LCSCNet: Linear Compressing Based Skip-Connecting Network for Image Super-Resolution

Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao

Abstract—In this paper, we develop a concise but efficient network architecture called linear compressing based skip-connecting network (LCSCNet) for image super-resolution. Compared with two representative network architectures with skip connections, ResNet and DenseNet, a linear compressing layer is designed in LCSCNet for skip connection, which connects former feature maps and distinguishes them from newly-explored feature maps. In this way, the proposed LCSCNet enjoys the merits of the distinguish feature treatment of DenseNet and the parameter-economic form of ResNet. Moreover, to better exploit hierarchical information from both low and high levels of various receptive fields in deep models, inspired by gate units in LSTM, we also propose an adaptive element-wise fusion strategy with multi-supervised training. Experimental results in comparison with state-of-the-art algorithms validate the effectiveness of LCSCNet.

Index Terms—Single-image super-resolution, deep convolutional neural networks, skip connection, feature fusion.

I. INTRODUCTION

SINGLE image super-resolution (SISR) is a classical but challenging ill-posed inverse problem in low-level computer vision, aiming at restoring a high-resolution (HR) image from a single low-resolution (LR) input image. It is widely used in various areas such as medical imaging, satellite imaging and security imaging [1], [2].

Early methods for SISR are mainly interpolation-based, including Bicubic interpolation [3] and Lanczos resampling [4]. Then more powerful reconstruction-based methods often adopt sophisticated prior knowledge to restrict the possible solution space, with the advantage of generating flexible and sharp details [5]–[8]. Learning-based methods are now mainstream algorithms for SISR, utilizing substantial data to learn statistical relationships between LR and HR pairs. Markov Random Field (MRF) [9] was firstly adopted by Freeman et al. to exploit the abundant real-world images to synthesize visually pleasing image textures. Neighbor embedding methods [10] proposed by Chang et al. took advantage of similar local geometry between LR and HR to restore HR image patches. Inspired by the sparse signal recovery theory, researchers applied sparse coding methods [11]–[15] to SR. Random forest [16] has also been used to improve the reconstruction performance.

Recently, remarkable performance has been achieved for SR by deep models, especially deep network architectures, which are elaborated for high-level tasks in computer vision. Notably, residual network (ResNet) and densely connected network (DenseNet) are two widely-used architectures, which use skip connections to alleviate gradient problems and degradation phenomena in training. Chen et al. [17] analyzed ResNet and DenseNet in the HORNN framework [18] and concluded that ResNet enables feature re-usage while DenseNet enables feature exploration, both important to learn powerful representations.

Through extensive experiments, [19] and [20] implied that ResNet shows an ensemble-like behavior within its structure. Yang et al. [21] showed that ResNet applied in SR would lead to output with a layer-by-layer progressive effect, and Huang et al. [22] argued that this might restrict ResNet from reaching more feasible solutions. Although DenseNet explores as many new features as possible by directly utilizing any former features, its excessive skip connections among intermediate layers increase the number of parameters and burden the hardware during training.

In this paper, we propose Linear Compressing Based Skip-Connecting Network (LCSCNet), as a framework for SR, which takes advantages of ResNet’s parameter-economic feature re-use and DenseNet’s distinguishing feature exploration, as well as mitigating difficulties of restricted structures of ResNet and parameter burden of DenseNet.

As the network depth grows, the features produced by different intermediate layers would be hierarchical with different receptive fields. Among deep SR models, DRCN [23] and MemNet [24] used these intermediate features with multi-supervised methods, in which each feature corresponded to a raw SR output, and then fused these intermediate SR outputs by a list of trained scalars. Such a fusion strategy has two flaws: 1) once the weight scalars are determined in training, it will not change with different inputs; 2) using a single scalar to weight SR output fails to take pixel-wise differences into consideration, i.e., it would be better to weight different parts distinguishingly in an adaptive way. To overcome these shortcomings, inspired by the gate units in LSTM [25], we develop an adaptive element-wise fusion strategy in a progressive constructive way.
to maintain the element-wise convex weighted pattern, aiming at making better use of hierarchical information with different receptive fields.

In the end, we composite the Basic LCSCNet architecture with the adaptive element-wise fusion strategy gracefully for SR. Analysis and experiments in the following sections will illustrate the rationality of the proposed methods.

The main contributions of this work are three-fold:

1) We propose an accurate and efficient Linear Compressing Based Skip-Connecting Network (LCSCNet) architecture, which inherits the advantage of DenseNet in treating features of different levels distinguishingly while reducing its parameter size by exploiting the parameter-economic strength of ResNet. Moreover, we develop an Enhanced LCSCNet (E-LCSCNet) to further alleviate difficulties of training large-scale networks.

2) Differently from the traditional stationary fusion strategy, we take the input differences as well as the element-wise variation into consideration and propose an adaptive element-wise fusion strategy to further utilize hierarchical information.

3) When compared with the state-of-the-art models trained on the widely-used 291 dataset and those light networks trained on the DIV2K dataset, our proposed framework achieves the state-of-the-art performance. When compared with large models trained on DIV2K, our E-LCSCNet is among the state-of-the-art with apparent parametric efficiency.

The rest of the paper is organized as follows. Section II reviews recent related work. Section III presents a detailed description of the proposed architecture, mainly on the configuration of Basic LCSCNet and the adaptive element-wise fusion algorithm. Section IV illustrates several intriguing properties of LCSCNet, which could explain the rationality of LCSCNet. Section V conducts ablation studies to further probe into the proposed framework. Section VI presents experimental results in comparison with other relevant methods. Section VII concludes the paper and envisages some future work.

II. RELATED WORK

Because our proposed methods include the Basic LCSCNet architecture and the adaptive element-wise fusion strategy, in this section we review related work mainly from the aspects of basic SISR reconstruction and sub-output fusion.

A. Basic SISR Reconstruction

Dong et al. pioneeredly proposed a three-layer super-resolution convolutional neural networks (SRCNN) [27], predicting the end-to-end nonlinear mapping between LR and HR spaces. This first trial significantly outperformed other algorithms at that time. To combine the benefits of the natural sparsity of images and deep neural network architectures, Wang et al. proposed the Cascaded Sparse Coding Network (CSCN) [28], which had a higher visual quality than previous work. After SRCNN, Dong et al. further proposed FSRCNN [29] improving SRCNN mainly by leveraging deconvolution layers, which reduced computation significantly by increasing the resolution only at the end of network. In the meantime, the Efficient Sub-Pixel Convolutional Neural Network (ESPCN) [26] was proposed by Shi et al., replacing the traditional deconvolution layer by an efficient sub-pixel convolution layer and further reducing computation.

Inspired by the success that very deep neural networks with sophisticated architectures and training strategies achieved in some high-level tasks in computer vision [30], Kim et al. employed the VGG architecture and high learning rate with gradient clipping to stack a very deep (20 layers) convolutional neural network (VDSR) [31] and gained a remarkable improvement. Mao et al. proposed a deep fully convolutional auto-encoder network with symmetric skip connections [32]. To handle the issue of large numbers of parameters brought by very deep architectures, Kim et al. proposed the Deeply-Recursive Convolutional Network (DRCN) [23], which was also 20-layer but with 16 recursions among its intermediate layers. To further exploit the advantages from deepening neural networks, motivated by the success of [33], Tai et al. proposed the Deep Recursive Residual Network (DRRN) [34], a 54-layer convolutional neural network for SR, in which they utilized the residual network architecture (ResNet) [35] in both global and local manners. Inspired by the Dense Connected Network (DenseNet) [22] proposed by Huang et al., Tong et al. introduced dense skip connections to their deep architecture [36]. Based on the correlations among the HR outputs with different scale factors and a heuristic methodology, Lai et al. proposed the Laplacian Pyramid Super-Resolution Network (LapSRN) [37] to progressively reconstruct the sub-band residuals of higher-resolution images, which was especially effective for large scale factors. Motivated by explicitly mining persistent memory through an adaptive learning process and further mitigating the difficulties of training deeper networks, Tai et al. proposed an 80-layer network for image restoration, named as Persistent Memory Network (MemNet) [24].

Very recently, to further explore the power of example-based SISR with abundant training data, a new dataset DIV2K [38] consisting of 800 2K resolution images was established. Based on this powerful dataset, many new architectures were proposed for performance improvement. Among them, by removing Batch-Normalization (BN) [39] and applying residual scaling, Lim et al. proposed the Enhanced Deep Residual Network (EDSR) [40], which significantly improved performance. Then the Deep Back-Projection Network (DBPN) [41] was proposed by Haris et al. to combine the merits of deep neural networks with the back-projection procedure, proven to be very effective for large scale factors. By making full use of local and global information from deep architectures, the Residual Dense Network (RDN) [42] proposed by Zhang et al. exhibits comparable performance to EDSR, with fewer parameters.

B. Sub-output Fusion

Features from different depths with different receptive fields specialize in different patterns in SISR. From the perspective of ensemble learning, a better result can be acquired by adaptively fusing the outputs from different-level features. Based on this concept, several fusion strategies were proposed. Among them, two representative weighted-summation methods were the vectorized weighted fusion strategy [23], [24] and MSCN [43]. In the vectorized weighted fusion, a trainable positive vector
Figure 1: The overall architectures of (a) Basic LCSCNet (E-LCSCNet) and (b) LCSCNet (E-LCSCNet). In (b), ⊗ means element-wise multiplication; \( Y_1, \ldots, Y_N \) are the intermediate HR outputs reconstructed from \( F_1, \ldots, F_N \). When E-LSCSNet is employed, red lined parts are activated. For fair comparison, the upsampling and reconstruction part of (Basic) LCSCNet varies with the training dataset: For models trained on the 291 dataset, this part is the traditional deconv layer consisting of “nearst-neighborhood upsampling + conv-ReLU + conv-ReLU + conv”; for models trained on the DIV2K dataset, we use ESPCN [26] instead. To be specific, we only use ESPCN as U&RNet in Section V-E and the E-LCSCNet in Table VIII.

whose \( \ell_1 \) norm is 1 is applied, and each element in this vector controls how much of the current sub-output contributes to the final one. To regularize each sub-output and stabilize training, multi-supervised training is adopted. In MSCN, an extra CNN module takes LR as input and outputs several tensors with the same shape as the HR. These tensors can be viewed as adaptive element-wise weights for raw HR outputs. Then the weight module and the basic SISR module are trained jointly by optimizing the fused results in an end-to-end manner.

Both of the two fusing strategies above have shortcomings. The vectorized approach does not take the diversity of input and pixel-wise differences into consideration, while in MSCN the summation of coefficients at each pixel is not normalized, which is incongruous. Therefore, in this paper we aim to propose a normalized adaptive element-wise fusion strategy to overcome the shortcomings of the two previous fusion methods.

The above-mentioned deep methods mainly minimized the mean squared error (MSE), which tended to be blurry, over-smoothing and perceptually unsatisfying, especially in the case of large scale factors. Recently, some inspiring deep learning-based works concentrated on the exploration of more effective loss functions for SR. In [44] and [45], the perceptual loss using high-level feature maps of VGG made HR outputs more visually pleasing; [46] introduced amortized MAP inference to the loss function to get more plausible results; [47] and [48] used the adversarial loss to produce photo-realistic HR outputs. Although these methods produced high-quality images with rich texture details, the details in their outputs may be quite different from original images. As we mainly aim to develop efficient deep models with fewer pixel-wise errors, our work does not belong to this group. Readers can refer to [49] for an elaborated survey on deep learning based SISR.

III. LINEAR COMPRESSING BASED SKIP-CONNECTING NETWORK (LCSCNet AND E-LCSCNet)

Our work has two main technical contributions: an (enhanced) linear compressing based skip-connecting structure for developing extremely deep efficient neural networks, and an adaptive fusion strategy for further utilizing intermediate features. In order to better clarify the contribution and function for each of them, here we briefly specify four architectures used in later discussions and ablation studies:
**Basic LCSCNet:** as shown in Fig.1(a) (without red-line parts), it firstly extracts features and then sends them to a series of LCSCBlocks, and the final results are obtained from the upsampling and reconstruction part;

**Basic E-LCSCNet:** quite similar to Basic LCSCNet except for the replacement of LCSCBlock by E-LCSCBlock (Fig.1(a)) and the extra additive skip connections with initial features;

**LCSCNet and E-LCSCNet:** applying the proposed adaptive fusion strategy to Basic LCSCNet and Basic E-LCSCNet respectively, as shown in Fig.1(b).

Because the structure of the above two basic networks are quite simple and we mainly use LCSCNet (E-LCSCNet) to compare with other state-of-the-art works, we will focus on the detailed descriptions on LCSCNet (E-LCSCNet).

As shown in Fig.1(b), our LCSCNet and E-LCSCNet both mainly consist of four parts: 1) a preliminary feature extraction (PFENet), 2) linear compressing based skip-connecting blocks (LCSCBlocks) or enhanced linear compressing based skip-connecting blocks (E-LCSCBlocks) for deep feature exploration, 3) a upsampling and reconstruction net (U&RNet), and 4) an adaptive element-wise fusion of all intermediate outputs. Many previous works [24], [31], [34], [37] learned the residue between HR and its bicubic interpolation and argued that this helps stabilize training and improves performance. When we compare LCSCNet with these works, as shown in Fig.1, the input $I_{in}$ is LR, and the output $I_{out}$ is the residue. Meanwhile, many recent works [40], [42] just learned the mapping between LR and HR. When we compare E-LCSCNet with these methods, we also follow this routine for fairness.

Our PFENet uses a single $3 \times 3$ convolution layer to conduct preliminary feature extraction:

$$F_0 = f_{PFEN}(I_{in}),$$

where $F_0$ denotes extracted features from the LR input.

![Figure 2: The configuration of LCSCUnit.](image)

**A. Configurations of LCSCUnit, LCSCBlock and E-LCSCBlock**

1) **LCSCUnit and LCSCBlock:** The features extracted by the PFENet are then transmitted to the second part of overall network, which uses LCSCBlocks to explore complicated features progressively. An LCSCBlock comprises a fixed number of linear compressing based skip-connecting units (LCSCUnit) with the same configuration. The basic configuration of the LCSCUnit is depicted in Fig.2, where $LU_{i,j}$ denotes the $j$-th unit in the $i$-th LCSCBlock, $Y_{in}$ denotes the input feature maps of this unit and $Y_{out}$ denotes the output feature maps, both maps with $n$ channels. In Fig.2, the upper convolution operator named as linear compressing (LC) layer is of size $1 \times 1 \times n_1$ with $n_1$ output channels. We denote the LC layer in $LU_{i,j}$ as $K_{i,j}^L$.

Motivated by [35], the nonlinear operator in the lower part of Fig.2 consists of two parts: ReLU and the convolution operator denoted as $K_{i,j}^{NL}$ of size $3 \times 3 \times n_2$ output channels. Here the superscripts $L$ and $NL$ denote the convolution kernels for linear and nonlinear transformations, respectively. Then the output of $K_{i,j}^L$ and $K_{i,j}^{NL}$ are concatenated to form a $n$-channel output feature maps. For simplicity, bias is omitted and convolution is replaced by matrix multiplication$^1$, then the whole process of LCSCUnit can be formulated as

$$Y_{out} = \text{concat}(K_{i,j}^L Y_{in}, \ K_{i,j}^{NL} \text{ReLU}(Y_{in})).$$

Furthermore, features and convolution kernels in LCSCUnits can be separated by their properties. As for features, $Y_{out}$ can be divided into $n_1$-channel $Y_{1, out}$ and $n_2$-channel $Y_{2, out}$, where superscripts $L$ and $NL$ in features denote features produced by linear and nonlinear operations, respectively. For convolution kernels, $K_{i,j}^L$ can be divided according to the output channel into $K_{i,j}^{L \cdot}$ and $K_{i,j}^{NL \cdot}$, where superscript $L \cdot$ means the part of the linear-transforming kernel $K_{i,j}^L$ operating on $Y_{1, in}$ and $L \cdot NL$ means the part operating on $Y_{2, in}$.

Notably, although the LC layer with $1 \times 1$ convolution resembles the bottleneck layer that is widely used to reduce dimensions of feature maps [22], [50], the main difference between them is that the bottleneck layer is placed before the nonlinear operator in a cascading manner, while the LC layer parallels the nonlinear operator.

Skip connections in a neural network structure create short paths from early layers to latter layers, which are considered as an effective way to ease the difficulties in training deep neural networks. In all LCSCNet, we implement skip connections mainly by the LC layer in each basic unit. In LCSCUnit, there is a parameter which controls the proportion of the number of linear output channel $n_1$ and the nonlinear output channel $n_2$. This parameter, which can affect the performance of the network, is defined as

$$\rho = \frac{n_2}{n_1 + n_2}.$$  

We find that a fixed $\rho$ for each LCSCUnit throughout the network can already offer a quite good performance. Alternatively, we can set LCSCUnits with different $\rho$, and the LCSCUnits with the same $\rho$ are connected consecutively and can be divided into different LCSCBlocks. For simplicity, we let each LCSCBlock contain the same number of LCSCUnit.

Suppose there are $N$ LCSCBlocks stacked to explore deep features, and $M$ LCSCUnits in an LCSCBlock. Let $LB_{d}^{out}$ denote the $d$-th LCSCBlock with specific $\rho$, $F_{d-1}$ denote its input features and $F_d$ its output features. The mapping of LCSCUnits in this block are denoted by $\{LU_{d,1}(\cdot), LU_{d,2}(\cdot), \ldots, LU_{d,M}(\cdot)\}$, then the whole process of this block can be formulated as

$$F_d = LB_{d}^{out}(F_{d-1}) = LU_{d,M}(LU_{d,M-1}(\cdots (LU_{d,1}(F_{d-1}) \cdots)),$$

$$1 \text{e.g. the convolution operation X+Y is rewritten as XY for simplicity.}$$
and it follows that
\[ F_N = LB_{N}^{ρ_N} (LB_{N-1}^{ρ_{N-1}} (\ldots (LB_{1}^{ρ_{1}} (F_0) \ldots ))). \]  

Furthermore, we investigate how the ordinal position of blocks with different \( ρ \) effects the final performance. Detailed discussions and relative comparative experiments will be demonstrated in Section V-C.

2) E-LCSCBlock: As mentioned in [40], the simplest way to enhance performance via increasing the number of parameters is to increase the width of deep architectures. However, a deep wide network is extremely hard to train. Inspired by the long-term memory connection in [24], we find that if we further concatenate the input and the output of LCSCBlock and then use a \( 1 \times 1 \) bottleneck layer to maintain the compactness of the output channel, it will alleviate the difficulty of training a large LCSCNet. We denote the LCSCBlock with such a long-term memory connection as E-LCSCBlock. Compared with (4), the mapping of E-LCSCBlock can be written as
\[ E LB_d^{ρ_d} (F_{d-1}) = bottle(concat(F_{d-1}, LB_d^{ρ_d} (F_{d-1}))), \]  

where \( ELB_d^{ρ_d} \) denotes the \( d \)-th E-LCSCBlock with specific \( ρ_d \), and \( bottle(\cdot) \) denotes the \( 1 \times 1 \) bottleneck layer. Moreover, we find that deep models of moderate scales using E-LCSCBlocks also perform slightly better than the ones using LCSCBlocks. Further discussions and ablation studies on E-LCSCBlock will be presented in Section V-E.

B. Upsampling and Reconstruction Net (U&RNet)

Sajjadi et al. [48] reported that adding convolution layers after the nearest-neighbor upsampling layer can help alleviate artifacts in SR. We follow this way in our models trained on the 291 dataset using the nearest-neighbor upsampling layer followed by three \( 3 \times 3 \) convolution kernels (except the last one) with ReLU. When we develop models aiming to compare with models trained on DIV2K, we use ESPCN as the U&RNet, as EDSR and RDN did, for fair comparison.

In LCSCNet, the deep features \( \{F_1, F_2, \ldots, F_N\} \), explored hierarchically in its second part by LCSCBlocks, are then sent to U&RNet \( UR(\cdot) \), which maps feature \( F_d \) to output \( Y_d \):
\[ Y_d = UR(F_d), \quad 1 \leq d \leq N. \]  

In E-LCSCNet, like EDSR and RDN, even without directly learning the residue between the HR and its bicubic version, the global residual learning is implemented by adding initial features \( F_0 \) to \( F_d \) before upsampling. That is, the U&RNet \( EUR(\cdot) \) in E-LCSCNet has input and output as
\[ Y_d = EUR(F_d + F_0), \quad 1 \leq d \leq N. \]  

C. Adaptive Element-wise Fusion Strategy

Feature maps of different receptive fields are sensitive to features of different sizes, which are often fused to enhance the performance in various computer vision tasks. In our case, we develop an adaptive element-wise fusion strategy. With \( N \) intermediate results \( \{Y_1, Y_2, \ldots, Y_N\} \) mapped from \( \{F_1, F_2, \ldots, F_N\} \) through U&RNet, a list of weight tensors \( \{W_1, W_2, \ldots, W_N\} \) with the same size of output are determined by \( \{Y_1, Y_2, \ldots, Y_N\} \), which control how much of each raw result contributes to the final fused output. Here the adaptive weight tensors satisfy two traits:
- Trait 1: Each adaptive tensor is determined by all intermediate outputs together, which can be formulated as
  \[ W_i = f_i(Y_1, Y_2, \ldots, Y_N), \quad i = 1, 2, \ldots, N, \]  
  where \( f_i \) is the mapping from \( \{Y_1, Y_2, \ldots, Y_N\} \) to \( W_i \);
- Trait 2: The value of each point in the weight tensor is between 0 and 1. The updating step (Step 6) ensures the output to be a
  \[ \sum_{i=1}^{N} W_i = I, \]  
  where \( I \) is the tensor with all elements being 1.

The final fused output \( M \) is a convex weighted average of intermediate outputs \( \{Y_1, Y_2, \ldots, Y_N\} \):
\[ I_{out} = M = \sum_{d=1}^{N} W_d Y_d. \]  

Inspired by the gate unit in LSTM, by adopting a series of \( 1 \times 1 \) convolution kernels followed by sigmoid activation functions, we develop a heuristic algorithm to construct the fused output \( M \), in which weight tensors satisfy the above two traits, as summarized in Algorithm 1. A sketch for Algorithm 1 is plotted in Fig.3, in which intermediate variable tensor \( \alpha_i \) \( (i = 1, \ldots, N - 1) \) is generated progressively given current SR outputs \( \{Y_1, \ldots, Y_i\}, 1 \times 1 \) convolution kernel \( C_i \) and sigmoid activation function. The use of sigmoid activation functions ensures the element-wise value of \( \alpha_i \) to be between 0 and 1. The updating step (Step 6) ensures the output to be a convex weighted average of current inputs \( \{Y_1, \ldots, Y_i\} \). From Algorithm 1, \( \{W_1, W_2, \ldots, W_N\} \) can be obtained as
\[ W_k = \begin{cases} \prod_{i=1}^{k-1} \alpha_i, & k = 1; \\ (I - \alpha_{k-1}) \left( \prod_{i=k}^{N-1} \alpha_i \right), & 2 \leq k < N; \\ I - \alpha_{N-1}, & k = N, \end{cases} \]  

where \( \alpha_{N-1} \) contains the information of \( \{Y_1, Y_2, \ldots, Y_N\} \). As (12) shows that every \( W_k \) contains \( \alpha_{N-1} \), the first trait in (9) is satisfied; with simple algebra, the second trait in (10) is also...
verified, and hence the rationality of the proposed methods.

D. Loss Function for Training

During training, we minimize the $\ell_1$ loss $L_1(x, y) = |x - y|$ over the training set of $M$ samples. Let $X^{(i)}$ denote the $i$-th ground-truth HR label in the training set and $I_{out}^{(i)}$ denote the corresponding output of network. Then the loss function $l$ is

$$l(I_{out}, X) = \frac{1}{M} \sum_{i=1}^{M} L_1(I_{out}^{(i)}, X^{(i)}). \quad (13)$$

When we apply the adaptive element-wise fusion strategy, we use the multi-supervised methods mentioned in [23], [24] to train our model. The loss function of multi-supervised LCSCNet can be formulated as

$$L(\Theta) = l(I_{out}, X) + \beta \sum_{d=1}^{N} l(Y_d, X), \quad (14)$$

where $I_{out}$ and $\{Y_1, \ldots, Y_N\}$ are defined in (11), and $\beta$ is a trade-off parameter.

IV. DISCUSSIONS

In this section, we mainly discuss the motivation and characteristics of Basic LCSCNet by showing its connections to DenseNet and its differences from ResNet and DenseNet.

A. Basic LCSCNet as an Efficient Variant of DenseNet

In this sub-section, we illustrate that Basic LCSCNet can be transformed from DenseNet with small changes on topology: we first show the redundancy in DenseNet and then introduce Basic LCSCNet as a remedy for this redundancy.

Skip connection in DenseNet is implemented by directly concatenating all former features to be the input of current layer. For illustration, a 4-layer DenseBlock is depicted in Fig.4(a), in which $Y_0$ is a $k$-channel input feature, $Y_i$ is the newly-explored feature after nonlinear mapping $f_i$ ($i = 1, 2, 3$), where $f_i$ consists of a ReLU followed by a $3 \times 3$ convolution kernel with $k_0$ output channels ($k_0$ is also called growth rate in DenseNet). The last nonlinear mapping $f_4$ acts as a transition layer; $C$ means a concatenation operator in the channel dimension.

To have a better understanding of DenseBlock in Fig.4(a), we can simplify Fig.4(a) into its equivalent form in Fig.4(b), where $Y_i = \text{concat}(Y_0, \ldots, Y_i)$ ($i = 0, 1, 2, 3$). By denoting the concatenation of former features as $Y_i$, excessive skip connections in Fig.4(a) are simplified into concise adjacent skip connections. For simplicity, unless otherwise specified, we take the DenseBlock in the form of Fig.4(b) as the basic DenseBlock structure.

As shown in Fig.4(a) and Fig.4(b), when depth increases, the number of parameters of the convolution kernel in DenseNet also increases. To reduce the parameter amount, the authors of DenseNet applied the bottleneck layer\(^2\) before every nonlinear mapping and called it B-DenseNet. Fig.4(c) is the bottleneck version of Fig.4(b), where $B_i$ ($i = 1, 2, 3$) is the $1 \times 1$ convolution kernel with $b$ output channels. We can see that the parameter amount of every convolution kernel $f_i$ for nonlinear mapping is a constant now, and only the parameter amount of bottleneck layer with fewer parameters increases with depth.

Although B-DenseBlock has reduced the parameter amount to a great extent, the parameter amount of each basic unit in B-DenseBlock still increases with depth. To further control the parameter amount, we make the parameter amount of each unit in B-DenseBlock a constant. One simple but effective solution is to move forward the bottleneck layer in each unit, reducing the number of channels of input feature to $b$ by the bottleneck layer before they are sent to the concatenation part, as shown in Fig.4(d). We can set $k = b + k_0$ to make channels of each feature unchanged. In this case, if we re-depict Fig.4(d) by allocating the bottleneck layer to each branch and using a nonlinear mapping $f'_i$ to replace $B_i \circ f_i$, the structure of Basic LCSCNet reemerges, as shown in Fig.4(e).

\(^2\)Many works add nonlinear activation before a $1 \times 1$ convolution kernel to make the bottleneck layer; here we take the $1 \times 1$ convolution kernel as the bottleneck layer.
of each LCSCUnit is already a constant, it is unnecessary to compress features again. From this perspective, BC-DenseNet can also be transferred to Basic LCSCNet in a similar way. Now look back into Fig.2: the nonlinear output channel is just the growth rate in DenseNet, denoting how many new features are explored, and $1 - \rho = \frac{n_1}{n_1 + n_2}$ acts as some kind “compress ratio” denoting how many former features have flowed to the current stage through skip connections.

B. Differences from ResNet and DenseNet

This sub-section aims to illustrate the differences between Basic LCSCNet and ResNet/DenseNet as well as the novelty of our proposed network. It is still an open problem to compare different deep architectures. When different ways of skip connections are employed to alleviate training difficulties, the output features explored by nonlinear mapping with different skip connections have different constitutions. We suppose that by comparing different constitutions of the feature maps, we could get some useful information about the properties of different skip-connection architectures.

1) Feature maps of ResNet: We use the structure in [35]. Let $Y_k$ and $Y_{k+1}$ denote the input and output of block $k$, respectively, and let $f^R_k(\cdot)$ denote the nonlinear transformation in block $k$. Then the mathematical formulation of block $k$ is

$$Y_{k+1} = Y_k + f^R_k(Y_k) = Y_j + \sum_{i=j}^k f^R_i(Y_i), \quad 1 \leq j \leq k. \quad (15)$$

From (15), we can see that in ResNet, skip connection is implemented by element-wise summation between adjacent features. Compared with traditional plain architecture, any former maps $Y_j (j = 1, \ldots, k)$ can be added to the current state $Y_{k+1}$, creating many short paths for more “smooth” gradient flow during back-propagation. Moreover, it is extremely concise because no extra parameter is required for this skip connection.

2) Feature maps of DenseNet: We employ the structure shown in Fig.4(b) to illustrate the properties of feature maps in DenseNet. Let $Y'_k$ and $Y'_{k+1}$ denote the input and output of unit $k$ in a DenseBlock, respectively, and $f^D_k(\cdot)$ the nonlinear transformation. Then the formulation of unit $k$ is

$$Y'_{k+1} = \text{concat}(Y'_k, f^D_k(Y'_k)), \quad (16)$$

and it follows that

$$Y'_{k+1} = \text{concat}(Y'_j, f^D_j(Y'_j), \ldots, f^D_k(Y'_k)), \quad 1 \leq j \leq k. \quad (17)$$

Like ResNet, all the former features in DenseNet can be fused into the current stage, but instead of summation, all the feature maps are concatenated in the channel dimension. Such a skip connection has both advantages and disadvantages compared with ResNet. One obvious advantage is that when features produced in DenseNet are sent to follow-up convolution kernels to explore new features, the features from different stages use different convolution kernels, while in ResNet the reused parts and newly-explored ones share the same convolution kernel. From this perspective, connecting features by element-wise summation may restrict a network from reaching better solutions in some cases. As for disadvantage, concatenating features need more following convolution kernels. As shown in Fig.4(a) and Fig.4(b), the parameter amount of DenseUnit
increases with depth. When a DenseNet is very deep, even a small growth rate may lead to a large parameter amount.

3) Feature maps of Basic LCSCNet: From the analysis above, we can conclude that the feature re-usage of ResNet benefits from its concise skip connection between adjacent basic blocks and the new feature exploration of DenseNet mainly benefits from its little relevance between newly-explored feature maps and former ones. We have already seen that in Basic LCSCNet, former features are firstly compressed and then concatenated with the newly-explored features. Now we examine how the former features are combined in the current stage. Let \( Y_k \) and \( Y_{k+1} \) denote the input and output of the \( k \)-th LCSCUnit, and \( K^L_k \) and \( K^{NL}_k \) its convolution kernels. From Fig.2 and (2), we can derive the formulation of \( 1 \times 1 \) convolution in the LC layer as

\[
Y_{k+1}^L(c_o) = \sum_{c_i=1}^{n} K^L_k(c_o, c_i)Y_k(c_i) \\
= \sum_{c_i=1}^{n} K^L_k(c_o, c_i)Y_k^L(c_i) + \sum_{c_i=n_1+1}^{n} K^L_k(c_o, c_i)Y_{NL}^L(c_i - n_1) \\
= \sum_{c_i=1}^{n_2} K^{NL,LL}_k(c_o, c_i)Y_k^L(c_i) + \sum_{c_i=1}^{n_2} K^{NL,L}_k(c_o, c_i)Y_{NL}^L(c_i),
\]

(18)

where \( c_i \) denotes the input channel and \( c_o \) the output channel. For simplicity, (18) can be rewritten as

\[
Y_{k+1}^L = K^{L,L}_kY_k^L + K^{NL,LL}_kY_{NL}^L.
\]

(19)

Applying the same approach to the convolution kernel \( K^{NL}_k \) in nonlinear transformation, we have

\[
Y_{k+1}^{NL} = K^{NL,L}_kReLU(Y_k^L) + K^{NL,LL}_kReLU(Y_{NL}^L),
\]

(20)

where \( K^{NL,L}_k \) is the part of \( K^{NL}_k \) only operating on \( Y_k^L \) and \( K^{NL,LL}_k \) only on \( Y_{NL}^L \).

A ‘global’ form of (19) is

\[
Y_{k+1}^L = P_{k+1}^L + P_{k+1}^{NL},
\]

(21)

\[
P_{k+1}^L = \prod_{i=1}^{k} K^{L,L}_iY_1^L,
\]

(22)

\[
P_{k+1}^{NL} = \sum_{i=1}^{k-1} (K^{L,LL}_i \prod_{j=i+1}^{k} K^{L,L}_j)Y_i^{NL} + K^{NL,LL}_kY_{NL}^L.
\]

(23)

From (19), we can see that \( Y_{k+1}^L \) restores the information of all former feature maps in the form of weighted summation. From (20), we can see that \( Y_{k+1}^{NL} \) is the new features explored by new nonlinear transformation. Among the deep features produced by deep architectures, newly-explored parts are thought to be more important. In Basic LCSCNet, we concatenate newly-explored features with the former ones like DenseNet, ensuring that features of different kinds can be treated differently. Meanwhile, as former features in the current stage are mainly aimed to create paths for training deep networks, instead of concatenating each former features separately, we compress all the former features and then concatenate them with the newly-explored ones, making it quite parameter-economic like ResNet.

V. ABLATION STUDIES

A. Comparison with ResNet and DenseNet

In this sub-section, we replace LCSCUnit in our basic LCSCNet by ResBlock or DenseUnit with the bottleneck layer. The three networks for comparison are all 34-layer, where Basic LCSCNet and DenseNet both have 30 units while ResNet has 15 blocks. As discussed above, the growth rate in DenseNet plays a similar role to the channel number of the nonlinear output. To compare fairly, if we set all output feature channels to 64 and \( \rho \) of every LCSCUnit to 0.5, then the growth rate of DenseNet is 32 and the output channel of the bottleneck is 64. As for BC-DenseNet, for example, BC-Dense_B3_U10 means dividing the network into 3 blocks uniformly and add a compressing layer at the end of each block, whose output channel is 64. Here we train the above three models with the 291 dataset for \( \times 3 \) scale and the results are shown in Table I. We use PSNR/SSIM [51] to measure reconstruction, and parameter amounts to measure storage efficiency. We can see that Basic LCSCNet has the least parameters and the competitive performance to DenseNet both better than ResNet.

B. Efficiency Brought by the LC layer

Here we discuss the rationale behind implementing the LC layer with \( 1 \times 1 \) convolution and its advantage on parameter efficiency. It is known that increasing receptive fields is essential for exploring deeper features. From Section IV we can see that the LC layer helps transport the previous features and does not produce newly-explored features directly. Hence, we do not need to use \( 3 \times 3 \) convolution to increase receptive fields in the LC layer and \( 1 \times 1 \) convolution is sufficient. To support
this view, we apply $3 \times 3$ convolution to the LC layer of the Basic LCSCNet mentioned in Section V-A. From Table II, we can see that the LC layer with $3 \times 3$ convolution indeed does not achieve better performance.

The usage of $1 \times 1$ convolution as the LC layer also makes the proposed architecture more parameter-economic compared with ResNet and DenseNet. Firstly we compare Basic LCSCNet with ResNet and DenseNet. A basic unit in ResNet with $n_1$ input channels, $n_2$ output channels and a $k \times k$ nonlinear transformation convolution kernel has $n_1n_2k^2$ parameters. The number of parameters of a basic unit in Basic LCSCNet, with $n_1$ input channels, $n_2$ output channels, a $k \times k$ nonlinear transformation convolution kernel and parameter $\rho_0$, is $n_1n_2(k^2 \rho_0 + 1 - \rho_0)$. The ratio of parameter amounts of these two units with the same $n_1$, $n_2$ and $k$ is

$$p_{L/R}(n_1, n_2, k) = \frac{\rho_0 + 1}{k^2}(1 - \rho_0). \quad (24)$$

As illustrated before, good performance can be obtained when $\rho_0$ is around 0.5. In practice, the size of a convolution kernel for feature extraction is usually an odd bigger than 3. So when $\rho_0$ is 0.5, $p_{L/R}(n_1, n_2, k) < 55.7\%$, which means the parameter amount of Basic LCSCNet is just half of the ResNet’s.

As for DenseNet, the parameter amount of a basic unit increases with depth. We take B-DenseNet as example: if the output channel of nonlinear mapping in LCSCUnit and DenseUnit is both $n_2$, the output channel of $1 \times 1$ compressing layer is both $n_1$, the nonlinear kernel size is $k \times k$, then the parameter amount of LCSCUnit is always $(n_1 + n_2)(k^2n_2 + n_1)$, while the parameter amount of the $p$-th DenseUnit is $(pn_2n_1 + k^2n_1n_2)$. If such Basic LCSCNet and DenseNet both have $L$ nonlinear mapping layers, the ratio of parameter amounts of the two networks with the same $n_1$, $n_2$ and $k$ is

$$p_{L/D}(L; n_1, n_2, k) = \frac{2}{2k^2 + L + 1} \left( \frac{1}{\rho_0} + \frac{k^2}{1 - \rho_0} \right). \quad (25)$$

From (25), we can see the advantage of Basic LCSCNet is more remarkable when the network goes deeper. When we compare Basic LCSCNet with an $L$-layer B-DenseNet of $N$ blocks, if the transition layer is omitted for simplicity, the ratio can be obtained by replacing $L$ with $\frac{L}{N}$ in (25).

C. Investigation into Parameter $\rho$

1) Fixed $\rho$ throughout the network: In this situation, we find when $\rho$ is around 0.5, the best performance could be achieved. Table III shows 34-layer Basic LCSCNets for $\times 3$ scale with different fixed $\rho$. As here we mainly focus on the effect of $\rho$, the experiments are conducted without adaptive element-wise fusion. Firstly, we consider two special cases of $\rho$. When $\rho$ is 0, the feature exploration part is a linear transformation; if the upsampling and reconstruction part is taken into account, the whole network has just two nonlinear convolution layers, whose fitting capacity for complex functions is relatively poor. In contrast, when $\rho$ is 1, Basic LCSCNet becomes the traditional feedforward neural network without skip connections, which is difficult to train. Hence $\rho$ balances the fitting capacity and the training ease of Basic LCSCNet. As Table III shows, when $\rho = 0.25$, the performance is suboptimal because of the restricted fitting capacity; when $\rho = 0.75$, the performance is suboptimal mainly because the LC layers output fewer feature maps. As we discussed before, the output feature maps of LC layers restore the information of former features, insufficiency of which leads to insufficient skip connections and thus training difficulty increase and performance decline.

2) Different $\rho$ throughout the network: In this situation, the LCSCUnits with the same $\rho$ form an LCSCBlock and different LCSCBlocks have different $\rho$s. We find that as depth increases, the $\rho$ of an LCSCBlock should be decreased slightly to improve performance. Table IV shows the relevant experimental results on the 34-layer Basic LCSCNets with different ordinal positions of $\rho$ list for $\times 3$ scale. One possible reason for this phenomenon is that as depth increases, exploring higher-level features becomes harder, so there is less room for newly-explored features. Meanwhile, as information on former feature maps accumulates, more room is needed for reusing former features.

D. Ablation Studies on Different Fusion Strategies

Table V compares properties of the vectorized fusion [23], MSCN [43] and our proposed fusion method. We can see that our method incorporates the advantages of the vectorized fusion and MSCN. We also compare these fusion strategies quantitatively. We train 34-layer LCSCNets for $\times 2$ scale with $\rho$ list [0.75, 0.6875, 0.625, 0.5625, 0.5], and every six LCSCUnits with the same $\rho$ form a LCSCBlock. In Table VI, Basic LCSCNet, LCSCNet_S, LCSCNet_M and LCSCNet denote the LCSCNet without any fusion, with vectorized fusion, with MSCN and with our proposed method, respectively. Here we note that our implementation of MSCN is slightly different from the original one. In the original MSCN, the input of the weight module is the bicubic of LR, while in our LCSCNet we use the upsampled LR input. This small difference should have little influence on final results. As Table VI shows, when combined with Basic LCSCNet, our fusion strategy performs better than the other two fusion benchmarks.

E. Ablation Studies on LCSCBlock and E-LCSCBlock

Firstly, we show that when we use LCSCUnits to construct deep models of moderate scales, the advantage of E-LCSCBlock is mild. With the 291 dataset, we train 34-layer, 44-layer and 54-layer Basic LCSCNets for $\times 3$ scale, $\rho$ of each unit is 0.5, and the number of feature channels is 64. For comparison, we use 10 LCSCUnits to constitute an E-LCSCBlock and train 37-layer, 48-layer and 59-layer Basic E-LCSCNets, respectively. As Table VII shows, every Basic E-LCSCNet performs better than its corresponding Basic LCSCNet.
Table III: Average $\times 3$ PSNR/SSIM for Basic LCSCNets with different $\rho$ on the Set5, Set14, BSD100 and Urban100 datasets, respectively. \textit{Red} color indicates the best performance.

| $\rho$ | 0 | 0.25 | 0.375 | 0.5 | 0.625 | 0.75 | 1 |
|-------|---|------|-------|----|-------|-----|---|
| Set5  | 32.66 | 9.103 | 33.86 | 9.229 | 33.97 | 9.242 | 33.99 | 9.241 | 33.94 | 9.241 | 33.92 | 9.237 | 31.78 | 8.894 |
| Set14 | 29.27 | 8.208 | 29.82 | 8.330 | 29.90 | 8.337 | 29.87 | 8.37 | 29.93 | 8.340 | 29.85 | 8.333 | 28.57 | 8.012 |
| BSD100| 28.41 | 7.858 | 28.85 | 7.984 | 28.88 | 7.997 | 28.87 | 7.994 | 28.87 | 7.994 | 28.85 | 7.990 | 27.92 | 7.648 |
| Urban100| 26.21 | 8.011 | 27.16 | 8.296 | 27.25 | 8.329 | 27.24 | 8.324 | 27.24 | 8.321 | 27.20 | 8.312 | 25.50 | 7.761 |

Table IV: The effect of ordinal position of block with different $\rho$ on average $\times 3$ PSNR/SSIM for the Set5, Set14, BSD100 and Urban100 datasets. Each block has the same number of LCSCUnits.

| $\rho$ | [0.5,0.75] | [0.75,0.5] | [0.5,0.625,0.75] | [0.75,0.625,0.5] |
|-------|-------------|-------------|-----------------|-----------------|
| Set5  | 33.97/0.9240 | 34.02/0.9244 | 33.89/0.9234 | 33.95/0.9239 |
| Set14 | 29.91/0.8341 | 29.91/0.8343 | 29.86/0.8332 | 29.86/0.8336 |
| BSD100| 28.88/0.7998 | 28.89/0.8001 | 28.86/0.7993 | 28.88/0.7994 |
| Urban100| 27.28/0.8336 | 27.31/0.8343 | 27.20/0.8317 | 27.25/0.8323 |

Table V: Brief comparisons among different fusion strategies.

| Adaptiveness | Pixel-wise | Normalization | Multi-supervised training |
|--------------|------------|---------------|--------------------------|
| Vectorized fusion [23] | MSCN [43] | Our method | Vectorized fusion [23] | MSCN [43] | Our method |
| $\times$ | $\checkmark$ | $\checkmark$ | $\times$ | $\checkmark$ | $\checkmark$ |

Table VI: Average $\times 2$ PSNR/SSIM for LCSCNets with different fusions on the Set5, Set14, BSD100 and Urban100 datasets, respectively. \textit{Red} indicates the best results.

| Basic LCSCNet | LCSCNet_S | LCSCNet_M | LCSCNet |
|--------------|----------|----------|--------|
| Set5 | 37.77 | 0.9558 | 37.80 | 0.9560 | 37.79 | 0.9559 | 37.84 | 0.9559 |
| Set14 | 33.23 | 0.9140 | 33.26 | 0.9144 | 33.25 | 0.9142 | 33.31 | 0.9144 |
| BSD100 | 32.06 | 0.8980 | 32.05 | 0.8981 | 32.07 | 0.8981 | 32.08 | 0.8984 |
| Urban100 | 31.15 | 0.9182 | 31.26 | 0.9197 | 31.23 | 0.9190 | 31.31 | 0.9200 |

Then we show that when we aim to develop an extremely deep and wide network, E-LCSCBlock can make up the deficiencies of LCSCBlock. With the DIV2K dataset we train a Basic LCSCNet for $\times 2$ scale of $\rho$ list $[0.75, 0.71875, 0.6875, 0.65625, 0.625, 0.59375, 0.5625, 0.53125, 0.5]$, every sixteen LCSCUnits with the same $\rho$ form a LCSCBlock, and the output channel of each feature is 128. Its convergence curve is the blue one in Fig.5, indicating quite poor performance. For comparison, we train the LCSCNet with the same setting, and its performance (the green curve in Fig.5) is significantly better than Basic LCSCNet. We contribute this obvious improvement to the extra short paths created by the adaptive fusion strategy, which suggests that more short paths may further help in this case. The experimental results shown in Fig.5 also support this view: when we evolve (Basic) LCSCNet into (Basic) E-LCSCNet, the performance of deep architecture booms.

Figure 5: Convergence comparison between deep wide (Basic) LCSCNet and (Basic) E-LCSCNet on the DIV2K validation set for scale $\times 2$.

VI. EXPERIMENTAL RESULTS

A. Comparison with State-of-the-Art Models

It is well known that the training set and the parameter amount largely influence the final performance of a model. To compare with various representative models fairly, we divide these models into three categories: models trained on the 291 dataset [11], [52], light models (Params $< 2M$) trained on the DIV2K dataset [38] and large models (Params $> 10M$) on DIV2K. When compared with models on the 291 dataset such as VDSR [31], DRCN [23], LapSR [37], DRRN [34] and MemNet [24], we train a 76-layer LCSCNet with the proposed fusion strategy, $\rho$ list is also $[0.75, 0.71875, 0.6875, 0.65625, 0.625, 0.59375, 0.5625, 0.53125, 0.5]$ but every eight units with the same $\rho$ form a block, denoted by LCSC$_{76}$291. When compared with light models on DIV2K and similarly large datasets such as SelNet [53], SRDenseNet [36], CARN [54] and FALSR-A [55], because the fusion part is quite computation-consuming, our light models was developed just based on Basic E-LCSCNet. Our light models share the same $\rho$ list with LCSC$_{76}$291, but every six units with the same $\rho$ form a block, denoted by BE-LCSC$_L$. When compared with large models on DIV2K such as EDSR [40] and RDN [42], the E-LCSCNet mentioned in Section V-E is adopted.

Quantitative comparisons on BE-LCSC$_L$ are listed in Table VIII. Because operations in neural networks for SISR are mainly multiplication along with addition, we use the number of composite multiply-accumulate operations in CARN, denoted by Mult&Adds, to measure computational efficiency, and we also assume that the HR image is $1280 \times 720$. From Table VIII, we can see that among the models trained on 291,
Table VII: Average ×3 PSNR/SSIM for Basic LCSCNet and its corresponding Basic E-LCSCNet on Set5, Set14, BSD100 and Urban100.
All the models are of moderate scales (Parameter amount < 150K).

|           | LC_34  | E-LC_37 | LC_44  | E-LC_48 | LC_54  | E-LC_59 |
|-----------|--------|---------|--------|---------|--------|---------|
| Set5      | 33.99/0.9241 | 34.01/0.9248 | 34.02/0.9244 | 34.05/0.9251 | 34.03/0.9244 | 34.08/0.9248 |
| Set14     | 29.87/0.8337 | 29.92/0.8349 | 29.85/0.8334 | 29.90/0.8345 | 29.86/0.8340 | 29.89/0.8339 |
| BSD100    | 28.87/0.7994 | 28.89/0.8002 | 28.87/0.7996 | 28.90/0.8004 | 28.89/0.7998 | 28.89/0.7998 |
| Urban100  | 27.24/0.8324 | 27.28/0.8340 | 27.23/0.8326 | 27.29/0.8343 | 27.27/0.8330 | 27.32/0.8347 |

LCSC_76_291 achieves better accuracy than MemNet. As for efficiency, MemNet has fewer parameters due to its recursive structure, but LCSC_76_291 is more computation-efficient than MemNet. When compared with SelNet and CARN, BE-LCSC_L is moderately computation-consuming but achieves obvious improvement. Among large models on DIV2K, our E-LCSCNet has the fewest Params for every scale. For ×2 scale, our E-LCSCNet holds the same level with RDN but with a clear advantage in Mult&Adds. For ×3 and ×4 scale, our E-LCSCNet performs better than EDSR and RDN, but is somehow more computation-consuming than RDN due to its fusion part.

Representative qualitative comparisons are shown in Figs.6-8. In Fig.6, our model restores the grid structure more precisely with fewer artifacts than other models. In Fig.7, compared with blurry characters generated by other models, our result has sharper edges. In Fig.8, compared with EDSR and RDN, E-LCSCNet recovers the line with the least blurry.

B. Implementation Details
For training LCSCNet, we augment data (90°, 180° and 270° rotation), and then downsample the LR input with the desired scaling factor. Like many methods trained on 291, we only take the luminance component for training. Ground truth for training is the residue between the bicubic of LR image and the original HR image, and all inputs are scaled into [-1, 1]. When trained on 291, training images are split into patches of sizes $18^2/36^2$, $12^2/36^2$ and $21^2/84^2$, respectively. We initialize all the convolution kernels as suggested by [56]. All intermediate feature maps have 64 channels. For optimization, we use Adam [57] with its default settings. Learning rate is initialized as $10^{-4}$, and is divided by 10 every 15 epochs over the whole augmented dataset and the training is stopped after 60 epochs. For training, we use Keras [58]; for testing, we use MatConvNet [59].

The training of BE-LCSC_L and E-LCSCNet is based on the PyTorch [60] version of EDSR with the same setting of EDSR except that the batch size is 32, and training is terminated after 650 epochs. The codes are available from https://github.com/XuechenZhang123/LCSC.
Table VIII: Quantitative comparisons among mainstream deep models for SISR. To compare fairly, we divide models into three categories: models trained on 291, light models (Params < 2M) trained on DIV2K, and large models trained on DIV2K. For each scale, we compare the models within the same category, and the best performance is highlighted in Red. In DRCN, MemNet, LCSC_76_291 and E-LCSCNet, extra Multi&Adds of the multi-supervised fusion part are added after the Multi&Adds of the basic structure.

| Scale | Model       | Training data | Params | Multi&Adds | Set5     | Set14    | BSD100 | Urban100 |
|-------|-------------|---------------|--------|------------|----------|----------|--------|----------|
| 2     | VDSR        | 291           | 665K   | 612.6G     | 37.53/0.9587 | 33.03/0.9124 | 31.90/0.8960 | 30.76/0.9140 |
|       | DRCN        | 291           | 1774K  | 9243.0G+8731.3G | 37.63/0.9588 | 33.04/0.9118 | 31.85/0.8942 | 30.75/0.9133 |
|       | LapSRN      | 291           | 813K   | 29.9G      | 37.52/0.9591 | 33.08/0.9130 | 31.80/0.8950 | 30.41/0.9101 |
|       | DRRN        | 291           | 297K   | 6769.9G    | 37.74/0.9591 | 33.23/0.9136 | 32.05/0.8973 | 31.23/0.9188 |
|       | MemNet      | 291           | 667K   | 2261.8G+3.2G | 37.78/0.9597 | 33.28/0.9142 | 32.08/0.8978 | 31.31/0.9195 |
|       | LCSC_76_291 | 291           | 1844K  | 407.8G+616.3G | 37.86/0.9600 | 33.34/0.9146 | 32.10/0.8985 | 31.34/0.9204 |

| 2     | SelNet      | DIV2K         | 974K   | 225.7G     | 37.89/0.9598 | 33.61/0.9160 | 32.08/0.8984 | -        |
|       | CARN        | DIV2K         | 1592K  | 222.8G     | 37.76/0.9590 | 33.52/0.9166 | 32.09/0.8978 | 31.92/0.9256 |
|       | FALSR-A     | DIV2K         | 1021K  | 234.7G     | 37.82/0.9595 | 33.55/0.9168 | 32.12/0.8987 | 31.93/0.9256 |
|       | BE-LCSC_L   | DIV2K         | 1552K  | 358.6G     | 38.01/0.9600 | 33.67/0.9160 | 32.23/0.9002 | 32.31/0.9297 |
| 3     | EDSR        | DIV2K         | 40.7M  | 9379.4G    | 38.11/0.9602 | 33.92/0.9195 | 32.12/0.9013 | 32.93/0.9351 |
|       | D_DBPN      | DIV2K+Flickr  | 5876.3K| 3429.0G    | 38.09/0.9600 | 33.87/0.9191 | 32.27/0.9000 | 32.55/0.9324 |
|       | RDN         | DIV2K         | 22.1M  | 5096.2G    | 38.24/0.9614 | 34.01/0.9212 | 32.34/0.9017 | 32.89/0.9353 |
|       | E-LCSCNet   | DIV2K         | 14.2M  | 3126.4G+1251.7G | 38.23/0.9608 | 33.85/0.9180 | 32.36/0.9018 | 32.93/0.9351 |

| 3     | SelNet      | DIV2K         | 1159K  | 120.0G     | 34.27/0.9257 | 30.30/0.8399 | 28.97/0.8025 | -        |
|       | CARN        | DIV2K         | 1592K  | 118.8G     | 34.29/0.9255 | 30.29/0.8407 | 29.06/0.8034 | 28.06/0.8493 |
|       | BE-LCSC_L   | DIV2K         | 1736K  | 179.1G     | 34.39/0.9265 | 30.33/0.8395 | 29.12/0.8065 | 28.25/0.8540 |
| 4     | EDSR        | DIV2K         | 43.7M  | 4471.8G    | 34.65/0.9280 | 30.52/0.8462 | 29.25/0.8093 | 28.80/0.8653 |
|       | D_DBPN      | DIV2K+Flickr  | -      | -          | -            | -           | -        | -        |
|       | RDN         | DIV2K         | 22.3M  | 2284.7G    | 34.71/0.9296 | 30.57/0.8468 | 29.26/0.8093 | 28.80/0.8653 |
|       | E-LCSCNet   | DIV2K         | 14.9M  | 1389.5G+1251.7G | 34.71/0.9286 | 30.56/0.8460 | 29.27/0.8104 | 28.83/0.8658 |

| 4     | SRDenseNet  | ImageNet Subset | 2015K | 389.9G    | 32.02/0.8934 | 28.50/0.7782 | 27.53/0.7337 | 26.05/0.7819 |
|       | SelNet      | DIV2K         | 1417K  | 83.1G      | 32.00/0.8931 | 28.49/0.7783 | 27.44/0.7325 | -        |
|       | CARN        | DIV2K         | 1592K  | 90.9G      | 32.13/0.8937 | 28.60/0.7806 | 27.58/0.7349 | 26.07/0.7837 |
|       | BE-LCSC_L   | DIV2K         | 1699K  | 124.8G     | 32.22/0.8948 | 28.66/0.7806 | 27.62/0.7390 | 26.22/0.7908 |
| 3     | EDSR        | DIV2K         | 43.1M  | 2890.0G    | 32.46/0.8968 | 28.80/0.7876 | 27.71/0.7420 | 26.64/0.8033 |
|       | D_DBPN      | DIV2K+Flickr  | 10.3M  | 5715.4G    | 32.47/0.8980 | 28.82/0.7860 | 27.72/0.7400 | 26.38/0.7946 |
|       | RDN         | DIV2K         | 22.6M  | 1300.7G    | 32.47/0.8990 | 28.81/0.7871 | 27.72/0.7419 | 26.61/0.8028 |
|       | E-LCSCNet   | DIV2K         | 14.8M  | 781.6G+1700.7G | 32.51/0.8984 | 28.81/0.7871 | 27.73/0.7433 | 26.64/0.8033 |
VII. CONCLUSION AND FUTURE WORKS

In this paper, we propose the linear compressing based skip-connecting network (LCSCNet) for image SR, which combines the merits of the parameter-economic form of ResNet and the effective feature exploration of DenseNet. Linear compressing layers are adapted to implement skip connections, connecting former features and separating them from the newly-explored features. Compared with previous deep models with skip connections, our LCSCNet can explore relatively more new features with lower computational costs. Based on LCSCNet, to improve the performance of extremely deep and wide networks, the Enhanced LCSCNet is developed. An adaptive element-wise fusion strategy is also proposed, not only for further exploiting hierarchical information from diverse levels of deep models, but also for stabilizing the training deep models by adding extra paths for gradient flows. Comprehensive experiments and discussions are presented in this paper and demonstrate the rationality and superiority of the proposed methods.

Future work can be mainly explored from the following two aspects: 1) it would be worthwhile to try to apply LCSCNet and E-LCSCNet or their basic units to other computer vision tasks; 2) in terms of Multi&Add in Table VIII, we can see the computational cost for this part is still somewhat high despite that we have managed to control its complexity; therefore, further efforts can be made to further improve its efficiency.

ACKNOWLEDGMENT

We would like to thank the authors of [23], [24], [31], [34], [37], [40], [42], [54] for releasing their source codes and models for comparison. We would also like to thank the Associate Editor and anonymous reviewers for their selfless dedication and constructive suggestions.

REFERENCES

[1] C.-Y. Yang, C. Ma, and M.-H. Yang, “Single-image super-resolution: A benchmark,” in Proceedings of the European Conference on Computer Vision, 2014, pp. 372–386.
[2] S. C. Park, M. K. Park, and M. G. Kang, “Super-resolution image reconstruction: a technical overview,” IEEE Signal Processing Magazine, vol. 20, no. 3, pp. 21–36, 2003.
[3] R. Keys, “Cubic convolution interpolation for digital image processing,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, no. 6, pp. 1153–1160, 1981.
[4] C. E. Duchon, “Lanczos filtering in one and two dimensions,” Journal of Applied Meteorology, vol. 18, no. 8, pp. 1016–1022, 1979.
[5] S. Dai, M. Han, W. Xu, Y. Wu, Y. Gong, and A. K. Katsaggelos, “SoftCuts: a soft edge smoothness prior for color image super-resolution,” IEEE Transactions on Image Processing, vol. 18, no. 5, pp. 969–981, 2009.
[6] J. Sun, Z. Xu, and H.-Y. Shum, “Image super-resolution using gradient prior,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[7] Q. Yan, Y. Xu, X. Yang, and T. Q. Nguyen, “Single image superresolution based on gradient profile sharpness,” IEEE Transactions on Image Processing, vol. 24, no. 10, pp. 3187–3202, 2015.
[8] A. Marquina and S. J. Osher, “Image super-resolution by TV regularization and Bregman iteration,” Journal of Scientific Computing, vol. 37, no. 3, pp. 367–382, 2008.
[9] W. T. Freeman, T. R. Jones, and E. C. Pasztor, “Example-based super-resolution,” IEEE Computer Graphics and Applications, vol. 22, no. 2, pp. 56–65, 2002.
[10] H. Chang, D.-Y. Yeung, and Y. Xiong, “Super-resolution through neighbor embedding,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2004, pp. 275–282.
[11] J. Yang, J. Wright, T. S. Huang, and Y. Ma, “Image super-resolution via sparse representation,” IEEE Transactions on Image Processing, vol. 19, no. 11, pp. 2861–2873, 2010.
[12] R. Zeyde, M. Elad, and M. Protter, “On single image scale-up using sparse-representations,” in Proceedings of the International Conference on Curves and Surfaces, 2010, pp. 711–730.
[13] R. Timofte, V. De, and L. Van Gool, “Anchored neighborhood regression for fast example-based super-resolution,” in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1920–1927.
[14] R. Timofte, V. De Smet, and L. Van Gool, “A+: Adjusted anchored neighborhood regression for fast super-resolution,” in Proceedings of the Asian Conference on Computer Vision, 2014, pp. 111–126.
[15] W. Yang, Y. Tian, F. Zhou, Q. Liao, H. Chen, and C. Zheng, “Consistent coding scheme for single-image super-resolution via independent dictionaries,” IEEE Transactions on Multimedia, vol. 18, no. 3, pp. 313–325, 2016.
[16] S. Schulte, C. Leistner, and H. Bischof, “Fast and accurate image upsampling with super-resolution forests,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3791–3799.
[17] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, “Dual path networks,” in Proceedings of the Advances in Neural Information Processing Systems, 2017, pp. 4470–4478.
[18] R. Soltani and H. Jiang, “Higher order recurrent neural networks,” arXiv preprint arXiv:1605.00064, 2016.
[19] A. Veit, M. J. Wilber, and S. Belongie, “Residual networks behave like ensembles of relatively shallow networks,” in Proceedings of the Advances in Neural Information Processing Systems, 2016, pp. 550–558.
[20] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, “Deep networks with stochastic depth,” in Proceedings of the European Conference on Computer Vision, 2016, pp. 646–661.
[21] W. Yang, J. Feng, J. Yang, F. Zhao, J. Liu, Z. Guo, and S. Yan, “Deep edge guided recurrent residual learning for image super-resolution,” IEEE Transactions on Image Processing, vol. 26, no. 12, pp. 5955–5967, 2017.
[22] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, “Densely connected convolutional networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[23] J. Kim, J. Kwon Lee, and K. Mu Lee, “Deeply-recurvontional network for image super-resolution,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1637–1645.
[24] Y. Tai, J. Yang, X. Liu, and C. Xu, “MemNet: A persistent memory network for image restoration,” in Proceedings of International Conference on Computer Vision, 2017, pp. 4539–4547.
[25] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[26] F. Shi, J. Caballero, F. Huszár, 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1874–1883.
[27] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in Proceedings of the European Conference on Computer Vision, 2014, pp. 184–199.
[28] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, “Deep networks for image super-resolution with sparse prior,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 370–378.
[29] C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network,” in Proceedings of the European Conference on Computer Vision, 2016, pp. 391–407.
[30] A. Veit, M. J. Wilber, and S. Belongie, “Residual networks behave like ensembles of relatively shallow networks,” in Proceedings of the Advances in Neural Information Processing Systems, 2016, pp. 550–558.
[31] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.
[32] J. Kim, J. Kwon Lee, and K. Mu Lee, “Accurate image super-resolution using very deep convolutional networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1646–1654.
[33] X. Mao, C. Shen, and Y.-B. Yang, “Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections,” in Proceedings of the Advances in Neural Information Processing Systems, 2016, pp. 2802–2810.
[34] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[35] Y. Tai, J. Yang, and X. Liu, “Image super-resolution via deep recursive residual network,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3147–3155.
