IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 9, 2019, accepted February 27, 2019, date of current version March 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2903582

End-to-End Image Super-Resolution via Deep and Shallow Convolutional Networks

YIFAN WANG, LIJUN WANG, HONGYU WANG, (Member, IEEE), AND PEIHUA LI, (Member, IEEE)

School of Information and Communication Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China

Corresponding author: Hongyu Wang (whyu@dlut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61471082 and Grant 61671103.

ABSTRACT In this paper, we propose a new image super-resolution (SR) approach based on a convolutional neural network (CNN), which jointly learns the feature extraction, upsampling, and high-resolution (HR) reconstruction modules, yielding a completely end-to-end trainable deep CNN. However, directly training such a deep network in an end-to-end fashion is challenging, which takes a longer time to converge and may lead to sub-optimal results. To address this issue, we propose to jointly train an ensemble of deep and shallow networks. The shallow network with weaker learning capability restores the main structure of the image content, while the deep network with stronger representation power captures the high-frequency details. Since the shallow network is much easier to optimize, it significantly lowers the difficulty of deep network optimization during joint training. To further ensure more accurate restoration of HR images, the high-frequency details are reconstructed in a multi-scale manner to simultaneously incorporate both short- and long-range contextual information. The proposed method is extensively evaluated on widely adopted data sets and compares favorably against state-of-the-art methods. In-depth ablation studies are conducted to verify the contributions of different network designs to image SR, providing additional insights for future research.

INDEX TERMS Super-resolution, deep and shallow convolutional networks, end-to-end training, multi-scale reconstruction.

I. INTRODUCTION

Single image super-resolution (SR) aims at restoring the high resolution (HR) image with abundant high-frequency details from the low resolution (LR) observation. Given that multiple HR images can be down-sampled into the same LR image, SR as the reverse problem is inherently ill-posed with insufficient knowledge.

Recently, learning-based methods have attracted increasingly more attention and delivered superior performance in image SR. The basic idea is to learn the mapping function from the LR image to the HR counterpart using auxiliary data [1]–[6] (Fig. 1 (a)). A variety of machine learning algorithms, *e.g.*, sparse coding [4], [5], [7], anchored neighbor [8]–[11], regression trees or forests [12]–[14], have been adopted to learn the mapping function. Some recent efforts [15]–[21] have also been made to apply CNNs to image SR, and deliver impressive performance.

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

On popular idea for image SR with CNNs focuses on learning the residual between the HR image and the bicubic-interpolated LR image [20], assuming that the target HR image shares the similar main structure to the bicubic upsampled LR version (Fig. 1 (b)). However, the hand-crafted bicubic interpolation is not specifically designed for this purpose [22] and may hinder the final performance.

As opposed to the above CNN with bicubic interpolation based approaches, our method learns a direct mapping from LR to HR images with CNNs (Fig. 1 (c)). However, our preliminary experiments suggest that training a sophisticated deep network in such an end-to-end fashion is challenging, leading to sub-optimal results. To address this issue, we propose to jointly train an ensemble of deep and shallow networks (Fig. 2). Specifically, the shallow network is lightweight (*e.g.*, only 3 convolutional layers) and easier to optimize, while the deep network is elaborately designed and consists of three major procedures. Firstly, feature extraction is performed to map the original LR image into a deep feature space. The deep features are then upsampled to the

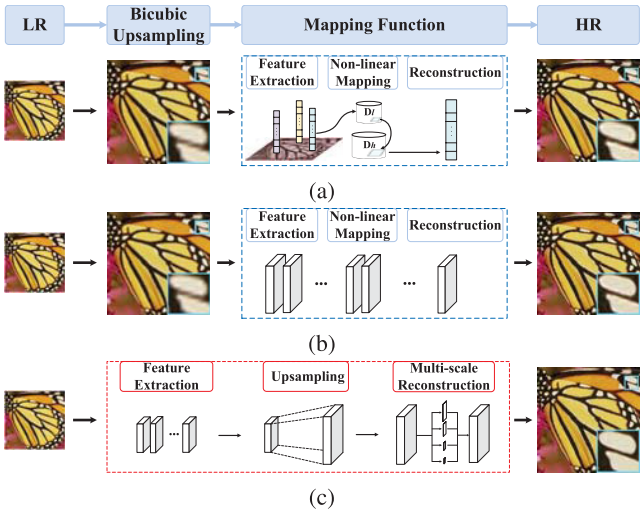

FIGURE 1. Overview of learning based SR methods. (a) Prior learning based methods with hand-designed interpolations, features and shallow models. (b) Prior deep learning based methods comprising hand-designed interpolations and automatically learned deep features. (c) A direct end-to-end mapping from LR to HR images with CNNs.

target spatial size with learned filters. Finally, the HR image is reconstructed by considering multi-scale context of the upsampled deep features. During joint training, the shallow network converges quickly and captures the major structure of the HR image, *i.e.*, mostly low-frequency content. As a consequence, the deep network is only responsible to restore the high-frequency details based on the main image structure, which effectively lowers the difficulty of deep network training.

The proposed network ensemble is similar to the above CNN with bicubic interpolation based approaches [21], [23], [24] in that the deep network is designed to learn the high-frequency residual content. However, different from these approaches, our method replaces the bicubic interpolation with a shallow network, allowing fully end-to-end trainable.

It has also been shown in [25] and [26] that reconstructing a pixel may depend on either short- or long-range contextual information. Some CNN-based approaches [16], [18], [19] rely on small image patches to predict the central pixel value, which is less effective for SR with large upscaling factors. In light of this observation, we propose to perform HR reconstruction in a multi-scale manner to simultaneously incorporate both short- and long-range contextual information, rendering more accurate HR image content.

The contributions of this paper are summarized as follows:

- We propose a shallow and deep network ensemble for SR, which is fully end-to-end trainable and effectively facilitates network optimization, which learns all procedures in an end-to-end manner and performs HR image reconstruction in multiple scales.
- We design a multi-scale HR image reconstruction module to simultaneously aggregate both short- and long-range contextual information, yielding more superior results.
- Extensive evaluations have been conducted to verify the above contributions. In-depth analysis on network architectures is performed from the perspective of SR.

The remainder of this paper is organized as follows. We review the related CNN based approaches in Section II. The proposed method is described in Section III. The experimental results are reported in Section IV. Section V concludes this paper.

II. RELATED WORK

Image SR can be generally classified into three categories, *i.e.*, interpolation-based [27], [28], reconstruction based [29]–[31], and learning-based methods [6], [9], [13]. Among them, learning-based methods become a hot research point in the field of image SR in recent years, whose basic idea is to formulate image SR as a nonlinear mapping from LR to HR images and learn the mapping using auxiliary data in a supervised manner.

The opening work is proposed by Freeman *et al.* [2], which employs Markov Random Field (MRF) and patch-based external examples to produce effective magnification. Inspired by [2], various methods have been developed subsequently. One of the representative methods is based on the sparse representation algorithm, which ensures that HR patches have a sparse linear representation over an over-complete dictionary of patches randomly sampled form similar images. Yang *et al.* [4] train LR and HR dictionaries jointly with the constraint that LR patches and the corresponding HR counterparts share the same sparse representation. This work is developed by [5] which employs K-SVD to train the coarse dictionary and Orthogonal Matching Pursuit (OMP) to solve the decomposition problem.

Based on the neighbor embedding algorithm, works of [8] and [32] super-resolve LR images with the assumption that LR and HR patches lie on low-dimensional nonlinear manifolds with locally similar geometry. To further improve computational efficiency, some techniques are put forward. Yang and Yang [9] cluster LR feature space into numerous subspaces and learn simple mapping functions for each subspace. Timofte *et al.* [10], [11] propose to use a number of linear regressors to locally anchor the neighbors. With the precalculated anchors and regressors, "$A+$" [11] increases SR performance both in terms of accuracy and speed.

Based on the regression trees or forests algorithm, another line of image SR technique [12]–[14] is proposed, which builds on linear multivariate regression models using leaf nodes and locally linearizes the mapping from LR to HR patches around centroids.

Deep learning based methods have recently been applied to image SR and delivered compelling performance [19], [22], [23], [33]. In [15], a CNN comprising three convolution layers is proposed for image SR. Later on, [18], [19] reformulate traditional sparse coding based method as deep networks and achieve promising results. Reference [34] restores the HR images using a Gibbs distribution as the conditional model, with its sufficient statistics predicted by a CNN.

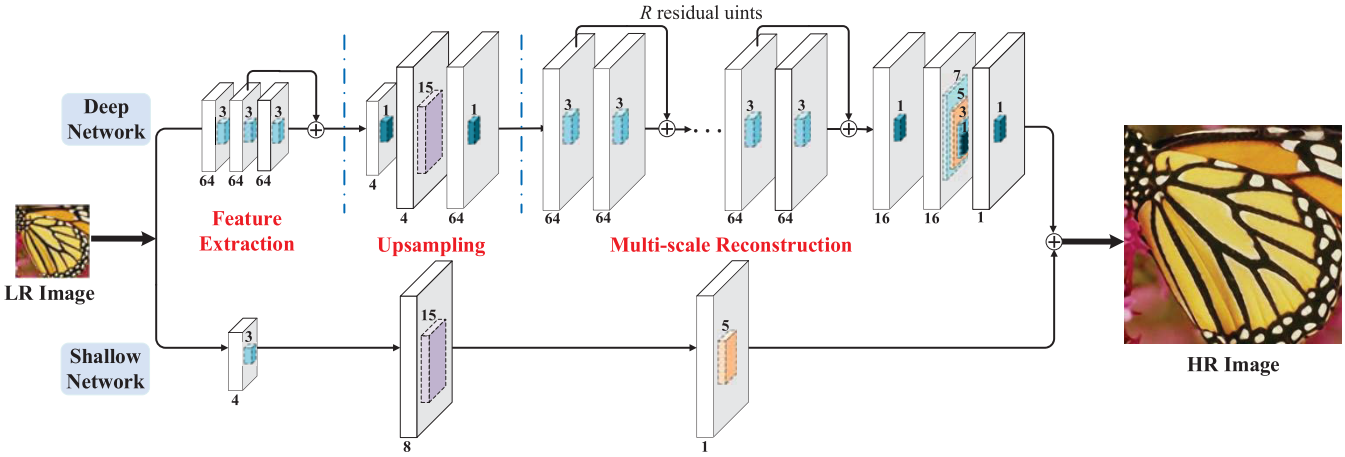

FIGURE 2. Network architecture of the proposed end-to-end deep and shallow (EEDS) networks.

Inspired by the residual prediction based methods [5], [10], [11], Kim *et al.* [21] propose a deep network with 20 convolutional layers to learn the residual between HR and LR images, which boosts performance by a large margin. The authors also present a deeply-recursive convolutional network to restore the HR images [20]. References [22] and [35] propose to extract feature maps in the LR space and learn to increase the resolution only at the very end of the network, which shows that the learned upscaling filters can further increase the accuracy of prediction. Subsequently, many other CNN-based techniques are applied in image SR, such as densely connected network [23], [24], recursive network [36] and cascade upsampling network [33], [37] and so on. Compared with the above works, we propose a fully end-to-end trainable system which adopts an ensemble of deep and shallow networks. In addition, a multi-scale HR image restoration module is also designed to aggregate both short- and long-range contextual information. These techniques have not been simultaneously explored in existing methods.

III. ARCHITECTURE

In this section, we introduce the proposed EEDS (End-to-End Deep and Shallow networks) method for image SR. Fig.2 overviews the architecture of the network ensemble comprising a deep and a shallow CNN. The deep CNN can be further divided into three modules: feature extraction, upsampling and multi-scale reconstruction. The complex architecture enables more accurate restoration of detailed HR image content, but makes the training more challenging. The shallow CNN with a more simple architecture is easier to converge, which aims at stabilizing the training process of the ensemble. We begin with the description of the deep CNN, and then introduce the architecture of the shallow one.

A. FEATURE EXTRACTION

In order to extract local features of high-frequency content, traditional shallow methods perform feature extraction by computing the first and second order gradients of the image patch, which is equivalent to filtering the input image with hand-designed, high-pass filters. Rather than manually designing these filters, deep learning based methods automatically learn these filters from training data. However, some works [11], [15], [21], [23] extract features from the coarse HR images, which is obtained by upsampling the LR images to the HR size with bicubic interpolation. We argue that the bicubic interpolation is not specifically designed for this purpose, and even damages important LR information that may play a central role in restoring the HR counterparts. Therefore, the proposed method adopts an alternative strategy [22], [35] and performs feature extraction directly on the original LR images with convolution layers.

Our feature extraction module consists of three convolution layers interleaved by Rectified Linear Unites (ReLUs) acting as nonlinear mappings. A shortcut connection with identity mapping is used to add the input feature map of the second layer to the output of the third layer, which is formulated as a ''residual unit''. As justified by [38], such residual unit can effectively facilitate gradients flow through multiple layers, thus accelerating deep network training. Similar structures have also been used in our reconstruction module (Section III-C). All three convolution layers have the same kernel size of 3×3 and generate feature maps of 64 channels. Zero padding is adopted to preserve the spatial size of the output feature maps.

B. UPSAMPLING

Given the extracted features from the original LR images, upsampling operation is performed to increase their spatial span to the target HR size. Instead of using hand-designed interpolation methods, we prefer a learning based upsampling operation, giving rise to an end-to-end trainable system. To this end, we consider two different strategies widely adopted in CNN for upsampling, *i.e.*, unpooling and deconvolutions. As opposed to pooling layers, the unpooling operation with an upscaling factor s replaces each entry in the

input feature map with a $s \times s$ block, where the top left element in the block is set to the value of the input entry and the others to zero. The unpooling operation yields enlarged yet sparse output feature maps. The sparsely activated output values can then be propagated to local neighborhoods by subsequent convolution layers. The deconvolution layer upscales the input feature maps by s-fold through reversing the forward and backward propagation of convolution layers with an output stride of s. Although unpooling and deconvolution resort to different implementations, they are essentially similar in upscaling feature maps and both are well suited to our task. We adopt the deconvolution layer and achieve promising performance.

The upsampling module connects the feature extraction and reconstruction modules and plays a key role in the proposed SR method. Our experiments empirically show that properly increasing the kernel size of the deconvolution layer can enhance the upsampling quality, leading to an improvement of the final performance. This may be attributed to the fact that a larger deconvolution kernel size allows the upsampling operation to consider a larger input neighborhood and better enforces spatial consistency. However, when the kernel size is sufficiently large, the improvement becomes marginal, while the computational overhead is significantly increased. For efficiency, two 1×1 convolution layers are conducted before and after the expensive deconvolution layer to further reduce the computational complexity, where the first convolution layer performs dimension reduction by mapping the 64-channel input feature maps to the 4-channel output feature maps for upsampling, and the last convolution layer then restores the upsampled feature maps back to 64 channels. In such a way, the deconvolution operation is performed in a reduced dimension. A ReLU layer is added to the end of the upsampling module to increase non-linearity.

C. MULTI-SCALE RECONSTRUCTION

Since similar image patterns may recur across different scales in different images of both training and test sets, accurate inference of the input image should be highly invariant to image scale variations and may rely on the aggregation of multi-scale contextual information. This insight has been intensively studied and verified in vision related problems, like image object detection [39], scene recognition [40], *etc.* From the perspective of image SR, some prior methods [25], [26] have also confirmed that multi-scale context can effectively benefit HR image reconstruction.

Considering that HR image restoration may rely on both short- and long-range contextual information, we propose to perform HR reconstruction with multi-scale convolutions to explicitly encode multi-context information.

The input of our HR reconstruction module firstly go through R residual units. Then a dimension reduction layer is followed that consists of a 1×1 convolution, mapping the input feature map of 64 channels to the output 16 channels. The subsequent multi-scale convolution layer comprises 4 convolution operations of 1×1, 3×3, 5×5, and 7×7 kernel sizes, respectively. All four convolutions are simultaneously conducted on the input feature map and produce four feature maps of 16 channels. The feature maps are then concatenated into a single 64-channel feature map, such that features encoding contextual information in different scales are fused together. The concatenated feature map is then fed into another 1×1 convolution layer, which serves as a weighted combination of multi-context feature and reconstructs the final HR images.

D. COMBINING DEEP AND SHALLOW NETWORKS

As opposed to prior decoupled SR methods, which firstly upsample LR images using bicubic interpolation and then generate HR outputs with CNNs, the proposed method formulates image SR as an end-to-end trainable system by directly mapping original LR images to the HR ones with CNNs. The end-to-end trainable system takes full advantage of the strong learning power of deep networks, meanwhile makes deep network training even more challenging. Our experiments show that the proposed deep CNN takes much longer training time to converge than a relatively shallow one. During the training process, the intermediate output of the deep CNN often suffers from slight but visible illumination shifts. Similar phenomenon is not observed in training shallow networks or training deep networks in the decoupled manner.

Through a more in-depth analysis of our preliminary experiments, we further observe that although the deep CNN suffers from illumination shifts, the high-frequency image content is mostly restored. In comparison, the shallow CNN is able to accurately restore the overall illumination but fails to capture high-frequency details. To combine the best of both worlds, we jointly train an ensemble comprising the proposed deep CNN and another shallow CNN. The shallow CNN serves as an anchor to predict the major component of HR images and facilitate faster optimization, while the deep CNN restores high-frequency details and corrects errors of the shallow CNN.

The shallow network consists of three trainable layers corresponding to the three modules of the proposed deep network. The first layer takes the original LR image as input and conducts 3×3 convolutions, producing a feature map of 4 channels. The second layer is a deconvolution layer which upsamples the input feature map to the target spatial size. The final layer reconstructs the HR image from the upsampled feature maps by 5×5 convolutions.

The deep and shallow networks do not share weights. Both of them independently conduct image SR by taking the same original LR image as input and can be viewed as an ensemble of networks. One simple strategy to combine the two networks for the final results is through addition as follows,

$$\hat{Y} = H_D(X, \theta_D) + H_S(X, \theta_S), \tag{1}$$

where X denotes the input LR image; $H_D(\cdot, \theta_D)$ and $H_S(\cdot, \theta_S)$ indicate the HR outputs of deep and shallow networks

parameterized by θ_D and θ_S, respectively; $\hat{Y}$ is the final HR image predicted by the ensemble. Apart from the simple addition, we also investigate a more general form of combination method as follows,

$$\hat{Y} = p(X)H_D(X, \theta_D) + (1 - p(X))H_S(X, \theta_S), \quad (2)$$

where $p(X)$ denotes the combination weight to balance the deep and shallow networks. We predict the weight value for the input image using another CNN, which is jointly trained with the deep and shallow networks, such that the weights of deep and shallow networks are adaptively set according to the input image. Similar idea can also be found in the Highway networks [41] and the Long Short-Term Memory (LSTM) recurrent networks [42].

Both the above strategies work well in practice. For simplicity, we adopt the addition based method. Detailed analysis and comparison results of the two combination strategies are reported in Section IV-C.

E. TRAINING

Given N training image pairs $\{X_i, Y_i\}_{i=1}^N$, the proposed deep and shallow networks are jointly learned by minimizing the Euclidean loss between the predicted HR image $\hat{Y}$ and the ground truth Y:

$$\min \frac{1}{2N} \sum_{i=1}^{N} \|\hat{Y}_i - Y_i\|_2^2 + \eta R(\theta_D, \theta_S), \quad (3)$$

where $\hat{Y}$ is the predicted HR image computed as (1) or (2), and $R(\theta_D, \theta_S)$ denotes the weight decay imposed on network parameters with a small trade off η.

The optimization is conducted by the mini-batch stochastic gradient descent method with a batch size of 64, momentum of 0.9, and weight decay of $1e-4$. All the filters in convolution layers are randomly initialized from a zero-mean Gaussian distribution with standard deviation 0.01. The filters in deconvolution layers are initialized from bilinear interpolation kernels. The learning rate is initially set to $1e-4$ and decreased by a factor of 0.1 when the validation loss is stabilized.

IV. EXPERIMENTS

A. SETUP

For fair comparisons with existing methods, we use the same training sets, test sets and protocols which are widely-adopted [15], [18], [19]. We evaluate the performance of upscaling factors 2, 3 and 4 on three public datasets: Set5 [32], Set14 [5] and BSD100 [43], which contain 5, 14, and 100 images, respectively. We train three deep models with deconvolution kernel sizes 14×14, 15×15, and 16×16, respectively, for the upscaling factors 2, 3 and 4. Our models are trained using 91 images proposed in [4]. Following existing CNN based methods, data augmentation techniques including rotation and flipping are performed to reduce overfitting, yielding a training set of 728 images and a validation set of 200 images. For each upscaling factor (*i.e.*, 2, 3 or 4), 96×96 patches are randomly cropped from training images as the ground truth HR examples, which are then downsampled using the bicubic interpolation to generate the corresponding LR training samples. The proposed models are trained using the Caffe framework [44] on a workstation with a Intel 3.6 GHz CPU and a NVIDIA GTX TITANX GPU. The training takes approximately 95 epochs to converge for each model.

We utilize PSNR and SSIM [45] metrics for quantitative evaluation, which are widely used in the image SR literature. At inference, our model takes the original LR image of arbitrary size as input and directly reconstructs the corresponding HR image. Since humans are more sensitive to changes of luminance than color, we follow most existing methods and only super-resolve the luminance channel in YCbCr color space. For the purpose of displaying, the other two chrominance channels are simply upsampled by bicubic interpolation.

B. COMPARISON WITH STATE-OF-THE-ARTS

We compare the proposed EEDS with state-of-the-art methods using either the results or the publicly available codes provided by the authors. The compared methods include the traditional bicubic interpolation, 4 shallow model based methods (SUSR [5], A+ [11], ARFL [12], NBSRF [14]) and 8 deep CNN based methods (SRCNN [15], SRCNN-L [16], CSC [19], CSCN [18], ESPCN [22], VDSR [21], FSRCNN [35] and ESCN [46]). Among others, all 4 shallow models and 3 CNN-based methods, including SRCNN, CSC, CSCN and ESCN, are trained using 91 training images, while SRCNN-L, VDSR and FSRCNN are trained on additional images apart from the 91 images. ESPCN provides the results with both 91 training images and additional training images. Based on their training images, we divide all the methods into two groups, *i.e.*, without or with additional training images, and independently compare methods within each group. As our baseline, we set the number of residual units $R = 2$ in the multi-scale reconstruction module and train our EEDS using 91 training images is compared with methods in the first group (without additional training images). For fair comparison against methods in the second group, we keep all the other settings unchanged and make two modifications on the original EEDS model following VDSR [21]: a) augmenting the 91 training images with additional 200 images from the BSD200 data set [47] and b) setting $R = 6$ to increase the depth of the deep CNN. The improved model is named as EEDS+. It should be note that increasing R will further improve the performance, however, it is not our main contribution.

Tab. 1 summarizes the quantitative performance of compared methods measured by average PSNR and SSIM. In the first group using 91 training images, the proposed EEDS method consistently outperforms the other methods across three test sets for all upscaling factors. As demonstrated in the last line of the first group, our method improves the performance over the second best method (CSCN) by a considerable margin in terms of both PSNR and SSIM.

TABLE 1. Average PSNR(SSIM) comparison on three test datasets among different methods. Red and blue colors indicate the best and the second best performance.

Dataset	Set5			Set14			BSD100		
Scale	×2	×3	×4	×2	×3	×4	×2	×3	×4
Performance with 91 training images									
Bicubic	33.66 (0.9299)	30.39 (0.8682)	28.42 (0.8104)	30.24 (0.8687)	27.55 (0.7736)	26.00 (0.7019)	29.56 (0.8431)	27.21 (0.7385)	25.96 (0.6675)
SUSR [5]	35.78 (0.9493)	31.90 (0.8968)	29.69 (0.8428)	31.81 (0.8988)	28.67 (0.8075)	26.88 (0.7342)	30.40 (0.8682)	27.15 (0.7695)	25.92 (0.6968)
A+ [11]	36.55 (0.9544)	32.59 (0.9088)	30.29 (0.8603)	32.28 (0.9056)	29.13 (0.8188)	27.32 (0.7491)	30.78 (0.8773)	28.18 (0.7808)	26.77 (0.7085)
ARFL [12]	36.71 (0.9548)	32.57 (0.9077)	30.21 (0.8565)	32.36 (0.9059)	29.12 (0.8181)	27.31 (0.7465)	31.26 (0.8864)	28.28 (0.7825)	26.79 (0.7066)
NBSRF [14]	36.76 (0.9552)	32.75 (0.9104)	30.44 (0.8632)	32.45 (0.9071)	29.25 (0.8212)	27.41 (0.7511)	31.30 (0.8876)	28.36 (0.7856)	26.88 (0.7110)
SRCNN [15]	36.34 (0.9521)	32.39 (0.9033)	30.09 (0.8530)	32.18 (0.9039)	29.00 (0.8145)	27.20 (0.7413)	31.11 (0.8835)	28.20 (0.7794)	26.70 (0.7018)
CSC [19]	36.62 (0.9548)	32.66 (0.9098)	30.36 (0.8607)	32.31 (0.9070)	29.16 (0.8209)	27.30 (0.7499)	31.27 (0.8876)	28.31 (0.7853)	26.83 (0.7101)
ESPCN [22]	–	32.55 (–)	–	–	29.08 (–)	–	–	–	–
CSCN [18]	36.93 (0.9552)	33.10 (0.9144)	30.86 (0.8732)	32.56 (0.9074)	29.41 (0.8238)	27.64 (0.7578)	31.40 (0.8884)	28.50 (0.7885)	27.03 (0.7161)
ESCN [46]	37.14 (0.9571)	33.28 (0.9173)	31.02 (0.8774)	32.67 (0.9093)	29.51 (0.8264)	27.75 (0.7611)	31.54 (0.8909)	28.58 (0.7917)	27.13 (0.7197)
EEDS	37.29 (0.9579)	33.47 (0.9191)	31.14 (0.8783)	32.81 (0.9105)	29.60 (0.8284)	27.82 (0.7626)	31.64 (0.8928)	28.64 (0.7925)	27.11 (0.7200)
Performance with additional training images									
SRCNN-L [16]	36.66 (0.9542)	32.75 (0.9090)	30.49 (0.8628)	32.45 (0.9067)	29.30 (0.8215)	27.50 (0.7513)	31.36 (0.8879)	28.41 (0.7863)	26.90 (0.7103)
ESPCN (ImageNet) [22]	–	33.13 (–)	30.90 (–)	–	29.49 (–)	27.73 (–)	–	–	–
FSRCNN [35]	37.00 (0.9558)	33.16 (0.9140)	30.71 (0.8657)	32.63 (0.9088)	29.43 (0.8242)	27.59 (0.7535)	31.50 (0.8906)	28.52 (0.7893)	26.96 (0.7128)
VDSR [21]	37.53 (0.9587)	33.66 (0.9213)	31.35 (0.8838)	33.03 (0.9124)	29.77 (0.8314)	28.01 (0.7674)	31.90 (0.8960)	28.82 (0.7976)	27.29 (0.7251)
EEDS+	37.78 (0.9609)	33.81 (0.9252)	31.53 (0.8869)	33.21 (0.9151)	29.85 (0.8339)	28.13 (0.7698)	31.95 (0.8963)	28.88 (0.8054)	27.35 (0.7263)

It should be noted that the CSCN method adopts a cascaded strategy to conduct SR for the upscaling factors 3 and 4 (*i.e.*, by super-resolving the LR image twice with a factor of 2), which is shown to improve the final performance. Further performance improvements of our method can also be expected when using the cascaded strategy. In the second group with additional training images, the proposed EEDS+ improves the performance of SRCNN-L, ESPCN and FSRCNN with a considerable margin across all the data sets and upsampling factors. The performance gain may be attributed to the fact that EEDS+ adopts a much deeper network, allowing stronger learning capability. The VDSR is arguably one of the best performing SR methods with very deep networks. Our EEDS+ adopts a CNN with the same depth of VDSR and compares favorably against VDSR, yielding higher performance across all the compared data sets for three upscaling factors. The comparison results verify the effectiveness of the proposed method.

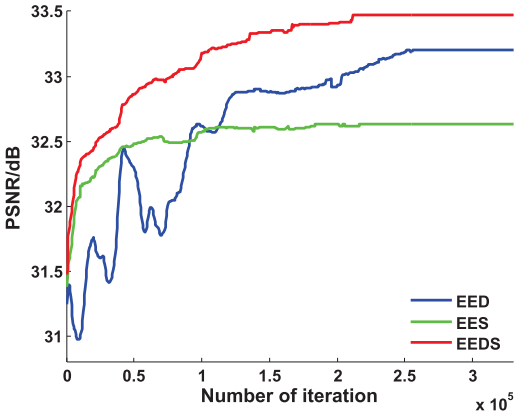

FIGURE 3. Convergence plots of different CNN architectures on the set5 data set with an upsampling factor 3.

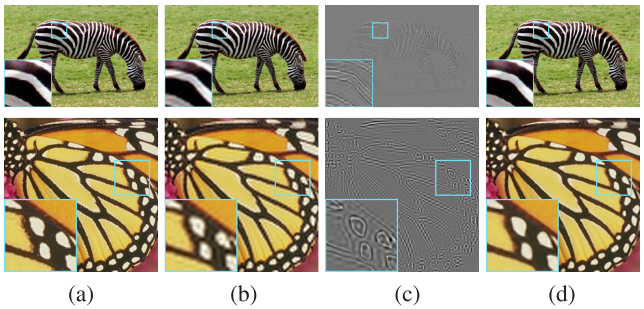

FIGURE 4. Output of the proposed EEDS model and its subnetworks with an upscaling factor 3. (a) Ground truth, (b) output of the shallow network of EEDS, (c) output of the deep network of EEDS, (d) final result of EEDS.

TABLE 2. Comparison of DCNN and DSCNN by average PSNR (dB) on three data sets with an upscaling factor 3.

Model	Set5	Set14	BSD100
DCNN	32.46	29.11	28.26
DSCNN	32.71	29.34	28.37

Fig. 5 and Fig. 6 illustrate some sampled results generated by the compared methods. The HR images restored by the proposed EEDS method are perceptually more plausible with relatively sharp edges and little artifacts.

C. ARCHITECTURE ANALYSIS

To gain further insights of our contributions, we conduct additional evaluations on different variants of the proposed EEDS method. Unless stated otherwise, we strictly follow the implementation settings in Section IV-A to train all the methods.

1) COMBINING DEEP AND SHALLOW NETWORKS THROUGH ADDITION

Our method jointly trains a deep and a shallow network as an ensemble. To investigate the impact of the two networks on the final performance, we split the two networks and obtain two variants of the proposed EEDS model, namely, EED (end-to-end deep network) and EES (end-to-end shallow network), respectively. Fig. 3 depicts the convergence plots of all three models on the Set5 data set. EES with a shallow network takes less time to converge. However, limited by its capacity, the final performance of EES is relatively low. In contrast, EED is more difficult to train. The training process is very unstable with oscillation in training loss. Upon convergence, EED achieves higher PSNR than EES, but is still unsatisfactory. This may be attributed to the fact that directly mapping LR images to HR ones is a very complex task and EED may converge to some local minimum.

The proposed EEDS method mitigates this issue by combining deep and shallow networks as an ensemble. At joint training, the shallow network still converges much faster and dominates the performance at the very beginning (Fig. 3). After the shallow network has already captured the major components of the HR images, the difficulty of direct SR has been significantly lowered. The deep network then starts to focus on the high-frequency details and learns to correct the errors made by the shallow network. As shown in Fig. 3, the EEDS method is much faster to converge than EED and achieves the best performance among all three methods. Upon convergence, the prediction made by the shallow network of EEDS restores most content with blur and artifacts (Fig. 4 (b)), whereas the deep network of EEDS learns to predict the residual between the HR image and the output of the shallow network, mostly containing high-frequency content (Fig. 4 (c)). The behavior of deep and shallow networks combined through simple addition is supported by and further confirms the key findings of deep residual networks [38], indicating that deep residual learning can be achieved through addition of subnetworks and makes deep networks more easier to optimize. Meanwhile, the addition of deep and shallow networks is also consistent to prior SR methods [5], [10], [11], where SR is conducted by learning the residual between HR image and the bicubic interpolated LR input. As opposed to these approaches, our EEDS method replaces the fixed bicubic interpolation with a shallow network and jointly trains the deep and shallow networks, making the residual prediction based method a special case of our method.

To study the impact of combining deep and shallow CNNs on other network architectures, we compare an eight-layer baseline deep CNN (denoted as DCNN) that has similar architecture to SRCNN [15] against the combination of the deep CNN and a 3-layer shallow CNN (denoted as DSCNN). As shown in Tab. 2, DSCNN consistently outperforms DCNN across all the data sets, suggesting that the benefits of combining deep and shallow networks can generalize to other network architectures.

2) ADAPTIVE COMBINATION OF DEEP AND SHALLOW NETWORKS

We also investigate the adaptive combination strategy defined in (2). To predict the combination weight, we jointly train a 4-layer convolution network with the shallow and deep

TABLE 3. Comparison of the combination of deep and shallow networks by using average PSNR (dB) evaluation.

Dataset	Set5			Set14			BSD100		
Scale	x2	x3	4	x2	x3	x4	x2	x3	x4
EES	36.93	33.06	30.56	32.55	29.38	27.50	31.23	28.47	26.89
EED	37.06	33.20	30.73	32.62	29.46	27.59	31.28	28.53	26.96
EED-BN	37.11	33.29	30.81	32.66	29.49	27.63	31.31	28.55	27.00
EEDS	37.29	33.47	31.14	32.81	29.60	27.82	31.64	28.64	27.11
EEDS-BN	37.33	33.50	31.17	32.83	29.63	27.84	31.66	28.65	27.13
EEDS-A	37.32	33.49	31.16	32.82	29.61	27.82	31.65	28.64	27.12

TABLE 4. Average PSNR (dB) of different upsampling strategies on set5 with an upscaling factor 3.

Model	EED-ND	EED	EEDS-ND	EEDS
Set5	33.01	33.20	33.22	33.47
Set14	29.35	29.46	29.47	29.60
BSD100	28.46	28.53	28.51	28.64

TABLE 5. Average PSNR (dB) of different kernel sizes for deconvolution layer on set5 with an upscaling factor 3.

Model	EEDS-D7	EEDS-D21	EEDS-D25	EEDS
Set5	33.45	33.47	33.48	33.47
Set14	29.57	29.61	29.61	29.60
BSD100	28.63	28.64	28.64	28.64

networks using the loss function defined in (3). The first three layers are convolution layers and has the same architecture as the corresponding layers of AlexNet [48] except that we use 1×1 stride and perform zero-padding for each layer to maintain the spatial size. The output feature map of the third layer is aggregated through the global average pooling [49] and then fed into the last fully connected layer followed by a sigmoid layer to generate a scalar weight value. We name the variant of our method using the adaptive combination strategy as EEDS-A. For comparison, we also explore the batch normalization technique [50] in the single deep network EED and the proposed EEDS. The corresponding variants are named as EED-BN and EEDS-BN, respectively. Tab. 3 shows the comparison results between EES, EED, EED-BN, EEDS, EEDS-BN and EEDS-A on all the compared data sets and upsampling factors. Though EED-BN achieves better results than EED, its overall performance is still unsatisfactory. In contrast, EEDS consistently outperforms EES, EED and EED-BN across all the evaluations with a considerable margin, which justifies the effectiveness of combining deep and shallow networks through addition. By using batch normalization, EEDS-BN can slightly improve the performance of EEDS, suggesting that the contributions of the proposed EEDS and the batch normalization to the final performance do not strongly overlap with each other. The performances of EEDS and EEDS-A are comparable, suggesting that both the addition-based and the adaptive combination strategy work well in our setting. In comparison, the simple addition based combination requires less network parameters and training time, thus is more suitable for our task.

3) UPSAMPLING ANALYSIS

To justify the effectiveness of learning based upsampling module over the bicubic-interpolated based approach, we compare EEDS with three variants: EED (End-to-End Deep network), EED-ND (EED with no deconvolution), and EEDS-ND (EEDS with no deconvolution). The EED-ND model is obtained by substituting the deconvolution layer of the deep network with a convolution layer producing the same number of channels. Similarly, the EEDS-ND model is obtained by replacing the deconvolution layers of both deep and shallow networks in EEDS with convolution layers. Correspondingly, both EED-ND and EEDS-ND take as input the LR images that have been upscaled to the desired sizes by bicubic interpolation.

Tab. 4 reports the average PSNR of the compared methods on three test sets with an upsampling factor 3. EEDS and EED considerably improves the performance of EEDS-ND and EED-ND, respectively, confirming that our learning based upsampling strategy in an appropriate feature space is more effective than directly upscaling the LR image by bicubic interpolation in the original color space.

Furthermore, since the key parameter in the upsampling module is the kernel size of the deconvolution layer, additional evaluations are also conducted to study the performance of different kernel sizes. While keeping the basic settings unchanged, we only modify the kernel size of deconvolution layer from the default value 15 to 7, 21 and 25, and denote their corresponding performance as EEDS-D7, EEDS-D21 and EEDS-D25, respectively.

Results in Tab. 5 show that the performance can be further improved by increasing of the kernel size, which suggests that the contextual information is beneficial for the task of SR. However, when the kernel size is sufficiently large (15×15 in this case), the performance becomes saturated. Considering that larger kernel sizes entail more computational overhead, we choose the size of 15 as a trade-off for both efficiency and effectiveness.

TABLE 6. Evaluation of the multi-scale layer for reconstruction by using average PSNR (dB) index.

Dataset	Set5			Set14			BSD100		
Scale	x2	x3	4	x2	x3	x4	x2	x3	x4
EEDS-SS1	37.01	33.14	30.91	32.63	29.46	27.65	31.51	28.52	26.68
EEDS-SS3	37.03	33.18	30.95	32.66	29.48	27.68	31.51	28.54	26.99
EEDS-SS5	37.09	33.23	30.97	32.67	29.48	27.68	31.53	28.55	27.01
EEDS-SS7	37.11	33.27	31.01	32.70	29.49	27.72	31.54	28.55	27.03
EEDS	37.29	33.47	31.14	32.81	29.60	27.82	31.64	28.64	27.11

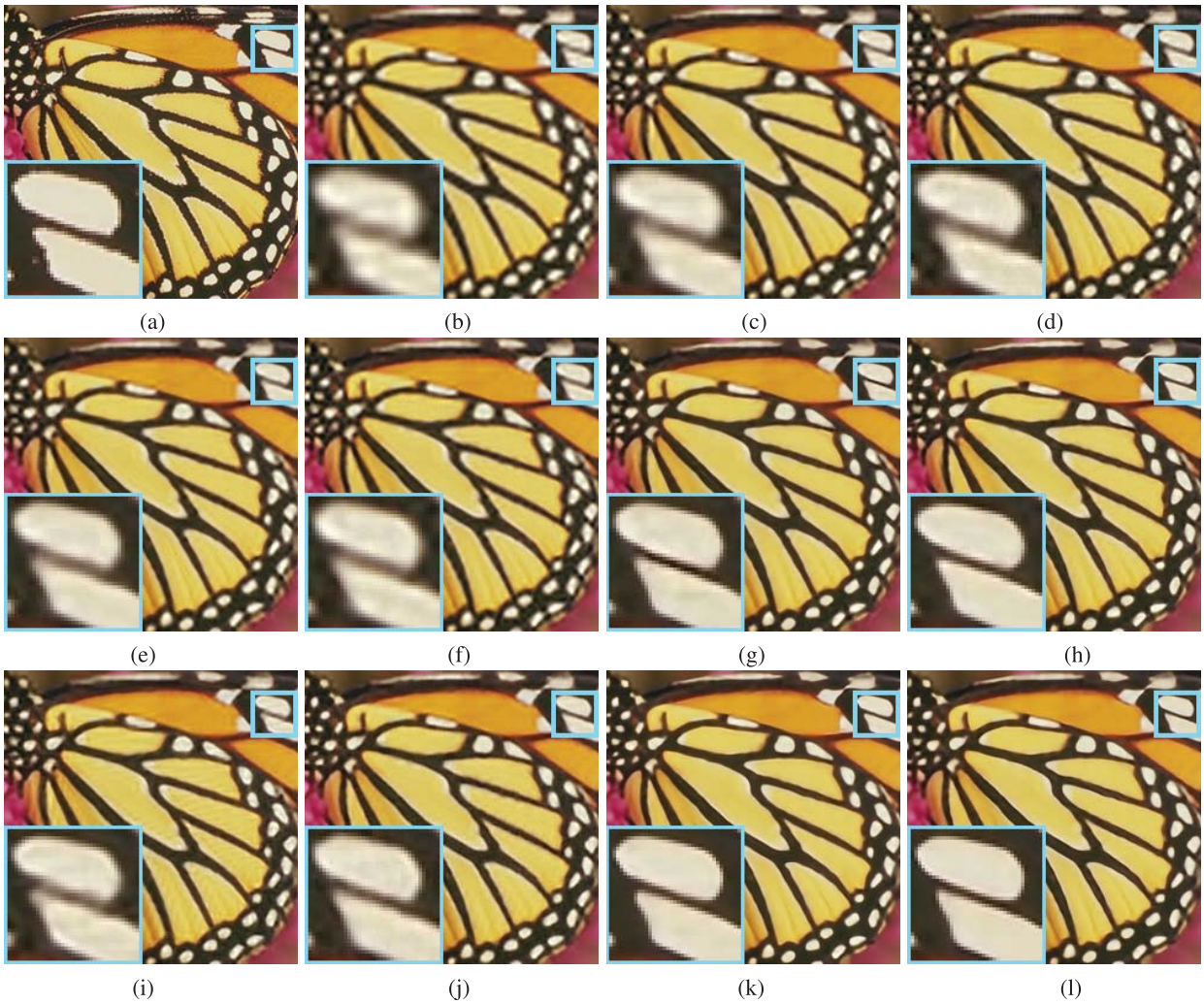

FIGURE 5. The "butterfly" image from set5 with an upscaling factor 4. (a) Ground truth / PSNR. (b) SUSR [5] / 23.59dB. (c) A+ [11] / 24.45dB. (d) ASRF [12] / 24.61dB. (e) NBSRF [14] / 25.03dB. (f) CSC [19] / 24.45dB. (g) CSCN [18] / 26.18dB. (h) EEDS / 26.55dB. (i) SRCNN-L [16] / 25.07dB. (j) FSRCNN [35] / 25.66dB. (k) VDSR [21] / 27.29dB. (l) EEDS+ / 27.41dB.

4) MULTI-SCALE ANALYSIS

In the reconstruction module of our EEDS model, the multi-scale convolution layer consists of four scales (kernel sizes): 1, 3, 5, 7. To verify the effect of the multi-scale strategy for the image SR task, we compare the proposed multi-scale EEDS model with variants using single-scale (denoted as SS) reconstruction modules. We set all the kernel sizes of the reconstruction module into the same sizes: 1×1, 3×3, 5×5, and 7×7, and obtain four variants denoted as EEDS-SS1, EEDS-SS3, EEDS-SS5 and EEDS-SS7, respectively.

The performance of each scale (EEDS-SS) and multi-scale (EEDS) are reported in Tab. 6, indicating that a large scale has slightly better performance than a small scale, due to the fact that large patches contain more contextual information than small ones. Moreover, when fusing the four scales together for reconstruction, EEDS considerably improves the average

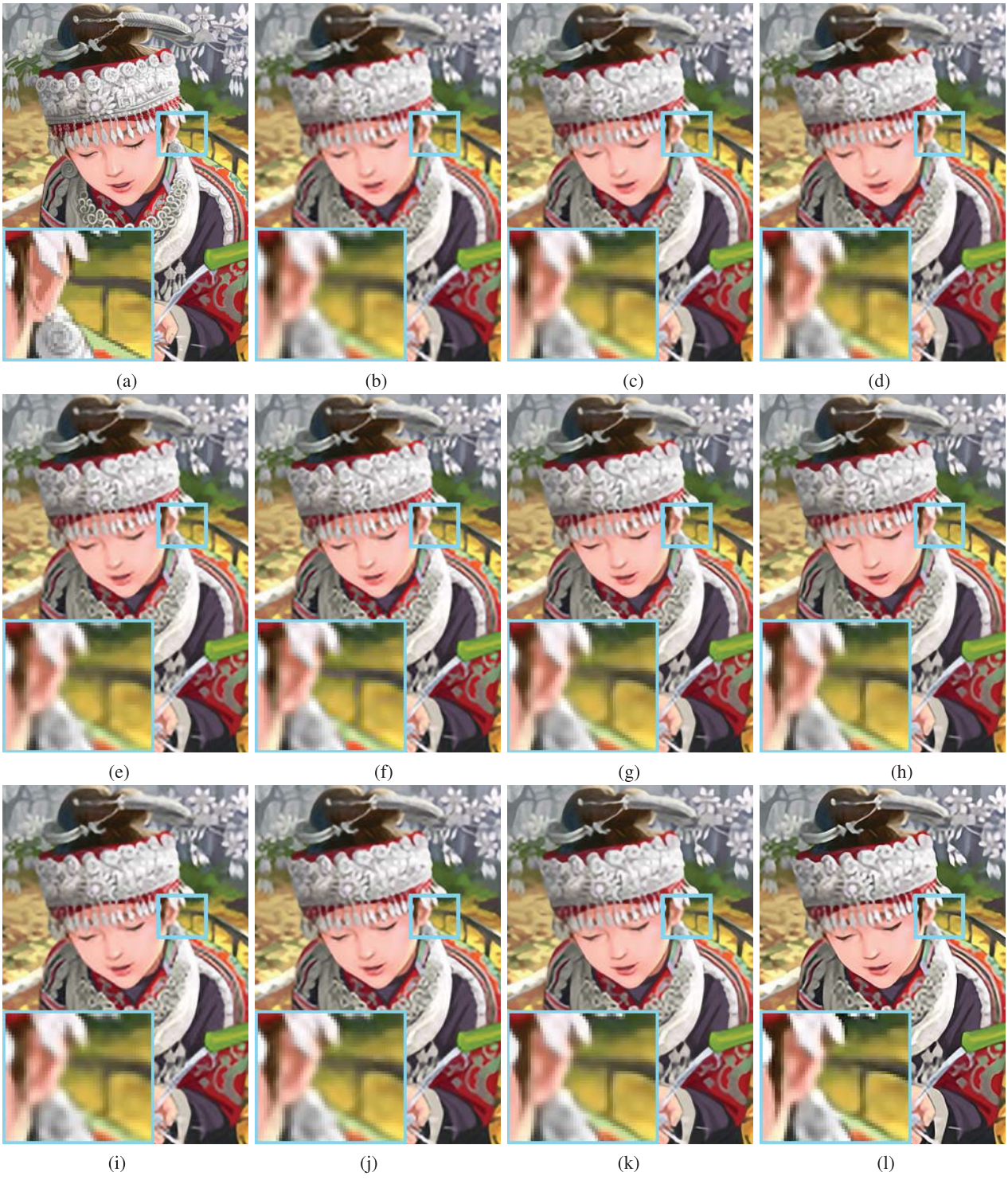

FIGURE 6. The "comic" image from Set14 with an upscaling factor 3. (a) Ground truth / PSNR. (b) SUSR [5] / 23.97dB. (c) A+ [11] / 24.40dB. (d) ASRF [12] / 24.40dB. (e) NBSRF [14] / 24.51dB. (f) CSC [19] / 24.43dB. (g) CSCN [18] / 24.70dB. (h) EEDS / 24.92dB. (i) SRCNN-L [16] / 24.56dB. (j) FSRCNN [35] / 24.72dB. (k) VDSR [21] / 25.12dB. (l) EEDS+ / 25.28dB.

PSNR of single scale variants across all the SR factors, which validates that combining both short- and long-range contextual information can significantly benefit the ill-posed detail recovery problem.

5) MORE ABLATION STUDIES ON DEEP NETWORKS

Tab. 4 demonstrates the impact of deconvolution on the deep network (*i.e.*, EED). To gain a comprehensive understanding of the proposed techniques, we further study the

TABLE 7. Average PSNR (dB) on three data sets with an upscaling factor 3.

Model	EED	EED-NSC	EED-NDR	EED-SS
Set5	33.20	33.16	33.22	33.08
Set14	29.46	29.44	29.45	29.37
BSD100	28.53	28.52	28.53	28.47

contributions of short-cut connection, multi-scale reconstruction, and dimension reduction on the single deep network. To this end, we compare EED with its variants EED-NSC, EED-NDR, EED-SS. Among others, EED-NSC and EED-NDR remove the short-cut connection and dimension reduction from EED, respectively. EED-SS replaces the multi-scale reconstruction of EED with a single convolution layer of 7×7 kernel size. The comparison results on all the data sets with an upsampling factor 3 is shown in Tab. 7. The performance gain yielded by multi-scale reconstruction is more significant than the other two techniques, suggesting that the benefit of multi-scale reconstruction is universal in the SR problem. With short-cut connection, EED performs slightly better and converges faster than EED-NSC, confirming the key findings in [38]. The performances of EED and EED-NDR are comparable. This makes sense since the architecture difference between EED and EED-NDR is minor and the goal of dimension reduction is only to reduce computational intensity at a minimum performance loss.

V. CONCLUSION

This paper proposes a fully end-to-end trainable system for single image SR using an ensemble of deep and shallow networks. The shallow network with a lightweight architecture is easy to optimize and learns to render the major structure of the HR image, while the deep network with a stronger learning capability is only responsible to capture the high frequency details. As such, jointly training the network ensemble can significantly lower the difficulty of network training and gives rise to more superior performance. To ensure more accurate restoration of HR images, the HR reconstruction is performed in a multi-scale manner to simultaneously incorporate both short- and long-range contextual information. Experiments confirm that the proposed method performs favorably against state-of-the-art approaches. In-depth ablation studies are also conducted to verify the contributions of different network designs to image SR, providing additional insights for future research.

REFERENCES

[1] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *Int. J. Comput. Vis.*, vol. 40, no. 1, pp. 25–47, 2000.

[2] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, Mar./Apr. 2002.

[3] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 2009, pp. 349–356.

[4] 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.

[5] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surf.* Berlin, Germany: Springer, Jun. 2010, pp. 711–730.

[6] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Trans. Graph.*, vol. 30, no. 2, p. 12, 2011.

[7] 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.

[8] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun./Jul. 2004, pp. 275–282.

[9] C.-Y. Yang and M.-H. Yang, "Fast direct super-resolution by simple functions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 561–568.

[10] R. Timofte, V. Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1920–1927.

[11] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 111–126.

[12] S. Schulter, C. Lesistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 184–199.

[13] J.-J. Huang, W.-C. Siu, and T.-R. Liu, "Fast image interpolation via random forests," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3232–3245, Oct. 2015.

[14] J. Salvador and E. Perez-Pellitero, "Naïve Bayes super-resulution forest," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 325–333.

[15] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.

[16] 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.

[17] C. Osendorfer, H. Soyer, and P. van der Smagt, "Image super-resolution with fast approximate convolutional sparse coding," in *Neural Information Processing*, 2014, pp. 250–257.

[18] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 370–378.

[19] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, and L. Zhang, "Convolutional sparse coding for image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1823–1831.

[20] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1637–1645.

[21] 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.

[22] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1874–1883.

[23] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4549–4557.

[24] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[25] 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.

[26] Z. Wang, Y. Yang, Z. Wang, S. Chang, J. Yang, and T. S. Huang, "Learning super-resolution jointly from external and internal examples," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4359–4371, Nov. 2015.

[27] R. G. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 6, pp. 1153–1160, Dec. 1981.

[28] C. E. Duchon, "Lanczos filtering in one and two dimensions," *J. Appl. Meteorol.*, vol. 18, no. 8, pp. 1016–1022, 1979.

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

[30] 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.
[31] 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.
[32] 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, pp. 1–10.
[33] Y. Zhao, G. Li, W. Xie, W. Jia, H. Min, and X. Liu, "GUN: Gradual upsampling network for single image super-resolution," *IEEE Access*, vol. 6, pp. 39363–39374, 2018.
[34] J. Bruna, P. Sprechmann, and Y. LeCun. (2015). "Super-resolution with deep convolutional sufficient statistics." [Online]. Available: https://arxiv.org/abs/1511.05666
[35] 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.
[36] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2790–2798.
[37] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 624–632.
[38] 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.
[39] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[40] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 2169–2178.
[41] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2377–2385.
[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[43] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Jul. 2001, pp. 416–423.
[44] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[46] L. Wang, Z. Huang, Y. Gong, and C. Pan, "Ensemble based deep networks for image super-resolution," *Pattern Recognit.*, vol. 68, pp. 191–198, Aug. 2017.
[47] D. R. Martin, C. C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jul. 2001, pp. 416–425.
[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Process. Syst.*, 2012, pp. 1097–1105.
[49] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2921–2929.
[50] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

YIFAN WANG received the B.S. and M.Sc. degrees in electronic engineering from the Dalian University of Technology, in 2013 and 2015, respectively, where she is currently pursuing the Ph.D. degree in signal and information processing. Her research interests include image super-resolution, image enhancement, and deep learning.

LIJUN WANG received the B.E. degree from the Dalian University of Technology, Dalian, China, in 2013, where he is currently pursuing the Ph.D. degree in signal and information processing. His current research interests include visual saliency, object tracking, image super-resolution, and deep learning.

HONGYU WANG received the M.Sc. degree in electronic engineering from the Graduate School, Chinese Academy of Sciences, Changchun, China, in 1993, and the Ph.D. degree in precision instrument and optoelectronics engineering from Tianjin University, Tianjin, China, in 1997. He was an Assistant Professor with the Department of Electronic Engineering, Zhejiang University, Zhejiang, China, from 1997 to 2004. He is currently a Professor with the School of Information and Communication Engineering, Dalian University of Technology. His research interests include mobile multimedia communications, and image and video processing. In recent years, he focuses on high-spectral image processing, image enhancement, video stability, and video surveillance.

PEIHUA LI received the Ph.D. degree from the Harbin Institute of Technology, in 2002. He was a recipient of the honorary nomination of National Excellent Doctoral Dissertation, in 2005. He was supported by the Program for New Century Excellent Talents in University of the Ministry of Education of China, in 2011. He is currently a Professor with the School of Information and Communication Engineering, Dalian University of Technology. He has published over 50 papers in referred conferences and journals. His current research interests include image classification and search using theoretical and computational methods of information geometry.