Building Detection from Satellite Imagery Using a Composite Loss Function

Sergey Golovanov
Neuromation OU, Tallinn, Estonia
sergey.golovanov@neuromation.io

Rauf Kurbanov
Neuromation OU, Tallinn, Estonia
JetBrains Research, St. Petersburg, Russia
rauf.kurbanov@neuromation.io

Aleksey Artamonov
Neuromation OU, Tallinn, Estonia
aleksey.artamonov@neuromation.io

Alex Davydow
Neuromation OU, Tallinn, Estonia
alex.davydov@neuromation.io

Sergey Nikolenko
Neuromation OU, Tallinn, Estonia
snikolenko@neuromation.io

Abstract

In this paper, we present a LinkNet-based architecture with SE-ResNeXt-50 encoder and a novel training strategy that strongly relies on image preprocessing and incorporating distorted network outputs. The architecture combines a pre-trained convolutional encoder and a symmetric expanding path that enables precise localization. We show that such a network can be trained on plain RGB images with a composite loss function and achieves competitive results on the DeepGlobe challenge on building extraction from satellite images.

1. Introduction

Satellite imagery is an important class of imaging data that has remained largely underutilized by modern computer vision researchers. Government agencies such as NASA or ESA and companies such as DigitalGlobe [15] have access to terabytes of satellite images.

However, satellite imagery has not yet become the target of much research in computer vision and deep learning. There are few large-scale publicly available datasets, and data labeling is always a bottleneck for segmentation tasks. The DeepGlobe Challenge at CVPR 2018 is designed to bridge this gap, bringing high-quality and at the same time labeled satellite imagery.

In this work, we have made an attempt to get accurate instance level prediction for the building detection task while addressing challenging examples to the loss function.

Name	Band	Name	Band
Coastal	400 - 450 nm	Red	630 - 690 nm
Blue	450 - 510 nm	Red Edge	705 - 745 nm
Green	510 - 580 nm	Near-IR1	770 - 895 nm
Yellow	585 - 625 nm	Near-IR2	860 - 1040 nm

Table 1. Multispectral channels of the WorldView-3.

2. Dataset and Evaluation Metric

Satellite images for the Building Extraction Challenge were selected from the *SpaceNet* dataset [8]. Images have 30cm per pixel resolution and have been gathered by the WorldView-3 satellite [1]. The organizers chose several cities such as Las Vegas, Paris, Shanghai and Khartoum.

Apart from traditional RGB images, *SpaceNet* also contains 8 additional spectral channels that are listed in Table 1. The training set contains 10593 11-bit 650×650 images in TIFF format, and each of the two test sets contains 3526 images. This sets include all of the above-mentioned cities. Every image in the training set is accompanied by building labels provided as polygons on the image.

For quality estimation, the Building Extraction Challenge uses the F1 score by a proposed buildings which have IoU (Intersection over Union) with ground truth greater than 0.5.

3. Methods

3.1. Model architecture

We used a modified version of LinkNet [5] as the basic underlying model for the building segmentation problem.

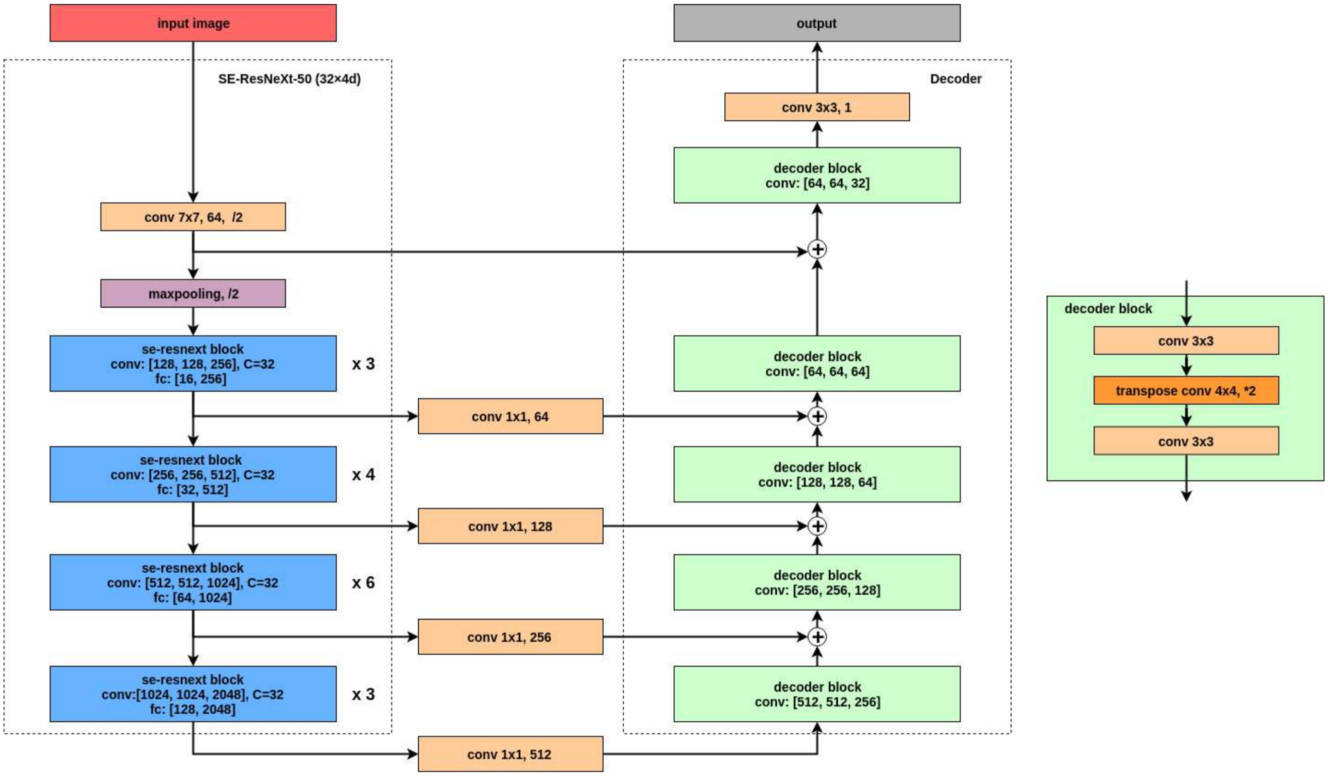

Figure 1. Modification of LinkNet with SE-ResNeXt-50.

The network architecture is shown on Figure 1: it consists of an encoder that extract multi-scale features and a decoder with skip-connections from the encoder for a more accurate localization of object boundaries.

In our solution, we propose to use the *SE-ResNeXt-50* architecture [12] pre-trained on the *ImageNet* 1000 classes subset [9] as the encoder.

This architecture achieves high accuracy in classification and does not require much memory, which allows to keep the batch size high enough for batch normalization to be applicable.

The decoder consists of five blocks, each of which is a 3×3 convolution, a transposed convolution that increases the size of the feature map by a factor of two, and a more 3×3 convolution. After each of these elements in the block, we applied batch normalization [14] and ELU non-linearity [7]. The input of each block receives the sum of the result of the previous block and the corresponding feature map from the encoder compressed by a 1×1 convolution.

The final output is obtained by applying the last 3×3 convolution. We used He normal initialization [11] for the weights in the network.

3.2. Loss function

As the loss function we used a weighted sum of the binary cross-entropy BCE combined with the Lovász hinge loss LH [3] and MSE for watershed energy:

$$Loss = \alpha \cdot \mathrm{BCE} + \beta \cdot \mathrm{LH} + \gamma \cdot MSE. \qquad (1)$$

Adding LH into BCE loss allows to increase the connectivity of pixels within the object area, and also accelerates the convergence of the model. However, for large values of β, the separation of closely located objects degrades.

Based on local validation, we chose the optimal values of loss parameters: $\alpha = 0.8, \beta = 0.2, \gamma = 10$.

3.3. Separating the buildings

One of the most important problems specific for the building detection challenge is *separation*: basic segmentation models find it hard to separate different buildings that are close to each other. To cope with this problem, we added weights for the pixels in the binary cross-entropy loss, which has led to significant improvements in separation. For the weight $\omega(x)$ of a pixel x, we build on the weight map proposed for U-Net [16]:

$$\omega_{\mathrm{UNet}}(x) = \omega_c(x) + \omega_0 e^{-\frac{1}{2\sigma^2}(d_1(x)+d_2(x))^2}. \qquad (2)$$

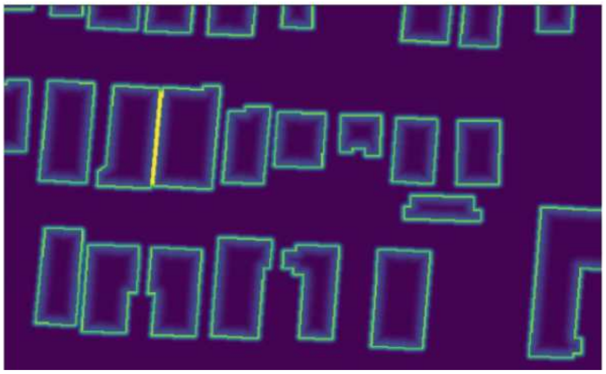

Figure 2. A heatmap of weights $\omega(x)$ in an example of closely located buildings.

City	Las Vegas	Paris	Shanghai	Khartoum	Total
Without watershed	**0.8816**	0.7535	0.6081	0.5564	0.7572
With watershed	0.8765	0.7544	0.6359	0.5707	0.7607
With watershed + TTA	0.8805	**0.7687**	**0.6440**	**0.5838**	**0.7679**

Table 2. Multispectral channels of the WorldView-3.

But unlike [16], we also increased weights for the areas inside the masks, which makes the boundaries more rectangular (a desired effect since we are looking for buildings). In addition to these weights, we increased weights of the areas located on narrow strips of land separating individual buildings. Problematic separating walls were found with the following simple algorithm: (i) given a mask M, apply morphological closing with a disk of radius r_d pixels; denote the result as $\bar{M}$; (ii) distort $\bar{M}$ back by r_d pixels, getting $\bar{\bar{M}}$ similar to M but with narrow borders eliminated; (iii) subtract $\bar{\bar{M}}$ from $\bar{M}$, getting the map of narrow separating borders B.

After this transformation, we assign weight map to the BCE loss as

$$\omega(x) = \omega_c(x) + \omega_0 e^{-\frac{1}{2\sigma^2}(d_1(x)+d_2(x))^2} + w_0 \frac{d_3}{2r_d}, \qquad (3)$$

where ω_c is the weight of the class, d_1 is the distance to the boundary of the nearest building, d_2 is the distance to the boundary of the second nearest building, d_3 is the distance from the narrow separating borders to the border, ω_0, σ and r_d are hyperparameters. In our final model, we used $\omega_c = 1$, $\omega_0 = 10$, $\sigma = 5$, $r_d = 5$. This change to the basic formula from [16] increased the F1 score on local validation; we illustrate this idea with a sample heatmap of the weights presented at Figure 2.

3.4. Preprocessing and training

Satellite images provided in the competition already have a sufficiently fine resolution of 30cm per pixel, but some of them had a very dense buildings arrangement, which made it difficult to separate buildings. We already described an approach to this problem with loss function weights above, but besides that we also increased the resolution of all images by a factor of two, which allowed us to build more detailed masks. This transformation also allowed the encoder to capture small details better, since the architecture of SE-ResNeXt-50 decreases resolution very rapidly with 2-step convolution and max-pooling at the very beginning.

The network was trained on 256×256 crops. We normalized the input images, and applied standard augmentations: rotations by multiples of 90 degrees, flips, random scaling, random shift, changes in brightness and contrast. For input RGB images we used batch size 32. Images for every batch were sampled with probabilities proportional to the weights. The weight for each image $\tau(I)$ was initialized in units and updated after every epoch. For an image I we updated the weight $\tau(I)$ smoothly:

$$\tau_{k+1}(I) = (1-\alpha) \cdot \tau_k(I) + \alpha \cdot L_{\mathrm{Crop}(I)}, \qquad (4)$$

where ω_k is the previous weight value, $L_{\mathrm{Crop}(I)}$ is the value of the loss function on the current crop of image I, and α is a hyperparameter; we used $\alpha = 0.2$. This sampling method has allowed us to focus on images that presented the greatest challenges for segmentation.

We trained the network for 120 epochs using SGD [17] with Nesterov momentum equal to 0.95. The initial learning rate was set to 0.01 and reduced by a factor of 0.1 whenever the value of the loss function on the validation set did not drop significantly for 5 epochs.

As the output of the network we predict building masks accompanied with watershed energy and normalized L^2 distance to the edge of the corresponding building. To separate falsely united building predictions, we used this energy with a threshold of 0.25 to isolate seeds for the watershed algorithm [2]. As the “landscape” for the watershed algorithm, we used the negative distance transform of the binary mask. Binarization of masks was performed with a threshold of 0.5. All found polygons with an area of less than 110 pixels were discarded.

4. Results

The resulting solution, even without the use of ensembles, achieved a relatively high F1 score in such cities as Las Vegas and Paris, but the results were more modest in other cities; see Table 2 for a breakdown across the cities on the local validation set.

Due to the dense positioning and small size of buildings, accurate detection is difficult (see Figure 3 for an example). Moreover, in some of the images outlines of the buildings are hard to distinguish from the general background even manually, which also leads to errors in segmentation.

For further work, we suggest to use multispectral images to alleviate this problem; however, in our experiments

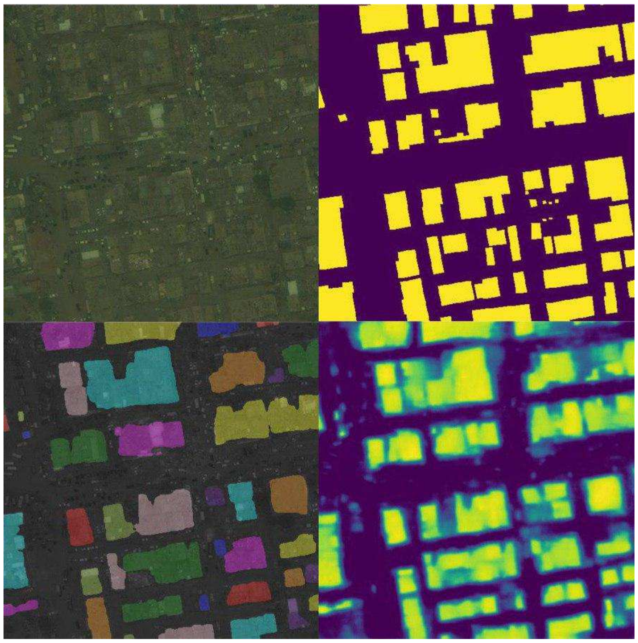

Figure 3. Example of building detection: original image (top left), ground truth mask (top right), predicted buildings (bottom left), building probability map (bottom right).

with additional channels during the competition we did not achieve any quality improvements. Other attempts to modify the architecture also have not resulted in any improvements in F1 score; in particular, we have tried to use the Atrous Spatial Pyramid Pooling module [6] and Squeeze-and-Excitation Blocks [13] in the decoder. However, some modifications have made it possible to accelerate convergence.

Therefore, we can conclude that in order to obtain an even better result, special attention in this kind of segmentation problems should be given to qualitative post-processing and methods of instance segmentation such as, for example, Mask R-CNN [10] or Discriminative Loss Function [4].

5. Conclusions

In this work, we have presented a building extraction approach from satellite imagery based on SE-ResNeXt-50 and LinkNet architecture. The characteristic features of our solution that most significantly contributed to the overall segmentation quality include:

(i) prediction energy to get better seeds for the watershed algorithm;

(ii) weight distribution for the loss function that allowed our solution to separate nearby buildings with morphological prepossessing;

(iii) the Lovász-Softmax loss function specifically designed to optimize IoU-based metrics together with the cross-entropy loss that makes it more robust.

References

[1] Worldview-3 scene. Longmont, CO: DigitalGlobe. Available: Brock University Local Access $\backslash DATA\backslash Remote_sensing\backslash WorldView\backslash$, 2015.

[2] R. Barnes, C. Lehman, and D. Mulla. Priority-flood: An optimal depression-filling and watershed-labeling algorithm for digital elevation models. *arXiv preprint arXiv:1511.04463*, 2015.

[3] M. Berman, A. Triki, and M. Blaschko. The lovasz-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. *arXiv preprint arXiv:1705.08790*, 2017.

[4] B. D. Brabandere, D. Neven, and L. V. Gool. Semantic instance segmentation with a discriminative loss function. *CoRR*, abs/1708.02551, 2017.

[5] A. Chaurasia and E. Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. *arXiv preprint arXiv:1707.03718*, 2017.

[6] L. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[7] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1502.01852*, 2015.

[8] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. *arXiv preprint arXiv:1805.06561*, 2018.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[10] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[11] K. He, Z. Xiangyu, R. Shaoqing, and S. Jian. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *arXiv preprint arXiv:1502.01852*, 2015.

[12] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 2017.

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

[14] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[15] F. Kruse, W. Baugh, and S. Perry. Validation of digitalglobe worldview-3 earth imaging satellite shortwave infrared bands for mineral mapping. *Journal of Applied Remote Sensing*, 9:9–17, 2015.

[16] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015.

[17] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, 2013.