Taylor Neural Network for Real-World Image Super-Resolution

Pengxu Wei, Ziwei Xie, Guanbin Li, and Liang Lin*, *Senior Member, IEEE*

***Abstract*—Due to the difficulty of collecting paired Low-Resolution (LR) and High-Resolution (HR) images, the recent research on single image Super-Resolution (SR) has often been criticized for the data bottleneck of the synthetic image degradation between LRs and HRs. Recently, the emergence of real-world SR datasets, e.g., RealSR and DRealSR, promotes the exploration of Real-World image Super-Resolution (RWSR). RWSR exposes a more practical image degradation, which greatly challenges the learning capacity of deep neural networks to reconstruct high-quality images from low-quality images collected in realistic scenarios. In this paper, we explore Taylor series approximation in prevalent deep neural networks for image reconstruction, and propose a very general Taylor architecture to derive Taylor Neural Networks (TNNs) in a principled manner. Our TNN builds Taylor Modules with Taylor Skip Connections (TSCs) to approximate the feature projection functions, following the spirit of Taylor Series. TSCs introduce the input connected directly with each layer at different layers, to sequentially produces different high-order Taylor maps to attend more image details, and then aggregate the different high-order information from different layers. Only via simple skip connections, TNN is compatible with various existing neural networks to effectively learn high-order components of the input image with little increase of parameters. Furthermore, we have conducted extensive experiments to evaluate our TNNs in different backbones on two RWSR benchmarks, which achieve a superior performance in comparison with existing baseline methods.**

***Index Terms*—Image Super-resolution, Real-World Super-resolution, High-order Information, Taylor Neural Network, Taylor Attention Map, Taylor Skip Connection.**

I. Introduction

IMAGE super-resolution (SR) aims to address an ill-posed problem of image degradation to reconstruct high-quality images from observed low-quality ones [1]–[4]. However, the SR research heavily relies on the synthetic degradation that hand-crafts paired images by down-sampling high-resolution (HR) images into their low-resolution (LR) counterparts. This inevitably suffers from a bottleneck of practical applications in the real world. Accordingly, with the establishment of real-world SR benchmarks, *e.g.*, RealSR and DRealSR, real-world image super-resolution (RWSR) has increasingly attracted a research interest on real heterogeneous image degradation [4]. To address this challenge, this inspires the further SR research on exploring how to reconstruct high-resolution images with more details from their low-resolution counterparts.

P. Wei, G. Li and L. Lin is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, P.R. China. E-mail: {weipx3, liguanbin}@mail.sysu.edu.cn, linliang@ieee.org. Z. Xie is with Tencent, Shenzhen, P.R. Chian.

Corresponding author: Liang Lin.

In the past years, many deep neural networks based SR approaches, *e.g.*, SRCNN [1], EDSR [2], ESRGAN [3], R-CAN [5], IPT [6] and SwinIR [7], have achieved remarkable progress with significant performance improvements over conventional methods. This progress, to some extend, is attributed to the emergence of various sophisticated deep neural networks as backbones, *e.g.*, Residual Network (ResNet) [8], Squeeze-and-Excitation Network (SENet) [9], Densely Convolutional Network (DenseNet) [10], *etc*. Those networks were originally proposed and successfully demonstrated in high-level tasks. However, few works have paid attention to the learning mechanism for low-level vision tasks, especially for (real-world) image super-resolution. Specifically, those networks process input images sequentially with stacked layers and their architectures essentially are limited for explicitly learning the high-order information of images. This would pose an obstacle to learn image contents with high-frequencies that are crucial for image super-resolution. This phenomenon is also demonstrated in [11]: the second derivative of networks with Rectified Linear Unit (ReLU) is zero everywhere, and they are thus incapable of modeling information contained in high-order derivatives of natural signals, which would lead to the sub-optimal results of feature learning and function approximation with deep neural networks. As shown in Fig.1, SRResNet [12] attends to flat regions that are easier to reconstruct than textures at shallow layers, and with the increase of layers, has fewer pixels that are activated, especially for textures. This would not guarantee the effective reconstruction of more image details [4].

To address this issue, this paper proposes a very general neural architecture derived from Taylor series approximation, named Taylor Neural Network (TNN), which is compatible with various existing backbone neural network architectures. Following the spirit of Taylor Series, we build features maps (*i.e.*, Taylor maps) with different order information of images at different layers and aggregate them together to approximate the feature projection function for image reconstruction. To incorporate the higher-order information, SIREN [11] utilizes periodic functions to replace ReLU, which essentially employs the nonlinearity of the activation function. But it is sensitive to the model initialization for training; it also does not sequentially build different high-order information at different layers similar to TNN. It is evidenced in our experiment that SIREN greatly underperforms TNN for image super-resolution. Compared to attention-based models [5], [13], one distinct difference is that attention-based models consider the attentive information independently at each layer, which is not explicitly learning of different high-order attentions. On the contrary, in our Taylor formulation, our TSC is employed to

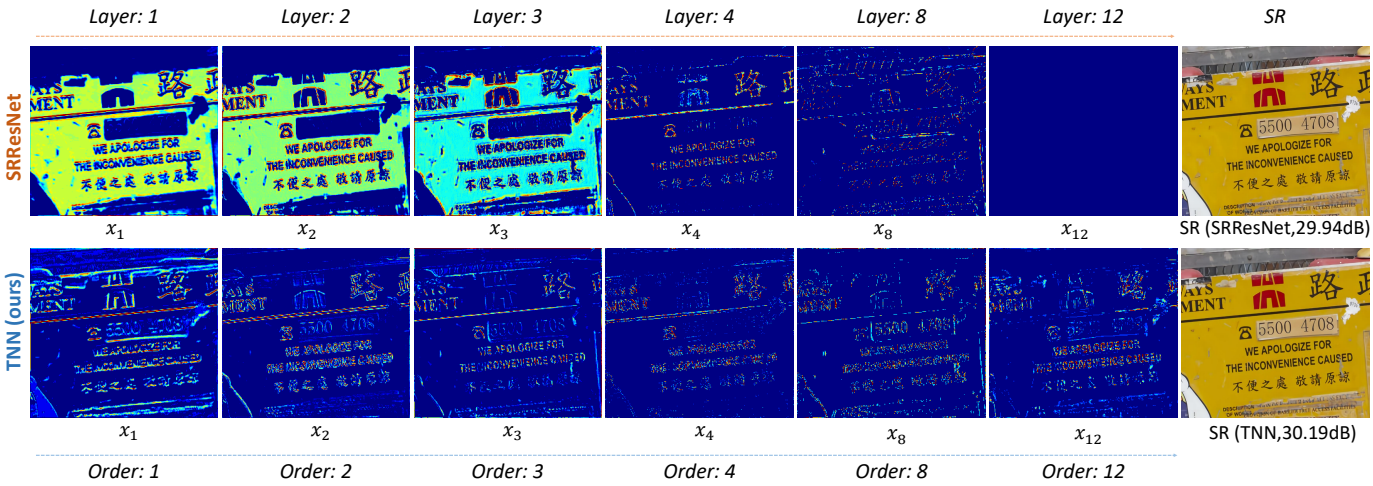


Fig. 1. Visualization of learned feature maps at different layers. At different layers, those maps produced by our TNN, also named as Taylor feature maps, have more pixels with large responses than those of SRResNet, which are prone to flat regions that are easier to reconstruct than textures. In particular, with the increase of layers, fewer pixels are activated in SRResNet. Instead, our TNN, in a Taylor architecture, tends to learn more image details.

connect each layer to build different high-order information, namely, different high-order attention maps. Particularly, our Taylor architecture sequentially builds and further organizes those high-order attention maps for reconstructing more image details.

Our Taylor architecture renews the conventional network formulation for feature learning and provides a principled way in a simple but subtle manner for real-world image super-resolution. This will be insightful to understand the feature learning for low-level vision tasks. The proposed Taylor architecture can be flexibly applied to various neural networks, *e.g.*, residual networks. Furthermore, we have evaluated our TNNs under different SR models on two representative RWSR datasets.

In brief, our contributions are summarized as follows:

- We propose a Taylor Neural Network (TNN) in a Taylor architecture, which introduces the high-order attention of the input following the spirit of Taylor series approximation. Taylor Neural Network is easily applied to various neural networks to learn high-order attention maps for feature learning.
- We propose Taylor Skip Connection to produce high-order attention feature maps at different layers, named Taylor maps, which are beneficial for the reconstruction of image details. TSC introduces the input connected directly with each layer to generate different high-order maps at different layers and aggregates the different order information at different layers.
- We conduct comprehensive experiments on two representative real-world SR datasets. Our experimental results demonstrate that the proposed TNNs in different backbones have a remarkable capacity of learning features for RWSR.

II. Related Work

With the emergence of LeNet [14] as the pioneering work, an unprecedented series of works on Convolutional Neural Networks (CNNs) [15] have achieved a great progress and established milestones of deep learning in computer vision. Residual Network (ResNet) [8] has been proposed to address the degradation issue of the model accuracy with the increase of network depth. Densely Convolutional Network (DenseNet) [10] connects each layer to other layers and combines features from all the preceding layers at each layer. Squeeze-and-Excitation Network (SENet) [9] investigates the channel attentions that model the relationship between channels to adaptively recalibrate channel-wise feature responses. Although they have been successfully demonstrated in high-level tasks, those seminal networks are also utilized in low-level vision tasks, greatly promoting the related research. Especially, recent years have witnessed an evolution of image super-resolution research with widely-explored deep learning, which has significantly improved the performance against traditional methods.

Deep learning based Image Super-Resolution. Dong et al. [1] proposed the first deep learning-based SR method. Subsequently, many CNN-based algorithms [16], [17] are proposed to take low-resolution images as inputs and use upsampling modules together with feature learning in deep neural networks. By using residual learning to ease the training of deeper networks, SRResNet [12] and EDSR [2] successfully build a deep network that further improve the SR performance. Based on traditional Laplacian pyramid algorithms, LapSRN [18] presents a novel multi-level super-resolution model which could generate multi-scale predictions from a pair of LR and HR images. RCAN [5] designs a deep residual channel attention architecture that involves residual in residual (RIR) blocks for constructing the network. To be specific, each residual in the residual module consists of several residual groups (RGs) as well as long skip connections (LSCs), and each RG contains some residual blocks and short skip connections (SSCs). RRDB and ESRGAN [3] utilized residual in residual dense blocks to enhance the learning ability of feature representations. Component Divide-and-Conquer network (CDC) [4] proposes an HGSR basemodel that leverages

stacked Hourglass blocks [19] to encode features.

Recent years have witnessed the prosperity of self-attention mechanism in low-level vision tasks, as it allows networks to further investigate the deep prior of images and thus boosts the global representation ability of networks. Inspired by the emergence of Transformer in NLP, IPT [6] demonstrates a novel self-attention based paradigm to solve multiple low-level vision tasks including denoising, deraining as well as super-resolution. To further reduce the parameters without sacrificing the performance, SwinIR [7] combines the basic swin transformer and residual learning to construct residual swin transformer blocks (RSTB) for deep feature extraction. To further explore the potential of self-attention mechanism in image restoration tasks, Restormer [20] focuses on capturing long-range pixel interactions for allowing the model to achieve higher performance on high-resolution images without suffering from computation bottleneck. And Uformer proposes a hierarchical network following the design of U-Net, which could combine multi-scale feature information for the image reconstruction.

Real-World Image Super-Resolution. Deep neural networks have significantly advanced the progress of image super-resolution, which cast the learning of single image SR models as the supervised learning task with synthesized paired LR-HR images [1], [2], [17]. Nevertheless, those datasets with synthetic image degradation between LRs and HRs suffer from a data bias in practical applications, which inevitably invites a distinct performance drop [21]. Until 2019, a real-world SR dataset, named RealSR [22], has been released by physically adjusting the camera focal length to collect paired LR-HR images. With a similar collection strategy, SR-Raw [23] and City100 [24] have been built for RWSR. However, those datasets only involve very limited cameras (*e.g.,* one or two cameras) and have very limited image diversity (*e.g.*, City100 is captured for printed posters in controlled laboratory conditions). Subsequently, Wei et al. [4] use five different cameras and build a large-scale Diverse Real-world SR dataset (DRealSR), as the largest one.

Cai et al. [22] propose a Laplacian Pyramid based Kernel Prediction Network (LP-KPN). LP-KPN inherits the spirit of KPN and learns per-pixel kernels to recover HR images [22]. Zhang et al. [23] propose a contextual bilateral loss (CoBi) to handle the slight misalignment between image pairs. However, in essence, image degradation in RWSR is heterogeneous for degradation kernels and is more complex among different camera devices [4]. To mitigate this issue, Wei et al. explore the learning bias in existing SR methods to those image regions that are easily reconstructed and propose Gradient-Weighted (GW) loss in the CDC model.

Nevertheless, the architectures of those widely-used neural networks sequentially process input images with stacked layers and are limited to well learn the high-order information of images explicitly and directly. Thus, in this work, we investigate the issue and propose Taylor Neural Network in a Taylor architecture to improve the feature learning for image super-resolution. Especially, our TNN can be implemented in different neural architectures by simple Taylor skip connections, with very little additional computation cost and additional parameter number, and further improves previous SR models.

III. Methodology

In this section, we first revisit the conventional CNNs in the SR task from Taylor series approximation, revealing the equivalence of our network and Taylor polynomial approximation. Then, we elaborate Taylor Neural Network and its connection to the Taylor expansion. Finally, we further introduce another two variants to explicitly learn high-order information under neural networks.

A. Revisiting Convolutional Neural Networks in SR

Given the input image $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$, a deep neural network with L layers is parameterized by parameters $\boldsymbol{\theta} = \{(\mathcal{W}_l, b_l), l = 1, ..., L\}$, and essentially learns a mapping function $\mathcal{F} : \boldsymbol{x} \mapsto \mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{x})$. In the l-th layer, its output is recursively computed with sequential layers as follows,

$$F_l = \sigma(\mathcal{W}_l ... \sigma(\mathcal{W}_1 * \boldsymbol{x} + b_1) + ...) + b_l), \tag{1}$$

where $\sigma(.)$ is the activation function of ReLU. With the convolutional filter $w^l_{ijc} \in \mathcal{W}_l$ ((i, j) is the 2-dimensional index of a filter for the channel c), its derived convolution response $F_l(s, t)$ in the image position (s, t) is simply formulated as follows,

$$F_l(s, t) = \sigma(\sum_c \sum_{i,j} w^l_{ijc} F_{l-1}(s + i, t + j, c) + b_{l-1}). \tag{2}$$

(1) CNN for high-resolution image reconstruction. Indicated in Equ.2, each element of F_l is locally determined with the receptive fields of the images. Due to the linearity of convolution operation and the piece-wise linearity of ReLU, together with Equ.1 and 2, *the derived mapping via the neural network is correlated with the input $\boldsymbol{x}$ in a first-order manner when σ is ReLU function*.

Based on this observation, we further analyze the neural networks and find that this first-order network architecture is not plausible for reconstructing high-quality images, especially those with complex textures or details that dominate the contents with high frequencies in image reconstruction problems. On the other hand, Equ. 1 and 2 indicate that the reconstructed signal (image) is simply regarded as the piecewise linear combination of the input. In the image super-resolution task, with this formulation, super-resolved images are derived from the interpolation of low-resolution images (inputs). Nevertheless, this strategy tends to fail to address the complex and heterogeneous image degradation in realistic scenarios [4].

(2) High-order information in CNN for image super-resolution. Essentially, the process of image super-resolution can be regarded as mapping low-quality images into high-quality images. However, *ReLU in networks is piecewise linear and would not provide the high-order information of input images.* ReLU networks are piecewise linear for which each pixel of an output image is a linear combination of the corresponding input image pixels. In other words, the ReLU network is essentially an adaptive interpolation method for pixels with different values and neighborhoods. In particular,

the second derivative of ReLU networks is zero everywhere, and they are thus incapable of modeling information contained in high-order derivatives of natural signals [11]. Considering the dilemma of the ReLU network, it is necessary to introduce the high-order information of images and to further improve the capability of deep neural networks.

(3) Connection with Taylor Expansion. On one hand, with the linearity of convolution and the piece-wise linearity of ReLU, $\mathcal{F}$ can be simply computed as follows,

$$\mathcal{F}(\boldsymbol{x}) = b + \sum_{(i,j)\in\mathcal{R}} \alpha_{ij}\boldsymbol{x}_{(i,j)}, \tag{3}$$

where, for simplicity, $\boldsymbol{x}_{(i,j)} \in \boldsymbol{x}$ represent the pixels of input images on the relative position with the convolutional filter, $\mathcal{R}$ represents the receptive field and α is determined by network parameters of different layers corresponding to the same receptive fields. On the other hand, a function can be approximated by the following Taylor expansion,

$$\mathcal{F}(\boldsymbol{x}) = b + \sum_{(i,j)} \nabla\mathcal{F}\boldsymbol{x}_{(i,j)} + \sum_{(i,j),(e,k)} \Delta\mathcal{F}\boldsymbol{x}_{(i,j)}\boldsymbol{x}_{(e,k)} + \cdots, \tag{4}$$

where $\nabla\mathcal{F}$ represents the first-order partial derivative $\frac{\partial\mathcal{F}}{\partial x_{(i,j)}}$ and $\Delta\mathcal{F}$ is the second-order partial derivative $\frac{\partial^2\mathcal{F}}{\partial x_{(i,j)}\partial x_{(e,k)}}$. Comparing with Equ.3 and 4, we can find that *the mapping function derived from Equ.1 can be simply regarded as to be identical to the first two terms of the Taylor expansion*, *i.e.*, *equivalent to the first-order network*. Thus, in a similar way, we can devise a high-order neural network. With the increase of layers, the network is trained via n-order Taylor expansion approximation.

B. Taylor Neural Network

Based on the aforementioned analyses, we aim to explore Taylor polynomial approximation in existing deep neural networks, to mitigate their inferior ability of learning the high-order information for reconstructing high-quality images.

Specifically, we propose a Taylor architecture with Taylor modules in neural networks for feature learning without bells and whistles, which is compatible with many existing deep neural networks and can be applied to widely-used neural network backbones, *e.g.*, convolutional neural networks and residual networks. The derived networks with Taylor architecture are named Taylor Neural Networks (TNNs). In this section, we will take plain convolutional networks as the backbone for example; for residual networks, convolutional layers are replaced with residual blocks in the Taylor architecture. TNN consists of multiple Taylor Modules at different layers. In each TM, to build the Taylor terms with high-order information, an additional branch, *i.e.*, TSC_c, is connected with each layer. With TSC_h at each layer, all the TM outputs F_i in Equ. 6 are summarized to generate the feature projection from the input, as indicated in Equ. 5.

1) Formulation: As claimed in Sec. III-A, a conventional neural network learns feature projections approximately with the first-order, which is approximatively equal to the summation of the first two terms of Taylor Series. Instead, in this paper, we formulate a general neural network architecture with Taylor Series for encoding/learning a feature representation/projection $\mathcal{F}$, which is employed in image reconstruction tasks to restore a high-quality image $\hat{Y}$,

$$\hat{Y} = up(\mathcal{F}(\boldsymbol{x})) = up(\sum\nolimits_{i=1}^{L} F_i) \tag{5}$$

$$F_i = \mathcal{W}_i * F_{i-1} \circ (\mathcal{W}_0 * \boldsymbol{x}), i = 2, 3, \cdots, L \tag{6}$$

$$F_1 = \mathcal{W}_1 * \boldsymbol{x}, \tag{7}$$

where $up(.)$ is the up-sampling function parameterized by $\{\mathcal{W}^R, b^R\}$, *i.e.*, PixelShuffle [16] for SR. At each layer, the feature map F_i is also named the *Taylor map* produced by the Taylor Module (TM). In this section, bias parameters and the activation function are omitted for simplifying the explanation. $\circ$ denotes element-wise product, and $\mathcal{W}_0$ is the parameter of an additionally introduced convolutional layer.

2) Taylor Skip Connection: At each layer, a TM outputs a Taylor map F_i, which is dominated by the input directly due to our proposed *Taylor Skip Connection (TSC)* which is different from other architectures with skip connections, *e.g.*, densely connected networks, because our architecture aggregates the information flows from different order contents of images at different layers. Specifically, the architecture of TNN introduces two types of Taylor Skip Connection, including Taylor-c skip connection (TSC_c) and Taylor-h skip connection (TSC_h). At the i-th layer, TM first obtains a response map ($\mathcal{W}_i\boldsymbol{x}$) by convoluting the output of the previous layer F_{i-1} as the input (the convolutional layer can be replaced with other layers or blocks *e.g.*, residual blocks [8], [12], residual in residual blocks [3]); then, TSC_c introduces the original input $\boldsymbol{x}$ into a TM with an element-wise product operation and derives a Taylor map; TSC_h aggregates the information (Taylor map) of this layer and all the previous layers with an addition operation.

For the i-th layer, we can recursively have the computation of Taylor map F_i (indicated in Equ. 6) as follows,

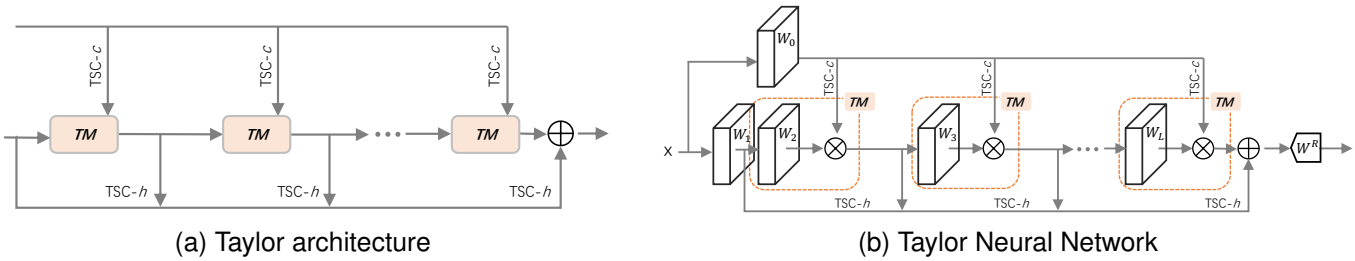
$$\begin{aligned} F_i &= \underbrace{\mathcal{W}_i * F_{i-1}}_{TSC_h} \circ \underbrace{(\mathcal{W}_0 * \boldsymbol{x})}_{TSC_c} \\ &= \underbrace{\mathcal{W}_i(\underbrace{\mathcal{W}_{i-1} * F_{i-2}}_{TSC_h} \circ \underbrace{(\mathcal{W}_0 * \boldsymbol{x})}_{TSC_c}) \circ \underbrace{(\mathcal{W}_0 * \boldsymbol{x})}_{TSC_c}}_{TSC_h} \\ &= \mathcal{W}_i(\mathcal{W}_{i-1} \cdots (\mathcal{W}_2(\mathcal{W}_1 * \boldsymbol{x}) \circ (\mathcal{W}_0 * \boldsymbol{x})) \circ \ldots (\mathcal{W}_0 * \boldsymbol{x})). \end{aligned} \tag{8}$$

As indicated in Equ. 8, the output feature map F_i is recursively related with preceding layers similar to conventional CNN, but introduces the high-order information with the aid of Taylor skip connections.

C. Connection to Taylor Expansion

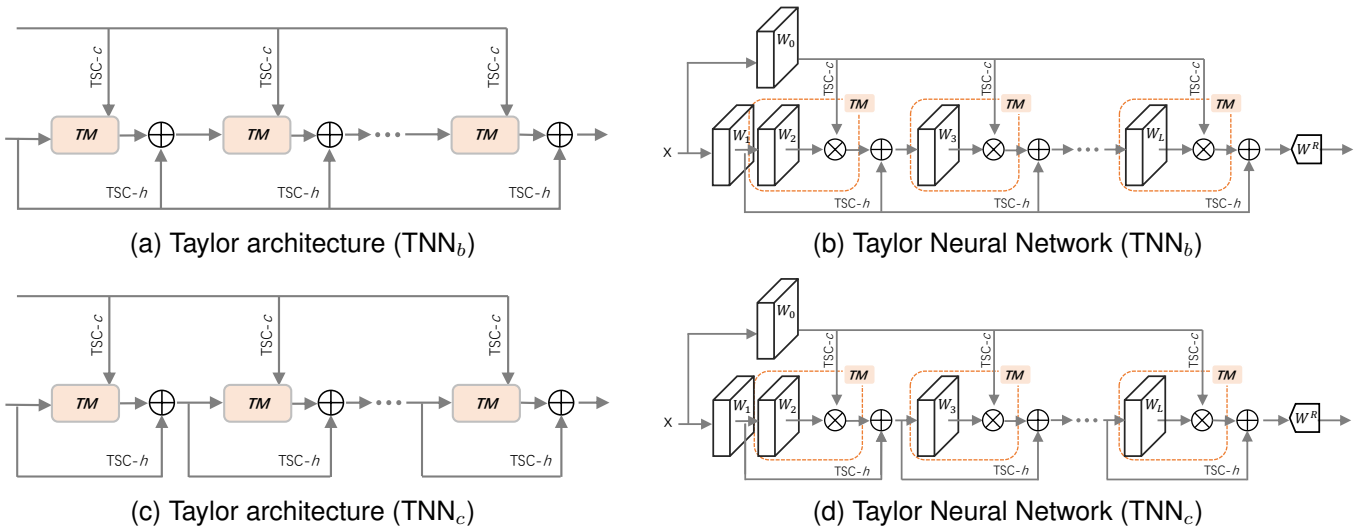
We will elaborately explain our Taylor neural network architecture from the view of Taylor series approximation.

1) The network with one layer: We first consider the Taylor Neural Network with only one layer in the backbone, whose parameters are $\{\mathcal{W}_1, b\}$. According to Equ. 7, the output feature map (Taylor map) of the backbone is $\mathcal{F} = b + \mathcal{W}_1 * \boldsymbol{x}$.



(a) Taylor architecture (b) Taylor Neural Network

Fig. 2. Illustration of the Taylor architecture. (a) is the Taylor architecture and (b) is its implementation. TNN is derived from our Taylor formulation in Equ. 5 and has sequential TMs with TSCs. TSC_c is utilized to introduce to derive the high-order terms and TSC_h aggregates different order information from different layers.



(a) Taylor architecture (TNN_b) (b) Taylor Neural Network (TNN_b)

(c) Taylor architecture (TNN_c) (d) Taylor Neural Network (TNN_c)

Fig. 3. Another two TNN variants. Following the TNN spirit (Equ.5-8), we provide three implements, including one in Fig. 2 (TNN_a) and these two additional variants (TNN_b and (TNN_c)). (a) and (b) show the architecture of (TNN_b) that summarizes the outputs from each TM in the tail of the last TM, and (c) and (d) show the architecture of (TNN_c) that accumulates the output of the previous layer to that of the next layer.

Accordingly, this derived feature map can also described as $\mathcal{F} = b + \sum \alpha_{ij} \boldsymbol{x}_{(i,j)}$.

2) The network with two layers: For a TNN with two layers in the backbone, whose parameters are $\{b, \mathcal{W}_0, \mathcal{W}_1, \mathcal{W}_2\}$, its output feature map (Taylor map) of the backbone is $\mathcal{F} = b + \mathcal{W}_1 * \boldsymbol{x} + \mathcal{W}_2(\mathcal{W}_1 * \boldsymbol{x}) \circ (\mathcal{W}_0 * \boldsymbol{x})$. Because of the linearity of the convolution operation, we have a simple alternative formulation to understand TNN,

$$\begin{aligned} \mathcal{F} &= b + \sum\nolimits_{(i,j)\in\mathcal{R}_1} \alpha_{ij} \boldsymbol{x}_{(i,j)} + \\ &\quad (\sum\nolimits_{(i,j)\in\mathcal{R}_{12}} w_{ij} \boldsymbol{x}_{(i,j)})(\sum\nolimits_{(e,k)\in\mathcal{R}_0} w_{ek} \boldsymbol{x}_{(e,k)}) \\ &= b + \sum \alpha_{ij} \boldsymbol{x}_{(i,j)} + \sum \beta_{ijek} \boldsymbol{x}_{(i,j)} \boldsymbol{x}_{(e,k)}. \end{aligned} \tag{9}$$

α_{ij} and β_{ijek} are the network parameters learned in the training process. $\mathcal{R}_0$, $\mathcal{R}_1$ and $\mathcal{R}_{12}$ represent the receptive field of one or two convolutional layers with parameters $\{\mathcal{W}_0, \mathcal{W}_1, \mathcal{W}_2\}$.

It is obvious that the final output feature of our proposed network is similar to Equ. 4, and the parameters of each layer just correspond to the partial derivatives of each order. It means that we can train the network from samples to learn the network parameters which can be analogous to the partial derivatives of each order for image super-resolution. It is the same case for TNN with more layers. Besides, the number of layers can be freely determined, which represents the order of Taylor expansion.

D. Two TNN variants

Following the spirit of TNN formulated in Equ. 5 $\sim$ 8, another two architectures are devised to implement the proposed TNN models, shown in Fig. 3.

Taylor Neural Network is formulated based on Taylor Series in Equ. 5 $\sim$ 8, which facilitates the deep feature learning to reconstruct a high-quality image in the image super-resolution task. At the i-th layer, on one hand, its output F_i is directly correlated with the input via TSC_c; on the other hand, TSC_h connects it with subsequent layers to aggregate different high-order information, which is analogous to Taylor expansion, to approximate the desired function. With the increase of layers, our Taylor architecture facilitates TNN progressively approximating the desired function.

Following this spirit, we propose another two variants of TNN, as shown in Fig. 3, which are all derived from our Taylor formulation in Equ. 5. For convenience, the version of TNN claimed in Sec.C is named as TNN_a; these two variants are named TNN_b and TNN_c, respectively. They have a stacked structure with sequential Taylor Modules together with the introduction of the input via TSC_c, while one distinct

difference is that they have different strategies to aggregate the outputs from different layers via TSC_h. TNN_a, shown in Fig. 2, summarizes the outputs from each TM in the tail of the last TM. It can be considered that each TM generates different high-order terms and the addition operation is conducted once. TNN_b, shown in Fig. 3, accumulates the output of the first layer to the outputs of other layers, respectively. TNN_c, shown in Fig. 3, accumulates the output of the previous layer to that of the next layer. Different from TNN_a, TNN_b and TNN_c conduct the addition operation at each layer, which is aggregating different high-order outputs from the first layer to the current layer. We mainly evaluate TNN_a in the paper.

IV. Experiments

In this section, we conduct comprehensive experiments to evaluate the proposed method on two representative real-world SR datasets. Our Taylor architecture of TNN can be applied in various deep neural networks. Thus, for each task, we choose those relevant state-of-the-art methods that are derived from representative neural networks. Besides, we also conduct experimental evaluations in the image enhancement task to further verify the proposed method.

A. Settings

Dataset. *1) Image super-resolution*: Real-world image super-resolution exhibits greater challenges of heterogeneous image degradation and complex degradation across different camera devices than synthetic image degradation [4]. Thus, in the paper, *all the SR methods are evaluated on more challenging real-world SR datasets, rather than synthetic SR datasets*. Two real-world SR datasets are considered, *i.e.*, RealSR [22] and DRealSR [4]. **RealSR** [22] has 595 HR-LR image pairs captured from two DSLR cameras. **DRealSR** has 894 image pairs captured from two DSLR cameras. Each training image is cropped in 192×192 patches.

Network Architecture. Several representative deep neural networks are considered as the backbones for three tasks, including conventional convolution networks, residual networks. To embed the proposed Taylor network structure into each backbone, a convolution layer is added to connect the input with each layer by element-wise product. In the tail of the network, an addition is performed to integrate the output information from all the layers. Compared with the original plain networks, the proposed Taylor structure just introduces the simple computation of one convolution layer and basic dot product with a trivial increase of parameters.

Implementation Details. In our proposed method, the size of the convolution layers is set to be 3×3. The convolutional layer which introduces the original input into TMs is the same size as the first convolutional layer of the backbone. In the case of RRDB with 23 Residual-In-Residual (RIR) dense blocks as the backbone, the first 12 RIR blocks are used to extract basic features and the subsequent 11 RIR blocks are transformed into Taylor blocks with TSCs.

In our proposed network, TMs are made up of different modules (*e.g.*, convolutional layers, residual blocks [8], dense blocks [10] and hourglass blocks [19]), respectively, which is determined by the backbone. In the case of HGSR [4], we used the first hourglass (HG) module to extract basic features and transformed the subsequent 5 HG blocks into Taylor Module with TSCs. To be applicable on various networks, we correspondingly introduce pooling layers and channel conversion layers to complete the element-wise product if the pooling layer is applied in the backbone or the channels of different layers are inconsistent.

TABLE II
Evaluation results of models with different layers.

Method	# Layer	PSNR	SSIM
SRResNet [12]	1	18.94	0.686
	2	19.24	0.704
	6	20.98	0.756
	11	25.49	0.809
	12	28.96	0.818
TNN (SRResNet)	1	26.58	0.746
	2	26.97	0.758
	6	27.56	0.777
	11	28.95	0.814
	12	**29.07**	**0.822**

B. Image Super-Resolution

1) Comparison with state-of-the-art methods: For image SR, we compare our method with several state-of-the-art methods, including SRResNet [12], EDSR [2], RRDB [3] and HGSR [4]. Notably, for a fair experimental comparison, all the methods have been trained following the same training settings, including training datasets. Similar to [2], EDSR has 256 filters per convolutional layer; other approaches use 64 filters. For a fair comparison, the plain base model of [4], HGSR without gradient weighted loss, is considered. With the same configuration, our TNNs are implemented by following the same configuration and only transforming their layers or blocks into our Taylor modules. Table I shows quantitative SR comparisons for $\times 2$, $\times 3$ and $\times 4$ on two real-world datasets. It is observed that our method achieves higher performance than existing models at all scales on both datasets. Especially, on $\times 2$ enlargement, the Taylor network derived from RRDB significantly improves the performance by 0.31dB on RealSR [22] and 0.56dB on DRealSR [4], which indicates the excellent effectiveness of our proposed method.

Comparison results of SR images on RealSR and DRealSR are illustrated in Fig. 4. It is observed that our proposed TNNs generate less over-smooth details and artifacts. For example, in the second row of Fig. 4, more sharp edges of the building are restored by our TNNs that take EDSR as the backbone, while EDSR with the plain architecture presents the over-fitting tendency. Particularly, false textures occur distinctly in the derived SR image from RRDB.

2) Model Analysis: Our detailed analyses of the proposed TNNs are conducted in the real-world image SR task.

Evaluation on the number of TMs. Considering that TNNs follow the spirit of Taylor Series, we provide the evaluation result on the number of TMs in Table II, to analyze the effects of TMs for TNN. This evaluation is taking SRResNet as the backbone. With the increase of the number of layers,

TABLE I
COMPARISON RESULTS OF IMAGE SUPER-RESOLUTION ON REALSR AND DREALSR. **VALUES** IN BOLD INDICATES THE HIGHEST PERFORMANCE AMONG ALL THE COMPARISON METHODS FOR A SCALE FACTOR.

Scale	Method	DataSet: RealSR [22]			DataSet: DRealSR [4]		
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
×2	SRResNet [12]	33.54	0.919	0.155	33.76	0.901	0.162
	TNN (SRResNet)	33.83	0.921	0.149	34.12	0.905	0.160
	HGSR [4]	33.62	0.920	0.142	33.97	0.901	0.157
	TNN (HGSR)	33.79	0.921	0.144	34.12	0.905	0.166
	EDSR [2]	33.88	0.920	0.145	34.24	0.908	0.155
	TNN (EDSR)	34.00	0.923	0.146	34.33	0.908	0.157
	RRDB [3]	33.80	0.922	0.146	33.89	0.906	0.155
	TNN (RRDB)	**34.11**	**0.925**	**0.139**	**34.45**	**0.911**	**0.151**
×3	SRResNet [12]	30.65	0.862	0.228	32.47	0.862	0.269
	TNN (SRResNet)	30.69	0.861	0.232	32.75	0.867	0.264
	HGSR [4]	30.68	0.863	0.227	32.54	0.865	0.252
	TNN (HGSR)	30.76	0.865	0.222	32.79	0.867	0.257
	EDSR [2]	30.86	0.867	0.219	32.93	**0.876**	0.241
	TNN (EDSR)	30.89	0.866	0.219	32.99	0.872	0.247
	RRDB [3]	30.72	0.866	0.219	32.79	0.873	0.242
	TNN (RRDB)	**30.92**	**0.867**	**0.218**	**33.03**	0.875	**0.241**
×4	SRResNet [12]	28.95	0.821	0.281	31.63	0.847	0.341
	TNN (SRResNet)	29.11	0.822	0.287	31.76	0.849	0.332
	HGSR [4]	29.12	0.824	0.284	31.79	0.850	0.314
	TNN (HGSR)	29.21	0.826	0.281	31.86	0.851	0.314
	EDSR [2]	29.09	0.827	0.278	32.03	0.855	0.307
	TNN (EDSR)	29.15	0.825	**0.273**	31.97	0.853	0.311
	RRDB [3]	29.15	0.826	0.279	31.92	0.857	0.308
	TNN (RRDB)	**29.27**	**0.828**	0.278	**32.14**	**0.857**	**0.305**

TABLE III
COMPUTATION COMPARISON WITH BASELINE METHODS ON THE REALSR DATASET.

×2								
Method	SRResNet	TNN (SRResNet)	HGSR	TNN (HGSR)	EDSR	TNN (EDSR)	RRDB	TNN (RRDB)
Params (M)	1.40	1.41	39.41	39.41	40.73	40.74	16.66	16.66
MACs (G)	72.56	73.13	322.35	322.44	2044.69	2045.05	847.83	847.92
×3								
Method	SRResNet	TNN (SRResNet)	HGSR	TNN (HGSR)	EDSR	TNN (EDSR)	RRDB	TNN (RRDB)
Params (M)	1.58	1.60	39.41	39.41	43.68	43.69	16.66	16.66
MACs (G)	85.73	86.3	341.35	341.44	2194.46	2194.82	866.83	866.92
×4								
Method	SRResNet	TNN (SRResNet)	HGSR	TNN (HGSR)	EDSR	TNN (EDSR)	RRDB	TNN (RRDB)
Params (M)	1.55	1.56	39.45	39.45	43.09	43.10	16.70	16.70
MACs (G)	111.58	112.15	375.36	375.46	2522.58	2522.94	900.85	900.94

TABLE IV
ABLATION STUDY ON DIFFERENT NONLINEAR ACTIVATION METHODS ON THE REALSR DATASET.

Method	Activation	PSNR	SSIM	LPIPS
SRResNet	ReLU	33.54	0.919	0.155
SRResNet	SiLU	33.75	0.904	**0.148**
SRResNet	ELU	33.78	0.901	0.149
SRResNet	Tanh	33.48	0.901	**0.148**
SRResNet	SIREN [11]	15.09	0.482	0.920
TNN (SRResNet)	ReLU	**33.83**	**0.921**	0.149

i.e., the number of TMs for TNN, a constant performance improvement is observed for TNN. In particular, with the same layers, TNN always outperforms the base model SRResNet in the metrics of PSNR and SSIM.

Analysis on the high-order in TNN. To investigate different order information at different layers, we visualize the generated SR images from the outputs of TMs at different layers in Fig.5. With the same backbone as SRResNet, our method recovers a relatively clear SR image when only using features at the first layer, while in SRResNet this phenomenon does not occur until the entire network is fully used. This can attribute to our TSC, which connects the final output with each layer and is beneficial for the gradient back-propagation to each layer directly. With the increase of layers and the order, the SR images of our method become better and better, which is analogous to the details brushed up layer by layer. Our TNNs provide an appealing strategy to learn high-order image information for image super-resolution. Furthermore, Compared with nonlinear activation methods to incorporate high-order information, TNN still presents a superior performance: 28.95dB (SRResNet) vs. 23.80dB (SIREN) vs 29.11dB (TNN), shown in Table IV. Instead, SRResNet with SIREN has a significant performance drop.

Analysis on the number of parameters and Multi-adds. Compared with baselines, our Taylor design has one additional convolutional layer with weights W_0 and introduces a very slight increase of parameter number. This is empirically demonstrated in Table III. For example, parameter numbers are

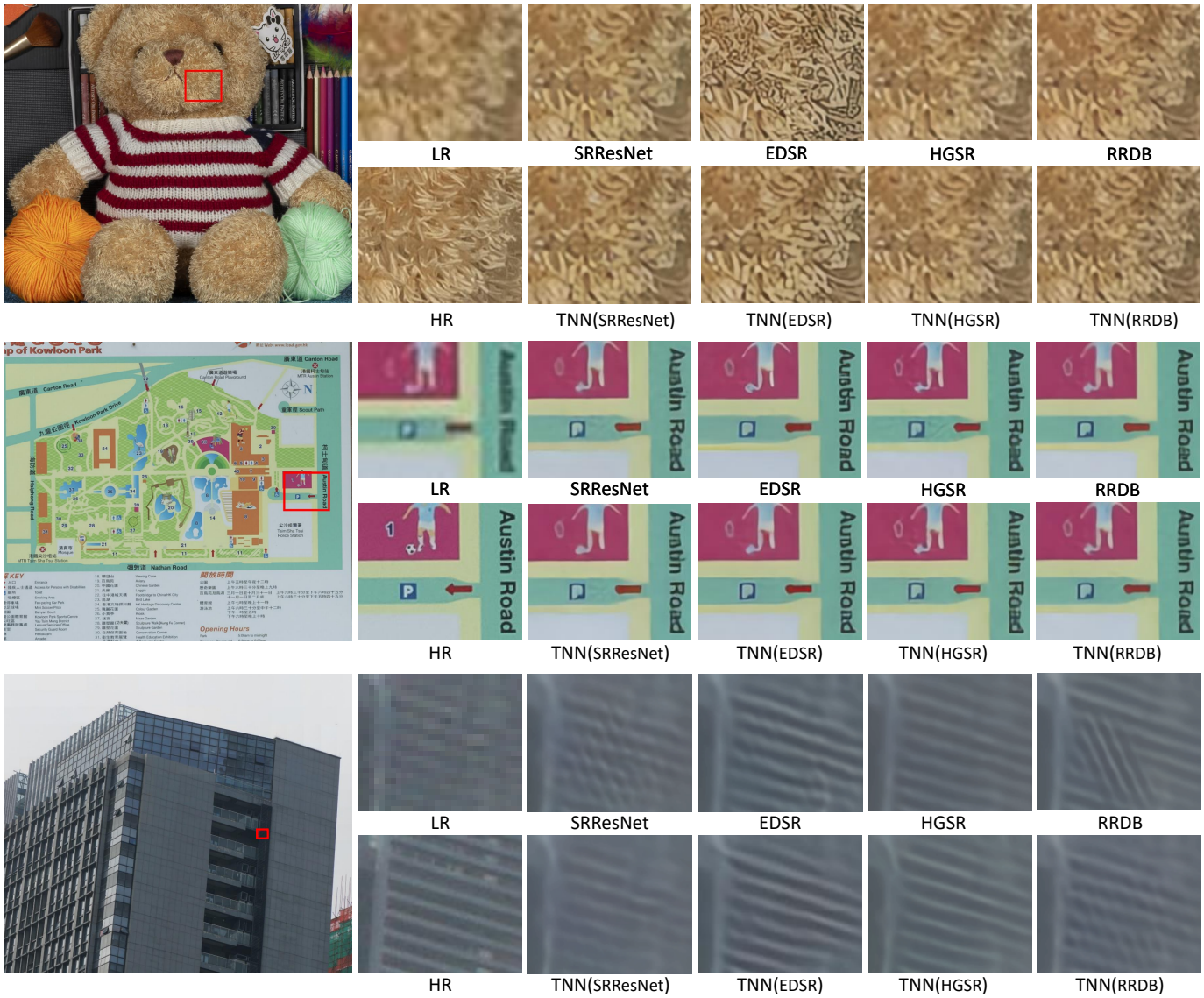


Fig. 4. Comparison of SR results with state-of-the-art methods on RealSR (the first image in each row) and DRealSR (the second image in each row).

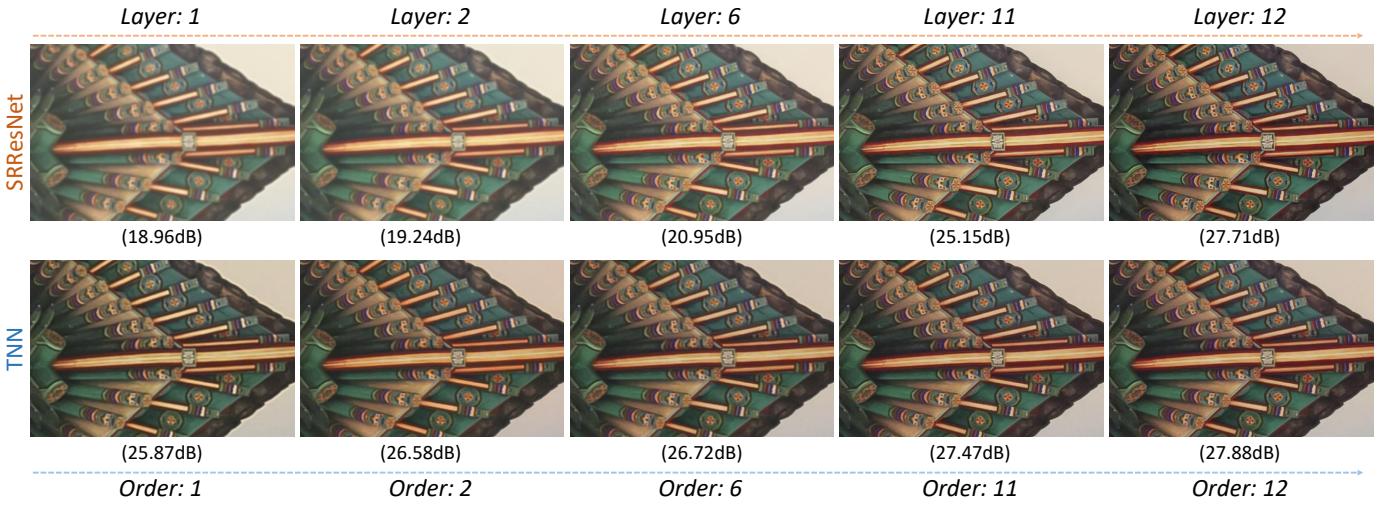


Fig. 5. Visualization of SR results at different layers. With the same backbone to SRResNet, our TNN reconstructs better and better SR results with the increase of layers and the learned high-order information is beneficial for image super-resolution.

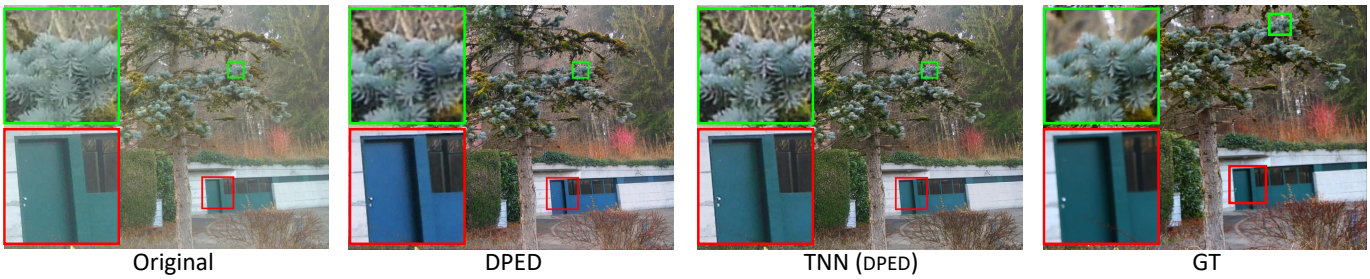


Fig. 6. Image enhancement results of DPED and our TNN with the same backbone as the DPED method.

TABLE V
PERFORMANCE OF THREE TNN VARIANTS.

TNN_a		TNN_b		TNN_c	
PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
29.11	0.822	29.06	0.821	29.07	0.821

43.09M (EDSR) vs. 43.10M ($TNN_{(EDSR)}$) and 16.70M (RRDB) vs. 16.70M ($TNN_{(RRDB)}$).

Evaluation on different activation functions. As suggested, we have provided an ablation study on different activation functions, as shown in Table IV. In comparison with ReLU, it is observed that SiLU and ELU bring about 0.2dB performance improvement, Tanh has a slight decrease, but SIREN [11] presents a significant performance drop. The main reason is possibly that, although it is not linear, its periodic property is not beneficial to reconstruct image details under the complex image degradation in the image super-resolution task; or, it is not well deployed or behaved under SRResNet for the image super-resolution task.

Analysis on different TNN variants. In Table V, we provide the experimental comparison results of TNN variants. With the backbone of SRResNet, TNN_a outperforms the other TNN models on the RealSR dataset; TNN_b and TNN_c have a similar performance. The reason is that TSC_h in TNN_a facilities directly the loss back-propagation to each layer. Thus, TNN_a is adopted in our work.

Taylor Maps. We further analyze the learned Taylor maps at different layers, as shown in Fig. 1. The learned feature maps by SRResNet [12] are also provided in Fig. 1. It is observed that our Taylor feature maps present strong activations for image details at different layers. Instead, those from SRResNet are prone to flat regions that are easier to reconstruct and gradually activate fewer details with the increase of layers. This can attribute to the simply learning of image residual, which would drive the model to fit the easy visual components [4], and thus is incapable of effectively learning more details for difficult visual components, *e.g.*, textures.

C. Image Enhancement

To further evaluate our proposed TNN, we also conduct experiments of image enhancement comparing our method with the DPED method on DPED dataset. **DEPD** [25] is a large-scale dataset for image enhancement, aiming to improve low-quantity phone images with respect to high-quantity D-SLR images. DEPD is commonly used for NTIRE image enhancement challenges [26]. It has over 6K photos taken synchronously by a DSLR camera and 3 low-end cameras of smartphones in a wide variety of conditions.

TABLE VI
COMPARISON RESULTS OF IMAGE ENHANCEMENT ON THE DPED DATASET.

Dataset		Method			
		DPED [25]	TNN (DPED)	DPED8 [25]	TNN (DPED8)
Iphone	PSNR	20.08	**20.55**	20.47	**20.73**
	SSIM	0.920	**0.920**	0.920	**0.921**
Blackberry	PSNR	20.07	**20.23**	20.14	**20.26**
	SSIM	0.933	**0.933**	**0.934**	0.933
Sony	PSNR	21.81	**21.97**	21.92	**22.03**
	SSIM	0.944	**0.946**	0.945	**0.946**

PSNR and SSIM results are reported in Table VI. The backbone of DPED [25] is derived from a residual network with three residual blocks. Due to few works with typical architectures of the neural network on DPED dataset, for better comparison, we deepened the network of DPED into an eight-layer residual network, named DPED8. DPED8 and our TNN based on it are trained on the same settings as DPED. As shown in Table VI, our TNN achieves higher PSNR performance on three mobile phone datasets, *i.e.*, Iphone, Blackberry and Sony. On Iphone, our TNN achieves 0.47dB PSNR improvement over DPED. Similar improvements are also significant on the other two datasets.

Visualization results are shown in Fig. 6. It is observed that DPED changes the color in the image region marked in red while our TNN restores similar color to the ground-truth (GT). In addition, the enhanced images by our TNN have more reconstructed details of the image.

V. CONCLUSION

In this work, we establish a Taylor architecture to explore the high-order information of images in neural networks and propose Taylor Neural Networks for feature learning in real-world image super-resolution. Without any image processing for extracting high-order or high-frequency contents in images, our TNNs build Taylor modules for Taylor maps with the different high-order attention information, and leverage Taylor Skip Connections to aggregate those Taylor maps from different layers for reconstructing more image details. The proposed

Taylor architecture can be flexibly applied to various existing networks in a simple manner for the introduction of high-order information, which provides an insightful research topic for feature learning in image reconstruction. The proposed TNNs are evaluated under different existing networks as baselines on two real-world SR benchmarks, and comprehensive experimental results demonstrate the superiority of our TNNs.

ACKNOWLEDGMENTS

This work was supported in part by NSFC (No. U21A20470), and Science and Technology Program of Guangzhou, China (No. 202201011550).

REFERENCES

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*, 2014, pp. 184–199.

[2] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1132–1140.

[3] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 0–0.

[4] P. Wei, Z. Xie, H. Lu, Z. Zhan, Q. Ye, W. Zuo, and L. Lin, "Component divide-and-conquer for real-world image super-resolution," in *Proceedings of the European Conference on Computer Vision*, 2020.

[5] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 286–301.

[6] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.

[7] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.

[8] 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.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[11] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Advances in Neural Information Processing Systems*, 2020.

[12] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 105–114.

[13] H. Wang, Y. Fan, Z. Wang, L. Jiao, and B. Schiele, "Parameter-free spatial attention network for person re-identification," *arXiv preprint arXiv:1811.12150*, 2018.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[16] W. 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.

[17] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European conference on computer vision*, 2016, pp. 391–407.

[18] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 624–632.

[19] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*, 2016, pp. 483–499.

[20] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5728–5739.

[21] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4791–4800.

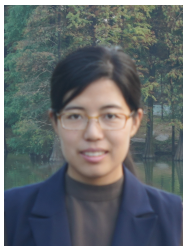
[22] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *International Conference on Computer Vision*, 2019.

[23] X. Zhang, Q. Chen, R. Ng, and V. Koltun, "Zoom to learn, learn to zoom," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3762–3770.

[24] C. Chen, Z. Xiong, X. Tian, Z. Zha, and F. Wu, "Camera lens super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1652–1660.

[25] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool, "Dslr-quality photos on mobile devices with deep convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3277–3285.

[26] A. Lugmayr, M. Danelljan, and R. Timofte, "Ntire 2020 challenge on real-world image super-resolution: Methods and results," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 494–495.



Pengxu Wei received the B.S. degree in computer science and technology from the China University of Mining and Technology, Beijing, China, in 2011 and the Ph.D. degree from University of Chinese Academy of Sciences in 2018. Her current research interests include computer vision and machine learning, especially benchmarks and solutions for real-world image super-resolution. She is currently a research scientist at Sun Yat-sen University.

Ziwei Xie received the B.S. degree in Electronic engineering and Information Science from University of Science and Technology of China, Hefei, China, in 2018 and the M.S. degree in Computer Science from Sun Yat-Sen University, Guangzhou, China, in 2021. He is currently working as a researcher in Tencent. His current research interests include computer vision and deep learning.

Guanbin Li (M'15) is currently an associate professor in School of Data and Computer Science, Sun Yat-sen University. He received his PhD degree from the University of Hong Kong in 2016. His current research interests include computer vision, image processing, and deep learning. He is a recipient of ICCV 2019 Best Paper Nomination Award. He has authorized and co-authorized on more than 100 papers in top-tier academic journals and conferences. He serves as an area chair for the conference of VISAPP. He has been serving as a reviewer for numerous academic journals and conferences such as TPAMI, IJCV, TIP, TMM, TCyb, CVPR, ICCV, ECCV and NeurIPS.

Liang Lin (M'09, SM'15) is a Full Professor of computer science at Sun Yat-sen University. He served as the Executive Director and Distinguished Scientist of SenseTime Group from 2016 to 2018, leading the R&D teams for cutting-edge technology transferring. He has authored or co-authored more than 200 papers in leading academic journals and conferences, and his papers have been cited by more than 17,000 times. He is an associate editor of IEEE Trans. Neural Networks and Learning Systems and IEEE Trans. Human-Machine Systems, and served as Area Chairs for numerous conferences such as CVPR, ICCV, SIGKDD and AAAI. He is the recipient of numerous awards and honors including Wu Wen-Jun Artificial Intelligence Award, the First Prize of China Society of Image and Graphics, ICCV Best Paper Nomination in 2019, Annual Best Paper Award by Pattern Recognition (Elsevier) in 2018, Best Paper Dimond Award in IEEE ICME 2017, Google Faculty Award in 2012. His supervised PhD students received ACM China Doctoral Dissertation Award, CCF Best Doctoral Dissertation and CAAI Best Doctoral Dissertation. He is a Fellow of IET/IAPR/AAIA.