Pattern Recognition 107 (2020) 107475

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Hierarchical dense recursive network for image super-resolution

Kui Jiang[a], Zhongyuan Wang[a,*], Peng Yi[a], Junjun Jiang[b]

[a] NERCMS, School of Computer Science, Wuhan University, China
[b] School of Computer Science and Technology, Harbin Institute of Technology, China

ARTICLE INFO

Article history:
Received 6 June 2019
Revised 17 March 2020
Accepted 23 May 2020
Available online 26 May 2020

Keywords:
Super-resolution
Hierarchical dense connection
Global feature fusion
Recursive network
Multi-scale

ABSTRACT

Image super-resolution (SR) techniques with deep convolutional network (CNN) have achieved significant improvements compared to previous shallow-learning-based methods. Especially for dense connection based networks, these methods have yielded unprecedented achievements but bring the higher complexity and more parameters. To this end, this paper considers both reconstruction performance and efficiency, and advocates a novel hierarchical dense connection network (HDN) for image SR. First of all, we construct a hierarchical dense residual block (HDB) to promote the feature representation while saving the memory footprint with a hierarchical matrix structure design. In this way, it can provide additional interleaved pathways for information fusion and gradient optimization but with a shallower depth compare to the previous networks. In particular, a group of convolutional layers with small size (1×1) are embedded in HDB, releasing the computational burden and parameters by rescaling the feature dimensions. Furthermore, HDBs are connected to each other in a sharing manner, thereby allowing the network to fuse the features in different stages. At the final, the multi-scale features from these HDBs are integrated into global fusion module (GFM) for a global fusion and representation, and then the final profile-enriched residual map is obtained by realigning and sub-pixel upsampling the fusion maps. Extensive experimental results on benchmark datasets and really degraded images show that our model outperforms the state-of-the-art methods in terms of quantitative indicators and realistic visual effects, as well as enjoys a fast and accurate reconstruction.

1. Introduction

Recovering the corresponding high-resolution (HR) image from the low-resolution (LR) counterpart is an important issue in image processing and computer vision tasks. Because of the high practical values in many fields, such as high-definition television (HDTV), video surveillance [1], satellite images [2,3], games and medical imaging [4], image super-resolution [5] has increasingly been studied and attracted great research interests in the past ten years.

In order to learn the map relations between LR and HR images and obtain a stable and accurate SR result, a variety of shallow-learning based algorithms have been constantly proposed so far, including dictionary learning [6,7], local linear regression [8], and random forest [9]. However, due to the over-reliance on the datasets and additional prior information, there are great limitations for these methods to meet the practical demands.

In recent years, CNNs [10,11] have drawn widespread attention because of its superior performance in image processing and analysis tasks. Especially for single-image super-resolution (SISR), CNN and its variants [12,13] are proposed to promote the reconstruction performance. Especially for the skip- or dense-connection based networks [14,15], these methods can aggregate the information from the preceding layers, thereby allowing a further selective fusion of features. For example, Tong et al. [16] introduced dense-connection manner into image SR and proposed SRDenseNet. In [16], the feature maps obtained from each dense block are propagated into the deconvolution layers to reconstruct SR images, providing an effective way to combine the low- and high-level features. Subsequently, Zhang et al. [17] came up with a residual dense block (RDB) to extract abundant local features, which allows direct connections from the state of preceding RDB to all the layers of current RDB, leading to favorable performance for image SR. These methods involved dense connection have achieved significant improvements of reconstruction quality, though they need to construct an extremely deep framework to provide enough linking pathways to maintain the modeling capabilities with the one-dimensional structure, thereby bringing the unacceptable calculation and memory cost. A natural question arises whether a more effective framework can provide more linking notes and pathways while with less parameters and shallower depth, and whether such

* Corresponding author.
E-mail address: wzy_hope@163.com (Z. Wang).

https://doi.org/10.1016/j.patcog.2020.107475
0031-3203/

framework can promote the understanding and representation of image contents?

It is a consensus that increasing the layer depth to obtain more link paths will lead to other problems, such as information loss, vanishing gradients, structural redundancy, and computational burden. To promote the reconstruction performance, we take both the effectiveness and efficiency of the feature representation into consideration. More specifically, we remain the dense connections within a residual block, but turn the one-dimensional structure into a hierarchical matrix structure by interleaved diagonal links to construct the hierarchical dense block (HDB). In HDB, the feature maps of different hierarchical layers are fused via small size convolutions (1×1) to represent the local feature through substantial vertical and diagonal connections. In this way, it allows the network to provide more opportunities for feature extraction and fusion with abundant linking pathways while alleviating the memory and calculation burden, contributing to the fusion and reuse of the features. Moreover, we propose a novel hierarchical dense recursive network (HDRN) for image SR reconstruction. It fully exploits the coarse-and-fine features all-throughout the network with the recurrent HDBs and the global transition layer. More specifically, the feature maps extracted from different HDBs are globally integrated and fused through a global fusion module (GFM) to exploit the coarse-and-fine features. With a sub-pixel upsampling operation, the features are then transformed into HR space with a rearrangement operation. Overall, we have formed a hierarchical dense recursive network (HDRN) for image SR, which is more effective for reconstructing the textural details, but with fewer convolutional layers, shallower depth, and less calculation cost at the same number of linking pathways.

The main contributions of this study are summarized as follows:

1. We advocate a novel hierarchical dense recursive network for image SR reconstruction. It fully exploits the coarse-and-fine features all-throughout the network with the recurrent HDBs and the global transition layer.
2. We propose a hierarchical residual block (HDB) for feature extraction and expression in an efficient information- and parameter-sharing fashion. The rich interleaved diagonal linking pathways in one HDB enable specific feature mapping for multi-level image components. As a consequence, it is more effective and efficient for reconstructing the textural details.
3. We conduct a global fusion module (GFM) to exploit the coarse-and-fine features from different HDBs to reconstruct realistic results faithful to the ground truth.

2. Related work

Deep CNNs [18–20] have been widely applied to the image processing and analysis tasks. A comprehensive review is beyond the scope of this work and we discuss the most related ones in this section.

2.1. Image super-resolution

In recent years, a variety of SR techniques for the restoration of HR images have been proposed [21,22]. For example, SRCNN [23], as a pioneering deep-learning based framework, learns a mapping relationship from LR to HR with a three-layer CNN structure by an end-to-end manner. After that, Shi et al. [24] developed a contextualized multi-task learning framework to address the SR problem. Then Sheng et al. [25] presented a deep Laplacian pyramid super-resolution network to reconstruct the sub-band residuals of HR images at multiple pyramid levels, in which a weight sharing mechanism is implemented in the same structure, thus greatly reducing the parameters and promoting the efficiency. More recently, Tong et al. [16] put forward a SISR network using dense skip connections (SRDenseNet), increasing feature maps progressively. Zhang et al. [17] came up with a residual dense block (RDB) to extract abundant local features, which allows direct connections from the state of preceding RDB to all layers of current RDB, leading to favorable performance for image SR.

2.2. Dense connection

The skip connection was first introduced for image SR by Kim et al. [13], who constructed a very deep CNN for accurate image SR (VDSR). Instead of learning the actual pixel values, the authors used the residual learning paradigm to predict the differences between the HR and the bicubic interpolated image, which makes the feature maps very sparse, enabling easy training and convergence. More recently, densely connected [26] strategy [16,27] is also adopted for image reconstruction by connecting the feature maps of the current layer to every subsequent layer in a feed-forward manner. For example, Tai et al. [28] integrated recursive learning and skip connections for image restoration tasks. They proposed long-term dense connections to recover much more high-frequency information. In [29], the authors exploited densely connected convolutional layers to maintain coarse and fine feature classification through a multi-scale network.

3. Proposed method

In this section, we first introduce the architecture of hierarchical dense recursive network (HDRN) and the corresponding objective function in detail. And then we individually illustrate the design details of two key modules in our proposed SR model, including hierarchical dense block (HDB) and the global fusion module (GFM).

3.1. Model optimization

As shown in Fig. 1, our proposed HDRN is a deep recursive neural network that can be roughly partitioned into three parts, namely initial feature extraction, local residual learning, and global feature fusion and reconstruction. In this study, I_{LR}, I_{SR}, and I_{HR} are considered as the input, SR output, and HR label, respectively. Given an LR RGB image, the aim of the first part is to transmit the image into feature space to obtain the initial features F_1 through only one convolutional layer. This procedure can be formulated as follows:

$$F_1 = H(I_{LR}), \tag{1}$$

where $H(\cdot)$ denotes the convolution operation, F_1 represents the initial feature maps extracted through the initial convolutional layers. After that, the shallow features F_1 are then transmitted into HDB based for deep feature extraction. In particular, the information extracted from the current HDB can be shared by the subsequent HDBs with substantial skip connections. Mathematically, the computation can be described in the following recurrent form:

$$B_1 = H_{HDB,1}(F_1), \tag{2}$$

$$B_i = H_{HDB,i}(B_{i-1}) + B_{i-1} \tag{3}$$

where $H_{HDB,1}$ and $H_{HDB,i}$ denote the convolution operation in the first and the ith HDB respectively. B_1 and B_i refer to the output of the first and the ith residual block. In HDRN, the current HDB can share the information from all the previous HDBs and serves as the input of the subsequent HDBs for a deep extraction. Thus the effective sharing mechanism between HDBs allows the network integrate multi-scale coarse-and-fine features in different stages, which

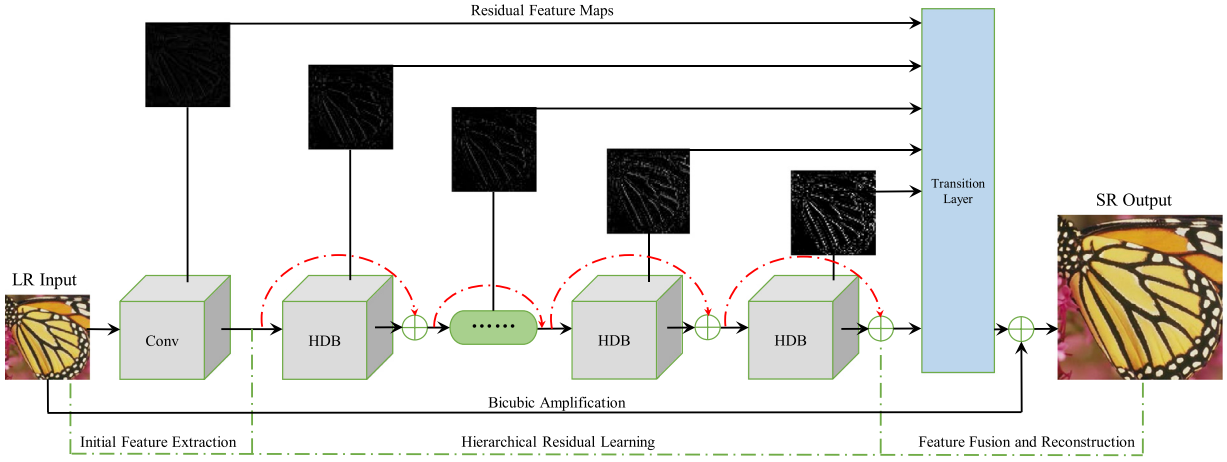


Fig. 1. The hierarchical dense recursive network. In the top, we show the details of a channel of the extracted feature maps by HDB, which show coarse-to-fine characteristics with the network propagation. Red dotted lines denote the skip connections between HDBs.

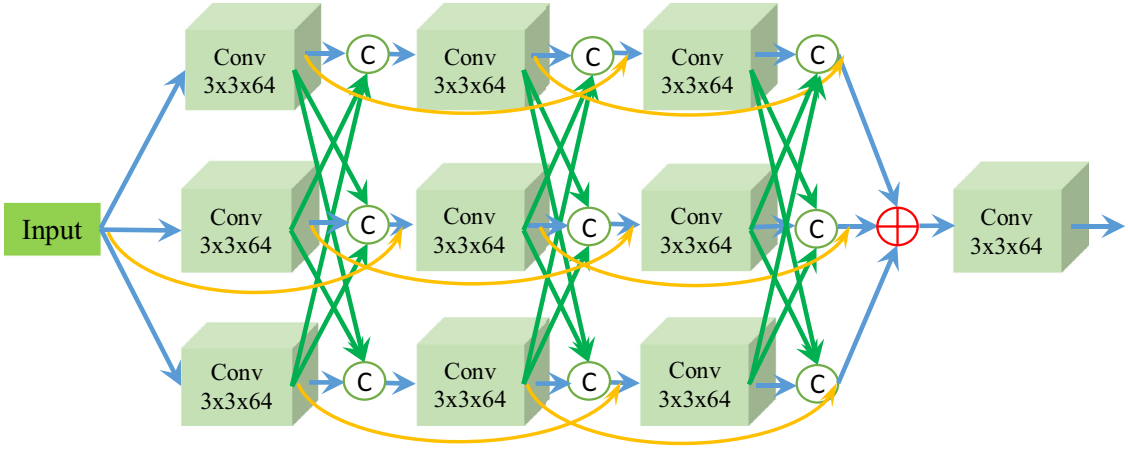


Fig. 2. The outline of the proposed HDB. "C" denotes the concatenation and transition operation with a 1×1 convolution layer. $\oplus$ refers to the summation.

can enforce information propagation and lead to fine feature representation.

In the final, the feature maps extracted by HDBs in different stages are concatenated and passed into the global fusion module (GFM) for further fusion. The features corresponding to the same spatial location at different levels are assembled together to preserve the most relevant components to HR image. This procedure can be described as follows:

$$F_{MS} = H_{GFM}([B_1, \ldots, B_g]), \tag{4}$$

where F_{MS} refers to the fused multi-scale features through GFM. $H_{GFM}(\cdot)$ is the function in GFM. $[B_1, \ldots, B_g]$ denotes the concatenation of feature maps produced by HDBs. And then, a sub-pixel realignment operation is used to match a certain point to obtain the structural information complementary residual image. In formal, it is expressed as follows:

$$I_{SR} = PS(F_{MS}) + I_B, \tag{5}$$

where I_B refers to the Bicubic interpolation image. $PS(\cdot)$ represents the reconstruction operation performing a sub-pixel rearranged alignment.

In practice, all the intermediate feature maps are simultaneously supervised during training by minimising the distance between I_{SR} and the ground truth I_{HR}. Generally, the constraint is imposed by minimizing the mean squared error (MSE) or maximizing the peak signal to noise ratio (PSNR). However, they lack the capability to capture perceptually relevant components. Enlighten by Lai et al. [25], we introduce Charbonnier penalty function to constrain the deviation of the prediction from the ground truth. Charbonnier loss is based on the l_1-norm and is favourable to realistic details faithful to the ground truth, expressed as

$$Loss(I_{SR}, I_{HR}, \theta) = \arg\min_{\theta} \sum \rho(I_{HR} - H_{HDRN}(I_{LR}, \theta)), \tag{6}$$

where θ denotes a set of model parameters to be optimized and $\rho(x) = \sqrt{x^2 + \varepsilon^2}$ represents the Charbonnier penalty function (a differentiable variant of l_1-norm). I_{SR} and I_{HR} refer to the predicted HR image and the ground truth.

3.2. Hierarchical dense block

It is acknowledged that rich dense connections and information sharing can promote feature expression [16,27]. Therefore, we design a hierarchical dense block (HDB) to infer and estimate local features. As shown in Fig. 2, the proposed HDB is composed of a plenty of convolutional layers stacked in a two-dimensional matrix form, connected to each other by horizontal or diagonal linking pathways to construct an interleaved connectivity structure. In particular, for each intersection node in matrix block, an individual convolutional layer with small size 1×1 is followed to integrate and select features extracted by the former column, and then combines the input of the former node for a deep extraction in next node. This paradigm can provide diverse short and long linking pathways and effective feature-sharing capability among the lay-

Table 1
The comparison of characteristics on different structures.

Framework	Conv Num	Channel	Para.	Links	Depth
Direct [23]	10(3 × 3)	64	**360Kb**	9	10
Skip [13]	10 (3 × 3)	64	360Kb	10	10
Dense [16]	10 (3 × 3)+ 1 (1 × 1)	64	2020Kb	45	11
HDB (Ours)	10 (3 × 3)+ 9 (1 × 1)	64	576Kb	**120**	**7**

ers, thereby allowing HDB fuse feature maps extracted from parallel multiple nodes, leading to a rich feature representation.

More specifically, *M, N* denote the horizontal and the vertical depth in an HDB block, respectively. Then there are total $M \times N$ intersection nodes containing $M \times N$ convolution layers (3×3), and additional $M \times N$ convolution layers (1×1). Let $F3_{x,y}$ be the output of the $[x, y]$th node and in HDB (x and y denote the indexes of raw and column respectively.). $F1_{x,y}$ refers to the output of transition layer after $F3_{x,y}$. In particular, 3 and 1 denote the convolutional kernel size. The computing procedures of the $[x, y]$th node can be formulated as follows:

$$\begin{aligned} F1_{x-1,y} &= H1_{x-1,y}(F3_{x-1,1}, \ldots, F3_{x-1,y}, \ldots, F3_{x-1,N}), \\ F3_{x,y} &= H3_{x,y}(F1_{x-1,y} + F1_{x-2,y}), \end{aligned} \tag{7}$$

In Eq. (7), $H1_{x-1,y}(\cdot)$ refers to the concatenation operation to integrate the feature maps from parallel multiple nodes in HDB, which is illustrated as an effective way to reduce the channel size and adaptively thin out the features maps, specially leading to an extremely cutting for the parameters and computational cost. Then the features are passed into the subsequent nodes for a deep extraction and representation.

We further illustrate the fact that our proposed HDB enjoys much richer connectivity paths, from simple to complex by a simple numerical comparison. As a specific example shown in Table 1, HDB offers the maximum number of link paths 120 for 10 identical convolution layers (3×3) while with much less parameters when compared with dense network [16]. Since HDB provides more candidate paths for feature extraction, uneven content complexity components (flat areas, textures, and edges) in the image will greatly benefit. At the same time, due to the effective hierarchical structure of HDB, the network requires less depth at the same number of nodes, which not only facilitates back-propagation iterative optimization, but also thins out the SR model. In particular, the sharing mechanism of parameters between HDBs further releases the computational and memory burden.

3.3. Feature fusion and reconstruction

The contents in an image are composed of features in different stages, ranging from low level to high level. Traditional neural networks learn coarse features in early layers and fine ones in later layers. However, in previous works [13,28], only the outputs of the final layer are utilized for reconstruction, thus they cannot make full utilization of information, hence hindering the representational power of CNNs. To address this issue, HDRN proposes to integrate the hierarchical features from different HDBs for a better reconstruction. Thus, in this work, we construct a global fusion module (GFM) to exploit the complementary information at multiple scales. The features at different levels in LR space are selected and fused by a group of convolution layers, formally as

$$P = H_{GFM}(C([B_1, \ldots, B_g])). \tag{8}$$

In Eq. (8), $C(\cdot)$ denotes the concatenation operation adopted to collect the feature maps B_i from HDB, $H_{GFM}(\cdot)$ refers to the complementary fusion by a group of convolution layers.

Following GFM, an upsampling operation is used to project the processed features P from LR space into HR space. The feature maps in different channels are rearranged from an $H \times W \times C \cdot r^2$ tensor into an $rH \times rW \times C$ tensor by a shuffling operator *PS*. Thus information in the same location for different channels is reassembled and placed into the corresponding location of HR space by a realignment. Therefore, we can obtain the reconstructed HR image from Cr^2 feature maps. Mathematically, the entire process can be formulated as

$$PS(T)_{x,y,c} = T_{\llcorner x/r \lrcorner, \llcorner y/r \lrcorner}(\mathrm{mod}(x, r), \mathrm{mod}(y, r)). \tag{9}$$

In Eq. (9), T indicates the fused information with the size of $W \times H \times Cr^2$, (x, y) denotes the output pixel coordinate in HR space, $(x/r, y/r)$ represents the pixel area of $r \times r$ in sub-pixel space, and $(\mathrm{mod}(x, r), \mathrm{mod}(y, r))$ refers to the pixel coordinate in LR space. Cr^2 channels of each pixel in the same location in LR space are rearranged into a region of $1r \times 1r \times C$, which corresponds to a subblock in HR image, and the feature image is rearranged into an HR image of $rW \times rH \times C$. The sub-pixel upsampling strategy not only achieves spatial resolution enhancement, but also eliminates the need to perform most SR operations in HR space, greatly reducing memory consumption and computational cost.

4. Experimental results

In this section, we first describe the experimental settings, including data collection and model parameters. Then, we conduct the comparison experiments with dense-based network to verify the effectiveness and efficiency of the proposed hierarchical dense block HDB. Subsequently, the contributions of different internal structures in our proposed method are investigated in comparison experiments. Furthermore, the comparison results with the state-of-the-art techniques (including A+ [8], SelfExSR [30], SRCNN [23], VDSR [13], DRCN [31], LapSRN [25], DRRN [19], SRDenseNet [16], and MemNet [28]) are given along with a comprehensive analysis.

4.1. Data collection

For general image SR, a large quantity of public training and evaluating datasets, such as DIV2K [32], BSD500 [30] and Yang291 [33], are used for comparison. In this work, we use the high quality 2K image dataset DIV2K for training, which is composed of 800 training images, 100 validation images, and 100 test images. In addition, the comparison models are evaluated with these public available benchmark datasets, including Set5 [34], Set14 [35], BSD100 [36], Urban100 [30] and Manga109 [37]. Among them, Set5, Set14, and BSDS100 consist of natural scenes; Urban100 contains urban scenes; and Manga109 is a dataset of Japanese manga. Similar to many previous works [28,31], we also adopt commonly used peak signal to noise ratio (PSNR) and structural similarity (SSIM) [38]) to quantitatively measure the image quality of the generated results on the *YCbCr* color. In particular, we have evaluated our model on a set of real-world images such as satellite images selected from the *Jilin*-1 satellite imagery and *Kaggle Open Source Datasets*, which are corrupted by unknown noises and blurring degradations.

4.2. Model parameters and experimental setup

In experiments, the original HR images are cropped with size of 96×96, and then downsized by Bicubic interpolation to generate LR images for training. We augment the training patches by horizontal or vertical flipping and rotating 90°. The learning rate is initialized to 10^{-3} for all layers and halved for every 2500 steps up to 10^{-5}. In our model, each convolution layer contains 64 filters with batch size of 16, followed by the PReLU with its parameter of 0.2. The depth of HDB in our baseline is set to 6, with M and

Table 2
The ablation study of model parameters on *Kaggle Open Source Dataset*.

Framework	Num	Image Size	Scale	Blocks	Parameters	Depth	PSNR/SSIM
DBN	30	720 × 720	4	8	2883Kb	63	31.81/0.887
HDRN (Ours)	30	720 × 720	4	6	**867Kb**	**45**	**32.12/0.893**

N being 3. In our experiments, training consumes approximately 10 h under the previously presented experimental settings (Only one NVIDIA Titan Xp GPU and an Intel I7-8700 CPU).

4.3. Effectiveness of the hierarchical dense connection

We estimate the effectiveness and efficiency of the proposed residual block HDB by designing two SR models. The first one is our baseline with the depth of HDB of 6, as well as $\boldsymbol{M}$ and $\boldsymbol{N}$ being 3 simultaneously. In addition, by replacing HDB with the dense block in SRDenseNet [16], we construct the another SR model (DBN). For the sake of fairness, we keep the same convolutional layers in these two modules. The comparison results of the characteristics (including reconstruction performance, parameters, time consuming, and depth) by two types of connection structures are reported in Table 2. From these results, it is obvious that HDB-based method exhibits the significant superiorities in terms of constructing an effective and efficient SR model. By utilizing HDB and parameter sharing mechanism, our proposed HDRN achieves the better construction performance while with less parameters and shallower depth when compared with dense-based SR model.

4.4. Influence of the parameters **M, N**, and **g**

In a further verification on HDB, we examine the effects of the raw M and the column N within HDB on the image reconstruction quality. We train the comparison models by training on 800 images and iterating for 10^5 steps. We test these models for scale of 4 on evaluation datasets and show the results in Fig. 3. Evidently, HDB with $\mathbf{M} = \mathbf{3}$ and $\mathbf{N} = \mathbf{3}$ shows the best convergent performance. Alternatively, the hierarchical dense structure with $\mathbf{M} = \mathbf{3}$ and $\mathbf{N} = \mathbf{3}$ can fully exploit the structural and content information from LR input. From the results, we can also see that the wider or deeper HDB's internal structure may lead to a slight drop for reconstruction performance, possibly due to structural redundancy of dense blocks.

In light of the observations in previous works [16,27], fine features can be well inferred from a deep CNN framework. Thus, we gradually increase the depth of the network by adding the number $\boldsymbol{g}$ of HDB. We assess the effects of different numbers (denoted by $\boldsymbol{g}$) within one HDB on reconstruction performance. In Fig. 4, we show the training details of HDRN with respect to the numbers of 15 and 33. A significant improvement can be clearly observed, where the model HDRN-33 surpasses HDRN-15 by approximately 0.11dB at the scale of 4. In justification, a more comprehensive and complex mapping between LR and HR images can be preserved by more hierarchical dense connections in the network, which produces a reasonable explanation for the phenomena in Fig. 4.

4.5. Comparison results on the benchmark datasets

We compare our best model HDRN (**B=33, M=3, N=3**) with other SISR methods, including bicubic, SRCNN [23], VDSR [13], DRCN [31], LapSRN [25], DRRN [19], and SRDenseNet [16], by the scaling factors of × 2, × 3, × 4, and × 8. The implementations of these anchor methods have been released online and can thus be conducted on the same evaluation datasets for fair comparisons. We select several different but representative scenarios from these test images. The reconstructed SR images are used to compute the PSNR and SSIM against the ground truth. The comparison results on the benchmark datasets are shown in the Fig. 5 and 6. It shows that the basic Bicubic interpolation method cannot

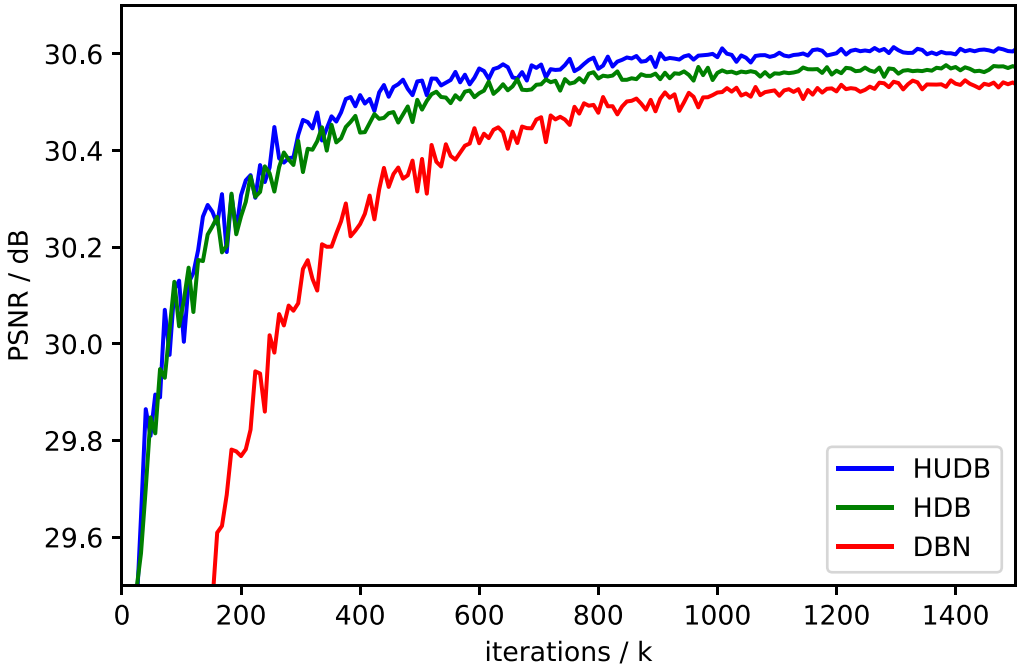


Fig. 3. The convergence curves for HDB in different depths and widths. g denotes the number of the blocks. M and N represent the raw and the column within HDB, respectively.

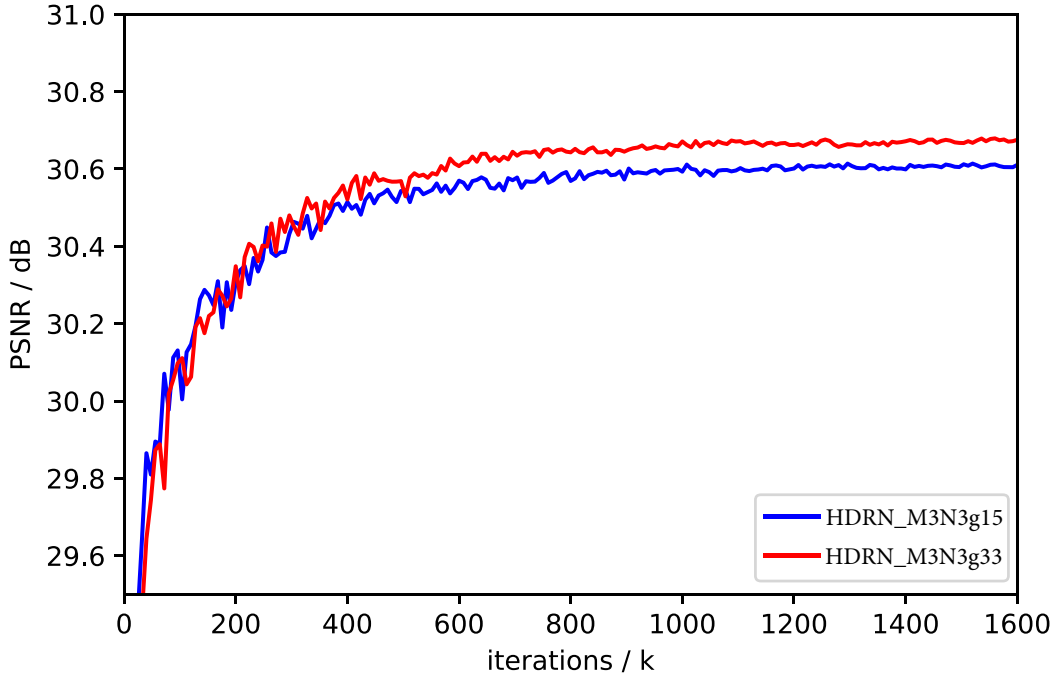


Fig. 4. Performance comparisons of HDRN in different depths. The symbol *g* refers to the number of HDB.

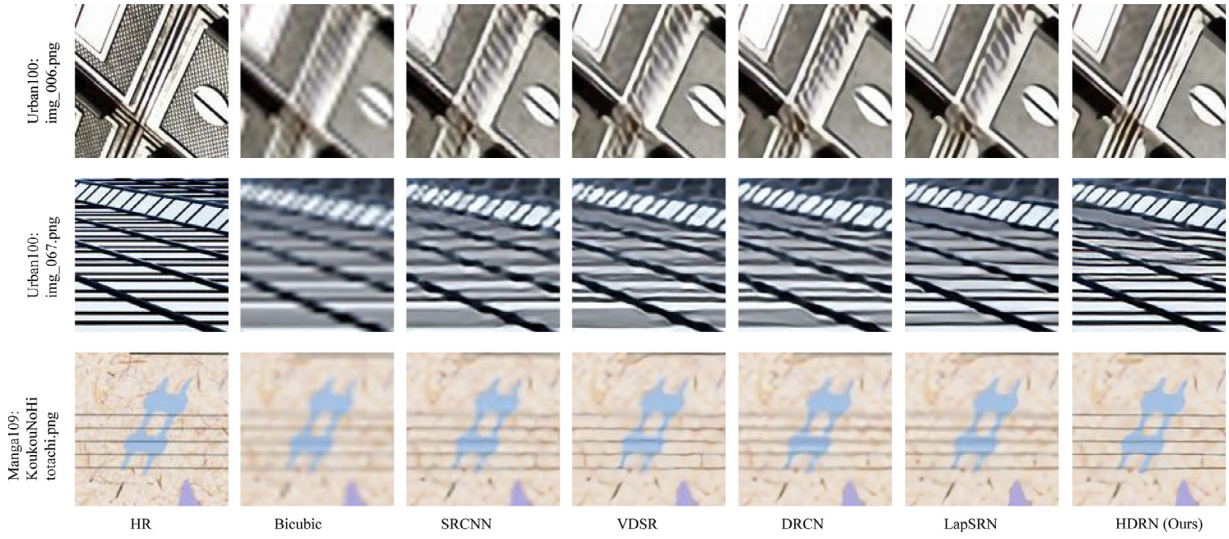


Fig. 5. Qualitative comparisons of our model with other methods by × 4 SR with cropped size of 120 × 120.

produce any extra details. As for the deep learning based technology, SRCNN can infer few authentic details. For VDSR, it can infer some global details, but results in blurry image contours due to their global optimization scheme and poor utilization of features. When compared with DRCN, which can be seen as the first recursive method for image SR, our results are still very competitive and much more realistic. As for LapSRN, the network generates an HR image by reconstructing the sub-band residuals of HR images at multiple pyramid levels, and learns the correspondences between LR and HR through exploring the multi-scale information. However, it fails and still generates blurry results. In addition, HDRN can generate results with sharp and clear details, more faithful to the ground truth, such as the "Butterfly", only our proposed HDRN yields the most accurate and realistic wing textures from visual effects. Moreover, as shown in Fig. 5, only our model restores the clear shape line and texture in the Urban100 [30] and Manga109 [37] datasets. Again, in Fig. 6, only HDRN reconstructs delicate texture details at the large upsampling scale of 8. Therefore, compared with the conventional direct- or skip-connection-based algorithms [13,23,25], the underlying hierarchical dense connection is really favourable to the SR reconstruction.

In addition, Table 3 tabulates the numerical results in terms of PSNR and SSIM. In most cases, our method achieves the best performance. Especially when the up-sampling factor becomes large, our method outperforms all state-of-the-art SR algorithms by a large margin. The superior visual effects and substantial quantitative gains jointly validate the advantages of hierarchical dense deep network in modelling the relations between LR and HR images, especially at large magnifications.

4.6. Super-resolution on the real-world images

Image reconstruction is devoted to restoring the realistic details from low-quality inputs. Especially for real-world scenarios

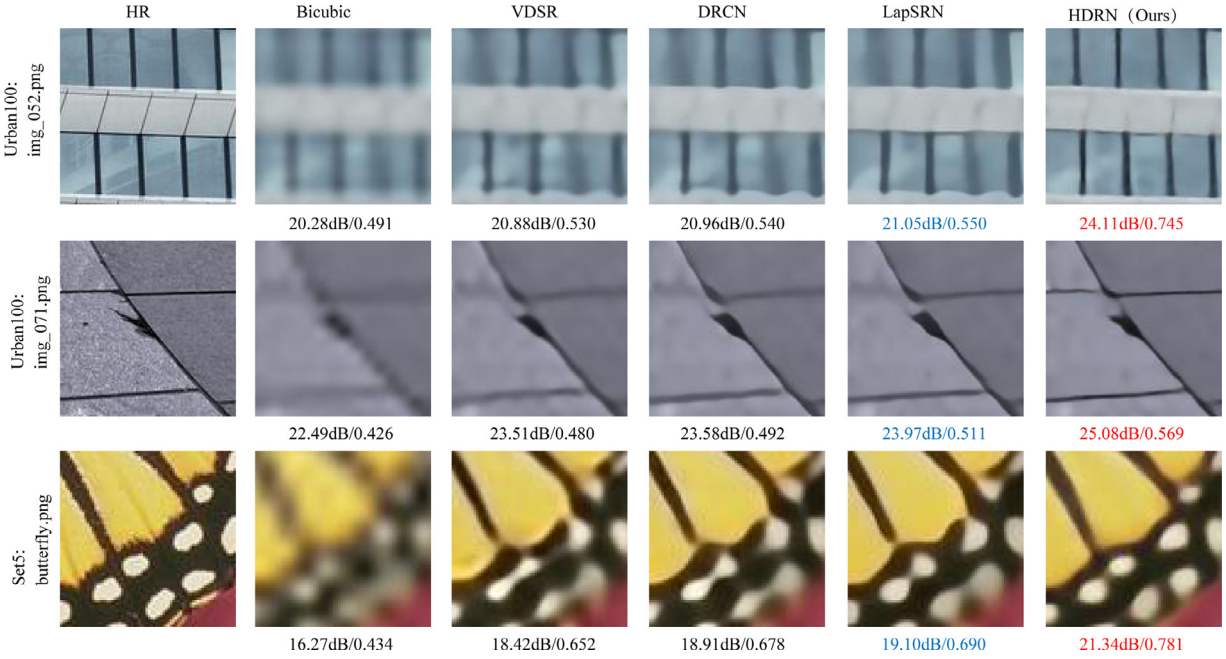


Fig. 6. Qualitative comparisons on Set5 and Urban100 with upsampling scale of 8.

Table 3
Quantitative evaluation of state-of-the-art SR algorithms on five benchmarks for scale factors × 2, × 3, × 4, and × 8. Red indicates the best and blue indicates the second best results.

Algorithms	Dataset Scale	Set5 [34] PSNR/SSIM	Set14 [35] PSNR/SSIM	BSDS100 [36] PSNR/SSIM	Urban100 [30] PSNR/SSIM	Manga109 [37] PSNR/SSIM
Bicubic	2	33.65/0.930	30.24/0.868	29.56/0.844	26.88/0.841	30.80/0.934
A+ [8]	2	36.54/0.954	32.40/0.906	31.22/0.887	29.23/0.894	35.33/0.967
SRCNN [23]	2	36.65/0.954	32.45/0.906	31.36/0.888	29.50/0.894	35.60/0.966
VDSR [13]	2	37.53/0.958	32.97/0.913	31.90/0.896	30.77/0.914	37.16/0.974
DRCN [31]	2	37.63/0.959	32.98/0.913	31.85/0.894	30.76/0.913	37.57/0.973
LapSRN [25]	2	37.52/0.959	33.08/0.913	31.80/0.895	30.41/0.910	37.27/0.974
DRRN [19]	2	37.74/0.959	33.23/0.913	32.05/0.897	31.23/0.919	37.92/0.976
MemNet [28]	2	37.78/0.959	33.28/0.914	32.08/0.897	31.31/0.919	37.72/0.974
HDRN (Ours)	2	37.75/0.959	33.49/0.915	32.03/0.898	31.87/0.925	38.07/0.977
Bicubic	3	30.39/0.868	27.55/0.774	27.21/0.738	24.46/0.735	26.95/0.855
A+ [8]	3	32.58/0.908	29.13/0.818	28.29/0.7835	26.03/0.797	-/-
SRCNN [23]	3	32.75/0.909	29.30/0.821	28.41/0.786	26.24/0.799	30.48/0.911
VDSR [13]	3	33.66/0.921	29.77/0.831	28.82/0.797	27.14/0.828	-/-
DRCN [31]	3	33.82/0.922	29.76/0.831	28.80/0.796	27.15/0.827	-/-
LapSRN [25]	3	33.82/0.922	29.79/0.832	28.82/0.797	27.07/0.827	32.19/0.933
DRRN [19]	3	34.03/0.923	29.96/0.835	28.95/0.800	27.53/0.837	32.42/0.935
MemNet [28]	3	34.09/0.924	30.00/0.835	28.96/0.800	27.56/0.837	32.51/0.936
HDRN (Ours)	3	34.24/0.924	30.23/0.840	28.96/0.804	27.93/0.849	33.17/0.942
Bicubic	4	28.42/0.810	26.10/0.702	25.96/0.667	23.15/0.659	24.92/0.789
A+ [8]	4	30.30/0.859	27.43/0.752	26.82/0.710	24.34/0.720	27.02/0.850
SRCNN [23]	4	30.49/0.862	27.61/0.754	26.91/0.710	24.52/0.722	27.58/0.855
VDSR [13]	4	31.35/0.882	28.03/0.770	27.29/0.726	25.18/0.753	28.82/0.886
DRCN [31]	4	31.53/0.884	28.04/0.770	27.24/0.724	25.14/0.752	28.97/0.886
LapSRN [25]	4	31.54/0.885	28.19/0.772	27.32/0.728	25.21/0.755	29.09/0.889
DRRN [19]	4	31.68/0.888	28.21/0.772	27.38/0.728	25.44/0.764	29.46/0.896
SRDenseNet [16]	4	32.02/0.894	28.50/0.778	27.53/0.733	26.05/0.782	-/-
MemNet [28]	4	31.74/0.889	28.26/0.772	27.40/0.728	25.50/0.763	29.42/0.894
HDRN (Ours)	4	32.23/0.896	28.58/0.781	27.53/0.737	26.09/0.787	30.43/0.908
Bicubic	8	24.39/0.657	23.19/0.568	23.67/0.547	20.74/0.516	21.47/0.647
A+ [8]	8	25.52/0.692	23.98/0.597	24.20/0.568	21.37/0.545	22.39/0.680
SRCNN [23]	8	25.33/0.689	23.85/0.593	24.13/0.565	21.29/0.543	22.37/0.682
VDSR [13]	8	25.72/0.711	24.21/0.609	24.37/0.576	21.54/0.560	22.83/0.707
DRCN [31]	8	25.93/0.723	24.25/0.614	24.49/0.582	21.71/0.571	23.20/0.724
LapSRN [25]	8	26.14/0.738	24.44/0.623	24.54/0.586	21.81/0.582	23.39/0.735
DRRN [19]	8	26.18/0.738	24.42/0.622	24.59/0.587	21.88/0.583	23.60/0.742
HDRN (Ours)	8	27.09/0.768	24.76/0.638	24.68/0.598	22.36/0.616	24.40/0.778

Table 4
The comparison results for AG, NIQE, and PI by scale factor of 4 on *Kaggle Open Source Datasets*.

Num	Metrics	Bicubic	SRCNN [23]	VDSR [13]	LapSRN [25]	HDRN (Ours)
30	AG	3.040	3.934	3.978	3.926	**4.011**
30	NIQE	8.213	6.240	6.006	6.185	**5.857**
30	PI	7.021	5.951	5.918	5.836	**5.825**

with unknown and complex degradations, whether previous works [23,31] are equally applicable to real-world SR tasks remains to be investigated. Because of complex motion blur, unknown degradation process and noise, emerging video satellite images reconstruction are therefore facing challenging. For a better and comprehensive evaluation of our proposed HDRN as well as other competing methods, we directly employ *Jilin-1* satellite imagery and *Kaggle Open Source Datasets* (covering agriculture, airport, buildings, warships, forest, freeway, parking lot, storage tanks, and harbor) as the LR inputs to conduct validations. Besides, since it lacks of the reference images, we additionally introduce Average gradient (AG) [39], Naturalness image quality evaluator (NIQE) [40], and Perceptual index (PI) [41] to perform evaluation. These indicators can reasonably assess image clarity since they sensitively reflect content sharpness, detail contrast and texture diversity. In contrast to PSNR and SSIM, PI particularly evaluates the image quality in a perceptual-quality aware manner, and is not based solely on distortion measurement. These evaluation indexes have a high consistency with the subjective quality and can effectively reflect the visual quality of images without reference. In particular, the larger scores of AG and the smaller values of NIQE/PI indicate better perceptual quality and clearer content.

In this experiment, the test satellite images are cropped to a specific size used as the LR input. After performing SR process, comparing with these state-of-the-art methods, such as SRCNN [23], VDSR [13], and LapSRN [25], we evaluate HDRN on visual performance and objective metrics (AG, NIQE, and PI). As shown in Fig. 7, two representative scenes ("urban" and "suburb") selected from *Kaggle Open Source Datasets* are used for a visual comparison. As a result, only our model has recovered the clear outline of the car and the realistic roof lines in the "urban" scene. In another scene, most of compared methods produce noticeable artifacts and blurred road signs, while our proposed HDRN yields better results with fewer jagged lines and ringing artifacts. In addition, the comparison results for the objective indicators are tabulated in Table 4. Again, our proposed HDRN method is highly competitive, achieving the highest AG and the lowest average values of NIQE/PI on Kaggle Open Source Datasets. These results show that our model is more robust than the comparison methods in coping with the SR tasks on real-world degradation scenes.

An additional experiment on *Jilin-1* satellite imagery is further conducted to illustrate the effectiveness and applicability of the proposed method. Compared with the first dataset *Kaggle Open Source Dataset*, the test images obtained from *Jilin-1* demonstrate lower quality (small ground objects and weak textures) but more realistic satellite imagery characteristics. Unlike the images in the training dataset, the test images take completely different imaging conditions, including ultra-high imaging distance, atmospheric scattering, relative motion between satellite and moving ground targets, and compression distortion. These complex imaging conditions place high demands on SR networks.

As shown in Fig. 8 (where two selected regions are specifically enlarged and displayed in the upper left and upper right corners for the convenience of inspection), most of the comparison methods produce noticeable artifacts and blurred edges, while HDRN still recovers sharp and clear edges, e.g., the outline of the build-

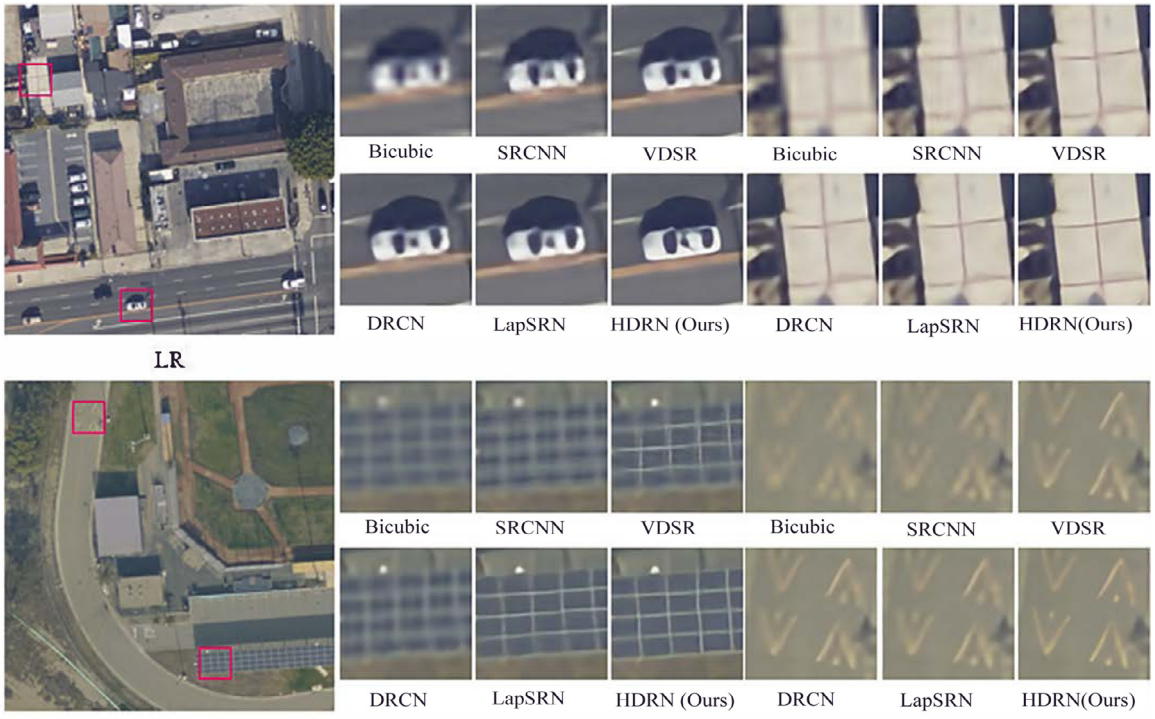


Fig. 7. The comparisons of our model with other methods on real-world satellite images with cropped size of 72 × 72 (by × 4).

Fig. 8. The reconstruction results on *Jilin-1* video satellite imagery by the scale of 4. We use a representative scenario (*i.e.*, city suburb) for comparisons. Two enlarged local regions are displayed in the upper left and upper right corners.

Table 5
Average time consumption in different test datasets with the scale of 4 on CPU I7-8700.

Datasets		Set5		Set14		BSDS100		Urban100	
Algorithms	Scale	Max size	time (s)	Max size	time (s)	Max size	time (s)	Max size	time (s)
VDSR [13]	4	512 × 512	2.20	720 × 516	3.46	321 × 481	1.45	1024 × 1024	15.51
LapSRN [25]	4	512 × 512	4.78	720 × 516	6.29	321 × 481	3.66	1024 × 1024	15.86
HDRN (Ours)	4	512 × 512	1.66	720 × 516	2.14	321 × 481	1.17	1024 × 1024	3.38

ings and roads. The experiment verifies the practicality and stability of HDRN on low-quality real video satellite imagery.

4.7. Comparison results on time complexity

Further, we have provided a comparison on model's efficient in terms of time complexity on different datasets (taking $\times$ 4 as an instance), as shown in Table 5. Comparably, our baseline (with $\mathbf{M} = \mathbf{3}$, $\mathbf{N} = \mathbf{3}$ and $\mathbf{g} = \mathbf{6}$) takes the least time consumption, but enables acceptable parameters. Especially for the large-size images, we can obviously see that HDRN shows a more obvious speed advantages when compared with VDSR [13] and LapSRN [25], extremely promoting the SR efficient. Therefore, HDRN is very suitable for batch-oriented video and image SR tasks in large resolution.

5. Conclusions

In this work, we have proposed an effective and efficient feature representation module, termed hierarchical dense block (HDB), and construct a practical image reconstruction algorithm, named hierarchical dense connection recursive network (HDRN). HDRN can effectively establish realistic mapping relationships between the LR and HR image with abundant interleaved diagonal linking pathways in HDB, which promotes the information interaction and representation. Additionally, HDB is elaborately designed to reduce the computational cost by embedding a large amount of small size convolutions(1×1) to rescale the feature channels. Moreover, we fully exploit multi-scale feature representation produced by HDBs at different stages, and further construct a global fusion module (GFM) to fuse all the complementary feature maps and reconstruct the HR residual image by realignment and sub-pixel upsampling operation. Extensive experiments on benchmark datasets demonstrate that the proposed approach substantially outperforms the state-of-the-art methods in terms of the quantitative metrics (PSNR and SSIM), the visual quality, as well as the model efficiency. Besides, under unknown and real-world degradations (for the real-world test samples, such as *Jilin-1* satellite imagery and *Kaggle Open Source Datasets*), the comparative advantages of our model become more appealing, surpassing these competing methods by a large margin in terms of reference-free indicators (AG, NIQE, and PI). In overall, an ambidextrous design of HDB not only

inherits the superiority of dense connection but also leads to a compact and practical network. It allows the network to infer and estimate the local feature maps through substantial vertical and diagonal connections between different hierarchical layers, thereby providing the opportunities for feature extraction and fusion with rich linking pathways. In particular, these strategies are reasonable to be applied into other computer vision tasks (such as image denoising, deraining, and inpainting), and serve as a basic component for designing networks.

Declaration of Competing Interest

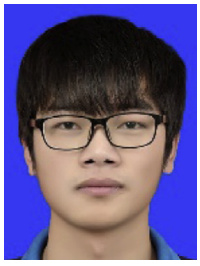
The authors declare that they do not have any financial or non-financial conflict of interests

Acknowledgement

This work is supported by National Key R&D Project (2016YFE0202300) and National Natural Science Foundation of China (U1903214, 61671332, U1736206, 41771452, 41771454, 61971165), and Hubei Province Technological Innovation Major Project (2019AAA049, 2018CFA024).

References

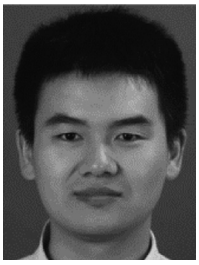
[1] J. Meng, A. Wu, W.-S. Zheng, Deep asymmetric video-based person re-identification, Pattern Recognit. 93 (2019) 430–441.
[2] K. Nogueira, O.A. Penatti, J.A. dos Santos, Towards better exploiting convolutional neural networks for remote sensing scene classification, Pattern Recognit. 61 (2017) 539–556.
[3] Y. Li, W. Xie, H. Li, Hyperspectral image reconstruction by deep convolutional neural network for classification, Pattern Recognit. 63 (2017) 371–383.
[4] V.R.V. Kumar, A. Vidya, M. Sharumathy, R. Kanizohi, Super resolution enhancement of medical image using quaternion wavelet transform with SVD, in: 2017 Fourth International Conference on ICSCN, 2017, pp. 1–7.
[5] X. Hu, P. Ma, Z. Mai, S. Peng, Z. Yang, L. Wang, Face hallucination from low quality images using definition-scalable inference, Pattern Recognit. 94 (2019) 110–121.
[6] J. Liu, W. Yang, X. Zhang, Z. Guo, Retrieval compensated group structured sparsity for image super-resolution, IEEE Trans. Multimed. 19 (2) (2017) 302–316.
[7] J. Yang, J. Wright, T. Huang, Y. Ma, Image super-resolution as sparse representation of raw image patches, in: IEEE Conference on CVPR, 2008, pp. 1–8.
[8] R. Timofte, V.D. Smet, L.V. Gool, A+: adjusted anchored neighborhood regression for fast super-resolution, in: 2014 Conference on ACCV, Springer International Publishing, 2015, pp. 111–126.
[9] S. Schulter, C. Leistner, H. Bischof, Fast and accurate image upscaling with super-resolution forests, in: IEEE Conference on CVPR, 2015, pp. 3791–3799.
[10] L. Zhu, Y. Chen, P. Ghamisi, J.A. Benediktsson, Generative adversarial networks for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. (2018) 1–18.
[11] S. Xie, H. Hu, Y. Wu, Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition, Pattern Recognit. 92 (2019) 177–191.
[12] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al., Recent advances in convolutional neural networks, Pattern Recognit. 77 (2018) 354–377.
[13] J. Kim, J.K. Lee, K.M. Lee, Accurate image super-resolution using very deep convolutional networks, in: IEEE Conference on CVPR, 2016, pp. 1646–1654.
[14] K. Jiang, Z. Wang, P. Yi, et al., ATMFN: Adaptive-threshold-based multi-model fusion network for compressed face hallucination[J], IEEE Trans. Multimed. (2019).
[15] Z. Wang, P. Yi, K. Jiang, et al., Multi-memory convolutional neural network for video super-resolution[J], IEEE Trans. Image Process. 28 (5) (2018) 2530–2544.
[16] T. Tong, G. Li, X. Liu, Q. Gao, Image super-resolution using dense skip connections, in: IEEE Conference on ICCV, 2017, pp. 4809–4817.
[17] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: IEEE Conference on CVPR, 2018, pp. 2472–2481.
[18] Z. Wu, C. Shen, A. Van Den Hengel, Wider or deeper: revisiting the resnet model for visual recognition, Pattern Recognit. 90 (2019) 119–133.
[19] Y. Tai, J. Yang, X. Liu, Image super-resolution via deep recursive residual network, in: IEEE Conference on CVPR, 2017, pp. 2790–2798.
[20] B. Lim, S. Son, H. Kim, S. Nah, K.M. Lee, Enhanced deep residual networks for single image super-resolution, in: IEEE Conference on CVPRW, 2017, pp. 1132–1140.
[21] C. Dong, C.L. Chen, X. Tang, Accelerating the super-resolution convolutional neural network, in: 2016 Conference on ECCV, 2016, pp. 391–407.
[22] X. Yang, H. Mei, J. Zhang, K. Xu, B. Yin, Q. Zhang, X. Wei, DRFN: Deep recurrent fusion network for single-image super-resolution with large factors, IEEE Trans. Multimed. 21 (2) (2018) 328–337.
[23] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2016) 295–307.
[24] Y. Shi, K. Wang, C. Chen, L. Xu, L. Lin, Structure-preserving image super-resolution via contextualized multitask learning, IEEE Trans. Multimed. 19 (12) (2017) 2804–2815.
[25] W.S. Lai, J.B. Huang, N. Ahuja, M.H. Yang, Deep laplacian pyramid networks for fast and accurate super-resolution, in: IEEE Conference on CVPR, 2017, pp. 5835–5843.
[26] K. Jiang, Z. Wang, P. Yi, et al., Edge-enhanced GAN for remote sensing image superresolution[J], IEEE Trans. Geosci. Remote Sens. 57 (8) (2019) 5799–5812.
[27] G. Huang, Z. Liu, L. v. d. Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: IEEE Conference on CVPR, 2017, pp. 2261–2269.
[28] Y. Tai, J. Yang, X. Liu, C. Xu, MemNet: a persistent memory network for image restoration, in: IEEE Conference on ICCV, 2017, pp. 4549–4557.
[29] G. Huang, D. Chen, T. Li, F. Wu, L.V.D. Maaten, K.Q. Weinberger, Multi-scale dense networks for resource efficient image classification, in: Conference on ICLR.
[30] J.B. Huang, A. Singh, N. Ahuja, Single image super-resolution from transformed self-exemplars, in: IEEE Conference on CVPR, 2015, pp. 5197–5206.
[31] J. Kim, J.K. Lee, K.M. Lee, Deeply-recursive convolutional network for image super-resolution, in: IEEE Conference on CVPR, 2016, pp. 1637–1645.
[32] R. Timofte, E. Agustsson, L.V. Gool, etc., Ntire 2017 challenge on single image super-resolution: methods and results, in: IEEE Conference on CVPRW, 2017, pp. 1110–1121.
[33] J. Yang, J. Wright, T.S. Huang, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (11) (2010) 2861–2873.
[34] M. Bevilacqua, A. Roumy, C. Guillemot, A. Morel, Low-complexity single-image super-resolution based on nonnegative neighbor embedding, in: BMVC, 2012, pp. 1–10.
[35] R. Zeyde, M. Elad, M. Protter, On single image scale-up using sparse-representations, in: International Conference on Curves and Surfaces, 2012, pp. 711–730.
[36] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: IEEE Conference on ICCV, vol. 2, 2001, pp. 416–423.
[37] S. Ross, D. Munoz, M. Hebert, J.A. Bagnell, Learning message-passing inference machines for structured prediction, in: IEEE Conference on CVPR, 2011, pp. 2737–2744.
[38] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[39] A.A. Chen, X. Chai, B. Chen, R. Bian, Q. Chen, A novel stochastic stratified average gradient method: convergence rate and its complexity, in: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–8.
[40] A. Mittal, R. Soundararajan, A.C. Bovik, Making a “completely blind” image quality analyzer, IEEE Signal Process. Lett. 20 (3) (2013) 209–212.
[41] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, L. Zelnik-Manor, The 2018 PIRM challenge on perceptual image super-resolution, in: Conference on ECCVW, 2018. pp. 0-0



Kui Jiang received the B.S. degree from College of Chemistry and Chemical Engineering, Xinjiang University, Urumchi, China, in 2017. He is currently pursuing the Ph.D. degree under the supervision of Prof. Zhongyuan Wang in the School of Computer, Wuhan University, Wuhan, China. His research interests include image/video processing and computer vision.

Zhongyuan Wang received the Ph.D. degree in communication and information system from Wuhan University, Wuhan, China, in 2008. Dr. Wang is now an associate professor with School of Computer, Wuhan University, Wuhan, China. He is currently directing three projects funded by the National Natural Science Foundation Program of China. His research interests include video compression, image processing, and multimedia communications, etc.

Peng Yi received the B.S. degree in Faculty of Electronic Information and Electrical Engineering from Dalian University of Technology, Dalian, China, in 2017. He is currently working toward the Ph.D. degree under the supervision of Prof. Zhongyuan Wang in the School of Computer, Wuhan University. His research interests include image/video processing and computer vision.



Junjun Jiang received the B.S. degree from the Department of Mathematics, Huaqiao University, Quanzhou, China, in 2009, and the Ph.D. degree from the School of Computer, Wuhan University, Wuhan, China, in 2014. From 2015 to 2018, he was an Associate Professor with China University of Geosciences, Wuhan. He was a Project Researcher with the National Institute of Informatics, Tokyo, Japan, from 2016 to 2018. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. He won the Finalist of the World's FIRST 10K Best Paper Award at ICME 2017, the Best Student Paper Runner-up Award at MMM 2015, the Best Paper Award at IFTC 2018. He received the 2016 China Computer Federation (CCF) Outstanding Doctoral Dissertation Award and 2015 ACM Wuhan Doctoral Dissertation Award. His research interests include image processing and computer vision.