Fusion of Heterogeneous Data in Convolutional Networks for Urban Semantic Labeling

(Invited Paper)

Nicolas Audebert*†, Bertrand Le Saux*
*ONERA, *The French Aerospace Lab*
F-91761 Palaiseau, France
{nicolas.audebert,bertrand.le_saux}@onera.fr

Sébastien Lefèvre†
†Univ. Bretagne-Sud, UMR 6074, IRISA
F-56000 Vannes, France
sebastien.lefevre@irisa.fr

***Abstract*—In this work, we present a novel module to perform fusion of heterogeneous data using fully convolutional networks for semantic labeling. We introduce residual correction as a way to learn how to fuse predictions coming out of a dual stream architecture. Especially, we perform fusion of DSM and IRRG optical data on the ISPRS Vaihingen dataset over a urban area and obtain new state-of-the-art results.**

I. Introduction

Following the take over of deep larning over the computer vision field, deep convolutional neural networks (CNN) propagated to remote sensing image processing. Deep networks are now state-of-the-art for object detection and classification, but also for semantic labeling, both in everyday images, e.g. PASCAL VOC2012 [1], and Earth Observation data, e.g. ISPRS Vaihingen 2D Semantic Labeling Challenge [2]. However, these deep networks have been originally designed for everyday RGB images. On the contrary, remote sensing data is rarely limited neither to RGB, nor to optical data, and often combines several heterogeneous sensors. In scene understanding of Earth Observation images, data fusion can therefore significantly improve a statistical model's accuracy by combining specific information from the different sensors. For example, hyperspectral and LiDAR sensors have different spatial resolution and do not share the same physical properties, although both the spectrum and the measured height can be relevant features for classification. In this work, we present a new residual correction module designed to perform efficient data fusion using CNN. We apply this technique to the IRRG images and DSM data of the ISPRS Vaihingen dataset and obtain new state-of-the-art results.

II. Related Work

Most works related to deep learning for urban semantic labeling use 3-channels networks designed for RGB (and sometimes IRRG), fine-tuned from a model trained on the ImageNet dataset [3], [4], [5]. Dual stream neural networks for data fusion have been introduced in [6] in an unsupervised framework for joint audio-video representation learning, using a dual stream auto-encoder. The same principles have been transposed to supervised learning in [7] for classification of RGB-D data.

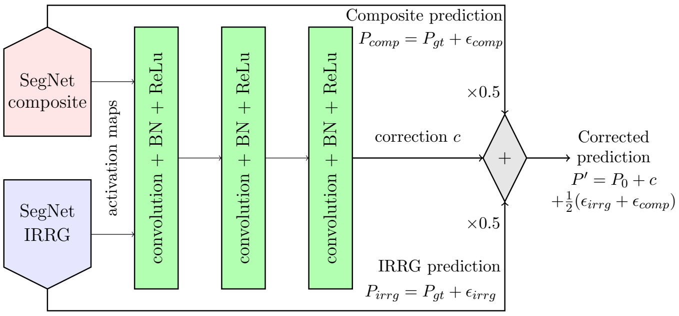


Figure 2: Fusion network to correct predictions with information from complementary SegNets using heterogeneous data.

Data fusion using CNN for classification of remote sensing images has also been explored in the Data Fusion Contest (DFC) 2015, where CNN have been used for multimodal and multi-scale feature extraction in combination with a SVM classifier [8]. Semantic labeling on the ISPRS Vaihingen dataset was further improved using fusion of CNN-based and expert-crafted features with random forests [3]. In the DFC 2016, semantic labeling based on a high resolution multispectral image and tracklet analysis on a spaceborne video were combined for traffic density and activity analysis [9].

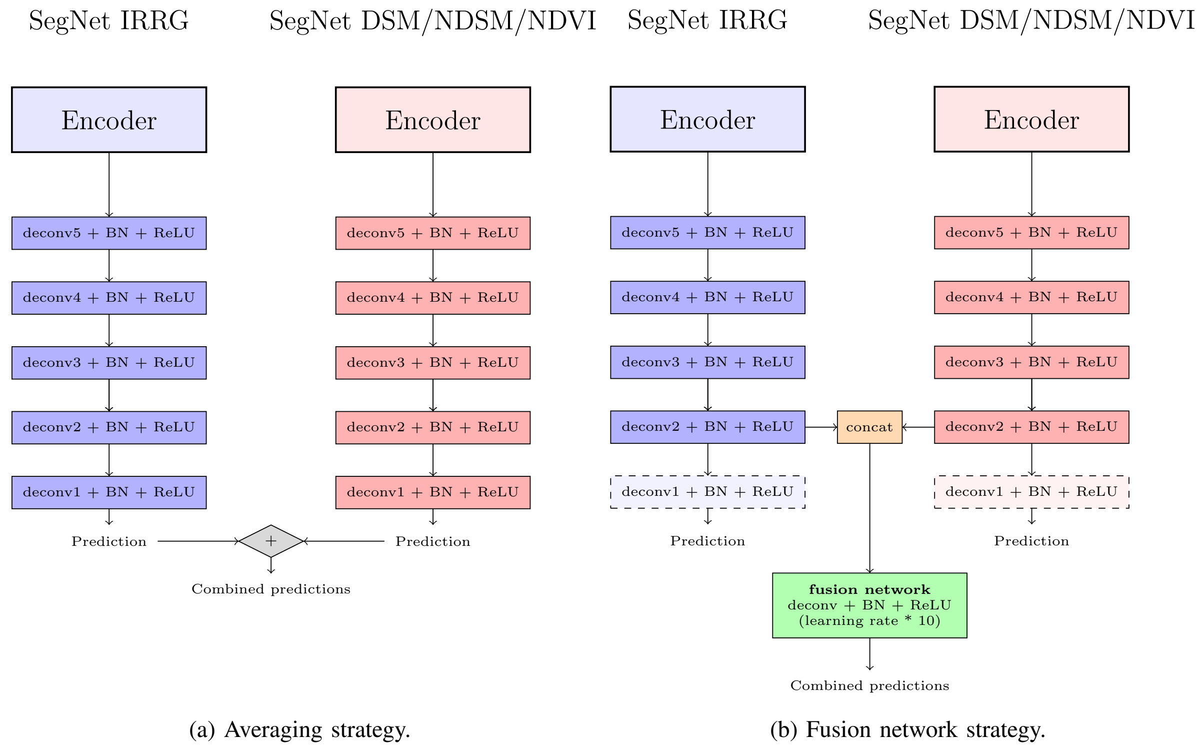
Finally, residual learning [10] was introduced with the idea that deep networks have trouble learning the identity function. Using bypass connections, the network would only have to learn a residual addition to the input, which would be easier.

Building on these works, our residual correction is a generic module fully integrated in the CNN pipeline that can be added on any multiple stream architecture. Moreover, it uses recent advances in deep learning by linking residual learning to the signal processing viewpoint on error correction. Especially, we integrate the fusion with the recent fully convolutional networks (FCN) [11] that are able to perform end-to-end dense semantic labeling.

III. Heterogeneous Data Fusion with Residual Correction

A naive approach to data fusion using deep networks would be to concatenate all channels (e.g. RGB and depth) and use

978-1-5090-5808-2/17/$31.00 © 2017 IEEE



(a) Averaging strategy.

(b) Fusion network strategy.

Figure 1: Fusion strategies of our dual-stream SegNet architecture.

it as the input. However, preliminary experiments showed that this actually degrades the network accuracy as heterogeneous data sources require different processing. Luckily, prediction oriented fusion has been proven effective using dual stream networks both in unsupervised [6] and supervised [7] settings. Therefore, we first train two 3-channels SegNet [12] for semantic labeling using our two data sources. Then, we perform prediction fusion at end of the networks by merging the two parallel streams. As a baseline, we just perform simple averaging after the softmax (Fig. 1a). To improve the fusion, we introduce a fusion neural network that learns to improve the average prediction by using data specific information.

Building on the idea of residual deep learning [10], we propose a fusion network based on residual correction. We define a convolutional residual block using the same parameters as the rest of the SegNet network (3×3 convolutions and 1 pixel of padding). Intermediate feature maps from the decoding parts of the two SegNets are fed as inputs to a 3-convolution layers “correction” network (cf. Fig. 1b. The output of the residual block is then summed in the residual fashion with the average of the two predictions, as illustrated by Fig. 1b. Residual learning fits this use case, as the average score is already a close estimate of the truth. To improve the results, we aim to use the complementary channels to correct small errors in the prediction maps. In this context, the residual can be seen as a corrective term for our predictive model. This module is trained using backpropagation on the standard softmax loss. Learning rates for the input SegNets are set to zero, as this considerably speeds up the training without significant loss.

Assuming that P_0 is the ground truth tensor and P_i is the predicted output of the i^{th} stream, we have :

$$P_i = P_0 + \epsilon_i \text{ where } |\epsilon_i| \ll |P_i| \tag{1}$$

ϵ_i is an error term that is small if the prediction P_i is accurate enough. We expect the network to learn to estimate the errors and to infer when and how to merge the streams.

Let R be the number of outputs on which to perform residual correction. We predict P', the sum of the averaged predictions and the correction term c:

$$P' = P_{avg} + c = \frac{1}{R}\sum_{i=1}^{R} P_i + c = P_0 + \frac{1}{R}\sum_{i=1}^{R} \epsilon_i + c \tag{2}$$

As our residual correction module is optimized to minimize the loss, we enforce:

$$\|P' - P_0\| \to 0 \tag{3}$$

which translates into a constraint on c and ϵ_i:

$$\|\frac{1}{R}\sum_{i=1}^{R} \epsilon_i - c\| \to 0 \tag{4}$$

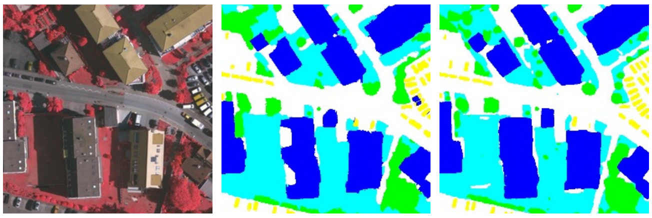
This can be seen as learning a model of the average error based on the feature maps. Indeed, at training time, the ground truth P_0 is known and the residual correction learns how to infer $\sum_{i=1}^{R} \epsilon_i$. The residual learning framework suits well this idea of error correction, as the residual is expected to be of a small amplitude compared to the main identity (or “bypass”) signal. The residual correction module is detailed in Fig. 2.

Table I: Results on the validation set.

Type/Stride (px)	128	64	32
Single stream (IRRG)	87.8%	88.3%	88.8%
Fusion (average)	88.2%	88.7%	89.1%
Fusion (correction)	88.6%	89.0%	89.5%

Table II: ISPRS 2D Semantic Labeling Vaihingen results.

Method	imp surf	building	low veg	tree	car	OA
RF + CRF (“HUST”)[14]	86.9%	92.0%	78.3%	86.9%	29.0%	85.9%
CNN+RF+CRF (“ADL_3”)[3]	89.5%	93.2%	82.3%	88.2%	63.3%	88.0%
FCN (“DLR_2”)[4]	90.3%	92.3%	82.5%	89.5%	76.3%	88.5%
FCN+RF+CRF (“DST_2”)	90.5%	93.7%	83.4%	89.2%	72.6%	89.1%
Ours (IRRG only)	**91.5%**	94.3%	82.7%	89.3%	**85.7%**	89.4%
Ours (fusion)	91.0%	**94.5%**	**84.4%**	**89.9%**	77.8%	**89.8%**



IRRG data “DLR” (FCN)[4] Ours (SegNet)

Figure 3: Comparison of our method with a FCN on the ISPRS Vaihingen benchmark. Building detection is not impeded by shadows anymore and cars are more finely segmented.
(white: roads, blue: buildings, cyan: low vegetation, green: trees, yellow: cars)

IV. Experiments

A. Experimental Setup

We test our method on the ISPRS Vaihingen 2D Semantic Labeling dataset comprised of IRRG images over an urban area. The 3 channels (i.e. near-infrared, red and green) are processed as an RGB image in the first stream of our dual SegNet architecture. The dataset also includes a Digital Surface Model (DSM) acquired with a Lidar and the Normalized Digital Surface Model (NDSM) from [13]. We compute the Normalized Difference Vegetation Index (NDVI) from the near-infrared and red channels, which is an indicator for vegetation ($NDVI = (IR - R)/(IR + R)$).

For each tile, we aggregate DSM, NDSM and NDVI into a composite image used in the second stream of our architecture. The two streams use mostly heterogeneous data (height versus optical data). The composite image also includes redundant optical data (the NDVI) so that it contains relevant for information for all the classes (e.g. height helps discriminate road vs building while NDVI helps find the vegetation). Therefore, both the IRRG and composite images can be used to infer segmentations with similar accuracies, which will ease the predictions fusion.

We process the tiles using a 128×128 sliding window with a 32px stride. We split the tiles with the public ground truth into a training set (12 tiles) and a validation set (4 tiles). We train separately the two SegNets for 10 epochs with a learning rate of 0.1, divided by 10 after 5 epochs. For our baseline, we compute the average prediction during testing. We fine-tune the residual correction module for 1 epoch, as longer training does not improve convergence.

B. Results

Our best model achieves state-of-the art results on the ISPRS Vaihingen dataset (cf. Table II) [1]. Fig. 3 illustrates a qualitative comparison between our SegNet-based residual correction and a traditional fully convolutional architecture on an extract of the Vaihingen testing set. The provided metrics are the global pixel-wise overall accuracy (OA) and the F1 score on each class:

$$F1_i = 2\frac{precision_i \times recall_i}{precision_i + recall_i} \tag{5}$$

$$recall_i = \frac{tp_i}{C_i},\ precision_i = \frac{tp_i}{P_i}\ , \tag{6}$$

where tp_i the number of true positives for class i, C_i the number of pixels belonging to class i, and P_i the number of pixels attributed to class i by the model. These metrics are computed using an alternative ground truth in which the borders have been eroded by a 3px radius circle.

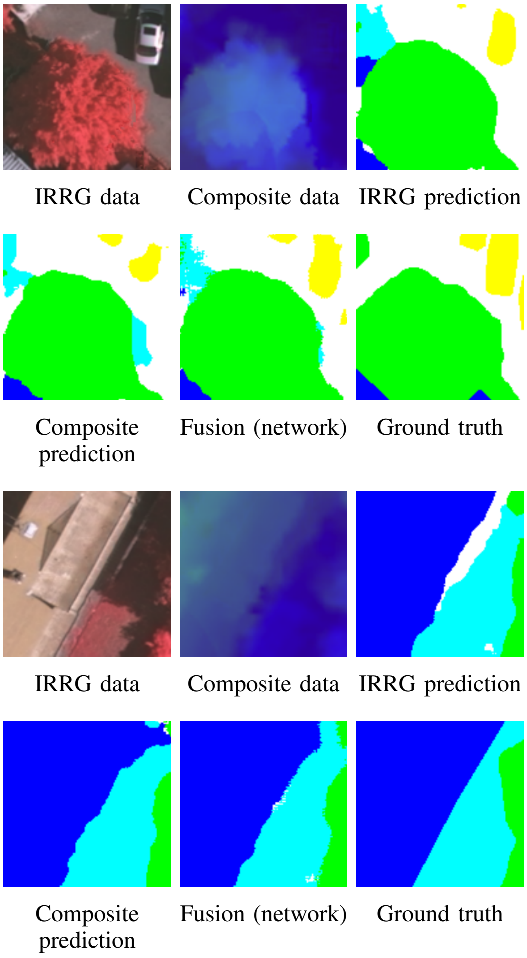
C. Analysis

On the validation set, naive fusion by averaging the maps boosts the OA by 0.4%, and the residual correction improves it further by an additional 0.4%. As illustrated in Fig. 4, the fusion manages to correct errors in one model by using information from the other source. The residual correction network generates more visually appealing predictions, as it learns which network to favor for each class. For example, the IRRG data is often right when predicting car pixels, therefore the correction network trusts the IRRG prediction about cars more often. However the composite data has the advantage of the DSM to help distinguishing between low vegetation and trees. Thus, the correction network gives more weight to the predictions of the “composite SegNet” for these classes. Interestingly, if m_{avg}, m_{corr}, s_{avg} and s_{corr} denote the respective mean and standard deviation of the activations after averaging and after correction, we see that $m_{avg} \simeq 1.0$, $m_{corr} \simeq 0$ and $s_{avg} \simeq 5$, $s_{corr} \simeq 2$. We conclude that the network actually learnt how to apply small corrections to achieve a higher accuracy, which is in phase with both our expectations and theoretical developments [10].

This approach obtains state-of-the-art results on the ISPRS Vaihingen 2D Labeling Challenge at 89.8% [2] (cf. Table II). F1 scores are significantly improved on buildings and vegetation, thanks to the discriminative power of the DSM and NDVI. However, even though the F1 score on cars is competitive, it is lower than expected. This is partly due by poor position and

[1] http://www2.isprs.org/vaihingen-2d-semantic-labeling-contest.html

[2] “ONE_7”: https://www.itc.nl/external/ISPRS_WGIII4/ISPRSIII_4_Test_results/2D_labeling_vaih/2D_labeling_Vaih_details_ONE_7/index.html



IRRG data | Composite data | IRRG prediction

Composite prediction | Fusion (network) | Ground truth

IRRG data | Composite data | IRRG prediction

Composite prediction | Fusion (network) | Ground truth

Figure 4: Effects of our fusion strategies on selected patches.

height values of cars in the DSM, making fusion harder for the network. We wish to investigate this issue further, e.g. by incorporating hard-negative mining to help the fusion module learn how to merge very hetereogeneous predictions. Nonetheless, we argue that such our residual correction module can help improve exploit the complementarity of heteregeneous inputs when one has trained an ensemble of classifiers.

V. Conclusion

In this work, we presented a residual correction neural network designed to perform prediction-oriented data fusion of heterogeneous sources. On top of parallelized deep fully convolutional networks, the residual correction improved our semantic labeling model using sensor specific information. Especially, our experimental study showed that the residual correction is able to accurately identify which stream to trust for the different classes. We validated the residual correction technique on Earth Observation data, specifically the ISPRS 2D Vaihingen Semantic Labeling challenge, on which we fused IR/R/G, height information and NDVI data and improved the state-of-the-art by 1%.

Future work involves using residual correction to merge streams coming from networks with different topologies and making the fusion more aware of the early layers, in order to benefit from a mix of low, medium and high level features. We also would like to show that this solution generalizes to other use cases of fusing predictions from several classifiers.

Acknowledgment

The Vaihingen dataset was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [15]: http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html. Nicolas Audebert's work is supported by the Total-ONERA research project NAOMI.

References

[1] M. Everingham, S. M. A. Eslami, L. Gool, C. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jun. 2014.

[2] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf, "The ISPRS benchmark on urban object classification and 3d building reconstruction," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci*, vol. 1, p. 3, 2012.

[3] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van Den Hengel, "Effective semantic pixel labelling with convolutional networks and Conditional Random Fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2015, pp. 36–43.

[4] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic Segmentation of Aerial Images with an Ensemble of CNNs," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 3, pp. 473–480, 2016.

[5] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 105–109, Jan. 2016.

[6] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[7] A. Eitel, J. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proceedings of the International Conference on Intelligent Robots and Systems*. IEEE, 2015, pp. 681–687.

[8] A. Lagrange, B. Le Saux, A. Beaupere, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, and M. Ferecatu, "Benchmarking classification of earth-observation data: From learning explicit features to convolutional networks," in *IEEE International Geosciences and Remote Sensing Symposium (IGARSS)*, Jul. 2015, pp. 4173–4176.

[9] L. Mou and X. Zhu, "Spatiotemporal Scene Interpretation of Space Videos via Deep Neural Network and Tracklet Analysis," in *IEEE International Geosciences and Remote Sensing Symposium (IGARSS)*, Jul. 2016.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[12] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[13] M. Gerke, "Use of the Stair Vision Library within the ISPRS 2d Semantic Labeling Benchmark (Vaihingen)," International Institute for Geo-Information Science and Earth Observation, Tech. Rep., 2015.

[14] N. T. Quang, N. T. Thuy, D. V. Sang, and H. T. T. Binh, "An Efficient Framework for Pixel-wise Building Segmentation from Aerial Images," in *Proceedings of the Sixth International Symposium on Information and Communication Technology*. ACM, 2015, p. 43.

[15] M. Cramer, “The DGPF test on digital aerial camera evaluation – overview and test design,” *Photogrammetrie – Fernerkundung – Geoinformation*, vol. 2, pp. 73–82, 2010.