LKD-Net: Large Kernel Convolution Network for Single Image Dehazing

Pinjun Luo 1  Guoqiang Xiao 1  Xinbo Gao 2  Song Wu 1∗
1 College of Computer and Information Science, Southwest University, China
2 Chongqing University of Posts and Telecommunications, China
zhongjiuhuan@email.swu.edu.cn, gqxiao@swu.edu.cn, gaoxb@cqupt.edu.cn, songwuswu@swu.edu.cn

Abstract—The previous deep CNN-based single-image dehazing methods are devoted to improving the performance by increasing the network's depth and width. In this paper, a novel Large Kernel Convolution Dehaze Network (LKD-Net) is proposed to enhance the network's performance by increasing the size of the convolutional kernel. The main module in LKD-Net is the designed Large Kernel Convolution Dehaze Block (LKD Block), which consists of the Decomposition deep-wise Large Kernel Convolution Block (DLKCB) and the Channel Enhanced Feed-forward Network (CEFN). DLKCB is designed to reduce the massive amount of computational overhead and parameters of the large kernel by splitting the deep-wise large kernel convolution into a smaller depth-wise convolution and a depth-wise dilated convolution. Meanwhile, CEFN is designed to enhance the robustness of the Feed-forward Network by exploiting significant channels. The extensive experiments on RESIDE dataset demonstrate that our LKD-Net outperforms the state-of-the-art with far less computational overhead and parameters.

Index Terms—Dehazing, large convolutional kernel

I. INTRODUCTION

Single-image dehazing aims to estimate the latent haze-free image from the observed hazy image, leading to information degradation such as color distortion and reduced visibility. Such degradation can have severe consequences in certain scenarios, such as autonomous driving and adversarial attacks, making the removal of haze from a single image a challenge in computer vision.

Single-image dehazing aims to estimate the latent haze-free image from the observed hazy image. Recent deep learning-based approaches have achieved success in single image dehazing, and the backbone networks of these approaches can mainly be divided into two categories: those based on convolutional neural networks (CNNs) and those based on Vision Transformers (ViTs). Thanks to the multi-head attention mechanism (MHSA), Transformer-based models [1], [2] have demonstrated excellent results in single image dehazing. However, MHSA is a double-edged sword and has some drawbacks, such as high model complexity and the need for more training data. Some studies [3], [4] have also shown that the success of the Transformer should be mainly attributed to its larger effective receptive fields (ERFs) compared to CNNs, since larger ERFs capture more structured information in the learned latent domain space. If the ERFs of CNNs are increased to the same level as those of the Transformer, they can achieve comparable performance. One of the most effective ways to increase the ERFs of CNNs is to directly increase the size of the convolutional kernel. However, expanding the size of the convolutional kernel results in more computational overhead and parameters due to the quadratic computational complexity with kernel size.

This paper proposes a novel Large Kernel Convolution Dehaze Block (LKD Block) to address the aforementioned limitations. Specifically, motivated by the convolution decomposition mechanism [5], [6], we design a Decomposition of the Large Kernel Convolution Block (DLKCB) to replace the MHSA in ViTs. DLKCB increases the effective receptive field and builds long-range information between features without adding more computational overhead and parameters by decomposing the traditional large depth-wise kernel convolution into a combination of a small depth-wise convolution and a depth-wise dilated convolution. Furthermore, we design a Channel Enhanced Feed-forward Network (CEFN), which integrates a Channel Attention mechanism [7] into the conventional Feed-forward Network (FN) to improve the efficiency of network optimization by exploiting significant and critical channels in FN. By combining DLKCB and CEFN, we create our Large Kernel Convolution Dehaze Block (LKD Block), which can be treated as a plug-in and added to the deep architecture of CNNs and ViTs for both high-level and low-level computer vision tasks. In this paper, we add the LKD Block to multiple U-Net-like dehazing networks for high-
Figure 1 displays a comparison between LKD-Net and state-of-the-art methods on the SOTS indoor set. Our LKD-Net outperforms the previous Swin Transformer-based approach [2] on the SOTS [8] indoor set with significantly fewer computational overhead and parameters. Additionally, LKD-Net has the same scale-up capability as the Transformer-based method. The main contributions of our LKD-Net can be summarized as follows:

We propose LKD-Net, a highly efficient end-to-end multiple U-Net-like deep architecture for single-image dehazing that outperforms state-of-the-art methods with significantly fewer parameters and lower computational overhead. We design the Large Kernel Convolution Dehaze Block (LKD Block), which can be used as a plug-in module to enhance the performance of both CNNs and Transformer architectures. Meanwhile, the LKD Block is more efficient and effective for the single image dehazing task than Transformer-based methods. We design the Decomposition of Large Kernel Convolution Block (DLKCB), which decomposes the large depth-wise convolution into a small depth-wise convolution and a depth-wise dilated convolution to increase the effective receptive field without significantly increasing parameters and computational overhead. We design the Channel Enhanced Feed-forward Network (CEFN), which can effectively explore and integrate channels with more critical information in FN, further improving the robustness and efficiency of network optimization.

II. RELATED WORK

We classify deep learning-based single-image dehazing methods into two categories based on their backbone networks: CNNs and ViTs.

a) CNN-based Image Dehazing Methods: Due to their high semantic abstraction capability, translation invariance property, parameter sharing, and local perception, CNNs are well-suited for low-level tasks such as dehazing. For instance, DehazeNet [9] employs CNNs to estimate the medium transmission map and restore hazy images by the atmosphere scattering model. FFA-Net [10] utilizes CNNs to build a feature fusion attention mechanism to process different channel information and pixel information flexibly. AECR-Net [11] improves on FFA-Net by using downsampling and contrastive learning. YOLY [12] separates hazy images into scene radiance layer, transmission map layer, and atmospheric light layer through an unsupervised approach, and then obtains hazy-free images using an atmospheric scattering model [13], [14]. ZID [15] adopts a zero-shot approach to disentangle a given hazy image into its haze-free version, transmission map, and atmospheric light through three joint subnetworks. However, these CNN-based methods primarily focus on increasing the depth and width of the network instead of on the kernel size. This is mainly because directly expanding the size of the convolutional kernel results in introducing more computational overhead and parameters, considering the quadratic computational complexity with the size of the kernel.

b) Transformer-based Image Dehazing Methods: Since Dosovitskiy et al. [16] introduced Transformer to computer vision, ViTs have surpassed CNN-based methods in various tasks, including single image dehazing. For instance, Song et al. [1] proposed Dehazeformer, which uses the Swin Transformer as the backbone and outperforms all previous CNN-based methods by a large margin on the SOTS [8] dataset. Some methods [3], [4] transform CNNs into Transformer-like CNN architectures and have also yielded promising results in various fields. However, these methods usually spend a lot of resources on processing tokens and overlook the fact that different channels have entirely different information weights in FN, resulting in a less efficient network.

III. METHOD

A. Overall Architecture

Our main objective is to develop an efficient deep learning model that can restore hazy observation images to haze-free images. As depicted in Figure 2 (a), LKD-Net is a U-Net-like architecture that contains multiple LKD Blocks, forming a multi-scale hierarchical framework. This framework has the significant advantage of improving performance while maintaining no additional computational cost. The downsample layer uses non-intersecting convolution to divide the image into small patches with non-intersecting and increase the number of channels. The upsample layer uses Pixelunshuffle [17] to aggregate the corresponding downsample layer patches and reduce the number of channels. SK Fusion [1] replaces the concatenation fusion layer and uses a channel attention mechanism to fuse the features of different branches. The Soft Reconstruction [1] layer replaces global residual learning and introduces a weakly constrained prior to global residual learning, resulting in better network performance.

B. Large Kernel Convolution Dehaze Block

As illustrated in Figure 2 (b), our LKD Block mainly contains two modules: the designed DLKCB and CEFN. DLKCB processes the spatial dimension information, which benefits the network in retaining more spatial structure information by increasing the effective receptive field. CEFN processes information in the channel dimension and improves the efficiency of network optimization by using channel attention [7]. CEFN treats different channels unequally and makes the network pay more attention to the channels with more critical information. Moreover, our LKD Block can be considered a Transformer-like CNN plug-in, using the designed DLKCB to replace the MHSAs in Transformer to enhance performance, and using the CEFN to replace the FN in CNNs to improve efficiency. Thus, LKD Block can be used as a plug-in module in both CNNs and ViTs for high-level and low-level computer vision tasks. Furthermore, our DLKCB architecture has the same scale-up capability as Transformer-based [16], [18] methods, enabling our network to be more adaptable to devices with different computational performances than traditional CNNs. Detailed experimental results will demonstrate the efficiency of our LKD Block compared to other architectures.
C. Decomposition Large Kernel Convolution Block

A large receptive field is crucial for capturing structured information in the feature domain space [11], which is essential for image dehazing. Increasing the depth of the network by stacking several small convolutions (e.g., 3 × 3 convolutions) is the most popular approach to increase the receptive field. However, this approach can increase the theoretical receptive field but is limited in increasing the effective receptive fields (ERFs). Recent work [20], [21] has demonstrated that different channel features in neural networks possess distinct weightings, indicating that certain channel features are less critical for network optimization. If these channel features are treated equally, it may lead to the allocation of excessive resources to less important information, which could negatively impact the network’s performance. Additionally, when replacing the Multi-Head Self-Attention (MHSA) module with the Decomposed Large Kernel Convolution Block (DLKCB) module, the network loses the ability to dynamically adjust the weights, potentially affecting the model’s generalization ability. To remedy this issue, we propose the Channel-wise Enhanced Feed-forward Network (CFN), illustrated in Figure 2 (b). CFN incorporates channel attention into the conventional feed-forward network (FN), enabling it to re-weight different channel features. Furthermore, inspired by [5], [22], we integrate a 3 × 3 depth-wise convolution into the conventional FN to capture the spatial correlations between neighboring feature pixels. The formula for CFN is presented as follows:

$$K_{DWDC} = \left\lfloor \frac{K}{d} \right\rfloor \times \left\lfloor \frac{K}{d} \right\rfloor$$

$$K_{DWC} = (2d - 1) \times (2d - 1)$$

Here, $K$ denotes the size of the convolution kernel being decomposed, and $d$ is the dilation rate. $K_{DWDC}$ denotes the depth-wise dilated convolution, and $K_{DWC}$ denotes the depth-wise convolution. The number of parameters, $P(K, d)$, and the number of floating-point operations (FLOPs), $F(K, d)$, for the decomposition of a large depth-wise convolution are expressed as follows:

$$P(K, d) = C\left(\left\lfloor \frac{K}{d} \right\rfloor^2 \times C + (2d - 1)^2\right)$$

$$F(K, d) = P(K, d) \times H \times W.$$

Where $K$ denotes kernel size and $d$ is dilation rate.

As shown in Figure 5, we find that the decomposition of large kernel depth-wise convolution obtains larger ERFs in practice. The detail can be seen in the ablation study.
(5)

\[ \hat{X} = X + \text{norm}(\text{norm}(F_{\text{N}}(X) \odot CA(X) \cdot \alpha)), \]

Where \( \odot \) denotes element-wise multiplication. \( X \in \mathbb{R}^{H \times W \times C} \) and \( \hat{X} \in \mathbb{R}^{H \times W \times C} \) are the input and output feature maps. \( \text{norm} \) is batch normalization. \( \alpha \) is a learnable scaling parameter. \( F_{\text{N}} \) is the Feed-forward Network. \( CA \) denotes channel attention, and its formula is expressed as follows:

\[ CA(X) = \sigma(\text{Linear}(\text{ReLU}((\text{Linear}(\text{GAP}(X)))))), \]

Here, \( \sigma \) is the sigmoid function, and \( \text{GAP} \) denotes the global average pooling operation.

IV. EXPERIMENTS

A. Implementation Details

We classify our LKD-Net into LKD-T, LKD-S, LKD-B, and LKD-L according to the number of their parameters and computational overhead, which correspond to tiny, small, basic, and large, respectively. The source code of our LKD-Net is available at GitHub\(^1\). Table I lists the configuration details of these variants. All models are implemented with PyTorch 1.10.1 on two NVIDIA TITAN Xp GPUs. The AdamW \([24]\) optimizer is utilized to optimize our LKD-Net with exponential decay rates \( \beta_1 \) and \( \beta_2 \) equals to 0.9 and 0.999, respectively. The initial learning rate is set to 0.0002, and the cosine annealing strategy is used to adjust the learning rate. The batch size is set to 16, and the patch size is set to \( 256 \times 256 \) with a random crop. We only use \( L_1 \) loss to optimize our LKD-Net. We decompose the \( 21 \times 21 \) convolution by default, which is proven to have the best parameter-performance trade-off in the work \([5]\).

![Fig. 4. Qualitatively comparing image dehazing methods on SOTS set (zooming in for a better view). The first row is SOTS indoor images, and the last row is outdoor images.](Image)

TABLE I

| Model | Num. Blocks | Embedding Dims | MLP Ratio |
|-------|-------------|----------------|-----------|
| LKD-T | [1, 1, 2, 1, 1] | [24, 48, 96, 48, 24] | [4, 4, 4, 4, 4] |
| LKD-S | [2, 2, 4, 2, 2] | [24, 48, 96, 48, 24] | [4, 4, 4, 4, 4] |
| LKD-B | [4, 4, 4, 4, 4] | [24, 48, 96, 48, 24] | [4, 4, 4, 4, 4] |
| LKD-L | [8, 8, 16, 8, 8] | [24, 48, 96, 48, 24] | [4, 4, 4, 4, 4] |

\(1\)https://github.com/SWU-CS-MediaLab/LKD-Net

B. Datasets and Evaluation Metrics

RESIDE dataset \([8]\) is used for the evaluation, which contains five subsets: Indoor Training Set (ITS), Outdoor Training Sets (OTS), Synthetic Objective Testing Set (SOTS), Real World task-driven Testing Set (RTTS), and Hybrid Subjective Testing Set (HSTS). Following the objective evaluation protocol \([1], [10], [11]\), Our LKD-Net is trained on ITS for 300 epochs and OTS for 30 epochs, respectively, and evaluated on the SOTS subset. Meanwhile, the Peak Signal Noise Ratio (PSNR) and the Structure Similarity index (SSIM) are used to evaluate the performance of our LKD-Net and the compared state-of-the-art methods.

C. Results on RESIDE Dataset

We quantitatively compare the performance of our LKD-Net and the state-of-the-art image dehazing methods, including the DCP \([25]\), DehazeNet \([9]\), AOD-Net \([26]\), GFN \([27]\), and GridDehazeNet \([28]\), MSBDN \([29]\), PFNet \([30]\), FFA-Net \([10]\), AECR-Net \([11]\), UDN \([31]\), Dehamer \([2]\), MAXIM \([32]\). The comparison results are shown in Table II. It can be seen that our LKD-L outperforms all methods on the SOTS indoor dataset. It outperforms the previous best method UDN by 0.86 dB in the PSNR evaluation metric with only 56% of the number of parameters. In particular, compared with the Transformer-based method Dehamer, LKD-L outperforms Dehamer by 3.08 dB in the PSNR evaluation metric with only 1.79% of the number of parameters and 48.9% of the FLOPs, which demonstrates that the Transformer-based method may not be the optimal option for low-level computer vision tasks. In addition, benefiting from the fantastic architecture of LKD-Block, all variants of LKD-Net have achieved good performance. Thus, LKD-Net is a scalable method, making it adaptable to various devices with different performances. On the SOTS outdoor dataset, our LKD-T outperforms all methods with much less #Param and FLOPs, indicating that LKD-Net has a faster convergence speed.

The qualitative evaluation is also implemented, and the visualization results are shown in Figure 4. We can see that GridDehazeNet and FFA-Net cannot successfully remove the haze from the images. Dehamer performs well in outdoor scenes. However, it also suffers from color distortion in indoor scenes (e.g., the background color of the characters changes in the first row). In contrast, MAXIM performs well in indoor scenes. Therefore, our experiments demonstrate that our LKD-Net is more effective than the FN applied to Transformers \([1], [16], [18]\), and MLPs \([23]\).
scenes. However, color oversaturation occurs in outdoor scenes (e.g., oversaturation of train lights in the second column). Compared to them, the recovered images from our LKD-Net are significantly closer to Ground Truth.

Fig. 5. The Effective Receptive Fields (ERFs) visualization results in different kernels. (a) 9 × 9 large kernel depth-wise convolution. (b) 21 × 21 large kernel depth-wise convolution. (c) 21 × 21 decomposition of large kernel depth-wise convolution. A more widely distributed red area indicates a larger ERF. Compared to LK 9 × 9 and LK 21 × 21, our DLK 21 × 21 obtains larger ERFs.

D. Ablation Study

Ablation studies are conducted to demonstrate the effectiveness of each proposed component in LKD-Net, and the minimal model LKD-T is used for the ablation analysis. We first construct a Base network as our baseline, which is implemented by replacing 21 × 21 decomposition of depth-wise convolutions with 7 × 7 depth-wise convolutions, replacing CEFN with the regular Feed-forward Network (FN), replacing SK Fusion with concatenation, and replacing SR with global residual learning.

Subsequently, we replace different modules into base network construct four different variants: (1) Base+SF: Replace concatenation in the Base with SK Fusion. (2) Base+SF+SR: Replace global residual learning in the Base+SF with Soft Reconstruction. (3) Base+SF+SR+DLK: Replace 7 × 7 depth-wise convolutions in the Base+SF+SR with 21 × 21 decomposition of depth-wise convolutions. (4) Base+SF+SR+CEFN: Replace the regular Feed-forward Network (FN) in the Base+SF+SR with CEFN. (5) Ours: Replace 7 × 7 depth-wise convolutions in the Base+SF+SR+CEFN with 21 × 21 decomposition of depth-wise convolutions. These models are trained on ITS dataset and evaluated on SOTS indoor set. The performance of these models is summarized in Table III.

Effectiveness of DLK. Compared to Base+SF+SR, the DLK can significantly increase the PSNR by 1.95dB and SSIM by 0.007, while only introducing 0.022M #Param and 0.33G FLOPs. We further performed ablation experiments with other conventional large depth-wise convolutions to demonstrate the advantages of our DLK. Table IV shows that our DLK achieves better performance using less computational overhead and parameters than 9 × 9 depth-wise convolution and 21 × 21 depth-wise convolution. We argue that the main reason for the high effectiveness of DLK is that it can capture larger ERFs compared to conventional large depth-wise convolutions. To support our viewpoint, we use the public tool (code is available at [4]) to visualize the ERF of the feature map centroids of the LKD-Net L₄ output. As shown in Figure 5, we can observe that the distribution of red dots in DLK 21 × 21 is larger and more widely than LK 9 × 9 and LK 21 × 21, which indicates that DLK 21 × 21 obtains larger ERF indeed. We also present quantitative analysis in Table V, where t stands a threshold.

### Table II

| Methods | ITS | OTS | Overhead |
|---------|-----|-----|---------|
|         | SOTS indoor | SOTS outdoor | #Param | FLOPs |
| DCP (TPAMI’10) | 16.62 | 0.818 | - | - |
| DehazeNet (TIP’16) | 21.14 | 0.847 | - | - |
| AOD-Net (ICCV’17) | 19.06 | 0.850 | - | - |
| GFN (CVPR’18) | 22.30 | 0.880 | - | - |
| GridDehazeNet (ICCV’19) | 32.16 | 0.984 | - | - |
| MSBDN (CVPR’20) | 33.67 | 0.985 | - | - |
| PFDN (ECCV’20) | 32.68 | 0.976 | - | - |
| FFA-Net (AAAI’20) | 36.39 | 0.989 | - | - |
| AEKR-Net (CVPR’21) | 37.17 | 0.990 | - | - |
| UDN (AAA’22) | 38.62 | 0.991 | - | - |
| Dehame (CVPR’22) | 36.36 | 0.988 | - | - |
| Maxim (CVPR’22) | 38.11 | 0.991 | - | - |

**Table III**

**Ablation study on LKD-T with different architectures.**

| Model | PSNR | SSIM | #Param | FLOPs |
|-------|------|------|--------|-------|
| Base  | 29.37 | 0.998 | 0.010M | 2.98G |
| Base+SF | 31.74 | 0.977 | 0.314M | 3.06G |
| Base+SF+SR | 33.83 | 0.934 | 0.315M | 3.07G |
| Base+SF+SR+DLK | 33.78 | 0.985 | 0.334M | 3.40G |
| Base+SF+SR+CEFN | 33.38 | 0.983 | 0.334M | 3.27G |
| Ours  | 34.77 | 0.987 | 0.343M | 3.41G |
For example, if $t = 20\%$ and $r = 4.4\%$ that means 20% of the pixel contributions reside in 4.4% total pixel area. We can see that DLK $21 \times 21$ has a smoother distribution of high contributing pixels compared to LK $9 \times 9$ and LK $21 \times 21$, which proves our viewpoint that the main reason for the high effectiveness of DLK is the effectively captured larger ERFs.

### TABLE IV

| Kernel size | PSNR | SSIM | #Param | FLOPs |
|-------------|------|------|--------|-------|
| LK $9 \times 9$ | 33.93 | 0.985 | 0.342M | 3.45G |
| LK $21 \times 21$ | 34.31 | 0.987 | 0.466M | 5.430G |
| DLK $21 \times 21$ | 34.77 | 0.987 | 0.343M | 3.14G |

### TABLE V

Quantitative analysis on the ERF with the high-contribution area ratio $t$. A larger $r$ indicates a smoother distribution of high-contribution pixels. Hence, larger ERFs.

| Kernel size | $t = 20\%$ | $t = 30\%$ | $t = 50\%$ | $t = 90\%$ |
|-------------|-------------|-------------|-------------|-------------|
| LK $9 \times 9$ | 4.1% | 7.3% | 17.1% | 95.3% |
| LK $21 \times 21$ | 6.6% | 12.4% | 27.4% | 95.4% |
| DLK $21 \times 21$ | 6.6% | 13.5% | 30.8% | 98.4% |

**Effectiveness of CEFN.** Compared to Base+SF+SR, CEFN significantly improves performance with a 1.55 dB increase in PSNR and a 0.005 increase in SSIM and only introduces 0.019M #Param and 0.2G FLOPs. We believe that the main reason for the high effectiveness of CEFN is that the channel attention mechanism [7] allows CEFN to focus more on the channels with important information.

### V. CONCLUSION

This paper proposes a novel LKD-Net for high-performance single image dehazing. The designed DLKCB can effectively capture ERFs and model long-range information, and the designed CEFN can effectively enhance channel dimension features in FN. Evaluation results show that LKD-Net outperforms the state-of-the-art and dramatically outperforms the Transformer-based method Dehamer. Thus, we argue that our LKD-Net is an effective and universal end-to-end image restore method, and can be used for video dehazing and other low-level computer vision tasks such as image denoising, rain removal, deblurring, super-resolution, etc. Moreover, the decomposition of depth-wise convolutions in DLKCB may be used in CNNs and ViTs to enhance the performance of both low-level and high-level vision tasks.

### REFERENCES

[1] Y. Song, Z. He, H. Qian, and X. Du, “Vision transformers for single image dehazing,” arXiv preprint arXiv:2204.03883, 2022.

[2] C.-L. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, “Image dehazing transformer with transmission-aware 3d position embedding,” in CVPR, 2022, pp. 5812–5820.

[3] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A convnet for the 2020s,” in CVPR, 2022, pp. 11,976–11,986.

[4] X. Ding, X. Zhang, J. Han, and G. Ding, “Scaling up your kernels to 31x31: Revisiting large kernel design in cnns,” in CVPR, 2022, pp. 11,963–11,973.

[5] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, “Visual attention network,” arXiv preprint arXiv:2202.09741, 2022.

[6] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, “Restorner: Efficient transformer for high-resolution image restoration,” arXiv preprint arXiv:2111.09881, 2021.