How Effective Is Super-Resolution to Improve Dense Labelling of Coarse Resolution Imagery?

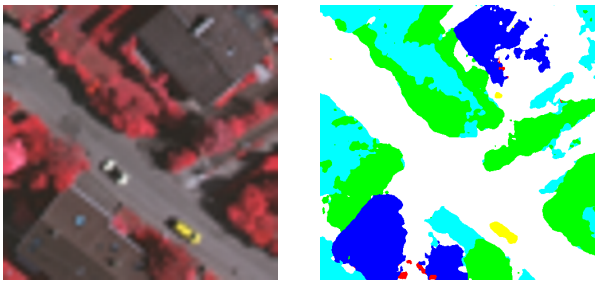
Matheus B. Pereira, Jefersson A. dos Santos
Department of Computer Science
Universidade Federal de Minas Gerais, Brazil
Belo Horizonte, Minas Gerais, 31270-901
Email: {matheuspereira, jefersson}@dcc.ufmg.br

***Abstract*—Coarse resolution remote sensing images, such as LANDSAT and MODIS are easily found in public open repositories and, therefore, are widely used in many studies. But their use for automatic creation of thematic maps is very restrict since most of the deep-based semantic segmentation (a.k.a dense labelling) approaches are only suitable for subdecimeter data. In this paper, we design a straightforward framework in order to evaluate the effectiveness of deep-based super-resolution in the semantic segmentation of low-resolution remote sensing images. We carried out an extensive set of experiments on three remote sensing datasets with distinct nature/properties. The results show that super-resolution is effective to improve semantic segmentation performance on low-resolution aerial imagery. It not only outperforms unsupervised interpolation but also achieves semantic segmentation results comparable to high-resolution data.**

I. Introduction

High-end satellites and drones are, nowadays, two of the main ways of directly acquiring high-resolution (HR) aerial images [1], which are important to monitor short and long-term changes and the impact of human activities on the environment [2]. In reality, however, HR image data is not always employable or accessible. In order to map a large area, for instance, drones lack enough autonomy: if using only one (or a few of them), the time required to map the whole area would be high, while using many of them would increase the cost and the number of people involved in the process. HR satellites provide a more autonomous way, therefore capable of overcoming the problems presented by drones, but their images are expensive and often present low temporal resolution. With all these issues in mind, an alternative for remote sensing applications is to get their data from cheap, low-resolution (LR) satellite imagery, which also usually has a long history of acquisition.

HR aerial images are essential for many remote sensing applications, as they provide a finer representation of spatial boundaries [3], more precise textures and can even display small objects that are barely visible in an LR representation. However, due to data unavailability or high-cost reasons, the use of LR images is often adopted in replacement of the HR ones. LANDSAT data, for example, is publicly available[1], with a long record of images since 1984. If, on the one hand, these data provide multispectral information, on the other hand, pattern recognition algorithms can have their performances compromised due to the lack of spatial information. As observed by [4], semantic segmentation is one of the computer vision applications that are more severely affected by the input of LR images. This application aims at predicting a class for each pixel of the image. Thus, if the objects of a class are way too small or have similar textures, the low-resolution will eventually cause many cases of mislabeling, dropping the accuracy of the algorithm. An example of this problem is shown in Figure 1.



(a) LR image (b) Semantic Segmentation

Fig. 1. Semantic segmentation (b) obtained from an LR (a) input image with a SegNet [5]. The image was up-sampled from its original 60×60 dimension to eight times more (480×480) with bicubic interpolation. Note that there are three cars (yellow class) in the image, but only one of them was correctly labeled by the network, while the other two could not be found. Also, it is possible to see that the texture of the buildings (dark blue class) was often mislabeled to impervious surfaces (white class).

Single image super-resolution (SISR) aims to construct an HR image from a single LR input and it is used as a major tool to restore the quality of degraded images. Lately, deep neural networks have been successfully applied to this problem with large improvements in accuracy. This makes SISR a viable choice for enhancing the performance of pattern recognition algorithms as a pre-processing step by restoring high-frequency details from LR inputs. A few works [4], [6], [7] have analyzed the effectiveness of SISR when applied with different tasks, however, most of them tackle the object detection problem, which needs less spatial information in comparison to semantic segmentation. Furthermore, to the best of our knowledge, only a few of them, such as [7], explore the application to aerial imagery, although not for semantic segmentation. Therefore, the contribution of our paper is the evaluation of the performance of a framework that

[1]https://www.usgs.gov/land-resources/nli/landsat

unites super-resolution and semantic segmentation in order to generate high-quality thematic maps from LR remote sensing imagery. We analyze the situations in which super-resolution considerably improves the generated thematic maps compared to the use of a bicubic interpolated LR input, and under which degradation factors the improvements are more relevant. We perform this evaluation on three remote sensing datasets with highly distinct properties.

The remainder of this paper is organized as follows. Section II presents the literature review. Section III details our methodology and the the proposed framework. Section IV and V describe and discuss the experimental analysis, respectively. Finally, Section VI concludes the paper.

II. Related Work

We organize this section into three parts. The two first are focused on the most relevant works for deep-based super-resolution and semantic segmentation tasks, respectively. The last part regards the works which proposed approaches that employ super-resolution as part of other image analysis problems, such as detection or scene recognition.

A. Super-resolution methods

The first deep-based super-resolution method (SRCNN) was proposed by Dong *et al.* [8], which is a convolutional neural network (CNN) capable of learning an end-to-end mapping between low and high-resolution images with only a few layers. In [9], the authors proposed the method named VDSR, which was the first one to use residual learning. Lim *et al.* [10] optimized the residual modules in existing conventional networks and proposed two methods, EDSR and MDSR, which ranked, respectively, first and second places on the NTIRE2017 Super-Resolution Challenge.

More recently, [11] were the winners of the $8\times$ classic bicubic-down-sampled track of the NTIRE2018 competition with their method named D-DBPN, which construct mutually-connected up and down-sampling stages. WDSR [12] is another method that achieved impressive results in the same competition, although in tracks in which the up-scaling factor is $4\times$, but that also include different types of degradation on the input images. This makes the super-resolution task more difficult.

Although less frequently, some papers have addressed the SISR problem for satellite imagery. Most of them only modify or apply similar networks to those already well stabilized on the literature. [13] proposed two methods based on EDSR [10] for the Sentinel-2 satellite. Another method [14], named DDRN, used a recursive strategy and ultra-dense-connections, similar to [11] on the Kaggle Open Source Dataset and Jilin-1 video satellite imagery.

B. Semantic segmentation methods

Like super-resolution, semantic segmentation also had the state-of-the-art changed by the introduction of deep learning. Long *et al.* [15] adapted classification networks, such as AlexNet, VGG net, and GoogLeNet into FCNs and fine-tuned their learned representations to the segmentation task.

Ronneberger *et al.* [16] proposed U-Net, a method that works with very few training images by relying on a strong use of data augmentation. Their method resembles an FCN, with the difference that they use skip connections to link low and high-level feature maps across resolutions.

As the previous works mainly focused on adapting deep architectures designed for classification to pixel-wise labeling, [5] proposed a method (Segnet) that was created to be more optimized for the semantic segmentation task, while also being efficient both in terms of memory and computational time. Segnet stores the max-pooling indices of the feature maps and uses them in its decoder network to achieve good performance, instead of storing the encoder network feature maps in full [5].

Semantic segmentation has also been applied on remote sensing data. Many methods base themselves on already consolidated networks and propose some modifications or adaptations. [17], for example, proposed methods based on fully convolutional networks [15], while [18] and [19] proposed methods based on Segnet [5].

C. Super-resolution for image analysis improvement

Despite the growing interest in super-resolution and semantic segmentation, no in-depth study has yet been made evaluating the performance of methods for both problems together. Dai *et al.* [4] evaluated super-resolution methods for other vision tasks, which were edge detection, semantic segmentation, digit recognition, and scene recognition. Their experiments showed that applying super-resolution to input images of other vision systems does improve their performance when the input images are of low-resolution. Although having a similar purpose to us, it is important to remark that not only they did not make an evaluation on aerial imagery, but they also applied methods for both super-resolution and semantic segmentation that are no longer close to the state-of-the-art.

Haris *et al.* [6] proposed a more elaborated framework, named Task-Driven Super-Resolution. This framework unifies both super-resolution and object detection tasks in one end-to-end training, which incorporates a tradeoff between detection and reconstruction losses. For this purpose, they used D-DBPN [11] for super-resolution and SSD [20] for object detection. Like in [4], the tests were not conducted on aerial images.

In [7], the authors used super-resolution to assist object detection performance in aerial imagery. They employed SSD [20] for detection and VDSR [9] for super-resolution. Shermeyer & Van Etten [7] verified that there is less gain at coarser resolutions and justified that this happens because the algorithm is unable to find enough unique discriminating features to adequately reconstruct an HR image.

III. Methodology

In this section, we present the proposed framework, which is composed of two main blocks: a super-resolution network (D-DBPN [11]) and a semantic segmentation network (Segnet [5]). The proposed framework is presented in Figure 2.

Fig. 2. Overview of the proposed framework. The pipeline is straightforward. First, an LR image, from which we desire to generate a thematic map, is processed by the super-resolution network. The output from this first step is a super-resolved version of the LR input that, ideally, has more details and helpful features for the following network. The second step consists of inputting the super-resolved image in the semantic segmentation network. The final output, therefore, is the thematic map classifying each pixel of the super-resolved image. As the resolution of this image is higher than the original input, the final thematic map should be more accurate than one generated by directly inputting the LR image into the semantic segmentation network.

Although we employed two specific methods for super-resolution and semantic segmentation, it is important to note that any other one (deep-based or not) with the same inputs and outputs could replace them according to the task or application. The choice for D-DBPN comes from the fact that this method was able to win the NTIRE2018 challenge on the high-scale restoration (up to 8 times) track. Being able to recover better details for such a high factor is especially important for aerial images, since they can present ground sample distance (GSD, which is the distance measured on the ground between two pixels) varying from a few centimeters (drone images, for example) to many meters (such as in the LANDSAT-8 satellite).

Regarding the existing methods for semantic segmentation, we selected Segnet [5] because it not only has been successfully used as a base approach for methods focused on aerial data [18], [19], but it also presents an efficient architecture in terms of memory and computational time. This is especially important for satellite imagery, which is usually composed of huge images that can easily excess hardware limits.

Differently from [6], our framework is not trained in an end-to-end manner. Thus, the two networks are trained separately, as detailed in the next two subsections. The main disadvantage of this approach is that it is not possible to use the semantic segmentation loss to bias the super-resolution network into creating an output that is more easily segmented by the other method. On the other hand, since the training of the semantic segmentation network is performed apart, any available data that does not have a corresponding thematic map can be used to train the super-resolution network. This is especially useful in the context of aerial imagery, which has less labeled data available when compared to normal images. This happens because the labeling of aerial data is usually more difficult and expensive, as many classes require the knowledge of a specialist, such as different types of vegetation. Creating data for the super-resolution network, however, is easier, as it is only necessary to degrade HR images and use them as input to the network. This way, we will not limit our super-resolution training only to labeled images for semantic segmentation. The code for the proposed framework is available at https://github.com/matheusbarrosp/sr-semseg-pytorch.

A. The super-resolution network

We employed D-DBPN [11] as the super-resolution network responsible for recovering the details of an LR input image. This method focuses on projecting the HR features back to the LR spaces using down-sampling layers. Each up and down-sampling stage represent different types of degradation and HR components [6].

In order to train the super-resolution network, we need pairs of corresponding low and high-resolution images. Two sensors of different image quality or in different heights can be used to get those pairs, but it is also possible to automatically generate the LR images by degrading the HR ones. Therefore, we follow the same approach used by the track in which D-DBPN [11] won the NTIRE2018 challenge and that was also used by [6] and [4], which is a bicubic kernel. Thus, we apply bicubic interpolation on the HR image with the desired down-scaling factor (4 and 8 times in our case). More complex degradation schemes may also be applied, such as in [7].

We use the same default network configuration proposed in the original paper [11] for D-DBPN. Thus, for $4\times$ enlargement we use 8×8 convolutional layer with four striding and two padding, while for $8\times$ enlargement we use 12×12 convolutional layer with eight striding and two padding. As the final network in [11], we set the number of back-projection stages to 7. We train the model for 300 epochs and randomly extract a 32×32 random patch for input from the low-resolution image on each iteration. The learning rate is initialized to $1e-4$ and is decayed by a factor of 10 at half of the total epochs. For optimization, we use Adam with 0.9 momentum and $1e-4$ weight decay. We refer to [11] for more details about the network.

B. The semantic segmentation network

The semantic segmentation network is responsible for classifying each one of the pixels from an input image in one of the possible classes from the problem it was trained for. A thematic map is the output of this process.

Segnet [5] has an encoder-decoder architecture that is followed by a pixelwise classification layer. The encoder network consists of the first 13 convolutional layers of the VGG16 network [21] for object classification. In order to train Segnet, we only use available HR data. On the other hand, during testing, we input the super-resolved images generated from the LR ones. The reason is that, sometimes, only a small amount of data is available for training, but we need to perform the segmentation on LR images. Thus, the idea is that it is possible to achieve better segmentation results by inputting a regenerated version of the LR images, instead of their original state.

The configuration of the Segnet in this work follows the same that was proposed in the original paper [5]. We train the model for 500 epochs with inputs of size 480×480 (as explained later in Section V). The learning rate is initialized to $1e-4$. We use Adam optimizer with 0.9 momentum and $5e-4$ weight decay. We also refer to [5] for more specific details related to the network.

IV. Experimental Setup

The objective of the experiments is to evaluate how well can super-resolution be used to improve semantic segmentation by using LR images as input. We simulate this situation by evaluating the proposed framework effectiveness with images in different levels of resolution degradation. Subsection IV-A presents the datasets we applied in the experiments. In subsection IV-B, we describe implementation details of the experimental protocol.

A. Datasets

In order to evaluate our framework, we selected three distinct remote sensing datasets. The first one is an agricultural dataset composed of scenes containing coffee and non-coffee areas. The second one is the Vaihingen dataset, provided by the International Society for Photogrammetry and Remote Sensing (ISPRS) Commission for the 2D Semantic Labeling Contest, which contains urban scenes with six different pixel classes. The third one is the 2014 IEEE GRSS Data Fusion Contest dataset, that also contains urban scenes and seven thematic classes.

1) Coffee Dataset: this dataset contains images from three different Brazilian cities from the state of Minas Gerais: Monte Santo, Guaranésia and Guaxupé. They are composed of green, red, and near-infrared bands with over 6000×10000 pixels each. Although having only 2 classes (coffee and non-coffee), this is a challenging dataset, as noted by [22], since it contains high intraclass variance, scenes with distinct plant ages and images with spectral distortions caused by shadows.
2) Vaihingen Dataset: this dataset contains 33 high-resolution images and six different thematic labels: impervious surfaces, building, low vegetation, tree, car, clutter/background. Similarly to the coffee dataset, the images are composed of near-infrared, red and green bands. From all the images, only 16 of them have ground-truth available for the semantic segmentation task. As mentioned in Section III, our framework pipeline allows us to train the super-resolution network with the non-labeled images, while only using the 16 labeled ones for the semantic segmentation task.
3) Thetford Dataset: this dataset contains one image (divided into two RGB sub-images) from an urban area near Thetford Mines (Quebec, Canada) and seven thematic labels: trees, vegetation, road, bare soil, red roof, gray roof, and concrete roof, along with the unclassified pixels.

The selected datasets fit well for our purpose since they present different characteristics that can be explored by the networks. The coffee dataset, for example, requires a lot of texture information to be able to distinguish coffee crops from non-coffee areas, while the urban datasets contain small objects (such as cars) that need to be visible enough for the network to classify their pixels correctly.

For the coffee dataset, we employed a protocol in which we train the networks on the images of two cities and test them on the remaining city. Thus, we trained the models on Montesanto and Guaxupé and tested on Guaranésia. By separating the images of a whole distinct city for test, we simulate a real-world scenario in which we have HR images taken from different cities, but need to apply the semantic segmentation on LR data from a new location. This case could not be reproduced with fidelity by simply selecting for test random crops of all the available data.

Labeled ground truth is provided for only one part of the Vaihingen dataset, since the remaining scenes were not released and require submission to the benchmark test organizers to be evaluated. Therefore, we trained and tested our framework using only the publicly available images. We applied the same division of the data as in [22]: areas 11, 15, 28, 30 and 34 are used as test, while the remaining areas are used as training and validation.

For the Thetford dataset, we use the same parts of the image selected by the contest for training and test. The original dataset contains seven classes, but one of them (bare soil, yellow label) is only present in the training part of the full data. As this sub-image is also used to train D-DBPN, it would not be fair to compare the performance with bicubic interpolation, since we would be super-resolving the same data used to train. Thus, we do not considerate the bare soil class in our results.

B. Implementation Details

We evaluate our method under two scaling factors of degradation: $4\times$ and $8\times$. Our objective is to compare how much the super-resolution network can help in each case. Intuitively speaking, low amounts of degradation ($\times 2$, for example) should not present enough difference from the LR image to compensate the need of a super-resolution network, while high amounts of degradation (such as more than $8\times$) should make it almost impossible to recover enough information when creating the super-resolved image.

The evaluation for super-resolution is different from semantic segmentation. We evaluate the quality of regenerated images in terms of Peak Signal-to-Noise Ratio (PSNR), which is the default metric for most of the super-resolution methods. Since two of the datasets we selected are not RGB, we evaluate the PSNR over all the three channels of the inputs (instead of only the Y channel of YCbCr). For the semantic segmentation, we selected four metrics: overall accuracy (acc), normalized accuracy ($norm.acc$), mean intersection over union (IoU) and Cohen's kappa coefficient ($Kappa$). These are the final values we use to evaluate how much the super-resolution interfered on the result, as the PSNR is just a measure of how well an image was regenerated.

We applied a similar experimental protocol for each one of the datasets. First of all, we divide the training and testing HR images in crops of size 480×480, from which we create the low-resolution inputs of size 120×120 and 60×60 for, respectively, $4\times$ and $8\times$ up-scaling factors. The choice for this dimension comes from the fact that the original Segnet paper [5] uses inputs of size 360×480. Furthermore, we separate 20% of the training data for validation. Due to the lack of

abundant training data in the Thetford dataset, we cropped the training images with 50% overlap, while also excluding the crops with more than 66% of the pixels labeled as unclassified. This left us with 53 images for training and 142 for test. The weights of D-DBPN are initialized with the pre-trained model provided by the original Github repository of the paper [11]. Similarly, Segnet is fine-tuned with the VGG16 trained weights.

Following the pipeline explained in Section III, we start off by training the super-resolution network, D-DBPN, for the desired up-scaling factor using the pair of original HR data and the generated LR images. Regardless of this step, we also need to train the semantic segmentation network with the available HR data and the thematic maps. During these two stages, we make sure the images used as test on the semantic segmentation task will not be included in the training set of the super-resolution network. We only want to apply super-resolution on the images that none of the networks saw during training, or else we would be creating a bias towards the reconstruction of the LR image.

After training both networks separately, the final evaluation is performed on the output of the semantic segmentation network for two versions of each image from the test set: the super-resolved images generated after the first step from the pipeline and the LR inputs, which are up-sampled back to their original size by using bicubic interpolation.

V. Results and Discussion

We conducted an extensive series of experiments in order to answer the following research questions: (1) How effective is deep-based super-resolution to different levels of degradation for remote sensing semantic segmentation tasks? (2) How deep-based super-resolution compares to classical unsupervised interpolation? (3) Is deep-based super-resolution able to reconstruct small object and, consequently, contribute to semantic segmentation improvement?

A. Effectiveness to different levels of degradation

Table I shows the performance of the proposed framework for different levels of resolution degradation in the task of semantic segmentation.

TABLE I
Semantic Segmentation Performance of the Proposed Framework for Different Degradation Factors and Ground Truth (gt)

Dataset	Deg.	Acc	Norm. acc	IoU	Kappa
Coffee	$8\times$	0.7715	0.7434	0.6046	0.4995
	$4\times$	0.7965	0.7672	0.6379	0.5516
	gt	0.8187	0.8062	0.6784	0.6128
Vaihingen	$8\times$	0.7447	0.5931	0.4762	0.6621
	$4\times$	0.7912	0.6369	0.5256	0.7234
	gt	0.8479	0.6833	0.5909	0.7984
Thetford	$8\times$	0.5444	0.6000	0.2916	0.4065
	$4\times$	0.7178	0.6665	0.4268	0.5897
	gt	0.8452	0.8184	0.6463	0.7636

As expected, the results show, for all datasets, that image resolution has a high impact on semantic targeting results. In general, the lower the resolution, the worse the result. However, the impact of the degradation rate impacts differently for each dataset.

Concerning coffee, the segmentation quality loss is relatively low for all metrics, ranging, for example, from mean 0.80 to 0.74 in the case of normalized accuracy from the original HR image to the same image with $8\times$ resolution degradation factor. It may indicate that for cropping the use of deep-based super-resolution could improve results. In the case of the urban datasets, the impact of the loss of resolution was greater than for coffee crops. Regarding the Thetford dataset, in particular, the normalized accuracy was reduced from 0.81 to 0.60 for $8\times$ degradation. For the Vaihingen dataset, the normalized accuracy was reduced from 0.68 to 0.59. The main explanation for the effect is that the Coffee dataset has only two classes and, in general, the coffee crops are relatively large areas. They also depend more on the texture than in the shape. In the case of the urban scenes, the accuracy was reduced mainly due to classes such as trees and cars that are composed of small regions which are difficult to recover given the strong loss of information.

Even though the urban datasets were more affected by the degradation than the coffee dataset, the difference for Thetford was way higher than for Vaihingen. This can be explained by the high amount of data that is available for the Vaihingen dataset, especially to train D-DBPN. As mentioned before, some images from the dataset were not labeled for semantic segmentation, but were used to train the super-resolution network. The consequence for this is that the quality of the reconstructed images is much higher compared to Thetford (which contains a small quantity of training data), thus helping more the semantic segmentation task. This indicates that deep-based super-resolution can increase the semantic segmentation results relatively close to a native HR data given enough training.

B. Comparison to bicubic interpolation

Table II presents results for semantic segmentation by using bicubic interpolation and super-resolution with D-DBPN. It also reports the reconstruction rate with PSNR.

Regarding the reconstruction, as expected, the PSNR is higher when applying D-DBPN as an up-scaling method instead of a simple bicubic interpolation. This means that the super-resolved output contains more visually appealing, high-frequency details than an interpolated image.

As Table II shows, the use of super-resolution improved the results of all the metrics for all datasets and degradation factors. This is especially true for higher degradation factors ($8\times$), since the loss of information is more considerable and, thus, deep-based super-resolution can learn to recover the details much better than a simple interpolation. As the semantic segmentation metrics are better with higher PSNR values, we can also verify that this metric correlates well with how better a reconstructed image can be segmented.

TABLE II
COMPARISON BETWEEN THE PERFORMANCE OF BICUBIC INTERPOLATION AND SUPER-RESOLUTION BY D-DBPN.

Dataset	Deg.	Method	PSNR (dB)	Norm. acc	Kappa	IoU
Coffee	4×	Bicubic	24.2429	0.5396	0.0962	0.3654
		D-DBPN	25.7454	0.7672	0.5516	0.6379
	8×	Bicubic	20.7929	0.5003	0.0007	0.3134
		D-DBPN	21.2265	0.7434	0.4995	0.6046
Vaihingen	4×	Bicubic	28.7458	0.5741	0.6417	0.4526
		D-DBPN	31.1974	0.6369	0.7234	0.5256
	8×	Bicubic	25.3886	0.4747	0.5281	0.3449
		D-DBPN	27.4540	0.5931	0.6621	0.4762
Thetford	4×	Bicubic	26.8292	0.5776	0.4271	0.2906
		D-DBPN	31.0294	0.6665	0.5897	0.4268
	8×	Bicubic	23.3354	0.4660	0.1672	0.1408
		D-DBPN	26.3173	0.6000	0.4065	0.2916

The big difference in the semantic segmentation metrics shows that super-resolution allowed the network to predict with more precision the classes of each dataset when compared to interpolated LR inputs. This is another clear evidence of the capability of super-resolution to recover important visual details for a semantic segmentation algorithm. The most impressive improvement can be noted in the coffee dataset results. Using bicubic interpolation, the normalized accuracy is close do 50%, the $Kappa$ values are close to zero and the IoU is way smaller than employing D-DBPN. This happened because the loss of texture information was so severe, that Segnet was incapable of detecting most of the coffee areas, thus classifying almost 100% of the images as non-coffee. Employing super-resolution, on the other hand, allowed the coffee areas to be segmented with much more precision.

Another important improvement can be noted for the Thetford dataset. Being able to increase the performance with deep-based super-resolution even when it contains a small amount of training data (a bad scenario for a deep learning algorithm), shows that this approach is more reliable than interpolation.

C. Robustness to small object segmentation

In order to verify the effectiveness of the proposed framework in the segmentation of small objects, we analyzed the results obtained by class for datasets Vaihingen and Thetford. It can be seen in Figures 3 and 4, respectively. Visual segmentation results are shown in Figures 5 and 6.

For the Vaihingen dataset, one can note in Figure 3 that super-resolution greatly improved the segmentation of the car class: for $8\times$ degradation, bicubic up-sampled inputs could predict correctly only 19% of the car pixels, while super-resolved inputs increased this value to 58% (only 11% less than high-resolution inputs). This confirms that super-resolution is capable of turning visible (for a machine) objects that are too small in an LR representation. The results also present a great improvement for the building class, which was highly mislabeled as impervious surfaces by the LR representation, but that was better segmented on super-resolved inputs, increasing the value from 31% to 68% for $8\times$ degradation. In this dataset, super-resolution lost to interpolation only on the tree class (5% on $\times 8$ degradation). However, looking at the low values of accuracy and IoU presented in Table II, the reason for this is that the network simply classified a higher amount of pixels as tree, what increases the chances of predicting this class, but causes many cases of mislabeling for the other labels. This is true especially for low vegetation, which was mislabeled as tree only 9.2% of the time on super-resolved inputs, but 16% on bicubic up-sampled images. We point out that even though this dataset contains six classes, we excluded from the heatmaps the clutter/background (red label) results. The reason for this is that the class represents less than 1% of the dataset (being highly mislabeled even with high-resolution inputs), since it is designated to unclassified or rejected objects on the scene, and is also not considered in some similar works, such as the ones proposed in [22], [23].

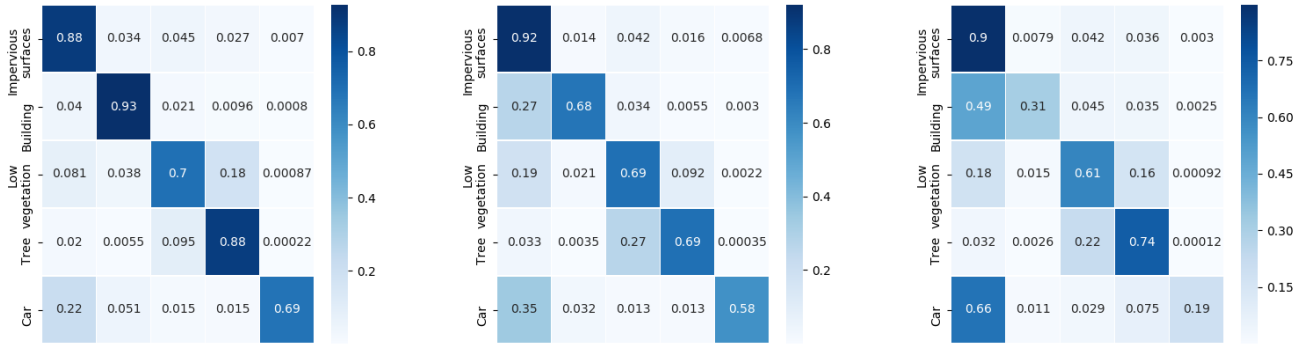
For the Thetford dataset, applying semantic segmentation on high-resolution data resulted in accurate predictions, which can be confirmed by the high diagonal numbers in Figure 4. Super-resolution and bicubic interpolation, on the other hand, mislabeled trees by vegetation in almost all cases. However, high super-resolution improvements can be seen especially on the road and grey roof classes. While interpolation mislabeled roads in 97% of the cases, super-resolution decreased this error to 50% on $8\times$ up-scaling. For grey roofs, super-resolution improved the precision from 4% to 28%. Like in the Vaihingen Dataset, bicubic interpolation performing better on some classes can be explained by the higher amount of pixels labeled as them, which decreases the precision on different labels and is reflected in the small numbers from Table II.

VI. CONCLUSION

In this paper, we presented a framework that generates more accurate semantic segmentation thematic maps for LR remote sensing inputs with the employment of super-resolution. We evaluated its performance on three highly different aerial datasets and under two degradation factors, comparing the results with LR bicubic up-sampled inputs and native HR data.

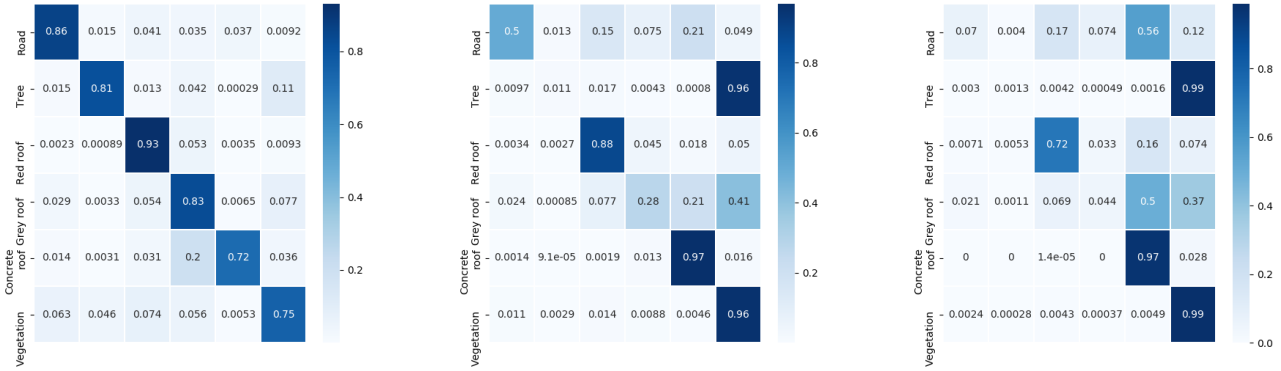
Super-resolution was confirmed to be a viable strategy to recover important texture and object details for semantic segmentation. With enough training data, the recovered texture information greatly helps the semantic segmentation not to mislabel similar classes. Small objects, such as the cars in the Vaihingen dataset, which are not detected in LR representations, can become visible with the employment of super-resolution.

For either quantitative or qualitative evaluation, super-resolved inputs surpassed the LR bicubic up-sampled ones in all cases. This improvement was more significant on $8\times$ down-sampling factors, since the amount of information loss in the LR representation is vast enough to negatively affect the generation of thematic maps, but also small enough to allow the reconstruction by the super-resolution network. Also, deep-based super-resolution allows the performance of the semantic



(a) Original HR Image (b) 8× D-DBPN reconstrution (c) Bicubic interpolation

Fig. 3. Accuracy heatmaps (confusion matrix) for semantic segmentation on the Vaihingen dataset with 8× up-scaling.



(a) Original HR Image (b) D-DBPN reconstrution (c) Bicubic interpolation

Fig. 4. Accuracy heatmaps (confusion matrix) for semantic segmentation on the Thetford dataset with 8× up-scaling.

segmentation to get close to HR images under proper training conditions.

For future work, the employment of an end-to-end framework which trains both networks at the same time, and while sharing their losses (similarly to the task-driven approach proposed in [6]), can bring more improvements to the semantic segmentation task, but with the drawback of not being able to train the super-resolution network with unlabeled data for semantic segmentation by normal means. Also, there is space to study how different visual benchmarks can help to enhance the semantic segmentation performance more than PSNR, such as adversarial losses from GANs.

Acknowledgment

The authors would like to thank NVIDIA for the donation of the GPUs that allowed the execution of all experiments in this paper. We also thank CAPES, CNPq, and FAPEMIG for the financial support provided for this research project.

References

[1] V. V. Klemas, "Coastal and environmental remote sensing from unmanned aerial vehicles: An overview," *Journal of Coastal Research*, vol. 31, no. 5, pp. 1260–1267, 2015.

[2] R. A. Schowengerdt, *Remote sensing: models and methods for image processing*. Elsevier, 2006.

[3] D. Pouliot, R. Latifovic, J. Pasher, and J. Duffe, "Landsat super-resolution enhancement using convolution neural networks and sentinel-2 for training," *Remote Sensing*, vol. 10, no. 3, 2018.

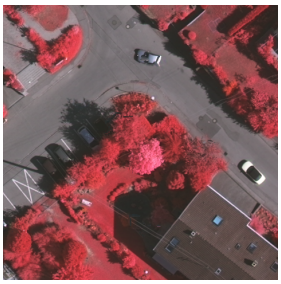
[4] D. Dai, Y. Wang, Y. Chen, and L. Van Gool, "Is image super-resolution helpful for other vision tasks?" in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–9.

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, 2017.

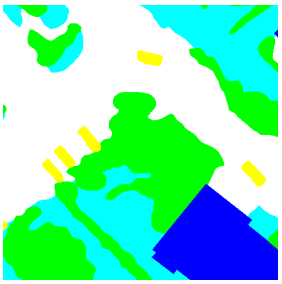
[6] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," *arXiv preprint arXiv:1803.11316*, 2018.

[7] J. Shermeyer and A. Van Etten, "The effects of super-resolution on object detection performance in satellite imagery," *arXiv preprint arXiv:1812.04098*, 2018.

[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

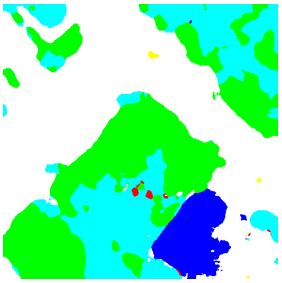


(a) Original HR image

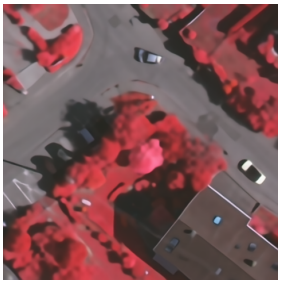


(b) Ground truth

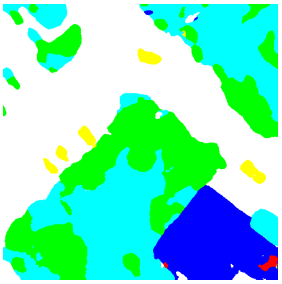
(c) 8× bicubic interpolation



(d) 8× bicubic + segmentation



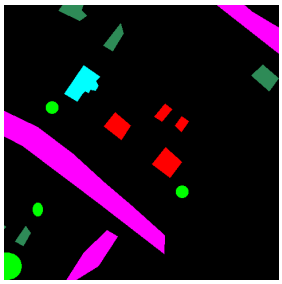
(e) 8× D-DBPN reconstrution



(f) 8× D-DBPN + segmentation

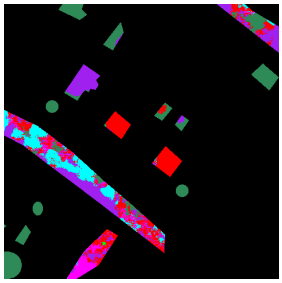
Fig. 5. Semantic segmentation results for the Vaihingen dataset

(a) Original HR image



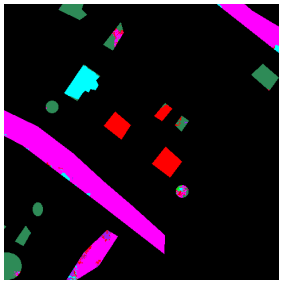
(b) Ground truth

(c) 8× bicubic interpolation



(d) 8× bicubic + segmentation

(e) 8× D-DBPN reconstruction



(f) 8× D-DBPN + segmentation

Fig. 6. Semantic segmentation results for the Thetford dataset

[9] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[10] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*, vol. 1, no. 2, 2017, p. 4.

[11] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep backprojection networks for super-resolution," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, and T. Huang, "Wide activation for efficient and accurate image super-resolution," *arXiv preprint arXiv:1808.08718*, 2018.

[13] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler, "Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network," *arXiv preprint arXiv:1803.04271*, 2018.

[14] K. Jiang, Z. Wang, P. Yi, J. Jiang, J. Xiao, and Y. Yao, "Deep distillation recursive network for remote sensing imagery super-resolution," *Remote Sensing*, vol. 10, no. 11, 2018.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, 2015, pp. 234–241.

[17] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," *arXiv preprint arXiv:1606.02585*, 2016.

[18] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Computer Vision – ACCV 2016*. Springer International Publishing, 2017, pp. 180–196.

[19] D. Marmanis, K. Schindler, J. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 158 – 172, 2018.

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 21–37.

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

[22] K. Nogueira, M. Dalla Mura, J. Chanussot, W. R. Schwartz, and J. A. dos Santos, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–18, 2019.

[23] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution aerial image labeling with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7092–7103, 2017.