block of our encoding network, and we fuse the edge and body features into a final footprint output that encloses both body and edges as shown in Figure 3. We apply a final convolutional block on each of the edge, body, and fused feature maps to produce the final prediction map of the model.

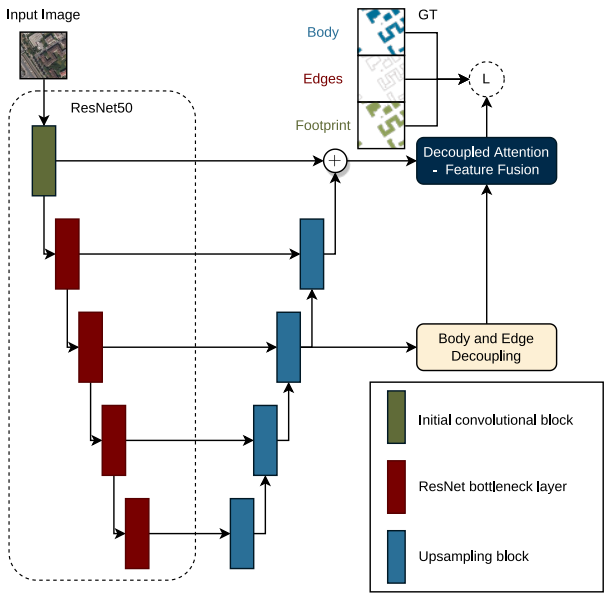

Figure 4. Proposed architecture.

The body, edge and the fused output streams are all trained in a supervised manner. Given the original 'complete' building masks of the dataset one can easily extract the edge and body masks that

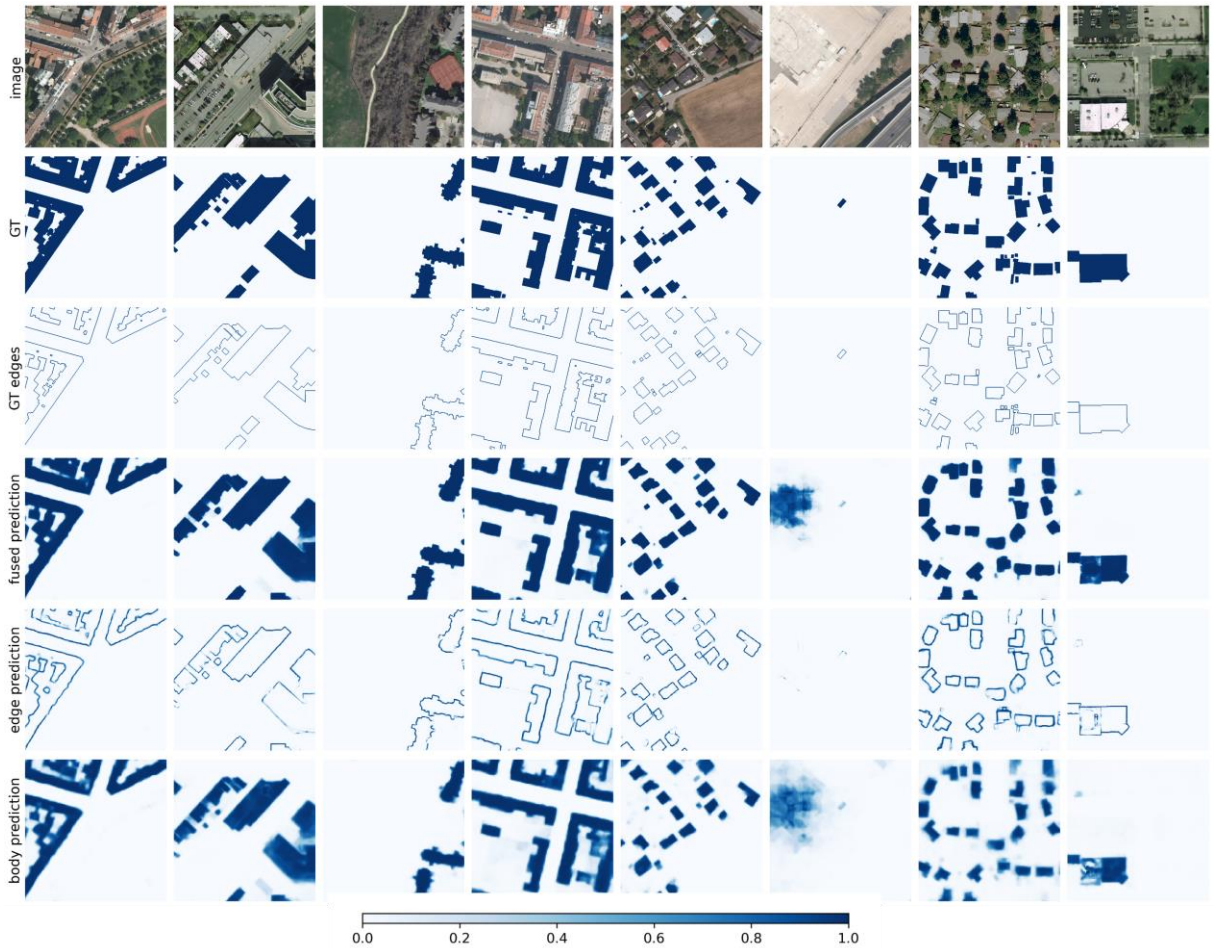

Figure 5. Random image samples together with the ground truth building footprints (row 2) and building boundaries (row 3) as well as the proposed model predictions.

can then be used to train the network (see Section 3.4). An overview of the proposed approach is given in Figure 4.

3.3 Loss Functions

Our loss function consists of 3 separate components: the first component, l_f, relates to the final/fused result compared to the building footprint mask, a second term, l_e, measures the agreement between the predicted and the ground truth edges and finally the l_b term relates to the body part of the predicted objects. We have used a combination of the Binary Cross Entropy (BCE) loss and the dice coefficient (Eq. 1) to model the l_f and l_e components and the binary cross entropy loss (Eq. 2) to model the body component l_b of the loss function. The total loss is an aggregate of the three losses with the building loss given a smaller weight than the other two components (Eq. 3).

$$l_f = -\frac{1}{n}\cdot\sum_{i=1}^{n}(\lambda_1\cdot(y_i\cdot\log(\hat{y}_i)+(1-\hat{y}_i)\cdot\log(1-y_i))+\lambda_2\cdot\frac{2\cdot y_i\cdot\hat{y}_i}{y_i+\hat{y}_i}) \quad (1)$$

$$l_b = -\frac{1}{n}\cdot\sum_{i=1}^{n}(y_i\cdot\log(\hat{y}_i)+(1-\hat{y}_i)\cdot\log(1-y_i)) \quad (2)$$

$$L = l_f + l_e + 0.5\cdot l_b \quad (3)$$

where y_i is the flattened model's predictions for image i (could be referring to either body or edge of final prediction depending on the context), $\hat{y}_i$ is the flattened ground truth values for image i, n is the number of images per batch and λ_1 and λ_2 are weighting parameters that were set to 0.5 and 1 respectively for all training experiments for both the fused and the edge components of the loss function.

3.4 Dataset

For our experiments we used the Inria Aerial Image Labelling dataset (Maggiori et al., 2017), which consists of aerial orthorectified RGB imagery of 0.3m ground resolution and the corresponding binary masks classifying each pixel as either building or not building. The dataset covers an area of 810 km^2 of dissimilar urban settlements from different regions around the globe captured with varying illumination conditions and at different seasons, aiming to assess the generalization capacity of the applied models.

As mentioned in Section 3.2 we extracted the boundaries of each building from the dataset's binary building masks to create the edge ground truth masks. We then subtracted the edges from the edge areas from the original building footprints to create the body ground truth masks (Figure 6). For our experiments we have randomly split the 180 images of the training dataset into a training and a validation set consisting of 145 and 35 images respectively.

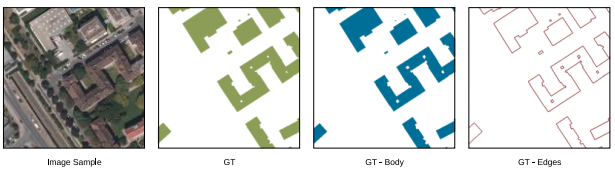

Figure 6. Image and ground truth masks sample.

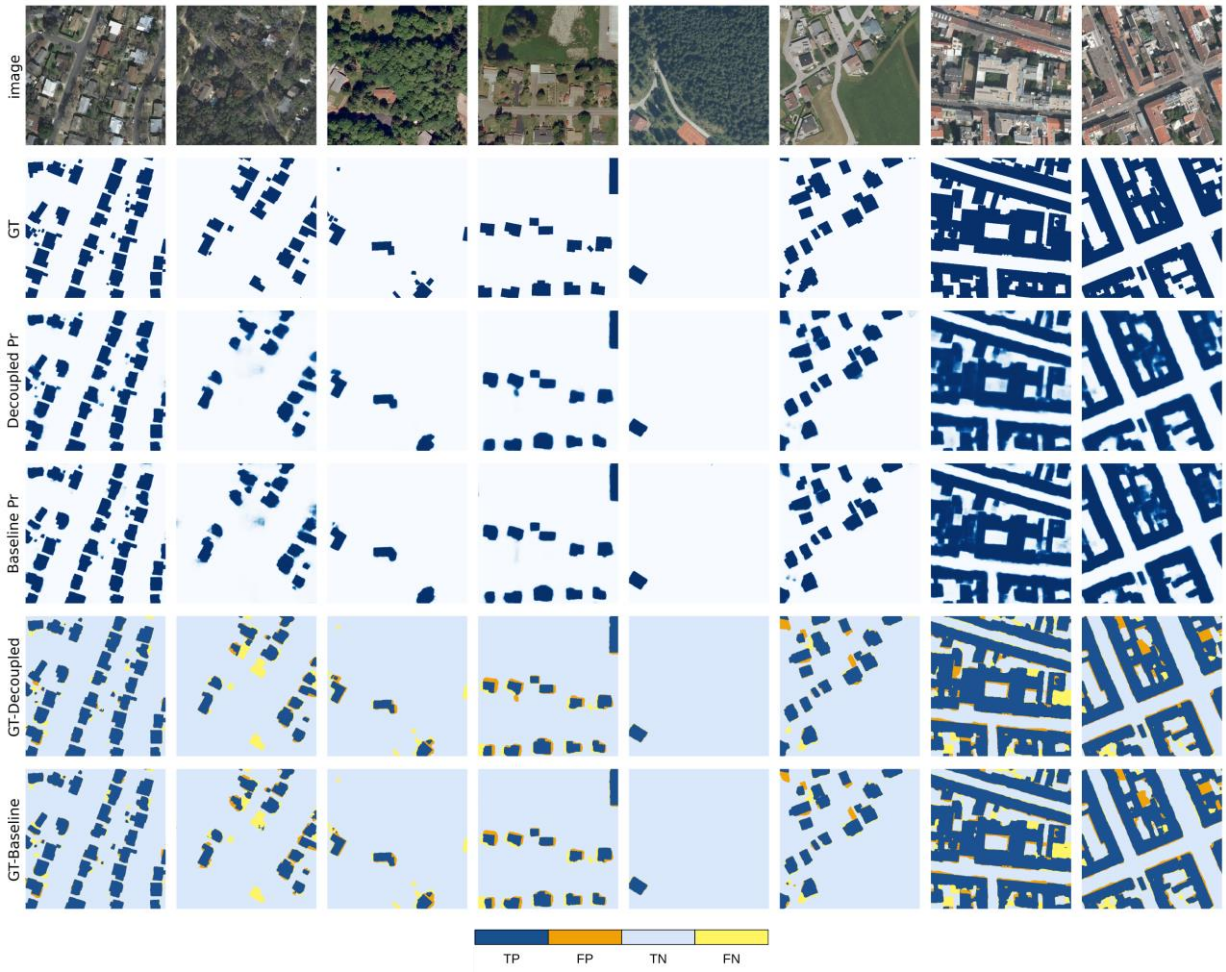

Figure 7. Comparison between the predicted results and the ground truth data on random image samples.

3.5 Model training

For our encoder network we have used the ResNet50 architecture pretrained on ImageNet. Using an encoder with pretrained weights on a very big dataset like ImageNet is a simple and common transfer learning technique that significantly reduces training time by taking advantage of the expressive features learned on the big dataset, especially the features of the first few layers which correspond to simple visual building blocks like edges, corners and simple blob-like structures that are generally task-independent.

For the training of all models, we used the Adam optimizer with the default parameters for the coefficients of the running average (0.9 and 0.999) and without weight decay. We used a batch size of eight and a learning rate of 0.0003 for the first 54000 iterations which was then reduced to 0.0001 for the final 18000 iterations. The training was performed using the PyTorch framework (Paszke et al., 2019) on an NVIDIA Quadro RTX 5000 GPU.

3.5.1 Data Augmentation

Both for training and validation we used 512×512 pixels image samples. Since the original image size is 5000×5000 pixels for each training image we randomly crop 512×512 sample windows that we use as input to our models. We also randomly apply horizontal and vertical flips and 90º rotations to the sampled images as well as random shifts in the interval [-20,20] to the RGB values of the image samples to help the model generalize better and reduce overfitting to the training data. For the validation set, we simply divide each image into 100 smaller images (with a very small overlap between image samples) in order to be able to monitor the training progress of a model in a consistent way.

4. RESULTS

In this section we present the results for the proposed model that incorporates the decoupled body and edge module, referred as ResUNet[1] Decoupled, and compare them to the results of the two baseline approaches, referred as Plain ResUNet and Plain ResUNet with edges respectively.

Figure 5 presents 8 random examples of ResUNet Decoupled predictions together with the ground truth building footprints (row 2) and boundaries (row 3). The fused building footprint predictions are presented in row 4 and the edge and body predictions in rows 5 and 6 respectively. The values of the color scale correspond to the confidence of the model that a pixel belongs to the building class with 1 meaning that the pixel is most certainly a part of a building (footprint or edge) and 0 meaning that the pixel is part of the background. A visual inspection of the examples suggests that the model predicts most of the building footprints and the building boundaries successfully. In certain

[1] Network architecture described in Section 3.1. Not to be confused with ResUNet by Zhang et al. (2018)

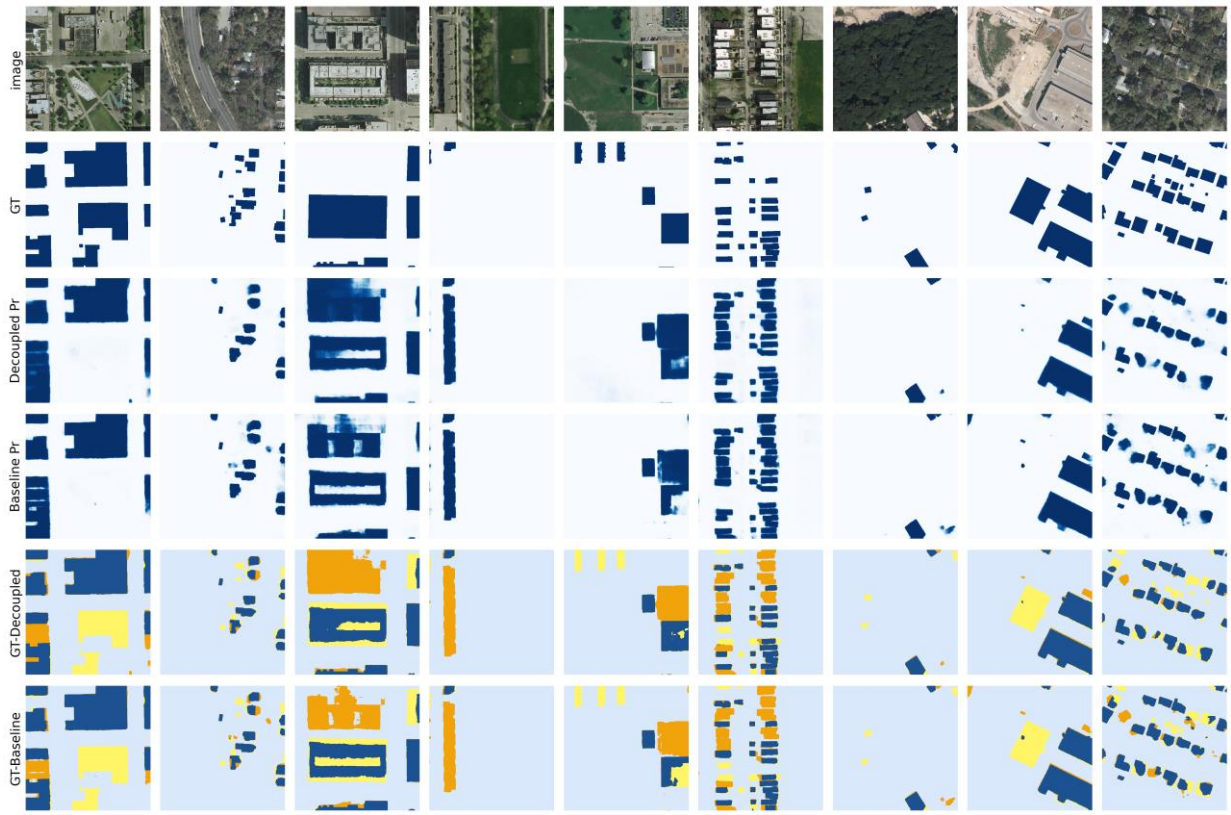

Figure 8. Examples of poor ground truth building masks.

more challenging cases where the buildings might:
- not be fully visible,
- be located near the image boundaries (columns 2 and 4),
- have intricate structures (column 4)
- have strong shadows (column 2)

then the predictions of the fused building footprint can be blurry, and the boundaries not so well defined for certain parts of the building.

For the quantitative evaluation of the models' performance, we used the average per image values of precision, recall, F1 score and Intersection over Union (IoU) for the validation set (Equations 4 to 7). TP (True Positive) refers to pixels that were correctly identified as buildings, TN (True Negative) refers to pixels that were correctly classified as background, FP (False Positive) are the pixels that were erroneously classified as buildings and FN (False Negative) are pixels that should have been classified as buildings but were instead classified as background.

$$precision = \frac{TP}{TP + FP} \quad (4)$$

$$recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 = 2\frac{precision \cdot recall}{precision + recall} \quad (6)$$

$$IoU = \frac{TP}{TP + FN + FP} \quad (7)$$

Table 1 summarizes the mean evaluation metrics per image for the validation set for all three models. We can see that for most metrics (precision, F1 score and IoU) the proposed approach (ResUNet – Decoupled) outperforms the baseline models by a small margin (from 1.1 to 2.7%). However, the plain ResUNet has a slightly higher recall rate (less than 0.3%) compared to the proposed approach. A visual comparison of the Decoupled semantic segmentation model's performance compared to the two baseline models is illustrated in Figure 9.

Model	Precision	Recall	F1	IoU
Plain ResUNet	0.8457	**0.8862**	0.8453	0.7632
Plain ResUNet & edges	0.8344	0.8828	0.8314	0.7493
ResUNet – Decoupled	**0.8617**	0.8837	**0.8565**	**0.7754**

Table 1. Mean evaluation metrics per image for the validation dataset.

Contrary to our expectations the introduction of edges as an extra supervised task in the Plain ResUNet & edges model did not lead to improvements compared to the Plain ResUNet model but instead led to a small deterioration of the values for all four metrics. One possible explanation of this deterioration might be that the additional edge supervision makes the model less robust to misclassification errors that are present in the dataset and leads to this small degradation of the average metrics' values.

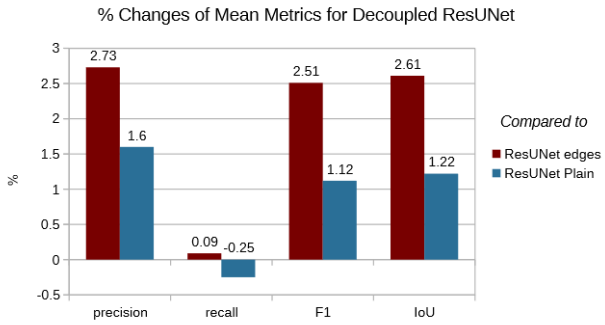

Figure 9. Comparison of average evaluation metrics between the proposed architecture and the baseline models.

Figure 7 presents the final predictions of the ResUNet Decoupled model and the Plain ResUNet with edges model and a comparison between the predicted and the ground truth masks. The second row presents the ground truth mask for each image. The third and fourth rows present the predicted (fused) building footprints for the proposed model (ResUNet Decoupled) and the Plain ResUNet with edges model respectively. The last two rows present a per pixel comparison between the predicted and the ground truth binary masks. The color scale for row 3 and 4 is the same with the one of Figure 5.

Based on a visual comparison of the results of the two models (Fig. 7) there are no distinguishable benefits when using the ResUNet Decoupled architecture. It seems that the small quantitative improvements on the average evaluation metrics are not strongly reflected in the visual inspection. Both models seem to perform relatively well on predicting regularly shaped buildings, and both have small discrepancies from the ground truth when predicting the footprints of certain complicated building structures such as the examples of the last two columns. Regarding the example on the second column, we can see that the two building masks that have been classified as FN in the comparison rows for both models do not actually correspond to a building in the image but are a result of misclassification errors in the ground truth masks. Another building on the top middle of the same image is partly misclassified as a result of tree canopy occlusions. The last two observations introduce another important aspect of the validation process which relates to the validity of the ground truth data predictions.

In Figure 8 we present examples that highlight cases with inconsistencies in the ground truth masks. The organization of the results and the color bar used for the comparisons are the same with Figure 7. Examples in columns 1, 5 and 8 illustrate cases where the ground truth mask includes building footprints for buildings that are not present in the images. In all 3 cases both models correctly classified these regions as background. In examples 3,4 and 6 we can see cases of missing footprints from the ground truth masks. Finally, in examples 2,7 and 9 we can see cases of certain small buildings being partly or entirely occluded by the tree canopy. Although most ground truth building masks are valid there are multiple cases of inconsistencies between ground truth building footprints and the corresponding images, which cannot be easily quantified. Since most of the GT misclassification cases have been (at least partly) correctly classified by the models, an interesting future application might be the creation of a semi-automatic process for the refinement of the dataset labels.

5. CONCLUSION

In this work we investigated the benefits of leveraging semantic information of objects' boundaries in CNN architectures to improve the localization accuracy of building footprints semantic segmentation from aerial imagery. We proposed a new end-to-end UNet/ResNet based approach that incorporates an edge and body decoupling module and showed that for most evaluation metrics it outperforms by a small margin equivalent architectures that do not explicitly model the objects' body and edge.

ACKNOWLEDGEMENTS

This work is financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery grant) and York University.

REFERENCES

Bousias Alexakis, E., & Armenakis, C. (2021). Performance Improvement of Encoder/Decoder-Based CNN Architectures for Change Detection from Very High-Resolution Satellite Imagery. *Canadian Journal of Remote Sensing*, *47*(2), 309–336. https://doi.org/10.1080/07038992.2021.1922880

Chatterjee, B., & Poullis, C. (2019). *Semantic Segmentation from Remote Sensor Data and the Exploitation of Latent Learning for Classification of Auxiliary Tasks*. https://arxiv.org/abs/1912.09216v1

Chen, L.-C., Barron, J. T., Papandreou, G., Murphy, K., & Yuille, A. L. (2016). *Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform*. 4545–4554. https://doi.org/10.1109/CVPR.2016.492

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848. https://doi.org/10.1109/tpami.2017.2699184

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation. *ArXiv:1706.05587 [Cs]*. http://arxiv.org/abs/1706.05587

Girard, N., Smirnov, D., Solomon, J., & Tarabalka, Y. (2021). Polygonal Building Segmentation by Frame Field Learning. *ArXiv:2004.14875 [Cs, Eess]*. http://arxiv.org/abs/2004.14875

He, C., Li, S., Xiong, D., Fang, P., & Liao, M. (2020). Remote Sensing Image Semantic Segmentation Based on Edge Information Guidance. *Remote Sensing*, *12*(9), 1501. https://doi.org/10.3390/rs12091501

He, C., Li, S., Xiong, D., Fang, P., & Liao, M. (2020). Remote Sensing Image Semantic Segmentation Based on Edge Information Guidance. Remote Sensing, 12(9), 1501. https://doi.org/10.3390/rs12091501

He, H., Li, X., Cheng, G., Shi, J., Tong, Y., Meng, G., Prinet, V., & Weng, L. (2021). *Enhanced Boundary Learning for Glass-Like Object Segmentation*. 15859–15868. https://openaccess.thecvf.com/content/ICCV2021/html/He_Enhanced_Boundary_Learning_for_Glass-Like_Object_Segmentation_ICCV_2021_paper.html

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Ji, S., Wei, S., & Lu, M. (2019). Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(1), 574–586. https://doi.org/10.1109/TGRS.2018.2858817

Jung H, Choi H-S, Kang M (2022) Boundary Enhancement Semantic Segmentation for Building Extraction From Remote Sensed Image, IEEE Transactions on Geoscience and Remote Sensing, Vol. 60, 2022

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, 448–456.

Li, X., Li, X., Zhang, L., Cheng, G., Shi, J., Lin, Z., Tan, S., & Tong, Y. (2020). Improving Semantic Segmentation via Decoupled Body and Edge Supervision. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 435–452). Springer International Publishing. https://doi.org/10.1007/978-3-030-58520-4_26

Liu, S., Ding, W., Liu, C., Liu, Y., Wang, Y., & Li, H. (2018). ERN: Edge Loss Reinforced Semantic Segmentation Network for Remote Sensing Images. *Remote Sensing*, *10*(9), 1339. https://doi.org/10.3390/rs10091339

Long, J., Shelhamer, E., & Darrell, T. (2014). *Fully Convolutional Networks for Semantic Segmentation*. https://arxiv.org/abs/1411.4038v2

Lyu, H., Fu, H., Hu, X., & Liu, L. (2019). Esnet: Edge-Based Segmentation Network for Real-Time Semantic Segmentation in Traffic Scenes. *2019 IEEE International Conference on Image Processing (ICIP)*, 1855–1859. https://doi.org/10.1109/ICIP.2019.8803132

Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 3226–3229. https://doi.org/10.1109/IGARSS.2017.8127684

Marmanis, D., Schindler, K., Wegner, J. D., Galliani, S., Datcu, M., & Stilla, U. (2018). Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, *135*, 158–172. https://doi.org/10.1016/j.isprsjprs.2017.11.009

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28

Shao, Z., Tang, P., Wang, Z., Saleem, N., Yam, S., & Sommai, C. (2020). BRRNet: A Fully Convolutional Neural Network for Automatic Building Extraction From High-Resolution Remote Sensing Images. *Remote Sensing*, *12*(6), 1050. https://doi.org/10.3390/rs12061050

Soria, X., Riba, E., & Sappa, A. (2020). Dense Extreme Inception Network: Towards a Robust CNN Model for Edge Detection. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1912–1921. https://doi.org/10.1109/WACV45572.2020.9093290

Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., & Zhang, L. (2020). Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sensing of Environment, 237, 111322. https://doi.org/10.1016/j.rse.2019.111322

Xu, Y., Wu, L., Xie, Z., & Chen, Z. (2018). Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters. *Remote Sensing*, *10*(1), 144. https://doi.org/10.3390/rs10010144

Yuan, J. (2018). Learning Building Extraction in Aerial Scenes with Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(11), 2793–2798. https://doi.org/10.1109/TPAMI.2017.2750680

Zhang, Z., Liu, Q., & Wang, Y. (2018). Road Extraction by Deep Residual U-Net. IEEE Geoscience and Remote Sensing Letters. https://doi.org/10.1109/LGRS.2018.2802944

Zhao, K., Kamran, M., & Sohn, G. (2020). Boundary Regularized Building Footprint Extraction from Satellite Images Using Deep Neural Networks. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-2–2020*, 617–624. https://doi.org/10.5194/isprs-annals-V-2-2020-617-2020

Zhao, K., Kang, J., Jung, J., & Sohn, G. (2018). Building Extraction from Satellite Images Using Mask R-CNN with Building Boundary Regularization. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 242–2424. https://doi.org/10.1109/CVPRW.2018.00045

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *ArXiv:1807.10165 [Cs, Eess, Stat]*. http://arxiv.org/abs/1807.10165

Zhu, Y., Sapra, K., Reda, F. A., Shih, K. J., Newsam, S., Tao, A., & Catanzaro, B. (2019). Improving Semantic Segmentation via Video Propagation and Label Relaxation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8848–8857. https://doi.org/10.1109/CVPR.2019.00906