On Building Classification from Remote Sensor Imagery Using Deep Neural Networks and the Relation Between Classification and Reconstruction Accuracy Using Border Localization as Proxy

Bodhiswatta Chatterjee and Charalambos Poullis
Immersive and Creative Technologies Lab, Department of Computer Science and Software Engineering
Gina Cody School of Engineering and Computer Science, Concordia University, Montreal, Quebec, Canada
bodhiswattachatterjee@gmail.com, charalambos@poullis.org

***Abstract*—Convolutional neural networks have been shown to have a very high accuracy when applied to certain visual tasks and in particular semantic segmentation. In this paper we address the problem of semantic segmentation of buildings from remote sensor imagery. We present ICT-Net: a novel network with the underlying architecture of a fully convolutional network, infused with feature re-calibrated Dense blocks at each layer. Uniquely, the proposed network combines the localization accuracy and use of context of the U-Net network architecture, the compact internal representations and reduced feature redundancy of the Dense blocks, and the dynamic channel-wise feature re-weighting of the Squeeze-and-Excitation(SE) blocks. The proposed network has been tested on INRIA's benchmark dataset and is shown to outperform all other state-of-the-art by more than 1.5% on the Jaccard index.**

Furthermore, as the building classification is typically the first step of the reconstruction process, in the latter part of the paper we investigate the relationship of the classification accuracy to the reconstruction accuracy. A comparative quantitative analysis of reconstruction accuracies corresponding to different classification accuracies confirms the strong correlation between the two. We present the results which show a consistent and considerable reduction in the reconstruction accuracy. The source code and supplemental material is publicly available at http://www.theICTlab.org/lp/2019ICTNet/.

***Keywords*-building classification; building reconstruction; classification accuracy; reconstruction accuracy; relationship between classification and reconstruction accuracy;**

I. Introduction

Reconstructing large-scale urban areas is an inherently complex problem which involves a number of vision tasks. Typically, the first step is classification where the objective is to label each pixel into an urban feature type e.g., building, road, tree, car, ground, vegetation, etc. Next, the pixel-level labels are used to cluster the pixels into contiguous groups corresponding to instances of the urban features they represent. Finally, the reconstruction is performed on each cluster. A reconstruction algorithm is applied on each cluster according to the urban feature type the cluster corresponds to. In the case of clusters corresponding to buildings, a boundary refinement process is typically performed prior to extruding the building facades.

The objectives and contributions of this paper are twofold. Firstly, we address the problem of the classification of buildings in remote sensor imagery. We investigate a number of state-of-the-art deep neural network architectures and present a comparative study of the results along with a reasoned justification on the design decisions for the proposed network named ICT-Net: a novel network with the underlying architecture of a fully convolutional network infused with Dense feature re-calibrated blocks at each layer. We demonstrate that this combination of components leads to superior performance. The proposed network is ranked first at an international benchmark competition organized by INRIA with more than 1.5% difference in terms of performance from the second best and other ensemble networks.

Secondly, we address the problem of reconstruction of the classified buildings and in particular study the relationship between the classification accuracy and reconstruction accuracy. We perform a comparative quantitative analysis on the reconstructions corresponding to classifications of different accuracies and report the results. Due to the lack of depth information, reconstructing 3D models is not feasible therefore the accuracy of the border localization is used as proxy for the evaluation since it is tightly coupled to the reconstruction accuracy i.e. buildings are extruded using their boundaries. As anticipated there is a strong correlation between the classification accuracy and the accuracy of the reconstruction however the analysis has shown that there is a consistent and considerable decrease in the reconstruction accuracy in terms of the per-pixel and per-building Jaccard indices. To the best of our knowledge this is the first time a quantitative analysis is performed in order to establish how the classification accuracy relates to the accuracy of the reconstruction as determined by the accuracy of the border localization.

Paper organization: The paper is organized as follows: Section II presents an overview of state-of-the-art in the area of classification of urban features in satellite images using deep neural networks. In Section III we summarize the work reported in the paper. The proposed neural network ICT-Net is explained in Section IV including a reasoned justification of the design decisions, and details on the training and testing of the network. Section V presents a quantitative analysis of the reconstruction accuracies resulting from different classification accuracies, and Section VI concludes the work and discusses future directions.

II. Related Work

Over the years object recognition has become one of the most addressed vision challenges and as a result a plethora of work has already been proposed. Initially the goal was on image classification where the entire image was classified according to the single object it contained; with some of the most important work in this area being [15], [26], [27]. More recently, the objective has shifted towards the semantic object segmentation or semantic labeling where multiple objects contained in a single image are labeled according to their class at the pixel level. Some of the most important work in this area is the work in [16], [1], [23]. Recent techniques using deep neural networks have demonstrated excellent results. Below we provide a brief overview of the state-of-the-art related to the area of semantic labeling with a particular focus on remote sensor imagery. A comprehensive review of neural network architectures for semantic segmentation can be found in [4].

Typical semantic segmentation architectures comprise of a down-sampling path responsible for feature extraction and an up-sampling path to restore the resolution of the semantic labels. Skip connections between the two paths help to have a smooth gradient back propagation and fast training of the network. The U-Net [23] architecture was able to achieve end-to-end semantic labeling with high accuracy in the field of medical image segmentation. Since then the U-Net [23] architecture has been extensively used and adapted to many other domains especially labeling of buildings from aerial imagery as in [14], [6], [11].

At the same time, deeper networks [27] have demonstrated the capacity to extract better features from images. Skip connections have been shown [7] to play a critical role in the training of very deep networks as they facilitate very good gradient propagation. There has been a lot of work on the pattern of skip connections with a very promising pattern known as Dense blocks proposed in [10] for the problem of image classification. In a Dense block every layer is connected to every other layer in a feed forward fashion. This provides implicit deep supervision and feature reuse which in turn improves the feature extraction power without making it difficult to train the network. The Tiramisu network architecture proposed in [13] extended the use of Dense blocks for semantic segmentation and was able to outperform state-of-the-art on two benchmark data sets: Gatech and CamVid.

Most deep neural networks for object recognition consider all extracted features at each layer to be of equal importance. This was until the method proposed in [8] showed that feature re-calibration i.e. weighing of the features, can be used effectively to model inter-dependencies between channels and produce even better performance with little computational overhead. Along a similar direction, in [24] the authors have shown that feature re-calibration combined with well known FCN networks perform well for medical image segmentation.

With respect to urban reconstruction, the extraction of urban geospatial features such as buildings from remote sensor imagery has also been an area of research interest for a very long time [25], [28], [5]. Automatic reconstruction of 3D models from the extracted features is extremely useful for many applications ranging from urban and community planning, development and architectural design, training of emergency response personnel, military personnel, etc. In [21] the authors propose a novel, robust, automatic segmentation technique based on the statistical analysis of the geometric properties of the data as well as an efficient and automatic modeling pipeline for the reconstruction of large-scale areas containing several thousands of buildings. With the recent advances in deep neural network architectures the pipeline has been upgraded to feature extraction using a semantic labeling CNN followed by clustering the points based on their label, and specialized processing for each of the labels of geospatial objects as proposed in [3].

Recently there has been a lot of interest for semantic labeling of buildings [18], [11], [14] fueled by the release of very large datasets like INRIA Aerial Image Labeling dataset [17], and SpaceNet where a corpus of commercial satellite imagery with labeled training data was made publicly available for use in machine learning research. In [11] the authors use a variant of the aforementioned U-Net network architectures replacing the VGG11 [26] encoder with a more powerful activated Batch Normalized [2] WideResnet-38 [30] in the context of instance segmentation of buildings for DeepGlobe-CVPR 2018 building detection sub-challenge, and were able to get very good results.

In this work, we propose ICT-Net: a novel network architecture that combines the strengths of deep neural network architectures (UNet) and building blocks (DenseNet block, SE block) which when applied to the problem of semantic labeling of buildings is proven to achieve better classification accuracy than state-of-the-art on the INRIA Aerial Image Labeling dataset. As of writing this manuscript the proposed network is top ranked on the competitions' leaderboard with more than 1.5% difference from the second best entry.

III. System Overview

Figure 1 summarizes the pipeline followed in the paper. Firstly, an orthorectified RGB image is fed forward into the neural network to produce a binary (building/non-building) classification map. Next, the binary classification map becomes the input to the reconstruction process. Due to the fact that it is extremely difficult to acquire building blueprints or CAD models for such large areas, and depth/3D information is not available for the images of the benchmark we posit that the building boundaries extracted from the binary classification map and refined, can serve as a proxy to the accuracy of the reconstruction. This is justified since the extracted boundaries are extruded in order to create the 3D models for the buildings. Therefore, the building boundaries are extracted, refined, and are used for the comparative analysis and evaluation of the accuracy of the reconstruction.

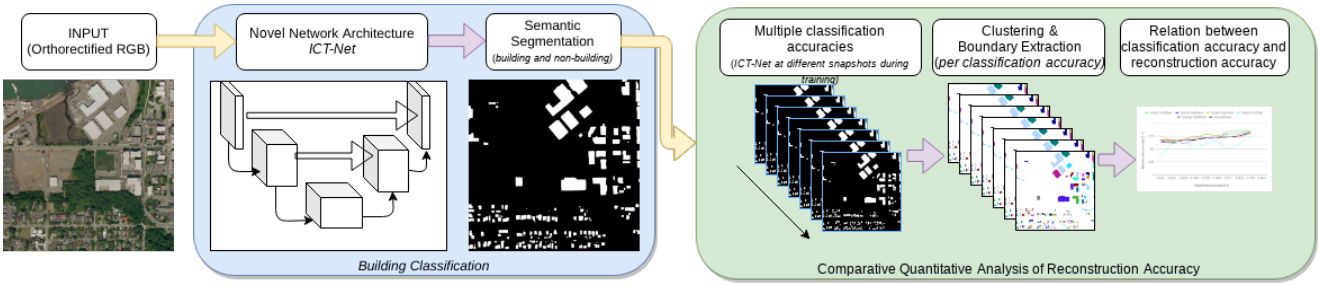


Figure 1: The diagram summarizes the work presented in this paper. Firstly, we focus on the building classification and propose a novel network architecture which outperforms state-of-the-art on benchmark datasets and is currently top-ranking. Secondly, we investigate the relation between the classification accuracy and the reconstruction accuracy and conduct a comparative quantitative analysis which shows a strong correlation but also a consistent and considerable decrease of the reconstruction accuracy when compared to the classification accuracy.

IV. Building Classification

In this section we describe the details of the proposed neural network architecture including information about the dataset, training/validation, and testing, as well as the justification for all design decisions.

A. Dataset

The training of the network is performed using the INRIA Aerial Image labeling dataset [17] which consists of pixelwise labeled aerial imagery for building classification. The dataset covers $810km^2$ area across 10 different cities with spatial resolution of $30cm$, and is split into two equal sets ($405Km^2$ each) for training and testing. The dataset consists of 3-band orthorectified RGB images and the training labels consist of ground truth data for two semantic classes: building and non-building. The training data covers parts of the cities of Austin, Chicago, Kitsap county, western Tyrol, and Vienna. The test data covers parts of the cities of Bellingham, Bloomington, Innsbruck, San Francisco and Eastern Tyrol. There are 36 tiles with resolution of 5000×5000 pixels for each city, each tile covering $1500 \times 1500m^2$ area on the ground. The training data is further divided into two sets: (1) the validation set which comprises of the first 5 tiles of each city, and (2) the training set which consists of the rest of the tiles as suggested in [17]. An example image from the dataset can be seen in Figure 1.

We have chosen the INRIA benchmark dataset over other available options because it uniquely offers two significant advantages. Firstly, the training and testing datasets are from *completely different cities* with no overlap i.e. *all* images of 5 cities (Austin, Chicago, Kitsap, Western-Tyrol, Vienna) are provided for training, and *all* images of another 5 different cities (Bellingham, Bloomington, Innsbruck, San Francisco, Eastern-Tyrol) are used for testing. Secondly,the dataset covers *dissimilar urban settlements* e.g., European, American, etc, with large variations in building density, architecture, and overall characteristics e.g., red shingles, flat roofs, etc. For these reasons, we have chosen this benchmark dataset because it is ideal for assessing the *generalization capacity of the network*.

B. Network Architecture

A vast number of networks has been proposed for image classification and semantic labeling. State-of-the-art performance is generally achieved with deep networks however these are difficult to train due to vanishing or exploding gradients. Many networks [7], [10], [23], [13] have shown skip connections play an important role in having good gradient propagation through the network. In our work, as part of the network design process, we first identified the requirements for the particular task at hand i.e. semantic segmentation of buildings from remote sensor images, and then decisions were made to address these:

- **Requirement 1:** An important aspect of semantic segmentation of buildings is to have high localization accuracy and take into account as much context information as possible. This is necessary in order to address the wide variability in buildings typically relating to their function e.g., shape, size, color and/or region they appear in e.g., density in urban/rural, etc.
  Decision: To that end, the U-Net architecture [23] takes into account spatial information and combines it with contextual information via the direct downsampling-upsampling links.
- **Requirement 2:** In order to be able to process large chunks of data at a time it is imperative that the network contains as few parameters as possible.
  Decision: Dense blocks connect every layer to every other layer in a feed-forward fashion. Along with good gradient propagation they also encourage feature reuse and reduce the number of parameters substantially as there is no need to relearn the redundant feature maps. At the end of every Dense block all the extracted features accumulate creating a very diverse set of features. As a result of this feature redundancy there is a substantial reduction in the network parameters leading to faster training times. This allows the processing of larger patch (and batch) sizes (which also addresses Requirement 1) therefore allowing additional contextual information during each feed-forward pass.
- **Requirement 3:** The contribution of the feature maps at each layer to the output must depend on their importance.

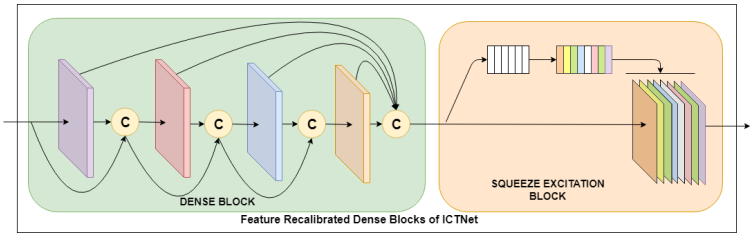


Figure 2: Proposed feature recalibrated Dense block with 4 convolutional layers and a growth rate $\kappa = 12$ used by the ICT-Net. c stands for concatenation.

Decision: Using the Squeeze-and-Excitation (SE) blocks the dynamic channel-wise feature re-weighting mechanism provides a way to upweigh important feature maps and downweigh the rest. In [8] authors show adaptive recalibration of channel-wise feature responses by explicitly modelling inter-dependencies between channels using squeeze and excitation block on existing architectures [7], [27], [29] results in improved performance.

The proposed network architecture is distinct and combines the strengths of the U-Net architecture, Dense blocks, and Squeeze-and-Excitation (SE) blocks. This results is improved prediction accuracy and it has been shown to outperform other state-of-the-art network architectures such as the ones proposed in [9] which have a much higher number of learning parameters on the INRIA benchmark dataset. Figure 2 shows a diagram of the proposed feature recalibrated Dense block with 4 convolutional layers and a growth rate $\kappa = 12$ used by the ICT-Net. The proposed network has 11 feature recalibrated dense blocks with [4,5,7,10,12,15,12,10,7,5,4] number of convolutional layers in each dense block.

Perhaps the closest architecture to the one proposed was discussed in [13] which uses 103 convolutional layers. If SE blocks are introduced at the output of every layer this will cause a vast increase in the number of parameters which will hinder the training. In contrast, in our work we have chosen to include an SE block only at the end of every Dense block in order to re-calibrate the accumulated feature-maps of all preceding layers. Thus, the variations in the information learned at each layer - in the form of the features maps - are weighted by the SE block according to their importance as determined by the loss function.

Discussion: To verify the validity of the above design decisions we performed a comparative study involving a number of state-of-the-art architectures and blocks. Following the same training procedure for all architectures reported, and without any data augmentation the ICT-Net was compared with U-Net [23] and Tiramisu-103 [13]. The results on the validation dataset are shown in Table I where it is evident that the proposed architecture outperforms both U-Net and Tiramisu-103.

Paper	Method	Overall IoU (%)	Overall Accuracy (%)
[23]	UNet	70.86	95.51
[13]	Tiramisu-103	73.91	95.71
Ours	ICT-Net	**75.5**	**96.05**

Table I: Performance evaluation of SOTA architectures (U-Net [23] and Tiramisu-103 [13]) on the validation dataset

C. Training and Validation

The network is trained on 155 tiles each with resolution 5000×5000 from the available training data with their corresponding ground truth. The training is performed for 100 epochs on a single nVidiaGTX 1080Ti. We used Tensorflow API for the development and training/testing of the network. Due to the large size of the dataset it requires approximately 6 hours to complete 1 epoch of training. Every epoch was divided into 31 sub-epochs each consisting of 5 tiles (1 from each city). Limited by GPU memory we had to choose a small batch size of 4 to have a comparatively larger patch size of 256×256 as we observed context is very important for semantic labeling of buildings.

Implementation details: The network was trained using cross-entropy loss with RMSProp Optimizer with an initial learning rate of 0.001 and decay of 0.995 for the first 50 epochs. After the 50^{th} epoch the learning rate was reduced to 0.0001 and trained for another 50 epochs. Instead of using dropout as a regularization technique we applied a large number of data augmentations in order to restrict the network from overfitting to the training dataset.

Data input: Our network takes in patches of 256×256 out of the entire tile with 50% overlap. The patches are selected sequentially for every odd epoch and the same number of patches is selected randomly for every even epoch during the training. We use the alternating patch generation strategy to restrict the network from overfitting while still having the opportunity to learn all the features from every tile. At testing the input patch size in increased to 768×768 (the maximum that could fit in the GPU memory) so that we are able to increase the context for large building in every patch. During testing, the patches are selected using 50% overlap similar to what is done during training.

Network output: The output produced by the network is a 1-channel gray-scale image of the same size as the input image where each pixel has a probability score of being a building in the range $[0, 1]$. We convert the probability map into a binary mask by thresholding. We conducted an empirical study on the validation dataset and have chosen $\tau = 0.4$ as the optimal threshold value for converting the gray-scale image to a binary map as shown in Figure 3. The output patches are then assembled into tiles of size 5000×5000 by weighted average and overlapping areas near the edges are down-weighted. During the testing, the standard test time augmentations are applied to each tile and they are merged back using an average of the

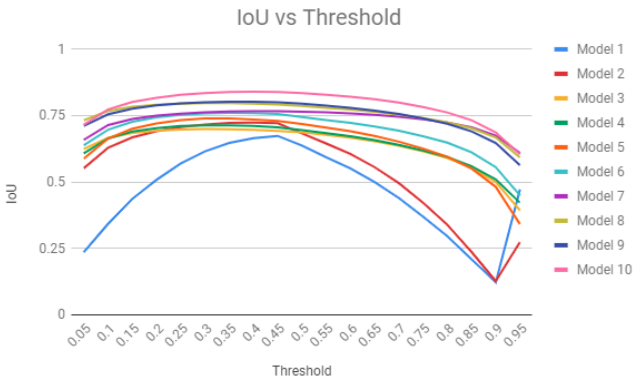


Figure 3: Empirical study to determine the optimal thresholding value for converting the grayscale classification map produced by the network to a binary map. The models shown correspond to the same network ICT-Net at different training snapshots for which the classification accuracy (i.e. IoU in the graph) was calculated **after** the thresholding at every 0.05 intervals as shown. The optimal threshold value is $\tau = 0.4$.

probability scores.

Data augmentations: Based on the validation results we used the pretrained weights and trained our network with the following data augmentations with a probability of 70% to be applied to every patch: random rotations in the range $[0°, 360°]$ using reflection padding, random flip, random selection of a patch in the range of $[0.75, 1.25]$ of the image patch size and re-size it to original patch size of 256. Data augmentations significantly improved the performance of the network in terms of accuracy.

D. Evaluation - Test dataset

The INRIA dataset uses two main performance measures: Intersection over Union (Jaccard index) and Accuracy. Intersection over Union (IoU) is defined as the number of pixels labeled as buildings in both the prediction and the reference, divided by the number of pixels labeled as buildings in the prediction or the reference. Accuracy is defined as the percentage of correctly classified pixels.

The measures are calculated by the organizers of the competition and involve the classification of 5 cities for which no images have been used for training and validation, and for which no ground truth is available to the participants. As of writing this manuscript the proposed architecture is ranked as the top performing in terms of both IoU (80.32%) and accuracy (97.14%) on the competition's leaderboard [1]. Figure 4 shows an example of a result for a small area of an image from the test dataset (top left). The probability image produced by the network is shown as a heat map (bottom right) overlaid on top of the RGB image (bottom left). The binary map resulting after the thresholding is shown in the top right image.

[1] https://project.inria.fr/aerialimagelabeling/leaderboard/

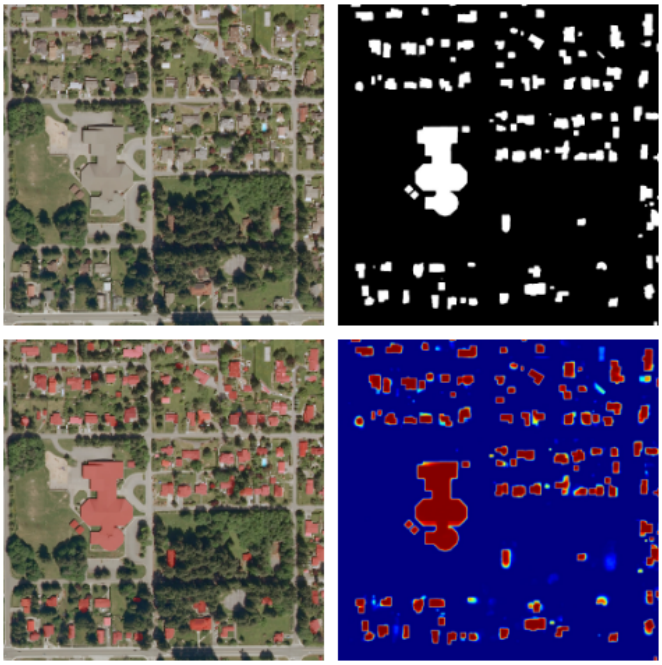
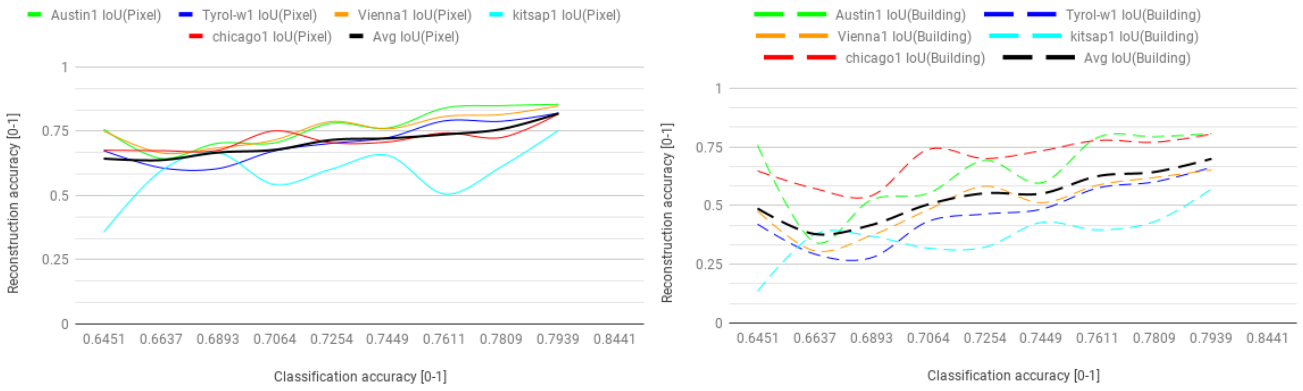


Figure 4: Building Classification of Bellingham city, tile 17. Result for an image from the test dataset. Top left: A closeup of a small area of an input image. Bottom left: The probability image overlaid on top of the RGB image. Bottom right: The probability image shown as a heat map. Top right: The binary map resulting after the thresholding of the probability image.

As previously mentioned, the proposed network is currently ranked as the top performing network with the second best having more than 1.5% difference in terms of the IoU. The authors in [9] provide details of the next 4 top performing techniques on the INRIA aerial image labeling benchmark dataset. All 4 methods are Convolutional Neural Networks(CNNs), among which 3 of them are based on U-Net architecture. Table II shows a quantitative comparison between the proposed network ICT-Net and these other techniques on the test dataset as reported by the competition organizers.

Paper	Method	Overall IoU	Overall Accuracy
[9]	Raisa	69.57	95.30
[9]	ONERA	71.02	95.63
[9]	NUS	72.45	95.90
[9]	AMLL	72.55	95.91
[12]	N/A	78.80	96.91
Ours	ICT-Net	**80.32**	**97.14**

Table II: Performance evaluation of the top 5 performing networks on the test dataset. ICT-Net outperforms all others with more than 1.5% difference in terms of the IoU.



(a) (b)

Figure 5: (a) Reconstruction vs Classification accuracy. The classification accuracy ranges from [0.6451, 0.8441] as calculated on the validation test. The different classification accuracies correspond to the same architecture (ICT-Net) but at different snapshots during training. The binary classification map produced at each snapshot: (i) is refined using conditional random fields (CRF), (ii) building boundaries are extracted, (iii) building boundaries are refined using the Douglas-Pecker algorithm, (iv) converted back to a binary classification map, and (v) compared with the ground truth. Two metrics are used: per-pixel IoU and per-building IoU (with a threshold of 75% overlap for true positive). There is an average decrease of 4.43% $\pm$ 1.65% (confidence level 95%) in per-pixel IoU (a) of the reconstruction accuracy; and an average decrease of 21.7%$\pm$ 4.21% (confidence level 95%) in per-building IoU (b) of the reconstruction accuracy. The reported averages are calculated across the accuracy levels.

V. Comparative Quantitative Analysis of Reconstruction Accuracies

As previously stated the objectives and contributions of our work are two-fold. In Section IV we proposed a novel, top ranking architecture for classifying buildings from remote sensor imagery. This binary classification map is typically used as a first step to the reconstruction process since it allows the application of specialized reconstruction algorithms according to the classified type of the pixels. In this section we focus on the equally important aspect of the relation between the classification accuracy and the accuracy of the reconstruction. Since it is extremely difficult to acquire building blueprints or CAD models for such large areas, and no 3D/depth information is available as part of the benchmark dataset we posit that the building boundaries extracted from the classification binary map can serve as a *proxy* to the quality of the reconstruction since the boundaries are typically extruded in order to create the 3D models corresponding to the buildings. More specifically, the procedure for quantitatively evaluating the accuracy of the reconstruction is as follows:

- Building boundaries B_g are extracted from the ground truth provided as part of the training dataset.
- The RGB image corresponding to the ground truth above is used as input to the ICT-Net. The binary classification map C_b resulting from feeding forward the RGB image classifies pixels into buildings and non-buildings.
- The binary classification map C_b is refined $C_b^{refined}$ using a CRF-based technique where an energy function is minimized via graph-cut optimization for finding an optimal labeling f_p for every pixel p such that $f_p \rightarrow l$, where l is the new label. The data term of the energy function of a pixel p with label l_{p_i} is defined as,

$$E_d = \begin{cases} 10, & \text{if } f(p_i) \neq l_{p_i} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The smoothness term of the energy function of two neighbouring pixels p_1 and p_2 with labels l_{p_1} and l_{p_2} respectively is defined as,

$$E_s = \begin{cases} 20, & \text{if } l_{p_1} == l_{p_2} \text{and } f(p_1) \neq f(p_2) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The values of 10 and 20 in the equations were selected such that smoothness is favored over the observed data.
- Building boundaries B_b are extracted from the refined classification map $C_b^{refined}$. A simplification process i.e. Douglas-Pecker approximation with a tolerance of $\tau = 0.5$, is applied to the boundaries. This simplification process is a step applied to the building boundaries prior to extruding the 3D model if 3D/depth information is available [22], [19], [20].
- The simplified boundaries B_b^{approx} are finally converted back to a binary classification map and quantitatively compared to the ground truth B_g. This comparison involves IoU metrics on (i) a per-pixel and (ii) a per-building bases. In the case of the per-building IoU metric, a true positive is considered only if a building has at least 75% of its pixels overlap the pixels of the same building in the ground truth.

The procedure described above is followed for all input images with no changes to the values and thresholds used; the only varying condition is the classification accuracy. In our experiments, the input images are processed by the proposed ICT-Net at different training snapshots having

different classification accuracies. Thus, multiple binary classification maps were produced each with a different classification accuracy.

Table III shows the quantitative results of the comparison. A total of 5 cities were processed using the aforementioned procedure. Figures 5a and 5b show the relation between the reconstruction accuracy with respect to the classification accuracy. We have used increasing classification accuracies based on the same architecture (ICT-Net) at different snapshots during the training. Using the binary classification maps we have followed the aforementioned procedure which is typical to the reconstruction process. Two metrics have been used to assess the reconstruction accuracy, namely per-pixel IoU and per-buildng IoU (with 75% threshold for being considered a true positive). As expected, the graph shows a strong correlation between the classification accuracy and the reconstruction accuracy. However the reconstruction accuracy is consistently lower than the classification accuracy by an average of 4.43% $\pm$ 1.65% (confidence level 95%) on the per-pixel IoU and an average of 21.7%$\pm$ 4.21% (confidence level 95%) on the per-building IoU. This discrepancy can be attributed to the fact that the ground truth images used for training the network may contain errors and are in most cases manually created which results in much higher classification accuracy than the reconstruction accuracy. Moreover, the high discrepancy on the per-building IoU can be attributed to the fact that a threshold must be used i.e. 75%, when calculating the true positives.

The results of this analysis clearly indicate *that high classification accuracy does not translate into high reconstruction accuracy*. More importantly though, the results of the analysis clearly indicate that the reconstruction accuracy must be taken into account as part of the loss function along with the classification accuracy during the training of the network.

Figure 6 shows an example of the downtown Montreal. The building classification is generated with the proposed ICT-NET network and refined as explained above. In this example, LiDAR information was available which after resampling at the same resolution as the orthorectified image was used to extrude the 3D buildings from the extracted boundaries. The result shown is fully automated and no post-processing was performed. It should be noted that no images of the city of Montreal have been used in the training. We have manually evaluated the result by counting the number of buildings and confirming that all of them have been classified correctly by the network and therefore reconstructed. The accuracy of the classification is also evident from the fact that there is no "bleeding" between the buildings and any other urban features e.g., roads, trees, cars, etc in the final result.

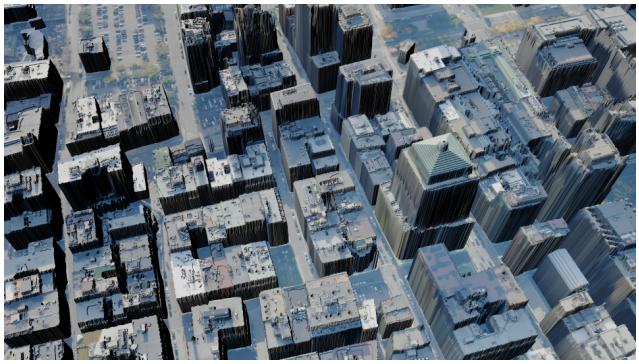


Figure 6: A fully automated result without any post-processing. Downtown Montreal for which no training images were used and no ground truth is available. Classification by ICT-Net and reconstruction by extruding the extracted boundaries of the buildings using the LiDAR pointcloud corresponding to the same area. The elevation of all non-building points is set to zero. All buildings have been manually verified that they are correctly classified. The accuracy of the classification can also be visually verified since there is no "bleeding" between the buildings and any other urban features e.g., roads, trees, cars, etc. The orthophoto RGB image is courtesy of Defence Research and Development Canada and Thales Canada.

VI. Conclusion

We have presented a novel network which combines the strengths of state-of-the-art techniques like Dense blocks in fully convolutional networks and feature recalibration using SE blocks. We have identified the requirements for the particular task and based our decisions on the actual characteristics and observations. We have shown that the proposed architecture outperforms other state-of-the-art including ensemble techniques.

Furthermore, we investigated the relation between the classification accuracy and the reconstruction accuracy. Due to the extreme difficulty of acquiring blueprints for such large areas and the unavailability of 3D information we have used the building boundaries as a proxy to the reconstruction accuracy. The proposed ICT-Net at different training snapshots was used to generate binary maps of different classification accuracies which were then used for extracting the boundaries. We presented a comparative quantitative analysis which shows a strong correlation between the two but also a consistent and considerable decrease of the reconstruction accuracy when compared to the classification accuracy.

With respect to the future work, we plan on extending this work to (i) the classification of multiple urban feature types and not just buildings, (ii) conduct a comparative quantitative analysis using ground-truth 3D information acquired by LiDAR and manually processed, and (iii) design a loss function which takes into account the reconstruction accuracy in addition to the classification accuracy.

Acknowledgement

This research is based upon work supported by the Natural Sciences and Engineering Research Council of Canada Grants DG-N01670 (Discovery Grant) and DND-N01885 (Collaborative Research and Development with the Department of National Defence Grant). The authors would like to thank Jonathan Fournier from Valcartier

Classification Accuracy	Austin1 IoU		Tyrol-W1 IoU		Vienna1 IoU		Kitsap1 IoU		Chicago1 IoU		Average IoU	
	per-pix.	per-bldg	per-pix.	per-bldg	per-pix.	per-bldg	per-pix.	per-bldg	per-pix.	per-bldg	per-pix.	per-bldg
0.6451	0.7038	0.5004	0.4683	0.1887	0.7291	0.3880	0.1063	0.02384	0.6445	0.4952	0.5304	0.3192
0.6637	0.7583	0.7583	0.6749	0.4213	0.7514	0.4776	0.3575	0.1354	0.6765	0.6481	0.6437	0.4881
0.6893	0.6443	0.3451	0.6084	0.2944	0.6660	0.3080	0.5949	0.3770	0.6747	0.5734	0.6377	0.3796
0.7064	0.7034	0.5240	0.6046	0.2780	0.6855	0.3728	0.6671	0.3704	0.6760	0.5420	0.6673	0.41745
0.7254	0.7049	0.5520	0.6735	0.4325	0.7160	0.4820	0.5432	0.3194	0.7516	0.7386	0.6778	0.5049
0.7449	0.7812	0.6926	0.7032	0.4643	0.7881	0.5829	0.6026	0.3230	0.7056	0.7011	0.7162	0.5528
0.7611	0.7630	0.5976	0.7256	0.4850	0.7597	0.5128	0.6561	0.4286	0.7088	0.7336	0.7226	0.5515
0.7809	0.8408	0.7914	0.7907	0.5756	0.8078	0.5879	0.5059	0.3973	0.7436	0.7782	0.7378	0.6261
0.7939	0.8498	0.7936	0.7891	0.6016	0.8153	0.6202	0.6131	0.4328	0.7259	0.7703	0.7586	0.6437
0.8441	0.8549	0.8073	0.8212	0.6634	0.8490	0.6519	0.7541	0.5714	0.8179	0.8050	0.8194	0.6998

Table III: The ICT-Net at different training snapshots having different classification accuracy vs the reconstruction accuracy measured using two metrics: per-pixel IoU, and per-building IoU (with a threshold of 75% overlap for true positives)

DRDC, and Hermann Brassard, Sylvain Pronovost, Dave Lajoie, and Xian Wang from Presagis Inc Canada, for their invaluable discussions and assistance in processing maps for Montreal. The authors would also like to thank Defence Research and Development Canada and Thales Canada for providing orthophoto RGB images used for testing, and the reviewers for their comments and suggestions.

REFERENCES

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561, 2015. 2

[2] S. R. Bulò, L. Porzi, and P. Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. *CoRR*, abs/1712.02616, 2017. 2

[3] T. Forbes and C. Poullis. Deep autoencoders with aggregated residual transformations for urban reconstruction from remote sensing data. *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 23–30, 2018. 2

[4] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. G. Rodríguez. A review on deep learning techniques applied to semantic segmentation. *CoRR*, abs/1704.06857, 2017. 2

[5] T. L. Haithcoat, W. Song, and J. D. Hipple. Building footprint extraction and 3-d reconstruction from lidar data. In *Remote Sensing and Data Fusion over Urban Areas, IEEE/ISPRS Joint Workshop*, pages 74–78. IEEE, 2001. 2

[6] R. Hamaguchi and S. Hikosaka. Building detection from satellite imagery using ensemble of size-specific detectors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 2

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 2, 3, 4

[8] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017. 2, 4

[9] B. Huang, K. Lu, N. Audebert, A. Khalel, Y. Tarabalka, J. Malof, A. Boulch, B. Le Saux, L. Collins, K. Bradbury, et al. Large-scale semantic classification: outcome of the first year of inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium–IGARSS 2018*, 2018. 4, 5

[10] G. Huang, Z. Liu, and K. Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. 2, 3

[11] V. I. Iglovikov, S. S. Seferbekov, A. V. Buslaev, and A. Shvets. Ternausnetv2: Fully convolutional network for instance segmentation. *CoRR*, abs/1806.00844, 2018. 2

[12] Inria Aerial Image Labeling Benchmark. Inria Aerial Image Labeling Benchmark LeaderBoard. `https://bit.ly/2GC88nr` (last accessed 18th Feb. 2019). 5

[14] A. Khalel and M. El-Saban. Automatic pixelwise object

[13] S. Jégou, M. Drozdzal, D. Vázquez, A. Romero, and Y. Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. *CoRR*, abs/1611.09326, 2016. 2, 3, 4

labeling for aerial imagery using stacked u-nets. *CoRR*, abs/1803.04953, 2018. 2

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc. 2

[16] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014. 2

[17] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017. 2, 3

[18] V. Mnih. *Machine learning for aerial image labeling*. University of Toronto (Canada), 2013. 2

[19] C. Poullis. A framework for automatic modeling from pointcloud data. *IEEE transactions on pattern analysis and machine intelligence*, page 1, 2013. 6

[20] C. Poullis. Large-scale urban reconstruction with tensor clustering and global boundary refinement. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2019. 6

[21] C. Poullis and S. You. Automatic reconstruction of cities from remote sensor data. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2775–2782, June. 2

[22] C. Poullis and S. You. Automatic creation of massive virtual cities. In *2009 IEEE Virtual Reality Conference*, pages 199–202. IEEE, 2009. 6

[23] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 2, 3, 4

[24] A. G. Roy, N. Navab, and C. Wachinger. Concurrent spatial and channel squeeze & excitation in fully convolutional networks. *CoRR*, abs/1803.02579, 2018. 2

[25] A. K. Shackelford, C. H. Davis, and X. Wang. Automated 2-d building footprint extraction from high-resolution satellite multispectral imagery. In *Geoscience and Remote Sensing Symposium, 2004. IGARSS'04. Proceedings. 2004 IEEE International*, volume 3, pages 1996–1999. IEEE, 2004. 2

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 2

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. 2, 4

[28] O. Wang, S. K. Lodha, and D. P. Helmbold. A bayesian approach to building footprint extraction from aerial lidar data. In *3DPVT, Third International Symposium on*, pages 192–199. IEEE, 2006. 2

[29] S. Xie, R. B. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431, 2016. 4

[30] S. Zagoruyko and N. Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016. 2