Efficient Lightweight Image Denoising with Triple Attention Transformer

Yubo Zhou[1*], Jin Lin[1*], Fangchen Ye[1], Yanyun Qu[1†], Yuan Xie[2†]

[1]School of Informatics, Xiamen University, Fujian, China
[2]School of Computer Science and Technology, East China Normal University, Shanghai, China
ybzhou@stu.xmu.edu.cn, yxie@cs.ecnu.edu.cn, yyqu@xmu.edu.cn

Abstract

Transformer has shown outstanding performance on image denoising, but the existing Transformer methods for image denoising are with large model sizes and high computational complexity, which is unfriendly to resource-constrained devices. In this paper, we propose a Lightweight Image Denoising Transformer method (LIDFormer) based on Triple Multi-Dconv Head Transposed Attention (TMDTA) to boost computational efficiency. LIDFormer first implements Discrete Wavelet Transform (DWT), which transforms the input image into a low-frequency space, greatly reducing the computational complexity of image denoising. However, the low-frequency image lacks fine-feature information, which degrades the denoising performance. To handle this problem, we introduce the Complementary Periodic Feature Reusing (CPFR) scheme for aggregating the shallow-layer features and the deep-layer features. Furthermore, TMDTA is proposed to integrate global context along three dimensions, thereby enhancing the ability of global feature representation. Note that our method can be applied as a pipeline for both convolutional neural networks and Transformers. Extensive experiments on several benchmarks demonstrate that the proposed LIDFormer achieves a better trade-off between high performance and low computational complexity on real-world image denoising tasks.

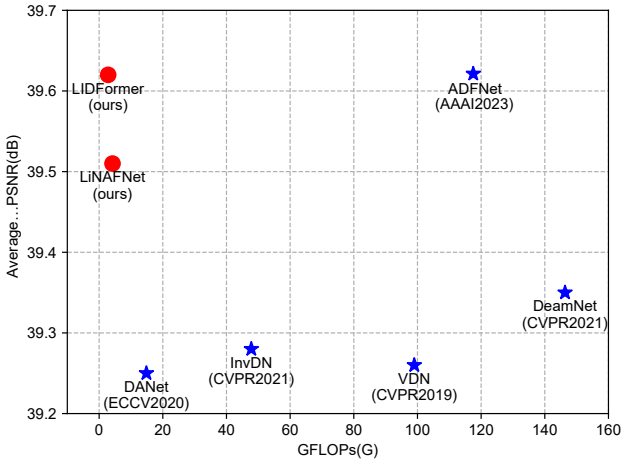

Figure 1: Performance and FLOPs cost of LIDFormer compared to other popular efficient and lightweight denoising methods on SIDD. The LiNAFNet is formed by applying the module from LIDFormer to NAFNet (Chen et al. 2022a). Our method achieves a better trade-off between performance and FLOPs cost on image denoising tasks.

Introduction

Image denoising is an important task in image restoration and is widely applied to many scenarios (Anwar, Khan, and Barnes 2020). With the rise of deep learning, image denoising methods have made great progress (Tai et al. 2017; Chen and Pock 2016; Zhou et al. 2020; Mao, Shen, and Yang 2016; Ulyanov, Vedaldi, and Lempitsky 2018; Cheng et al. 2021). However, the existing models mostly require high computational complexity in order to obtain good performance, which may hinder the widespread application of methods on resource-limited devices such as mobile phones, robotics, and some edge devices. Efficient and lightweight denoising methods attract more and more attention.

With the rising up of deep learning, convolutional neural networks are used for image denoising. The method (Xu, Yang, and Jiang 2017; Yuan, Liu, and Liang 2023) reduces computations and storage costs by utilizing the sparse nature of images. It has achieved remarkable results in denoising effect and computational speed. Moreover, the approach (Yu et al. 2018) based on the joint loss function is a new idea proposed in recent years, which improves the denoising effect by simultaneously considering the local and global information of the image. Meanwhile, the method (Jin et al. 2019) based on depthwise separable convolution is widely used in image denoising tasks. It improves the denoising efficiency and accuracy by separating spatial and channel dimensions while reducing model parameters and computational costs. This method has been shown to be effective in many image denoising tasks, especially for practical applications and large-scale data.

Although the above methods have accelerated the denoising process in different aspects, the computational efficiency of current lightweight image denoising models still has resource barriers compared with advanced semantic tasks such

*These authors contributed equally.
†Corresponding authors.

as image classification. Therefore, in order to narrow the gap with advanced semantic tasks and realize the computational efficiency of image denoising algorithms adapted to practical devices, image denoising methods with less than 5 GFLOPs of computational cost are worth exploring and designing.

In response to the above problems, we propose a Lightweight Image Denoising Transformer method (LIDFormer) based on Discrete Wavelet Transform (DWT) (Mallat 1989) and Triple Multi-Dconv Head Transposed Attention (TMDTA), which aims to produce excellent performance while being computationally efficient. To be specific, our proposed lightweight feature module utilizes DWT (Mallat 1989) to losslessly transform the input image into a low-resolution space composed of high-frequency and low-frequency information sets. Notably, DWT (Mallat 1989) is an established lossless frequency-domain transformation function that is not involved in model training, so it can be considered a non-computationally consuming module.

Moreover, Complementary Periodic Feature Reusing (CPFR) is introduced to mitigate the loss of information due to low resolution. Through continuous complementary residual connection, CPFR combines the historical feature with the current feature in a weighted and complementary way. It also avoids the discarding of valid features due to the refinement of the feature information as the network level goes deeper. In particular, the complementary residual connections are learnable channel attention functions.

From another point of view, the multi-head self-attention (MHSA) proposed by Transformer (Vaswani et al. 2017) can effectively refine characteristic information and overcome the "short-range" effect of local convolution. However, since the global pixel-based computation of self-attention is too large (proportional to the resolution of the features), it is usually not directly applicable to image restoration tasks. The feature lightweighting strategy propounded by LIDFormer allows all the computations of features to be performed in low-scale space, thus making global self-attention possible for resource-constrained devices. Based on the above discussion, LIDFormer introduces TMDTA, namely horizontal self-attention, vertical self-attention and channel-wise self-attention, for collaborative computing. Finally, LIDFormer achieves a computational cost close to image classification with 2.8 GFLOPs. More intuitively, as shown in Fig. 1, LIDFormer significantly outperforms the majority of popular efficient image denoising methods and has much lower computational complexity than these approaches.

We summarize the main contributions of this work as follows:

- We propose an efficient and lightweight image denoising method based on DWT and TMDTA (namely LIDFormer). Our LIDFormer provides a novel pipeline to reduce computational complexity, and it is a universal and generalizable efficient method.
- We design the Complementary Periodic Feature Reusing module (CPFR), which can effectively overcome the problem of compact and insufficient feature information caused by feature lightweighting. The reused effect can solve the issue of catastrophic forgetting to a certain extent and effectively retain low-frequency information.
- We introduce a Triple Multi-Dconv Head Transposed Attention module (TMDTA) to improve the performance of conventional multi-head self-attention based on feature pixels in a multi-dimensional and lightweight manner.
- Extensive experiments demonstrate that our LIDFormer achieves a better trade-off between performance and computational complexity. The pipeline can also be generalized to different image denoising methods.

Related Work

Deep Learning-based Image Denoising

Image denoising tasks aim to restore a high-quality image from the noisy observation (Chen et al. 2022a). In recent years, with the rise of deep learning technology, CNN-based network architectures (Tai et al. 2017; Chen and Pock 2016; Zamir et al. 2020, 2021; Zhang et al. 2020, 2017; Cheng et al. 2021) have achieved significant success in the field of image denoising, and their performance is far superior to that of traditional restoration methods (Dabov et al. 2008; Gu et al. 2014; Xu et al. 2017; Yair and Michaeli 2018; He, Sun, and Tang 2010). These deep networks have different characteristics in their designs, and most of them (Wang et al. 2022; Yue et al. 2020; Zamir et al. 2021; Zhang et al. 2021) are based on the UNet (Ronneberger, Fischer, and Brox 2015) architecture, which uses skip-connections to fuse the pixel-level features of the image with semantic-level features for better restoration results.

As Transformer-based models (Vaswani et al. 2017; Fedus, Zoph, and Shazeer 2022; Radford et al. 2018) have achieved excellent performance in the NLP domain, more and more vision applications, both high-level tasks (Graham et al. 2021; Liu et al. 2021b; Carion et al. 2020; Xie et al. 2021) and low-level tasks (Liang et al. 2021a; Kumar, Weissenborn, and Kalchbrenner 2020; Zamir et al. 2022; Wang et al. 2022), have tried to introduce it recently due to its strong capability of modeling long-range relations. Most of them have achieved better results compared to convolutional networks. The Vision Transformer (ViT) (Dosovitskiy et al. 2020) divides an image down into a series of patches (local windows) and discovers how they relate to one another. Benefiting from the powerful multi-head self-attention mechanism, its ability to calculate long-distance information interaction is particularly outstanding. Some existing works (Zamir et al. 2022; Chen et al. 2021, 2022b) have achieved promising performance by applying the ViT architecture to image denoising while alleviating the prohibitively expensive training complexity. Vision Transformers have shown their strong potential as an alternative to the previously dominant CNNs (Liang et al. 2021b). Recently, Restormer (Zamir et al. 2022) is proposed as a high-performance Transformer model for image denoising. It introduces a gating mechanism based on depth-wise convolutions to perform controlled feature transformation. Although this method achieves state-of-the-art denoising performance, it also sacrifices a large amount of computational cost. In this paper, we propose a computationally friendly method named

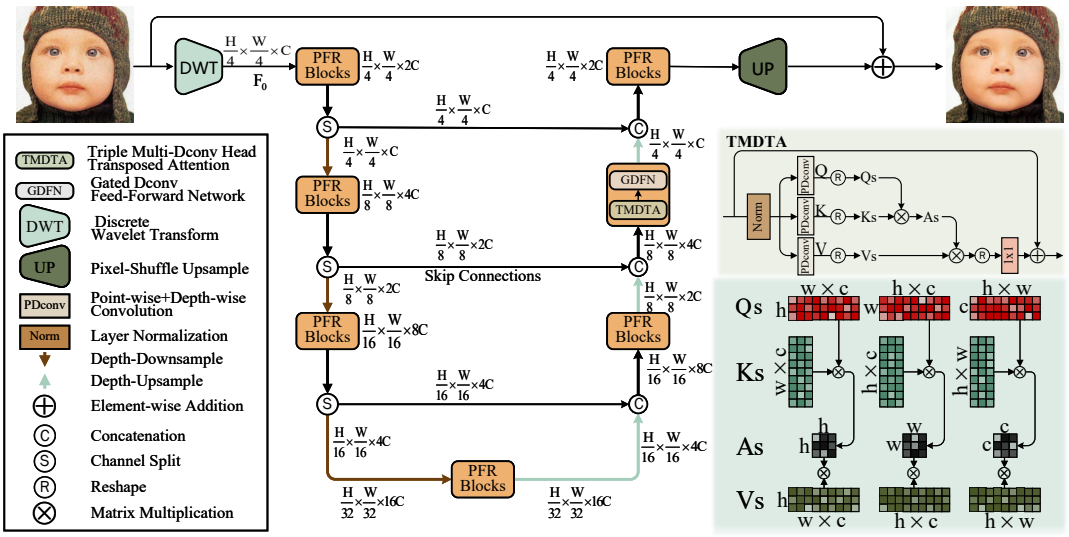

Figure 2: Illustration of our proposed LIDFormer. First, the input image is transformed into a low-resolution frequency-domain space using DWT. Then, the CPFR module is used to combine features from historical and current periods, effectively multiplexing the features and avoiding the issue of shallow features being forgotten due to the filtering of depth information. Additionally, LIDFormer incorporates TMDTA to capture global feature information in three dimensions, which approximates traditional high-computation full-pixel self-attention.

LIDFormer for image denoising. Our method reduces the computational workload of existing models without compromising the capability of denoising.

Efficient and Lightweight Image Denoising

Although the performance of the image denoising methods mentioned above improves significantly, they mostly suffer from high computational costs, which do not favor resource-constrained devices such as smart phones. To relieve the computation burden and improve efficiency, there are emerging efforts to design efficient and lightweight image denoising approaches. Zhang *et al*. (Zhang, Zuo, and Zhang 2018) propose a new CNN model based on DnCNN (Zhang et al. 2017) , namely FFDNet, for rapid, effective, and adjustable discriminative denoising. FFDNet uses downsampled sub-images, which significantly speeds up training and testing while also expanding the receptive area. Yue *et al*. (Yue et al. 2019) utilize a new variational inference method (VDN) to fast infer both the underlying clean image and the noise distribution from an observed noisy image in a unique Bayesian framework. DANet (Yue et al. 2020) approximates the joint distribution from two different factorized forms in a dual adversarial manner. The joint distribution theoretically contains more complete information underlying the data set, which significantly reduces the time required to collect clean-noisy image pairs. Zou *et al*. (Zou et al. 2023) make contributions for efficient image denoising by a lightweight network and a novel distillation algorithm with retargeting supervision. Another related work is Thunder (Zhou et al. 2022), which leverages the RGB thumbnail instead of the feature subspace to accelerate the denoising process. More specifically, it adopts the subspace projection method to guarantee the denoising effect while refining the thumbnail. Unlike Thunder (Zhou et al. 2022), our method yields better denoising effects with faster calculation efficiency by incorporating the DWT module and the TMDTA module.

Method

Overview of LIDFormer

As shown in Fig. 2, LIDFormer consists of three main components: (1) A feature lightweighting module based on DWT, which maps a given noisy image x from RGB space to low-resolution frequency-domain space through a double discrete wavelet transform (DWT); (2) A Complementary Periodic Feature Reusing (CPFR) module, which performs non-linear operations on low-resolution frequency-domain features; (3) A Triple Multi-Dconv Head Transposed Attention (TMDTA) module, which introduces three-dimensional co-computation of horizontal, vertical and channel self-attention. First of all, the input image is transformed into a low-resolution frequency-domain space using DWT to alleviate the computational bottleneck. Then, the function of feature multiplexing is realized through the CPFR module. CPFR can effectively combine the features of different periods and avoid the problem of shallow features being forgotten due to the filtering of depth information. Besides, TMDTA is utilized to obtain the global information of features in three dimensions and approximately replaces the traditional high-computation full-pixel self-attention. Note that the upsampling module is implemented directly by using the conventional non-computationally intensive "pixel-shuffle" operation, which compresses the feature channel, and then the compressed part is filled in the channel to achieve lossless amplification feature resolution. The specific process is

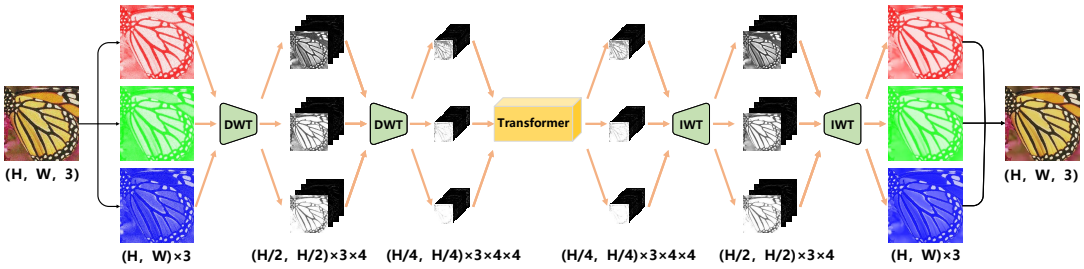

Figure 3: Illustration of the double DWT feature lightweighting network. Here, we choose Transformer-based models as the backbone network. A color image with three RGB channels is used as the initial input. The channel of middle input is increased by 16 times compared to the original image by a double DWT, while the resolution is decreased by 16 times.

expressed as: (B, C $\times$ γ^2 , H, W) $\longrightarrow$ (B, C, H $\times$ γ , W $\times$ γ). The above process can be expressed as:

$$\begin{aligned} F_0 &= f_{DWT}(x), \\ F_n &= f_{Unet_{CPFR}}(F_0), \\ I_{Restored} &= x + f_{UP(F_n)}. \end{aligned} \tag{1}$$

Among them, $f_{DWT}(\bullet)$ means that the double discrete wavelet transform performs frequency-domain compression on the input noise image x, and $f_{Unet_{CPFR}}(\bullet)$ means the noise extraction function for low-resolution features.

DWT-based Feature Lightweighting

At present, the denoising models based on the deep network mainly rely on the noise extraction method for image denoising. The details are as follows:

$$y = x + \eta, \tag{2}$$

where y denotes the noisy image, x denotes the denoised clear image, both represented as vectors, and η is the noise distribution of the noisy image. Moreover, in most deep learning-based models, η is usually learned from the noisy image in an end-to-end image denoising task, as follows:

$$\eta = F(y), \tag{3}$$

where $F(\bullet)$ denotes the noise extractor in the denoising process. As shown in Fig. 2, before noise extraction, the original input image will be converted to the frequency-domain space through the DWT-based resolution compression module, and the input image resolution will be compressed to $\frac{1}{16}$ of the original size by the double DWT, which can greatly reduce the computational complexity while keeping the number of channels and model parameters unchanged. The specific low-resolution frequency-domain compression module can be described as:

$$\eta = F(DWT(y)), \tag{4}$$

where $DWT(\bullet)$ denotes the double DWT transform function, which is an established lossless frequency-domain transform function. What needs to be emphasized here is that our method uses the classic Haar (Mallat 1989) as the discrete wavelet and aims to decompose the input image $X \in \mathbb{R}^{H\times W\times 3}$ into 48 low-resolution frequency-domain sub-features $f_i \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}}, i \in [1, \ldots, 48]$.

The image denoising process based on the DWT makes features lightweight, as shown in Fig. 3. The input image is a color one with three RGB channels. Through two first-order DWT transformations (i.e., double DWT), the feature channels of middle input are expanded by 16 times compared to the original image, while the feature resolution is reduced by 16 times, achieving a lossless feature lightweighting effect.

Complementary Periodic Feature Reusing

Although feature lightweighting based on the double DWT greatly reduces the computational complexity of the model, the intermediate feature is compressed 16 times compared to the original model without lightweighting, resulting in a serious shortage of feature information. Therefore, LIDFormer proposes the Complementary Periodic Feature Reusing (CPFR) module, which aims to reuse historical features and fill in the shallow historical feature information lost during the learning process of compact features. First, the compact features are expanded by a factor of two in the channel dimension, and a simple linear feature embedding is done by using a generalized 3 $\times$ 3 convolution (CONV3 below) to double the information space of compact features:

$$F_0 = \mathrm{CONV}\,3(DWT(x)). \tag{5}$$

Then, as shown in Fig. 4, in order to make full use of the extended feature information and perform effective historical feature reusing, our method calculates the features of the next stage while retaining the feature information of the previous stage. The simple CPFR is expressed as follows:

$$f_n = \begin{cases} T \cdot F_n(f_{n-1}) + (1-T)\cdot f_{n-1}, & n = 2\times k \\ f_{n-1}, & else \end{cases} \tag{6}$$

where f_n denotes the upper or lower half of the extended feature, $F_n(\bullet)$ represents the processing unit (i.e., TMDTA), and T is the complementary coefficient. In the experiment of this method, the value of T is set to 0.5.

Complementary Adaptive Channel Attention

Since the information of each feature is constantly changing and the information of deep and shallow features is not uniform at each pixel, it is not friendly to set the value of the complementary coefficient rigidly. To address the above

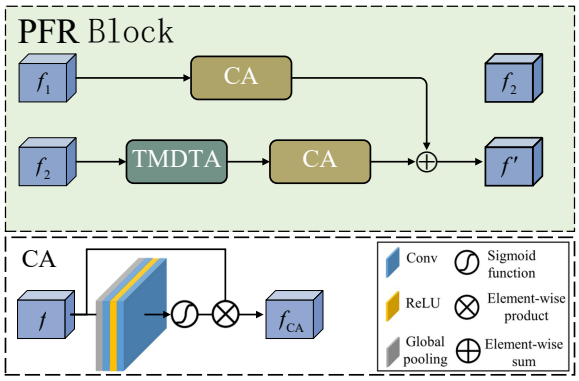

Figure 4: Illustration of the Complementary Periodic Feature Reusing (CPFR) module. CA denotes the channel attention function. f_1 and f_2 denote the historical feature and the current feature, respectively.

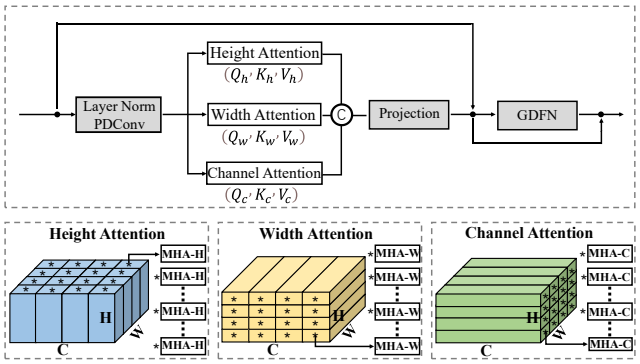

Figure 5: Illustration of the Triple Multi-Dconv Head Transposed Attention (TMDTA) moudle. The attention of characteristic pixels is decomposed into three directions of self-attention for cooperative computation: horizontal self-attention, vertical self-attention, and channel self-attention.

problem, CPFR uses Complementary Adaptive Channel Attention (CACA) to compensate for the deficiency of hard complementary coefficients, which is represented as follows:

$$g_n = F_n(f_{n-1})$$
$$f_n = \begin{cases} F^{CA}_{n_1}(g_n) \cdot g_n + F^{CA}_{n_2}(f_{n-1}) \cdot f_{n-1}, & n = 2 \times k \\ f_{n-1}, & else \end{cases} \tag{7}$$

where F^{CA}_n represents the channel attention function, as shown in the CA module in Fig. 4, which consists of simple convolution, activation, pooling, and other basic constructions. More importantly, the convolution calculation is processed on the pooled single-point multi-channel features; that is, the overall calculation is done with the feature resolution of only one, which is almost negligible compared to the overall feature calculation.

In addition, CPFR imposes a complementary constraint on adaptive channel attention and construes this part with a simple MSE loss. The effectiveness of the complementary constraint has been demonstrated through experiments. The complementary constraint is shown below:

$$L_{CA} = \sum_{i=1}^{n} \left\| F^{CA}_{i_1}(F_i(f_{i-1})) + F^{CA}_{i_2}(f_{i-d}) - ONEs \right\|_2, \tag{8}$$

where n denotes the number of computing units and $ONEs$ denotes a matrix that elements with values of one in the same dimension as the outputs of $F^{CA}(f_{i-d})$, achieving pixel-level complementarity constraints.

Triple Multi-Dconv Head Transposed Attention

In addition to the above approaches, LIDFormer considers a very significant issue: the limitation of Transformer in image restoration lies in the huge computational complexity caused by the demand to complete high-resolution correlation calculations between various pixels. As shown in Fig. 3, the pixel magnification of intermediate features has been scaled by 16 times, and the computation can be reduced by 256 times if the traditional self-attention mechanism of full pixels is adopted. Even so, when it comes to higher resolution images, there is still the problem of "high computational complexity". To this end, considering the information redundancy of full-pixel self-attention, LIDFormer proposes Triple Multi-Dconv Head Transposed Attention (TMDTA). It decomposes the attention of characteristic pixels into three directions of self-attention for cooperative computation: horizontal self-attention, vertical self-attention, and channel self-attention.

As shown in Fig. 5, the input features first pass through the "Layer Norm + PDConv" layer to generate the locally enriched $query(Q)$, $key(K)$ and $value(V)$. The Layer Norm (LN) denotes the regular layer normalization, and the PDConv denotes the combination of Pointwise Convolution (PWConv) and Depthwise Convolution (DWConv). Then, the $query(Q)$ and $key(K)$ are reshaped in three-dimensional directions, resulting in the horizontal $query_H(Q_H)$ and $key_H(K_H)$, the vertical $query_W(Q_W)$ and $key_W(K_W)$, and the $query_C(Q_C)$ and $key_C(K_C)$, respectively. Then, matrix multiplication is performed on them respectively to generate three transposed attention matrices with sizes of $\mathbb{R}^{H\times H}$, $\mathbb{R}^{W\times W}$ and $\mathbb{R}^{C\times C}$, instead of the regular attention matrix $\mathbb{R}^{HW\times HW}$ of characteristic pixels (Vaswani et al. 2017; Dosovitskiy et al. 2020). It is worth noting that all three processes are transformed from $query(Q)$, $key(K)$ and are synergistically related to each other. In general, the process definition of TMDTA is as follows:

$$\begin{aligned} \mathbf{X}' &= W_p\,\mathrm{Atention}(\mathbf{Qs}, \mathbf{Ks}, \mathbf{Ys}) + \mathbf{X}, \\ \mathrm{Atention}(\mathbf{Qs}, \mathbf{Ks}, \mathbf{Vs}) &= \mathrm{Concat}(\mathbf{A_H}, \mathbf{A_W}, \mathbf{A_C}), \\ \mathbf{A_H} &= \mathbf{V_H} \times \mathrm{Softmax}(\mathbf{K_H} \times \mathbf{Q_H}/\alpha_H), \\ \mathbf{A_W} &= \mathbf{V_W} \times \mathrm{Softmax}(\mathbf{K_W} \times \mathbf{Q_W}/\alpha_W), \\ \mathbf{A_C} &= \mathbf{V_C} \times \mathrm{Softmax}(\mathbf{K_C} \times \mathbf{Q_C}/\alpha_C), \end{aligned} \tag{9}$$

where $\mathbf{X}$ and $\mathbf{X}'$ denote the input and output features; $\mathbf{Q}_i \in (\mathbb{R}^{WC\times H}, \mathbb{R}^{HC\times W}, \mathbb{R}^{HW\times C})$, $\mathbf{K}_i \in (\mathbb{R}^{H\times WC}, \mathbb{R}^{W\times HC}, \mathbb{R}^{C\times HW})$, $\mathbf{V}_i \in (\mathbb{R}^{WC\times H}, \mathbb{R}^{HC\times W}, \mathbb{R}^{HW\times C})$ denotes the horizontal,

Baseline	DWT	CPFR	CACA	TMDTA	GFLOPs	PSNR
✓	×	×	×	×	140	40.02
✓	✓	×	×	×	8.75	39.55
✓	✓	✓	×	×	2.82	39.55
✓	✓	✓	✓	×	2.83	39.58
✓	✓	✓	✓	✓	2.83	39.62

Table 1: Ablation experiments are conducted with different modules of the LIDFormer.

vertical, and channel reshaping by the generated $query(Q)$, $key(K)$ and $value(V)$, respectively; α_i denotes a learnable scaling parameter to control the size of the dot product of $\mathbf{Q}_i$ and $\mathbf{K}_i$ before applying the activation function. In the above expression, $i \in [\mathbf{H}, \mathbf{W}, \mathbf{C}]$.

Experiments

Implementation Details

To ensure the fairness of the comparison between methods, our method and conventional denoising methods adopt the same classic denoising dataset SIDD (Abdelhamed, Lin, and Brown 2018) for model training. Moreover, the trained model is evaluated on two publicly available datasets, SIDD (Abdelhamed, Lin, and Brown 2018) and DND (Plotz and Roth 2017). In our work, the Adam optimizer with β_1 = 0.9, β_2 = 0.999 and L1 loss are utilized to train the model. The training process takes 300K iterations with the learning rate being initially set to $3e$-4. And the learning rate will gradually decrease to $1e$-6 by using cosine annealing technique (Loshchilov and Hutter 2016). For iterative learning, 128 $\times$ 128 image patches with RGB channels are used to train a lightweight denoising model. The mini-batch size is set to 16. Besides, the resolution of image patches and the batch size are updated at iteration numbers of 92k, 156K, 204K, 240K, and 276K to (160^2, 8), (192^2, 6), (256^2, 4), (320^2, 2), and (384^2, 1), respectively. Horizontal and vertical flipping are implemented for data augmentation.

Evaluation Metrics

Objective criteria, i.e., peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), are adopted to evaluate the performance of denoising models. The two metrics are both calculated on the Y channel of the YCbCr space. Besides, Giga Floating-point Operations Per second (GFLOPs) is used as the efficient evaluation criterion for the denoising model in our work.

Ablation Study

We conduct ablation studies to validate the effect of each component in our proposed method. All the experiments use Restormer (Zamir et al. 2022) as the baseline model. The quantity results are shown in Table 1.

Effectiveness of Double DWT. As shown in Table 1, using the double DWT is able to compress the original model by 16 times with high performance. It can be seen from the table that as the computational complexity of the model reduces, the performance also decreases slightly. Therefore, in order to verify whether the method of feature lightweighting is feasible and universal, we have carried out corresponding experiments on different methods, as shown in Table 2. It can be observed from the table that the feature lightweighting method will cause a very serious decline in the performance of the model, especially in the experiment on CBDNet (Guo et al. 2019) (dwtCBDNet in the table), where the performance of the model encountered a catastrophic mutation. From this, the corresponding experimental conclusion can be drawn: feature lightweighting by using DWT can significantly reduce the computational complexity of the model and alleviate the computational pressure, but it cannot guarantee the performance of the lightweight model.

Effectiveness of CPFR. As shown in Table 1, we aim to explore efficient image denoising methods whose computational complexity approximates image classification tasks (i.e., below 5 GFLOPs). By reducing the number of modules in the original model and reusing historical features, this deliberate architectural refinement achieves an efficient denoising model of 2.82 GFLOPs in the table. It is worth noting that the utilization of CPFR has greatly saved the performance of dwtCBDNet in Table 2. The results show that our proposed CPFR module further reduces the computational complexity of the efficient denoising model, which has the same performance advantages as the model of 8.75 GFLOPs, verifying the effectiveness of this module.

Effectiveness of CACA. Since the complementary coefficient set in the simple CPFR module is a constant value (i.e., 0.5), the flexibility of feature learning is limited. Therefore, Complementary Adaptive Channel Attention (CACA) is proposed to release the pressure of the given value, making CPFR adaptively complementary. To combine historical features and current deep features, adaptive channel attention considers freely choosing summation coefficients. As shown in Table 1, compared with the simple CPFR module, the introduction of the CACA module has a certain improvement effect. In addition, the adaptive learning method can enhance the generalization of our method and avoid the discomfort of the given value in other methods.

Effectiveness of TMDTA. The channel-wise multi-head self-attention designed in the original Restormer (Zamir et al. 2022) effectively overcomes the inadequacy of the Transformer's (Vaswani et al. 2017) full-pixel self-attention in dense prediction tasks. However, channel-wise multi-head self-attention cannot completely replace the role of full-pixel self-attention because channel global information is unable to represent spatial global information. Therefore, as shown in Table 1, TMDTA is more effective than the original local-global representation learning by aggregating spatial global information and channel global information.

Application to Other Image Denoising Models

To demonstrate the versatility of the proposed lightweight framework (LIDFormer), a generalization analysis of our method is performed on three representative image denoising approaches: Restormer (Zamir et al. 2022), CBDNet (Guo et al. 2019) and NAFNet (Chen et al. 2022a). All these denoising models are retrained under the conditions of the original model. The results are presented in Table 2. It is shown that the proposed pipeline is generally applicable to

Methods	GFLOPs	SIDD PSNR / SSIM	DND PSNR / SSIM
Restormer	140.00	40.02 / 0.9600	40.03 / 0.9560
dwtRestormer	8.75	39.55 / 0.9326	39.65 / 0.9417
LIDFormer	2.83	39.62 / 0.9557	39.76 / 0.9558
CBDNet	34.00	39.30 / 0.9214	39.35 / 0.9351
dwtCBDNet	2.20	27.68 / 0.7214	27.54 / 0.7134
LiCBDNet	2.00	39.01 / 0.9014	39.06 / 0.9117
NAFNet	65.00	39.77 / 0.9524	39.81 / 0.9561
dwtNAFNet	4.00	39.43 / 0.9317	39.48 / 0.9342
LiNAFNet	4.20	39.51 / 0.9437	39.62 / 0.9525

Table 2: Generalization results of the efficient framework in LIDFormer for different image denoising methods. Among them, Restormer is based on Transformer while CBDNet and NAFNet are based on convolutional neural networks.

Methods	GFLOPs / Params	SIDD PSNR / SSIM	DND PSNR / SSIM
DnCNN	- / 0.56M	23.66 / 0.5830	32.43 / 0.7900
FFDNet	- / 0.48M	- / -	34.40 / 0.8474
CBDNet	34 / 4.34M	33.28 / 0.8680	38.06 / 0.9421
RIDNet	196.52 / 1.49M	- / -	39.26 / 0.9528
VDN	99.00 / 7.81M	39.26 / 0.9550	39.38 / 0.9518
DANet	14.85 / 9.15M	39.25 / 0.9160	39.47 / 0.9548
DeamNet	146.36 / 2.23M	39.35 / 0.9550	39.63 / 0.9555
InvDN	47.80 / 2.64M	39.28 / 0.9550	39.57 / 0.9522
Thunder	18.81 / 2.68M	39.47 / 0.9570	39.57 / 0.9526
ADFNet	117.32 / 7.65M	**39.63** / **0.9580**	**39.87** / 0.9555
LIDFormer	**2.83** / 2.72M	39.62 / 0.9575	39.76 / **0.9558**

Table 3: Quantitative comparison of LIDFormer with other efficient and lightweight denoising methods. The best performance is bolded and the second is underlined. "GFLOPs" presents the computational cost for peer 256 × 256 images. "Params" means the number of model parameters.

existing denoising methods, both convolutional neural networks and Transformers. Compared with the original model, the performance of the model after computational complexity reduction has a slight decrease, but the computational complexity has been optimized by more than 16 times, indicating that our pipeline is an effective and universally efficient method.

Comparison with State-of-the-Art Methods

We compare the proposed LIDFormer with popular state-of-the-art efficient and lightweight methods for real-world image denoising, including DnCNN (Zhang et al. 2017), FFDNet (Zhang, Zuo, and Zhang 2018), CBDNet (Guo et al. 2019), RIDNet (Anwar and Barnes 2019), VDN (Yue et al. 2019), DANet (Yue et al. 2020), DeamNet (Ren et al. 2021), InvDN (Liu et al. 2021a), Thunder (Zhou et al. 2022) and ADFNet (Shen, Zhao, and Zhang 2023). The compared results are shown in Table 3.

From the table, it can be concluded that our LIDFormer achieves the best results in terms of computational complexity and performance compromise. In particular, the performance of ADFNet (Shen, Zhao, and Zhang 2023) is slightly better than our method, but the FLOPs cost is more than ×

Figure 6: Visual comparison of LIDFormer with other efficient and lightweight denoising methods on SIDD.

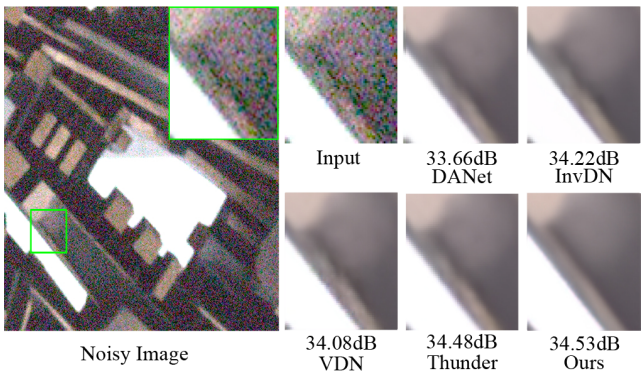

Figure 7: Visual comparison of LIDFormer with other efficient and lightweight denoising methods on DND.

40 ours. Therefore, LIDFormer achieves a better trade-off between high performance and low computational complexity. Moreover, the visual comparisons of our proposed LIDFormer with other methods are given in Fig. 6 and Fig. 7. Our proposed method is not inferior to other efficient and lightweight denoising methods in terms of visual effect.

Conclusion

In this paper, we propose an efficient and lightweight image denoising method named LIDFormer. LIDFormer includes three parts: feature lightweighting based on double Discrete Wavelet Transform (DWT), Complementary Periodic Feature Reusing (CPFR) and Triple Multi-Dconv Head Transposed Attention (TMDTA). Among them, the feature lightweighting based on the double DWT is used to transform the input image into a low-resolution space for low-computing operation; the CPFR module effectively amplifies feature information in low-resolution space and alleviates catastrophic forgetting; the TMDTA mechanism enhances the interaction of feature information and relieves the computational complexity of full-pixel self-attention. The qualitative and quantitative experimental results indicate that LIDFormer can achieve a "low computational complexity" level close to advanced semantic tasks while maintaining high performance. Moreover, the efficient framework in LIDFormer can be generalized to other image denoising methods for effective optimization.

Acknowledgments

This work was supported by the National Key Research and Development Program of China No.2020AAA0108301, the National Natural Science Foundation of China under Grant No.62176224, No.62222602, No.62176092, the Natural Science Foundation of Chongqing under No.CSTB2023NSCOJOX0007, and the CCF-Lenovo Blue Ocean Research Fund.

References

Abdelhamed, A.; Lin, S.; and Brown, M. S. 2018. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1692–1700.

Anwar, S.; and Barnes, N. 2019. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3155–3164.

Anwar, S.; Khan, S.; and Barnes, N. 2020. A deep journey into super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 53(3): 1–34.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 213–229. Springer.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12299–12310.

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022a. Simple baselines for image restoration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, 17–33. Springer.

Chen, Y.; and Pock, T. 2016. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 39(6): 1256–1272.

Chen, Z.; Zhang, Y.; Gu, J.; Kong, L.; Yuan, X.; et al. 2022b. Cross Aggregation Transformer for Image Restoration. *Advances in Neural Information Processing Systems*, 35: 25478–25490.

Cheng, S.; Wang, Y.; Huang, H.; Liu, D.; Fan, H.; and Liu, S. 2021. Nbnet: Noise basis learning for image denoising with subspace projection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4896–4906.

Dabov, K.; Foi, A.; Katkovnik, V.; and Egiazarian, K. 2008. Image restoration by sparse 3D transform-domain collaborative filtering. In *Image Processing: Algorithms and Systems VI*, volume 6812, 62–73. SPIE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1): 5232–5270.

Graham, B.; El-Nouby, A.; Touvron, H.; Stock, P.; Joulin, A.; Jégou, H.; and Douze, M. 2021. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12259–12269.

Gu, S.; Zhang, L.; Zuo, W.; and Feng, X. 2014. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2862–2869.

Guo, S.; Yan, Z.; Zhang, K.; Zuo, W.; and Zhang, L. 2019. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1712–1722.

He, K.; Sun, J.; and Tang, X. 2010. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12): 2341–2353.

Jin, Y.; Jiang, X.-B.; Wei, Z.-k.; and Li, Y. 2019. Chest X-ray image denoising method based on deep convolution neural network. *IET Image Processing*, 13(11): 1970–1978.

Kumar, M.; Weissenborn, D.; and Kalchbrenner, N. 2020. Colorization Transformer. In *International Conference on Learning Representations*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021a. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Liang, Y.; Chongjian, G.; Tong, Z.; Song, Y.; Wang, J.; and Xie, P. 2021b. EViT: Expediting Vision Transformers via Token Reorganizations. In *International Conference on Learning Representations*.

Liu, Y.; Qin, Z.; Anwar, S.; Ji, P.; Kim, D.; Caldwell, S.; and Gedeon, T. 2021a. Invertible denoising network: A light solution for real noise removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13365–13374.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2016. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations*.

Mallat, S. G. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7): 674–693.

Mao, X.; Shen, C.; and Yang, Y.-B. 2016. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in neural information processing systems*, 29.

Plotz, T.; and Roth, S. 2017. Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1586–1595.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Ren, C.; He, X.; Wang, C.; and Zhao, Z. 2021. Adaptive consistency prior based deep network for image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8596–8606.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Shen, H.; Zhao, Z.-Q.; and Zhang, W. 2023. Adaptive dynamic filtering network for image denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2227–2235.

Tai, Y.; Yang, J.; Liu, X.; and Xu, C. 2017. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, 4539–4547.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9446–9454.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17683–17693.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Xu, J.; Zhang, L.; Zhang, D.; and Feng, X. 2017. Multi-channel weighted nuclear norm minimization for real color image denoising. In *Proceedings of the IEEE international conference on computer vision*, 1096–1104.

Xu, S.; Yang, X.; and Jiang, S. 2017. A fast nonlocally centralized sparse representation algorithm for image denoising. *Signal Processing*, 131: 99–112.

Yair, N.; and Michaeli, T. 2018. Multi-scale weighted nuclear norm image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3165–3174.

Yu, Y.; Chang, M.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2018. Image denoising algorithm based on adversarial learning using joint loss function. In *Fifth Conference on Frontiers in Optical Imaging Technology and Applications*, volume 10832, 204–210. SPIE.

Yuan, W.; Liu, H.; and Liang, L. 2023. Joint group dictionary-based structural sparse representation for image restoration. *Digital Signal Processing*, 104029.

Yue, Z.; Yong, H.; Zhao, Q.; Meng, D.; and Zhang, L. 2019. Variational denoising network: Toward blind noise modeling and removal. *Advances in neural information processing systems*, 32.

Yue, Z.; Zhao, Q.; Zhang, L.; and Meng, D. 2020. Dual adversarial network: Toward real-world noise removal and noise generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, 41–58. Springer.

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

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2020. Learning enriched features for real image restoration and enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, 492–511. Springer.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14821–14831.

Zhang, K.; Li, Y.; Zuo, W.; Zhang, L.; Van Gool, L.; and Timofte, R. 2021. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6360–6376.

Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; and Zhang, L. 2017. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7): 3142–3155.

Zhang, K.; Zuo, W.; and Zhang, L. 2018. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, 27(9): 4608–4622.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2020. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7): 2480–2495.

Zhou, Y.; Jiao, J.; Huang, H.; Wang, Y.; Wang, J.; Shi, H.; and Huang, T. 2020. When awgn-based denoiser meets real noises. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13074–13081.

Zhou, Y.; Xu, X.; Liu, S.; Wang, G.; Lu, H.; and Shen, H. T. 2022. Thunder: Thumbnail based Fast Lightweight Image Denoising Network. *arXiv preprint arXiv:2205.11823*.

Zou, B.; Zhang, Y.; Wang, M.; and Liu, S. 2023. Toward Efficient Image Denoising: A Lightweight Network with Retargeting Supervision Driven Knowledge Distillation. In *Advances in Computer Graphics: 39th Computer Graphics International Conference, CGI 2022, Virtual Event, September 12–16, 2022, Proceedings*, 15–27. Springer.