Spatial Transformer Generative Adversarial Network for Image Super-Resolution

Pantelis Rempakos[1], Michalis Vrigkas[2][0000−0001−5888−6949], Marina E. Plissiti[1][0000−0002−2344−9404], and Christophoros Nikou[1][0000−0003−1388−6915]

[1] Department of Computer Science and Engineering, University of Ioannina, 45110
Ioannina, Greece
{cs04279, marina, cnikou}@uoi.gr
[2] Department of Communication and Digital Media, University of Western
Macedonia, 52100 Kastoria, Greece
mvrigkas@uowm.gr

Abstract. High-resolution images play an essential role in the performance of image analysis and pattern recognition methods. However, the expensive setup required to generate them and the inherent limitations of the sensors in optics manufacturing technology leads to the restricted availability of these images. In this work, we exploit the information retrieved in feature maps using the notable VGG networks and apply a transformer network to address spatial rigid affine transformation invariances, such as translation, scaling, and rotation. To evaluate and compare the performance of the model, three publicly available datasets were used. The model achieved very gratifying and accurate performance in terms of image PSNR and SSIM metrics against the baseline method.

Keywords: Image super-resolution · Spatial transformer · VGG · SR-GAN

1 Introduction

The prospect of obtaining detailed digital high-resolution (HR) images from a set of low-resolution (LR) observations has been a topic of great interest in both the fields of signal and image processing [15, 21]. In the last few decades, recent developments in convolutional neural networks and advancements in GPU technology have further lowered the barrier to accumulating high-resolution images and videos. In addition, state-of-the-art machine learning models have made major breakthroughs in conventional computer vision tasks [12, 17].

Despite the aforementioned improvements, the methods of image and video super-resolution (SR) available today can be unsatisfying, in the sense that they fail to match expectations in perceptual quality and computational efficiency [14, 11]. Nonetheless, the need for high-quality images has remained at large an essential need for human interpretation of information and machine perception. In this work, we seek to achieve a feature fine, realistic, and computationally efficient super-resolution method that brings about quality image enhancement.

One of the main challenges when it comes to SR methods is the task of generating high-quality images that are realistic and sensible to human perception. This problem is difficult in the sense that the quality of the generated images is affected by multiple factors such as the quality of the input image taken by the image extracting sensor type, the image spectral and spatial resolution, and light variations. As a consequence, images generally tend to appear distorted, and blurry, with noise which SR models seek to adverse. In our approach, we aspire to develop a robust model that reliably understands this problem and generates HR images that are invariant to large geometric aberrations.

More specifically, in our approach, we construct a novel robust model that reliably generates high-resolution images that are invariant to these geometric aberrations. We call this model, the ST-SRGAN model, which essentially accounts for the fusion of spatial transformer networks (STN) alongside a super-resolution generative adversarial network (SRGAN) [11]. Our model exhibits remarkable performance in common publicly available datasets, as was verified by the experimental results.

2 Related Work

In the past, the consensus was to approach the SR problem with either statistical, prediction, or patch-based methods [16, 20, 19]. We can briefly summarize the related super-resolution methods in two major categories namely (i) learning-based and (ii) reconstruction-based methods.

Learning-based methods approximate HR images by using neighbor embedding, sparse coding, pixel-based, and example-based methods [24, 6, 7]. On the other hand, reconstruction-based methods use prior retrieved information to determine the HR limitations such as edge sharpening, regularization, and deconvolution methods [3, 1, 18]. Nowadays, researchers have substantially suppressed the limitations of the SR problem, leading to the development of cost-effective systems that allow researchers to make better use of big data.

Learning-based methods. A case in point is the super-resolution convolutional neural network (SRCNN) [4], which is regarded as the earliest CNN super-resolution model. The model structure consists of three parts. The first part relates to the extraction of data from the LR image. The second part implements non-linear mapping, a dimension-reducing method that attempts to retain the distances between data points as well as possible. Finally, in the third part, the model super resolves the image and reconstructs its high-resolution counterpart.

Kim *et al.* [9] showed that increasing network depth resulted in significant improvements in model accuracy. The VDSR network architecture consists of 20 weight layers which is much deeper than its SRCNN counterpart which only has three layers. By cascading small filters many times in a deep network structure, contextual information over large image regions is exploited efficiently. However using very deep networks, convergence speed would become a critical issue during training. To counter this problem the model would learn only residuals and use extremely high learning rates (104 times higher than SRCNN) enabled by adjustable gradient clipping.

The limitation of these methods is that increasing the resolution of LR images before the image-enhancing step may lead to high computational complexity. This is an especially problematic state for CNNs, where processing speed depends directly on the resolution of the input image. Furthermore, interpolation methods that are typically used to accomplish this task (e.g. bicubic interpolation) do not bring the additional information required to tackle the ill-posed nature of the SR reconstruction problem.

Reconstruction-based methods. Dong *et al.* [5] proposed an improvement to the SRCNN model that uses a post-upsampling reconstruction technique called FSRCNN. In this approach, feature extraction is performed in the low-resolution space. In addition, FSRCNN also uses a 1×1 convolutional layer after feature extraction to reduce the computational complexity cost by reducing the number of channels required. FSRCNN has a relatively shallow network which makes it easier to learn about the effect of each component. This model is even faster with better-reconstructed image quality than the previous SRCNN.

Nonetheless, for problems where LR images need to be upscaled by large factors (i.e., $8\times$), regardless of whether the upsampling is complete before or after passing through the deep SR network, the results are bound to be suboptimal. It makes more sense to progressively upscale the LR image in such cases to finally meet the spatial dimension criteria of the HR output rather than upscaling by $8\times$ in one shot. To this end, Lai *et al.* [10] proposed that the sub-band residuals of HR images can progressively be reconstructed. Sub-band residuals refer to the differences between the upsampled image and the ground truth HR image at the respective level of the network.

3 Method

Image SR methods aim to reconstruct high-resolution images given a set of low-resolution observations. In this section, we elaborate the details of the proposed ST-SRGAN method and describe the architectural units and training objectives.

3.1 Image Model Formulation

Since digital imaging systems are subject to hardware limitations, images are often degraded due to these limitations. The captured image may often be distorted by motion blur or additive noise because of the limited time window during the sensor of the image-capturing system is open. This problem is posed in its linear form as:

$$\mathbf{y}_k = \mathbf{W}_k \mathbf{z} + \mathbf{n}\,, \tag{1}$$

where $\mathbf{y}_k \in \mathbb{R}^{M \times N}$ is the *k-th* LR image, with $k = 1, \ldots, p$. The desired HR image $\mathbf{z} \in \mathbb{R}^{r_1 M \times r_2 N}$, where r_1 and r_2 are the up-scale factors in the horizontal and vertical directions, respectively. The degradation matrix $\mathbf{W}_k = \mathbf{D}_k \mathbf{B}_k \mathbf{M}_k$

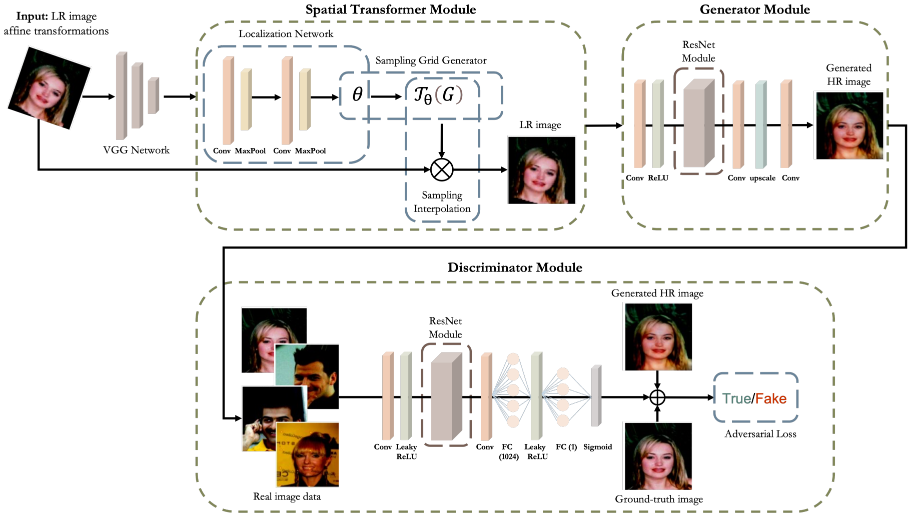

Fig. 1. Overview of the proposed ST-SRGAN architecture. First, the LR images are passed over the VGG network to extract a feature map and a spatial-transformer network is used to assess the affine transformations. Once the images have been aligned correctly, the generator module generates the estimated HR image. Finally, the generated HR image is detected as real or fake by the discriminator network based on real high-resolution images.

for the k-th frame performs the operations of (i) a motion matrix $\mathbf{M}$ that includes rigid transformation parameters such as rotation angle and translation vector, (ii) a blurring matrix $\mathbf{B}$, and (iii) a sub-sampling matrix $\mathbf{D}$. Finally, $\mathbf{n}$ is additive Gaussian noise. Note that all images are ordered lexicographical order.

3.2 Proposed Approach

We propose a spatial transformer-enhanced SRGAN network (ST-SRGAN) to address the aforementioned limitations of CNN-based methods. More specifically, we design a generator network to model the relationship among different views of an LR image to capture the relevant perceptual information in a robust and geometrically invariant way. Compared to traditional CNN-based methods, this model can discriminately incorporate information from multiple angular views, noisy, and in general spatially distorted images. The transformer works as an added self-attention mechanism to the generator network. The architecture of the proposed ST-SRGAN network is depicted in Fig. 1.

The LR images are first examined by the VGG network which extracts high-quality feature maps from LR images, that may be used to generate their high-resolution counterparts. The localization network takes the input maps retrieved by the VGG network and outputs the parameters of the affine transformations that should be applied to the feature maps. Following this procedure, the spatial-transformer processes these feature maps. Then, the grid generator proceeds to generate a grid of (x, y) coordinates using the parameters of the affine transformation that correspond to a set of points where the input feature map should be sampled to produce the transformed output feature map followed by a bilinear interpolation. The correctly aligned images are then fed into the generator, which then generates the estimated HR image. Using real high-resolution images and the generated HR images, the discriminator network predicts whether an image given by the generator is real or fake.

The generator and discriminator units work similarly to the original SRGAN architecture. However, compared to traditional GAN approaches it should be noted that the capabilities of the proposed discriminator are hampered by adding Gaussian noise layers in between the original layers of the model. This improves the performance of the model as a strong discriminator model is proven to work as a step function that hinders the result by producing no useful gradients to update the generator. Finally, dropout layers are also used.

Spatial-Transformer Module. Existing super-resolution methods do not consider that, when image transmission is over noisy channels, the effect of any possible geometric transformations could incur significant quality loss and distortions. To address this problem, the proposed model is formulated as a fusion of the SRGAN [11] and the spatial-transformer network [8]. This allows the development of a robust, spatially-transformed deep learning framework that is able to simultaneously perform both geometric transformations and image super-resolution. The reason for using the spatial-transformer network is that it provides model invariance when it comes to spatial transformations of LR images such as rotation, translation, and scaling.

More specifically, the spatial-transformer module consists of three main components, namely, (i) the localization network, (ii) the sampling grid generator, and (iii) bilinear interpolation. The localization network input corresponds to a $4D$ tensor representation of a batch of LR images $\mathbf{y}_k \in \mathbb{R}^{M \times N \times C}$, where C is the number of channels. The network contains a few convolutional layers and a few dense layers. Its output prediction consists of the parameters of transformation matrix $\mathbf{W}_k$. These parameters are used to determine the input feature map transformations that the network must estimate, such as the rotation angle of the input LR images, the amount of translation, and the scaling factor required to focus on the region of interest in the input feature map.

Then, the sampling grid generator predicts the transformation parameters which are in turn used in the form of an affine transformation matrix of size 2×3 for each LR image in the batch. Thus, we obtain a sampling grid of transformed indices:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \mathcal{T}_\theta \left(G_i \right) = \mathrm{A}_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} \; \theta_{12} \; \theta_{13} \\ \theta_{21} \; \theta_{22} \; \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} , \tag{2}$$

where $\mathcal{T}_\theta \left(G_i \right)$ represents the transformation of grid $G_i = (x_i^t, y_i^t)$ of the target coordinates of the regular grid in the output feature map and (x_i^s, y_i^s) are the source coordinates in the input feature map. Matrix A_θ corresponds to the affine transformation and the parameters θ represent the rotation angles, translation and scale parameters of each LR image in the batch. Finally, a bilinear interpolation on transformed indices is performed to estimate the pixel value at the transformed point (x_i^t, y_i^t) using the four nearest pixel values. For example, a point $(1, 1)$ after a counter clockwise 45° rotation of its axes, becomes $(2, 0)$.

Residual Network Module. In our approach, layers are reformulated as learning residual functions with reference to the layer inputs instead of learning unreferenced functions. Each residual block can be expressed as a sequence of the following equations:

$$y_l = h\left(x_l\right) + F\left(x_l, W_l\right) \ , \tag{3}$$

$$x_{l+1} = f\left(y_l\right) \ , \tag{4}$$

where x_l and x_{l+1} are the input and the output of the l-th block, F is a residual function, $h(x_l)$ is an identity mapping function and f is a ReLU function. The main idea behind this sequence is to learn the additive residual function F with respect to the $h(x_l)$, taking advantage of an identity mapping function $h(x_l) = x_l$. To formulate an identity mapping $f(y_l) = y_l$, activation functions ReLU and batch normalization are considered as the "preactivation" of the weight layers, while traditional techniques considered them as "post-activation".

Generator and Discriminator Module. The generator contains the residual network module, instead of deep convolution networks because residual networks are easy to train and allow to be substantially deeper to generate better results. During training, an HR image is down-sampled to a LR image. The generator unit then tries to up-sample the image from low to high resolution. After the image is passed into the discriminator, the latter tries to distinguish between a ground-truth super-resolution and the estimated HR image and generates the adversarial loss which is then backpropagated into the generator unit.

The discriminator unit implements LeakyReLU as activation. The network contains eight convolutional layers with of 3×3 filter kernels, increasing by a factor of two from 64 to 512 kernels. Strided convolutions are used to reduce the image resolution each time the number of features is doubled. The resulting 512 feature maps are followed by two dense layers with a leakyReLU applied between those two layers and a sigmoid activation function is used to obtain a probability for sample classification.

3.3 Training Objective

Content loss: In this work, we used two types of content losses. The first one is the pixel-wise MSE loss $\mathcal{L}_{MSE}^{SR}$ of the residual network module, which is the most common MSE loss for image SR. However, MSE loss is not able to deal with high-frequency content in the image which resulted in producing overly smoother images.

$$\mathcal{L}_{MSE}^{SR} = \frac{1}{r_1 M r_2 N} \sum_{i=1}^{r_1 M} \sum_{j=1}^{r_2 N} \left(z_{i,j} - G_{\theta G}\left(y_{i,j}\right)\right)^2 \ . \tag{5}$$

The second loss is the VGG loss, which is based on the ReLU activation layer of the pre-trained VGG-19 network. Here the VGG network works as a feature extractor and the feature map $\phi(\cdot)$ that is extracted is used in the loss function.

$$\mathcal{L}_{VGG}^{SR} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(\phi\left(z\right)_{i,j} - \phi\left(G_{\theta G}\left(y\right)\right)_{i,j} \right)^2 . \tag{6}$$

Adversarial Loss: The adversarial loss forces the generator to generate an image more similar to the HR image by using the discriminator to differentiate between ground-truth and the estimated HR image.

$$\mathcal{L}_{G}^{SR} = \sum_{i=1}^{N} -\log D_{\theta D}\left(G_{\theta G}\left(y\right)\right) . \tag{7}$$

The total loss is computed as the sum of all the individual losses:

$$\mathcal{L}_{tot}^{SR} = \mathcal{L}_{MSE}^{SR} + \mathcal{L}_{VGG}^{SR} + \lambda \mathcal{L}_{G}^{SR} . \tag{8}$$

where $\lambda = 10^{-3}$ controls the importance of the $\mathcal{L}_{G}^{SR}$ term in the total loss. This loss is preferred over the mean-squared error loss because we do not care about the pixel-by-pixel comparison of the images. We are mostly concerned about the improvement in the quality of the images. Hence, by using this loss function in the ST-SRGAN model, we are able to achieve high-quality results.

4 Experiments

Data Selection In this study, the CelebA Dataset is used [13] for training purposes. This dataset is a large-scale face attributes collection with more than 200K celebrity photographs, each with 40 attribute annotations. The images cover background clutter and large pose variations. Furthermore, Set5 [2] and Set14 [23] datasets were also used for testing.

Evaluation Metrics To evaluate the SR reconstruction results, (i) the peak-signal-to-noise-ratio (PSNR), (ii) the mean squared error (MSE) of the original image and the degraded image, and (iii) the structural similarity index (SSIM) [22] were used.

Training Details and Parameters The network was trained on an NVIDIA Titan XP GPU using a random sample of 10K images from the CelebA dataset. These images are distinct from the testing images. To obtain the 64×64 LR and 256×256 HR images, we down-sampled and up-sampled respectively using the OpenCV library. The same applies to different upscaling factors (i.e., 2×, 3×, and 4×) used for the evaluation of the proposed method.

The LR and HR input image range is scaled to $[-1, 1]$ because we are using the *tanh* activation function. The MSE loss was thus calculated on images of intensity range $[1, 1]$. Note that the VGG-22 network is used to extract the feature maps of the dataset images. Also, batch sizes of 1, 4, and 8 were used

Table 1. Experimental results for *Config 1* variants (i.e., *c1-a*, *c1-b*, and *c1-c* with different up-scaling factors).

Up-scale factor	4×			3×			2×		
Configuration	*c1-a*	*c1-b*	*c1-c*	*c1-a*	*c1-b*	*c1-c*	*c1-a*	*c1-b*	*c1-c*
Batch size	$B = 1$	$B = 4$	$B = 8$	$B = 1$	$B = 4$	$B = 8$	$B = 1$	$B = 4$	$B = 8$
PSNR ↑	22.43	27.32	29.46	24.36	25.83	28.81	19.98	22.81	25.94
SSIM ↑	0.88	0.92	0.94	0.89	0.91	0.93	0.87	0.88	0.90
MSE ↓	0.28	0.24	0.22	0.38	0.32	0.23	0.23	0.21	0.16

and Adam optimizer is used with a learning step of 2×10^{-4} and decay rates $\beta 1 = 0.5$ and $\beta 2 = 0.999$, respectively.

The generator network is comprised of 16 residual blocks. In addition, dropout and Gaussian noise layers were added to the discriminator to avoid the vanishing gradient problem. The spatial-transformer module was applied prior to the convolutional layer of the generator and after receiving the extracted feature maps using the VGG-22 architecture. The localization network to the identity transformation of the spatial-transformer module was initialized before the training process and while building the generator network. Note that the MSE-based SR-ResNet network was employed as initialization to the generator network to avoid undesired local optima. The variant configurations were trained with 5×100 iterations in which the generator and discriminator have been trained alternatively between iterations.

4.1 Experimental Results

Image reconstruction measurements are accomplished via the PSNR, MSE (%), and SSIM (%) metrics. Parameter B represents the batch size of the experiment. In the tables below, we affirm the experimental results of various configurations made to evaluate our model i.e., namely *Config 1* and *Config 2* and their variations. These values are obtained through 5 random realizations of the experiment in each case.

- *Config 1*. This configuration consists of three variants namely, *c1-a*, *c1-b*, and *c1-c* that represent the behavior of the model when the input images suffer from blurring and additive Gaussian white noise for batch sizes of 1, 4, and 8. Finally, for each configuration, different up-scaling factors of 4×, 3×, and 2× were applied, respectively. The performance of the different configurations is shown in Table 1.
- *Config 2*. This configuration comprises also three variants namely, *c2-a*, *c2-b*, and *c2-c* that represent the behavior of the model when the input images suffer from blurring, additive Gaussian white noise, and spatial translations and rotations with batch sizes of 1, 4, and 8. For each configuration, different up-scaling factors of 4×, 3×, and 2× were employed, respectively. Quantitative results are summarized in Table 2.

Table 2. Experimental results for *Config 2* variants (i.e., *c2-a*, *c2-b*, and *c2-c* with different up-scaling factors).

Up-scale factor	4×			3×			2×		
Configuration	*c2-a*	*c2-b*	*c2-c*	*c2-a*	*c2-b*	*c2-c*	*c2-a*	*c2-b*	*c2-c*
Batch size	$B = 1$	$B = 4$	$B = 8$	$B = 1$	$B = 4$	$B = 8$	$B = 1$	$B = 4$	$B = 8$
PSNR ↑	17.85	18.38	20.84	18.28	19.84	22.74	19.01	21.85	21.75
SSIM ↑	0.86	0.87	0.90	0.85	0.92	0.92	0.87	0.92	0.92
MSE ↓	0.28	0.26	0.21	0.28	0.24	0.19	0.19	0.15	0.16

Table 3 shows a comparison of the nearest neighbor, bicubic, and SRGAN with the proposed method on benchmark data. The results confirm that the proposed ST-SRGAN methods outperform all reference methods regarding to the evaluation metrics. The values in bold indicate the best-quality reconstructed images. However, it is worth noting that visual inspection remains the main method to perform assessment for SR methods. Finally, visual results of the reconstructed HR images with 4× up-scaling are depicted in Fig. 2. As it can be observed, from LR images (first column), our method produces high-quality images (last column), which are a reliable and accurate approximation of the original image (middle column). Furthermore, our method overcomes both the limitations of blurring/noising and spatial translations and rotations.

Table 3. Comparison of NN, bicubic, and SRGAN [11] with 4× up-scaling.

	Config 1				*Config 2*			
	NN	Bicubic	SRGAN [11]	ST-SRGAN	NN	Bicubic	SRGAN [11]	ST-SRGAN
PSNR↑	26.26	28.43	29.40	**29.46**	17.14	19.01	20.80	**20.84**
SSIM ↑	0.76	0.82	0.85	**0.94**	0.67	0.72	0.90	**0.90**

5 Conclusion

We introduced a novel spatial transformer GAN network to solve the problem of image super-resolution. In this work, we aimed at showing the robustness of pairing a SR network with the spatial transformer network for estimating transformation parameters between LR images. CNN-based SR algorithms, such as the SRGAN can be massively improved when paired with spatial transformers. The spatial transformations applied to three publicly available image datasets were successfully learned by the network and this is further evaluated by performance metrics. In our experiments, 64×64 distorted face images were up-sampled in various degrees of up-scaling (e.g., 2×, 3×, and 4×). Moreover, the proposed method is not limited to its use in face SR but other image datasets. The robustness of

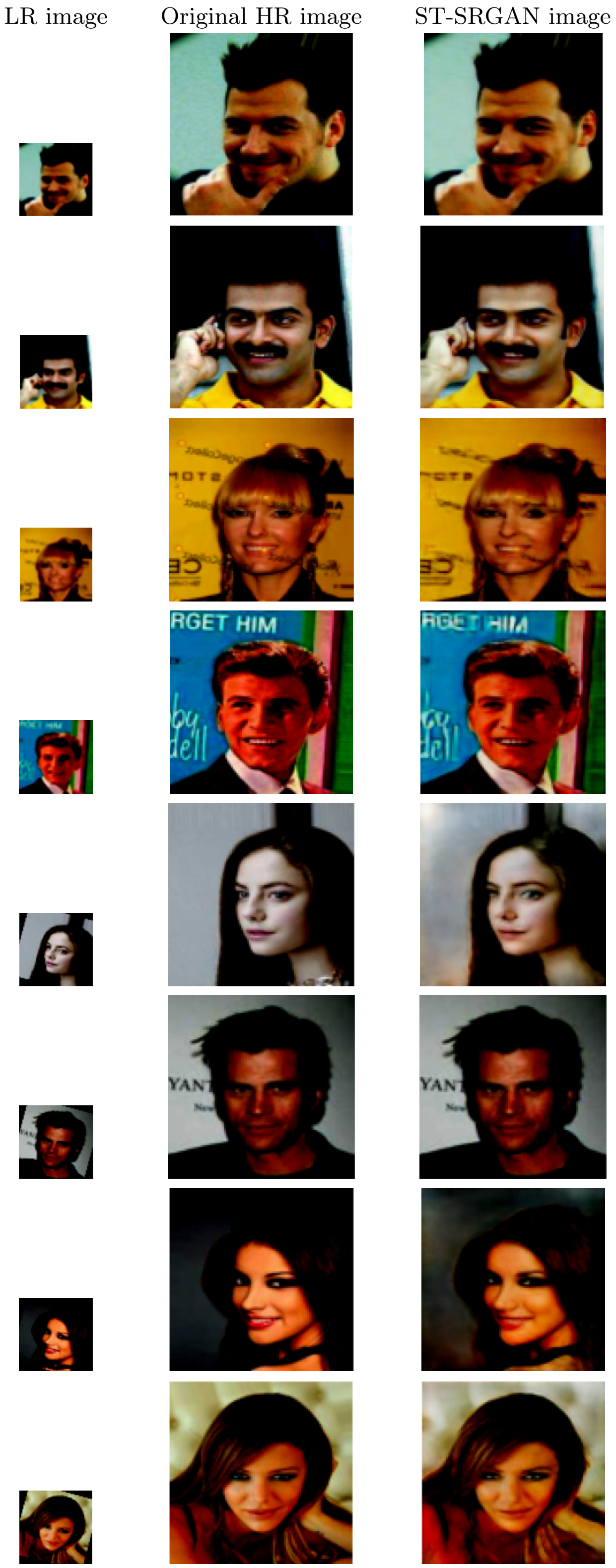

Fig. 2. Reconstructed HR images with 4× up-scaling.

the proposed model in spatial invariance is the reason behind its supremacy compared to baseline and state-of-the-art methods in super-resolution.

Furthermore, the cost of adding a spatial transformer model to our network is negligible. There are almost no extra computational costs in time and the size of the information required to process their trainable variables. The spatial transformer module has proved to be very powerful and very useful and its total potential is yet to be realised. In future work, we intend to extend the model for video super-resolution and exploit the spatiotemporal information found in between concurrent frames.

Acknowledgments The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan XP GPU used for this research. We acknowledge support of this work by the projects “Dioni: Computing Infrastructure for Big-Data Processing and Analysis” (MIS No. 5047222), which is implemented under the Action “Reinforcement of the Research and Innovation Infrastructure”, and “Bessarion” (T6YB -00214), which is implemented under the call “Open Innovation in Culture”, both funded by the Operational Program “Competitiveness, Entrepreneurship and Innovation” (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the sponsors.

References

1. Aly, H.A., Dubois, E.: Image up-sampling using total-variation regularization with a new observation model. IEEE Transactions on Image Processing **14**(10), 1647–1659 (Oct 2005)
2. Bevilacqua, M., Roumy, A., Guillemot, C., Morel, M.L.A.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: Proc. British Machine Vision Conference. No. 135, British Machine Vision Association (2012)
3. Dai, S., Han, M., Xu, W., Wu, Y., Gong, Y.: Soft edge smoothness prior for alpha channel super resolution. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (Jun 2007)
4. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(2), 295–307 (Feb 2016)
5. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Proc. European Conference on Computer Vision. pp. 391–407. Springer International Publishing (2016)
6. Freeman, W., Jones, T., Pasztor, E.: Example-based super-resolution. IEEE Computer Graphics and Applications **22**(2), 56–65 (2002). https://doi.org/10.1109/38.988747, https://doi.org/10.1109/38.988747
7. Gao, X., Zhang, K., Tao, D., Li, X.: Image super-resolution with sparse neighbor embedding. IEEE Transactions on Image Processing **21**(7), 3194–3205 (Jul 2012)
8. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. arXiv (Jun 2015)

9. Kim, J., Lee, J., Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. pp. 1646–1654. IEEE Computer Society (jun 2016). https://doi.org/10.1109/CVPR.2016.182
10. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5835–5843 (2017). https://doi.org/10.1109/CVPR.2017.618
11. Ledig, C., Theis, L., Huszar, F., Caballero, J., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. arXiv **abs/1609.04802** (2016), http://arxiv.org/abs/1609.04802
12. Liang, J., Zeng, H., Zhang, L.: Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5657–5666 (June 2022)
13. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. arXiv **abs/1411.7766** (2014), http://arxiv.org/abs/1411.7766
14. Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., Zeng, T.: Transformer for single image super-resolution. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 457–466 (June 2022)
15. Maral, B.C.: Single image Super-Resolution methods: A survey. arXiv (2022). https://doi.org/10.48550/ARXIV.2202.11763
16. Michaeli, T., Irani, M.: Nonparametric blind super-resolution. In: Proc. IEEE International Conference on Computer Vision. pp. 945–952 (2013). https://doi.org/10.1109/ICCV.2013.121
17. Pesavento, M., Volino, M., Hilton, A.: Attention-based multi-reference learning for image super-resolution. In: Proc. IEEE/CVF International Conference on Computer Vision. pp. 14697–14706 (October 2021)
18. Shan, Q., Li, Z., Jia, J., Tang, C.K.: Fast image/video upsampling. ACM Transactions on Graphics **27**(5) (dec 2008). https://doi.org/10.1145/1409060.1409106, https://doi.org/10.1145/1409060.1409106
19. Vrigkas, M., Nikou, C., Kondi, L.P.: On the improvement of image registration for high accuracy super-resolution. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 981–984. Prague, Czech Republic (May 2011)
20. Vrigkas, M., Nikou, C., Kondi, L.P.: A fully robust framework for map image super-resolution. In: Proc. IEEE International Conference on Image Processing. pp. 2225–2228. Orlando, FL (September 2012)
21. Wang, Y., Zhao, L., Liu, L., Hu, H., Tao, W.: URNet: A u-shaped residual network for lightweight image super-resolution. Remote Sensing (Basel) **13**(19), 3848 (Sep 2021)
22. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
23. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Proc. International Conference on Curves and Surfaces, pp. 711–730. Lecture notes in computer science, Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
24. Zhang, K., Gao, X., Tao, D., Li, X.: Single image super-resolution with non-local means and steering kernel regression. IEEE Transactions on Image Processing **21**(11), 4544–4556 (Nov 2012)