Mapping the Unseen: Exploiting Super-Resolution for Semantic Segmentation in Low-Resolution Images

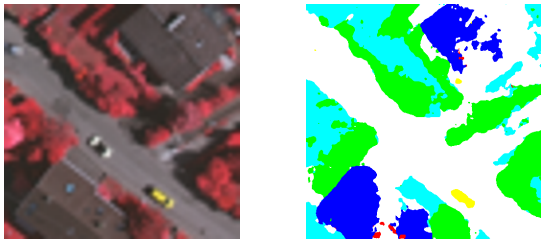
Matheus B. Pereira, Jefersson A. dos Santos
Department of Computer Science
Universidade Federal de Minas Gerais, Brazil
Belo Horizonte, Minas Gerais, 31270-901
Email: {matheuspereira, jefersson}@dcc.ufmg.br

***Abstract*—High-resolution aerial images are usually not accessible or affordable. On the other hand, low-resolution remote sensing data is easily found in public open repositories. The problem is that the low-resolution representation can compromise pattern recognition algorithms, especially semantic segmentation. In this M.Sc. dissertation[1], we design two frameworks in order to evaluate the effectiveness of super-resolution in the semantic segmentation of low-resolution remote sensing images. We carried out an extensive set of experiments on different remote sensing datasets. The results show that super-resolution is effective to improve semantic segmentation performance on low-resolution aerial imagery, outperforming unsupervised interpolation and achieving semantic segmentation results comparable to high-resolution data.**

I. Introduction

High-resolution (HR) aerial images are essential for many remote sensing applications, as they provide a finer representation of spatial boundaries [1], more precise textures and can even display small objects that are barely visible in a low-resolution (LR) representation. High-end satellites and drones currently are two of the main ways of acquiring HR data. In reality, however, this type of data is not always employable or accessible: drones lack autonomy and are not suitable for large scale problems, and data from HR satellites is expensive while often presenting low temporal resolution. Thus, relying on these options is often impracticable.

Due to data unavailability or high-cost reasons, the use of LR images is often adopted in replacement of the HR ones. An alternative for remote sensing applications is to get their data from LR satellite imagery, which is cheap (or free) and present a long history of acquisition. But a main problem arises from the use of LR images: the amount of important information compressed into one single pixel can compromise machine learning algorithms to detect or segment objects. As observed by [2], semantic segmentation is one of the computer vision applications that is more severely affected by the input of LR images. If the objects are way too small or have similar textures the low-resolution may cause cases of mislabeling, dropping the accuracy of the algorithm. In Figure 1 we can see an example of this situation: there are three cars (yellow class), but only one of them was correctly labeled by a semantic segmentation network. Also, it is possible to see that big parts of the buildings (dark blue class) were mislabeled for impervious surfaces (white class).



(a) LR image (b) Predicted Thematic Map

Fig. 1. Thematic map (b) generated from an LR (a) input image (up-sampled to eight times more with bicubic interpolation) with a SegNet [3].

Considering this situation, a natural question to arise is: how can we effectively reconstruct low-resolution remote sensing imagery in order to improve semantic segmentation? In this work, we provide ways of achieving this goal with the use of super-resolution.

Single image super-resolution (SR) aims to construct a HR image from a single LR input. In this work, we study the performance of two frameworks that unite SR and semantic segmentation methods to generate high-quality thematic maps for LR remote sensing images. The first one is a two stage framework that uses SR as a pre-processing step for a semantic segmentation task, training both networks separately. The second one is an end-to-end framework that trains both networks at the same time.

The remainder of this document is structured as follows. Section II presents similar works that evaluated SR for the improvement of different pattern-recognition tasks. Section III introduces the proposed frameworks. Section IV presents the experimental setup and datasets. In Section V we show and discuss the results. In Section VI we conclude our work and discuss future possibilities.

[1]This work relates to a M.Sc. dissertation.

II. Related Work

Despite the growing interest in SR and semantic segmentation, almost no study has yet been made evaluating the performance of methods for both problems together. [2] evaluated SR methods for several vision tasks: edge detection, semantic segmentation, digit recognition, and scene recognition. Their experiments showed that applying SR to input images of other vision systems improves their performance when the input images are of low-resolution and that standard perceptual criteria used to evaluate SR methods (such as PSNR) correlate quite well with the usefulness for the vision tasks. Although having a similar purpose to us, not only they did not evaluate on aerial imagery, but they also applied methods for SR and semantic segmentation that are no longer close to the state-of-the-art, all in a superficial manner.

[4] proposed a framework that unifies SR and object detection tasks in an end-to-end training, which incorporates a trade-off between detection and reconstruction losses. They used D-DBPN [5] for SR and SSD [6] for object detection. In terms of detection performance, their results surpassed the mean average precision (mAP) of bicubic interpolation more than three times for the test without blur or noise and up to around eight times for the test with noise on $8\times$ up-scaling. Like in [2], the tests were not conducted on aerial images.

In [7] and [8], the authors used SR to assist object detection performance in aerial imagery. Both used SSD [6] for detection, but while the former applied SRGAN [9] as the SR method, the latter chose VDSR [10]. [8] verified that super-resolving native $30cm$ ground sampling distance (GSD) imagery to $15cm$ yields a 16 to 20% improvement in detection mAP on the xView Dataset. [7] applied their tests on the VEDAI dataset and achieved around 27% more mAP in $4\times$ up-scaling compared to an LR input.

Another recent work has evaluated the use of SR for the improvement of semantic segmentation on remote sensing imagery [11]. They use ESPCN [12] for SR and U-Net [13] for semantic segmentation. This method inputs in the testing phase a panchromatic image from a different sensor of lower resolution. This work applies SR as a pre-processing step for semantic segmentation and trains the semantic segmentation network with HR images. In our work, we also study the case in which we do not have access to HR data even for training. Moreover, our frameworks do not require a panchromatic image as input.

III. Methodology

In this section, we introduce the two frameworks we have proposed in this work. Both of them are composed of two main blocks: a SR network and a semantic segmentation network. The first framework uses SR as a pre-processing step for semantic segmentation, training the two networks separately. The second framework is an end-to-end approach that trains both networks at the same time, taking into consideration the loss of the two tasks. We first introduce each individual network (for SR and semantic segmentation) and then we present the proposed frameworks.

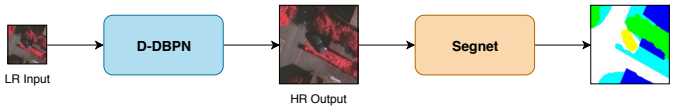


Fig. 2. Overview of the two stage framework, which applies SR on LR images before sending them to a semantic segmentation network.

A. The Super-resolution Network

D-DBPN [5] was the chosen SR network to be employed in the frameworks. The main characteristic of D-DBPN is the error feedback mechanism, in which the method projects the HR features back to the LR spaces using down-sampling layers [5]. This allows the network to guide the image reconstruction by calculating the projection error from the up and down-sampling blocks. The different ways of projecting back to another LR representation enrich the knowledge of the network, which learns different ways of up-sampling the features. We refer to the original paper [5] for further details.

In order to train the SR network, we need pairs of corresponding low and high-resolution images. It is possible to automatically generate LR images by degrading the HR ones. We use the same default network configuration proposed in D-DBPN's original paper [5] in terms of kernel size, striding, padding and number of back-projection stages. This network is trained with the L1 loss.

B. The semantic segmentation network

For the semantic segmentation task, we employ SegNet [3], which 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 [14] for object classification. Each decoder layer has a corresponding encoder from which it receives max-pooling indices to perform non-linear upsampling of their input feature maps [3]. We also refer to the original paper [3] for further details. The training of the Segnet is performed with the use of a pixelwise cross-entropy loss.

C. The Two Stage Framework

The two stage framework is presented in Figure 2. It was first used similarly in [2], but, as explained in Section II, the evaluation was highly superficial and not for remote sensing images. The pipeline of such framework is straightforward. First, an LR image, from which we desire to generate a thematic map, is processed by the SR network. The output from this first step is a reconstructed version of the LR input. 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 this image. As the resolution and quality of the reconstructed image are 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 D-DBPN and SegNet, any other method with the same input and output configuration could replace them according to the task or preference.

As mentioned, the two networks are trained separately. The main disadvantage of this approach is that it is not possible to use the semantic segmentation loss to bias the SR 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 SR network. This is especially useful in the context of aerial imagery, which has less labeled data available when compared to normal images.

For D-DBPN, we train the model for 300 epochs and randomly extract a 32×32 random patch for input from the LR 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.

For SegNet, we also follow the same approach that was proposed in its original paper [3]. We train the model for 500 epochs with inputs of size 480×480. The learning rate is initialized to $1e-4$. We use Adam optimizer with 0.9 momentum and $5e-4$ weight decay. Also, in order to train SegNet in this framework, we only use available HR data. However, during testing, we input the reconstructed images generated from the LR ones. The motivation for this is that in many cases, only a few amounts of data are available for training, but in practice, we often need to perform semantic segmentation on LR images. Our objective is to demonstrate that it is possible to achieve more accurate semantic segmentation results by inputting a reconstructed version of the LR images, instead of the degraded images themselves.

D. The End-to-end Framework

The end-to-end framework is capable of training the SR and the semantic segmentation network at the same time. This is an interesting approach as it allows the semantic segmentation network to guide the SR reconstruction in a way that is more beneficial for its own vision. When using the two stage framework, the SR method does not take anything into consideration apart from the network's loss and the quality of reconstruction criterion (PSNR). By allowing the semantic segmentation loss to be also used in the training procedure together with the SR loss, we are letting it bias the reconstruction procedure in a way that makes the image features more easily segmented.

Our proposed framework for this case is based on the task-driven architecture proposed in [4] and can be seen in Figure 3. It works as follows: first, the LR input image is sent to the framework, where it will first be processed by the SR module. The result of this process will be a super-resolved image that will be used both to calculate the $L1$ loss (SR loss) and as input to the semantic segmentation part. After being processed by the SegNet, the final output of the framework will be an HR thematic map made from the LR input. This thematic map will also serve to calculate the semantic segmentation loss (cross-entropy). The unified loss (ξ) of the framework is calculated

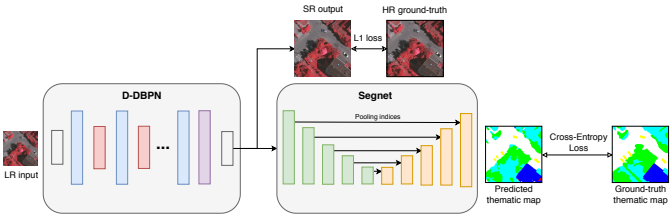


Fig. 3. Overview of the end-to-end framework, which trains the SR and semantic segmentation networks at the same time.

as in Equation 1, similarly to how [5] applied it to the object detection task.

$$\xi = \alpha L1(I_{HR}, SR(I_{LR})) + \beta Ce(y_{HR}, Seg(SR(I_{LR}))), \quad (1)$$

where $L1(.)$ represents the SR loss, and $Ce(.)$ the cross-entropy loss for semantic segmentation. I_{HR} and I_{LR} represent, respectively, the HR ground truth image and the LR input. y_{HR} is the ground-truth thematic map. $SR(.)$ and $Seg(.)$ are, respectively, the SR and semantic segmentation networks. Finally, α and β are pre-defined values that represent the balance between the SR and semantic segmentation losses.

The definition of the α and β values is the key to defining how biased the outputs will be for human or machine perception. With an α value higher than β, the network will prioritize the SR reconstruction over the result of the semantic segmentation. However, by setting a β value higher than α, the framework will penalize more the semantic segmentation error and consider less how the image reconstruction is being performed.

This framework is trained for 300 epochs with inputs of size 480×480. The learning rate is initialized to $1e-5$ and is decayed by a factor of 10 at half of the total epochs. Each individual network inside the framework (D-DBPN and SegNet) is optimized under the same conditions as they do in the two stage framework.

IV. Experimental Setup

A. Datasets

In order to evaluate our frameworks, we selected three distinct remote sensing datasets. The first one is the Brazilian Coffee Scenes Dataset [15], which is an agricultural dataset composed of scenes containing coffee and non-coffee areas from three cities located in Minas Gerais (Brazil): Guaranésia, Guaxupé and Monte Santo de Minas. 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 (Thetford dataset), that also contains urban scenes and seven thematic classes.

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. Later, the results for this dataset will be

reported in terms of the mean and standard deviation of these three cases.

Labeled ground truth is provided for only one part of the Vaihingen dataset. Thus, we trained and tested our framework using only the publicly available images. We applied the same division of the data as in [16]: areas 11, 15, 28, 30 and 34 are used as test, while the remaining areas are seen during training. We excluded from the results the clutter/background class, since it represents less than 1% of the dataset and is designated to unclassified or rejected objects on the scene.

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) is only present in the training part of the full data. Thus, we do not considerate this class in our results, as it takes part in the SR training set.

B. Implementation Details

We evaluate our frameworks under two scaling factors of degradation: $4\times$ and $8\times$. Our objective is to compare how much the SR network can help in each case.

We evaluate the quality of regenerated images in terms of Peak Signal-to-Noise Ratio (PSNR). We evaluate the PSNR over all the three channels of the inputs. For the semantic segmentation, we used four metrics: pixel accuracy (acc), normalized accuracy ($norm.acc$), mean intersection over union (IoU) and Cohen's Kappa coefficient ($Kappa$).

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 LR inputs. To create the LR images, we follow the same approach used by [4] (on their first experiment), [2] and [7]: we apply bicubic interpolation on the HR image with the desired down-scaling factor.

The weights of D-DBPN are initialized with the pre-trained model provided by the original Github repository of the paper [5]. Similarly, Segnet is fine-tuned with the VGG16 trained weights for image classification. For the end-to-end framework, we also initialize each one of these corresponding blocks in the same way.

For the two stage framework, we start by training the SR network for the desired up-scaling factor using the pairs of original HR data and the generated LR images. During this stage we can also use data that does not contain a corresponding thematic map in order to improve the SR training. We also train the semantic segmentation network with the available HR data and the thematic maps. This procedure works as a simulation of a real-world scenario, in which we have already trained the models with the few available HR data and now we need to apply the semantic segmentation on a new LR image. The final evaluation is performed on the output of the semantic segmentation network.

For the training of the end-to-end framework, the SR and semantic segmentation networks are trained together in a single step. Thus, unlabeled data for semantic segmentation will not be used in the framework. Differently from the two stage framework, the end-to-end framework simulates the whole process of training and testing already expecting an LR image as input. Therefore, this framework is less versatile than the previous (detaching the semantic segmentation block will not allow it to perform well in HR data), but it is more powerful on LR images.

We experimented with many different losses configurations for the end-to-end framework by changing the α and β values of Equation 1. As our objective is to improve semantic segmentation results, we aim for higher β values. We tested the framework on the Vaihingen dataset with α values from the set $\{0.001, 0.1, 1\}$, and β values from the set $\{1, 10, 100, 1000, 10000, 100000\}$. We observed that lower β values achieved results that were similar to the two stage framework and that the higher the β values, the higher was the number of artifacts created in the reconstructed image. Under these circumstances, the best results were achieved with the $0.1/1000$ configuration for α and β, respectively. For now on, the results reported next for the end-to-end framework are all using this same configuration.

V. Results and Discussion

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

A. Effectiveness to different levels of degradation

Table I presents the performance of the two stage framework with $4\times$ and $8\times$ degradation factor.

TABLE I
Semantic Segmentation Performance of the Two Stage Framework for Different Degradation Factors

Dataset	Deg.	Acc	Norm. acc	IoU	Kappa
Coffee	8×	0.763 ± 0.011	0.720 ± 0.030	0.581 ± 0.030	0.463 ± 0.047
	4×	0.802 ± 0.005	0.772 ± 0.003	0.645 ± 0.006	0.562 ± 0.009
	1×	0.833 ± 0.013	0.816 ± 0.009	0.697 ± 0.017	0.639 ± 0.024
Vaihingen	8×	0.744	0.593	0.476	0.662
	4×	0.791	0.636	0.525	0.723
	1×	0.847	0.683	0.590	0.798
Thetford	8×	0.544	0.600	0.291	0.406
	4×	0.717	0.666	0.426	0.589
	1×	0.845	0.818	0.646	0.763

The results show for all datasets that image resolution has a high impact on semantic targeting results. 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. It is an indication that for cropping, the use of deep-based SR can improve the results. In the case of the urban datasets, the impact of the loss of resolution was greater than for coffee crops. 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. 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 that are difficult to recover given the strong loss of information. The difference for Thetford was higher than for Vaihingen because of the high amount of data that is available for the Vaihingen dataset, especially to train D-DBPN. This indicates that deep-based SR can increase the semantic segmentation results relatively close to a native HR data given enough training.

TABLE II
SEMANTIC SEGMENTATION PERFORMANCE OF THE END-TO-END FRAMEWORK FOR DIFFERENT DEGRADATION FACTORS

Dataset	Deg.	Acc	Norm. acc	IoU	Kappa
Coffee	$8\times$	0.800 ± 0.025	0.778 ± 0.024	0.647 ± 0.034	0.565 ± 0.052
	$4\times$	0.820 ± 0.012	0.809 ± 0.009	0.680 ± 0.014	0.616 ± 0.020
	$1\times$	0.833 ± 0.013	0.816 ± 0.009	0.697 ± 0.017	0.639 ± 0.024
Vaihingen	$8\times$	0.828	0.662	0.565	0.773
	$4\times$	0.829	0.663	0.569	0.773
	$1\times$	0.847	0.683	0.590	0.798
Thetford	$8\times$	0.860	0.856	0.698	0.788
	$4\times$	0.873	0.841	0.711	0.799
	$1\times$	0.845	0.818	0.646	0.763

Table II shows the semantic segmentation results of the end-to-end framework. It is possible to see that the impact of $8\times$ degradation factor while using the end-to-end approach is not as considerable compared to $4\times$ as in the two stage framework. In the Thetford dataset, the normalized accuracy was even higher when inputting $8\times$ degraded images, while the remaining metrics also stood close. This indicates that the end-to-end framework is more capable of dealing with higher degradation factors without losing too much semantic segmentation accuracy. This is due to the fact that this framework can change the reconstructed image with information that is more easily discernible for the semantic segmentation task. When relying only on the SR loss, there is no assurance that similar textures will be reconstructed in a way that highlights the differences among them. By letting the semantic segmentation task guide the super-resolution, we are allowing this highlighting to occur automatically. Another important reason that makes the end-to-end framework perform better is that it is trained with LR data as input, while the Segnet of the two stage framework is trained with HR data. Therefore, the difference in the degrading factors impacts more a network trained with HR data only than a network trained with that specific degradation as input.

We can also see that even the difference to native HR data ($1\times$ degradation) is smaller in the end-to-end framework. The most interesting and noticeable change is in regard to the Thetford dataset. The end-to-end framework managed to achieve better results with LR images than the semantic segmentation trained on HR data. One of the reasons that made this happen is the low amount of training data for this dataset. The lack of training images compromised the SegNet to differentiate similar classes and deal with the intra-class variance. The framework is capable of differentiating these cases due to the way it is trained. Also, considering that the low-resolution aspect diminishes the intra-class variance, the results of the framework ended up being better than expected.

What the results show is that the end-to-end framework is more effective than applying SR as a pre-processing step for semantic segmentation trained with HR data. This was an expected conclusion, since the semantic segmentation network of the two stage framework is trained with HR data and, therefore, is not being tested with the same resolution. This does not change the fact that SR can indeed improve the results for LR data when the training is performed on HR images.

B. Comparison to bicubic interpolation

TABLE III
COMPARISON BETWEEN THE PERFORMANCE OF BICUBIC INTERPOLATION AND THE PROPOSED FRAMEWORKS.

Dataset	Deg.	Method	PSNR (dB)	Norm. acc	Kappa	IoU
Coffee	$4\times$	Bicubic	25.209 ± 0.797	0.566 ± 0.021	0.161 ± 0.050	0.401 ± 0.027
		Two stage	27.278 ± 1.114	0.772 ± 0.003	0.562 ± 0.009	0.645 ± 0.006
		End-to-end	26.055 ± 0.369	0.809 ± 0.009	0.616 ± 0.020	0.680 ± 0.014
	$8\times$	Bicubic	21.660 ± 0.694	0.501 ± 0.001	0.003 ± 0.001	0.317 ± 0.003
		Two stage	22.833 ± 1.182	0.720 ± 0.030	0.463 ± 0.047	0.581 ± 0.030
		End-to-end	21.272 ± 1.349	0.778 ± 0.024	0.565 ± 0.052	0.647 ± 0.034
Vaihingen	$4\times$	Bicubic	28.745	0.574	0.641	0.452
		Two stage	31.197	0.636	0.723	0.525
		End-to-end	26.369	0.663	0.773	0.569
	$8\times$	Bicubic	25.388	0.474	0.528	0.344
		Two stage	27.454	0.593	0.662	0.476
		End-to-end	22.619	0.662	0.773	0.565
Thetford	$4\times$	Bicubic	26.829	0.577	0.427	0.290
		Two stage	31.029	0.666	0.589	0.426
		End-to-end	29.818	0.841	0.799	0.711
	$8\times$	Bicubic	23.335	0.466	0.167	0.140
		Two stage	26.317	0.600	0.406	0.291
		End-to-end	25.592	0.856	0.788	0.698

Table III presents the results for semantic segmentation by using bicubic interpolation and the proposed frameworks. It also reports the reconstruction rate with PSNR. As the table shows, the use of SR improved the semantic segmentation results of all the metrics for all datasets and degradation factors. An important improvement can be noted for the Thetford dataset. Being able to increase the performance with deep-based SR even when it contains a small amount of training data shows that our frameworks are more reliable than interpolation.

Regarding the reconstruction, the PSNR is higher when applying D-DBPN directly as an up-scaling method (as in the two stage framework) instead of a simple bicubic interpolation. This means that the super-resolved output contains more visually appealing, high-frequency details than an interpolated image. However, the PSNR in the end-to-end approach does not follow the same pattern. Even though the reconstruction metric is not as high as the other options in some cases, the semantic segmentation results are better. That happens because the supervision of the semantic segmentation network in the training of the SR method allows the framework to change the visual characteristics of the reconstructed image. This makes the PSNR drop since the SR output will present details that are inexistent in the ground-truth HR image, but those details are exactly what makes the performance of the SegNet improve.

C. Robustness to small object segmentation

In order to verify the effectiveness of the two frameworks in the segmentation of small objects, we analyzed the accuracy obtained by class for the Vaihingen dataset. Visual results for SR and semantic segmentation are shown in Figure 4.

Considering the car class, a Segnet trained with HR data achieves 69% accuracy. The end-to-end framework managed to stay close to this value even under $8\times$ degradation: it achieved 65%. The use of LR data compromises a lot the performance: bicubic up-sampled inputs could predict correctly only 19% of the car pixels. The two stage framework increased this value to 58%. This confirms that SR and both frameworks are capable of making more discernible objects that are too small in an LR representation. In Figure 4 we can see an example of a segmentation that missed most of the car information due to the LR representation, but that was successfully recovered with the use of both frameworks. The results also presented a great improvement for the building class, which achieved only 31% accuracy with LR inputs, but 68% and 89% with the two stage and end-to-end frameworks, respectively (HR inputs achieved 93%).

Finally, by observing the visual results of the reconstructed images from the end-to-end framework in Figure 4, it is possible to see the different textures employed by the semantic segmentation network that helped it to more accurately classify the pixels.

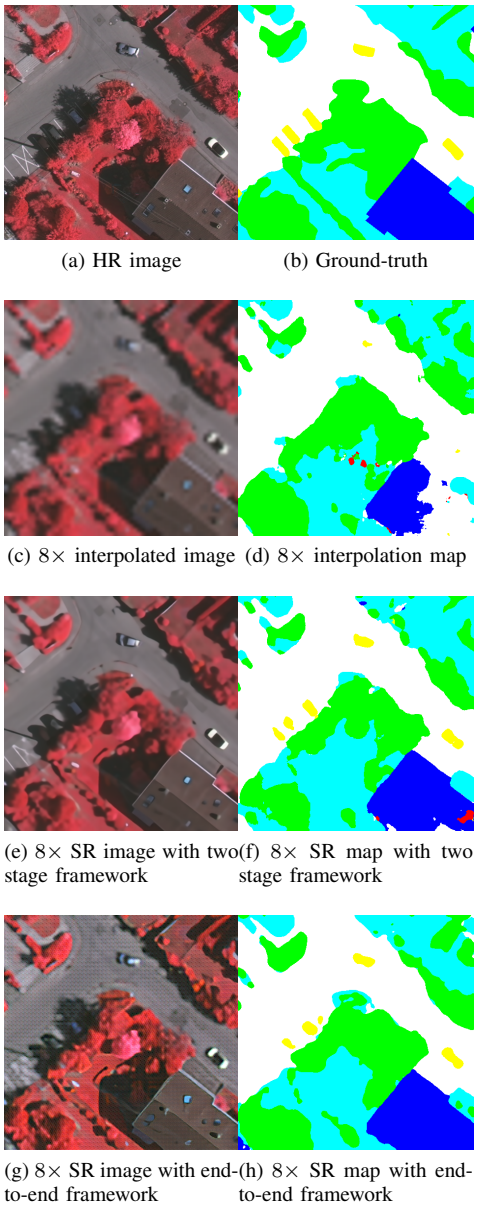
VI. Conclusion

In this work, we presented two frameworks that generate more accurate semantic segmentation thematic maps for LR remote sensing inputs with the use of super-resolution. The first one uses SR as a pre-processing, while the second one trains one single network that shares the loss of both tasks. We evaluated their performances on three different aerial datasets and under two degradation factors, comparing the results with bicubic up-sampled inputs.

SR was confirmed to be a viable strategy to recover important texture and object details for semantic segmentation. The recovered texture information greatly helps not to mislabel similar classes. Small objects, such as the cars in the Vaihingen dataset, which are not easily detected on LR representations, can become more discernible with the employment of SR.

The end-to-end framework allows the semantic segmentation loss to coordinate the image reconstruction. The reconstructed image for this framework presents artifacts generated by the semantic segmentation network that help the task to be performed.

For future work we plan to evaluate different experiments, such as the performance of a semantic segmentation network when trained and tested with LR data compared to the use of SR. We can also evaluate the performance with only reconstructed data. There is also space to study how different reconstruction/visual benchmarks can help to enhance the semantic segmentation performance more than PSNR, such as adversarial losses from GANs.



(a) HR image (b) Ground-truth

(c) $8\times$ interpolated image (d) $8\times$ interpolation map

(e) $8\times$ SR image with two stage framework (f) $8\times$ SR map with two stage framework

(g) $8\times$ SR image with end-to-end framework (h) $8\times$ SR map with end-to-end framework

Fig. 4. Example results for the Vaihingen dataset with $8\times$ up-scaling factor.

Publications

The work presented in this paper (which is related to a M.Sc. dissertation) allowed the publication of two papers: (i) "How Effective Is Super-Resolution to Improve Dense Labelling of Coarse Resolution Imagery?", accepted in the main track of the 32nd Conference on Graphics, Patterns and Images (SIBGRAPI 2019) [17], and (ii) "An End-to-end Framework for Low-Resolution Remote Sensing Semantic Segmentation", accepted in the main track of the 2020 Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS 2020) [18].

References

[1] 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.

[2] 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.

[3] 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.

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

[5] ——, "Deep backprojection networks for super-resolution," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] 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.

[7] S. N. Ferdous, M. Mostofa, and N. M. Nasrabadi, "Super resolution-assisted deep aerial vehicle detection ," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, T. Pham, Ed., vol. 11006, International Society for Optics and Photonics. SPIE, 2019, pp. 432 – 443.

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

[9] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network." in *CVPR*, vol. 2, no. 3, 2017, p. 4.

[10] 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.

[11] Z. Guo, G. Wu, X. Song, W. Yuan, Q. Chen, H. Zhang, X. Shi, M. Xu, Y. Xu, R. Shibasaki, and X. Shao, "Super-resolution integrated building semantic segmentation for multi-source remote sensing imagery," *IEEE Access*, vol. 7, pp. 99 381–99 397, 2019.

[12] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] 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.

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

[15] O. A. Penatti, K. Nogueira, and J. A. Dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 44–51.

[16] 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.

[17] M. B. Pereira and J. A. dos Santos, "How effective is super-resolution to improve dense labelling of coarse resolution imagery?" in *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2019, pp. 202–209.

[18] M. B. Pereira and J. A. d. Santos, "An end-to-end framework for low-resolution remote sensing semantic segmentation," *arXiv preprint arXiv:2003.07955*, 2020.