Scale-wise Convolution for Image Restoration

Yuchen Fan¹, Jiahui Yu¹, Ding Liu², Thomas S. Huang¹

¹University of Illinois at Urbana-Champaign, ²Bytedance Inc.
yuchenf4@illinois.edu, jyu79@illinois.edu, liuding@bytedance.com, t-huang1@illinois.edu

Abstract

While scale-invariant modeling has substantially boosted the performance of visual recognition tasks, it remains largely under-explored in deep networks based image restoration. Naively applying those scale-invariant techniques (e.g., multi-scale testing, random-scale data augmentation) to image restoration tasks usually leads to inferior performance. In this paper, we show that properly modeling scale-invariance into neural networks can bring significant benefits to image restoration performance. Inspired from spatial-wise convolution for shift-invariance, "scale-wise convolution" is proposed to convolve across multiple scales for scale-invariance. In our scale-wise convolutional network (SCN), we first map the input image to the feature space and then build a feature pyramid representation via bi-linear down-scaling progressively. The feature pyramid is then passed to a residual network with scale-wise convolutions. The proposed scale-wise convolution learns to dynamically activate and aggregate features from different input scales in each residual building block, in order to exploit contextual information on multiple scales. In experiments, we compare the restoration accuracy and parameter efficiency among our model and many different variants of multi-scale neural networks. The proposed network with scale-wise convolution achieves superior performance in multiple image restoration tasks including image super-resolution, image denoising and image compression artifacts removal. Code and models are available at: https://github.com/ychfan/scn_sr.

1 Introduction

The exploitation of scale-invariance in computer vision and image processing has greatly benefited feature engineering (Lowe 2004), image classification (Szegedy et al. 2015), object detection (Cai et al. 2016; Lin et al. 2017; Fu et al. 2017; Singh and Davis 2018; Yu et al. 2016), semantic segmentation (Ronneberger, Fischer, and Brox 2015; Badrinarayanan, Kendall, and Cipolla 2017) and the training of convolutional neural networks (Szegedy et al. 2015; Singh, Najibi, and Davis 2018). For example, Lowe proposed the scale-invariant features (SIFT) (Lowe 2004) that are consistent across a substantial range of affine distortion. Due to its robustness, the SIFT features have been shown undeniably successful and ubiquitously used for image matching, image classification, object detection and many other computer vision tasks. More recently, the importance of scale-invariance is demonstrated in object detection using convolutional neural networks (Singh and Davis 2018), which is further improved by an efficient multi-scale training method (Singh, Najibi, and Davis 2018) of object detectors.

While modeling scale-invariance has greatly improved visual recognition performance, few efforts have succeeded in utilizing scale-invariant features for image restoration tasks based on convolutional neural networks. For example, random-scale data augmentation during training and multi-scale testing are proven beneficial to a wide range of image recognition problems like classification (Szegedy et al. 2015) and detection (Fu et al. 2017). However, naively applying those techniques usually lead to worse performance for image restoration like super-resolution. It is likely due to the fact that, unlike image recognition problems, the performance of low-level image restoration tasks is very sensitive to naive scaling of input images. Is modeling scale-invariance unnecessary for image restoration tasks?

In this work, we demonstrate that properly modeling scale-invariance into convolutional neural networks can bring significant gain for image restoration tasks. We draw inspiration from spatial-wise convolution for shift-invariance, and propose the "scale-wise convolution" that is designed to convolve across multiple scales to achieve scale-invariance. In our scale-wise convolutional network (SCN), we first map the input image to the feature space and then build a feature pyramid representation via bi-linear down-scaling progressively. Based on the feature pyramid, we apply the same function (several shared layers) to process different scales and aggregate the neighborhoods. After several residual blocks with scale-wise convolutions, we take the feature output of the largest scale and predict the results with a linear convolution. The proposed scale-wise convolution learns to dynamically activate and aggregate features from different input scales in each residual building block, in order to exploit contextual information on multiple scales.

We compare our proposed network to several multi-scale convolutional networks with different design choices (shared or un-shared weights, different layer connectivity, etc). We show in our experiments that these multi-scale convolutional networks are inferior than our proposed scale-wise convolutional network. We also study the way of
feature fusing across different scales: i.e., (1) convolution and deconvolution with stride, (2) average pooling down-sampling and nearest neighbour up-sampling, and (3) bilinear re-sampling (both down-sampling and up-sampling). We further study the number of scales and the scale factor in our scale-wise convolutional networks and analyze their performances.

To demonstrate the effectiveness of our proposed scale-wise convolutional networks, we conduct extensive experiments on image super-resolution, denoising and compression artifacts removal. Our experiments reveal that: (1) Multi-scale models significantly improve the performance compared with single-scale models. (2) Scale-wise convolution for cross-scale modeling is more parameter-efficient than general multi-scale models, and achieves better efficiency-accuracy trade-offs. (3) Scale-wise convolution benefits from more scales in feature pyramid and proper scaling ratios in-between. Moreover, our proposed SCN can achieve superior results to state-of-the-art methods with even fewer parameters on image super-resolution, denoising and compression artifacts removal.

2 Related Work

2.1 Scale-Invariant Modeling in Visual Perception

The prior of scale-invariance in image processing has motivated many deep neural network architectures and models. For example, a scale-invariant convolutional neural network (SiCNN) proposed by (Xu et al. 2014) is designed to incorporate multi-scale feature exact and classification into the multi-column architecture. These columns share a same set of weights which are transformed for different scales. SiCNN improves the performance of image classification by learning the feature in different scales in different columns. Multi-scale architectures like U-Net (Ronneberger, Fischer, and Brox 2015) and SegNet (Badrinarayanan, Kendall, and Cipolla 2017) incorporate multiple scales in different network staged and inter-connects them, which are proven to have higher performance on image segmentation tasks. Feature Pyramid Network (Lin et al. 2017) proposed a top-down architecture with lateral connections for building high-level semantic feature maps at all scales. Moreover, during the training and testing of deep neural networks, multiple scales can also be useful for a wide range of applications like classification and detection.

2.2 Multi-Scale Architectures in Image Restoration

Several approaches have explored involving multiple scales into neural network architecture for image restoration tasks. The Dual-State Recurrent Network (DSRN) (Han et al. 2018) proposed a neural network to jointly utilize signals on both low-resolution scale and high-resolution scale for image super-resolution. Specifically, recurrent signals in DSRN are exchanged between these two scales in both directions via delayed feedback. The Deep Back-Projection Networks (Haris, Shakhnarovich, and Ukita 2018) exploited multiple iterative up-sampling and down-sampling layers, which provides an error feedback mechanism for projection errors at each stage of image super-resolution. Balanced two-stage network (Fan et al. 2017) is decoupled into scales along layers. Multi-scale residual network (MSRN) is proposed to fully exploit the image features (Li et al. 2018), in which the convolution kernels of different sizes can adaptively detect the image features in different scales.

2.3 Weight Sharing in Image Restoration

For image restoration tasks, the receptive field of a convolutional network has a critical role as it determines the amount of contextual information that can be used for prediction. However, naively increasing the number of layers leads to computationally expensive models with low parameter efficiency. To make a model compact, recursive neural networks (Kim, Kwon Lee, and Mu Lee 2016b; Tai, Yang, and Liu 2017) are proposed by sharing weights repeatedly among different recurrent stages. The recurrent architecture has also been used in (Han et al. 2018; Liu et al. 2018).

3 Scale-wise Convolutional Networks

In this section, we introduce our approach in the following manner. We first describe the scale-wise convolution operator and discuss its properties in details. Then the scale-wise convolutional networks are developed based on repeatedly constructing residual blocks with scale-wise convolution. Finally we discuss a number of design choices and construct several multi-scale network variants as our baselines.

3.1 Scale-wise Convolution

We propose the scale-wise convolution operator that convolves features along the scale dimension from a multi-scale feature pyramid. Figure 1 visually illustrates the idea of “convolution” across scales.

Formally, the scale-wise convolution \( \mathcal{F} \) takes a multi-scale feature pyramid \( X^l = \{x^l_s\}_{s=1}^N \) from layer \( l \) with multiple scales from 1 to \( N \), and generates the next feature pyramid with the same size \( X^{l+1} = \mathcal{F}(X^l) \) by “convolution” across scales. The output feature \( x^{l+1}_{s} \) is transformed from features of previous neighbouring scales \( \{x^l_{s-k}, \ldots, x^l_{s}, \ldots, x^l_{s+k}\} \). Specifically, the output of scale-
The pro-
Unified Architecture for Image Restoration.

networks for image super-resolution (Yu et al. 2018; Fan, Yu, and Huang 2018; Fan et al. 2019), by adding multi-scale features and scale-wise convolutions in residual blocks.

Residual Blocks with Scale-wise Convolution. Based on the feature pyramid, we apply the same function (i.e., several shared layers) to process different scales. Each residual block has a spatial convolution layer for expanding features, a ReLU activation layer and another spatial convolution layer for mapping feature back to original feature size. Afterwards we aggregate the neighbouring scales, as shown in Figure 1. The proposed scale-wise convolution learns to dynamically activate and aggregate features from different input scales in each residual building block.

Wide-activated residual networks (Yu et al. 2018) with wider features before ReLU activation achieve significantly better performance for image and video super-resolution, with the same parameters and computational budgets. The resulted residual network has a slim identity mapping pathway with wider channels before activation in each residual block.
Comparison to Multi-Scale Architectures. Our proposed SCN, at the first glance, is similar to multi-scale neural networks. However, they are fundamentally different in two aspects.

First, the scale-wise convolution applies same weights to different scales, while other multi-scale networks usually aggregate features in difference scales with specific parameters. For example, U-Net style (Ronneberger, Fischer, and Brox 2015) models in Figure 3 progressively down-sample features through explicit stages with multiple network layers in encoder and linearly composite multi-scale features by skip connections in decoder. Moreover, our proposed scale-wise convolution also adapts a unified operator to fuse between different scales and thus is more compact.

Second, the proposed scale-wise convolution aggregates multi-scale features gradually and locally, while other multi-scale networks fuse features at specific layers. For example, PSPNet style (Zhao et al. 2017) models in Figure 3 independently execute over scales and combine multi-scale features in the final layers. The scale-wise convolution fully utilizes the similarity prior of images across scales and is more robust to scale variance. Thus, the proposed operator can be viewed as “convolution” across scales.

Our experiments show that the proposed SCN achieved better performance than previous multi-scale architectures.

4 Experimental Results

We perform our experiments on image super-resolution, denoising and compression artifact removal to show the significance of our proposed SCN for image restoration.

4.1 Experimental Settings

Dataset. Multiple datasets are used for different image restoration tasks.

For image super-resolution, models are trained on DIV2K (Timofte et al. 2017) dataset with 800 high resolution images since the dataset is relatively large and contains high-quality (2K resolution) images. The default splits of DIV2K dataset consist 800 training images and 100 validation images. The benchmark evaluation sets include Set5 (Bevilacqua et al. 2012), Set14 (Zeyde, Elad, and Protter 2010), BSD100 (Martin et al. 2001) and Urban100 (Huang, Singh, and Ahuja 2015) with three upscaling factors: x2, x3 and x4.

For image denoising, we use Berkeley Segmentation Dataset (BSD) (Martin et al. 2001) 200 training and 200 testing images for training purpose, as (Zhang et al. 2017). The benchmark evaluation sets are Set12, BSD64 (Martin et al. 2001) and Urban100 (Huang, Singh, and Ahuja 2015) with noise level of 15, 25, 50.

For compression artifacts removal, we use 91 images in (Yang et al. 2010) and 200 training images in (Martin et al. 2001) for training. The benchmark evaluation sets are LIVE1 and Classic5 with JPEG compression quality 10, 20, 30 and 40.

Training Setting. Experiments are conducted in patch-based images and their degraded counterparts. Data augmentation including flipping and rotation are performed online during training, and Gaussian noise for image denoising tasks is also online sampled. 1000 patches randomly sample per image and per epoch, and 40 epochs in total. All models are trained with L1 distance through ADAM optimizer, and learning rate starts from 0.001 and halves every 3 epochs after the 25th. We use deep models with 8 residual blocks, 32 residual units and 4x width multiplier for most experiments, and 64 blocks and 64 units for super-resolution benchmarks.

4.2 Main Results

In this part, we compare our models with the state-of-the-art methods on image super-resolution, image denoising and image compression artifact removal. The results are shown in Table 1 (Single Image Super-Resolution), Table 2 (Image Denoising) and Table 3 (Image Compression Artifact Removal).

Image super-resolution. We compare our SCN with state-of-the-art single image super-resolution methods: A+ (Timofte, De Smet, and Van Gool 2014), SRCNN (Dong et al. 2014), VDSR (Kim, Kwon Lee, and Mu Lee 2016a), EDSR (Lim et al. 2017), WDSR (Yu et al. 2018). Self-ensemble strategy is also used to further improve our SCN, which we denote as SCN+.

As shown in Table 1, our proposed SCN and SCN+ achieves best results on all benchmark dataset and across all three upscaling factors. It indicates that properly modeling scale-invariance into neural networks, specifically the scale-wise convolution, can bring significant benefits to image super-resolution.

Furthermore, we visually compare our super-resolved images with other baseline methods. As shown in the Figure 4, our method produces higher quality images (for example, sharper edges) than our baseline methods: SR-CNN (Dong et al. 2014), FSRCNN (Dong, Loy, and Tang 2016), VDSR (Kim, Kwon Lee, and Mu Lee 2016a), LapSRN (Lai et al. 2017), MemNet (Tai et al. 2017) and EDSR (Lim et al. 2017).

Image denoising. We compare our SCN with state-of-the-art image denoising methods: BM3D (Dabov, Foi, and
Table 1: Public image super-resolution benchmark results and DIV2K validation results in PSNR / SSIM. Red indicates the best performance and blue indicates the second best. + indicates results with self-ensemble.

| Method | Set5 | B100 | Urban100 | DIV2K validation |
|--------|------|------|----------|-----------------|
| Scale 2x | 3.66 / 0.9299 | 2.96 / 0.8431 | 2.82 / 0.8124 | 2.82 / 0.8124 |
| Scale 3x | 3.66 / 0.9282 | 2.96 / 0.8431 | 2.82 / 0.8124 | 2.82 / 0.8124 |
| Scale 4x | 3.66 / 0.9282 | 2.96 / 0.8431 | 2.82 / 0.8124 | 2.82 / 0.8124 |

Table 2: Benchmark image denoising results. Training and testing protocols are followed as in (Zhang et al. 2017). Average PSNR/SSIM for various noise levels. The best results are in bold.

| Method | LIVE1 | BD100 | Urban100 |
|--------|-------|-------|----------|
| Noisy 20 | 29.86 / 0.9248 | 29.86 / 0.9248 | 29.86 / 0.9248 |
| Noisy 40 | 29.86 / 0.9248 | 29.86 / 0.9248 | 29.86 / 0.9248 |

4.3 Ablation Study

In this section, we conduct ablation study to justify the significance of proposed scale-wise convolution. The experiments are on image super-resolution with x2 up-scaling.

Multi-scale architectures. We compare three representative design choices of multi-scale network structures discussed in previous section. The U-Net style model (Ronneberger, Fischer, and Brox 2015) has 2 scales in both encoder and decoder, and 2 residual blocks per scale, then 8 blocks in total. The PSPNet style model (Zhao et al. 2017) has 2 scales and 8 residual blocks per scale, and features from scales are fused in output layer. Our SCN also has 2 scales and 8 residual blocks. The total numbers of parameters of the models are almost the same. Table 4.3 shows the significance of proposed scale-wise convolution. The experiments are on image super-resolution with x2 up-scaling.

Parameter sharing across scales. We study whether the parameters for different scales can be shared or not. In table 5 we compare (1) single-scale baseline (Baseline), (2) multi-scale architecture without sharing weights (Unshared), (3) multi-scale architecture with sharing weights (Shared) and (4) a larger model (more parameters) multi-scale architecture without sharing weights. We report their number of parameters in the Table.

Table 5 shows that under same number of parameters, shared version has much better performance. To achieve a similar performance, the un-shared version has to involve more parameters. The results prove the scale-invariance of our proposed SCN and its effectiveness.

Resampling in Scale-wise Convolution. Scale-wise convolution aggregates information from nearby input scales. A resampling operation is required to fuse features of different scales into the same scale, as shown in Figure 1. In this part,
Figure 4: Visual comparison results of ×4 image super-resolution on Urban100 datasets. SCN achieved qualitative results. More visual comparisons are shown in the supplementary materials.

Figure 5: SCN for JPEG compression artifact removal. From left to right, we show ground truth image, compressed image, and our restored image with scale-wise convolutional network.
Table 4: Ablation study on multi-scale network structures.

| Method       | PSNR  |
|--------------|-------|
| U-Net style  | 34.56 |
| PSPNet style | 34.58 |
| SCN          | 34.67 |

Table 5: Ablation study on parameter sharing across scales.

| Method     | # Params | PSNR  |
|------------|----------|-------|
| Baseline   | 1.2M     | 34.74 |
| Un-shared  | 1.2M     | 34.68 |
| Shared     | 1.2M     | 34.77 |
| Un-shared Large | 2.1M | 34.77 |

Table 6: Ablation study on re-sampling methods.

| Down-sample | Up-sample | PSNR  |
|-------------|-----------|-------|
| Conv        | Deconv    | 34.77 |
| AvgPool     | Nearest   | 34.78 |
| Bilinear    | Bilinear  | 34.80 |

Figure 6: Ablation study on number of scales and scaling ratios.

For resampling, there exists several methods: (1) down-sampling with strided convolution, up-sampling with strided deconvolution, (2) down-sampling with average pooling, up-sampling with nearest neighbor sampling and (3) down-sampling with bilinear sampling, up-sampling with bilinear sampling. We note that for the first two methods, the scaling ratio is limited to 2, 4, etc, because the kernel size have to be integer numbers. Our results in Table 5 indicate that bilinear has superior performance.

Scales in Scale-wise Convolution. In this part, we study the number of scales and the scaling ratios between neighborhood scales. From the results in Figure 6, we have several observations. (1) The scaling ratio plays an important role in the performance: a small scaling ratio (e.g., \(\frac{1}{2}\)) may lead to worse performance as the input feature map being too small, a large scaling ratio (e.g., comparing the curve of \(\frac{3}{4}\) and \(\frac{1}{2}\)) impedes the growth of receptive fields. (2) Multi-scale modeling is essential for image restoration and leads to much better performance compared with single-scale (34.74 in PSNR). Meanwhile more scales may lead to slightly inferior performance, which is likely due to the noise from the smallest scale. For example, in the setting of 6 scales with \(\frac{1}{2}\) scaling ratio, the smallest image has resolution \(\frac{1}{64}\) of the original resolution.

Figure 7: Ablation study on different number of scales for evaluation.

5 Conclusions

In this work we have presented scale-wise convolution that models scale-invariance in deep neural networks. The scale-wise convolutional network significantly improves the predictive accuracy on several image restoration datasets including image super-resolution, image denoising and image compression artifact removal. We also conducted experiments to proof the scale-invariance of proposed SCN and show its advances over other multi-scale models.
