The AAAI-17 Workshop on
Artificial Intelligence for Connected and Automated Vehicles
WS-17-03

Object Contra Context: Dual Local-Global Semantic Segmentation in Aerial Images

Alina Marcu,[1,2] Marius Leordeanu[1,2]
[1] University Politehnica of Bucharest, 313 Splaiul Indepententei, Bucharest, Romania
[2] Autonomous Systems, 22 Tudor Vladimirescu, Bucharest, Romania
alymarcu91@gmail.com, marius.leordeanu@cs.pub.ro

Abstract

The importance of visual context in object recognition has been intensively studied over the years. Along with the advent of deep convolutional neural networks (CNN), using contextual information with such systems starts to receive attention in the literature. Regardless of deep learning advances, aerial image analysis still poses many great challenges. Satellite images are often taken under poor lighting conditions and contain low resolution objects, many times occluded. For this particular task, visual context could be of great help, but there are still very few papers that consider context in aerial image understanding. Our work addresses the task of object segmentation in aerial images with a novel dual-stream deep convolutional neural network that integrates the local object appearance and global contextual information into a unified network. Our model learns to combine local object appearance and global semantic knowledge simultaneously and in a complementary way, so that together they form a powerful classifier. Experiments on the Massachusetts Buildings Dataset demonstrate the superiority of our model over state-of-the-art methods. We also introduce two new challenging datasets for the task of buildings and road segmentation. While our local-global model could also be useful in general recognition tasks, we clearly demonstrate the effectiveness of visual context in conjunction with deep nets in aerial image understanding.

Introduction

Aerial image understanding is enjoying a growing interest today, due to recent technological advancements in computer vision, along with important improvements in high performance, low-cost GPUs. The possibility of accurately recognizing different types of objects in aerial images (e.g. buildings, roads, vegetation etc.) could greatly help in many applications, such as creating and keeping up-to-date maps, improving urban planning, environment monitoring and disaster relief. Besides the practical need for accurate aerial image interpretation systems, this domain also offers specific scientific challenges to the computer vision domain. The local appearance of objects in aerial images is often degraded due to occlusions, illumination, shadows and distance, leading to poor resolution. In such cases, contextual cues provide semantic insights that improve object recognition. Our work demonstrates that visual context is vital for accurate recognition and plays a fundamental role in aerial image understanding.

Context could be understood in many forms and has been studied for quite some time in computer vision, with approaches going from reasoning about objects against the global scene to looking at more precise spatial and temporal relationships and interactions between different object categories (Torralba 2003; Oliva and Torralba 2007; Leordeanu et al. 2016; Stretcu and Leordeanu 2015; Hoiem, Efros, and Hebert 2008; Collins, Liu, and Leordeanu 2005; Rabinovich et al. 2007; Felzenszwalb et al. 2010; Desai, Ramanan, and Fowlkes 2011; Tu and Bai 2010; Yao, Fidler, and Urtasun 2012). One recent relevant example is the work of (Choi et al. 2010) that combines both spatial relations to other objects as well as global scene context. It is not yet known what is the best way to combine object relationships and global information for contextual reasoning. Deep neural networks are an interesting choice for modeling context. By reasoning in a hierarchical manner they also offer the possibility of integrating information from one level of abstraction as contextual input to the next, thus relating to approaches using autocontext (Tu and Bai 2010). Therefore, deep nets seem to offer the proper environment for designing effective architectures for using and studying visual context. Such systems, combining context with deep networks, were proposed for action classification (Gkioxari, Girshick, and Malik 2015), segmentation by modeling CRFs (Zheng et al. 2015) with recurrent networks and object detection by training contextual networks over nearby bounding box regions (Zhu et al. 2015; Gidaris and Komodakis 2015).

We propose a dual-stream approach using deep convolutional neural networks that combines the $local$ appearance of the object with $global$ information retrieved from a larger scene. Thus, the object is seen both as a separate entity from the perspective of its own appearance, but also as a part of a larger scene which acts as its complement and implicitly contains information about it. We formulate the problem as one of segmentation in the sense of finding an accurate shape for the object of interest. Our combined network is trained jointly, end-to-end. Different from (Zhu et al. 2015; Gidaris and Komodakis 2015) our proposed deep architecture is based on a dual-stream network, each pathway hav-

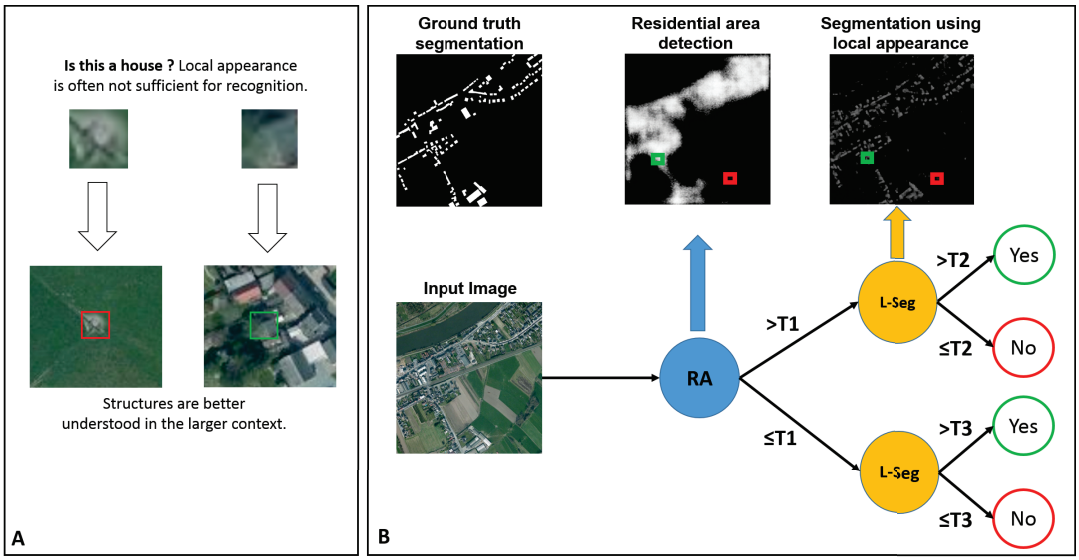

Figure 1: **A**: Local appearance is often not sufficiently informative for segmentation in low-resolution aerial images. The larger context could provide vital information even for highly localized tasks such as fine object segmentation: the exact shape of the house in the example on the right is better perceived when looking at the larger residential area, which contains other houses of similar shapes and orientations. Thus, local structure could be better interpreted in the context of the larger scene. **B**: Our initial model for residential area detection (RA) has poor localization but low false positive rate within a larger neighborhood. RA can be effectively combined, in a simple classification tree, with the local semantic segmentation model (L-Seg), which has higher localization accuracy, but relatively high false positive rate. Note how the output from RA can be used in order to filter out the houses hallucinated by the local L-Seg model.

ing its own different architecture, centered on the object but looking over different image areas.

Different from previous work, we study context in the domain of aerial imagery, where objects are relatively small and it is easy to include larger areas as input. There is also recent work (Mattyus et al. 2015) that combines satellite aerial images available online with ground truth labels from OpenStreeMap for learning, in order to enhance road maps. Authors use some weak context features based on differences in mean pixels intensities between the road area and its background, within a Markov Random Field formulation. Very few approaches in aerial image analysis use CNNs, with improved results (Mnih 2013a; Saito and Aoki 2015). Our main contribution over the prior work is to show that contextual information is important for accurate object recognition in aerial images and also provide a novel dual-stream architecture, based on deep convolutional neural networks, which learns in parallel to recognize objects from two complementary views, one from the local level of object appearance and the other from the contextual level of the scene.

Intuition and Motivation

Let us look at Figure 1 A. We present two local patches and their larger scene context. By looking at the patches only, it appears that local appearance is not sufficient for confidently recognizing the presence and the shape of a house. In fact, from the local patch alone, the example on the left seems to be more likely to belong to a house than the one on the right. When we consider the larger contextual neighborhood, the house roof is more clearly perceived in the second case, in which the larger residential area contributes in an important way to the local perception. Geometric grouping cues such as agreements of houses' orientations and similar appearances in the larger residential area increase the chance that we are indeed looking at a house and also help "seeing" its shape better. In the case on the left, the contextual alignment of the diagonals in the larger region of grass lowers the possibility that we are indeed looking at a house.

Buildings vs. Residential Regions: For better motivating our dual-stream CNNs presented in later sections, we first discuss the task of finding the shapes of buildings in an aerial image. We consider both their local appearance and the information from the larger scene containing them. We are interested to study the role of context on this task first, as buildings have various shapes and appearances and are representative for most aerial images. We employ two models based on CNNs. First, a local deep neural network, based on the state-of-the-art VGG-Net (Simonyan and Zisserman 2014), is trained to output 16×16 patches of pixel wise labels, with values between 0 and 1, in order to predict the presence or absence of a building at a given pixel. At test time the image is divided into a disjoint set of patches, on a grid, and each patch is classified independently. The end result becomes a segmentation of the entire image, with white areas belonging to building pixels. The input to the network is

a larger 64×64 patch that, in the case of smaller houses, often contains little surrounding background information. This network is thus trained to detect and segment houses (output their exact shapes) using mostly local information. We will refer to it as the local *L-Seg* network. In order to study the role of the larger context, we employ a wider (with larger filters and input) but shallower architecture based on AlexNet (Krizhevsky, Sutskever, and Hinton 2012), which takes as input a 256×256 image patch ($16\times$ larger in area than the input to L-Seg) centered at the same location. We trained this model to segment houses using global information (*G-Seg*). A different model (RA) based on the same architecture is not trained for accurate shape prediction, but only to output a single binary variable - whether the input patch belongs to a residential area or not. In our case, a large 256×256 patch is considered to be residential if it contains at least 15 houses. This is a moderate number for such patches in an image with 1 m^2 per pixel. For training, the non-residential patches were not allowed to contain any buildings.

The two models trained completely separately on two different tasks (one for accurate shape segmentation and the other for binary classification) can be effectively joined into a classifier tree (Figure 1), in which the residential area classifier acts as a filter for the local buildings shape segmenter. The tree model is built by placing the RA classifier as the root node and the L-Seg model at the leaves. Depending on how the first node classifies the patch, the leaves will classify it using different thresholds. Consequently, if a patch is classified as residential, the segmenter L-Seg will be more likely to detect buildings than otherwise. Thus, by combining RA with L-Seg we generally obtained an improvement in F-measure of about 1% over the L-Seg alone in our experiments. Some qualitative results can be seen in Figure 1. These results confirm our intuition and constitute a good motivation for the more complex models we present next.

Dual Local-Global Semantic Segmentation

We take the intuition and initial experiments from the previous section a step further and create an architecture that combines L-Seg model and G-Seg model into a single local-global deep network, termed *LG-Seg*. The two pathways process information in parallel, taking as input image patches of different sizes. Then, the last fully-connected (FC) layers of each individual network are concatenated and fed into two FC layers (the first FC layer with 4096 activation units followed by a layer with 256 units) that learn how to combine local and contextual information at the level of semantic interpretation. The final level of abstraction is the place where bottom-up and top-down reasoning about objects meet in order to resolve conflicts and reinforce agreements. Based on the experiments performed with the simple tree model we want to find whether the two pathways indeed learn categories at different levels, the local one focusing more on the exact shape of individual buildings and the other classifying larger residential areas with less focus on exact localization and delineation of buildings. We expect the single combined network trained end-to-end to be able to produce more accurate segmentations over the simple tree model. Note that the tree model usually does not

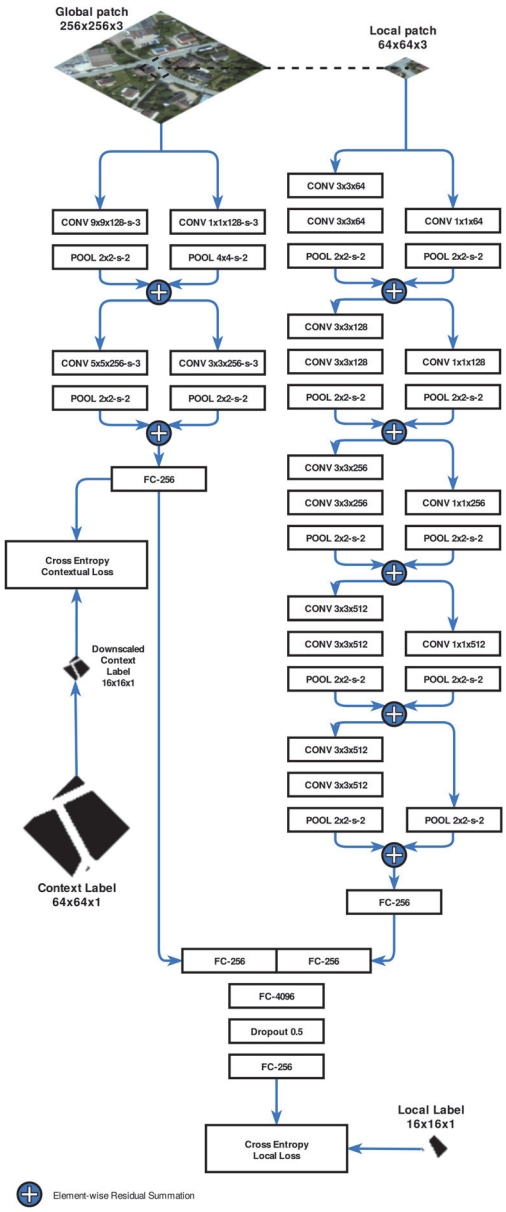

Figure 2: Our proposed local-global architecture with residuals and contextual intermediate loss. When stride s is not mentioned, we use a default value of 1.

improve the shape of the segmentation produced by the local network, but only changes the recognition confidence, using two different thresholds, over relatively large areas. In the classifier tree case, the residential area network outputs a single label per patch, while in the LG-Seg model they are jointly trained to segment objects.

Learning with Residuals and Contextual Loss: The LG-Seg model is a dual-stream combined architecture, an adapted VGG-Net and an adapted AlexNet joining into two last FC layers. Next, we present a second novel dual-stream

architecture for local-global segmentation which uses residual connections (He et al. 2015) and an intermediate contextual loss (Figure 2). The two local and global pathways take the same inputs as in the case of LG-Seg.

The residuals connections are capable of bypassing extra levels of depth and thus simultaneously combining shallow and deep pathways into a single multi-path net with filters of different sizes also acting in parallel. This model is superior to the more traditional LG-Seg and is able to handle better the difficult challenges present in aerial image recognition. While we let our initial LG-Seg architecture to learn by itself the complementarity between local and the global pathways of processing (see Figure 6), in this case we added an extra intermediate loss for the global pathway in order to enforce learning of global context in the pathway receiving larger image patches. Thus we hope to improve the training time and also the quality of segmentation in places where context matters more. We experiment with two variants of this network, one with intermediate loss and global processing (*LG-Seg-ResNet-IL*) and the other with no contextual loss (*LG-Seg-ResNet*). Our extensive experiments demonstrate that both the addition of residuals and the intermediate loss help improve both the training time and accuracy in the case of road detection, where context plays a strong role - as roads form structures that are better understood from a higher, more global level.

Problem formulation and learning: We formulate the object segmentation problem as a binary labeling task, where all pixels belonging to the object of interest are 1 and all the others are 0, in a similar way to the one proposed by Mnih et. all (Mnih and Hinton 2010).

Let $\mathbf{I}$ be the satellite aerial image and $\mathbf{M}$ the corresponding ground truth labeled map. The goal is to predict a labeled image $\hat{\mathbf{M}}$ from an input aerial image $\mathbf{I}$, that is to learn $P(M_{ij}|\mathbf{I})$ from data, for any location $p = (i, j)$ in the image. We train our network to predict a labeled image patch $W(\mathbf{M}, p, w_m)$, extracted from labeled map $\mathbf{M}$, centered at location p, of window width $w_m = 16$, from two aerial image patches $W(\mathbf{I}, p, w_l)$ and $W(\mathbf{I}, p, w_g)$, centered at the same location p, with a smaller size window width $w_l = 64$ for the local patch and a larger window width $w_g = 256$ for the global patch. We want to learn a mapping from raw pixels to pixel labels and use a loss function to minimize the total cross entropy between ground truth patches and predicted label patches.

Given a set of N examples let $\hat{\mathbf{m}}^{(n)}$ be the predicted label patch for the n^{th} training case and $\mathbf{m}^{(n)}$ the ground truth patch. Then our loss function L is:

$$L = -\sum_{n=1}^{N}\sum_{p=1}^{w_m^2}(m_p^{(n)} \log \hat{m}_p^{(n)} + (1 - m_p^{(n)}) \log(1 - \hat{m}_p^{(n)})) \quad (1)$$

The minimization of this loss is solved using stochastic gradient descent. All our models were trained end-to-end using the same hyperparameters: starting with a learning rate of 0.0001 and L_2 weight decay of 0.0005, momentum set to 0.9 and weight initialization using (Glorot and Bengio 2010). Our models were implemented, trained and tested using Caffe (Jia et al. 2014) on a GeForce GTX 970 GPU.

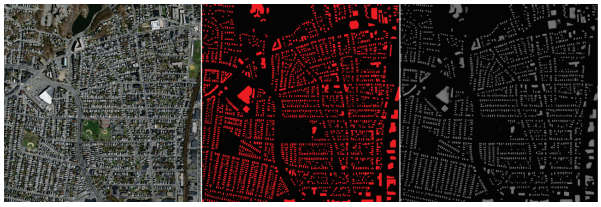

Figure 3: Example of buildings detection results on the Massachusetts Dataset. Note the high level of regularity of buildings and roads, which look very similar to each other. This permits the deep nets to almost match human performance. From left to right, in order, we present the satellite input image, the corresponding ground truth label and the result of our LG-Seg model.

Experimental Analysis

We test our models on three different datasets from US and Europe, on the tasks of detecting buildings and roads. The datasets vary greatly in terms of structure and complexity. The European Datasets are new and publicly available for download [1].

We measured performance in the same way as our competitors. We used a relaxed version of precision and recall (Wiedemann et al. 1998), usually applied in recognition from aerial imagery: a positive pixel is considered correctly labeled if it is within ρ pixels from a positively labeled pixel in the ground truth. In our experiments we set $\rho = 3$ pixels.

Detection of Massachusetts Buildings: We start by experimenting with the relatively recent Massachusetts Buildings Dataset (Mnih 2013b). It consists of 151 high quality aerial RGB images of the Boston area. They are of size 1500×1500, at resolution 1 m^2 pixel, and represent mostly urban and suburban areas, containing larger buildings, individual houses and sometimes even garages. The entire dataset covers roughly 340 km^2. It is randomly divided in a set of 137 images used for training, 4 used for the validation of the model and 10 images for testing. Our qualitative and quantitative results are presented in Figure 3 and Table 1. We densely sample our patches from a larger map on a grid. We extract a 16×16 label, a 64×64 local patch and a 256×256 global patch, centered at the same location. Therefore, from each map of size 1500×1500, we sample about 10k patches using a stride of 16 in order to restore the map to its original size. It is worth noting the significant difference between our approach and the previous state-of-the-art on the Massachusetts Dataset in the high 90% F-measure regime. While the improvement between 2013 and 2015 was less than 0.5%, we brought a significant 2% improvement, from 92.3% to 94.3%, thus reducing the error rate by 25%.

[1] https://sites.google.com/site/aerialimageunderstanding/

Table 1: Results on Massachusetts Buildings Dataset

Method	(Mnih 2013a)	(Saito and Aoki 2015)	LG-Seg	LG-Seg-ResNet-IL
F-measure	0.9211	0.9230	0.9423	**0.9430**

Table 2: Quantitative results on the European Buildings Dataset

Method	G-Seg	L-Seg	LG-Seg	LG-Seg-ResNet-IL
F-measure	0.6271	0.8266	**0.8420**	0.8387

Detection of European Buildings: We have collected the European Buildings Dataset from Western European urban and suburban areas. They contain a lot more variation than in US, in terms of general urban structure, architectural style, layout of green spaces versus residential areas and geography. We have gathered 259 RGB satellite images from Google and Bing Maps, of size 1550×1600 pixels, with a spatial resolution of about 0.8 m^2 per pixel, with locations picked randomly from different Western European countries. Covering a total area of 348.5 km^2 of urban and rural regions spread across Europe, these images also had a lot more variation in illumination as compared to those from Boston. We randomly selected 144 images for training (about 198.2 km^2), 10 for validation (21.3 km^2) and 100 for testing (129 km^2). We automatically aligned the satellite images with their corresponding ground truth label map, generated using vector metadata from the OpenStreetMap (OSM). For training we extracted about 1 million patches. We tested four models (Table 2 and Figure 4): our full LG-Seg and LG-Seg-ResNet-IL nets, and models formed by keeping only one pathway, G-Seg with the adjusted AlexNet only and L-Seg formed by the adjusted VGG-Net only. We wanted to test the capabilities of each separately and study the potential advantage of combining them into a single LG-Seg. We also wanted to study the influence of the use of intermediate loss on this task. Interestingly, the LG-Seg performed, on average slightly better than the LG-Seg-ResNet-IL with intermediate contextual loss. However, qualitatively the network with residual connections and intermediate loss was able to segment finer level of detail. This capability is more visible in the next set of experiments, on road detection. All models were trained until complete convergence of the loss, with the G-Seg model taking 34 epochs, L-Seg model 23 epochs, LG-Seg 12 epochs, and LG-Seg-ResNet-IL converging the fastest, in only 6 epochs.

Detection of European Roads: The European Road Dataset was designed for road segmentation, a much more challenging task due to limitations and variations in the data. This particular image set offers large variations in the roadmap complexity and structure, roads shapes, width and length. We have collected 200 satellite images (aprox. 276.4 km^2) for training, 20 images (27.7 km^2) for validation and 50 images (70 km^2) for testing our models. The images were collected and aligned with the ground truth OpenStreetMap

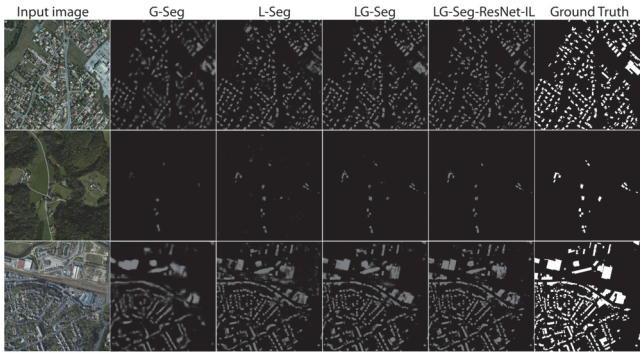

Figure 4: Qualitative performance comparison of our models on the European Buildings Dataset. Note that the models based on the dual local-global pathways perform significantly better, a fact validated by our quantitative results. LG-Seg and LG-Seg-ResNet-IL performed similarly, with the later paying slightly more attention to fine details.

Table 3: Results on European Roads Dataset

Method	LG-Seg	LG-Seg-ResNet	LG-Seg-ResNet-IL
F-measure	0.7046	0.7207	**0.7342**

in the same manner as the European Buildings Dataset. On this dataset again the models using both local and global information outperformed the others (Table 3 and Figure 5). It should be noted that on this dataset the network LG-Seg-ResNet-IL using residuals and the intermediate contextual loss significantly outperforms LG-Seg (3% improvement). Both the residual connections and the intermediate loss are important as results show in the table, where LG-Seg-ResNet is the residual network trained without the intermediate loss. This fact is very interesting as in the case of roads it is expected that context plays a more important role. Road pixels are part of much larger structures that form road maps at a higher level of interpretation, while covering large areas. This is different from the case of buildings, which are more local, occupying a limited enclosed region of space.

Our experiments on the three datasets, of different content type and structure, reveal once again the importance of data in learning. When the structures are regular and look very similar across images, such as it is the case with the Massachusetts Buildings, the performance reaches almost human level. However, as the variations in the data and frequency of occlusions increases, the performance starts degrading, dropping by almost 20% on the European roads. These results prove that aerial image understanding is far from being solved even in the context of state-of-the-art deep networks and that it remains a very challenging problem.

Discussion on Local-Global Complementarity: What are the two pathways learning in the case of LG-Seg, when no intermediate loss is used? In these experiments we highlight the individual role of each pathway in the combined output. When we first designed our combined LG-Seg net-

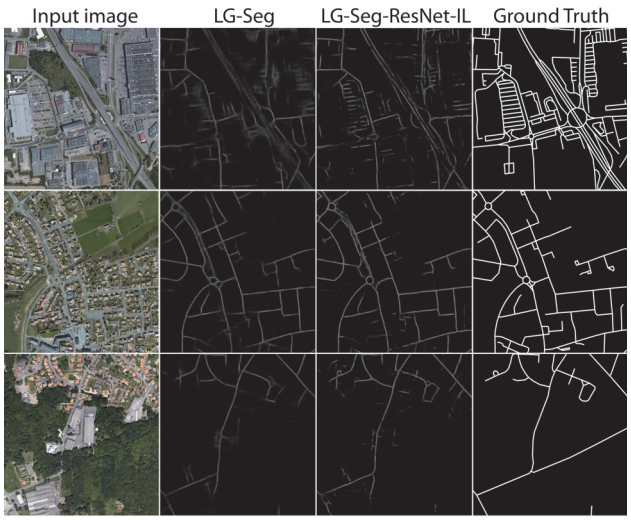

Figure 5: Example results on European Roads. Note how difficult the task is on these images, posing a real challenge even for humans.

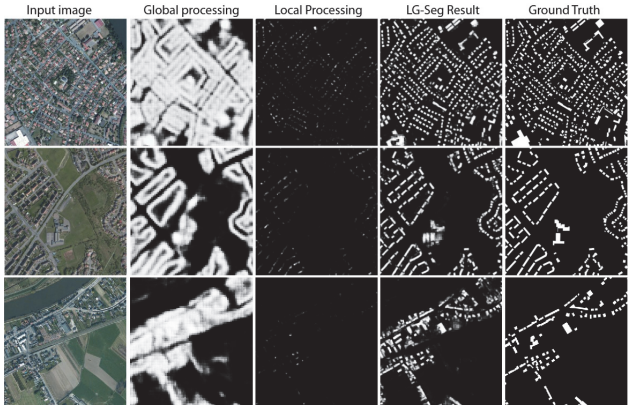

Figure 6: Proving Local-Global Complementarity: the second column shows results when only the global pathway is fed with real image signal, the other being given blank image as input. The third column shows the opposite case, when only the local pathway is given real information. The fourth column presents the output of the network running normally, with both pathways having image input. Note that the global subnet learns to detect residential areas similar to our initial classifier for such regions. The third, bottom example shows the results of our model on the same image as in Figure 1 B. Note that the residential area segmentation produced by the LG-Seg is superior to the one produced by the RA classifier, even though in the case of LG-Seg it was not asked to learn about residential areas. Also note that the local pathways focuses only on small, detailed structures. We can safely conclude that the two pathways learn to work together, in complementarity.

work we intentionally chose two different architectures: one deeper but narrower with smaller filter sizes and smaller input, better suited for more detailed local processing and one shallower with a larger input and filters encouraging contrasting learning along the two pathways. We performed a set of experiments in order to better understand the role of each processing branch. After training the full LG-Net, we performed the following: first, we ran the model over the test images by providing the local pathway with the original image input, whilst giving a blank image to the global pathway. The blank image was created by averaging each RGB channel separately. Then, we performed the opposite experiment and switched the inputs, by giving the original image to the global pathway and blocking the signal to the local one. The idea was to see how, in the fully trained model, each path contributes to the final decision. The qualitative results of this experiment, depicted in 6 are both very interesting and satisfying. When provided with information for local processing only, the network responds to small buildings with very clear structure, having crisp, very local responses over individual houses or buildings. On the other hand, when given information only to the global stream, the network produced a result that was closer to a soft residential area segmentation, in which individual buildings were undistinguishable from each other. This result is similar but of higher quality than our initial residential area detection based on the same adjusted AlexNet architecture. These qualitative experiments suggest that the intermediate loss is not always needed, as it is the case of detection of buildings - the network is able to learn by itself to process data in two complementary ways. However, when the role of context increases, as it is the case of road detection, then the intermediate loss plays an important role as the results on the European Roads show.

Conclusions

We have studied different ways of combining local appearance and global contextual information for semantic segmentation in aerial images and have proposed two novel dual local-global networks. The LG-Seg model learns completely by itself to look at objects from two complementary perspectives. When given the task of segmentation of buildings the network learns to treat each pixel, in parallel, both as part of a building and as part of a larger residential area. The second LG-Seg-ResNet-IL model, which uses residual connections and an intermediate contextual loss, is superior on the task of road segmentation, where context plays a more important role. We have performed extensive experiments, with several architectures, that study along several dimensions the role of context in aerial image understanding. Our results show that context is very important and that a dual, local-global approach is necessary in order to overcome the limitations of the local appearance alone. Consequently, we see our work as having the potential to influence future research that will shed new light on the understanding of context in vision.

Acknowledgements: This work was supported in part by UEFISCDI, under projects PCE-2012-4-0581 and PN-III-P4-ID-ERC-2016-0007.

References

Choi, M. J.; Lim, J. J.; Torralba, A.; and Willsky, A. S. 2010. Exploiting hierarchical context on a large database of object categories. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, 129–136. IEEE.

Collins, R. T.; Liu, Y.; and Leordeanu, M. 2005. Online selection of discriminative tracking features. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 27(10):1631–1643.

Desai, C.; Ramanan, D.; and Fowlkes, C. C. 2011. Discriminative models for multi-class object layout. *International journal of computer vision* 95(1):1–12.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2010. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32(9):1627–1645.

Gidaris, S., and Komodakis, N. 2015. Object detection via a multi-region and semantic segmentation-aware cnn model. In *Proceedings of the IEEE International Conference on Computer Vision*, 1134–1142.

Gkioxari, G.; Girshick, R.; and Malik, J. 2015. Contextual action recognition with r* cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 1080–1088.

Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, 249–256.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Hoiem, D.; Efros, A. A.; and Hebert, M. 2008. Putting objects in perspective. *International Journal of Computer Vision* 80(1):3–15.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Leordeanu, M.; Radu, A.; Baluja, S.; and Sukthankar, R. 2016. Labeling the features not the samples: Efficient video classification with minimal supervision. In *AAAI*.

Mattyus, G.; Wang, S.; Fidler, S.; and Urtasun, R. 2015. Enhancing road maps by parsing aerial images around the world. In *Proceedings of the IEEE International Conference on Computer Vision*, 1689–1697.

Mnih, V., and Hinton, G. E. 2010. Learning to detect roads in high-resolution aerial images. In *Computer Vision–ECCV 2010*. Springer. 210–223.

Mnih, V. 2013a. *Machine learning for aerial image labeling*. Ph.D. Dissertation, University of Toronto.

Mnih, V. 2013b. *Machine Learning for Aerial Image Labeling*. Ph.D. Dissertation, University of Toronto.

Oliva, A., and Torralba, A. 2007. The role of context in object recognition. *Trends in cognitive sciences* 11(12):520–527.

Rabinovich, A.; Vedaldi, A.; Galleguillos, C.; Wiewiora, E.; and Belongie, S. 2007. Objects in context. In *Computer vision, 2007. ICCV 2007. IEEE 11th international conference on*, 1–8. IEEE.

Saito, S., and Aoki, Y. 2015. Building and road detection from large aerial imagery. In *IS&T/SPIE Electronic Imaging*, 94050K–94050K. International Society for Optics and Photonics.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Stretcu, O., and Leordeanu, M. 2015. Multiple frames matching for object discovery in video. In *British Machine Vision Conference*.

Torralba, A. 2003. Contextual priming for object detection. *International journal of computer vision* 53(2):169–191.

Tu, Z., and Bai, X. 2010. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32(10):1744–1757.

Wiedemann, C.; Heipke, C.; Mayer, H.; and Jamet, O. 1998. Empirical evaluation of automatically extracted road axes. *Empirical Evaluation Techniques in Computer Vision* 172–187.

Yao, J.; Fidler, S.; and Urtasun, R. 2012. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 702–709. IEEE.

Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; and Torr, P. H. 2015. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 1529–1537.

Zhu, Y.; Urtasun, R.; Salakhutdinov, R.; and Fidler, S. 2015. segdeepm: Exploiting segmentation and context in deep neural networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4703–4711.