FUSENET: END-TO-END MULTISPECTRAL VHR IMAGE FUSION AND CLASSIFICATION

John Ray Bergado, Claudio Persello, Alfred Stein

Dept. of Earth Observation Science, ITC, University of Twente,
Enschede, The Netherlands
j.r.bergado@utwente.nl, c.persello@utwente.nl, a.stein@utwente.nl

ABSTRACT

Classification of very high resolution (VHR) satellite images faces two major challenges: 1) inherent low intra-class and high inter-class spectral similarities and 2) mismatching resolution of available bands. Conventional methods have addressed these challenges by adopting separate stages of image fusion and spatial feature extraction steps. These steps, however, are not jointly optimizing the classification task at hand. We propose a single-stage framework embedding these processing stages in a multiresolution convolutional network. The network, called *FuseNet*, aims to match the resolution of the panchromatic and multispectral bands in a VHR image using convolutional layers with corresponding downsampling and upsampling operations. We compared *FuseNet* against the use of separate processing steps for image fusion, such as pansharpening and resampling through interpolation. We also analyzed the sensitivity of the classification performance of *FuseNet* to a selected number of its hyperparameters. Results show that *FuseNet* surpasses conventional methods.

Index Terms— Convolutional networks, image fusion, land cover classification, VHR image, deep learning.

1. INTRODUCTION

Classification of very high resolution (VHR) satellite images presents two major challenges: 1) inherent low intra-class and high inter-class spectral similarities and 2) mismatching resolution of available bands. The first challenge is often addressed by extracting spatial-contextual features from the image such as texture-describing measures, e.g. gray level co-occurrence matrix (GLCM) and local binary patterns (LBP) [1] or products of morphological operators that are expected to reduce spectral class ambiguities. The second challenge is dealt with pansharpening and interpolation-based resampling techniques used to fuse images of different resolutions. A typical approach to classification of a multiresolution VHR satellite image would then be as shown in Figure 1 (a). These additional steps to address problems in classifying a multiresolution VHR satellite image are disjoint from the supervised classifier, and hence, not optimized for the task at hand.

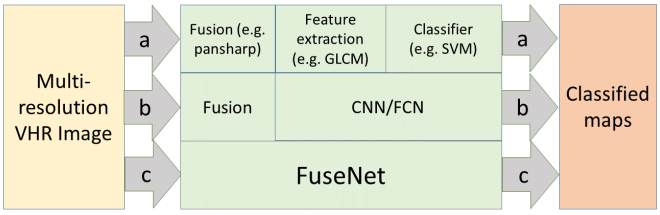

Fig. 1. Comparison of a standard (a), state-of-the-art (b), and proposed (c) piplelines for classifying multiresolution VHR images.

Deep learning offers a framework to build *end-to-end* classifiers by directly learning the predictions from the inputs with minimal or no separate pre-classification steps. Convolutional neural networks (CNN), for instance, integrate the feature extraction step within the training of the supervised classifier and have performed better than intermediate handcrafted features [2, 3]. Recently, a patch-based CNN [3] and a fully convolutional network (FCN) [4], utilizing pansharpening for image fusion, were used to detect informal settlements from a multiresolution VHR satellite image. Both works have addressed the classification challenges as in Figure 1 (b). In this paper, we present a novel single-stage network performing image fusion and classification of a multiresolution VHR satellite image in an end-to-end fashion as in Figure 1 (c).

2. METHODOLOGY

We propose a multiresolution convolutional network, called FuseNet, to perform an end-to-end image fusion and classification of a multi-resolution VHR satellite image. FuseNet is built on top of a fully convolutional network architecture learning to: 1) fuse panchromatic (PAN) and multispectral (MS) bands of a VHR satellite image, 2) extract spatial feature, and 3) classify land cover classes.

FuseNet is specifically designed for VHR satellite images with PAN band and MS bands having a ground sampling distance ratio of four (e.g. Quickbird, Worldview 2/3, Pleiades, Ikonos). This architecture can be generalized to fuse any number of images with different spatial resolutions and any

number of bands. It accepts two sets of input: an image patch of dimensions $N \times 1 \times 4M \times 4M$ taken from a PAN image and another patch of dimensions $N \times 4 \times M \times M$ taken from corresponding locations in the MS image. It performs two series of convolution, nonlinearity, and maximum pooling with downsampling to the PAN image patches such that the spatial dimensions of the intermediate feature maps match the spatial dimensions of the MS image patches. The nonlinear operations use an exponential linear activation function [5]. The second input is linearly projected in k dimensions using 1×1 convolutions such that k matches the number of intermediate feature maps extracted from the first set of input. This ensures that succeeding feature maps extract the same number of pattern variations from both sets of inputs. FuseNet merges the linear projection of the MS image patches with intermediate feature maps extracted from the PAN image patches via a concatenation operation.

Additional series of convolution, nonlinearity, and maximum pooling with downsampling operations are applied to the merged feature maps thus producing a set of feature maps with the smallest spatial dimensions—called a *bottleneck*. FuseNet then upsamples the bottleneck back to the resolution of the PAN input image patches using transposed convolutions. The resulting set of feature maps is linearly projected again using 1×1 convolutions such that the number of feature maps matches the number of classes C. FuseNet applies a softmax activation to calculate normalized class score maps and couples those with a cross-entropy loss function:

$$E_N = -\sum_{n=1}^{N} \mathbf{t}_n \bullet \log(\mathbf{y}_n) \tag{1}$$

where E is the loss function value evaluated over N samples, $\mathbf{t}_n$ is a binary vector encoding the the target class labels (with the index corrresponding to a class having a value of 1 and 0 otherwise), $\bullet$ denotes the dot product, and $\mathbf{y}_n$ is the class score maps of a sample n calculated using a *softmax* activation function.

This configuration of FuseNet is called FuseNet$_{low}$ because it performs fusion at the lower (MS image) resolution. We also tested a network, called FuseNet$_{skip}$, adding skip connections to lower-level feature maps of FuseNet$_{low}$ [6]. Additionally, we experimented with a version of FuseNet performing fusion at the resolution of the PAN image, called FuseNet$_{high}$ which is more similar to pansharpening as it upsamples the MS image patches first before fusing them with the PAN image patches. Table 1 shows details of the operations, including dimensions of intermediate output feature maps used by FuseNet$_{low}$.

Table 1. Detailed operations of FuseNet$_{low}$.

FuseNet$_{low}$	
$\mathbf{x}_{PAN}(1 \times 4M \times 4M)$	$\mathbf{x}_{MS}(4 \times M \times M)$
conv13-16 maxpool conv7-32 maxpool	conv1-32
IFM1 ($32 \times M \times M$)	IFM2 ($32 \times M \times M$)
concat	
IFM3 ($64 \times M \times M$)	
conv3-64 maxpool conv3-128 maxpool	
BFM ($128 \times M/4 \times M/4$)	
ups2-128 ups2-64 ups2-32 ups2-16 conv1-6	
IFM4 ($6 \times 4M \times 4M$)	
softmax	

Table format as in [7].
$\mathbf{x}_{PAN}$ and $\mathbf{x}_{MS}$ denote input patches from the PAN and MS images, respectively.
IFM and BFM corresponds to intermediate and bottleneck feature maps respectively.

3. EXPERIMENTAL RESULTS

We evaluated the proposed network for a land cover classification of a dataset covering Quezon City, Philippines. The dataset is composed of a Worldview-03 satellite image acquired on 17^{th} April 2016 and manually prepared reference images for five chosen tiles (subsets) of the satellite image. The satellite image has a PAN band of 0.3 m resolution and four MS bands (near-infrared, red, green, and blue) of 1.2 m resolution.

The satellite image was first divided into regularly-sized image tiles. PAN image tiles have a dimension of 3200×3200 pixels, while MS image tiles have a dimension of 800×800 pixels. Five non-adjacent tiles were sparsely labeled—annotating a pixel with a label belonging to one of the following six classes: 1) impervious surface, 2) building, 3) low vegetation, 4) tree, 5) car, and 6) clutter. Two of the five labeled tiles were used for training (tiles 100 and 105), one for validation (tile 45), and the remaining two for testing (tiles 78 and 82). Figure 2 shows the two tiles used for testing.

We compared FuseNet with two baseline methods: one using pansherpening and the second using bilinear interpolation to match the resolution of the MS image patches to the resolution of the PAN image patches, called Net$_{pansharp}$, SegNet [8], Net$_{pan-cnn}$, and Net$_{bilinear}$, respectively. Net$_{pansharp}$ applies the Gram-Schmidt pansharpening technique, SegNet uses the first three principal components of the inputs of Net$_{pansharp}$, while Net$_{pan-cnn}$ adapts the CNN-based

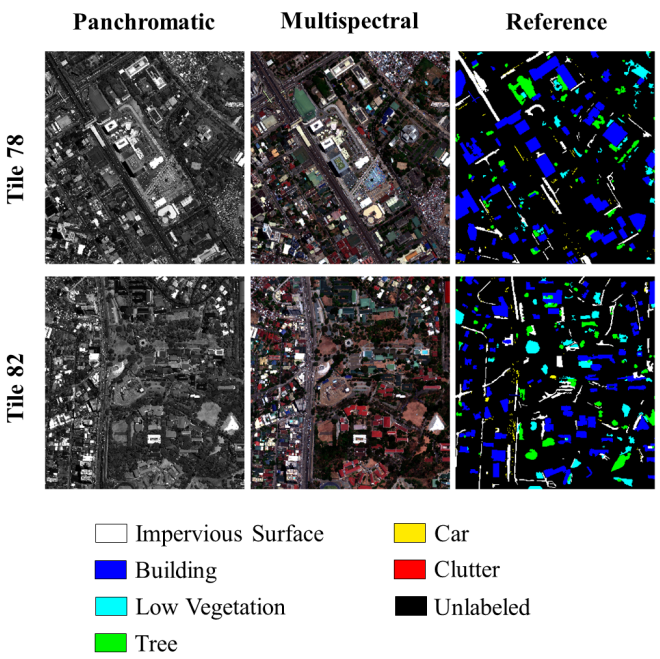

Fig. 2. PAN, MS, and reference images in the tiles used for testing. Corresponding legend is shown.

pansharpening method proposed in [9]. Only the pansharpened image is fed as an input into Net$_{pansharp}$, SegNet, and Net$_{pans-cnn}$. In contrast, Net$_{bilinear}$ upsamples the resolution of the MS image to match the resolution of the PAN image using bilinear interpolation. The upsampled MS images are then merged with the PAN image using concatenation. The architecture of the network after the fusion is kept the same to have a fair comparison among the different methods.

Table 2. Comparison of fusion approaches

Network	OA (%)	κ (%)	AA (%)	F1 (%)
Net$_{bilinear}$	84.76	78.70	81.99	77.48
Net$_{pansharp}$	86.87	81.53	82.76	77.86
Net$_{pan-cnn}$ [9]	87.88	82.69	84.58	72.45
SegNet [8]	88.11	83.17	83.96	77.01
FuseNet$_{high}$	88.03	83.18	89.79	79.06
FuseNet$_{low}$	91.63	88.03	92.91	**82.90**
FuseNet$_{skip}$	**91.90**	**88.43**	**93.46**	81.74

Table 2 shows the results of accuracies comparing different fusion approaches. FuseNet$_{skip}$ scores the highest in all the four numerical metrics, except for F1 where FuseNet$_{low}$ scores the highest. Correspondingly, FuseNet$_{low}$, the architecture from which FuseNet$_{skip}$ was derived, outperforms all the other networks, except for FuseNet$_{skip}$ itself. Observing each metric: FuseNet$_{low}$ gains about 3–6% in OA, 4–9% in κ, 3–10% in AA, and 3–10% in F1 against the other baselines. FuseNet$_{skip}$ further increases the numerical results of FuseNet$_{low}$ for the first three metrics by about 0.2–0.5%, but degrades the F1 by about 1.2%.

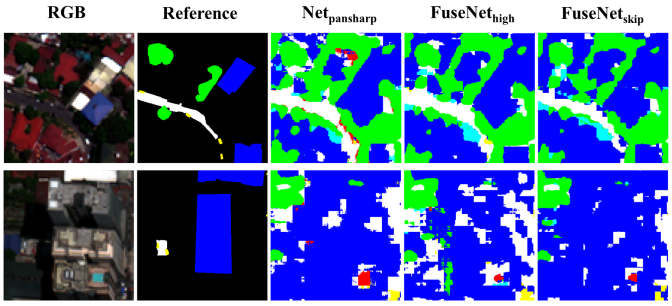

Fig. 3. Classification maps from selected FuseNet variants and baseline methods of a zoomed subset from the test tiles.

We notice that: 1) learning fusion can improve the classification of PAN and MS VHR images with different resolutions; 2) fusing at the scale of the image with lower resolution results in better classification than performing fusion at the scale of the image with higher resolution. The first point demonstrates our expected effectiveness of coupling and learning the fusion operation within a supervised classifier. Regarding the second point, introducing upsampling layers early in the network (FuseNet$_{high}$) may produce artifacts that can degrade its performance.

Figure 3 shows the classification maps from selected FuseNet variants and baseline methods. The most noticeable misclassifications are found in large and high-rise buildings and overpassing roads. The facades and rooftops of the buildings are often mistaken to be impervious surfaces by the classifiers, while overpassing roads are mistaken to be a building. These regions can appear to have similar spectral characteristics and can only be distinguished by presence of other indications such as appearing to be elevated. Manually distinguishing arguably vaguely-defined classes such as low-vegetation and impervious surface can also be problematic, especially in the PAN image, with the lack of ancillary information such as elevation. The cars are also generally misclassified by all the classifiers. This is, aside from being underrepresented in terms of the number of labeled pixels, due to the lack of spatial resolution of the MS bands and the spectral similarity of cars with other classes impervious surface and buildings in the PAN band. Overall, FuseNet$_{skip}$ has less errors in the facade of large buildings and provides a better delineation of classes with irregular boundaries such as trees and low-vegetation—providing the best classification results.

3.1. Sensitivity Analysis

Fig. 4 shows the results of a sensitivity analysis performed on four chosen hyperparameters of FuseNet: 1) bottleneck fea-

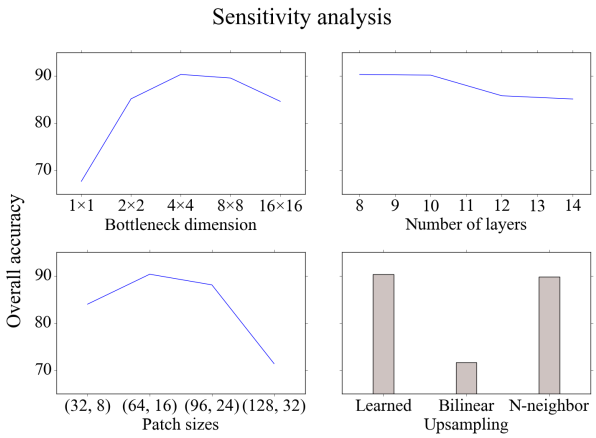

Fig. 4. Plots showing the results of sensitivity analysis. Patch sizes are written as "$\langle(4M, M)\rangle$". N-neighbor denotes nearest neighbor interpolation.

ture map dimensions, 2) number of convolutional layers in the downsampling part of the network, 3) input patch sizes, and 4) upsampling methods. We got the highest validation accuracy of 90.35% using a bottleneck feature map dimension of 4×4 pixels. Decreasing the dimension below its optimum severely degrades the classification resulting to large uniform areas producing stamp-like patterns especially at the 1×1 level. Increasing the dimensions produces much noisier classification. Fixing the bottleneck size dimension to 4×4 and further increasing the number of convolutional layers without downsampling did not produce any improvements in the validation accuracy. Hence, the results show that with only eight convolutional layers with downsampling, we can learn enough contextual information for the most accurate classification.

We found the optimal patch sizes to be equal to 64×64 for the PAN image patches and 16×16 for the MS image patches. Further increasing the patch sizes results in overclassification of a single class impervious surface. Increasing the patch size also increases the proportion of frequently occurring classes in the training sample, possibly resulting into overclassification. Lastly, we noted that the use of transposed convolution for learned upsampling performs better than the use of interpolation for fixed upsampling. This result supports the expected flexibility of empirically learning upsampling directly from the data.

4. CONCLUSION

In this paper, we presented a multiresolution convolutional network named FuseNet to classfiy a VHR satellite image. The operations for fusing the bands with different resolutions are learned within convolutional layers with corresponding downsampling and upsampling operations to match the resolution of the images. Results show the advantages of incorporating image resolution matching within the training of the classifier. To this end, we provided a single-stage classification pipeline incorporating image fusion and feature extraction combined in a convolutional network trained in an end-to-end manner.

5. REFERENCES

[1] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, Jul 2002.

[2] J. R. Bergado, C. Persello, and C. Gevaert, "A deep learning approach to the classification of sub-decimetre resolution aerial images," in *Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS) 2016*, 2016, vol. 2016-November, pp. 1516–1519.

[3] N. Mboga, C. Persello, J.R. Bergado, and A. Stein, "Detection of informal settlements from vhr images using convolutional neural networks," *Remote Sensing*, vol. 9, no. 11, pp. 1106, 2017.

[4] C. Persello and A. Stein, "Deep fully convolutional networks for the detection of informal settlements in vhr images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2325–2329, Dec 2017.

[5] D.A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *International Conference on Learning Representations*, 2016.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 3431–3440.

[7] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[9] Giuseppe Scarpa, Sergio Vitale, and Davide Cozzolino, "Target-adaptive cnn-based pansharpening," *CoRR*, vol. abs/1709.06054, 2017.