IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 13, 2020, accepted May 1, 2020, date of publication May 25, 2020, date of current version June 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2997028

Modified Dual Path Network With Transform Domain Data for Image Super-Resolution

DE-WEI CHEN AND CHIH-HUNG KUO, (Member, IEEE)
Department of Electrical Engineering, National Cheng Kung University, Tainan 70101, Taiwan

Corresponding author: De-Wei Chen (alcohol102501025@gmail.com.tw)

This work was supported in part by the Higher Education Support Project, Ministry of Education to the Headquarters of University Advancement, National Cheng Kung University, in part by the Ministry of Science and Technology of Taiwan under Grant MOST 107-2221-E-006-221, and in part by the National Cheng Kung University and Qualcomm Collaborating Research.

ABSTRACT Recently, studies on single image super-resolution using Deep Convolutional Neural Networks (DCNN) have been demonstrated to have made outstanding progress over conventional signal-processing based methods. However, existing architectures have grown wider and deeper, resulting in a large amount of computation and memory cost, but only a small improvement in performance. To address this issue, in this paper, we present a Wavelet- and Saak-transform Dual Path Network (WSDPN), which considers not only low-resolution images but also transform-domain information. The proposed network exploits the rich information extracted from the transform domain to reconstruct more accurate high-resolution images. In addition, to reap the benefits from both residual network (ResNet) and densely convolutional network (DenseNet) topologies, we use dual-path blocks as the basic building blocks which allow feature re-use while ensuring the ability to continue extracting new features. Thanks to extensive research on the attention mechanism, we further introduce spatial and self-attention blocks to refine features based on feature correlations at different layers. The experimental results show that our proposed approach achieves better performance on extensive benchmark evaluation than other state-of-the-art methods.

INDEX TERMS Super resolution, deep learning, convolutional neural network.

I. INTRODUCTION

Single Image Super Resolution (SISR) aims to leverage one Low-Resolution (LR) image to predict useful information to reconstruct a High-Resolution (HR) image while improving the image quality. SISR has been widely applied in many image processing fields, such as satellite image, surveillance, medical imaging, and remote sensing. Since a determined LR input can be obtained by adopting the same degradation process on many possible HR images, which is the reason why SISR has been an ill-posed problem despite decades of extensive research.

Interpolation-based methods, including nearest neighbor, bilinear, and bicubic interpolations, weight the adjacent pixels of an LR image to generate the HR image, which often accompanied by blurred artifacts. Reconstruction-based methods exploit complex prior knowledge to limit the desirable HR space to generate clear details. However, as the upscaling factor increases, a huge number of training samples are required to recover HR images at a visually satisfactory level which is very time-consuming. On the other hand, the learning-based or example-based methods use supervised learning to build a mathematical analysis between LR and corresponding HR patches from the sample data which is learned by either extracting internal similarities from the LR patch itself or the correspondence between external exemplar pairs. The neighbor embedding [1] method conducts manifold learning on multiple nearest neighbors in the training dataset to reconstruct HR patches. Sparse coding [2], [3] methods consider image patches as a sparse linear combination of elements from a compact dictionary. Nonetheless, due to over-reliance on the well-trained mappings and their associated weak representation capabilities, they are usually inefficient and thus show limited visual quality. Recently, Convolutional Neural Networks (CNNs) have been shown to exhibit superior performance as compared to prior models by their remarkable learning capabilities. The Super-Resolution Convolutional Neural Network (SRCNN) [4], known as the

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

first CNN model in super-resolution tasks, learn end-to-end mappings from LR to HR images through a fully convolutional network. It outperforms the classical non-deep learning method. However, most deep-learning SR methods are based on spatial domain input to reconstruct the output of the network.

In this work, as an alternative, we investigate the advantages of the data from the transform domain. More specifically, we attempt to capture image information in both the spatial and spectral domains to enhance SR quality. In addition, motivated by the promising performance of the residual network (ResNet) [5] and the densely convolutional network (DenseNet) [6] in classification tasks, we propose taking advantage of both networks and combining them with wavelet and Saak transforms [7]. Our network is trained with nine input channels, which comprise four sub-bands of the low-resolution wavelet coefficients, four sub-bands of the low-resolution Saak coefficients, and one LR image. The experiments show that using the transformed signals only increases the parameters by a small amount, but it greatly improves the quality of the reconstructed image. To further improve performance, we propose the Dual-Path Block (DPB), which inherits the advantages of the residual network (ResNet) and densely convolutional network (DenseNet), to facilitate the feature reuse within the network and to meanwhile obtain more compact and accurate representations that lead to more realistic visual effects. We also apply self and spatial attention blocks to consider the correlations between features at different levels and to recalibrate the feature maps with context information.

The rest of this paper is organized as follows: The related background is reviewed in Section II. Section III details the proposed network. Model comparisons and experimental results are presented in Section IV. Finally, conclusions are offered in Section V.

II. RELATED WORK

A. DEEP LEARNING BASED IMAGE SUPER-RESOLUTION

The strong feature extraction and data representation abilities in deep learning have led to a surge of research on convolutional neural networks for SISR. As the seminal SR method based on the convolutional structure, Dong *et al.* [4] proposed the SRCNN that learns the nonlinear mapping between LR and HR patches. However, the high computational cost still hampers it from being applied in real-time applications since the network takes upsampled images as input. To further improve accuracy, speed and memory efficiency, FSRCNN [8] and ESPCN [9] conduct feature extraction directly from LR images and adopt a post-upscaling scheme which employs a deconvolution layer or a sub-pixel convolution module at the tail of the network to upsample the spatial size. By so doing, as compared to SRCNN, they significantly reduce computations while still failing to construct a deeper model due to the difficulties of training. Also, the quality of the reconstructed images may be degraded if there are no sufficient layers in the network. In order to relieve the training burden of deep networks, Kim *et al.* [10] proposed VDSR, which introduces global residual learning so that the network only needs to learn the residuals between HR and LR patches, and the residuals are then added to the original images to recover the SR results.

To avoid the checkboard artifacts produced by the deconvolutional operation, we choose the sub-pixel convolution module, which is regarded as a common convolution in LR space followed by a periodic shuffling, for the last stage of the network to upscale the spatial size. Furthermore, we apply the residual learning mechanism to help the convergence of training deep networks.

B. DEEP LEARNING IN TRANSFORM DOMAIN

A discrete wavelet transform analyzes an image by decomposing it into sub-bands that can capture textural and contextual information in both the frequency and location domains. A super-resolution algorithm with the wavelet transform is implemented to estimate the missing coefficient, where the LR image is considered to be the low-frequency subband of the HR image. The difficulty lies in predicting the unknown coefficients of the lost high-frequency subbands. DASR [11] combines both interpolated LR images and high-frequency subband images acquired by the discrete wavelet transform to fulfill reconstruction with high-quality in the spatial domain. Nguyen and Milanfar [12] established an interlaced sampling structure in training data for the purpose of efficiently calculating the wavelet coefficients. In addition to the wavelet-based constraint, Jiji *et al.* [13] used a smooth prior to determine the appropriate wavelet interpolation. Sparse coding was integrated to design different interpolation methods in [14]–[16]. Kinebuchi *et al.* [17] exploited hidden Markov trees to interpolate wavelet coefficients. DWSR [18] combined CNN and a wavelet transform, which benefits from the sparsity of wavelet residuals, to recover missing details and achieves competitive performance. However, due to the lack of training and the use of simple interpolation approaches, the above methods failed to prove their superiority over recent deep learning based models.

Kuo [19], [20] proposed the RECOS (REctified-COrrelations on a Sphere) transform to explain CNN in a mathematical model. This is a multi-layer transform whose forward process maps three-dimensional data into one-dimensional rectified spectral vectors. In order to reduce the defects in the inverse process, Kuo *et al.* proposed a Subspace approximation with an augmented kernel (Saak) transform [7], which adopted the Karhunen-Loève (KL) basis as the basic kernel. By using its negative vector to augment the transform kernel and performing the sign-to-position format conversion which is equivalent to the ReLU activation, the Saak transform can solve the sign confusion problem when multi-level transforms are cascaded. There is no need to train the transform kernels through back propagation, and it is possible in the meantime to minimize the transform losses. The Saak coefficients represent the spectral

components in the corresponding spatial area and thus offer a joint spatial-spectral representation. In addition, the Saak transform can apply Principal Component Analysis (PCA) technique to reach energy compaction. The small perturbation in the test data would not affect the leading coefficients, and thus makes the Saak transform a robust process. Saak transform is a data-driven approach, so it can easily adapt to any task as a feature extraction technique or an unsupervised dimension reduction procedure.

C. DUAL PATH NETWORK

Taking advantage of cutting-edge neural network architectures to design a novel one is the most intuitive and effective way to enhance the model learning ability in a variety of tasks. Recent studies show that scaling up networks [21] has been widely adopted to improve performance of neural networks. However, deep neural networks will encounter degradation problems in which the network accuracy begins to saturate. To solve this issue, recent works have focused on helping the flow of information and gradients in the network to avoid problems such as vanishing gradients and the curse of dimensionality by modifying the network structure.

He *et al.* [5] proposed a residual network that introduced identity mapping and shortcut connections to ease optimization issues. In addition, due to the sparsity of input and output signals, the networks are more robust and can be trained more easily to further construct deep neural networks with hundreds of layers. Recently, densely convolutional networks [6] were proposed, where skip connections are introduced to concatenate from the input of convolutional layers to the output to facilitate training by strengthening feature propagation and encouraging feature reuse. Nevertheless, the method used to fuse features is not an adding process but rather is a concatenation, which results in the width of the densely connected path to increase linearly as the depth rises. This hampers building a deeper and wider network due to a large number of parameters and a large amount of GPU memory cost. Chen *et al.* [22] proposed the Dual Path Network (DPN), which is a compound network design intended to follow the core idea of both residual and densely connected networks. The DPN took the residual network as its backbone and attached a thin, densely connected path to construct the dual path network. However, in order to avoid high redundancy and relieve the computational burden, DPN used the grouped convolution to reduce the number of parameters caused by the densely connected paths, which resulted in additional hyperparameters. Also, in the DPN, shortcut connections both in the residual path and densely connected path are only applied between different blocks, making it difficult to share information and improve gradient flow across layers. In addition, in the two path topologies, skip connections do not take the importance of different features into account, where they are simply fused through feature adding and concatenation. Different from the DPN, we use the gating mechanism to limit the growth of the number of the feature maps and simultaneously merge the information of the two paths where the importance of the features in both paths will be adaptively considered.

III. PROPOSED NETWORK

In this section, we introduce the proposed model in detail, including the transformed inputs, the design of the dual-path block, and then the overall network architecture.

A. TRANSFORM DOMAIN DATA

Image super-resolution technology can be divided into two categories: frequency domain methods and spatial domain methods. In this work, we propose the use of transform domain signals to enhance SR quality. Here, the data are first transformed to the frequency domain and they are then combined together to capture both spatial and spectral information. After processing by CNNs, the signals are inverse transformed into the spatial domain to reconstruct a super-resolved image.

Over the past few years, the common Fourier transform was gradually replaced with the wavelet transform in image and signal processing. In the study of Fourier theory, intricated but periodic signals are represented as the sum, theoretically infinite, of sine and cosine waves. Though it can decompose the analyzed signal into the frequency information, it does not provide any time or location details that may benefit SR applications. To address this issue, wavelet transform applies different versions of basis function to analyze a signal in the time domain which offers both the frequency and location information. [23] Besides, wavelets allow rapid and efficient transform algorithms that need to be considered when the training of deep networks becomes a burden. Similarily, the Saak coefficients collect the spectral component in the corresponding spatial area and thus can provide a joint spatial-spectral representation. As stated above, we chose wavelet and Saak algorithms [7] to transform the input image to offer additional information.

1) WAVELET-TRANSFORM INPUT

In order to obtain the wavelet subbands, LR training images $\boldsymbol{X}$ are upsampled using bicubic interpolation. Then we generate four LR wavelet sub-bands by conducting a Haar wavelet on bicubic interpolated images $\boldsymbol{X}_{bic}$, which can be denoted as:

$$\{LL, LH, HL, HH\} = 2dDWT\{\boldsymbol{X}_{\text{bic}}\}$$

where the LL, LH, HL, and HH are four subbands of the bicubic-interpolated image, respectively. Note that 2dDWT$\{\cdot\}$ denotes the 2D discrete wavelet transform.

2) SAAK-TRANSFORM INPUT

As in the above procedure, LR training images $\boldsymbol{X}$ are upscaled first using a bicubic interpolation. Then, we reshape enlarged LR images $\boldsymbol{X}_{bic}$ into a one-dimensional vector $\boldsymbol{f}$ by scanning the grid points in a fixed order, after which we can calculate the correlation matrix of $\boldsymbol{f}$ and take the eigenvectors of the correlation matrix as the Karhunen-Loève (KL) basis $\boldsymbol{b}_k$, for $k = 1, \cdots, K$.

In summary, the anchor vectors can be denoted as

$$A = \{a_0, a_1, \cdots, a_k, \cdots, a_K\},$$

where $K = L_k - 1$ and L_k denotes the number of the spectral dimensions. We then separate the anchor vectors into two types, DC and AC vectors. The DC anchor vector is

$$a_0 = \frac{1}{\sqrt{N}}(1, 1, \cdots, 1)^T.$$

The AC anchor vectors are the remaining anchor vectors $a_1 \cdots a_K$. A basic way to obtain AC anchor vectors is to train a convolutional neural network by backpropagation. Instead, the Saak transform takes the KLT's kernel vectors as the AC anchor vectors and first augments the k-th KLT kernel vector as

$$a_{2k-1} = b_k, \quad a_{2k} = -b_k.$$

Then, it projects the input vector $\boldsymbol{f}$ onto the set of augmented kernels to obtain $p_k = a_k^T \boldsymbol{f}$. Finally, the projection $\boldsymbol{p}$ is reshaped back into a 2D Saak feature map.

B. DUAL PATH BLOCK

As shown in Fig. 1, there are two paths, a residual path and a densely connected path, in our dual-path block. We modify the basic structure of ResNet [5] by passing the information for the preceding layer to each layer in the residual path. In order to make the network to adaptively consider the importance of features at different levels while avoiding the instability that may occur when training deep networks, we set learnable weights to adjust the path of each skip connection. Let B_{i-1} and B_i be the input and output of the i-th DPB, respectively. The residual path output $R_{o,i}$, which has two convolution layers, can be formulated as

$$R_{o,i} = W_{i,2}^{res} F_{i,1}^{res} + \sigma(W_{i,1}^{res} B_{i-1}) + B_{i-1},$$

where $F_{i,1}^{res} = \sigma(W_{i,1}^{res} B_{i-1})$, $W_{i,c}^{res}$ is the weights of the c-th convolution layer, and σ denotes the ReLU activation function. Note that the bias term is omitted for simplicity.

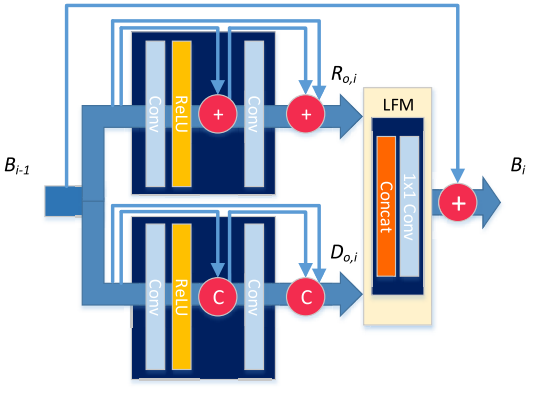

FIGURE 1. The architecture of the dual-path block. Node "+" denotes the element-wise addition, and "c" denotes the concatenation operation.

Similar to the residual path, the densely connected path connects all layers within the blocks in a feed-forward design. Dense connections can improve information and gradients flow throughout the network. Besides, concatenating feature maps attained by other layers provides more information in the input of following layers and improves performance. The dense connected path output $D_{o,i}$, which consists of two convolution layers, can be formulated as

$$D_{o,i} = [W_{i,2}^{dense} F_{i,1}^{dense}, \sigma(W_{i,1}^{dense} B_{i-1}), B_{i-1}],$$

where $F_{i,1}^{dense} = \sigma(W_{i,1}^{dense} B_{i-1})$, $W_{i,c}^{dense}$ is the weights of the c-th convolution layer, σ denotes the ReLU activation function, and $[\cdots, \cdots]$ refers to the concatenation operation. The Local Fusion Module (LFM) is then applied to adaptively fuse the features from both paths. Since we want to fuse the information extracted from both paths, and the feature maps of the densely connected path are directly preserved in a concatenative manner, we first concatenate the feature maps of the two paths and then adopt a 1×1 convolutional layer to fuse the information and to adaptively control the width increment of the dual-path block as well as the memory cost. We can formulate the operation of Local Fusion Module (LFM) as

$$F_{i,LFM} = LFM_i([R_{o,i}, D_{o,i}]),$$

where $LFM_i(\cdot)$ denotes the 1×1 convolutional layer. To further enhance the information flow, improve the network representation ability and increase performance, local residual learning is applied. Thus, the final output can be reached by

$$B_i = B_{i-1} + F_{i,LFM}.$$

C. FEATURE FUSION MODULE (FFM)

We constructed the Feature Fusion Module (FFM) in our feature mapping sub-network to make full use of the information obtained from each DPB and preserve persistent memory. As shown in Fig. 2, the FFM is composed of a dual path blockchain and a Mid-range Fusion Module (MFM). A series of continuous DPBs are stacked into a chain structure to form the dual-path blockchain for the purpose of performing further feature extraction at multiple levels. The Mid-range Fusion Module (MFM) is attached at the end of each FFM to merge the information from the preceding FFM and from the current blockchain to keep information. Similar to the Local Fusion Module (LFM), MFM first concatenates the features obtained by the previous FFM and by the current blockchain and then passes through a convolutional layer that serves as a gating mechanism to screen out the output information.

Inspired by [24] and [25], we also adopted the spatial attention block attached at each DPB to learn the correlations between hierarchical features as shown in Fig. 3. The spatial attention first divides the input data into three parts by three 1×1 convolution layers and then performs the dot-product operation in pairs to compute the similarity between two feature maps at different levels. Note that there is a shortcut

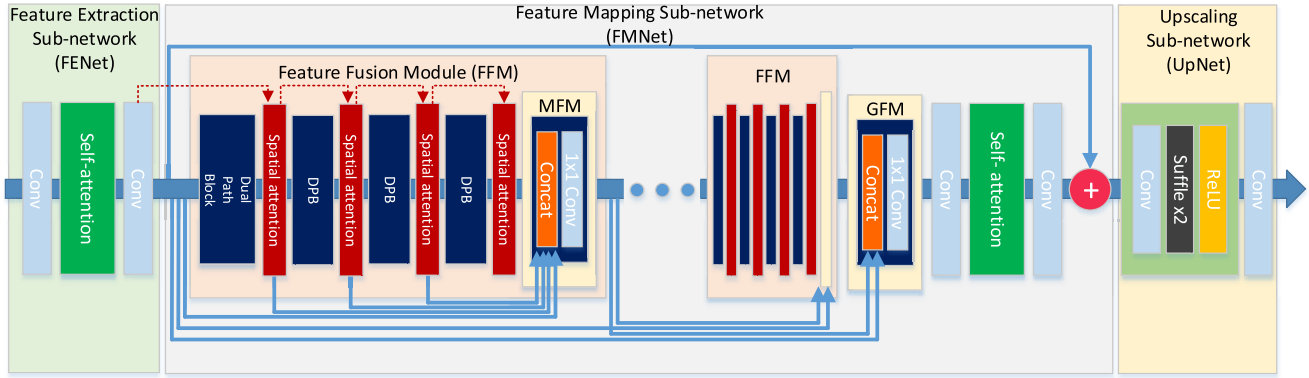

FIGURE 2. The framework of the proposed network.

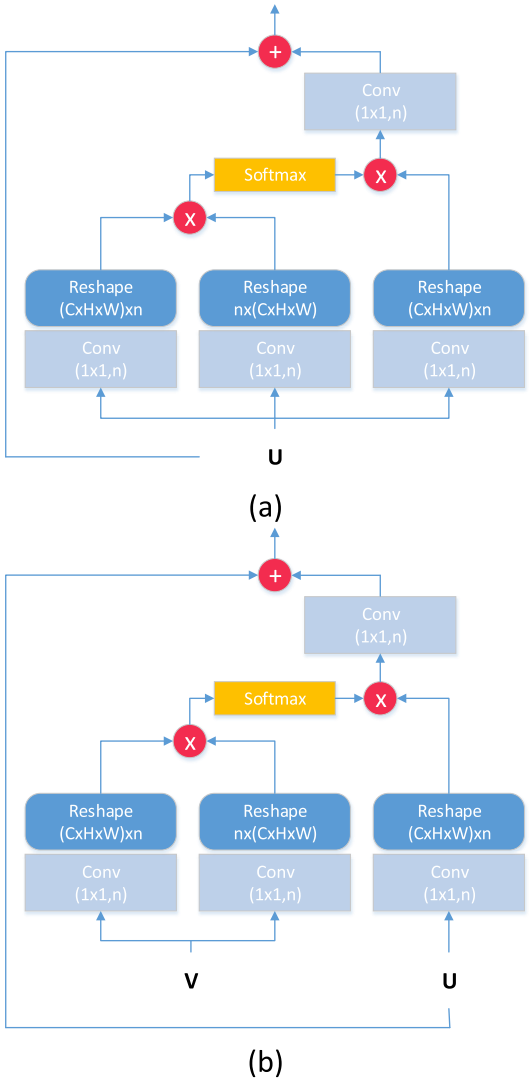

FIGURE 3. The modules of (a) self-attention and (b) spatial attention. *U* and *V* denote the current and previous output feature maps, respectively. *n*, *C*, *H*, and *W* refer to the filter number, the channel number, the height and width of the images, respectively.

connection between input and output. Therefore, the attention model only needs to learn the residual mapping to fine-tune the feature maps. Since the spatial attention takes the outputs of both the previous and the current DPB as input, it can comprehensively take the contextual information into consideration.

In summary, in the dual path blockchain, the stacked DPBs expand the receptive field of the network to extract deep feature representations, and the use of spatial attention considers hierarchical features to obtain more precise information. In MFM, multiple skip connections facilitate feature reuse and improve information flow across blocks.

D. NETWORK ARCHITECTURE

The proposed network for SISR, which is demonstrated in Fig. 2, contains a feature extraction sub-network (FENet), a feature mapping sub-network (FMNet), and an upscaling sub-network (UpNet). The FENet extracts the feature maps from the transform domain data. The FMNet is then applied to learn finer features by multiple stacked feature fusion modules. The learned features are used to generate the final SR result in the UpNet. Specifically, in FENet, we adopt a convolutional layer to extract the initial features from the concatenation of LR input images and the transformed inputs. The self-attention block proposed in [25] is located at the end of FENet to recalibrate the features. Note that the self-attention takes only the current input for computation, as shown in Fig. 3. For the FMNet, we stack multiple FFMs, which are designed to refine the features yielded from the feature extraction sub-network. The Global Fusion Module (GFM) is utilized to integrate the global features and avoid long-term information loss by fusing hierarchical features from all the FFMs. After integrating the highly informative features, we adopt another self-attention block to further adjust features for subsequent global residual learning. Finally, we utilize the sub-pixel convolution layer [9] to upsample the spatial resolution for the purpose of reconstructing HR images in UpNet.

E. LOSS FUNCTION

Given N training sample pairs $\{\boldsymbol{X}^n, \boldsymbol{Y}^n\}_{n=1}^N$ from the dataset, the proposed network is optimized to minimize the L_1 loss

function

$$\mathcal{L}(\Theta) = \frac{1}{N}\sum_{n=1}^{N} ||\boldsymbol{Y}^n - D_\Theta(\boldsymbol{X}^n)||_1,$$

where Θ and $D_\Theta(\cdot)$ denotes the parameter set and the output, respectively, of the network.

IV. EXPERIMENTS AND DISCUSSION

In this section, we first present the training data setting and provide implementation details, including the model hyper-parameters. Then, we analyze the influence of different composing units in the proposed model using ablation studies. Finally, comparisons with other state-of-the-art methods on several publicly available benchmark datasets are made to prove the superiority of our proposed network.

A. DATA AND SIMILARITY MEASURES

For the evaluation, the proposed method was compared on four standard benchmark datasets, Set5 [26], Set14 [27], BSD100 [28], and Urban100 [29]. The Set5, Set14, and BSD100 consist of human images and natural scenes, while the Urban100 contains the urban view. For training, we chose the DIV2K dataset [30], which consists of 800 high-quality (2K resolution) images for image restoration tasks. We performed random horizontal flipping and 90 degree rotation to augment the training data. We used the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) [31] index as the measurement metrics. For a fair comparison, we only take the luminance channel in YCbCr space into consideration to calculate PSNR (dB) and SSIM index and all images were center-cropped and a 4-pixel wide stripe was removed from each border, which is a common practice in SISR.

B. IMPLEMENTATION DETAILS

Based on existing studies [32], [33], we used the MATLAB [34] bicubic kernel to downsample HR images into LR images. For training, we randomly cropped the 32×32 LR patches from the LR images as the inputs and set the mini-batch size to 16. For optimization, we used an Adam optimizer [35] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate was initialized to 10^{-4}, which was decayed by a factor of 2 at every 2×10^5 iterations. For model training and testing, we used PyTorch [36] on an NVIDIA GTX 1080Ti GPU, and it took about four days to train our proposed WSDPN model. In the proposed network, we set 64 filters for all convolutional layers and the kernel sizes were all 3×3 with the exception of the 1×1 convolutional layers. In order to improve the learning capabilities of the network and control the computational costs, we adopted 24 DPBs in the FMNet. Meanwhile, we conducted zero-padding at the boundaries of each feature-map to keep the spatial size after the convolutional operation.

C. MODEL DISCUSSION

In this subsection, we discuss the influence of the different components making up our model through ablation experiments.

Considering different combinations of the types of input data, we examined several settings for the proposed network in Table 1. For quick validation, we used the original residual block [37] as the building block and removed the feature fusion module. Among the different combinations of input data types, it could be observed that residual blocks with both wavelet and Saak coefficients as input outperformed those with only LR images based on the PSNR gains of 0.05dB. Besides, using multiple inputs only increased the number of parameters by 0.3%. This demonstrates that leveraging transform domain input, which yields more information, indeed benefits super-resolution.

TABLE 1. Ablation study on the effects of different input combinations in residual blocks. Average PSNRs for a scale factor ×2 on the Set14 dataset are reported.

	Combinations of the types of input data			
LR	✓	✓	✓	✓
Wavelet coeff.	×	✓	×	✓
Saak coeff.	×	×	✓	✓
PSNR(dB)	33.58	33.6	33.6	33.63
Parameter	-	+0.16%	+0.16%	+0.3%

TABLE 2. Ablation study on the effects of different path combinations in dual-path blocks. Average PSNRs for a scale factor ×2 on the Set14 dataset are reported.

	Combinations of path topologies			
(a) Original residual path	✓	✓	×	✓
(b) Densely convolutional path	✓	×	✓	✓
(c) Modified residual path	×	✓	✓	✓
PSNR(dB)	33.73	33.73	33.79	33.77

We also explored the effects of different path topologies on the dual-path block. Fig. 4 shows the three path topologies used for the comparison: (a) the original residual path in SRResNet [37] and EDSR [38], which removes all the

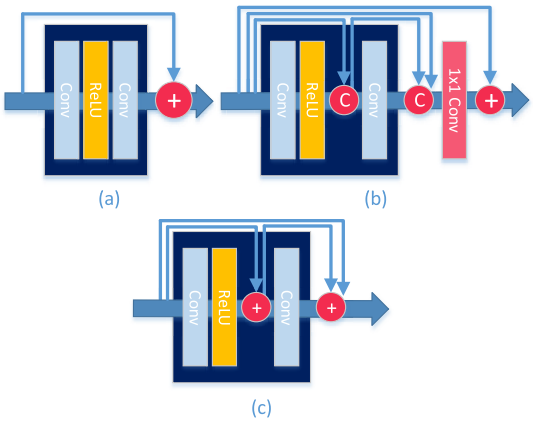

FIGURE 4. Three path topologies for dual-path blocks. Node "+" denotes the element-wise addition, and "c" denotes the concatenation operation.

TABLE 3. Quantitative evaluations of state-of-the-art SR methods. The best and the second best results are marked in red and blue, respectively.

Scale	Method	Set5		Set14		BSD100		Urban100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
2x	Bicubic	33.66	0.9299	30.24	0.8688	29.56	0.8431	26.88	0.8403
	SRCNN	36.66	0.9542	32.42	0.9063	31.36	0.8879	29.50	0.8946
	FSRCNN	37.05	0.9560	32.66	0.9090	31.53	0.8920	29.88	0.9020
	VDSR	37.53	0.9587	33.03	0.9124	31.90	0.8960	30.76	0.9140
	LapSRN	37.52	0.9590	33.08	0.9130	31.80	0.8950	30.41	0.9100
	DRCN	37.63	0.9588	33.04	0.9118	31.85	0.8942	30.75	0.9133
	DRRN	37.74	0.9591	33.23	0.9136	32.05	0.8973	31.23	0.9188
	D-DBPN	38.09	0.9600	33.85	0.9190	32.27	0.9000	33.02	0.9310
	EDSR	38.11	0.9601	33.92	0.9195	32.32	0.9013	32.93	0.9351
	RDN	38.24	0.9614	34.01	0.9212	32.34	0.9017	32.89	0.9353
	WSDPN	38.16	0.9606	33.97	0.9201	32.34	0.9016	32.99	0.9356
	WSDPN+	38.24	0.9610	34.07	0.9209	32.40	0.9021	33.18	0.9370
3x	Bicubic	30.39	0.8682	27.55	0.7742	27.21	0.7385	24.46	0.7349
	SRCNN	32.75	0.9090	29.28	0.8209	28.41	0.7863	26.24	0.7989
	FSRCNN	33.18	0.914	29.37	0.8240	28.53	0.7910	26.43	0.8080
	VDSR	33.66	0.9213	29.77	0.8314	28.82	0.7976	27.14	0.8279
	LapSRN	33.82	0.9220	29.87	0.8320	28.82	0.7980	27.07	0.8280
	DRCN	33.82	0.9226	29.76	0.8311	28.80	0.7963	27.15	0.8276
	DRRN	34.03	0.9244	29.96	0.8349	28.95	0.8004	27.53	0.8378
	D-DBPN	-	-	-	-	-	-	-	-
	EDSR	34.65	0.9282	30.52	0.8462	29.25	0.8093	28.80	0.8653
	RDN	34.71	0.9296	30.57	0.8468	29.26	0.8093	28.80	0.8653
	WSDPN	34.69	0.9288	30.57	0.8467	29.27	0.8097	28.86	0.8662
	WSDPN+	34.78	0.9294	30.68	0.8484	29.34	0.8109	29.07	0.8688
4x	Bicubic	28.42	0.8104	26.00	0.7027	25.96	0.6675	23.14	0.6577
	SRCNN	30.48	0.8628	27.49	0.7503	26.90	0.7101	24.52	0.7221
	FSRCNN	30.72	0.8660	27.61	0.7550	26.98	0.7150	24.62	0.7280
	VDSR	31.35	0.8838	28.01	0.7674	27.29	0.7251	25.18	0.7524
	LapSRN	31.54	0.8850	28.19	0.7720	27.32	0.7270	25.21	0.7560
	DRCN	31.53	0.8854	28.02	0.7670	27.23	0.7233	25.14	0.7510
	DRRN	31.68	0.8888	28.21	0.7720	27.38	0.7284	25.44	0.7638
	D-DBPN	32.47	0.8980	28.82	0.7860	27.72	0.740	26.08	0.7950
	EDSR	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033
	RDN	32.47	0.8990	28.81	0.7871	27.72	0.7419	26.61	0.8028
	WSDPN	32.51	0.8975	28.82	0.7879	27.73	0.7424	26.69	0.8044
	WSDPN+	32.66	0.8989	28.95	0.7903	27.81	0.7440	26.91	0.8088

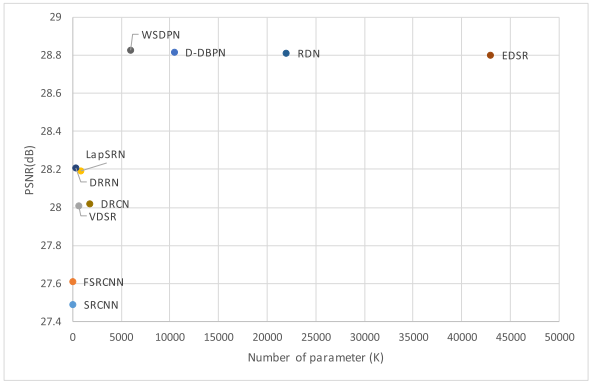

FIGURE 5. PSNR performance v.s. the number of parameters. The results are evaluated on the Set14 dataset for a scale factor of 4×.

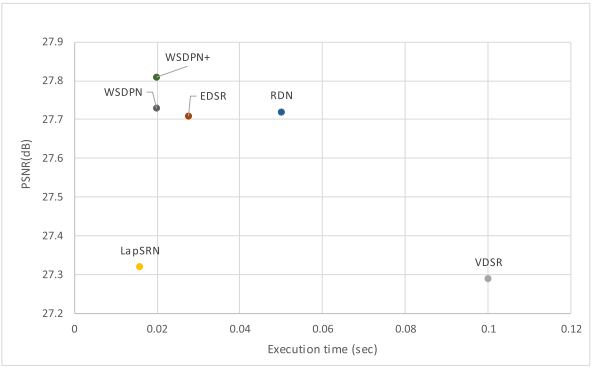

FIGURE 6. Running time and accuracy trade-off. The results are evaluated on the BSD100 dataset for a scale factor of 4×.

batch normalization layers in ResNet [5] for reducing memory consumption, (b) the densely convolutional path, which concatenates the feature maps of each layer to every other layer within the blocks, and (c) the modified residual path that adds the information of the preceding layer to each layer of the residual path. Table. 2 provides all possible combinations of the DPB topology. For simple, quick validation, we removed the attention modules and used only 16 DPBs as the baseline model. It was observed that applying the densely convolutional path and the modified residual path in the DPBs led to the best performance. This was because it can reserve the shallow features and continue discovering the finer ones. If we take three path topologies at the same time, the extracted features will interfere with each other, and the PSNR performance will degrade slightly.

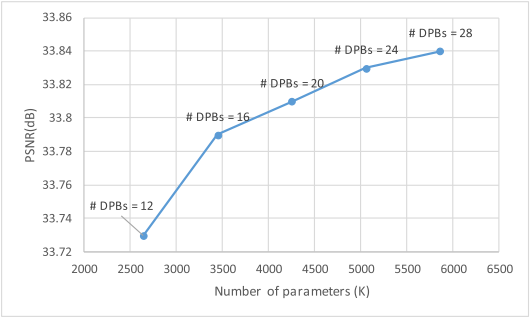

FIGURE 7. The relationship between the number of DPBs and the performance of the proposed model. Average PSNRs for a scale factor ×2 on the Set14 dataset are reported.

To show the tradeoff between performance and model size from the proposed network and existing SR networks, we made a comparison shown in Fig. 5. The results are evaluated on the Set14 dataset for a scaling factor of 4×. It can be observed that our model outperforms most state-of-the-art methods. It should be noted that our model shows higher PSNR values but with fewer parameters than EDSR and RDN. This evidence indicates that our network has a better trade-off between performance and the number of parameters. Fig. 6 demonstrates the trade-offs between the reconstruction accuracy and the execution time. In terms of running time, WSDPN runs faster than other SR methods and also obtains better PSNR results. It is obvious that our model strikes a good balance between the reconstruction accuracy and the running time.

To study the relationship between the number of DPBs in FENet and the reconstruction performance of the proposed model, we provide different number of DPBs and the corresponding PSNR results in Fig. 7. To save training time and perform simple and quick validation, we removed the attention mechanism. It can be noticed that though the depth increases as we use more DPBs, causing the number of parameters to grow linearly, the increase in PSNR tends to saturate. Thus, we chose 24 DPBs as our final model for subsequent comparisons. The conclusion can be drawn that designing a delicate architecture will be more helpful for reconstruction accuracy than blindly increasing the depth of the network.

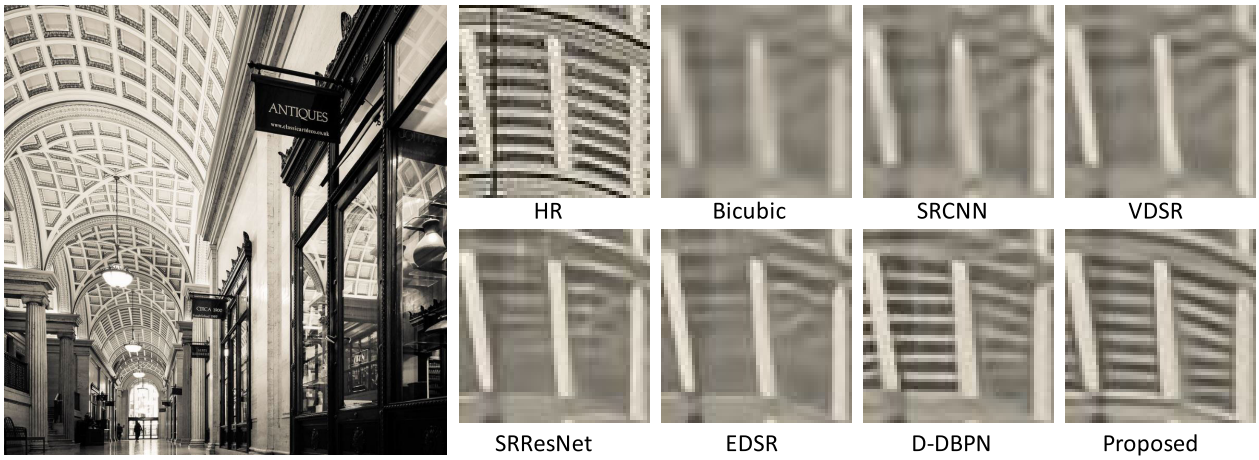

FIGURE 8. The reconstruction results for "img_083" from Urban100 with a scale factor of 4.

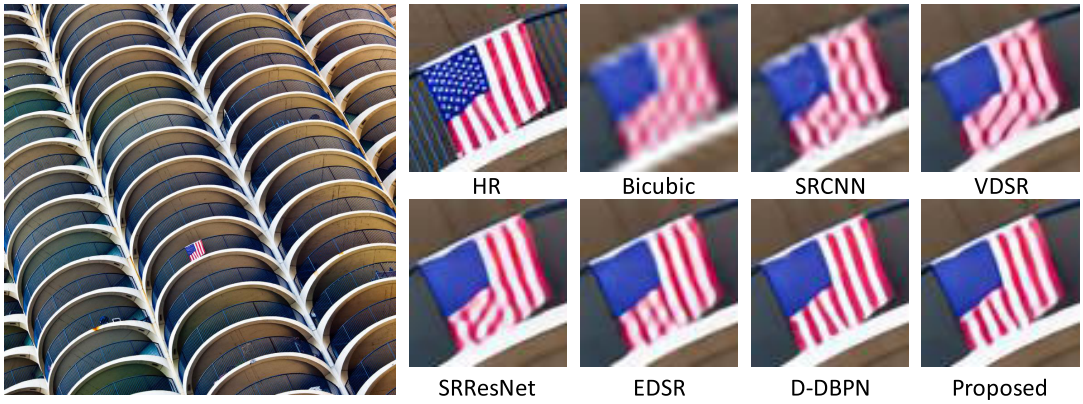

FIGURE 9. The reconstruction results for "img_100" from Urban100 with a scale factor of 4.

FIGURE 10. The reconstruction results for "img_820" from DIV2K with a scale factor of 4.

FIGURE 11. The reconstruction results for "img_842" from DIV2K with a scale factor of 4.

FIGURE 12. The reconstruction results for "img_900" from DIV2K with a scale factor of 4.

D. COMPARISON WITH STATE-OF-THE-ART METHODS

To show the effectiveness of our model, the average PSNR and SSIM results of several state-of-the-art SISR methods, including SRCNN [4], FSRCNN [8], VDSR [10], LapSRN [33], DRCN [39], DRRN [40], SRResNet [37], D-DBPN [41], EDSR [38], and RDN [42], are reported in the

form of a quantitative evaluation. In this work, the geometric self-ensemble [38] technique is also performed to obtain higher performance. Specifically, the test images are flipped and rotated to augment seven images from the original. We input these images into the network and perform an inverse transform on the output high resolution images. All of these images are then added together and averaged to obtain the final high-resolution output.This self-ensemble strategy has an edge over other ensemble methods in that it can be easily applied to various models without further training. Although the self-ensemble method does not require additional parameters, it can be noticed that this technique indeed increases the PSNR metric at approximately 0.1dB. The model with the self-ensemble method is denoted by adding "+" postfix to the model name. The quantitative evaluations for scales $\times 2$, $\times 3$ and $\times 4$ in the benchmark datasets are listed in Table 3. It can be seen that, compared to other methods, our model yields the best performance. In addition, the visual results of various methods for a scale factor $\times 4$ are presented in Fig. 8$\sim$ 12. It can be observed that the results of prior methods often include some distortions and artifacts, such as the stripes or the fur on animals, the word contour, and the lines of the buildings. By contrast, our method prevents such distortions, avoids the artifacts, and produces more realistic results. The proposed model sufficiently recovers the HR images with fine textures and thus demonstrates its superiority.

V. CONCLUSION

In this paper, we propose image super-resolution algorithms benefiting from the Saak and wavelet transforms. Our proposed multiple transform domain inputs extract rich information from the original LR image and thus can make our network learn the finer mapping between LR and HR pairs. Thanks to the robustness and efficiency of the residual network and the densely convolutional network, we apply the dual-path blocks as the basic architecture by which to construct our network. To further improve performance, we connect each layer within the dual-path block to increase information and gradient flows. In addition, we adopt self and spatial attention mechanisms, which aim to progressively recalibrate the learned feature maps, to improve the representational ability of the network. Compared with most state-of-the-art methods, WSDPN can achieve competitive or even better results under the premise that the number of parameters is economical.

REFERENCES

[1] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR*, vol. 1, Jul. 2004, p. 1.

[2] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[3] 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.

[4] 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, 2015.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[6] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[7] C.-C. Jay Kuo and Y. Chen, "On data-driven saak transform," *J. Vis. Commun. Image Represent.*, vol. 50, pp. 237–246, Jan. 2018.

[8] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 391–407.

[9] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[10] 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. (CVPR)*, Jun. 2016, pp. 1646–1654.

[11] G. Anbarjafari and H. Demirel, "Image super resolution based on interpolation of wavelet domain high frequency subbands and the spatial domain input image," *ETRI J.*, vol. 32, no. 3, pp. 390–394, Jun. 2010.

[12] N. Nguyen and P. Milanfar, "An efficient wavelet-based algorithm for image superresolution," in *Proc. Int. Conf. Image Process.*, vol. 2, Sep. 2000, pp. 351–354.

[13] C. V. Jiji, M. V. Joshi, and S. Chaudhuri, "Single-frame image super-resolution using learned wavelet coefficients," *Int. J. Imag. Syst. Technol.*, vol. 14, no. 3, pp. 105–112, 2004.

[14] S. Mallat and G. Yu, "Super-resolution with sparse mixing estimators," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2889–2900, Nov. 2010.

[15] M. F. Tappen, B. C. Russell, and W. T. Freeman, "Exploiting the sparse derivative prior for super-resolution and image demosaicing," in *Proc. IEEE Workshop Stat. Comput. Theories Vis.*, Jan. 2003, pp. 1–28.

[16] 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.

[17] K. Kinebuchi, D. D. Muresan, and T. W. Parks, "Image interpolation using wavelet based hidden Markov trees," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2001, pp. 1957–1960.

[18] T. Guo, H. S. Mousavi, T. H. Vu, and V. Monga, "Deep wavelet prediction for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 104–113.

[19] C.-C.-J. Kuo, "Understanding convolutional neural networks with a mathematical model," *J. Vis. Commun. Image Represent.*, vol. 41, pp. 406–413, Nov. 2016.

[20] C.-C.-J. Kuo, "The CNN as a guided multilayer RECOS transform [Lecture Notes]," *IEEE Signal Process. Mag.*, vol. 34, no. 3, pp. 81–89, May 2017.

[21] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*. [Online]. Available: https://arxiv.org/abs/1905.11946

[22] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4467–4475.

[23] B. Toufik and N. Mokhtar, "The wavelet transform for image processing applications," *Adv. Wavelet Theory Their Appl. Eng., Phys. Technol.*, vol. 17, pp. 395–422, Apr. 2012.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[25] Z.-S. Liu, L.-W. Wang, C.-T. Li, W.-C. Siu, and Y.-L. Chan, "Image super-resolution via attention based back projection networks," 2019, *rXiv:1910.04476*. [Online]. Available: https://arxiv.org/abs/1910.04476

[26] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. BMVA*, 2012, pp. 135.1–135.10.

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

[28] 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.
[29] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.
[30] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 451–466.
[31] 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.
[32] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Proc. Asian Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2014, pp. 111–126.
[33] 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. (CVPR)*, Jul. 2017, pp. 624–632.
[34] *MATLAB, 9.5.0.944444 (R2018b)*. Natick, MA, USA: The Math-Works Inc., 2018.
[35] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.
[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-highperformance-deep-learning-library.pdf
[37] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[38] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.
[39] 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. (CVPR)*, Jun. 2016, pp. 1637–1645.
[40] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3147–3155.
[41] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.
[42] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

DE-WEI CHEN was born in Changhua City, Taiwan, in 1994. He received the B.S. degree in electrical engineering from National Central University, Taoyuan City, Taiwan, and the M.S. degree in electrical engineering from National Cheng Kung University, Tainan, Taiwan. His research interests include deep learning and image processing.

CHIH-HUNG KUO (Member, IEEE) received the B.S. and M.S. degrees from National Tsing Hua University, Hsinchu, Taiwan, in 1992 and 1994, respectively, and the Ph.D. degree from the University of Southern California (USC), Los Angeles, CA, USA, in 2003, all in electrical engineering.

He was with the Computer and Communications Research Laboratories/Industrial Technology Research Institute (CCL/ITRI), Taiwan, as a DSP Design Engineer, from 1996 to 1998. From March 2004, he was a Senior Engineer with Winbond Electronics Corporation, Taiwan. In August 2004, he joined the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, as an Assistant Professor. He has been an Associate Professor, since February 2010. His current research interests include system-level designs for video processing and multimedia communications.