Deep CNN with Skip Connection and Network in Network for Super-Resolving Blurry Text Images

Hala Neji^{*†}, Mohamed Ben Halima[†], Javier Noguera-Iso[‡], Adel M. Alimi[†]

^{*}University of Gabes, Tunisia

[†]REGIM-Lab, University of Sfax, Tunisia

[‡]I3A, Universidad de Zaragoza, Spain

Abstract—Deep convolutional neural networks (Deep CNN) have achieved hopeful performance for single image super-resolution. In particular, the Deep CNN skip Connection and Network in Network (DCSCN) architecture has been successfully applied for natural images super-resolution. In this work we adopt DCSCN for deblurring and super-resolving single text images. We suggest a novel approach that performs jointly super-resolution and deblurring from a low blurry image. The experimental results have achieved state-of-the-art performance in PSNR and SSIM metrics. Thus, we confirm that DCSCN provides satisfactory results for enhancement tasks on low blurry images. The quantitative and qualitative evaluation on different datasets proves the high performance of our model to reconstruct high-resolution and sharp text images. In addition, in terms of computation time, our proposal improves the efficiency of comparable approaches.

Index Terms—Image super-resolution, Image deblurring, Deep CNN, Text images.

I. INTRODUCTION

Image quality enhancement is an essential step in image processing. In image processing there are several categories of image enhancement such as deblurring, denoising, dehazing and super-resolution. One of the hardest challenges in image processing is the super-resolution of blurry text images [1], [2], [3]. In this paper, we are interested in improving the quality of text images to facilitate document image analysis tasks such as optical character recognition (OCR) and further text processing. In recent years, many works have focused on the super-resolution of images. For instance, Deep convolutional neural networks (Deep CNN) have achieved hopeful performance for single image super-resolution. However, there is little research in the literature about the super-resolution of blurry text images, i.e address jointly the problems of deblurring and super-resolving text images.

The main objective of this paper is to achieve these two joint problems of deblurring and super-resolution of text images without having knowledge of the blur kernel. Our proposal is based on the Deep CNN skip Connection and Network in Network (DCSCN)[4] architecture, which has been successfully applied in the past for natural images super-resolution. Within the context of our proposal, we use DCSCN to recover a sharp high resolution image from a low blurred input image. Thus, no information about the blur kernel is required. As shown in Figure 1, DCSCN is a fully convolutional neural network. This model contains two sections. The first one is the

feature extraction network, which is in charge of extracting both the local and the global image features. The second one is the reconstruction network, which is responsible of reconstructing the image details. One of the main advantages of this model is that it provides an efficient computation for image reconstruction, achieving a computation time lower than comparable state-of-the-art approaches.

The remainder of this paper is organized as follows. We present the state-of-the-art methods in Section II. Section III presents the details about the proposed method. The experiment results and discussions are described in Section IV. Section V provides some concluding remarks.

II. RELATED WORK

A. Image Super-resolution

In recent years, there have been remarkable advances in image super-resolution. We can identify two main categories of super-resolution methods: exemplar-based methods and regression-based methods.

Regarding exemplar-based methods, Wang et al [5] proposed a method for learning a semi-coupled dictionary that consists of a pair of dictionaries (one for high resolution domain, and another for low resolution domain) and a mapping function. In the same line of using dictionaries, Yang et al [6] trained jointly two dictionaries for low and high resolution image patches and establish a similarity function between the sparse representations of low and high resolution image patch pairs. Zeng et al. [7] also proposed a single image super-resolution method that learns iteratively non-linear mappings between low and high resolution sparse representations. Focused on the problem of generating high resolution images from face images, Yang et al [8] proposed to learn statistical priors by exploiting local image structures (facial components, contours and smooth regions).

With respect to regression-based methods, we can cite the work of Zhang et al [9], who proposed a non-local kernel regression framework that takes profit of non-local self-similarity of image patches that tend to repeat themselves in natural images and the local structural regularity properties in image patches. Sun et al. [10] exploited the use of gradient profiles describing the shape and sharpness of image gradients to estimate a high-resolution image from a low resolution image. In the case of Yang et al. [11], they proposed a method that learns statistical priors by applying simple functions to

Fig. 1. Adjustment of DCSCN architecture proposed by Yamanaka et al [4].

exemplars, which are extracted by dividing the feature space in numerous subspaces. Another relevant work is the one proposed by Timofte et al [12], who combined the simple functions approach with the use of anchored neighborhood regression, an approach that learns sparse dictionaries and regressors anchored to the dictionary atoms.

Last, we must mention the existence of recent works that have successfully applied convolutional neural networks (CNN) for single image super-resolution. In particular, the work of Tran et al. [13] is especially relevant as they have focused on text image super-resolution by adapting the Deep Laplacian Pyramid Network.

B. Image deblurring

Image deblurring has been studied for a long time, and in recent years several works have specially focused on the problem of deblurring text images by means of Bayesian-based methods. For instance, Ljubenovic et al. [14] proposed the use of a dictionary-based prior for class-adapted blind deblurring of document images. Jjang et al. [15] have also proposed a method based on the two-tone prior for text image deblurring.

In addition, CNN networks have been also applied for text image deblurring. An example of this kind of works is the one proposed by Hradis et al. [16], an end-to-end method to generate sharp images from blurred ones using CNN. Their network, consisting of 15 convolutional layers, is trained on 3 classes (blur, sharp image and kernel blur).

C. Joint super-resolution and deblurring

There have been recent proposals to solve jointly the image super-resolution and the deblurring problems using deep learning.

Zhang et al. [1] suggested a deep encoder-decoder network (ED-DSRN) to solve the problem of blurry images degraded by Gaussian blur kernel. Zhang, Dong et al. [2] proposed a method for weighting several enhanced versions of an input image with the weights predicted by a CNN. Liang et al. [3] proposed a novel dual supervised network (DSN) to solve the deblurring and super-resolution problems.

Du et al. [17] suggested a novel approach based on CNN to reconstruct high resolution images from low-blurry ones. Liu et al. [18] proposed a deep decoupled cooperative learning based CNN deblurring model to achieve disentangling and synthesizing single image super-resolution and motion deblurring. Lumentut et al. [19] proposed a framework for the light field (LF) image enhancement using a deep neural net to super-solve the LF spatial deblurring and super-resolution under 6-degree-of-freedom camera motion. Albluwi et al. [20] also proposed a novel method for single image super-resolution to tackle blur with the down-sampling of images by using CNN.

Making profit from Generative Adversarial Networks (GAN), several works have been proposed in recent years. Yun and Park [21] used GAN to reconstruct high-resolution facial images by simultaneously generating a high-resolution image with and without blur. Li et al. [22] proposed a novel approach using GAN with Pixel and Perceptual Regularizations, denoted P2GAN, to restore single motion blurry and low-resolution images jointly into clear and high-resolution images. Du et al. [23] also proposed a method based on GAN for reconstructing clear high-resolution images directly from blurred low-resolution natural scene images. And focusing on the context of images combining faces and text, Xu et al. [24] trained a GAN network for super-resolving this kind of images.

III. PROPOSED METHOD

As shown in Figure 1, we employ the same network and we start from scratch to train the model. The model is a fully convolutional neural network. It consists of two parts: a feature extraction network and a reconstruction network. We try to adjust the network to optimize the results on text images by changing the hyper-parameters.

A. Feature extraction

In general, all deep learning methods have a pre-processing step which is the up-sampling of the images input. In these methods the single image super-resolution networks can be pixel-wise, which is the main reason to have a big network with a large CNN rather than the complexity of computation. The fact is that these methods require multiple GPU computing to accelerate the computation, but these computation resources are not always available. To solve this issue, the DCSCN model uses a feature extraction section to decrease the size of the network and have a fast response time. DCSCN uses the original image as input of the network and the up-sampling step is inside the architecture, which decreases the number of layers and obtains a better performance with faster computation.

As an initial configuration of our architecture we started with the same network used in DCSCN, which contains seven 3x3 CNNs with *ReLU* layer activator. Since the results were promising, we tried to optimize the network for text images and tested several activators such a *ReLU*, *LeakyReLU*, *PReLU*, *Sigmoid*, *Tanh* and *SELU*. Finally, we fixed the activator at *PReLU* because it obtained the best results to train the network.

In addition, to extract features efficiently, we evaluated different number of layers. On the one hand, we wanted to use the minimum layers to have a fast performance. On the other hand, we aimed to solve correctly the problem of deblurring and super-resolution at the same time. After many experiments, we fixed the number of filters of first feature-extraction CNNs to 196 and the number of filters of last feature-extraction CNNs to 32. Unlike to DCSCN, which has seven layers, we propose a cascade of eight 3x3 CNNs.

With respect to other parameters, we have decreased the filter decay gamma from 1.5 to 1.2. In addition, we have tested different initializers for weights: *Uniform*, *Stddev*, *Xavier*, *He*, *Identity* or *Zero*. At the end, we chose *He* initializer.

In summary, with this network the input image cascades to a set of CNN weights, biases and non-linear layers. After, all the extracted features are connected with skip connection to the next section of the network for reconstructing the high-resolution and sharp image.

B. Image reconstruction

In the reconstruction network, the CNN structure is parallelized in the same way as in the Network in Network architecture [25]. The main benefit of this structure is to reduce the dimensions of the previous layer for faster computation and

to add more non-linearity to enhance the potential representation of the network. Therefore, we can lessen the number of transposed CNN filters. In addition, it must be noted that 1x1 CNNs have a computation cost 9 times lower than 3x3 CNNs. Thus, the size of the network will be reduced. As it has been mentioned before, all the global/local features are extracted and concatenated at the input layer of the reconstruction network. The input data will be very large, so it is a good idea to use 1x1 CNNs to reduce the dimensions and generate the high resolution sharp image. As shown in figure 1, the dark blue CNN generate 4 outputs channels (in case the scale factor is equals to 2) and at the end of the model the low-resolution image is reshaped to a high-resolution and sharp one by adding the bicubic up-sampled original input image. We fixed the number of reconstruction layers to 3. The number of CNN filters in A1 at the reconstruction network is 64, and the number of CNN filters in B1 and B2 at the reconstruction network is set to 32.

IV. EXPERIMENT RESULTS

A. Datasets and environment configuration

For training the network we used the dataset proposed by Hradis et al. [16], which contains 66,742 paired images. We used only the sharp images as input to train the model. We cropped randomly the images with a patch size of 32x32, generating around 600,000 training images. For the generation of the training set, we fixed the pixel shuffler filters at 1. All the images, originally in RGB color, were converted to YCbCr images, and only the Y-channel was processed. As mentioned before, each training image was split into 32x32 patches and 20 patches were used as a mini-batch.

With respect to the network, the *He* initializer [26] was used for each CNN, and all biases and *PReLU*s were initialized to 0. We fixed the dropout to 0.8 used in each output of *PReLU* layers. To minimize the loss, we used the Adam optimizer with a learning rate equals to 0.002.

About the testing dataset, we decided to test the model on two different datasets: 100 paired images of the dataset proposed by Hradis et al.[16], which include both blurred and sharp low resolution images; and the dataset proposed by Tran et al. [13].

Figure 2 shows some representative examples of the training and testing datasets.

In addition, it must be noted that all the computation works were executed on an Ubuntu server with NVIDIA Quadro P6000 GPUs.

B. Quantitative Evaluation

We have evaluated quantitatively our method using two metrics: Structural Similarity Index (SSIM) and Peak Signal to Noise Ratio (PSNR). Table I shows the values obtained for these metrics according to different scales and selection of subsets extracted from the testing dataset. We make a distinction between blurred and sharp images selected from Hradis et al source [16], and the images taken from Tran et al source [13].

Fig. 2. Examples of the training and test dataset

TABLE I
QUANTITATIVE RESULTS OF OUR METHOD ON DIFFERENT TESTING DATASETS AND SCALES

Scale	Testing dataset	PSNR	SSIM
x4	Hradis et al., blurred	32.337284	0.744406996
x4	Hradis et al., sharp	34.32488	0.96228275
x4	Tran et al.	33.2383627	0.76454719
x3	Hradis et al., blurred	31.653434	0.581587
x3	Hradis et al., sharp	34.95663191	0.946485035
x3	Tran et al.	33.41604541	0.7966061
x2	Hradis et al., blurred	31.883465	0.67898
x2	Hradis et al., sharp	36.5258812	0.990195
x2	Tran et al.	35.70103685	0.933932

TABLE II
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS USING BLURRED TEXT IMAGES SELECTED FROM HRADIS ET AL SOURCE [16]

Algorithm	Scale	PSNR	SSIM
Xu et al.[24]	4	20.1289	0.8970
DCSCN[4]	4	32.5246	0.9475
Ours	4	34.3248	0.96228
Xu et al.[24]	2	22.8236	0.9038
DCSCN[4]	2	35.9654	0.96938
Ours	2	36.5258	0.990195

In addition, we have compared quantitatively our method with the original method of DCSCN [4] and the method proposed by Xu et al [24]. For this comparison we have used the blurred text images selected from Hradis et al source [16] and two different scales. As shown in table II, our model achieves good values of SSIM and PSNR, a bit higher than the original method of DCSN and clearly better than the results obtained by Xu et al.

Last, we have also compared the quality of the OCR output obtained from the improved images generated by our method and DCSCN[4], which behaves better than the method of Xu et al. [24] as confirmed in table II. After applying Tesseract software¹ to obtain OCR from the sharp images in the testing dataset (used as reference text) and the images returned by

¹<https://opensource.google/projects/tesseract>

TABLE III
QUANTITATIVE COMPARISON OF OCR OUTPUT USING IMAGES FROM HRADIS ET AL SOURCE [16]

Algorithm	Scale	Average WER	Average cosine similarity
DCSCN[4]	4	0.5775671	0.94491255029269
Ours	4	0.5772893	0.944915078771532
DCSCN[4]	2	0.5178426	0.961377834949894
Ours	2	0.5164537	0.961456816180844

the methods, we have computed the average Word Error Rate (WER) and the average cosine similarity between the character frequency vectors derived from each pair of corresponding OCR output files. As shown in table III, our method performs better in scale 2 and 4, i.e. the WER value is lower and the cosine similarity is higher. It must be noticed that the obtained OCR is not perfect, even for the sharp images in the testing dataset, because we are using cropped images with incomplete words and lines. In the case of character frequency vectors it must be also noted that we have considered only the detection of letters and digits within the range of printable characters in ASCII encoding.

C. Qualitative Evaluation

Figures 3, 4 and 5 show three examples of the reconstruction provided by our model. Our model succeeded to reconstruct high resolution clear images from low blurry inputs. Compared to the ground truth, our method generates images with much clearer characters.

It must be noted that figure 5 also includes a visual comparison with respect to the results obtained by the state-of-the-art methods. Apart from the quantitative improvement shown in table II, the visual result of our method provides a much clearer and sharper image.

V. CONCLUSIONS

In this work, we have demonstrated that DCSCN is a feasible approach to generate high-resolution clear images from blurred low-resolution text images. With respect to the original architecture [4], the main contribution of the paper

Fig. 3. Results of our model on low-blurry text images with scale=4 (blurred example taken from Hradis et al. test dataset).

Fig. 4. Results of our model on low-resolution text images with scale=2 (example taken from Tran et al. test dataset).

Fig. 5. Qualitative comparison with state-of-the-art methods with scale=4 (blurred example taken from Hradis et al. test dataset).

is the adjustment of parameters in DCSCN architecture to provide appropriate results for text images.

In addition, we have compared our proposal with two relevant state-of-the-art methods verifying that our approach performs better, not only in terms of quantitative measures, but also taking into account the visual results. Last, we must

remark that this proposal can be efficiently implemented and run with a quick response time.

ACKNOWLEDGMENT

The research leading to these results has received funding from the Ministry of Higher Education and Scientific Research

of Tunisia under the grant agreement number LR11ES48, from the Spanish Government through the project TIN2017-88002-R, and from the Aragon Government (Spain) through the project LMP19_18.

REFERENCES

- [1] X. Zhang, F. Wang, H. Dong, and Y. Guo, "A deep encoder-decoder networks for joint deblurring and super-resolution," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1448–1452.
- [2] X. Zhang, H. Dong, Z. Hu, W.-S. Lai, F. Wang, and M.-H. Yang, "Gated fusion network for joint image deblurring and super-resolution," *arXiv preprint arXiv:1807.10806*, 2018.
- [3] Z. Liang, D. Zhang, and J. Shao, "Jointly Solving Deblurring and Super-Resolution Problems with Dual Supervised Network," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 790–795.
- [4] J. Yamanaka, S. Kuwashima, and T. Kurita, "Fast and accurate image super resolution by deep CNN with skip connection and network in network," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 217–225.
- [5] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2216–2223.
- [6] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19 (11), p. 2861–2873, 2010.
- [7] K. Zeng, H. Zheng, Y. Qu, X. Qu, L. Bao, and Z. Chen, "Single Image Super-Resolution With Learning Iteratively Non-Linear Mapping Between Low-and High-Resolution Sparse Representations," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 507–512.
- [8] C.-Y. Yang, S. Liu, and M.-H. Yang, "Structured face hallucination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1099–1106.
- [9] H. Zhang, J. Yang, Y. Zhang, and T. S. Huang, "Non-local kernel regression for image and video restoration," in *European Conference on Computer Vision*. Springer, 2010, pp. 566–579.
- [10] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.
- [11] C.-Y. Yang and M.-H. Yang, "Fast direct super-resolution by simple functions," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 561–568.
- [12] R. Timofte, V. D. Smet, and L. V. Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian conference on computer vision*. Springer, 2014, pp. 111–126.
- [13] H. T. Tran and T. Ho-Phuoc, "Deep Laplacian Pyramid Network for Text Images Super-Resolution," in *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*. IEEE, 2019, pp. 1–6.
- [14] M. Ljubenovic, L. Zhuang, and M. A. Figueiredo, "Class-adapted blind deblurring of document images," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 721–726.
- [15] X. Jiang, H. Yao, and S. Zhao, "Text image deblurring via two-tone prior," *Neurocomputing*, vol. 242, pp. 1–14, 2017.
- [16] M. Hradiš, J. Kotera, P. Zemcik, and F. Šroubek, "Convolutional neural networks for direct text deblurring," in *Proceedings of BMVC*, vol. 10, 2015, p. 2.
- [17] B. Du, X. Ren, and J. Ren, "CNN-based Image Super-Resolution and Deblurring," in *Proceedings of the 2019 International Conference on Video, Signal and Image Processing*, 2019, pp. 70–74.
- [18] H. Liu, J. Qin, Z. Fu, X. Li, and J. Han, "Fast simultaneous image super-resolution and motion deblurring with decoupled cooperative learning," *JOURNAL OF REAL-TIME IMAGE PROCESSING*, 2020.
- [19] J. S. Lumentut and I. K. Park, "Deep neural network for joint light field deblurring and super-resolution," in *International Workshop on Advanced Imaging Technology (IWAIT) 2020*, vol. 11515. International Society for Optics and Photonics, 2020, p. 1151507.
- [20] F. Albluwi, V. A. Krylov, and R. Dahyot, "Super-resolution on degraded low-resolution images using convolutional neural networks," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.
- [21] J. U. Yun and I. K. Park, "Joint Face Super-Resolution and Deblurring Using a Generative Adversarial Network," *arXiv preprint arXiv:1912.10427*, 2019.
- [22] Y. Li, Z. Yang, X. Mao, Y. Wang, Q. Li, W. Liu, and Y. Wang, "GAN with Pixel and Perceptual Regularizations for Photo-Realistic Joint Deblurring and Super-Resolution," in *Computer Graphics International Conference*. Springer, 2019, pp. 395–401.
- [23] B. Du, X. Ren, S. Chen, J. Ren, and D. Cao, "Image Super-Resolution and Deblurring Using Generative Adversarial Network," in *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, 2019, pp. 266–271.
- [24] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 251–260.
- [25] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.
- [26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.