




Pyramid Attention Network for Image Restoration

Yiqun Mei¹ · Yuchen Fan² · Yulun Zhang³ · Jiahui Yu⁴ · Yuqian Zhou⁵ · Ding Liu⁶ · Yun Fu⁷ · Thomas S. Huang⁸ · Humphrey Shi⁹ 

Received: 24 August 2022 / Accepted: 14 June 2023 / Published online: 8 August 2023
© The Author(s) 2023

Abstract

Self-similarity refers to the image prior widely used in image restoration algorithms that small but similar patterns tend to occur at different locations and scales. However, recent advanced deep convolutional neural network-based methods for image restoration do not take full advantage of self-similarities by relying on self-attention neural modules that only process information at the same scale. To solve this problem, we present a novel Pyramid Attention module for image restoration, which captures long-range feature correspondences from a multi-scale feature pyramid. Inspired by the fact that corruptions, such as noise or compression artifacts, drop drastically at coarser image scales, our attention module is designed to be able to *borrow* clean signals from their “clean” correspondences at the coarser levels. The proposed pyramid attention module is a generic building block that can be flexibly integrated into various neural architectures. Its effectiveness is validated through extensive experiments on multiple image restoration tasks: image denoising, demosaicing, compression artifact reduction, and super resolution. Without any bells and whistles, our PANet (pyramid attention module with simple network backbones) can produce state-of-the-art results with superior accuracy and visual quality. Our code is available at <https://github.com/SHI-Labs/Pyramid-Attention-Networks>

Keywords Image restoration · Image denoising · Demosaicing · Compression artifact reduction · Super-resolution

Communicated by Chen Change Loy.

✉ Humphrey Shi
shi@gatech.edu

Yiqun Mei
ymei7@jhu.edu

Yuchen Fan
fyc0624@gmail.com

Yulun Zhang
yulun100@gmail.com

Jiahui Yu
jiahuiyu@google.com

Yuqian Zhou
zhouyuqian133@gmail.com

Ding Liu
liudingdavy@gmail.com

Yun Fu
yunfu@ece.neu.edu

Thomas S. Huang
t-huang1@illinois.edu

¹ Johns Hopkins University, Baltimore, MD, USA

1 Introduction

Image restoration algorithms aim to recover a high-quality image from the contaminated counterpart, and is viewed as an ill-posed problem due to the irreversible degradation processes. They have many applications depending on the type of corruptions, for example, image denoising Zhang et al. (2017a, 2019); Liu et al. (2018), demosaicing Zhang et al. (2017b, 2019), compression artifacts reduction Dong et al. (2015); Chen and Pock (2017); Zhang et al. (2017a), super-resolution Kim et al. (2016); Lai et al. (2017); Tai et al. (2017) and many others Li et al. (2017); He et al. (2010); Chen et

² Meta Reality Labs, Menlo Park, CA, USA

³ ETH Zürich, Zürich, Switzerland

⁴ Google Brain, Bellevue, WA, USA

⁵ Adobe, Seattle, WA, USA

⁶ ByteDance, Mountain View, CA, USA

⁷ Northeastern University, Boston, MA, USA

⁸ UIUC, Urbana-Champaign, USA

⁹ Georgia Tech & UIUC & UO & PicsArt, Atlanta, GA, USA

al. (xxx). To restore missing information in a contaminated image, a variety of approaches based on leveraging image priors have been proposed Buades et al. (2005); Zontak et al. (xxx); Roth and Black (2005); Zoran and Weiss (2011).

Among these approaches, the prior of self-similarity in an image is widely explored and proved to be important. For example, non-local mean filtering Buades et al. (2005) uses self-similarity prior to reduce corruptions, which averages similar patches within the image. This notion of non-local pattern repetition was then extended to across multiple scales and demonstrated to be a strong property for natural images Zontak and Irani (2011); Glasner et al. (2009). Several self-similarity based approaches Glasner et al. (2009); Freedman and Fattal (2011); Singh and Ahuja (2014) were first proposed for image super-resolution, where they restore image details by borrowing high-frequency details from self-recurrences at larger scales. The idea was then explored in other restoration tasks. For example, in image denoising, its power is further strengthened by observing that noise reduces drastically at coarser scales Zontak et al. (xxx). This motivates many advanced approaches Zontak et al. (xxx); Michaeli and Irani (2014) to restore clean signals by finding “noise-free” recurrences in a built image-space pyramid, yielding high-quality reconstructions. The idea of utilizing multi-scale non-local prior has achieved great successes in various restoration tasks Bahat and Irani (2016); Zontak et al. (xxx); Michaeli and Irani (2014); Lotan and Irani (2016).

Recently deep neural networks trained for image restoration have made unprecedented progress. Following the importance of self-similarity prior, most recent approaches based on neural networks Zhang et al. (2019); Liu et al. (2018) adapt non-local operations into their networks, following the *non-local neural networks* Wang et al. (2018). In a non-local block, a response is calculated as a weighted sum over all pixel-wise features on the feature map, thus it can obtain long-range information. Such a module was initially designed for high-level recognition tasks and proven to be also effective in low-level vision problems Zhang et al. (2019); Liu et al. (2018).

However, these approaches which adapt the naive self-attention module to low-level tasks have certain limitations. First, to our best knowledge, multi-scale non-local prior is never explored. It has been demonstrated in the literature that cross-scale self-similarity can bring impressive benefits for image restoration Zontak et al. (xxx); Bahat and Irani (2016); Michaeli and Irani (2014); Glasner et al. (2009). Unlike high-level semantic features for recognition which makes not too much difference across scales, low-level features represent richer details, patterns, and textures at different scales. Nevertheless, the leading non-local self-attention fails to capture the useful correspondences that occur at different scales. Second, pixel-wise matching used in the self-attention module is usually noisy for low-level vision

tasks, thus reducing performance. Intuitively, enlarging the searching space raises possibility for finding better matches, but it is not true for the existing self-attention modules Liu et al. (2018). Unlike high-level feature maps where numerous dimension reduction operations are employed, image restoration networks often maintain the input spatial size. Therefore, feature is only highly relevant to a localized region, making them easily affected by noisy signals. This is in line with conventional non-local filtering, where pixel-wise matching performs much worse than block matching Buades et al. (2011).

In this paper, we present a novel non-local pyramid attention as a simple and generic building block for exhaustively capturing long-range dependencies, as shown in Fig. 1. The proposed attention takes full advantages of traditional non-local operations but is designed to better accord with the nature of image restoration. Specifically, the original search space is largely extended from a single feature map to a multi-scale feature pyramid. The proposed operation exhaustively evaluates correlation among features across multiple specified scales by searching over the entire pyramid. This brings several advantages: (1) It generalizes existing non-local operation, where the original searching space is inherently covered in the lowest pyramid level. (2) The long-range dependency between relevant features of different sizes is explicitly modeled. Since the operation is fully differentiable, it can be jointly optimized with networks through back propagation. (3) Similar to traditional approaches Zontak et al. (xxx); Bahat and Irani (2016); Michaeli and Irani (2014), one may expect noisy signals in features can be drastically reduced via rescaling to coarser pyramid level via operations like bi-cubic interpolation. This allows the network to find “clean signal” from multi-scale correspondences. Next, we enhance the robustness of correlation measurement by involving neighboring features into computation, inspired by traditional block matching strategy. Region-to-region matching imposes additional similarity constraints on the neighborhood. As such, the module can effectively single out highly relevant correspondences while suppressing noisy ones.

We demonstrate the power of non-local pyramid attention on various image restoration tasks: image denoising, image demosaicing, compression artifacts reduction and image super-resolution. In all tasks, a single pyramid attention, which is our basic unit, can model long-range dependency without scale restriction, in a feed forward manner. With one attention block inserted into a very simple backbone network, the model achieves significantly better results than the latest state-of-the-art approach with well-engineered architecture and multiple non-local attention units. In addition, we also conduct extensive ablation studies to analyze our design choices. All these evidences demonstrate our module is a better alternative of current non-local operation and can

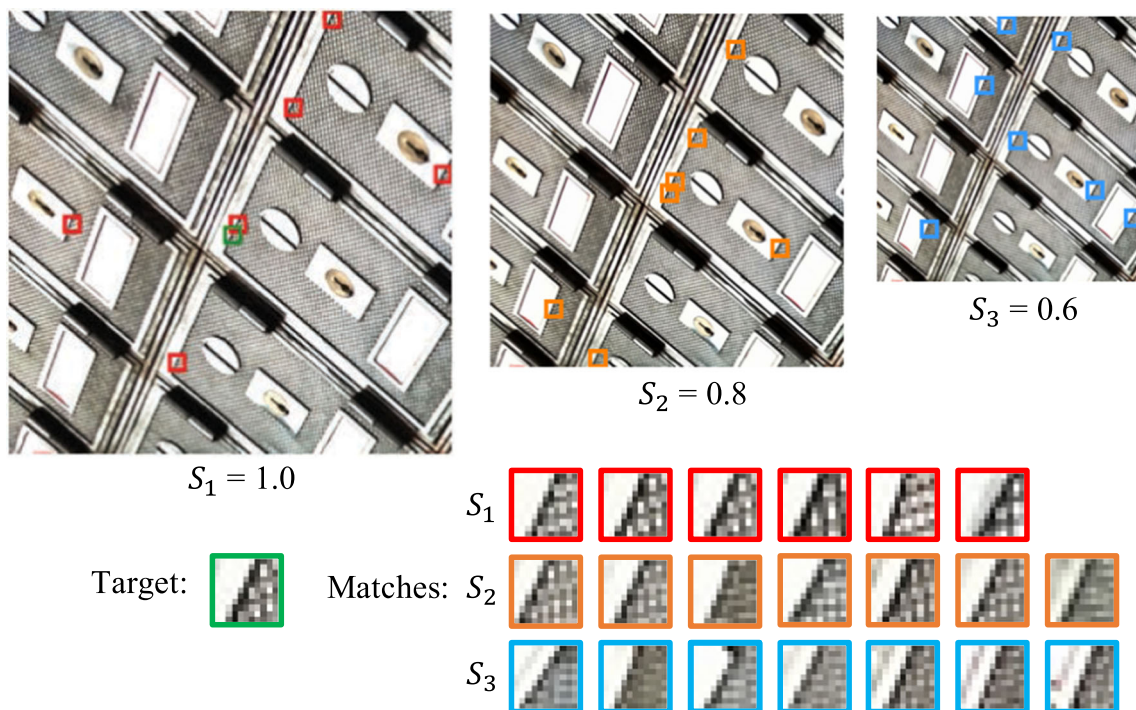


Fig. 1 Visualization of most correlated patches captured by our pyramid attention. Pyramid attention exploits multi-scale self-exemplars to improve reconstruction

be used as a fundamental unit in neural networks for generic image restoration.

In our previous conference version Mei et al. (2020), the original cross-scale non-local attention (CSNLA), although it makes some successful attempts, has limited ability to leverage cross-scale similarity. This is because the searching space is restricted to a scale specified by the SR task, and thus fails to fully utilize self-recurrences across multiple scales. On the other hand, it is designed as an upsampling operation, where it replaces a small patch (e.g. 3×3) with larger patches (e.g. 6×6) from the same feature map. This makes it inapplicable for general image restoration tasks, where the output image keeps the original resolution.

As discussed, our pyramid attention tackles these shortcomings by drawing inspirations from traditional self-similarity-based methods Zontak et al. (xxx); Michaeli and Irani (2014), where it has been demonstrated that corruptions can be effectively reduced by downscaling. Thus a cleaner image can be recovered by finding “noise-free” self-recurrences in a built image-space pyramid Zontak et al. (xxx). Inspired by these classical approaches, the proposed pyramid attention improves the searching range of CSNLA to a downscaled feature pyramid. This leaves the resolution unchanged while allowing the network to effectively leverage abundant multi-scale information, making it suitable for various image restoration tasks. As such, this work method-

ologically extends and significantly generalizes the previous conference version.

2 Related Works

2.1 Self-similarity Prior for Image Restoration

Self-similarity property that small patterns tend to recur within a image powers natural images with strong self-predictive ability Bahat and Irani (2016); Glasner et al. (2009); Zontak and Irani (2011), which forms a basis for many classical image restoration methods Zontak and Irani (2011); Zontak et al. (xxx); Bahat and Irani (2016); Michaeli and Irani (2014); Huang et al. (2015). The initial work, non-local mean filtering Buades et al. (2005), globally averages similar patches for image denoising. Later on, Dabov et al Dabov et al. (2007b) introduced BM3D, where repetitive patterns are grouped into 3D arrays to be jointly processed by collaborative filters. In LSSC Mairal et al. (2009), self-similarity property is combined with sparse dictionary learning for both denoising and demosaicing. This “fractal like” characteristic was further strengthened to **across different scales** and shown to be a very strong property for natural images Glasner et al. (2009); Zontak and Irani (2011). To enjoy cross-scale redundancy, self-similarity based approaches were proposed for image super-resolution

Glasner et al. (2009); Freedman and Fattal (2011); Huang et al. (2015), where high frequency information is retrieved uniquely from internal multi-scale recurrences. Observing that corruptions drop drastically at coarser scales, Zontak Zontak et al. (xxx) demonstrated that a clean version of noisy patches (99%) exists at coarser level of the original image. This idea was developed into their denoising algorithm, which achieved promising results. The cross-scale self similarity is also of central importance for many image deblurring Michaeli and Irani (2014); Bahat et al. (2017) and image dehazing approaches Bahat and Irani (2016).

2.2 Non-local Operation in Deep CNNs

Non-local operation in deep CNNs was initially proposed by Wang et al Wang et al. (2018) for video classification. In their networks, non-local units are placed on high-level, sub-sampled feature maps to compute long-range semantic correlations. By assigning weights to features at all locations, it allows the network to focus on more informative areas. Adapting non-local operation also showed considerable improvements in other high-level tasks, such as object detection Cao et al. (2019), semantic segmentation Fu et al. (2019) and person Re-id Xia et al. (2019). For image restoration, recent approaches, such as NLRN Liu et al. (2018), RNAN Zhang et al. (2019) and SAN Dai et al. (2019), incorporate non-local operations in their networks. However, without careful modification, their performances are limited by simple single-scale correlations and further reduced by involving many ill-matches during the pixel-wise feature matching in attention units.

Recently, CSNLN Mei et al. (2020) (the conference version) first extends non-local attention to model cross-scale correlation for image SR. The concurrent work IGNN Zhou et al. (2020) explores a similar idea but extracts cross-scale information in the LR image with a graph-based formulation. While being effective for image SR, existing cross-scale methods still suffer from certain limitations. First, they cannot benefit general image restoration tasks such as image denoising, compression artifacts reduction and demosaicing. This is because, by design, they are essentially upsample operations, where they replace a small patch (e.g. 3×3) with larger ones (e.g. 6×6). Moreover, the low-quality input image itself contains severe degradation and thus may not provide high-quality information to best facilitate image restoration, if directly utilize recurrences from the original feature map. In contrast, the proposed pyramid attention adopts a pyramid structure, where the downsampling operation can naturally reduce noise and corruption, a fact validated in many classical methods Zontak et al. (xxx); Michaeli and Irani (2014); Bahat and Irani (2016). By searching for clean patches of same size in a pyramid, our method effectively improves image restoration quality without changing the resolution. Further, they

have limited ability in exploring cross-scale self-similarity by restricting the search space to the single scale defined by the super-resolution task. On the other hand, it has been well-demonstrated that natural images are “fractal like” and small patches tend to repeatedly occur at multi-scales. The pyramid attention is designed to tackle these shortcomings by making full use of multi-scale image prior.

2.3 Deep CNNs for Image Restoration

Adopting deep-CNNs for image restoration has shown evident improvements by embracing their representative power. In the early work, Vincent et al Vincent et al. (2008) proposed to use stacked auto-encoder for image denoising. Later, ARCNN was introduced by Dong et al Dong et al. (2015) for compression artifacts reduction. Zhang et al Zhang et al. (2017a) proposed DnCNN for image denoising, which uses advanced techniques like residual learning and batch normalization to boost performance. In IRCNN Zhang et al. (2017b), a learned set of CNNs are used as denoising prior for other image restoration tasks. Recent extensive efforts have been spent into designing advanced architectures and learning methods, such as progressive structureLai et al. (2017); Zamir et al. (2021), residual Lim et al. (2017) and dense connection Zhang et al. (2018c); Liu et al. (2020), back-projection Haris et al. (2018), scale-invariant convolution Fan et al. (2019), channel Zhang et al. (2018b); Magid et al. (2021); Anwar and Barnes (2019); Zamir et al. (2020) and holistic attention Niu et al. (2020), look-up table Li et al. (2022); Jo and Kim (2021) context guided convolution Zhang et al. (2021b), structured pruning Zhang et al. (2021a), dropout mechanism Kong et al. (2022) and transformer models Liang et al. (2021); Zamir et al. (2022); Wang et al. (2022); Chen et al. (2021). We refer readers to recent literature survey for a more comprehensive review Anwar et al. (2020); Li et al. (2019); Tian et al. (2020). Most state-of-the-art approaches Liu et al. (2018); Zhang et al. (2019); Dai et al. (2019); Liang et al. (2021); Zamir et al. (2022); Wang et al. (2022) incorporate non-local attention into networks to boost representation ability. Although extensive efforts have been made in architectural engineering, existing methods relying on convolution and standard non-local operation can only exploit information at a same scale.

3 Pyramid Attention

Both convolution operation and non-local attention are restricted to same-scale information. In this section, we introduce the novel pyramid attention, which can deal with non-local dependency across multiple scales, as a generalization of non-local operations.

3.1 Formal Definition

Non-local attention calculates a response by averaging features over an entire image, as shown in Fig. 2a. Formally, given an input feature map x , this operation is defined as:

$$y^i = \frac{1}{\sigma(x)} \sum_j \phi(x^i, x^j)\theta(x^j), \tag{1}$$

where i, j are index on the input x and output y respectively. The function ϕ computes pair-wise affinity between two input features. θ is a feature transformation function that generates a new representation of x^j . The output response y^i obtains information from all features by explicitly summing over all positions and is normalized by a scalar function $\sigma(x)$. While the above operation manages to capture long-range correlation, information is extracted at a single scale. As a result, it fails to exploit relationships to many more informative areas of distinctive spatial sizes.

To break this scale constraint, we propose pyramid attention (Fig. 2c), which captures correlations across scales. In pyramid attention, affinities are computed between a target feature and regions. Therefore, a response feature is a weighted sum over multi-scale correspondences within the input map. Formally, given a series of scale factor $S = \{1, s_1, s_2, \dots, s_n\}$, pyramid attention can be expressed as

$$y^i = \frac{1}{\sigma(x)} \sum_{s \in S} \sum_j \phi(x^i, x_{\delta(s)}^j)\theta(x_{\delta(s)}^j). \tag{2}$$

Here $\delta(s)$ represents a s^2 neighborhood centred at index j on input x .

In other words, pyramid attention behaves in a non-local multi-scale way by explicitly processing larger regions with sizes specified by scale pyramid s at all position j . Note that when only a single scale factor $s = 1$ is specified, the proposed attention degrades to current non-local operation. Hence, our approach is a more generic operation that allows the network to fully enjoy the predictive power of natural images.

Finding a generic solution, which models cross-scale relationships, is a non-trivial problem and requires carefully engineering. In the following section, we first address the non-local operation between two scales and then extend it to pyramid scales.

3.2 Scale Agnostic Attention

Given an extra scale factor s , how to evaluate the correlation between x^j and $x_{\delta(s)}^j$ and aggregate information from $x_{\delta(s)}^j$ to form y^i are two key steps. Here, the major difficulty comes from misalignment in their spatial dimensions. Common sim-

ilarity measurements, such as dot product and embedded Gaussian, only accept features with identical dimensions, thus are infeasible in this case.

To mitigate the above problem, we propose to squeeze the spatial information of $x_{\delta(s)}^j$ into a single region descriptor. This step is conducted by down-scaling the region $x_{\delta(s)}^j$ in a pixel feature z^j . As we need search over the entire feature map, we can therefore directly down-scale the original input x ($H \times W$) to obtain a descriptor map z ($\frac{H}{s} \times \frac{W}{s}$). The correlation between x^i and $x_{\delta(s)}^j$ is then represented by x^i and the region descriptor z^j . Formally, scale agnostic attention (Fig. 2b) is formulated as

$$y^i = \frac{1}{\sigma(x, z)} \sum_j \phi(x^i, z^j)\theta(z^j), \tag{3}$$

where $z = x \downarrow s$.

This operation brings additional advantages. As discussed in Sect. 1, downscaling regions into coarser descriptors reduces noisy levels. On the other hand, since the cross-scale recurrence represents a similar content, the structure information will be still well-preserved after down-scaling. Combing these two facts, region descriptors can serve as a “cleaner version” of the target feature and a better alternative of noisy patch matches at the original scale.

3.3 Pyramid Attention

To make full use of self-predictive power, the scale agnostic attention can be extended to pyramid attention, which computes correlations across multiple scales. In such units, pixel-region correspondences are captured over an entire feature pyramid. Specifically, given a series of scales $S = \{1, s_1, s_2, \dots, s_n\}$, it forms a feature pyramid $\mathcal{F} = \{F_1, F_2, \dots, F_n\}$, where F_i ($\frac{H}{s_i} \times \frac{W}{s_i}$) is a region descriptor map of the input x , obtained by down-scaling operation. In such case, the correlations between any pyramid levels and the original input x can be seen as a scale agnostic attention. Therefore, the pyramid attention is defined as:

$$y^i = \frac{1}{\sigma(x, \mathcal{F})} \sum_{z \in \mathcal{F}} \sum_{j \in z} \phi(x^i, z^j)\theta(z^j). \tag{4}$$

The cross-scale modeling ability is due to the fact that region descriptor z_i at different levels summarizes information over regions of various sizes. When they are copied back to original position i , non-local multi-scale information is fused together to form a new response, which intuitively contains richer and more faithful information than the matches from a single scale.

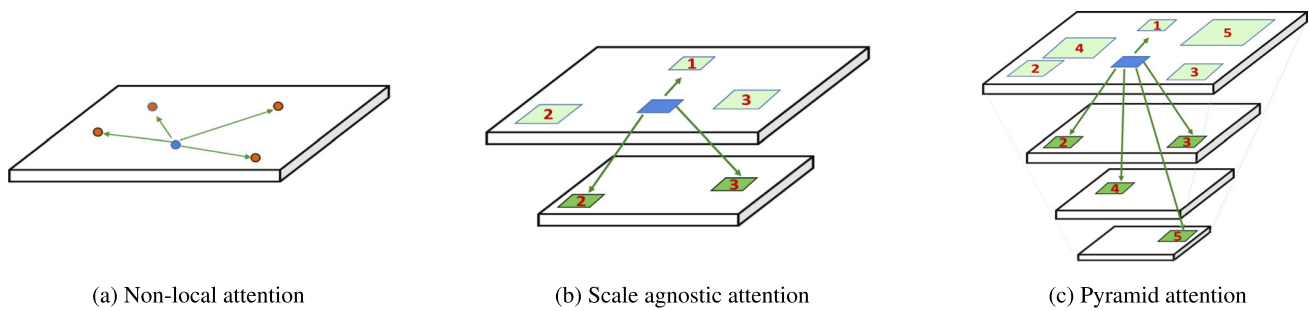


Fig. 2 Comparison of attentions. **a** Classic self-attention computes pair-wise feature correlation at scale. **b** Scale agnostic attention augments (a) to capture correspondences at one additional scale. **c** Pyramid attention generalizes (a) and (b) by modeling multi-scale non-local dependency

3.4 Instantiation

Choices of ϕ , θ and σ . There are many well-explored choices for pair-wise function ϕ Wang et al. (2018); Liu et al. (2018), such as Gaussian, embedded Gaussian, dot product and feature concatenation. In this paper, we use embedded Gaussian to follow previous best practices Liu et al. (2018): $\phi(x^i, z^j) = e^{f(x^i)^T g(z^j)}$, where $f(x^i) = W_f x^i$ and $g(z^j) = W_g z^j$.

For feature transformation function θ , we use a simple linear embedding: $\theta = W_\theta z^j$. Finally, we set $\sigma(x, \mathcal{F}) = \sum_{z \in \mathcal{F}} \sum_{j \in z} \phi(x^i, z^j)$. By specifying above instantiations, the term $\frac{1}{\sigma(x, \mathcal{F})} \sum_{x \in \mathcal{F}} \Phi(x^i, z^j)$ is equivalent to softmax over all possible positions in the pyramid.

Patch based region-to-region attention. As discussed in Sect. 1, information contained in features (for image restoration tasks) is very localized. Consequently, the matching process is usually affected by noisy signals. Previous approach relieves this problem by restriction search space to local region Liu et al. (2018). However, this also prevents them from finding better correspondences that are far away from current position.

To improve the robustness during matching, we impose extra neighborhood similarity, which is in line with classical non-local filtering Buades et al. (2005). As such, the pyramid attention (Eq. 3) is expressed as:

$$y^i = \frac{1}{\sigma(x, \mathcal{F})} \sum_{z \in \mathcal{F}} \sum_{j \in z} \phi(x_{\delta(r)}^i, z_{\delta(r)}^j) \theta(z^j), \quad (5)$$

where the neighborhood is specified by $\delta(r)$. This adds a stronger constraint on matching content that two features are highly correlated if and only if their neighborhood are highly similar as well. The block-wise matching allows the network to pay more attention on relevant areas while suppressing unrelated ones.

Implementation. The proposed pyramid attention is implemented using basic convolution and deconvolution operations, as shown in Fig. 3. According to Eq. 5, the pyramid attention is equivalent to first compute the S-A attention at each scale with the original feature map and then fuse the results later. The inner summation $\sum_{j \in z} \phi(x_{\delta(r)}^i, z_{\delta(r)}^j) \theta(z^j)$ corresponds to the scale agnostic attention between a down-scaled feature map $z \in \mathcal{F}$ and the original feature map x . The results are then aggregated over z and normalized by σ . Matching scores can be expressed as convolution over the input x using $r \times r$ patches extracted from the feature pyramid, followed by a softmax. To obtain a final response, we extract patches from the transformed feature map (by θ) to conduct a deconvolution over the matching score. Note that the proposed operation is fully convolutional, differentiable and accepts any input resolutions, which can be flexibly embedded into many standard architectures.

3.5 PANet: Pyramid Attention Networks

To show the effectiveness of our pyramid attention, we choose a simple ResNet as our backbone without any architectural engineering. The proposed image restoration network is illustrated in Fig. 3. We remove batch normalization in each residual block, following the practice in Lim et al. (2017). Similar to many restoration networks, we add a global pathway from the first feature to the last one, which encourages the network bypass low frequency information. We insert a single pyramid attention in the middle of the network.

Given a set of N paired images $I_{LQ}^k - I_{HQ}^k$, we optimize the $L1$ loss between I_{HQ}^k and recovered image I_{RQ}^k ,

$$\mathcal{L}_1 = \frac{1}{N} \sum_{i=1}^N \|I_{RQ}^k - I_{HQ}^k\|, \quad (6)$$

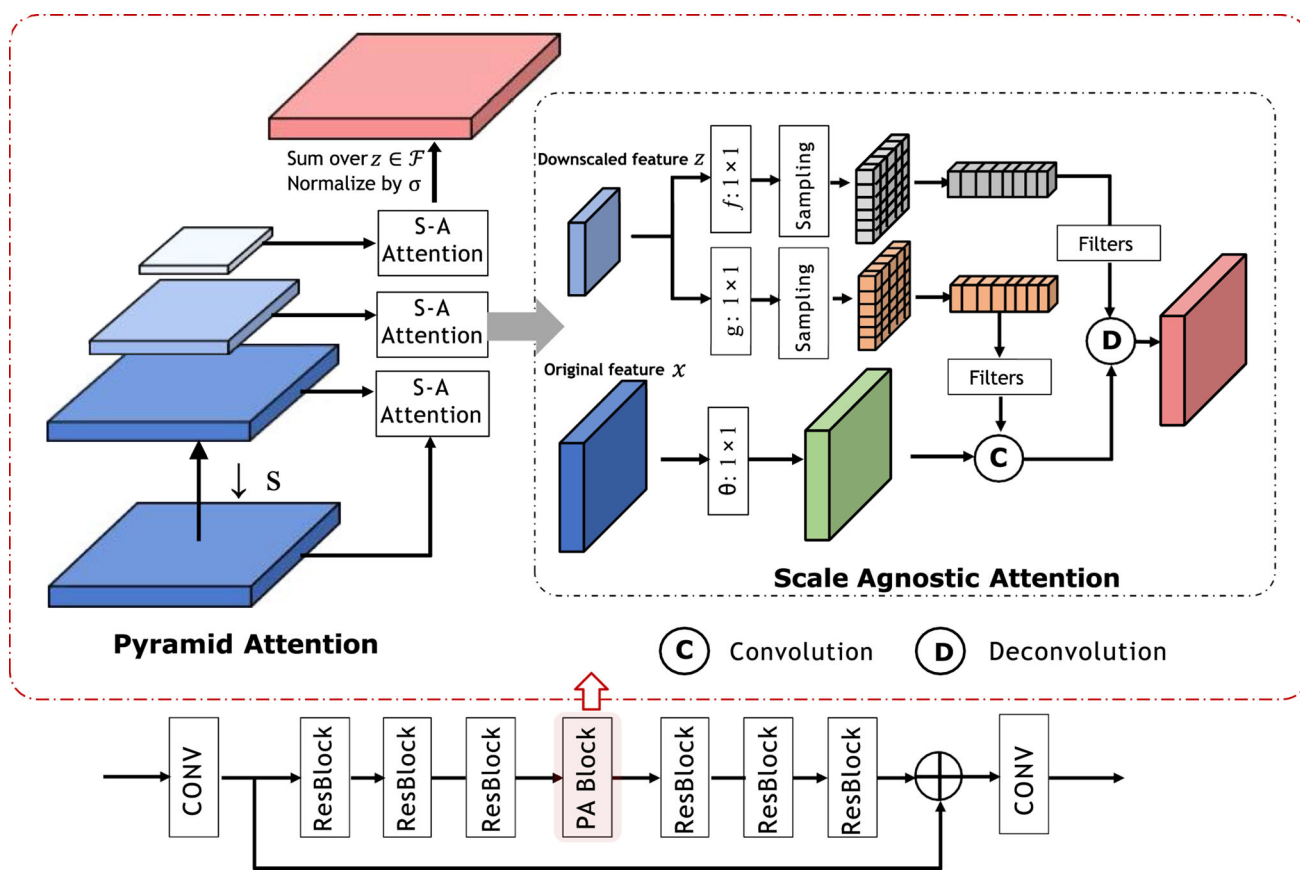


Fig. 3 PANet with the proposed pyramid attention (PA). Pyramid attention captures multi-scale correlation by computing Scale Agnostic (S-A) attention at each scale with the original feature map x (which corresponds to the inner summation $\sum_{j \in z} \phi(x_{\delta(r)}^j, z_{\delta(r)}^j) \theta(z^j)$ in eq. 5), and fusing the results over $z \in \mathcal{F}$

4 Experiments

4.1 Datasets and Evaluation Metrics

The proposed pyramid attention and PANet are evaluated on major image restoration tasks: image denoising, demosaicing and compression artifacts reduction and super-resolution. For fair comparison, we follow the setting specified by RNAN Zhang et al. (2019) for image denoising, demosaicing, and compression artifacts reduction. We use DIV2K Timofte et al. (2017) as our training set, which contains 800 high quality images. We report results on standard benchmarks using PSNR and/or SSIM Wang et al. (2004).

4.2 Implementation Details

For pyramid attention, we set the scale factors $S = \{1.0, 0.9, 0.8, 0.7, 0.6\}$, so that we construct a 5 level feature pyramid within the attention block. To build the pyramid, we use simple bi-cubic interpolation to rescale feature maps. While computing correlations, we use 3×3 small patches centered at target features. For fair comparison with repre-

sentative non-local approach RNAN, we adopt a backbone similar to theirs, but remove all engineered designs such as multi-scale and multi-branch, resulting in a plain ResNet. The proposed PANet contains 80 residual blocks with one pyramid attention module inserted after the 40-th block. All features have 64 channels, except for those used in embedded Gaussian, where the channel number is reduced to 32.

During training, each mini-batch consists of 16 patches with size 48×48 . We augment training images using vertical/horizontal flipping and random rotation of 90° , 180° , and 270° . The model is optimized by Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rate is initialized to 10^{-4} and reduced to a half after every 200 epochs. Our model is implemented using PyTorch Paszke et al. (2017) and trained on Nvidia TITANX GPUs.

4.3 Image Denoising

Following RNAN Zhang et al. (2019), PANet is evaluated on standard benchmarks for image denoising: Kodak24 (<http://r0k.us/graphics/kodak/>), BSD68 Martin et al. (2001), and Urban100 Huang et al. (2015). We create noisy images

Table 1 Quantitative evaluation of state-of-the-art approaches on color image denoising

Method	Kodak24				BSD68				Urban100			
	10	30	50	70	10	30	50	70	10	30	50	70
CBM3D	36.57	30.89	28.63	27.27	35.91	29.73	27.38	26.00	36.00	30.36	27.94	26.31
TNRD	34.33	28.83	27.17	24.94	33.36	27.64	25.96	23.83	33.60	27.40	25.52	22.63
RED	34.91	29.71	27.62	26.36	33.89	28.46	26.35	25.09	34.59	29.02	26.40	24.74
DnCNN	36.98	31.39	29.16	27.64	36.31	30.40	28.01	26.56	36.21	30.28	28.16	26.17
MemNet	N/A	29.67	27.65	26.40	N/A	28.39	26.33	25.08	N/A	28.93	26.53	24.93
IRCNN	36.70	31.24	28.93	N/A	36.06	30.22	27.86	N/A	35.81	30.28	27.69	N/A
FFDNet	<u>36.81</u>	31.39	29.10	27.68	36.14	30.31	27.96	26.53	35.77	30.53	28.05	26.39
RNAN	<u>37.24</u>	<u>31.86</u>	<u>29.58</u>	<u>28.16</u>	<u>36.43</u>	<u>30.63</u>	<u>28.27</u>	<u>26.83</u>	<u>36.59</u>	<u>31.50</u>	<u>29.08</u>	<u>27.45</u>
PANet	37.35	31.96	29.65	28.20	36.50	30.70	28.33	26.89	36.80	31.87	29.47	27.87

Best results are highlighted in bold

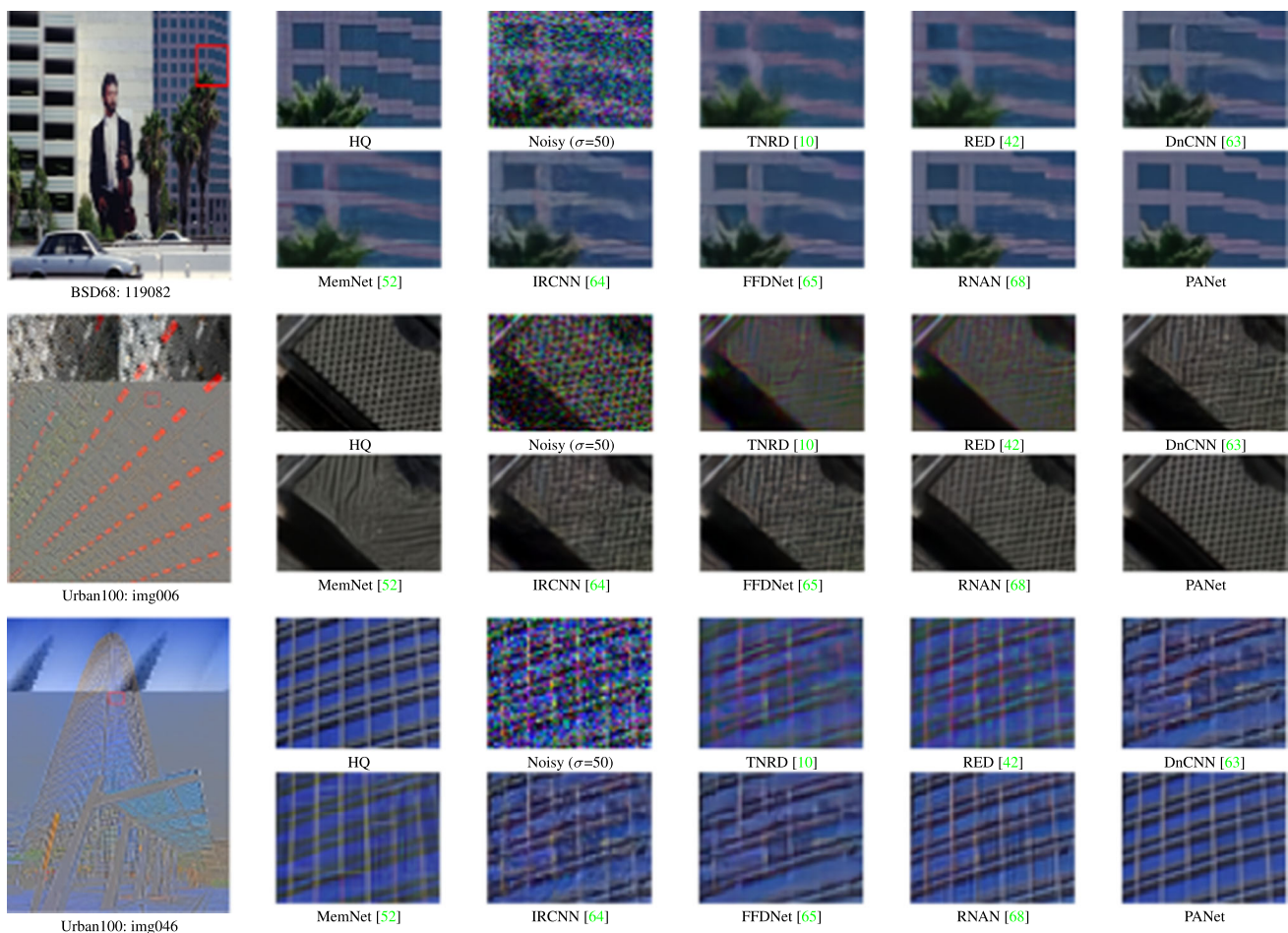
**Fig. 4** Visual comparison for color image denoising with noise level $\sigma = 50$

Table 2 Quantitative evaluation of state-of-the-art approaches on color image demosaicing

Method	McMaster18		Kodak24		BSD68		Urban100	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Mosaiced	9.17	0.1674	8.56	0.0682	8.43	0.0850	7.48	0.1195
IRCNN	37.47	0.9615	40.41	0.9807	39.96	0.9850	36.64	0.9743
RNAN	<u>39.71</u>	<u>0.9725</u>	<u>43.09</u>	<u>0.9902</u>	<u>42.50</u>	<u>0.9929</u>	<u>39.75</u>	<u>0.9848</u>
PANet	40.00	0.9737	43.29	0.9905	42.86	0.9933	40.50	0.9854

Best results are highlighted in bold

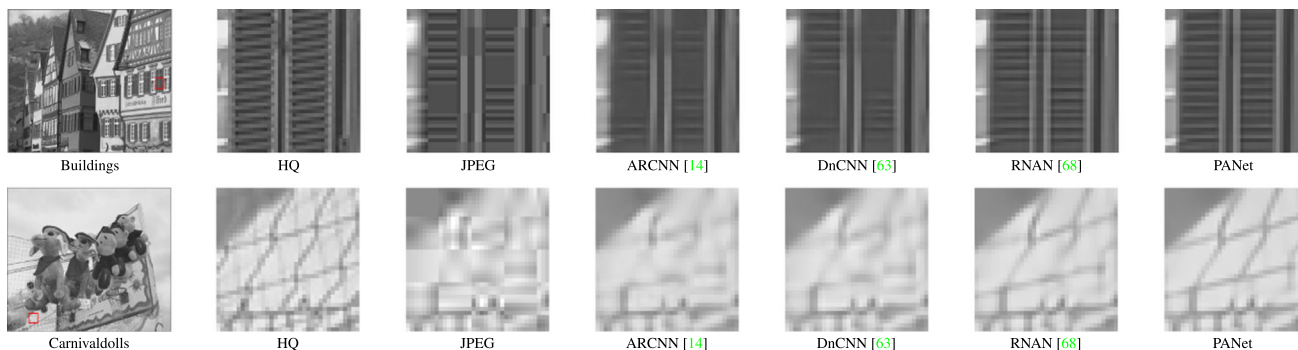


Fig. 5 Visual comparison for image CAR with JPEG quality $q = 10$

by adding AWGN noises with $\sigma = 10, 30, 50, 70$. We compare our approach with 8 state-of-the-art methods: CBM3D Dabov et al. (2007a), TNRD Chen and Pock (2017), RED Mao et al. (2016), DnCNN Zhang et al. (2017a), MemNet Tai et al. (2017), IRCNN Zhang et al. (2017b), FFDNet Zhang et al. (2017c), and RNAN Zhang et al. (2019).

As shown in Table 1, PANet achieved best performance on all datasets and noise levels. Our method surpassed FFDNet by around 0.6dB, 0.4dB and 1.3dB on three benchmarks respectively. PANet also yielded better results than prior state-of-the-art RNAN, which has well-engineered network and multiple non-local attention blocks. These results demonstrate that pyramid attention is indeed useful for image restoration. A single pyramid attention can drive the fair simple backbone to the state-of-the-art. One may further notice that PANet performs significantly well on Urban100 dataset, with more than 0.3 dB improvements over RNAN on all noise levels. This is because pyramid attention allows the network to explicitly capture abundant cross-scale self-exemplars in urban scenes. In contrast, traditional non-local attention, even with a multi-scale network structure, fails to explore those multi-scale relationships.

We further present qualitative evaluations on BSD68 and Urban100. The results are shown in Fig. 4. TNRD, RED, DnCNN and IRCNN cannot remove the noise pattern and create blur artifacts over high-frequency patterns. FFDNet and RNAN are able to reconstruct a clearer image but fail to recover the underlying textures. In contrast, by relying on a

single learned pyramid attention, PANet managed to produce the most accurate and faithful restoration results.

4.4 Image Demosaicing

For image demosaicing, we conduct evaluations on Kodak24, McMaster Zhang et al. (2017b), BSD68, and Urban100, following settings in RNAN Zhang et al. (2019). We compare our approach with recent state-of-the-arts IRCNN Zhang et al. (2017b) and RNAN Zhang et al. (2019). As shown in Table 2, mosaic corruption significantly reduced image quality in terms of PSNR and SSIM. RNAN and IRCNN can remove these corruptions to some degree and lead to relatively high-quality restoration. Our approach yields the best reconstruction, outperforming RNAN by 0.3dB, 0.2dB, 0.3dB and 0.7dB on four datasets respectively. These demonstrate advantages of exploiting multi-scale correlations.

4.5 Image Compression Artifacts Reduction

For image compression artifacts reduction (CAR), we compare our method with 5 most recent approaches: SADCCT Foi et al. (2007), ARCNN Dong et al. (2015), TNRD Chen and Pock (2017), DnCNN Zhang et al. (2017a), and RNAN Zhang et al. (2019). We present results on LIVE1 Sheikh et al. (2005) and Classic5 Foi et al. (2007), following the same settings in RNAN. To obtain the low-quality compressed images, we follow the standard JPEG compression process and use Matlab JPEG encoder with

Table 3 Quantitative evaluation of state-of-the-art approaches on compression artifacts reduction

Dataset	q	JPEG		SA-DCT		ARCNN		TNRD		DnCNN		RNAN		PANet	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LIVE1	10	27.77	0.7905	28.65	0.8093	28.98	0.8217	29.15	0.8111	29.19	0.8123	29.63	0.8239	29.69	0.8250
	20	30.07	0.8683	30.81	0.8781	31.29	0.8871	31.46	0.8769	31.59	0.8802	32.03	0.8877	32.10	0.8885
	30	31.41	0.9000	32.08	0.9078	32.69	0.9166	32.84	0.9059	32.98	0.9090	33.45	0.9149	33.55	0.9157
	40	32.35	0.9173	32.99	0.9240	33.63	0.9306	N/A	N/A	33.96	0.9247	34.47	0.9299	34.55	0.9305
Classic5	10	27.82	0.7800	28.88	0.8071	29.04	0.8111	29.28	0.7992	29.40	0.8026	29.96	0.8178	30.03	0.8195
	20	30.12	0.8541	30.92	0.8663	31.16	0.8694	31.47	0.8576	31.63	0.8610	32.11	0.8693	32.36	0.8712
	30	31.48	0.8844	32.14	0.8914	32.52	0.8967	32.78	0.8837	32.91	0.8861	33.38	0.8924	33.53	0.8939
	40	32.43	0.9011	33.00	0.9055	33.34	0.9101	N/A	N/A	33.77	0.9003	34.27	0.9061	34.38	0.9068

Best results are highlighted in bold

Table 4 Model size comparison

Methods	RED	DnCNN	MemNet	RNAN(1LB1NL)	RNAN	PANet-S	PANet
Parameters	4131K	672K	677K	1494K	7409K	665K	5957K
PSNR (dB)	26.40	28.16	26.53	28.36	29.15	28.80	29.47

Table 5 Quantitative results on SR benchmark datasets

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LapSRN Lai et al. (2017)	×2	37.52	0.9591	33.08	0.9130	31.08	0.8950	30.41	0.9101	37.27	0.9740
MemNet Tai et al. (2017)	×2	37.78	0.9597	33.28	0.9142	32.08	0.8978	31.31	0.9195	37.72	0.9740
SRMDNF Zhang et al. (2018a)	×2	37.79	0.9601	33.32	0.9159	32.05	0.8985	31.33	0.9204	38.07	0.9761
DBPN Haris et al. (2018)	×2	38.09	0.9600	33.85	0.9190	32.27	0.9000	32.55	0.9324	38.89	0.9775
RDN Zhang et al. (2018c)	×2	38.24	0.9614	34.01	0.9212	32.34	0.9017	32.89	0.9353	39.18	0.9780
RCAN Zhang et al. (2018b)	×2	38.27	0.9614	<u>34.12</u>	0.9216	32.41	0.9027	<u>33.34</u>	<u>0.9384</u>	39.44	<u>0.9786</u>
NLRN Liu et al. (2018)	×2	38.00	0.9603	33.46	0.9159	32.19	0.8992	31.81	0.9249	–	–
SRFBN Li et al. (2019)	×2	38.11	0.9609	33.82	0.9196	32.29	0.9010	32.62	0.9328	39.08	0.9779
OISR He et al. (2019)	×2	38.21	0.9612	33.94	0.9206	32.36	0.9019	33.03	0.9365	–	–
SAN Dai et al. (2019)	×2	<u>38.31</u>	0.9620	34.07	0.9213	<u>32.42</u>	0.9028	33.10	0.9370	39.32	0.9792
IGNN Zhou et al. (2020)	×2	38.24	0.9613	34.07	<u>0.9217</u>	32.41	0.9025	33.23	0.9383	39.35	<u>0.9786</u>
NSR Fan et al. (2020)	×2	38.23	0.9614	33.94	0.9203	32.34	0.9020	33.02	0.9367	39.31	0.9782
EDSR Lim et al. (2017)	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
PA-EDSR (ours)	×2	38.33	<u>0.9617</u>	34.22	0.9224	32.42	<u>0.9027</u>	33.38	0.9392	<u>39.37</u>	0.9782
LapSRN Lai et al. (2017)	×3	33.82	0.9227	29.87	0.8320	28.82	0.7980	27.07	0.8280	32.21	0.9350
MemNet Tai et al. (2017)	×3	34.09	0.9248	30.00	0.8350	28.96	0.8001	27.56	0.8376	32.51	0.9369
SRMDNF Zhang et al. (2018a)	×3	34.12	0.9254	30.04	0.8382	28.97	0.8025	27.57	0.8398	33.00	0.9403
RDN Zhang et al. (2018c)	×3	34.71	0.9296	30.57	0.8468	29.26	0.8093	28.80	0.8653	34.13	0.9484
RCAN Zhang et al. (2018b)	×3	34.74	0.9299	30.65	<u>0.8482</u>	29.32	0.8111	<u>29.09</u>	<u>0.8702</u>	<u>34.44</u>	<u>0.9499</u>
NLRN Liu et al. (2018)	×3	34.27	0.9266	30.16	0.8374	29.06	0.8026	27.93	0.8453	–	–
SRFBN Li et al. (2019)	×3	34.70	0.9292	30.51	0.8461	29.24	0.8084	28.73	0.8641	34.18	0.9481
OISR He et al. (2019)	×3	34.72	0.9297	30.57	0.8470	29.29	0.8103	28.95	0.8680	–	–
SAN Dai et al. (2019)	×3	<u>34.75</u>	<u>0.9300</u>	30.59	0.8476	<u>29.33</u>	<u>0.8112</u>	28.93	0.8671	34.30	0.9494
IGNN Zhou et al. (2020)	×3	34.72	0.9298	<u>30.66</u>	<u>0.8484</u>	29.31	0.8105	29.03	0.8696	34.39	0.9496
NSR Fan et al. (2020)	×3	34.62	0.9289	30.57	0.8475	29.26	0.8100	28.83	0.8663	34.27	0.9484
EDSR Lim et al. (2017)	×3	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
PA-EDSR (ours)	×3	34.84	0.9306	30.71	0.8488	29.33	0.8119	29.24	0.8736	34.46	0.9505
LapSRN Lai et al. (2017)	×4	31.54	0.8850	28.19	0.7720	27.32	0.7270	25.21	0.7560	29.09	0.8900
MemNet Tai et al. (2017)	×4	31.74	0.8893	28.26	0.7723	27.40	0.7281	25.50	0.7630	29.42	0.8942
SRMDNF Zhang et al. (2018a)	×4	31.96	0.8925	28.35	0.7787	27.49	0.7337	25.68	0.7731	30.09	0.9024
DBPN Haris et al. (2018)	×4	32.47	0.8980	28.82	0.7860	27.72	0.7400	26.38	0.7946	30.91	0.9137
RDN Zhang et al. (2018c)	×4	32.47	0.8990	28.81	0.7871	27.72	0.7419	26.61	0.8028	31.00	0.9151
RCAN Zhang et al. (2018b)	×4	32.63	0.9002	28.87	0.7889	<u>27.77</u>	<u>0.7436</u>	<u>26.82</u>	<u>0.8087</u>	<u>31.22</u>	<u>0.9173</u>
NLRN Liu et al. (2018)	×4	31.92	0.8916	28.36	0.7745	27.48	0.7306	25.79	0.7729	–	–
SRFBN Li et al. (2019)	×4	32.47	0.8983	28.81	0.7868	27.72	0.7409	26.60	0.8015	31.15	0.9160
OISR He et al. (2019)	×4	32.53	0.8992	28.86	0.7878	27.75	0.7428	26.79	0.8068	–	–
SAN Dai et al. (2019)	×4	<u>32.64</u>	<u>0.9003</u>	28.92	0.7888	27.78	<u>0.7436</u>	26.79	0.8068	31.18	0.9169
IGNN Zhou et al. (2020)	×4	32.57	0.8998	28.85	0.7891	<u>27.77</u>	0.7434	26.84	0.8090	<u>31.28</u>	<u>0.9182</u>
NSR Fan et al. (2020)	×4	32.55	0.8987	28.79	0.7876	27.72	0.7414	26.61	0.8025	31.10	0.9145
EDSR Lim et al. (2017)	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
PA-EDSR (ours)	×4	32.65	0.9006	<u>28.87</u>	0.7891	27.76	0.7445	27.01	0.8140	31.29	0.9194

Table 6 Benchmark results with **BD** and **DN** degradation models. Average PSNR/SSIM values for scaling factor $\times 3$

Dataset	Model	Bicubic	SRCNN Dong et al. (2014)	FSRCNN Dong et al. (2016)	VDSR Kim et al. (2016)	IRCNN Zhang et al. (2017b)	RDN Zhang et al. (2018c)	RCAN Zhang et al. (2018b)	PA-EDSR (ours)
Set5	BD	28.78/0.8308	32.05/0.8944	26.23/0.8124	33.25/0.9150	33.38/0.9182	34.58/0.9280	34.70/0.9288	34.82/0.9298
	DN	24.01/0.5369	25.01/0.6950	24.18/0.6932	25.20/0.7183	25.70/0.7379	28.47/0.8151	-/-	28.62/0.8194
Set14	BD	26.38/0.7271	28.80/0.8074	24.44/0.7106	29.46/0.8244	29.63/0.8281	30.53/0.8447	30.63/0.8462	30.77/0.8478
	DN	22.87/0.4724	23.78/0.5898	23.02/0.5856	24.00/0.6112	24.45/0.6305	26.60/0.7101	-/-	26.69/0.7148
B100	BD	26.33/0.6918	28.13/0.7736	24.86/0.6832	28.57/0.7893	28.65/0.7922	29.23/0.8079	29.32/0.8093	29.36/0.8113
	DN	22.92/0.4449	23.76/0.5538	23.41/0.5556	24.00/0.5749	24.28/0.5900	25.93/0.6573	-/-	25.99/0.6623
Urban100	BD	23.52/0.6862	25.70/0.7770	22.04/0.6745	26.61/0.8136	26.77/0.8154	28.46/0.8582	28.81/0.8647	29.15/0.8706
	DN	21.63/0.4687	21.90/0.5737	21.15/0.5682	22.22/0.6096	22.90/0.6429	24.92/0.7364	-/-	25.34/0.7522
Manga109	BD	25.46/0.8149	29.47/0.8924	23.04/0.7927	31.06/0.9234	31.15/0.9245	33.97/0.9465	34.38/0.9483	34.56/0.9500
	DN	23.01/0.5381	23.75/0.7148	22.39/0.7111	24.20/0.7525	24.88/0.7765	28.00/0.8591	-/-	28.24/0.8655

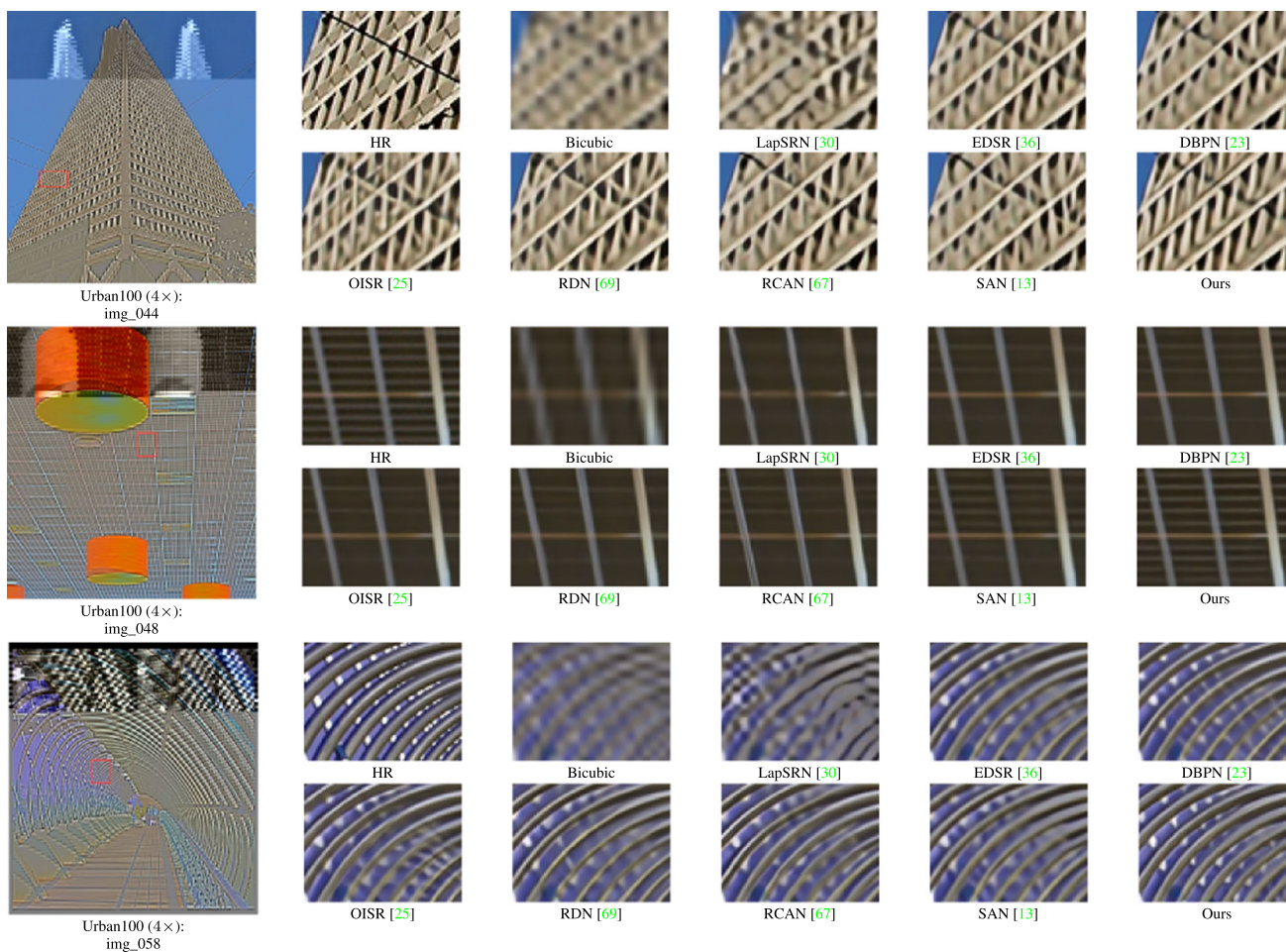


Fig. 6 Visual comparison for 4× SR on Urban100 dataset

quality $q = 10, 20, 30, 40$. For fair comparison, the results are only evaluated on Y channel in YCbCr Space.

The quantitative evaluation are reported in Table 3. By incorporating pyramid attention, PANet obtains best results on both LIVE1 and Classic5 with all quality levels. For example, on Classic5 and with $q = 20$, our approach achieves around 0.25dB and 0.73dB gains over RNAN and DnCNN respectively. Similar improvements can also be observed when comparing with other methods. These results shows the effectiveness of the proposed pyramid attention.

We further present visual comparisons on the most challenging quality level $q = 10$ in Fig. 5. One can see that the proposed approach successfully reduced compression artifacts and recovered the most image details. This is mainly because our PANet captures non-local relationships in a multi-scale way, helping to reconstruct more faithful details.

4.6 Model Size Analyses

We report our model size and compare it with other advanced image denoising approaches in Table 4. To compare with

light weight models, we also bulid a small PANet-S with only 8 residual blocks. One can see that PANet achieves the best performance with a lighter and much simpler architecture, as compared to the prior state-of-the-art approach RNAN. Similarly, PANet-S significantly outperforms other light weight models using only less than 50% parameters of RNAN (1LB+1NLB). Such observations indicate the great advantages brought by our pyramid attention module. In practice, our proposed pyramid attention module can be inserted in related networks.

4.7 Image Super Resolution

To further demonstrate the generality of pyramid attention, we present image super-resolution experiments. Here we follow Zhang et al. (2018c) and consider three different degradation models to simulate LR images. For **BI** model, we use bi-cubic downsampling to create LR images with scale factor $\times 2, \times 3$ and $\times 4$, by leveraging *matlab imresize* function. For **BD** setting, LR images are created by filtering HR images with a Gaussian blur kernel of size 7×7 before

downsampling. For **DN** model, images are first downsampled and then Gaussian noise with $\sigma = 30$ is added to the LR images. We evaluate methods with scale factor $\times 3$ for **BD** and **DN** settings.

To better show the effectiveness of the proposed pyramid attention, we choose EDSR Lim et al. (2017), the simplest network structure consisting of residual blocks and convolutions only, as our backbone. A single pyramid attention block is inserted after 16th residual block (denote as PA-EDSR). Network-level designs, such as back-projection (DBPN), dense connection (RDN) and channel attention (RCAN and SAN), which are perpendicular to our method, can also be easily combined with pyramid attention for superior performance.

4.7.1 Comparison with BI Degradation Model

We compare it with 11 state-of-the-art approaches: LapSRN Lai et al. (2017), MemNet Tai et al. (2017), SRMDNF Zhang et al. (2017b), EDSR Lim et al. (2017), DBPN Haris et al. (2018), RDN Zhang et al. (2018c), RCAN Zhang et al. (2018b), NLRN Liu et al. (2018), SRFBN Li et al. (2019), OISR He et al. (2019), SAN Dai et al. (2019), IGNN Zhou et al. (2020) and NSR Fan et al. (2020).

We report experiment results in Table 5. Without any architectural engineering, our simple PA-EDSR achieves best performance on almost all benchmarks and scales. In particular, our method outperforms a concurrent work IGNN on almost all entries, which is also built upon EDSR and replying on one cross-scale module to improve performances, indicating our design can make better use of self-similarity information. With a single pyramid attention, PA-EDSR also shows huge advantages over NLRN, which is the first non-local based approach for image SR and contains 12 standard non-local operations. It is worth noting that SAN is a very competitive approach, which contains multiple standard non-local attentions and more than 200 residual blocks, i.e., $\times 7$ deeper than ours. Even in this case, PA-EDSR still shows superior results on almost all entries. These results demonstrate the effectiveness of the proposed pyramid attention. When comparing with EDSR backbone, one can see that the additional pyramid attention brings constant improvements on all datasets, especially on Urban100 (0.4dB) and Manga109 (0.3dB). This is because images in these datasets contain abundant structural recurrences, such as edges and corners, which can more benefits from exploring cross-scale internal hints. We also observed considerable performance gains on natural image datasets: Set5 (0.2dB), Set14 (0.2dB) and B100 (0.1dB). This is accorded with previous observation that cross-scale self-recurrence is a common property for natural images Glasner et al. (2009). We claim that cross-scale intrinsic priors are indeed effective for a more faithful reconstruction.

Visual results are shown in Fig. 6. Our method perceptually outperforms other state-of-the-arts by a large margin. For these repeated high-frequency structures, PA-EDSR yields the most accurate reconstruction. In contrast, SAN with standard NL attention fails to handle these cases. This demonstrates that exploring internal HR hints from multi-scale self-recurrences indeed leads to a better local recovery.

4.7.2 Comparison with BD and DN Degradation Models

Following Zhang et al. (2018c), we report our results with **BD** and **DN** degradation models and compare it with SRCNN Dong et al. (2014), FSRCNN Dong et al. (2016), VDSR Kim et al. (2016), IRCNN Zhang et al. (2018a), RDN Zhang et al. (2018c) and RCAN Zhang et al. (2018b). Average PSNR and SSIM results on 5 benchmarks with scale factor $\times 3$ are shown in Table 6. Our method achieves the best performances for all entries. The constant performance gains over other methods indicate the proposed pyramid attention is indeed robust and powerful for **BD** and **DN** degradation models.

4.7.3 Performance on Lightweight Backbone

To better study the effectiveness of the proposed pyramid attention, we built a smaller model PANet-S with a model size comparable to DnCNN Zhang et al. (2017a). Specifically, PANet-S contains 8 ResBlocks with a channel number of 64. We insert one pyramid attention module after the 4-th blocks. As shown in Table 7, PANet-S achieves the best performances on all datasets with the smallest model size, demonstrating the performance of pyramid attention can be well-preserved on the lightweight backbone. Moreover, when comparing with RNAN(1NL), it can be seen that PANet-S with less than half of the parameters can still maintain 0.4dB performance gain. This shows that exploring of multi-scale similarity with pyramid attention is indeed more beneficial.

4.7.4 Performance on Lightweight Blind Image Denoising

To better demonstrate the effectiveness of our method, we conduct an experiment on blind image denoising with a lightweight backbone. We use the same training and test protocol in DnCNN and compare PANet-S with it. PANet-S has a similar number of parameters so that allows fair comparison. Results are reported in Table 8. The pretrained blind DnCNN model is derived from the official GitHub repository. One can see that our method outperforms DnCNN by a large margin, proving its effectiveness on blind image denoising.

4.7.5 Efficiency and Performance Analysis

Here we present the efficiency (FLOPs, running time, and peak memory consumption) and performance comparison

Table 7 Quantitative comparison on lightweight backbone

$\sigma = 50$ Size	DnCNN 672K	MemNet 677K	RNAN (1NL) 1494K	PANet-S (1PA) 655K
Kodak	29.16	27.65	–	29.38
CBSD68	28.01	26.33	–	28.15
Urban100	28.16	26.53	28.36	28.80

Table 8 Quantitative comparison on blind image denoising

Dataset	Kodak		CBSD68		Urban100	
	30	50	30	50	30	50
DnCNN	31.28	28.96	30.34	27.95	30.00	27.59
PANet-S	31.67	29.31	30.50	28.11	31.18	28.69

(on Set14 $\times 2$) with prior state-of-the-art SAN Dai et al. (2019) and our conference version CSNLN Mei et al. (2020). For efficiency comparison, models are evaluated at input size 100×100 . The running time is the average of 1K times on a single Nvidia RTX 2070 GPU.

We report results in Table 9. One may notice that PA-EDSR is considerably more efficient than its previous version CSNLN, with significant reductions in terms of running time, memory consumption and FLOPs. Specifically, PA-EDSR managed to reduce more than 60% running time and computational cost of CSNLN, while achieving comparable and better quantitative results. Moreover, we found it has comparable running time and better performance with SAN. Though PA-EDSR has more parameters due to the EDSR backbone (40.7M), the attention module itself is lightweight. The additional pyramid attention only cost 0.2% extra parameters in total. Therefore, we conclude that PA-EDSR achieves better trade-off between efficiency and performance.

A similar conclusion can be derived by evaluating the runtime on Urban100 dataset which contains large images (with an average size of around 1K resolution). It can be seen that our method achieves the best runtime speed. It is interesting to see that the runtime of SAN increases significantly on large images as it requires computing second-order statistics. In contrast, our method is 60% more efficient.

4.8 Visualization of Attention Map

To fully demonstrate that our pyramid attention captures multi-scale correlations, we visualize its attention map in Fig. 7. For illustration purpose, the selected images contain abundant self-exemplars at different locations and scales.

From Fig. 7, we find the attention maps follow distinct distributions over scales, demonstrating that our attention is able to focus on informative regions at multiple scales. It is interesting to point out, as level increases, the most engaged

patches move downwards. This is in line with that larger patterns, such as windows, appear at bottom in selected images. By capturing multi-scale correlations, the network managed to utilize these informative patches to improve restoration.

4.9 Ablation Study

4.9.1 Pyramid Attention Module

To verify the effectiveness of pyramid attention, we conduct control experiments on image denoising tasks ($\sigma = 30$). The baseline module is constructed by removing the attention block, resulting in a simple ResNet. We set the number of residual blocks $R = 16$ in this experiment. In Table 10, *baseline* achieves 30.86 dB on Urban100. To compare with standard non-local (self-) attention operations, we construct a non-local baseline by replacing the pyramid attention with non-local attention. We further construct a scale-agnostic (S-A) attention baseline, which is a special case of the proposed pyramid attention with only one additional pyramid level. From the result in column 2, we can see that single-scale non-local operation is able to bring improvements. Extending it to the scale-agnostic attention further brings about 0.09 dB improvement due to the exploration of information at another scale. However, the best performance is achieved by using the proposed pyramid attention, with brings 0.43 dB over the baseline, 0.15 dB over the standard non-local model and 0.06dB over the scale-agnostic attention. These results indicates the proposed pyramid attention can be served as a better alternative to model multi-scale long-range dependency than current non-local operation, which is of central importance for reconstructing more faithful images.

4.9.2 Matching: Pixel-Wise Versus Block-Wise

While classic non-local attentions compute pixel-wise (i.e. 1×1) feature correlation, we find block-wise matching yields much better restorations in practice. Because such design is perpendicular to the use of feature pyramid, to study its effectiveness, we build models upon the standard non-local operation and adopt different matching strategies, where the patch size is set to 1×1 (i.e. pixel-wise), 3×3 and 5×5 . As shown in Table 11, when using block matching, the performance is improve from 31.14 dB to 31.21 dB. This is because block-matching involves extra similarity constraint

Table 9 Efficiency comparison with attention-based methods

Method	PSNR	Para.	Urban100		100 × 100	
			Time (s)	Time (ms)	Mem (MB)	FLOPs (G)
SAN	34.07	15.7M	5.84	320	1188	1120
CSNLN	34.12	3.1M	16.5	900	982	7129
PA-EDSR	34.22	40.8M	2.35	327	859	2718

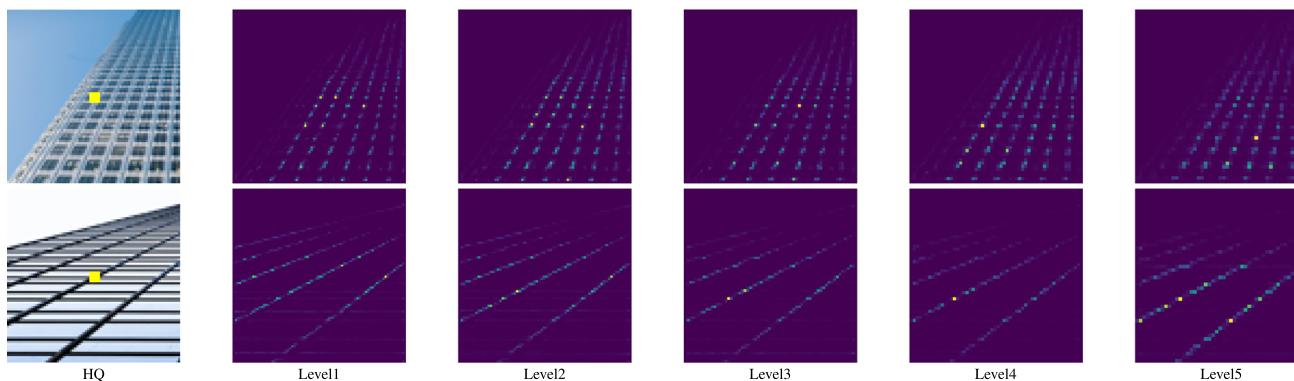


Fig. 7 Visualization of correlation maps of pyramid attention. Maps are rescaled to same size for visualization purpose. Brighter color indicates higher engagement. One can see that the attention focuses on different

locations at each scale, indicating the module is able to exploit multi-scale recurrences to improve restoration

Table 10 Comparison of different attention methods on Urban100

$\sigma = 30$	baseline	N-L (self-) attention	S-A attention	Pyramid attention
PSNR	30.86	31.14	31.23	31.29

Table 11 Comparison between pixel-wise matching and block-wise matching on Urban100. Results are based on the standard non-local attention (i.e. without pyramidal design)

$\sigma = 30$	1 × 1	3 × 3	5 × 5
PSNR	31.14	31.21	31.18

on nearby pixels, thus can better distinguish highly relevant correspondences from noisy ones. These results demonstrate that small patches are indeed more robust descriptors for similarity measurements. However, when further enlarging the patch size to 5 × 5, the performance begins to decrease. This is mainly because larger patches tend to impose an over-strong restriction on the content similarity, and therefore prevent many correlated patches from being leveraged by the network.

4.9.3 Feature Pyramid Levels

As discussed above, the key difference between classic non-local operation and pyramid attention is that our module allows the network to utilize correspondences at multiple scales. Here we investigate the influences of pyramid levels. We conduct control experiments by gradually adding

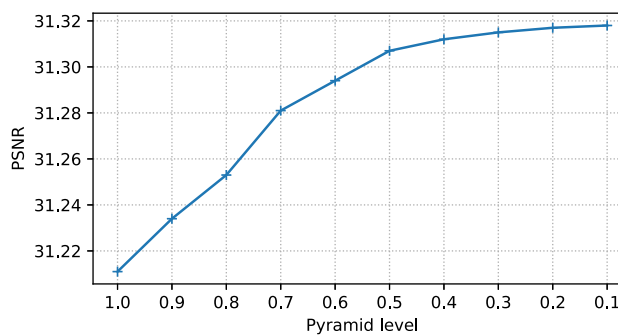


Fig. 8 Ablation study on pyramid levels

more levels to the feature pyramid until it covers the full possible range. The final pyramid consists of 10 layers with scale factors 1.0 to 0.1. As shown in Fig. 8, when more layers are added, we observe constant performance gains. The best performance is obtained when all levels are included. This is mainly because, as the search space is progressively expanded to more scales, the attention unit has higher possibilities to find more informative correspondences beyond the original image scale. Although higher levels only increase a small portion of the search space, thanks to the downscaling operation, patches at these higher levels contain more

Table 12 Results for models with pyramid attention inserted at different residual blocks on Urban100 ($\sigma = 30$)

Pre		✓			✓	✓		✓
Mid			✓		✓		✓	✓
Post				✓		✓	✓	✓
PSNR	30.86	31.07	31.29	31.18	31.33	31.33	31.39	31.48

Table 13 Effects of pyramid attention on different backbones ($\sigma = 30$)

$\sigma = 30$	Dense	PA-Dense	U-Net	PA-U-Net
Kodak24	31.45	31.66	31.54	31.67
BSD68	30.34	30.49	30.40	30.48
Urban100	30.72	31.21	30.81	31.23

"clean" information that could still benefit image restoration. This explains why searching at a very smaller scale (e.g., $s=0.2$) can still improve performance. These results indicate that modeling multi-scale correlation is indeed beneficial for improving restoration.

4.9.4 Positions in Neural Networks

Where should we add pyramid attention to the networks, in order to fully unleash its potential? Table 12 compares pyramid attentions inserted to different stages of a ResNet. Here we consider 3 typical positions: after the 1st residual block representing preprocessing, after the 8th residual block, which is the middle of the network, and after the last residual block representing post-processing. From the first 4 columns, we find that inserting our module at any stages bring evident improvements. The largest performance gain is achieved by inserting it at middle. Moreover, when multiple modules are combined, the restoration quality further boosts. The best result is achieved by including modules at all three positions.

4.9.5 Effects of Backbones

The proposed pyramid attention is a generic operation and its effectiveness is robust to specific architecture design. To demonstrate this, we evaluate the pyramid attention on DenseNet and U-Net, which are two commonly used network structures for image restoration. Here we construct a 18-layer DenseNet and a 3-level 26-layer U-Net with one additional pyramid attention at the end. Results are presented in Table 13. One can see that adding pyramid attention constantly improves the performances. It worth noting that U-Net inherently has multi-scale built in but pyramid attention can still bring considerable improvements. This is because U-Net can be seen as a specific instantiation of modeling multi-scale self-similarities, where only **in-place** self-similarities are fused together. In contrast, pyramid attention generalizes

this operation by exhaustively modeling multi-scale **non-local** correlations.

5 limitation and Future Work

While our method is capable to reconstruct clear and accurate image details, exhaustively computing the non-local correlations across scales adds extra computation burden. Therefore, how to further improve its efficiency for real-time inference is worth exploring. Recent research demonstrates Mei et al. (2021) that exploring sparsity in non-local operation can effectively reduce computational costs from quadratic to asymptomatic linear, and thus investigating sparse representation in pyramid attention may be a promising future direction. Since pyramid attention is a generic operation, it can be further applied to other image restoration tasks such as inpainting, deblurring and deraining or combined with recent vision transformers for superior performance. In addition, it is also interesting to combine PANet with adversarial training and perceptual loss to pursue more visual pleasing restoration.

6 Conclusion

In this paper, we proposed a simple and generic pyramid attention for image restoration. The module generalizes classic self-attention to capture non-local relationships at multiple image scales. It is fully differentiable and can be used into any architectures. We demonstrate that modeling multi-scale correspondences brings significant improvements for the general image restoration tasks of image denoising, demosaicing, compression artifacts reduction and super resolution. On all tasks, a simple backbone with one pyramid attention achieves superior restoration accuracy over prior state-of-the-art approaches. We believe pyramid attention should be used as a common building block in future neural networks.

Data Availability All training/testing data and code that support the findings of this study have been deposited in our GitHub repository <https://github.com/SHI-Labs/Pyramid-Attention-Networks>. Additional visual results and pre-trained models reported in the paper can be found through this [link](#).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Anwar, S., & Barnes, N. (2019). Real image denoising with feature attention. In *ICCV* (pp. 3155–3164).
- Anwar, S., Khan, S., & Barnes, N. (2020). A deep journey into super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 53(3), 1–34.
- Bahat, Y., Efrat, N., & Irani, M. (2017). Non-uniform blind deblurring by reblurring. In *ICCV* (pp. 3286–3294).
- Bahat, Y., & Irani, M. (2016). Blind dehazing using internal patch recurrence. In *ICCP* (pp. 1–9). IEEE.
- Buades, A., Coll, B., & Morel, J.M. (2005). A non-local algorithm for image denoising. In *CVPR*.
- Buades, A., Coll, B., & Morel, J. M. (2011). Non-local means denoising. *Image Processing On Line*, 1, 208–212.
- Cao, Y., Xu, J., Lin, S., Wei, F., & Hu, H. (2019). Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *ICCV Workshops* (pp. 0–0).
- Chen, C., Chen, Q., Xu, J., & Koltun, V. (xxxx). Learning to see in the dark.
- Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., & Gao, W. (2021). Pre-trained image processing transformer. In *CVPR* (pp. 12299–12310).
- Chen, Y., & Pock, T. (2017). Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. In *TPAMI*.
- Dabov, K., Foi, A., Katkovnik, V., & Egiazarian, K. (2007). Color image denoising via sparse 3d collaborative filtering with grouping constraint in luminance-chrominance space. In *ICIP*.
- Dabov, K., Foi, A., Katkovnik, V., & Egiazarian, K. (2007). Image denoising by sparse 3-d transform-domain collaborative filtering. In *TIP*.
- Dai, T., Cai, J., Zhang, Y., Xia, S.T., & Zhang, L. (2019). Second-order attention network for single image super-resolution. In *CVPR* (pp. 11065–11074).
- Dong, C., Deng, Y., Change Loy, C., & Tang, X. (2015). Compression artifacts reduction by a deep convolutional network. In *ICCV*.
- Dong, C., Loy, C. C., He, K., & Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In *ECCV*.
- Dong, C., Loy, C. C., & Tang, X. (2016). Accelerating the super-resolution convolutional neural network. In *ECCV*.
- Fan, Y., Yu, J., Liu, D., & Huang, T. S. (2019). Scale-wise convolution for image restoration. arXiv preprint [arXiv:1912.09028](https://arxiv.org/abs/1912.09028).
- Fan, Y., Yu, J., Mei, Y., Zhang, Y., Fu, Y., Liu, D., & Huang, T. S. (2020). Neural sparse representation for image restoration. *NeurIPS*, 33, 15394–15404.
- Foi, A., Katkovnik, V., & Egiazarian, K. (2007). Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images. In *TIP*.
- Freedman, G., & Fattal, R. (2011). Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2), 1–11.
- Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In *CVPR* (pp. 3146–3154).
- Glasner, D., Bagon, S., & Irani, M. (2009). Super-resolution from a single image. In *ICCV* (pp. 349–356). IEEE.
- Haris, M., Shakhnarovich, G., & Ukit, N. (2018). Deep back-projection networks for super-resolution. In *CVPR* (pp. 1664–1673).
- He, K., Sun, J., & Tang, X. (2010). Single image haze removal using dark channel prior. *TPAMI*, 33(12), 2341–2353.
- He, X., Mo, Z., Wang, P., Liu, Y., Yang, M., Cheng, J. (2019). Ode-inspired network design for single image super-resolution. In *CVPR* (pp. 1732–1741).
- Huang, J. B., Singh, A., & Ahuja, N. (2015). Single image super-resolution from transformed self-exemplars. In *CVPR* (pp. 5197–5206).
- Jo, Y., & Kim, S. J. (2021). Practical single-image super-resolution using look-up table. In *CVPR* (pp. 691–700).
- Kim, J., Kwon Lee, J., & Mu Lee, K. (2016). Accurate image super-resolution using very deep convolutional networks. In *CVPR*.
- Kong, X., Liu, X., Gu, J., Qiao, Y., & Dong, C. (2022). Re-flash dropout in image super-resolution. In *CVPR* (pp. 6002–6012).
- Lai, W. S., Huang, J. B., Ahuja, N., Yang, M. H. (2017). Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*.
- Li, B., Peng, X., Wang, Z., Xu, J., Feng, D. (2017). Aod-net: All-in-one dehazing network. In *ICCV* (pp. 4770–4778).
- Li, J., Chen, C., Cheng, Z., Xiong, Z. (2022). Mulu: Cooperating multiple look-up tables for efficient image super-resolution. In *European conference on computer vision* (pp. 238–256). Springer.
- Li, S., Araujo, I. B., Ren, W., Wang, Z., Tokuda, E. K., Junior, R. H., Cesar-Junior, R., Zhang, J., Guo, X., & Cao, X. (2019). Single image deraining: A comprehensive benchmark analysis. In *CVPR* (pp. 3838–3847).
- Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., & Wu, W. (2019). Feedback network for image super-resolution. In *CVPR* (pp. 3867–3876).
- Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). Swinir: Image restoration using swin transformer. In *ICCV* (pp. 1833–1844).
- Lim, B., Son, S., Kim, H., Nah, S., & Lee, K. M. (2017). Enhanced deep residual networks for single image super-resolution. In *CVPRW*.
- Liu, D., Wen, B., Fan, Y., Loy, C. C., & Huang, T. S. (2018). Non-local recurrent network for image restoration. In *NeurIPS*.
- Liu, J., Zhang, W., Tang, Y., Tang, J., & Wu, G. (2020). Residual feature aggregation network for image super-resolution. In *CVPR* (pp. 2359–2368).
- Lotan, O., & Irani, M. (2016). Needle-match: Reliable patch matching under high uncertainty. In *CVPR* (pp. 439–448).
- Magid, S. A., Zhang, Y., Wei, D., Jang, W. D., Lin, Z., Fu, Y., & Pfister, H. (2021). Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *ICCV* (pp. 4288–4297).
- Mairal, J., Bach, F., Ponce, J., Sapiro, G., & Zisserman, A. (2009). Non-local sparse models for image restoration. In *ICCV* (pp. 2272–2279). IEEE.
- Mao, X., Shen, C., & Yang, Y. B. (2016). Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *NeurIPS*.
- Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating

- segmentation algorithms and measuring ecological statistics. In *ICCV*.
- Mei, Y., Fan, Y., & Zhou, Y. (2021). Image super-resolution with non-local sparse attention. In *CVPR* (pp. 3517–3526).
- Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T. S., & Shi, H. (2020). Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *CVPR* (pp. 5690–5699).
- Michaeli, T., & Irani, M. (2014). Blind deblurring using internal patch recurrence. In *ECCV* (pp. 783–798). Springer.
- Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., Zhang, K., Cao, X., & Shen, H. (2020). Single image super-resolution via a holistic attention network. In *European conference on computer vision* (pp. 191–207). Springer.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in pytorch.
- Roth, S., & Black, M. J. (2005). Fields of experts: A framework for learning image priors. In *CVPR* (vol. 2, pp. 860–867). IEEE.
- Sheikh, H. R., Wang, Z., Cormack, L., & Bovik, A. C. (2005). Live image quality assessment database release 2.
- Singh, A., & Ahuja, N. (2014). Super-resolution using sub-band self-similarity. In *ACCV* (pp. 552–568). Springer.
- Tai, Y., Yang, J., Liu, X., & Xu, C. (2017). Memnet: A persistent memory network for image restoration. In *ICCV*.
- Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., & Lin, C. W. (2020). Deep learning on image denoising: An overview. *Neural Networks*, *131*, 251–275.
- Timofte, R., Agustsson, E., Van Gool, L., Yang, M. H., Zhang, L., Lim, B., Son, S., Kim, H., Nah, S., Lee, K. M., et al. (2017). Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*.
- Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008). Extracting and composing robust features with denoising autoencoders. In *ICML*.
- Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *CVPR*.
- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. In *TIP*.
- Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., & Li, H. (2022). Uformer: A general u-shaped transformer for image restoration. In *CVPR* (pp. 17683–17693).
- Xia, B. N., Gong, Y., Zhang, Y., & Poellabauer, C. (2019). Second-order non-local attention networks for person re-identification. In *ICCV* (pp. 3760–3769).
- Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M. H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *CVPR* (pp. 5728–5739).
- Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M. H., & Shao, L. (2020). Learning enriched features for real image restoration and enhancement. In *ECCV* (pp. 492–511). Springer.
- Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M. H., & Shao, L. (2021). Multi-stage progressive image restoration. In *CVPR* (pp. 14821–14831).
- Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. In *TIP*.
- Zhang, K., Zuo, W., Gu, S., Zhang, L. (2017). Learning deep cnn denoiser prior for image restoration. In *CVPR*.
- Zhang, K., Zuo, W., & Zhang, L. (2017). Ffdnet: Toward a fast and flexible solution for cnn based image denoising. arXiv preprint [arXiv:1710.04026](https://arxiv.org/abs/1710.04026).
- Zhang, K., Zuo, W., & Zhang, L. (2018). Learning a single convolutional super-resolution network for multiple degradations. In *CVPR* (pp. 3262–3271).
- Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. In *ECCV*.
- Zhang, Y., Li, K., Li, K., Zhong, B., & Fu, Y. (2019). Residual non-local attention networks for image restoration. In *ICLR*.
- Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2018). Residual dense network for image super-resolution. In *CVPR*.
- Zhang, Y., Wang, H., Qin, C., & Fu, Y. (2021). Aligned structured sparsity learning for efficient image super-resolution. *NeurIPS*, *34*, 2695–2706.
- Zhang, Y., Wei, D., Qin, C., Wang, H., Pfister, H., & Fu, Y. (2021). Context reasoning attention network for image super-resolution. In *ICCV* (pp. 4278–4287).
- Zhou, S., Zhang, J., Zuo, W., & Loy, C. C. (2020). Cross-scale internal graph neural network for image super-resolution. *NeurIPS*, *33*, 3499–3509.
- Zontak, M., & Irani, M. (2011). Internal statistics of a single natural image. In *CVPR* (pp. 977–984). IEEE.
- Zontak, M., Mosseri, I., & Irani, M. (xxxx). Separating signal from noise using patch recurrence across scales.
- Zoran, D., & Weiss, Y. (2011). From learning models of natural image patches to whole image restoration. In *ICCV* (pp. 479–486). IEEE.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.