ODAS 2016

TITLE: Learning Class-Contours of Very High Spatial Resolution Aerial Images Using Convolutional Neural Networks
NAME & First name of main author: Marmanis Dimitrios
NAME & First name of co-authors: **- Kerr, Grégoire H. G.** **- Wegner, Jan D.** **- Schindler, Konrad** **- Datcu, Mihai** **- Stilla, Uwe**
Address (Department/Institute): DLR, DFD-LAX,
E-mail address of main author: dimitrios.marmanis@dlr.de
Phone number of main author: + 49 176 68 150 343
Date of birth: **- 09.08.1986 (D. Marmanis)**

Abstract

One of the most challenging problems regarding Earth Observation data of very high spatial resolution (<10cm) is related to the construction of an accurate and fully-automatic semantic annotation workflow approach. The difficulty of this task is related to the increased intra-class variability and complexity of the depicted aerial urban scenes. Lately, interesting methods based on fully Convolutional Neural Networks have shown great potential for resolving this challenging problem. However, despite the overall accuracy of these methods, a precise localization of the class-contours still remains an unresolved issue, mainly due to the uncertainty of precisely localizing the class-boundary between two neighboring land-cover entities. In this paper, we propose learning the *Class-Contours* directly from the raw images by training a CNN regression model. This allows us get sharper, better localized edges by learning strong priors of particular object classes and enforce local smoothness. This visually improves class boundaries significantly and we improve our previously best-stated outcome from *85.1%* to *88.8 %* overall accuracy.

1. Introduction

The ever increasing amount of aerial and satellite data call for automated solutions for applications like map generation and update. Very high spatial resolution imagery (*< 10cm GSD*), one the hand, contains rich information about objects that allows visually distinguishing even fine object details, but, on the other hand, introduces new challenges due to the increased intra-class variability and decreased inter-class differences of the various depicted objects in the scene.

Semantic, pixel-wise segmentation of images can be viewed as an intermediate step between raw images and final maps. A class-label is assigned to each pixel based on evidence in the image that comprises its texture, color, context etc. which are usually encoded during a feature extraction step.

Ideally, a supervised classifier trained on representative training data should be able to adequately generalize the image information and even discard image details when such information is irrelevant to the task at hand. For example, on a flat house-roof, small patches of grass should possibly be ignored, if such information is not required in the final map. An ideal annotation algorithm would learn the hierarchical organization of objects in a scene.

In contrast to more traditional, rule-based techniques, current state-of-the-art approaches treat semantic segmentation as a supervised classification problem, where object evidence is learned from annotated ground truth. Using a large set of (manually) labeled data, we can train a statistical classifier to predict conditional probabilities based on a set of generated features derived directly from the raw data. Typical choices of such features are raw-intensities, image indexes, statistical parameters, and a plethora of texture filter banks [*Leung & Malik, 2001*], [Schmid, 2001], which are fed into a discriminative classifier (e.g. boosting, bagging, random-forest, etc.). In order to reduce a vast amount of potential features to a smaller, bette to handle but still expressive feature set, efficient feature selection strategies reduce redundancy [*Viola & Jones, 2001*], [*Dollar et al., 2009*]. For a more extensive overview over the semantic annotation problem, readers are referred to [*Thoma, 2016*].

In the last years, *deep-learning (DL)* models have revolutionized the field of image classification and semantic annotation by achieving state-of-the-art results over in basically all major fields of image analysis [*Krizhevsky et al., 2012*], [*Long et al, 2015*]. Such methods benefit from automatically building a set of discriminative features directly from raw images without any manual feature design that would potentially overlook important evidence. Network architectures with a large number of stacked layers allow the system to learn a whole range of abstraction levels from very low-level evidence like blobs and gradients to high-level object patterns like car wheels or faces.

Within the family of *deep-learning* methods, the most successful from a practical viewpoint are Convolutional Neural Networks (CNN). These models use a set of locally restricted, trainable weights, which are shared across the entire image extent. CNNs can learn complex object patterns over very large image extents, using a rather small set of weights. This latter aspect in addition to the excellent statistical outcomes have transformed CNNs into one of the most powerful approaches for image classification and annotation today [*Krizhevsky et al., 2012*], [*Farabet et al., 2013*], [*Pinheiro et al., 2014*], [*Long et al, 2015*], [*Noh et al., 2015*], [*Yu & Koltun, 2015*].

A common phenomenon if applied to to pixel-accurate semantic segmentation of CNNs is that these tend to slightly blur class boundaries. There are ample reasons for this effect, pooling operations that reduce spatial resolution and filter strides bigger than one are some of them. For our long-term goal of precise mapping, accurate class-contours (edges of semantic class) are, however, very important because these determine the final map accuracy. Due to sun shadow, fuzzy object boundaries between trees and roads, for example, manually drawing class boundaries is often hard to accomplish, too.

In this work, we present an idea to model class-contours of objects. We aim at minimizing the uncertainty of class-contours, allowing for *sharper* and *more distinct* transitions between the various object categories within a scene. We argue that focusing the learner on class transitions improves object boundaries visually and also boosts quantitative results on a challenging benchmark data set bringing us one step closer to fully automated map generation.

2. Previous Work

In the last few years there has been a great advance on the topic of pixel-wise semantic segmentation within the framework of DL by adaptatiing CNNs [*Farabet et al., 2013*], [*Pinheiro et al., 2014*], [*Long et al, 2015*], [*Noh et al., 2015*], [*Yu & Koltun, 2015*].

In order to get sharper object boundaries, generating high-accuracy class-contours have been of interest to several recent works that support boundary decisions by learning high level geometric priors of a scene [*Xie et al., 2015*], [*Kokkinos, 2015*]. According to these studies, CNNs are capable of directly modeling class-contours with great accuracy and in parallel suppress non-informative edges. In [*Pinheiro et al., 2016*] the authors propose a two-stage architecture, where top-down information is used for refining the standard bottom-up CNN architecture. Within this approach, authors do not explicitly model the class-contours, however, they improve annotation accuracy over the class-edges by considering global top-down information. [*Yang et al. 2016*] propose a combinatorial architecture for explicitly detecting the class-contours by merging a fully-convolutional CNN architecture (FCN), with a simplified encoder-decoder scheme, based on the previous work of [*Long et al., 2015*] and [*Noh et al., 2015*]. The main drawback of this method in relation to the HED model [*Xie et al., 2015*], which we will use here, is higher computational complexity due to multi-stage training of the proposed system.

To the best of our knowledge, in remote sensing the only work that explicitly trains CNNs for class-contours is the work of [*Malmgren-Hansen & Nobel, 2015*], where authors use a simple, shallow CNN architecture to directly learn the contour-edges in SAR images of three object categories (target, shadow, background). We follow a similar line of thought but consider a much larger dataset, more classes, a deeper and more powerful network architecture that is capable of learning complex priors by employing previously gained knowledge through the consideration of the pre-trained HED model (transfer learning).

Although not explicitly modeling edges, [*Marcu & Leordeanu, 2016*] propose a CNN for building and road extraction with initially two separate streams that is designed to learn object context at different scales, which they argue also enhances the *perception of shape.* Class-contours are not explicitly trained and thus, although overall results are promising, object boundaries are often fuzzy.

[*Basaeed et al., 2016*] propose a purely CNN-based segmentation algorithm as a general purpose remote sensing segmentation method, similar to the multi-resolution segmentation technique *of* [*Baatz M. & Schape, 2007*]. This CNN-based model is not constructed for extracting particular class-contours over VHSR data but for detecting variations and visual similarities within an RS images. From a model perspective, authors propose using a plethora of small CNN networks that cooperate cumulatively, whereas we build a single, larger (more weights per layer) and deeper pre-trained model.

[*Längkvist et al., 2016*] investigate the potential of CNNs for semantic segmentation using a set of small networks, initialized with k-means clustering. Within their detailed study they investigate the effect of introducing edge information in a post-processing step, where an object segmentation is derived using the SLIC algorithm [Achanta et al., 2010]. Instead of using a multi-step procedure with a heuristic, post-processing step, we propose learning class-contours and more general class evidence in a joint CNN directly from the data.

Our contribution in this work is the formulation of an end-to-end class-contour aware annotation framework, capable of directly modeling the class-edges to sharpen object boundaries. Furthermore, we propose a new way of modeling the class-contours, allowing for more flexibility over these much ambiguous regions by considering a buffer of uncertainty with additional weighting into the loss-function. We argue that reformulating the edge detection as a regression problem provides advanced flexibility in the overall modeling of the CNN training system and provides important quantitative improvements over a previous result.

3. The Methods

Constructing a class-contour model is far from trivial as edges may involve high localization-uncertainty and can follow very abstract patterns that do not explicitly agree with the image information. Class-contours are also much prone to the "*semantic-gap*" resulting from the difference of what humans perceive as a border between adjacent classes and the actual evidence contained within the RS image.

3.1 Definition of Class-Contours

Class edge uncertainty is a very common problem in RS image annotation. Various reasons such as illumination conditions, occlusion, acquisition angle, resolution or even imaging artifacts can strongly influence the precise localization of class-edges. Even if specific domain-knowledge is employed (manual human annotation) this process may still result in vague or partially wrong outcomes in RS and Vision related data as well [Yang, 2016].

It seems natural to account for edge localization uncertainty directly in our model. We tackle the class-contour delineation as a regression problem where the "*true edges"* have the largest score and this gradually decreases as we move away from this "*true" edge* location. After a small fixed distance the scoring values abruptly decrease, giving very little reward if the system predicts an edge far away from its actual location. This buffer around class-contours can handle noisy and even wrong edge-annotations.

We begin our model by segmenting the ground truth labels, resulting in a binary image of edge and non-edge regions. We dilate edges using a morphological operator and compute an Euclidean distance function over the dilated regions and the background (non-edge regions). For further enforcing the importance of edges and compensating for their rarity in the edge-image, all edges are weighted by a beta coefficient given by:

$$beta = non\text{-}edge\text{-}sum/tot\text{-}pixel \qquad (1)$$

Where *non-edge-sum* equals to the total number of pixels depicting non-edge data and tot-pixel equals to the sum of all pixels in the image. Similarly to (1) all non-edge pixels are multiplied with the opposite coefficient derived by:

$$1\text{-}beta + \epsilon \qquad (2)$$

Where epsilon adds a small constant (decimal) value to all non-edge location. This adaptation drives the optimization of the objective function to detect the rare but highly important class-contours and neglects the rest of the image.

3.2 Prediction of Class-Contours

Our model for inferring the class-contours is based on the approach of HED [*Xie et al., 2015*]. Our main contribution is that we do not tackle the class-contour problem as a classification problem, initially proposed in HED, but rather reformulate it as a regression problem with a mean squared loss-function. This allows us to better model contour uncertainty. Additionally, we use our dual-CNN structure of [*Marmanis et al., 2016*], where DEM and image components are separately processed initially and merged at a later, deeper stage of the network, allowing different input modalities to generate different types of features.

Our modified class-contour (CC) model was trained using standard Stochastic Gradient Descent with momentum with standard Backpropagation. The image stream weights were initialized using the HED weights and the DEM stream weights were initialized with *Xavier* weight initialization [*Glorot et al., 2010*]. Interestingly, when the standard HED weights were used for initializing the DEM CC stream, the system does not learn any meaningful representation. Contrary, when the *Xavier* initialization is employed the system seems to learn meaningful features.

Statistics over the validation class-contour dataset show that the model exhibits quite some overfitting, as the error on the training set is significant lower than the investigated validation set. This aspect signifies that is rather simple for such a model to learn the class-contours shape priors over RS data, where urban characteristics are rather interdependent according to some particular architectural and structural characteristics.

3.3 Complete Network Architecture

The second component of our model is the annotation network (AN) of [*Marmanis et al., 2016*]. Note that for the initial experiments presented here, we consider just a single annotation network architecture based on the VGG model (namely FCN-ImageNet in previous work) and not an ensemble of models as originally proposed in [*Marmanis et al., 2016*]. The annotation network has been separately trained on a set of annotation instances using an adaptation of [*Long et al., 2015*] where we also consider deep-supervision on various levels of the network [*Lee et al., 2014*].

For merging our two networks, namely the AN and CC there are two possible options, either allow the models to infer their respective representations individually and add a small merging-network on top, or place them in a sequential architecture where firstly the CC infers the class-contours and further propagate them to the AN network which produces the final annotation. In this work, we have decided upon the *sequential architecture* for its computational efficiency and due to the fact that through this approach, we avoid constructing a third merging network.

Furthermore, is important to underline that despite the fact that both CC and AN networks have converged before placing them into the sequential architecture (previously both models were trained individually), we have noted that the merged AN-CC architecture significantly decreases the error of the objective function when trained jointly for a few epochs. The reason for this is probably related with the fact that the merged system forces the AN network to heavily make use of the class-contours provided by the CC component part and improve final annotation.

4. Preliminary Experimental Results

4.1 Data set

We empirically validate our approach with experiments on a subset of the Vaihingen data set of the ISPRS 2D semantic labeling contest. Originally, the dataset comprises 33 tiles, varying in size, from an aerial orthophoto mosaic with three spectral bands (red, green, near-infrared), plus a digital surface model (DSM) of same resolution. The dataset contains roughly 1.7×10^8 pixels in total, but ground truth is only released for half of the tiles, which are designated for training and validation. The images are rich in detail, with a ground sampling distance of *9 cm*. Categories to be classified are *Impervious Surfaces, Buildings, Low Vegetation, Trees,* and *Cars*.

For our detailed experiments, we split the 16 tiles (along with provided labels), into a training subset *(tile numbers 1, 3, 11, 13, 15, 17, 21 ,26, 28, 32, 34, 37)* and a hold-out subset for validation *(tiles 5, 7, 23, 30)*. We randomly sample 12,000 patches of *259 × 259* pixels from the training subset for learning the CC-AN network parameters. Note that also at test time the network outputs labels for a complete patch of 259×259 pixels at once. To predict labels over an entire image, we run predictions on overlapping patches and average per-pixel class scores. The overlap employed between adjacent tiles is similar to [*Marmanis et al., 2016*], with values of *150, 200* and *220* pixel overlaps for breaking the aliasing effect.

4.2 Results

We provide quantitative and qualitative results on the hold-out validation set of four ISPRS Vaihingen images (tiles 5,7, 23, 30 of the benchmark). We report both qualitative, visual improvements of class-contours, as well as quantitative results (overall accuracies and confusion matrix). As a baseline for our comparison, we consider results from our previous work trained over the same data (with no class-contour component) as presented in [*Marmanis et al., 2016*].

It turns out that explicitly learning class-contours helps improving results. Compared to the *FCN-ImageNet* model in [Marmanis et al., 2016], overall accuracy increases from *85.1% to 88.8% (Table 1)*. Compared to the (more powerful) *CNN-ensemble* [*Marmanis et al., 2016*] that also applies a fully-connected CRF in a post-processing step, overall performance is still better (*88.8% vs. 86.0% - Table 1)*. These results show the critical role edges play in semantic segmentation. Learning these directly via CNN regression improves overall accuracy by *3.7* percent points compared to the VGG-based model and *2.8%* compared to the CNN-ensemble model. A per class evaluation (confusion matrix) is presented in *Table 2*, where the complete error-matrix of the current model is presented.

FCN-ImagenNet *[Marmanis et al., 2016]*	**Ensemble-CNN** *[Marmanis et al., 2016]*	**Ensemble-CNN+full-CRF** *[Marmanis et al., 2016]*	**CC-AN-CNN**
85.1 %	85.8 %	86.0 %	88.8%

Table 1. Overall Accuracies of the various considered models. Results from previous work as shown in columns 1-3 where results from this work are given in last column.

Imp. Surf.	**89.25 %**	3.75 %	5.75 %	1.25 %	0 %
Buildings	5.0 %	**93.5 %**	1.5 %	0 %	0 %
Low Veg.	7.75 %	2.25 %	**73.0 %**	16.75 %	0 %
Tree	2.75 %	0 %	6.5 %	**90.75 %**	0 %
Car	0.33 %	2.75 %	1.75 %	0.5 %	**62%**
Overall Accuracy : **88.84 %**					

Table 2. Confusion Matrix as evaluated on our validation set.

Qualitatively, we verify that the system accomplishes the task of learning the class-contours by visualizing the multi-scale learned edges over the training and validation set respectively (see *Figure 1* and *Figure 2*).

In *Figure 1* (second row), inferred class-contours follow a hierarchical approach where finer, insignificant details are eliminated sequentially at subsequent scale-levels resulting in a higher abstraction and more meaningful representation of the class-boundaries. As this particular image instance is retrieved from the training dataset, the system seems to predict the class-contours and end-annotation almost perfectly. On the other hand, in *Figure 2* where the class-contours are inferred over an instance from the validation set, the system fails to optimally compute the class-edges as the final contours are not complete (exhibit discontinuities). Nevertheless, even in this sub-optimal outcome, the general outline of the class-contours is visible. Hence the system exploits this evidence to improve its annotation accordingly.

Note that the CC network seems to overfit to the training data, which could be countered by using more training data, for example. Due to the somewhat limited size of the ISPRS Vaihingen data set, we employed strong "*Drop-Out" (*stochastically drop network connections of every training-epoch*)* over the network for trying to minimize this negative effect. Despite our efforts the network still suffers from quite some overfitting. Alternate strategies like data set augmentation is left for future work.

If visualizing feature maps at different levels of the merged CC-AN CNN network, we observe that the system keeps class-contours over all processing stages and finally enforces them to if evidence of the annotation component of the network is added. Such a property was not noticeable in our previous *Ensemble-CNN* network [*Marmanis et al., 2016*]. A qualitative example of this occurrence is shown in *Figure 3* below.

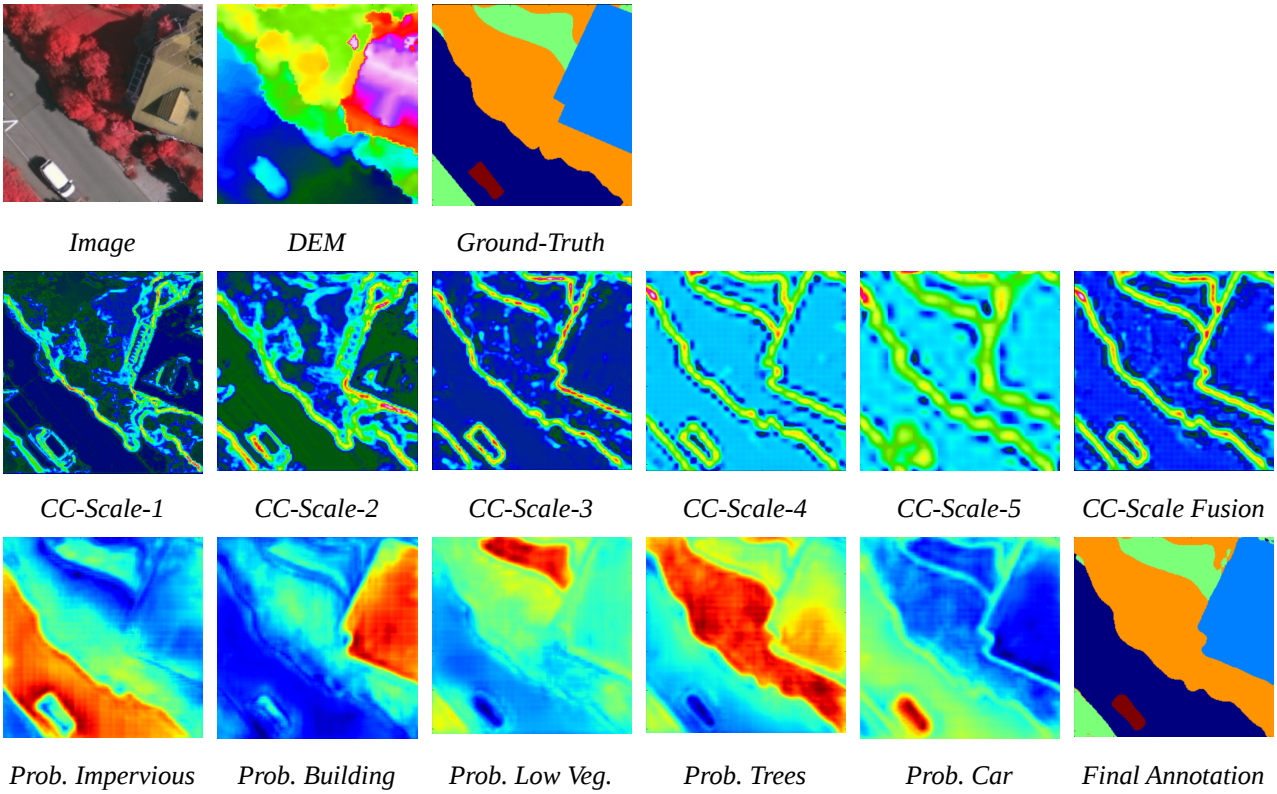

Figure 1. Outcomes of the merged CC-AN network over a sample from the training data set. In the first row, the input image, DEM and training annotation ground-truth are given respectively. In the second row, the various class-contours of the CC network are presented following the hierarchical nature of the CNN, as well as their cumulative fusion. In the third row, the per-class probability maps are depicted, as well as their final semantic annotation.

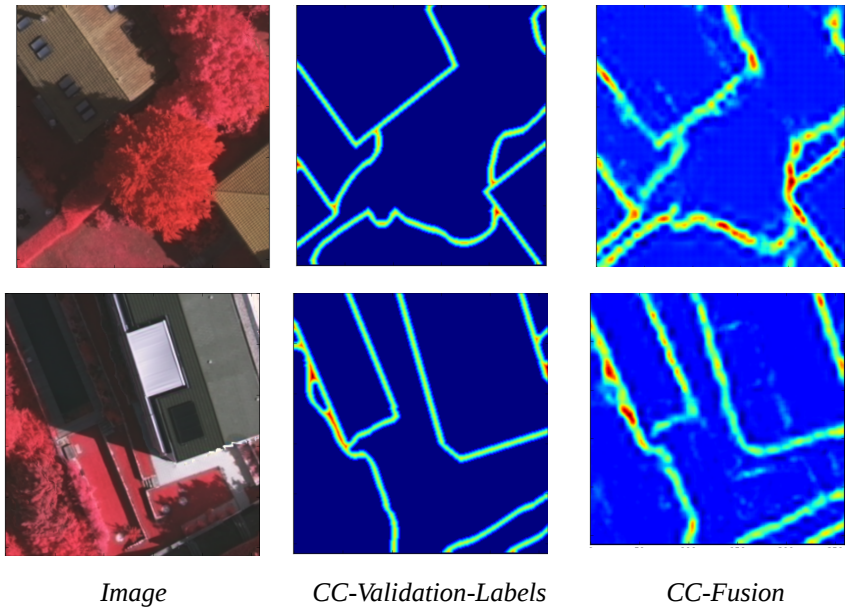

Figure 2. Class-Contour network inference over a few instances from the validation-set. On the left the provided input data to the network. In the middle the expected contour-classes labels. On the right the inferred class-contours. The network has never been trained on this particular examples as they are retrieved from the validation set.

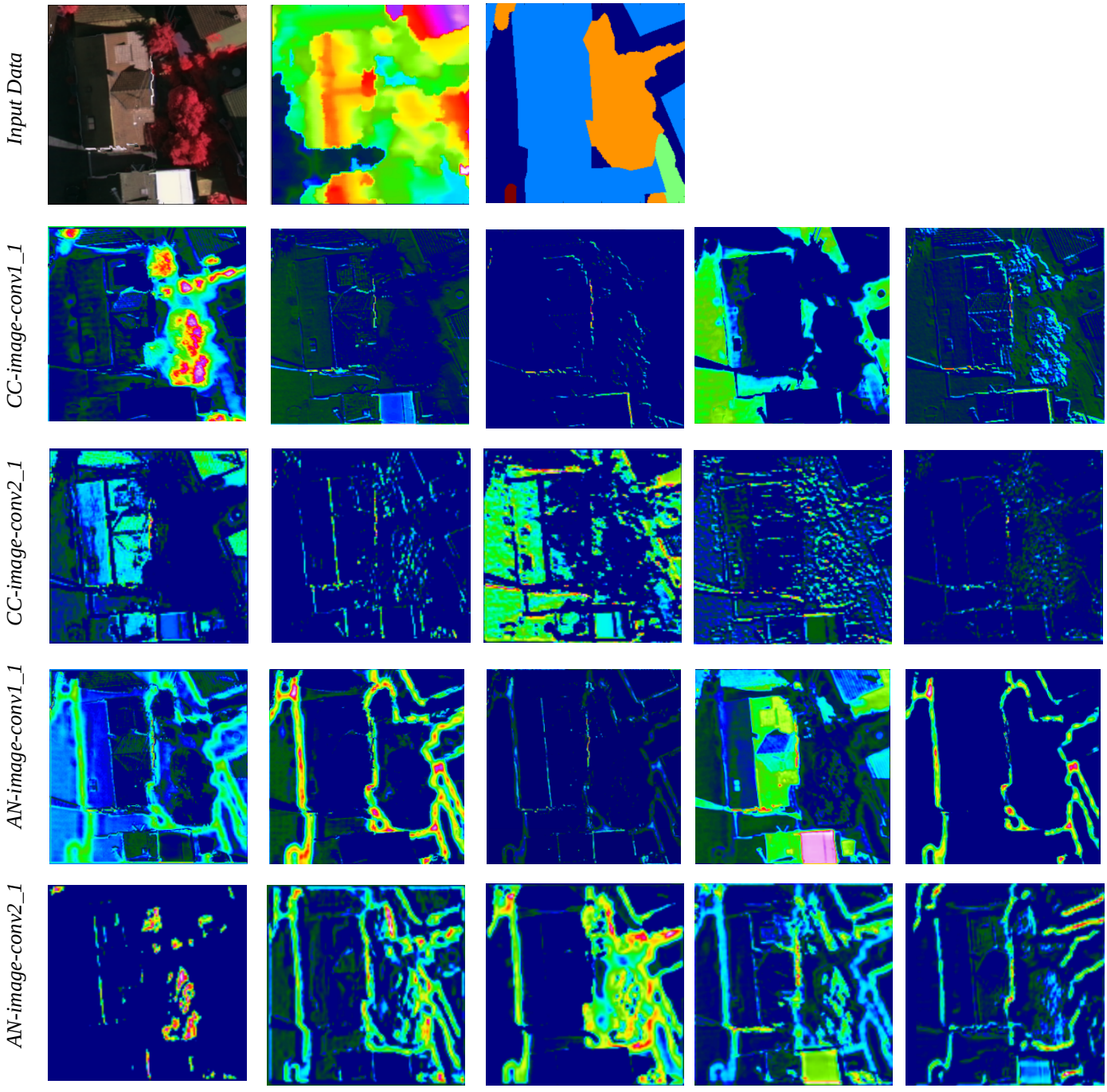

Figure 3. Generated feature maps on various depths of the merged CC-AN network. In the first row the input image, DEM and annotation ground-truth images respectively. On the second and third row, derived feature maps from the CC network component are presented. In the fourth and fifth row, feature maps from the AN network component are given. Thick class-contour edges seem to occur in the AN component, derived initially from the earlier CC part of the complete CC-AN network.

4.3 Discussion

Despite the interesting results achieved in this study, there is still much space for improvement. Introducing uncertainty in ground truth edges nicely accounts for small human labeling errors but also seems to cause contours to loose some of their sharpness. Furthermore, the current class-contour formulation is not specific to a transition between specific classes, that is, we do not

explicitly learn a typical "building-vegetation", for example. A CC network that would learn class-specific transitions separately could potentially collect further evidence and help better organizing it in the network together with the annotation component.

5. Conclusion

We have presented an end-to-end CNN semantic segmentation that explicitely learns to collect evidence from class-boundaries. The ensemble of such class-contours and standard pixel-wise annotation through a second CNN component results in significant improvements over our previously proposed plain annotation method.

References

Leung, T., & Malik, J. (2001). Representing and recognizing the visual appearance of materials using three-dimensional textons. International journal of computer vision, 43(1), 29-44.

Schmid, C., 2001. Constructing Models for Content-based Image Retrieval. In: Computer Vision and Pattern Recognition (CVPR).

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 1, pp. I-511). IEEE.

Dollár, P., Tu, Z., Perona, P. and Belongie, S., 2009. Integral channel features. In: British Machine Vision Conference.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

Pinheiro, P. O., Lin, T. Y., Collobert, R., & Dollár, P. (2016). Learning to refine object segments. arXiv preprint arXiv:1603.08695.

Yang, J., Price, B., Cohen, S., Lee, H., & Yang, M. H. (2016). Object Contour Detection with a Fully Convolutional Encoder-Decoder Network. arXiv preprint arXiv:1603.04530.

Malmgren-Hansen, D., & Nobel-J, M. (2015). Convolutional neural networks for SAR image segmentation. In 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) (pp. 231-236). IEEE.

Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1520-1528).

Farabet, C., Couprie, C., Najman, L., & LeCun, Y. (2013). Learning hierarchical features for scene labeling. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(8), 1915-1929.

Pinheiro, P. O., & Collobert, R. (2015). From image-level to pixel-level labeling with convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1713-1721).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3431-3440).

Yu, F., & Koltun, V. (2015). Multi-Scale Context Aggregation by Dilated Convolutions. arXiv preprint arXiv:1511.07122.

Xie, S., & Tu, Z. (2015). Holistically-nested edge detection. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1395-1403).

Kokkinos, I. (2015). Surpassing Humans in Boundary Detection using Deep Learning. arXiv preprint arXiv:1511.07386.

Marcu, A., & Leordeanu, M. (2016). Dual Local-Global Contextual Pathways for Recognition in Aerial Imagery. arXiv preprint arXiv:1605.05462.

Basaeed, E., Bhaskar, H., Hill, P., Al-Mualla, M., & Bull, D. (2016). A supervised hierarchical segmentation of remote-sensing images using a committee of multi-scale convolutional neural networks. International Journal of Remote Sensing, 37(7), 1671-1691.

Baatz M. & Schape A., (2007). Multiresolution Segmentationan Optimization Approach for High Quality Multi-scale Image Segmentation. http://www.definiens-imaging.com

Längkvist, M., Kiselev, A., Alirezaie, M., & Loutfi, A. (2016). Classification and Segmentation of Satellite Orthoimagery Using Convolutional Neural Networks. Remote Sensing, 8(4), 329.

Thoma, M. (2016). A survey of semantic segmentation. arXiv preprint arXiv:1602.06541.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2010). Slic superpixels (No. EPFL-REPORT-149300).

Marmanis, D., Wegner J., D., Galliani, S., Schindler, K., Datcu, M., Stilla, U. (2016). Semantic Segmentation of Aerial Images with an Ensemble of CNNs. ISPRS Annals (to appear)

Lee, C. Y., Xie, S., Gallagher, P., Zhang, Z., & Tu, Z. (2014). Deeply-supervised nets. arXiv preprint arXiv:1409.5185.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In International conference on artificial intelligence and statistics (pp. 249-256).