IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 18, 2019, accepted March 29, 2019, date of publication April 2, 2019, date of current version April 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2908996

A Very Deep Spatial Transformer Towards Robust Single Image Super-Resolution

JIANMIN JIANG[1], HOSSAM M. KASEM[1,2], AND KWOK-WAI HUNG[1]
[1]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[2]Faculty of Engineering, Tanta University, Tanta 31527, Egypt

Corresponding author: Kwok-Wai Hung (kwhung@szu.edu.cn)

This work was supported in part by the Natural Science Foundation China (NSFC) under Grant 61620106008 and Grant 61602312, and in part by the Shenzhen Commission for Scientific Research and Innovations under Grant JCYJ20160226191842793.

ABSTRACT In general, existing research on single image super-resolution does not consider the practical application 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, we present a new and robust super-resolution method in this paper, where a robust spatially-transformed deep learning framework is established to simultaneously perform both the geometric transformation and the single image super-resolution. The proposed seamlessly integrates deep residual learning based spatial transform module with a very deep super-resolution module to achieve a robust and improved single image super-resolution. In comparison with the existing state of the arts, our proposed robust single image super-resolution has a number of novel features, including 1) content-characterized deep features are extracted out of the input LR images to identify the incurred geometric transformations, and hence transformation parameters can be optimized to influence and control the super-resolution process; 2) the effects of any geometric transformations can be automatically corrected at the output without compromise on the quality of final super-resolved images; and 3) compared with the existing research reported in the literature, our proposed achieves the advantage that HR images can be recovered from those down-sampled LR images corrupted by a number of different geometric transformations. The extensive experiments, measured by both the peak-signal-to-noise-ratio and the similar structure index measurement, show that our proposed method achieves a high level of robustness against a number of geometric transformations, including scaling, translations, and rotations. Benchmarked by the existing state-of-the-arts SR methods, our proposed delivers superior performances on a wide range of datasets which are publicly available and widely adopted across relevant research communities.

INDEX TERMS Single image super-resolution, deep learning, spatial transform, geometric transformations.

I. INTRODUCTION

Over the past decades, the problem of image super-resolution has been extensively studied and numerous image SR methods have been reported to deal with this non-trivial problem [1]–[8]. In general, image SR reconstruction methods can be divided into two categories: multi-frame image SR (MISR) methods [9]–[11] and single-image SR methods (SISR) [12]–[14]. The MISR methods can be further divided into two categories: frequency domain methods and spatial domain methods [10]. The target of the frequency domain methods is to eliminate spectrum aliasing and reconstruct the high frequency information, including fourier transform-based methods [4], discrete cosine transform-based methods [15] and wavelet transform-based methods [16]. While the advantage of these methods is the computational efficiency, these methods have some limitations in incorporate the image prior knowledge [17]. The spatial domain methods such as non-uniform interpolation method [18], iterative back projection (IBP) method [19] and projection onto convex sets (POCS) [20]. While these methods generally have a good reconstruction ability, they are computational expensive [9]. Furthermore, it is not easy to obtain an sufficient number of LR images, in order to estimate a HR image from multiple blurred and noisy images [21]. Therefore, single-image SR methods are much more desirable in practical applications.

Technically, all the methods of single-image super-resolution (SISR) can be divided into: interpolation-based

The associate editor coordinating the review of this manuscript and approving it for publication was Yuming Fang.

methods [22], reconstruction-based methods [23], [24] and example learning-based methods [13], [25]. The concept of the interpolation-based method is to use the values of the neighbouring low-resolution pixels to estimate the value of the interpolated high-resolution pixel. The main advantages of the interpolation-based method are simple and relatively low computational complexity [1]. However, the interpolation-based methods often generate the reconstructed HR images with edge halos and artifacts [26].

The reconstruction-based methods recover the HR image based on the degradation model [24], [25]. As single image SR is highly ill-posed, image prior knowledge is often proposed to regularize the solution in this family of approaches. In order to obtain an effective image prior, it is of great importance to model the appropriate feature of natural images. Based on the max-posterior (MAP) theory which utilize some prior constraints into the data fidelity cost function, the SR problem turns to be well-posed. Therefore, the SR problem can be transformed into a minimization problem : $\hat{\mathbf{X}}_{HR} = \text{argmin}\left[L\left(\mathbf{X}_{HR}, \mathbf{Y}_{LR}\right) + \lambda U\left(\mathbf{X}_{HR}\right)\right]$, where $\hat{\mathbf{X}}_{HR}$ is the estimated HR image, $L(.)$ is the data fidelity term that represents the degree of consistency between the target HR image $\mathbf{X}_{HR}$ and the LR image $\mathbf{Y}_{LR}$, $U(.)$ is the regularization term describing the prior information of the original image, and λ is the regularization parameter, which is used to weigh the influence given by the prior $L(.)$ and $U(.)$ during the estimation. Many types of priors have been utilized into reconstruction- based methods, such as edge prior [27], gradient prior [28], and sparsity priors [29]–[33]. The learning-based methods estimate the HR image from the LR image by learning the relationship between the HR and LR image patches from the sample database. Inspired by the great success achieved by deep learning [34], people begin to use neural networks with deep architecture for image SR. Multiple layers are stacked together for robust learning of self-similar patches. Deep convolutional neural networks (CNN) [35] and deconvolutional networks [36] are designed to directly learn the non-linear mapping from LR space to HR space in a way similar to coupled sparse coding [37]. As these deep networks allow end-to-end training of all the model components between LR input and HR output, significant improvements have been observed over their shadow counterparts. Specifically, the feature spatial transform [38] tries to learn the texture transform from the segmentation maps that addresses the conventional super-resolution problem, which is fundamentally different from the proposed work to tackle geometric transformation and super-resolution simultaneously.

On the other hand, practical applications of single image super-resolution indicates that the real LR measurements usually suffer from various types of corruptions, such as geometric transformations, noise, and blurring. In this paper, we focus on dealing with the geometric transformation effects, for which the existing research has not properly addressed in the area of image super-resolution. Yet practical applications reveal that estimating HR images from transformed or distorted LR versions remains an important research topic demanded across a number of engineering sectors, such as self-driving vehicles, smart phones, and medical image analysis etc. One typical example is the classification of cancer types via LR images, in which various geometric transformations, such as scaling and rotation etc., could incur during the process of correlating different patterns across similar super-resolution images. Without the capability of dealing with these transformed LR images, the classification accuracy could be degraded significantly. Consequently, there are strong motivations to research on a robust super-resolution technique that can generate an HR image from the distorted and transformed LR image.

To alleviate the geometric transformation effects with resolution improvements, we propose a spatially transformed deep learning framework to achieve robust single image super-resolution in this paper. To our best knowledge, this is the first attempt on deep learning based super-resolution that simultaneously tackles the transformation effects and resolution enhancement, out of which our contributions can be highlighted as follows:

- Compared with all existing super-resolution deep learning networks, our proposed is a robust super-resolution network that can simultaneously handle both geometric transformations and resolution enhancement.
- In terms of structures, our proposed can seamlessly integrate a deep residual spatial transform network with a very deep super-resolution network to form a novel deep learning architecture.
- By replacing the original thick CNN with a new residual thin CNN, our proposed significantly improves the existing spatial transform network (STN) [39] in terms of functionality, performances, and efficiency (with less number of parameters.)
- In comparison with the existing state of the art SR methods reported in the literature, experimental results support and verify that our proposed ST-DISR achieves superior performances in terms of both PSNR and SSIM.

The rest of the paper is structured as follows. Section II describes our proposed framework via surveying the existing research across relevant areas, including both image super-resolution and spatial transformations. Section III presents experimental results and their analyses. Finally, section IV draws the concluding remarks and future research.

II. SPATIALLY TRANSFORMED DEEP FRAMEWORK FOR ROBUST IMAGE SUPER-RESOLUTION

A. RELATED WORK AND BACKGROUNDS

In general, the relationship between the original high-resolution and observed low-resolution images can be described by the following image observation model.

$$\mathbf{Y}_{LR} = \mathbf{DHX}_{HR} \tag{1}$$

where $\mathbf{H}$ is a degradation matrix that represents geometric transformations, and a down-sampling operator, $\mathbf{D}$, is applied

to the original image $\mathbf{X}_{HR}$ (i.e. HR image), to generate the observed image $\mathbf{Y}_{LR}$ (i.e. LR image). Due to the ill-condition of the SR problem, recovering HR image is not unique. The problem of image SR reconstruction can be generally modeled as: given the LR image $\mathbf{Y}_{LR}$, the objective is to estimate a HR image that is visually the same as the original HR image $\mathbf{X}_{HR}$. This problem can be formulated as

$$\hat{\mathbf{x}}_{HR} = \underset{\mathbf{x}_{HR}}{\operatorname{argmin}} \|\mathbf{Y}_{LR} - \mathbf{D}\mathbf{H}X_{HR}\|_2 \tag{2}$$

To solve this problem, an effective prior is necessary to turn the problem into a deterministic problem. Many solutions have been proposed to solve this ill-posed problem in the past decades [6]– [8]. Recently, there have been a large number of studies on deep learning for solving the image SR problem. As a result, we briefly review all the existing deep learning based SR techniques as follows in order to pave the way for our proposed framework.

As part of the pioneering exploration, super-resolution using deep convolutional networks (SRCNN) [8] is one of the state of the art for deep learning based SR methods, in which Dong et al. introduced a CNN constructing a mapping from the bicubic up-sampled LR space to HR space. Specifically, SRCNN utilizes the bicubic interpolation as its pre-processing step and then extracts the features of the overlapping patches using deep convolutional layers. At the final step, the extracted feature vectors are non-linearly mapped to each other and subsequently aggregated in the form of patches to form the reconstructed HR image.

In general, the essential advantage of the SRCNN is that only convolution layers are used. Consequently, the input image can be of any size and the algorithm can run on the input image in one pass. Although SRCNN claims efficiency in view of the its straightforward structure, it still has a number of limitations, in which the primary limitation is that the convergence of the network is too slow, and the network only works on a single scale.

Efforts were made by Dong *et al.* [40] to improve the efficiency of SRCNN, for which Dong *et al.* proposed fast SRCNN (FSRCNN). The reported FSRCNN [40] replaced the pre-processing Bicubic interpolation in SRCNN by a post-processing in the form of deconvolution. In addition, the FSRCNN has four convolution layers in the form of feature extraction, shrinking, mapping and expanding. That is, the mapping is preceded by shrinking feature dimensions and then expanding back at latter stages.

The success of CNN in SR often raises a question: whether a deeper network should be adopted in order to maximize the SR performances? Kim *et al.* [41] answered this question by proposing a very deep super-resolution (VDSR) network, which uses a very deep convolution network inspired by VGG-net used for ImageNet classification. The structure of VDSR contains 20 layers in a cascaded deep network. The filter size used in the network is 3-by-3 [41]. Kim *et al.* [41] utilized the residual learning to train the VDSR network, and the VDSR tackled the limitation of SRCNN by extending to multi-scales SR with a single network model.

Algorithm 1 Pseudocode of the Robust Single Image Super-Resolution Using Our Proposed Framework

Input: A LR image $\mathbf{Y}_{LR}$.
Output: An estimated HR image $\hat{\mathbf{X}}_{HR}$.
Initialization: Preparing the training dataset by distorted the HR image by bicubic downsampling and one or more of geometric transformation as in equation (1).
Steps of our proposed network:

- **First Module**: Deep residual learning based spatial transform module;
  - **Deep residual learning localization network**:
    - Extract the features of LR image through 20 stacked convolution layers as described in equation (6).
    - LR image version is added to the output of stacked layers.
    - The output is passed to the classifier to estimate the 6 geometric transformation parameters as given in equation (7).
  - **Grid Generator**: Estimate the sampling grid based on the estimated geometric transformation parameters.
  - **Sampler**: Interpolate the input LR image according to the sampling grid to alleviate the geometric transformation effects as described in equation (8).
- **Second Module**: Super-Resolution module: Refine the LR image to generate a HR image similar to the desired target.

Other researchers focus on utilizing different loss functions, instead of the mean square error (MSE), to generate HR images. Johnson *et al.* [42] utilized perceptual loss function to generate visually comforting results. Ledig *et al.* [43] employed a deep residual network (ResNet) as the discriminator to form the Super Resolution Generative Adversarial Network (SRGAN). The major problem of these networks is the difficult hyper-parameter tuning in the training process, including the weight initialization, the weight decay rate, and the learning rate, etc. A desirable property of an image processing system is to reason about the possible changes of object poses, and their relevant spatial transformations. A desirable SR network should be spatially invariant to the scaling and rotation effects incurred during transmission or hostile corruption. A spatial transformer network (STN) [39] is introduced to exploit the power of deep learning for dealing with spatial transformations, which is a dynamic mechanism that can spatially transform an image by producing an appropriate affine transformation for each input sample.

Specifically, a STN network takes an image or a feature map from a convolutional network as the input, and then an

affine transformation and bilinear interpolation is applied to produce the output of the STN. The affine transformation allows zoom, rotation and skew of the input, which enables the network to not only select regions of an image that are most relevant (attentive) but also to transform those regions to a canonical and expected pose, in order to simplify the recognition process in the following layers. Further, STN can be utilized with CNN to provide multi-functions, such as super-resolution as proposed in this paper.

Inspired by the spirit of both STN and residual learning, we propose a new spatial transformable deep super-resolution framework, to facilitate a robust super-resolution for LR input images. Extensive experiments support that our proposed framework is able to achieve expected robustness in comparisons with the existing state of the arts using deep learning for image super-resolutions.

B. OUR PROPOSED WORK

In this section, we illustrate the details of our proposed framework, including the deep residual learning based spatial transform module and the super-resolution module integrated to generate the final HR image and complete the robust single image super-resolution. The objective of our proposed framework is to create a robust SR network which is able to generate HR images from LR transformed images. The key aspect to achieving this goal is to be able to alleviate the effect of spatial transforms for corrupted LR images. The procedure of the proposed framework is summarized in **Algorithm** 1, which has the capability to tackle the transformation effects while refining and generating HR images to achieve robust super-resolution. Consequently, our proposed framework not only minimizes the effect of the geometric transformation effect, but also minimizes the difference between the estimated HR image and the HR image itself. Specific details are described as follows.

$$L(\theta) = \text{argmin} \left\| \hat{\mathbf{X}}_{HR} - \mathbf{X}_{HR} \right\|_2 \tag{3}$$

where $L(\theta)$ is a loss function (i.g. objective function), and θ represents the model parameters of the deep neural network. The above equation can be established via two operational steps. Firstly, we need to identify the affine transformation parameters, in order to capture any possible geometric transformation, through minimizing the errors between the HR and the distorted image. Secondly, we generate estimated HR image similar to the desired one through minimizing their corresponding MSEs. As a result, our loss function can be further formulated as follows:

$$L(\theta) = \underset{\theta_A}{\text{argmin}} \left\| \mathbf{Y}_{LR}(\theta_A) - \hat{\mathbf{X}}_T \right\|_2^2 + \underset{\theta_{DRLN}}{\text{argmin}} \left\| N(\hat{\mathbf{X}}_T, \theta_{DRLN}) - \mathbf{X}_{HR} \right\|_2^2 \tag{4}$$

where $\hat{\mathbf{X}}_T$ is the output image after performing the spatial transformation, and $\hat{\mathbf{X}}_{HR} = N(\hat{\mathbf{X}}_T, \theta_{DRLN})$ is the estimated HR image, θ_{DRLN} is the model parameters of the super-resolution neural network. $\hat{\mathbf{X}}_T$ is the output of a deep residual learning based spatial transform module, and θ_A represents the estimated geometric/affine transformation parameters.

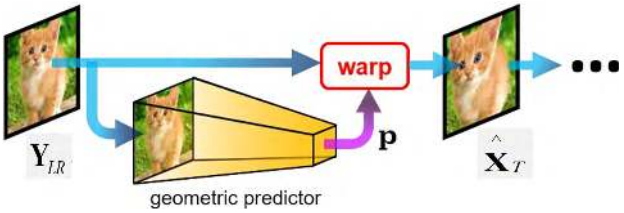

FIGURE 1. Spatial transform network [44].

The first part of Eq. (4), is to minimize the errors the error between the transformed LR image and the desired LR image for the super-resolution in the second part, in order to estimate the affine transformation parameters and mitigate the transformation effects. The second part of the equation is to minimize the errors between the output of the first module and HR image to obtain an image similar to the desired HR image. In practice, our network is trained end-to-end with one loss function only in Eq. (4), where the two sub-tasks (i.e. the first and the second part of Eq.(4)) are connected through the variable $\hat{\mathbf{X}}_T$, such that the weights of costs of two sub-tasks do not have any influences for the final results. In other words, given sufficient number of iterations during training, the network converges regardless of the weights of costs of two sub-tasks. Moreover, the back-propagation of network passes through the whole network due to the seamless integration of two sub-tasks in Eq. (4).

Over the recent years, deep neural networks have achieved huge success in resolving various computer vision problems. Nevertheless, there has not been an overwhelming solution for the problem of geometric variations upon the given datasets during the learning process. The recently introduced spatial transform network (STN) [39], which is able to perform spatial transformations on images with a differentiable module, has the function of reducing geometric variations of natural images and has attracted attentions across the deep learning community.

As shown in Fig.1, the STN warps an image conditioned on the input during the feed forward process, which can be formulated as:

$$\hat{\mathbf{X}}_T = \mathbf{Y}_{LR}(\theta_A), \ \texttt{where}\ \theta_A = f(\mathbf{Y}_{LR}) \tag{5}$$

where $\hat{\mathbf{X}}_T$ is the output of STN, $\mathbf{Y}_{LR}$ is the input image to STN and θ_A is the estimated geometric parameters.

STN contains three parts. The first part is a localization network, which takes the input image through convolutional neural network (CNN) and estimate the warping parameters from the input image. This part can be represented by a nonlinear function f, which is parametrized as a learnable geometric predictor. In the second part, the predicted transformation or warping parameters are utilized to form a sampling grid, which is implemented by a grid generator. In the third part, the input image and the sampling grid are taken as

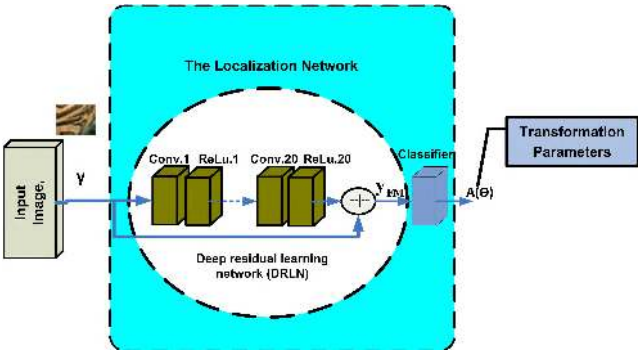

FIGURE 2. Deep residual learning network (DRLN) for localization network.

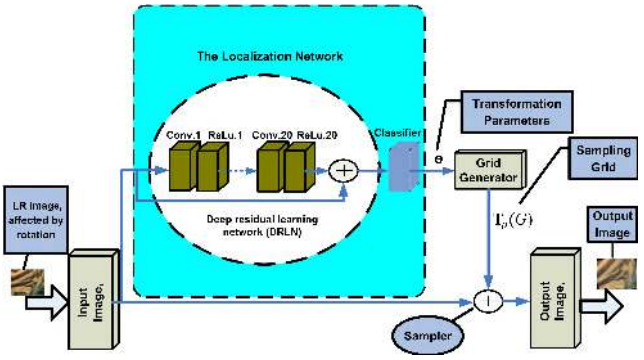

FIGURE 3. Illustration of our proposed very deep spatial transformer (VDST).

inputs to the sampler, in order to interpolate the output image. We note that the grid generator and the sampler from the original method can be combined to form a single warping function as shown in Fig.1.

Inspired by the idea of utilizing the residual learning for feature extraction [45], we propose to introduce a deeper residual learning localization network (DRLN) into the existing STN to create additional spaces and capabilities for exploiting wider contextual information inside the input images. Compared with the existing STN, our DRLN-based STN achieves two advantages: (i) increase the learning power to achieve more accurate estimation of the affine transform parameters; (ii) enable the STN not only to learn but also to optimize the capture of any geometric variation of input images in an adaptive manner against any possible content corruptions.

In the DRLN, we propose to stack 20 convolutional layers to extract the features of the input LR image. Compared with the existing localization using only two convolution layers with hundreds of feature maps [45]. the proposed DRLN is a deeper network with only 64 feature maps per layer, which is more powerful yet requires less parameters. However it will be more difficult to train such a deep network because of the vanishing gradients. This explains why we use such a residual learning framework, in order to overcome such a problem. The operation of DRLN can be formulated as:

$$\hat{\mathbf{X}}_{HR} = N(\hat{\mathbf{X}}_T, \theta_{DRRL}) + \hat{\mathbf{X}}_T \tag{6}$$

where $\hat{\mathbf{X}}_T$ and $\hat{\mathbf{X}}_{HR}$ are the input LR image and the output of the DRLN network respectively, θ_{DRRL} is the weights of convolutional layers, and the function $N(\hat{\mathbf{X}}_T, \theta_{DRRL})$ represents the estimated residual using the weights of the convolutional layers.

For the convenience of implementation, the feature maps that are extracted from DRLN are converted to one dimensional vectors using a fully connected layer, before being passed to the classifier that is able to estimate the geometric transformation parameters. The output of the classifier can be described as:

$$\theta_A = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \end{bmatrix} \tag{7}$$

In this way, more contextual information inside the LR input images can be exploited yet the number of learning parameters can also be reduced. Due to the deep structure of our proposed DRLN, the level of the feature is enriched and the accuracy of estimating the affine parameters is significantly improved in comparison with the existing STN [45]. Fig. 3 illustrates the overall structure of our proposed residual learning based spatial transform module.

As seen in Fig. 3, while the proposed DRLN is responsible for estimating the affine transformation parameters, the estimated affine parameters are utilized to form a mesh grid, which is implemented by grid generator. Both the input LR image and the grid are taken as inputs to the sampler in order to warp the input images. The full operation of our proposed transformer module can be described as follows:

$$\hat{\mathbf{X}}^c_{T_i} = \sum_n^H \sum_m^W \mathbf{Y}^c_{LR(n,m)} k\left(x_i - m; \Phi_x\right) k\left(y_i - n; \Phi_y\right), \quad \forall i \in \left[1....H'W'\right] \forall c \in [1....C] \tag{8}$$

where HW and $H'W'$ are the height and the width of the input and output images, respectively, Φ_x and Φ_y are the parameters of generic sampling kernel $k()$ which defines the image interpolation, $\mathbf{Y}_{LR(n,m)}$ is the value at location (n, m) of the input, $\hat{\mathbf{X}}_{T_i}$ is the output value for pixel i at location (x_i, y_i), and C represents the number of the channel of the input image.

By integrating the widely known VDSR [41] with the proposed deep residual learning based spatial transformer, we can construct a robust spatial-transformed deep image super-resolution network as shown in Fig. 4. As seen, the essential mechanism for achieving the robust single image super-resolution is the accurate estimation of spatial transform parameters, which simultaneously mitigate the geometric transformation effects and estimate the HR images via the very deep learning process.

III. EVALUATIONS AND EXPERIMENTAL RESULTS ANALYSIS

To evaluate our proposed very deep spatial transformer (VDST), we carry out extensive experiments and report our experimental results as well as their analyses in this section.

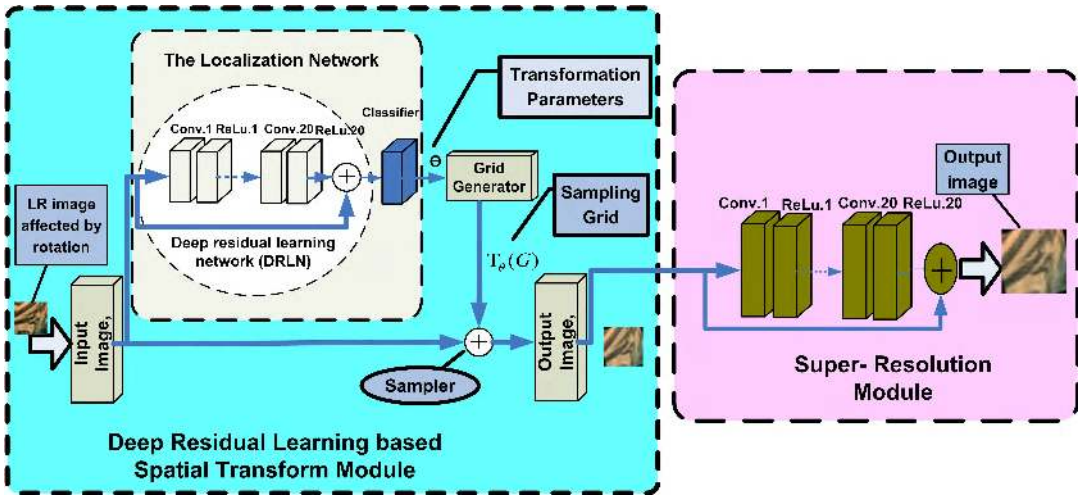

FIGURE 4. Illustration of the proposed framework for robust image super-resolution, where the input LR image is rotated by 20 degrees clock wise, and the HR image is generated with the corrected orientation after alleviating the geometric transformation effects at the output.

To make it convenient for benchmarking and comparative studies, we follow the experimental procedures as described in [41]. Firstly, we compare the performance of VDST with the existing spatial transform network [39]. For this comparison, we conduct experiments in order to evaluate the ability of each network to mitigate the geometric transformation effects and correct the orientation of the input image. Secondly, for the analysis purposes, we calculate number of the parameters needed in our proposed VDST and compare it with that of the exist spatial transform network.

To validate the effectiveness of our proposed VDST, extensive experiments are conducted on various standard benchmark datasets, which contain hundreds of natural images. To evaluate the performance of our proposed, a wide range of experiments are carried out under a number of different geometric transformations, including rotation (R), scaling (S), and translation (T), and we use these geometric transformations to simulate the possible geometric distortions likely incurred to the input images. The performance of our proposed framework is compared with the existing VDSR network to verify the effectiveness of our proposed. To validate that our proposed VDST outperforms the existing STN in terms of the robustness for single image super-resolution, we carry out further experiments against the existing STN followed by VDSR, for which we call STN-VDSR as an additional benchmark for assessing our proposed.

A. EXPERIMENT DESIGN AND SETUP

1) DATASET FOR TRAINING AND TESTING

We conduct the training experiments using 91 images from Yang *et al.* [29] and 200 images from the training set of Berkeley Segmentation dataset [46] as our training data. In each training batch, we randomly sample 25 patches. We augment the training data in three ways by following the protocol of the existing methods [41], and we generate the LR training patches using the bicubic down-sampling. For the convenience of benchmarking, we carry out experiments using 5 publicly available datasets, including: BSDS100 [46], SET5 [47], SET14 [48], URBAN100 [49] and MANGA109 [50].

2) IMPLEMENTATION DETAILS

To simulate the effect of the geometric transformations, we generate the transformed LR training image by five different transformations, including: (i) the rotation effect represented by R, in which the original image is rotated clockwise by 20 degrees; (ii) the scaling effect represented by S, in which the original image is scaled with a factor of 0.5; (iii) the effect of both rotation and scaling represented by RS; (iv) translation represented by T, in which the LR images are translated by 5 pixels in both X and Y directions; and finally (v) combinational effect of rotation, scaling and translation represented by RTS.

As explained in the spatial transformater [39], the spatial module can be inserted into any place of the existing network, and then the spatial network is trained together with the existing network. Hence, our combined network (VDST-VDSR) is trained end-to-end for a number of iterations until its convergence is reached. For the convenience of understanding and verification, we make our training codes and testing codes available on the Github: https://github.com/HossamMKasem/A-Very-Deep-Spatial-Transformer-Towards-Robust-Single-Image-Super-Resolution.

We train all experiments over 10 epochs with batch size 25, in which the learning rate is set to 0.01 and the learning rate decay is set to be 0. All the models that are utilized for comparison purposes are trained in an end-to-end manner. All the training is performed on NVIDIA Tesla P100. The evaluation results of all experiments are presented in term of two metrics widely used in the SR research community, which are Peak-signal-to-noise-ratio (PSNR) and Structural Similarity

FIGURE 5. Illustration of seven natural images used in our experiments.

TABLE 1. Comparisons of the network structure between the proposed DRLN and the existing STN.

Layers	STN [39]	No. of Parameters	DRLN	No. of Parameters
Prepossessing	Max Pooling (2,2)	/	Max Pooling (2,2)	/
Feature Extraction	1x Cov(5,200,3) ReLU MaxPooling(2,2)	15200	1x Cov(3,64,3) ReLU	1792
	1x Cov(5,300,200) ReLU MaxPooling(2,2)	15180300	18 x Cov(3,64,64) ReLU	664407
			1x Cov(3,64,3) ReLU	1731
Fully Connected layer	Linear (2700,200)	540000	Linear (432,30)	12960
Classifier	Linear (200,6)	1200	Linear (30,6)	1800
Total no. of parameters		16276700		682987

Index (SSIM). While the former metric is considered to be an objective measurement of the quality for generated HR image, the latter metric is considered to measure subjective quality of the generated HR image.

B. EXPERIMENTS MEASURED BY THE NUMBER OF PARAMETERS FOR DEEP-LEARNING NETWORKS

By using the number of parameters as the evaluating criteria, Table 1 summarizes the comparative results between our proposed VDST and the existing STN. In Table 1, the convolution layer is represented by Conv(k_i, n_i, c_i), where the variables k_i, n_i, c_i represent the filter size, the number of filters and the number of feature maps, respectively, and the linear layer is represented by Linear (m_i, o_i), where the variables m_i, o_i represent the size of the input vector and the size of the output vector, respectively. Note that the classifier works like a linear regression module to estimate the affine transformation parameters. It consists of one fully-connected layer where its network structure is described in Table 1.

As seen, the results given in Table 1 indicate that our proposed VDST is powerful in structure, cost-effective in learning, and capturing well the contextual information from the input images, leading to the improved performances.

Inspired by the existing work [51], which is a typical application using the STN, we follow their design to produce 200 feature maps as the first convolutional layer and 300 feature maps as the second layer, as shown in Table 1. For our proposed VDST, we set the first 19 layers to generate 64 feature maps and the last convolutional layer to generate 3 feature maps. Overall, our VDST is much deeper and thinner than the existing CNN for STN, paving the way for extracting deep contextual information.

TABLE 2. Experimental results (PSNR/SSIM) achieved by the existing various network structures compared with our proposed framework with the scaling factor ×2.

Transformation type	Test Dataset	VDSR [41]	STN-VDSR [39]	Our proposed VDST-VDSR
Rotation (R)	BSD	22.02/0.7240	22.57/0.8138	***24.06/0.8304***
	Manga	21.67/0.7819	23.05/0.8644	***24.18/0.8877***
	Set5	23.31/0.7896	25.98/0.8877	***27.07/0.9038***
	Set14	22.67/0.7646	23.28/0.8462	***25.19/0.8685***
	Urban	21.70/0.7277	22.63/0.8227	***24.23/0.8383***
Scaling (S)	BSD	25.71/0.8047	24.95/0.8230	***25.94/0.8277***
	Manga	26.36/0.8670	25.37/0.8740	***25.92/0.8783***
	Set5	27.83/0.8610	27.28/0.8294	***27.77/0.8939***
	Set14	26.19/0.8409	25.86/0.8544	***25.97/0.8640***
	Urban	25.79/0.8237	24.01/0.8134	***24.22/0.8179***
Rotation and scaling (RS)	BSD	21.86/0.8740	22.33/0.7823	***22.77/0.7851***
	Manga	22.25/0.7972	21.93/0.8340	***22.41/0.8392***
	Set5	22.12/0.8035	24.18/0.8486	***24.59/0.8533***
	Set14	21.60/0.7795	22.72/0.8151	***23.26/0.8339***
	Urban	20.69/0.7388	21.79/0.7701	***22.39/0.7770***
Translation (T)	BSD	17.14/0.6573	22.07/0.7524	***24.51/0.7971***
	Manga	18.08/0.6739	21.27/0.8039	***24.17/0.856***
	Set5	17.33/0.7020	23.15/0.8179	***26.97/0.8785***
	Set14	16.29/0.6740	21.11/0.7879	***25.83/0.8394***
	Urban	16.52/0.6356	21.45/0.7378	***24.13/0.8067***
RTS	BSD	18.17/0.6921	20.74/0.7521	***22.01/0.7753***
	Manga	16.34/0.7413	19.82/0.8009	***21.79/0.8370***
	Set5	18.71/0.7194	21.32/0.8110	***23.63/0.8399***
	Set14	17.58/0.7194	20.86/0.7843	***22.65/0.8165***
	Urban	17.45/0.6711	20.14/0.7346	***22.15/0.7818***

From Table 1, it can seen that the parameters of our proposed VDST is much less than that of the existing STN due to the fact that it uses less number of feature maps and much deeper layers in comparison with the existing STN.

C. EXPERIMENTS ON EFFECTIVENESS AND ROBUSTNESS OF OUR PROPOSED FRAMEWORK

To validate the effectiveness and the accuracy of the proposed deep spatial transformer using DRLN, we have carried out a range of experiments upon natural images to estimate the affine transformation parameters and mitigate the geometric transformation effects. Fig.5 illustrates seven samples of such natural images adopted as the test images in our experiments.

To evaluate the performance of our proposed, we transformed the original images using two different transformation effects. First, we translate the original images by 5 pixels in X and Y directions and explore the robustness of our proposed VDST against geometric transformations. For visual inspections and comparisons between our proposed VDST and the existing STN [39], Fig.6 shows five samples of the

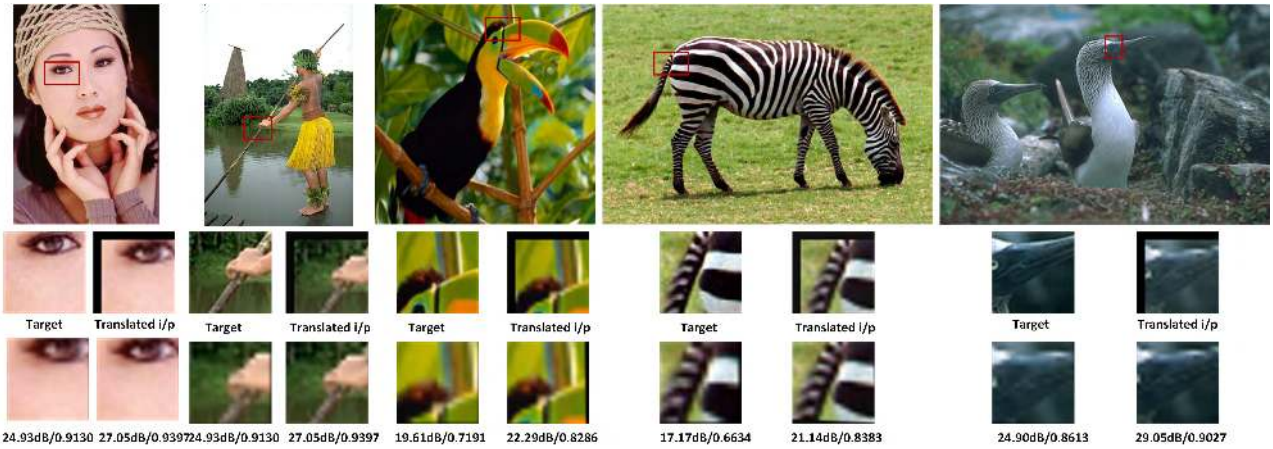

FIGURE 6. Visual comparison samples of our proposed VDST and existing STN: (i) the original images are translated in X, and Y direction by 5 Pixels. (ii) The target input patches and the rotated input patches are illustrated at the bottom of each original image. The output patches of our proposed framework and the existing STN are given on the last row of images underneath each original image, and the bottom row presents the corresponding outputs measured in PSNR/SSIM values. The lower row on the right is the output of the existing STN, and on the left is the output of our proposed VDST, respectively.

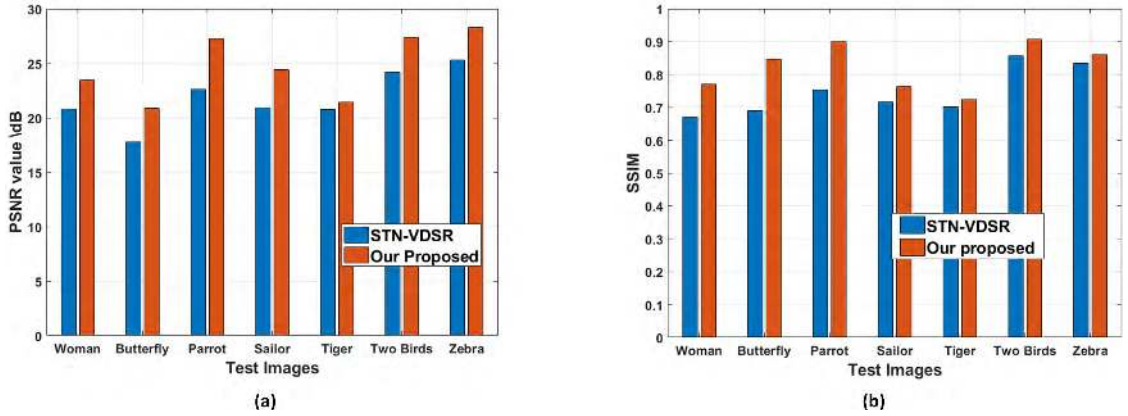

FIGURE 7. PSNR and SSIM values of our proposed framework and the combination of the existing STN and VDSR: The original images are translated in X, and Y direction by 5 pixels. The PSNR values are shown in (a), and SSIM values are shown in (b).

test images under the translation effect. It can be seen that VDST can estimate the affine parameters more accurately, and hence mitigate the transformation effects better than the existing STN. Further, the PSNR and SSIM values of our proposed spatial transformer is also much better than the exist STN.

Fig.7 shows the objective evaluation results in terms of PSNR/SSIM when the original images are translated in X and Y direction by 5 pixels, simulating another type of possible geometric distortion effect. As seen again, our proposed achieves higher PSNR/SSIM values over the existing STN. To test the robustness of our proposed against stronger transformation effects, we apply the combinational effect of rotation, translation and scaling, and the corresponding experimental results are illustrated in Fig. 8. As seen, the simulation results indicate that our proposed spatial transformer still outperforms the existing STN in estimating the affine parameters and mitigating the transformation effects.

As our proposed embeds a deeper DRLN and hence capable of exploiting more contextual information out of the input images, more levels of features can be provided for the classifier, and hence estimate the affine transformation parameters more accurately. Based on the deep estimation of parameters, our proposed VDST is able to mitigate the geometric transformation effects and produce an image in the same orientation as the desired image. Together with the experimental results summarized in Table 1, we can conclude that: (i) our proposed spatial transformer requires less number of network parameters than that of the existing spatial transformer, providing better cost-effectiveness; (ii) our proposed VDST outperforms the existing spatial transformer in estimating the affine parameters and hence achieving better mitigation for the possible geometric transformations.

To test the robustness of our proposed VDST against all the simulated geometric transformations, including rotation (R), scaling (S), rotation and scaling (RS), translation (T), and combination of rotation, scaling and translation (RTS), we carried out five experiments in total to evaluate the performances of our proposed in comparison with the existing state of the art VDSR [41], and illustrate how spatial

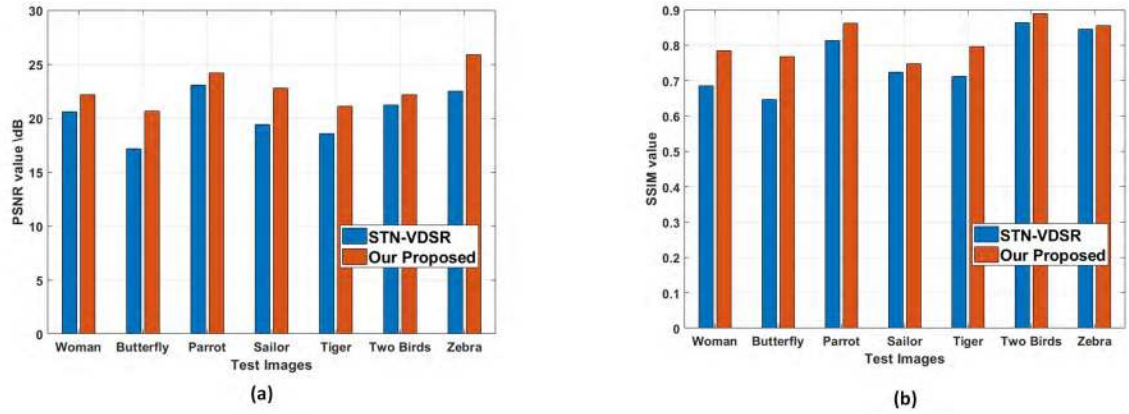

FIGURE 8. PSNR and SSIM values of our proposed framework and the combination of the existing STN and VDSR: The original images are rotated, translated and scaled (RTS). The PSNR values are shown in (a), and SSIM values are shown in (b).

TABLE 3. Experimental results (PSNR/SSIM) achieved by the existing various network structures compared with our proposed framework with the scaling factor ×3.

Transformation type	Test Dataset	VDSR [41]	STN-VDSR [39]	Our proposed VDST-VDSR
Rotation (R)	BSD	21.80/0.7380	22.48/0.8226	***23.77/0.8225***
	Manga	20.60/0.7608	23.67/0.8749	***23.80/0.8761***
	Set5	22.30/0.700	25.95/0.8975	***26.79/0.8976***
	Set14	21.72/0.7614	25.53/0.8591	***24.81/0.8605***
	Urban	20.70/0.720	22.35/0.8160	***23.77/0.8189***
Scaling (S)	BSD	24.37/0.8025	23.35/0.8144	***23.77/0.8225***
	Manga	25.75/0.8628	23.39/0.8647	***24.53/0.8764***
	Set5	26.43/0.8608	26.63/0.8864	***27.26/0.8936***
	Set14	25.38/0.8401	25.57/0.8562	***25.70/0.8600***
	Urban	23.62/0.7911	23.09/0.8010	***23.69/0.8164***
Rotation and scaling (RS)	BSD	21.61/0.7247	21.64/0.7786	***22.09/0.7779***
	Manga	20.48/0.7815	21.26/0.8261	***21.86/0.8298***
	Set5	22.05/0.7869	23.31/0.8581	***24.17/0.856***
	Set14	21.59/0.7632	22.05/0.8128	***22.75/0.8154***
	Urban	20.56/0.7244	20.99/0.7543	***21.64/0.7670***
Translation (T)	BSD	17.05/0.6546	20.60/0.7459	***24.20/0.7961***
	Manga	14.99/0.6700	19.93/0.7926	***23.91/0.8568***
	Set5	17.03/0.6972	21.22/0.8058	***26.68/0.8761***
	Set14	16.29/0.6716	20.59/0.7768	***24.80/0.8384***
	Urban	16.51/0.6324	20.23/0.7264	***23.72/0.7982***
RTS	BSD	18.10/0.6884	19.85/0.7499	***21.09/0.7730***
	Manga	16.34/0.7170	19.16/0.7954	***21.82/0.8329***
	Set5	18.63/0.7396	20.37/0.8073	***23.43/0.8256***
	Set14	17.49/0.7121	20.04/0.7801	***22.48/0.8135***
	Urban	17.45/0.6710	19.49/0.7295	***21.98/0.7728***

TABLE 4. Experimental results (PSNR/SSIM) achieved by the existing various network structures compared with our proposed framework with the scaling factor ×4.

Transformation type	Test Dataset	VDSR [41]	STN-VDSR [39]	Our proposed VDST-VDSR
Rotation (R)	BSD	21.69/0.7148	22.14/0.8127	***23.66/0.8136***
	Manga	20.59/0.7444	22.99/0.8612	***23.30/0.8620***
	Set5	22.29/0.7596	25.24/0.8859	***26.38/0.8877***
	Set14	21.68/0.758	23.26/0.8475	***24.60/0.8489***
	Urban	20.56/0.7163	21.89/0.7997	***23.28/0.8018***
Scaling (S)	BSD	24.04/0.8011	23.09/0.8026	***23.34/0.8163***
	Manga	25.54/0.8609	23.77/0.8620	***23.78/0.8640***
	Set5	26.23/0.8757	26.53/0.8767	***20.07/0.7845***
	Set14	25.20/0.8308	24.62/0.8485	***24.64/0.8486***
	Urban	23.39/0.7800	23.19/0.8006	***23.34/0.8008***
Rotation and scaling (RS)	BSD	21.49/0.707	21.64/0.7605	***21.96/0.7623***
	Manga	20.45/0.7680	21.11/0.8011	***21.74/0.8136***
	Set5	21.09/0.7623	23.31/0.8351	***23.99/0.8450***
	Set14	21.56/0.7612	22.02/0.8051	***22.53/0.8135***
	Urban	20.34/0.7112	20.79/0.7408	***21.33/0.7591***
Translation (T)	BSD	16.99/0.6501	15.53/0.7463	***19.32/0.7512***
	Manga	14.95/0.6670	18.26/0.7882	***19.69/0.7949***
	Set5	17.26/0.6930	19.01/0.7988	***19.76/0.8053***
	Set14	16.27/0.6691	18.93/0.7735	***19.41/0.7784***
	Urban	16.51/0.6304	18.57/0.7239	***19.05/0.7296***
RTS	BSD	18.00/0.6855	18.28/0.7487	***21.74/0.7755***
	Manga	16.32/0.7023	18.3/0.7905	***21.49/0.8292***
	Set5	18.35/0.7259	18.80/0.8014	***23.44/0.8442***
	Set14	17.12/0.7161	18.73/0.7762	***22.4170.8131***
	Urban	17.45/0.6703	18.41/0.7263	***21.06/0.7685***

transformation could be alleviated by our proposed framework. As our proposed framework is the first attempt for robust single image super-resolution and there exist no close work to compare, we combine the existing spatial transform network (STN) with the VDSR together to formulate a new benchmark, referred to as STN-VDSR.

Table 2, Table 3 and Table 4 show the comparative experimental results between the existing VDSR, STN-VDSR and our proposed VDST-VDSR under a number of the transformation effects with the scaling factors ×2, ×3, ×4, respectively. By comparing the results given in Table 2, we can observe that our proposed method significantly outperforms the VDSR, and STN-VDSR in terms of both PSNR and SSIM. From the simulation results for the rotation effect, as an example, it can be further seen that our proposed method has higher PSNR scores by at least 2dB in all test datasets across all rotation angles. In terms of SSIM, our proposed also achieves higher values than that of VDSR and STN-VDSR.

By comparing and examining the simulation results across Table 3 and Table 4 for the effect of RS, T, and RTS, it can also be seen that our VDST-VDSR outperforms the existing VDSR and STN-VDSR in both PSNR and SSIM values. Correspondingly, it can be concluded that our proposed successfully provides a well-validated solution for tackling the effects of geometric transformations, and achieve a robust single image super-resolution. This achieved improvement is due to the capability of our proposed method in exploiting the contextual information spread over the image regions as well as extraction of deeper features from the input images. In other words, the experimental results confirm that our proposed VDST-VDSR improves the accuracy of estimating the transformation parameters, leading to the superior performances for robust single image super-resolution.

To visually compare the experimental results between our proposed and the existing benchmarks, we illustrate a num-

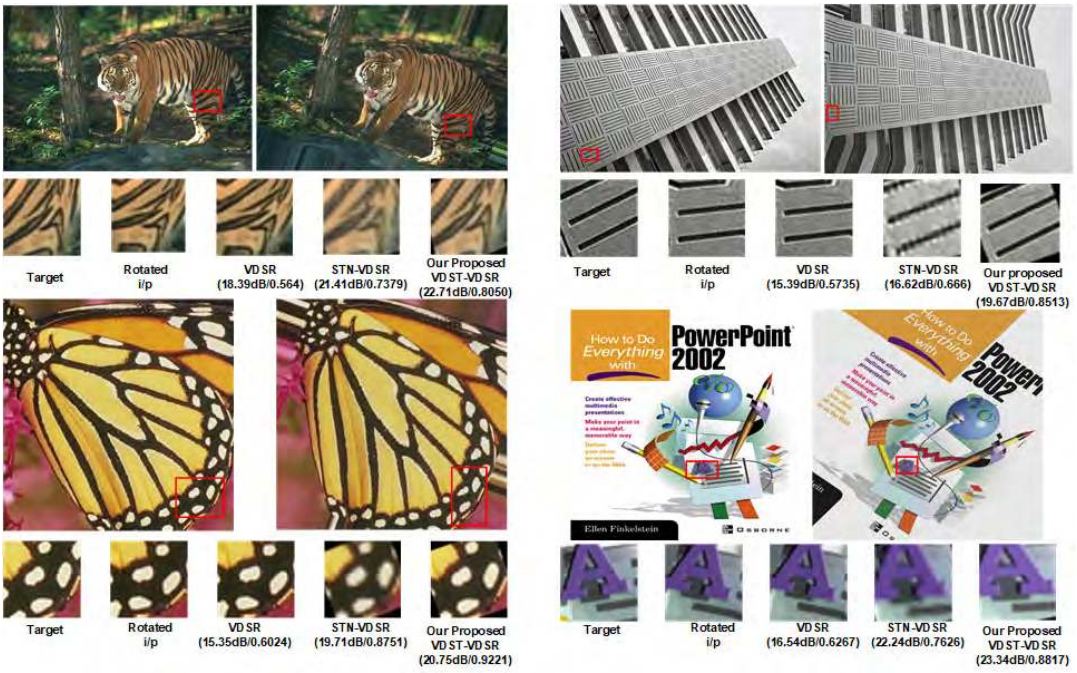

FIGURE 9. Visual comparison samples of our proposed VDST-VDSR, VDSR and STN-VDSR. Top left row: original image and rotated image by 20 degree. Under each row: The desired output, rotated input patch, VDSR output (PSNR/SSIM),STN-VDSR output (PSNR/SSIM), and our proposed framework (PSNR/SSIM), respectively.

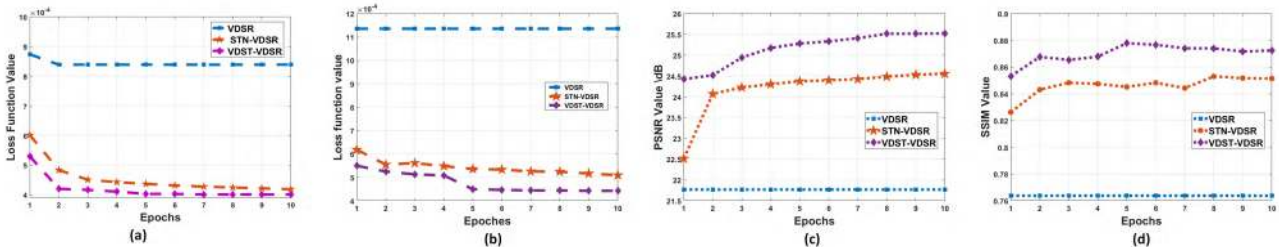

FIGURE 10. Convergence curves of the objective function, and the PSNR/SSIM values under the rotation effect: (a) Loss function values during the training phase. (b) Loss function values during the validation phase; (c) The PSNR values; and (d) The SSIM values.

ber of samples in Fig.9 for visual inspections and subjective assessments. In Fig.9, we show visual comparisons on BSDS 100, MANGA 109, URBAN 100 and SET 5 with a scaling factor of 2× under the rotation(R) effect. As seen, our proposed framework accurately mitigates the rotation effect, and reconstructs the straight lines and the grid patterns well, such as the stripes on the tiger, the window, and the lines on the butterfly. The visual inspection supports that our proposed VDST-VDSR outperforms both VDSR and STN-VDSR.

D. EXPERIMENT ON CONVERGENCE

We conduct an experiment to test the convergence of the proposed VDST-VDSR according to the loss function in (4) compared with STN-VDSR, and VDSR. All the results are illustrated as convergence curves during the training, and testing steps. In addition, we test the performance of our proposed framework by calculating the PSNR/SSIM values during each epoch, and hence compare these values with the existing STN.

We test the convergence of our proposed framework under the rotation(R) effect, the curves of the convergence, and PSNR/SSIM are presented in Fig. 10. From the results shown in Fig. 10, we can conclude that our proposed framework takes less number of epochs to reach the convergence point than that by STN-VDSR. our proposed framework takes no more than 5 epochs to reach the convergence point. Further, the PSNR/SSIM curves validate the superiority of our proposed VDST in terms of both PSNR and SSIM values. To provide a comprehensive assessment, we test the convergence of our proposed framework under a stronger transformation effect which is a combination of rotation, translation and scaling (RTS), and hence compare the proposed frame work convergence curves with STN-VDSR curves as shown in Fig. 11. The results in Fig. 11 confirm that our proposed VDST is able to reach the convergence point faster that than

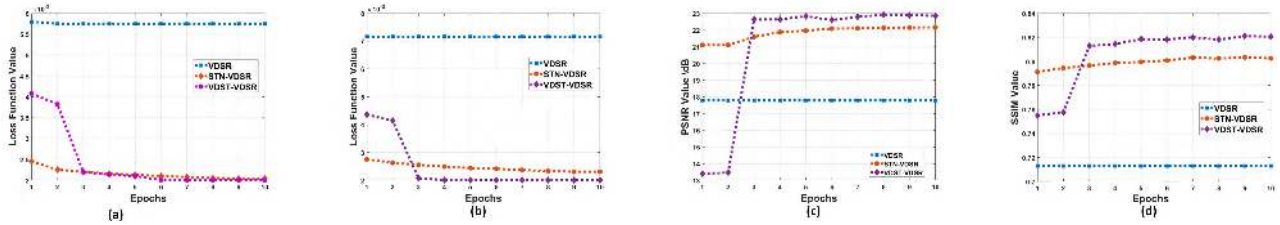

FIGURE 11. Convergence curves of the objective function, and the PSNR/SSIM values under RTS effect: (a) Loss function values during the training phase.(b) Loss function values during the validation phase; (c) PSNR values; and (d) SSIM values.

TABLE 5. Experimental results (PSNR/SSIM) achieved by the existing various network structures compared with our proposed framework with the scaling factor ×2.

Transformation type	Test Dataset	VDSR	VDSR-STN	Our Proposed VDSR-VDST
Rotation (R)	BSD	22.02/0.7240	23.57/0.8254	***24.46/0.8320***
	Manga	21.67/0.7819	23.04/0.8832	***24.77/0.8871***
	Set5	23.31/0.7896	26.35/0.8925	***26.87/0.8948***
	Set14	22.67/0.7646	23.64/0.8620	***25.26/0.8703***
	Urban	21.70/0.7277	23.62/0.8260	***25.04/0.8609***
Scaling (R)	BSD	25.71/0.8047	24.30/0.8092	***25.27/0.8187***
	Manga	23.36/0.8047	23.81/0.8638	***25.71/0.8734***
	Set5	27.83/0.8610	26.56/0.8842	***27.38/0.8821***
	Set14	26.19/0.8409	24.69/0.8451	***25.65/0.8456***
	Urban	25.79/0.8237	23.38/0.7986	***24.32/0.7992***
Rotation and scaling (RS)	BSD	21.86/0.8740	21.64/0.7615	***21.89/0.7725***
	Manga	22.25/0.7972	21.14/0.8132	***21.70/0.8319***
	Set5	22.12/0.8035	23.20/0.8283	***24.68/0.8452***
	Set14	21.60/0.7795	21.93/0.7974	***23.49/0.8146***
	Urban	20.69/0.7388	21.04/0.7463	***21.55/0.7689***
Translation (T)	BSD	17.14/0.6573	22.07/0.7524	***24.26/0.7585***
	Manga	18.08/0.6739	21.26/0.8039	***23.61/0.8357***
	Set5	17.33/0.7020	23.15/0.8179	***26.78/0.8301***
	Set14	16.29/0.6740	22.11/0.7879	***24.42/0.7889***
	Urban	16.52/0.6356	21.54/0.7378	***23.79/0.7533***
RTS	BSD	18.17/0.6921	20.74/0.7521	***21.87/0.7611***
	Manga	16.34/0.7413	19.81/0.8009	***20.39/0.8131***
	Set5	18.71/0.7194	21.31/0.8110	***22.87/0.8202***
	Set14	17.58/0.7194	20.85/0.7843	***22.26/0.7962***
	Urban	17.45/0.6711	20.13/0.7346	***21.58/0.7509***

of STN-VDSR, and meanwhile, our proposed framework still have higher PSNR and SSIM values than that of STN-VDSR.

E. EXPERIMENT ON CHANGING THE ORDER OF GEOMETRIC TRANSFORMATION AND SUPER-RESOLUTION

To further evaluate our proposed, we conduct another experiment to test the performance of the proposed VDST-VDSR in case it is required that the position of the spatial transform module and the VDSR module need to be changed. In other words, we connect the VDSR model to the input image to generate the HR image, followed by our VDST to mitigate the transformation effects. For fair comparison, we change the order of the existing STN and VDSR module as the comparing benchmark. Correspondingly, we refer our new method as VDSR-VDST and the existing method as VDSR-STN.

All the results are illustrated in Table 5. From the results shown in Table 5, we can see that our proposed VDSR-VDST outperforms the existing VDSR and VDSR-STN. We note that the two versions of proposed method (VDST-VDSR and VDSR-VDST) in Table 2 and Table 5 produce similar results in terms of PSNR and SSIM values (with 0.32dB change in PSNR on average). Our analysis shows that this phenomenon

TABLE 6. Experimental results (PSNR/SSIM) achieved by the existing various network structures compared with our proposed framework with the scaling factor ×2.

Image Number	Transformation Type	VDSR	STN-VDSR	Our proposed VDST-VDSR
Image_1	Rotation (R)	25.65/0.7687	25.97/0.7962	26.53/0.8250
	Scaling (S)	26.04/0.7808	26.93/0.8235	27.24/0.8246
	Rotation and Scaling (RS)	20.56/0.7716	22.68/0.8164	25.75/0.8180
	Translation (T)	14.84/0.6890	25.91/0.8028	27.61/0.8104
	RTS	17.39/0.7283	25.25/0.8017	25.63/0.8204
Image_2	Rotation (R)	20.03/0.6861	22.31/0.7845	22.94/0.7934
	Scaling (S)	21.89/0.7686	22.85/0.7961	23.82/.8250
	Rotation and Scaling (RS)	19.75/0.6897	21.10/0.7261	22.59/0.7452
	Translation (T)	15.32/0.6099	21.61/0.7145	23.450.7616
	RTS	16.96/0.6528	21.10/0.7159	21.60/0.7363
Image_3	Rotation (R)	23.46/0.7404	24.19/0.8260	25.22/0.8300
	Scaling (S)	23.09/0.8235	25.95/0.8473	26.40/0.8562
	Rotation and Scaling (RS)	19.18/0.7456	21.97/0.7856	23.92/0.7866
	Translation (T)	14.711/0.6601	22.94/0.7648	25.63/0.8109
	RTS	16.74/0.9687	22.51/0.7603	22.64/0.7851
Image_4	Rotation (R)	22.56/0.7897	25.21/0.8578	27.29/0.8659
	Scaling (S)	23.72/0.8310	25.29/0.8629	27.11/0.8623
	Rotation and Scaling (RS)	20.47/0.8018	22.04/0.8300	22.91/0.8320
	Translation (T)	14.59/0.6928	26.54/0.8222	28.21/0.8462
	RTS	16.68/0.7435	20.96/0.8196	23.49/0.8397
Image_5	Rotation (R)	20.87/0.6800	21.31/0.7947	21.70/0.8057
	Scaling (S)	21.79/0.7792	23.81/0.8111	24.32/0.8299
	Rotation and Scaling (RS)	18.73/0.6871	20.14/0.7284	20.63/0.7349
	Translation (T)	14.83/0.5870	20.43/0.7087	23.02/0.7701
	RTS	16.34/0.6503	19.85/0.7055	20.46/0.7651
Image_6	Rotation (R)	21.05/0.7097	22.88/0.8249	22.98/0.8305
	Scaling (S)	22.19/0.8036	23.20/0.8317	24.18/0.8328
	Rotation and Scaling (RS)	19.87/0.7154	20.88/0.7548	21.26/0.7626
	Translation (T)	15.36/0.6039	20.62/0.7340	23.53/0.7956
	RTS	16.84/0.6688	20.03/0.7294	21.50/0.7863
Image_7	Rotation (R)	22.03/0.7030	23.02/0.7949	23.58/0.8005
	Scaling (S)	21.25/0.7731	23.58/0.8044	23.91/0.8031
	Rotation and Scaling (RS)	14.72/0.6144	20.85/0.7298	22.88/0.7442
	Translation (T)	14.72/0.6144	21.81/7289	23.90/0.7732
	RTS	16.60/0.6693	21.59/0.7283	22.89/0.7701
Image_8	Rotation (R)	20.14/0.6794	23.25/0.7454	24.93/0.7559
	Scaling (S)	19.24/0.7070	20.95/0.7510	21.37/0.7508
	Rotation and Scaling (RS)	16.13/0.6926	17.728/0.7235	19.49/0.7231
	Translation (T)	13.25/0.6162	22.41/0.7035	23.42/0.7326
	RTS	15.27/0.6275	18.32/0.7150	20.53/0.7332
Image_9	Rotation (R)	21.96/0.7918	23.58/0.8904	24.11/0.8933
	Scaling (S)	21.62/0.8764	23.40/0.8723	23.95/0.8735
	Rotation and Scaling (RS)	18.98/0.8057	20.4/0.8184	20.96/0.8235
	Translation (T)	14.98/0.6751	20.55/0.8041	23.66/0.8345
	RTS	16.52/0.7273	19.28/0.7955	20.87/0.8395

is due to the fact that the spatial transformer is a linear module for affine transformation, yet the super resolution module is a nonlinear module (because of the ReLU nonlinear activation in the neural network). When both modules are interchanged, as a result, the nonlinearity of the super-resolution module inevitably causes the small changes in PSNR and SSIM values.

F. EXPERIMENTS ON REAL WORLD IMAGES

In real scenarios, the users can define the desired geometric transformations manually to achieve the simultaneous geometric transformation and super-resolution. To validate the effectiveness and the accuracy of our proposed method,

FIGURE 12. Illustration of nine real world images used in our experiments.

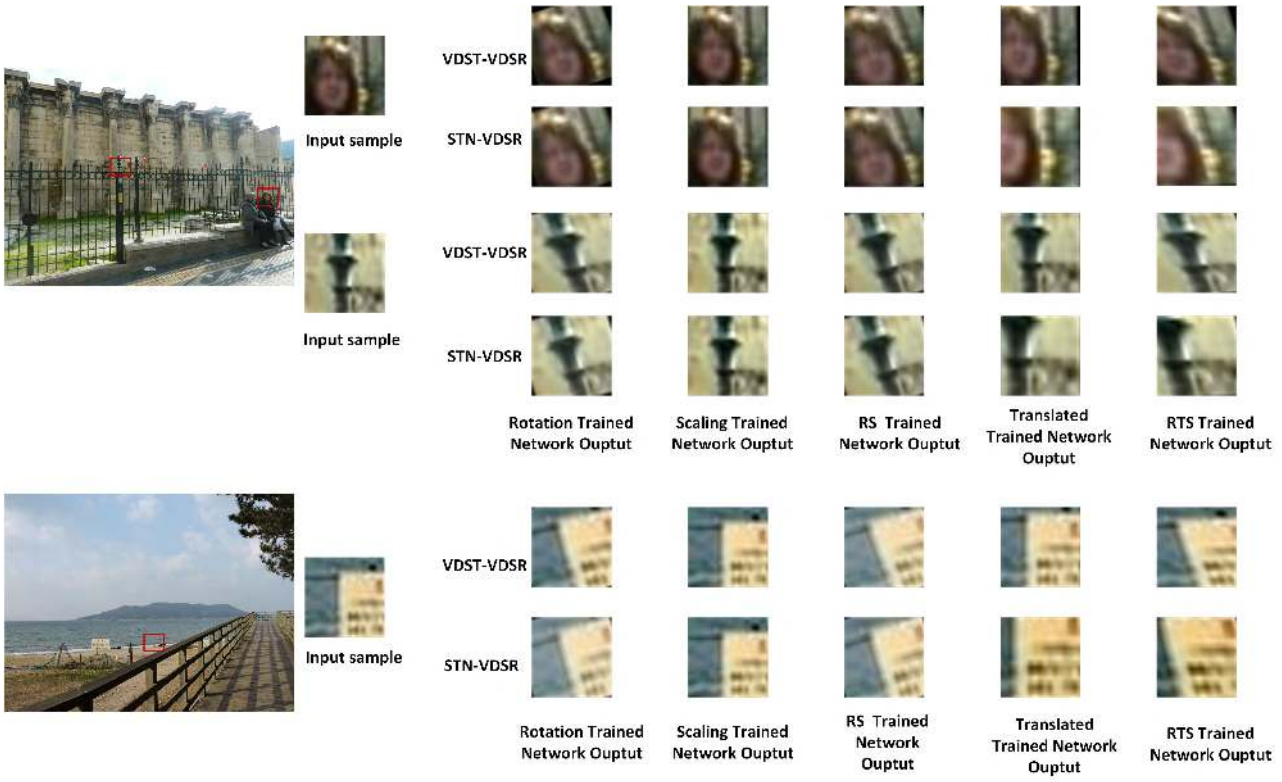

FIGURE 13. Illustration of real world images for geometric transformation and super-resolution.

we have conducted extensive experiments on real world images to estimate the affine transformation parameters and mitigate the geometric transformation effects. Fig.12 illustrates nine samples of real world images adopted as test images in our experiments. We have labelled the images used in the test by numbers from the top left to bottom right in Fig.12 (i.e. we give number 1 to the image in the left of the top row, number 2 to the next image in the top row, and number 6 to the left image in the bottom row and so on.). These images were captured using Nikon coolpix P500 camera. We carried out five experiments to evaluate the performances of our proposed method in comparison with the existing state of the art VDSR [41] and the combination of STN and VDSR. We tested the methods under five different transformation effects, including rotation (R), scaling (S), rotation and scaling (RS), translation (T), and combination of rotation, scaling and translation (RTS).

Table 6 shows the experimental results of the existing VDSR, STN-VDSR and our proposed VDST-VDSR under a number of various transformation effects with scale $\times 2$. By comparing the results shown in Table 6, we can conclude that our proposed method shows superior performances compared with the existing VDSR and STN-VDSR in terms of PSNR and SSIM across all the testing images. The results in Table 6 show that our proposed method produces higher PSNR and SSIM values for the images that contain more

foreground shapes. Compared with the benchmarks, our proposed achieves at least 1dB higher PSNR scores across all the tested images.

Fig.13 illustrates the results of our further experiments upon real-world images, where three input samples are shown, containing rich details such as face, fence, and shapes for detailed comparative analysis. As seen, the proposed VDST-VDSR can preserve the image details while successfully performing geometric transformation and super-resolutions. In contrast, the combination of STN and VDSR, i.e., STN-VDSR, fails to achieve the desired translation and RTS effects, where more blurry results are produced and noticeable. Therefore, such experimental results further validate the improved performances and advantages achieved by our proposed.

IV. CONCLUSION

In this paper, we proposed a novel spatially transformed deep framework to achieve robust single image super-resolution. Our proposed framework can simultaneously perform geometric corrections and super-resolution reconstruction, which, to the best of our knowledge, is the first deep learning-based method to tackle such problems. Our proposed framework is based on improving the existing STN with a deep residual learning network (DRLN) to extract deeper features and exploit contextual information against the process of possible geometric transformations. Extensive evaluations on widely used datasets with different transformation effects demonstrate that the proposed deep framework successfully mitigates the effect of the geometric transformations and achieves superior performances in comparison with the existing state of arts, including VDSR and STN-VDSR, the combination of the existing state of the arts for both spatial transformation and image super-resolution. Future research can be identified to consider multiple networks for handling wider range of transformations, leading to a further optimization of the SR performances and improvement upon its robustness against not only the geometric transformations but also distortions.

ACKNOWLEDGMENT

(Jianmin Jiang and Hossam M. Kasem are co-first authors.)

REFERENCES

[1] H. Wang, X. Gao, K. Zhang, and J. Li, "Single-image super-resolution using active-sampling Gaussian process regression," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 935–948, Feb. 2016.

[2] Y. Zhang, J. Liu, W. Yang, and Z. Guo, "Image super-resolution based on structure-modulated sparse representation," *IEEE Trans. Hum.-Mach. Syst.*, vol. 24, no. 9, pp. 2797–2810, Sep. 2015.

[3] Y. Li, Y. Wang, Y. Li, L. Jiao, X. Zhang, and R. Stolkin, "Single image super-resolution reconstruction based on genetic algorithm and regularization prior model," *Inf. Sci.*, vol. 372, pp. 196–207, Dec. 2016.

[4] T. Huang and R. Tsai, "Multi-frame image restoration and registration," in *Advances in Computer Vision and Image Processing*. Greenwich, CT, USA: JAI Press, 1984, pp. 317–339.

[5] R. Chao, X. He, and T. Q. Nguyen, "Single image super-resolution via adaptive high-dimensional non-local total variation and adaptive geometric feature," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 90–106, Jan. 2017.

[6] J. Jiang, X. Ma, C. Chen, T. Lu, Z. Wang, and J. Ma, "Single image super-resolution via locally regularized anchored neighborhood regression and nonlocal means," *IEEE Trans. Multimedia*, vol. 19, no. 1, pp. 15–26, Jan. 2017.

[7] L. J. Deng, W. Guo, and T.-Z. Huang, "Single image super-resolution by approximated Heaviside functions," *Inf. Sci.*, vol. 348, pp. 107–123, Jun. 2016.

[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.

[9] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[10] Q. Yuan, L. Zhang, and H. Shen, "Multiframe super-resolution employing a spatially weighted total variation model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 3, pp. 379–392, Mar. 2012.

[11] A. W. M. van Eekeren, K. Schutte, and L. J. van Vliet, "Multiframe super-resolution reconstruction of small moving objects," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2901–2912, Nov. 2010.

[12] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1127–1133, Jun. 2010.

[13] Z. Zhu, F. Guo, H. Yu, and C. Chen, "Fast single image super-resolution via self-example learning and sparse representation," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2178–2190, Dec. 2014.

[14] N. Qi, Y. Shi, X. Sun, W. Ding, and B. Yin, "Single image super-resolution via 2D sparse representation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jun./Jul. 2015, pp. 1–6.

[15] S. Rhee and M. G. Kang, "Discrete cosine transform based regularized high-resolution image reconstruction algorithm," *Opt. Eng.*, vol. 38, no. 8, pp. 1348–1356, 1999.

[16] H. Demirel, S. Izadpanahi, and G. Anbarjafari, "Improved motion-based localized super resolution technique using discrete wavelet transform for low resolution video enhancement," in *Proc. 17th Eur. Signal Process. Conf.*, Aug. 2009, pp. 1097–1101.

[17] L. Yue, H. Shen, J. Li, Q. Yuanc, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal Process.*, vol. 128, pp. 389–408, Nov. 2016.

[18] M. S. Alam, J. G. Bognar, R. C. Hardie, and B. J. Yasuda, "Infrared image registration and high-resolution reconstruction using multiple translationally shifted aliased video frames," *IEEE Trans. Instrum. Meas.*, vol. 49, no. 5, pp. 915–923, Oct. 2002.

[19] B. Wan, L. Meng, D. Ming, H. Qi, Y. Hu, and K. D. K. Luk, "Video image super-resolution restoration based on iterative back-projection algorithm," in *Proc. IEEE Int. Conf. Comput. Intell. Meas. Syst. Appl.*, May 2009, pp. 46–49.

[20] H. Stark and P. Oskoui, "High-resolution image recovery from image-plane arrays, using convex projections," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 6, no. 11, pp. 1715–1726, 1989.

[21] K. Zhang, X. Gao, D. Tao, and X. Li, "Single image super-resolution with non-local means and steering kernel regression," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4544–4556, Nov. 2012.

[22] T. Sigitani, Y. Iiguni, and H. Maeda, "Image interpolation for progressive transmission by using radial basis function networks," *IEEE Trans. Neural Netw.*, vol. 10, no. 2, pp. 381–390, Mar. 1999.

[23] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the nonlocal-means to super-resolution reconstruction," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 36–51, Jan. 2009.

[24] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[25] X. Li, H. He, R. Wang, and D. Tao, "Single image superresolution via directional group sparsity and directional features," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2874–2888, Sep. 2015.

[26] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, May 2003.

[27] Y.-W. Tai, S. Liu, M. S. Brown, and S. Lin, "Super resolution using edge prior and single image detail synthesis," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2400–2407.

[28] J. Sun, J. Sun, Z. Xu, and H.-Y. Shum, "Gradient profile prior and its applications in image super-resolution and enhancement," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1529–1542, Jun. 2011.

[29] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.
[30] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, Jul. 2011.
[31] T. Peleg and M. Elad, "A statistical prediction model based on sparse representations for single image super-resolution," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2569–2582, Jun. 2014.
[32] V. Papyan and M. Elad, "Multi-scale patch-based image restoration," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 249–261, Jan. 2016.
[33] W. Dong, L. Zhang, R. Lukac, and G. Shi, "Sparse representation based image interpolation with nonlocal autoregressive modeling," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1382–1394, Apr. 2013.
[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[35] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 8692, 2014, pp. 184–199.
[36] C. Osendorfer, H. Soyer, and P. van der Smagt, "Image super-resolution with fast approximate convolutional sparse coding," in *Proc. Int. Conf. Neural Inf. Process.*, 2014, pp. 250–257.
[37] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3467–3478, Aug. 2012.
[38] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 606–615.
[39] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
[40] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 391–407.
[41] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.
[42] J. Johnson, A. Alahi, and F. F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[43] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 105–114.
[44] C.-H. Lin and S. Lucey, "Inverse compositional spatial transformer networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2252–2260.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[46] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.
[47] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. BMVC*, 2012.
[48] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surfaces*. Berlin, Germany: Springer, 2010, pp. 711–730.
[49] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5197–5206.
[50] Y. Matsui *et al.*, "Sketch-based manga retrieval using manga109 dataset," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21811–21838, 2017.
[51] D. CireçAn, U. Meier, J. Masci, and J. Schmidhuber, "2012 special issue: Multi-column deep neural network for traffic sign classification," *Neural Netw.*, vol. 32, no. 1, pp. 333–338, 2012.

JIANMIN JIANG received the Ph.D. degree from the University of Nottingham, Nottingham, U.K., in 1994. From 1997 to 2001, he was a Full Professor of computing with the University of Glamorgan, Pontypridd, U.K. In 2002, he joined the University of Bradford, Bradford, U.K., as a Chair Professor of digital media and the Director of the Digital Media and Systems Research Institute. He was a Full Professor with the University of Surrey, Guildford, U.K., from 2010 to 2014, and a Distinguished Chair Professor (1000-Plan) with Tianjin University, Tianjin, China, from 2010 to 2013. He is currently a Distinguished Chair Professor and the Director of the Research Institute for Future Media Computing, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He has published around 400 refereed research papers. His current research interests include, image/video processing in compressed domain, digital video coding, medical imaging, computer graphics, machine learning and AI applications in digital media processing, and retrieval and analysis. He was a Chartered Engineer, a Fellow of IEE and RSA, a member of EPSRC College in the U.K., and an EU FP-6/7 Evaluator.

HOSSAM M. KASEM received the Ph.D. degree from the Egypt-Japan University of Science and Technology, Egypt, in 2015. From 2015 to 2017, he was an Assistant Professor with Faculty of Engineering, Tanta University, Egypt. Since 2017, he has been a Postdoctoral Fellow with the Digital Media and Systems Research Institute, Shenzhen University, China. His research interests include deep learning applications in digital multimedia analysis, wireless communication, and signal processing application in multimedia.

KWOK-WAI HUNG received the B.Eng. and Ph.D. degrees from The Hong Kong Polytechnic University, in 2009 and 2014, respectively. From 2014 to 2016, he was a Research Engineer with Huawei and ASTRI. Since 2016, he has been an Assistant Professor with the Research Institute for Future Media Computing, Shenzhen University, China. His research interests include deep learning applications in digital multimedia processing and signal processing applications in multimedia.