IEICE TRANSACTIONS

on Information and Systems

This advance publication article will be replaced by the finalized version after proofreading.

Vision Transformer with Key-select Routing Attention for Single Image Dehazing

Lihan TONG[†], **Weijia LI**[††], **Qingxia YANG**[†], **Liyuan CHEN**[†], *and* **Peng CHEN**[†a)], *Nonmembers*

SUMMARY   We present Ksformer, utilizing Multi-scale Key-select Routing Attention (MKRA) for intelligent selection of key areas through multi-channel, multi-scale windows with a top-k operator, and Lightweight Frequency Processing Module (LFPM) to enhance high-frequency features, outperforming other dehazing methods in tests.
key words: single image dehazing, Multi-scale Key-select Routing Attention Module, Lightweight Frequency Processing Module, vision transformer

1.   Introduction

Single image dehazing [3]–[5] aims to restore clear, high-quality images from hazy ones, essential for applications like object detection [2] and semantic segmentation [1]. Traditional methods [6], [8], [28] may not give ideal dehazing results because they can't cover all scenarios [15]. With the rise of deep learning, convolutional neural networks (CNNs) [4], [5], [16] have been widely applied to image dehazing and have achieved good results. However, because CNNs cannot capture long-range dependencies, this limits further improvement in dehazing effects. Recently, Transformers [17], [18], [20], [21] have been widely used in computer vision tasks because they can capture long-range dependencies. However, they have a problem where their computational complexity is proportional to the square of the image resolution. Many efforts [21]–[24] have been made to address this issue by introducing handcrafted sparsity. But the sparsity added by hand doesn't relate to the content, causing some loss of information.

We propose Ksformer, which is made up of MKRA and LFPM. MKRA estimates queries in windows of different sizes and then uses a top-k operator to select the most important k queries. This approach enhances computational efficiency and incorporates content-aware capabilities. Meanwhile, multi-scale windows adeptly manage blurs of varying sizes. LFPM employs lightweight parameters to extract spectral features. The contributions of this work are summarized as:

- Ksformer is content-aware, selecting key-value pairs with important information to minimize content loss, while also capturing long-range dependencies and reducing computational complexity.

- Ksformer extracts spectral features with ultra-lightweight parameters, performing MKRA in both spatial and frequency domains and then fusing them, which narrows the gap between clean and hazy images in terms of both space and spectrum.
- Ksformer achieves a PSNR of 39.4 and an SSIM of 0.998 with only 5.8M parameters, which is significantly better than other state-of-the-art methods.

2.   Method

2.1   Image Dehazing

We use three encoders and three decoders and downsample by 4×4 for a compact model. We use Multi-scale Key-select Routing Attention Module (MKRAM) only in the smaller dimensions to reduce computational complexity. To lower the difficulty of training [25], [26], we strengthen the exchange of information between layers and use skip connections at both the feature and image levels.

2.2   Multi-scale Key-select Routing Attention

MKRA uses a top-k operator to select the most important key-value pairs, balancing content awareness with lower computational complexity. For any given input feature map $X \in R^{H \times W \times C}$, first, we divide it into four parts along the channel dimension, with window sizes of 2×2, 4×4, 8×8, 64×64. Then, it is divided into $S \times S$ non-overlapping regions. Each region contains $\frac{HW}{S^2}$ feature vectors. After this step, X is reshaped into $X^r \in R^{S^2 \times \frac{HW}{S^2} \times \frac{C}{4}}$. Then, we use linear projection to weight and derive $Q, K, V \in R^{S^2 \times \frac{HW}{S^2} \times \frac{C}{4}}$.

$$Q = X^r W^q, K = X^r W^k, V = X^r W^v, \tag{1}$$

Here, $W^q, W^k, W^v \in R^{\frac{C}{4} \times \frac{C}{4}}$ represent the weights for Q, K, and V, respectively. We construct an Attention module to identify the areas where important key-value pairs are located. In simple terms, we use the average values of each region to derive region-level queries and keys, $Q_r, K_r \in R^{s^2 \times \frac{C}{4}}$. Then, we derive the region-to-region importance association matrix using the following formula.

$$A_r^{n \times n} = \text{Softmax}\left(\frac{Q_r (K_r)^T}{\sqrt{\frac{C}{4}}}\right), \tag{2}$$

[†]The authors are with the school of Ocean Information Engineering, Jimei University, Xiamen, China
[††]The author is with the school of Computer Science, Jimei University, Xiamen, China
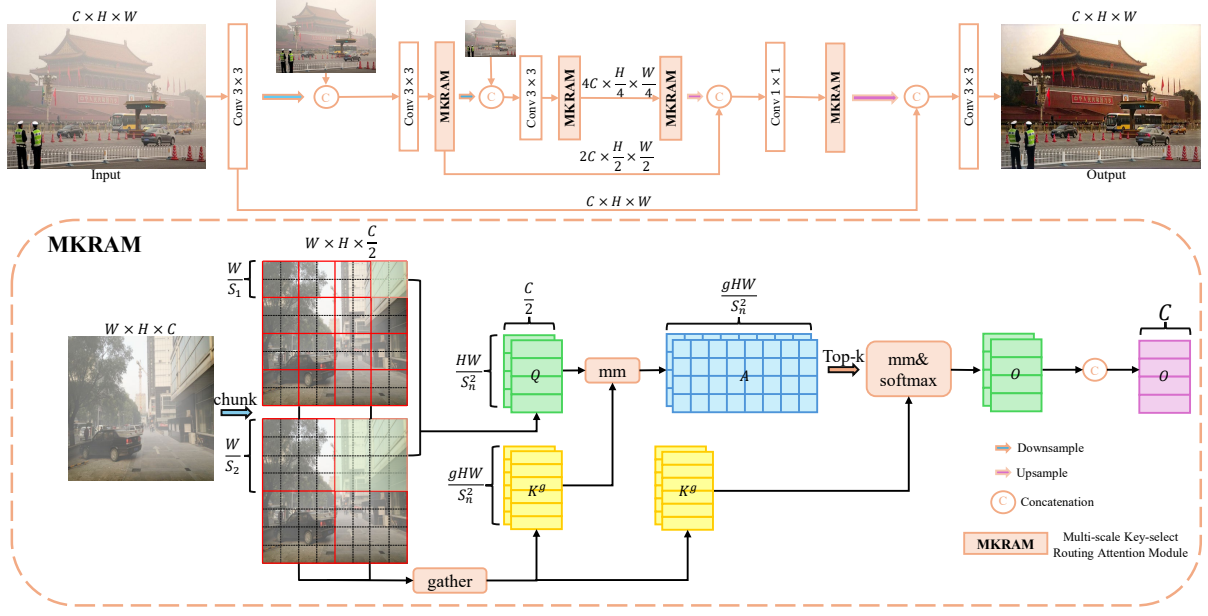a) E-mail: chenpeng@jmu.edu.cn(Corresponding author)

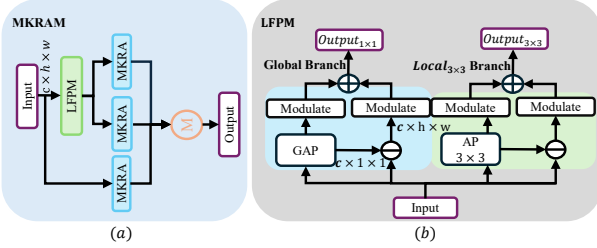**Fig. 1** The architecture of the proposed Ksformer.

Fig. 2 (a) is the architecture of the proposed MKRAM. (b) is the architecture of the proposed LFPM.

Here, $A_r^{n \times n} \in R^{S^2 \times \frac{C}{4}}$, represents the degree of association between two regions. $n \times n$ represents the size of the window. Next, we concatenate $A_r^{n \times n}$ along the channel dimension to obtain $A_r \in R^{S^2 \times C}$. Next, we retain the top k most important queries using the top-k operator and prune the association graph to derive the index matrix.

$$I_r = \text{topk}(A_r), \quad (3)$$

Here, $I_r \in R^{S^2 \times K}$. So, the i-th row of I contains the k indices of the most relevant regions for the i-th region. Using the importance index matrix I_r, we can capture long-range dependencies, be content-aware, and reduce computational complexity. For each query and token in region i, it will focus on all key-value pairs in the union of k important regions indexed by $I_r^{(i,1)}, I_r^{(i,2)}, \ldots, I_r^{(i,k)}$. We collect the key and value tensors.

$$K^g = \text{gather}(K, I_r), V^g = \text{gather}(V_r), \quad (4)$$

We collect the key and value tensors. Here $K^g, V^g \in R^{S^2 \times \frac{kWH}{S^2} \times C}$, Finally, we focus our attention on the collected key-value pairs.

Table 1 Quantitative comparisons with SOTA methods on the RESIDE-Indoor [34] and Haze4K [27] datasets.

Method	RESIDE-IN [34]		Haze4k [27]		# Param	# GFLOPs
	PSNR (dB)	SSIM	PSNR (dB)	SSIM		
(ICCV'17) AOD-Net [3]	19.82	0.818	17.15	0.83	0.002M	-
(ICCV'19) GridDehazeNet [7]	32.16	0.984	-	-	0.96M	-
(AAAI'20) FFA-Net [4]	36.39	0.989	26.96	0.95	4.68M	288.34
(CVPR'20) MSBDN [9]	33.79	0.984	22.99	0.85	31.35M	41.58
(CVPR'20) KDDN [13]	34.72	0.985	-	-	5.99M	-
(CVPR'21) AECR-Net [5]	37.17	0.990	-	-	2.61M	26.10
(CVPR'22) Dehamer [14]	36.63	0.988	-	-	-	59.14
(ECCV'22) PMNet [10]	38.41	0.990	33.49	**0.98**	18.9M	-
Ksformer(Ours)	**39.40**	**0.994**	**33.74**	**0.98**	5.8M	92.12

$$\text{Output} = \text{Attention}(Q, K^g, V^g). \quad (5)$$

3. Lightweight Frequency Processing Module

Past approaches often employed wavelet [29]–[32] or Fourier [33] transformations to divide image features into multiple frequency sub-bands. These approaches raised the computational load for reverse transformation and didn't boost key frequency elements. To solve this, we added a lightweight module for spectral feature extraction and modulation. It efficiently splits the spectrum into different frequencies and uses a small number of learnable parameters to emphasize the most informative ones.

4. Multi-scale Key-select Routing Attention Module

As shown in Fig.2, to improve model efficiency, our MKRAM module processes the spatial domain, low frequencies, and high frequencies in parallel and then fuses the three outputs.

| AOD-Net | FFA-Net | MSBDN | KDDN | Dehamer | ours | hazy |

Fig. 3 Visual results comparisons on RTTS [34] dataset. Zoom in for best view.

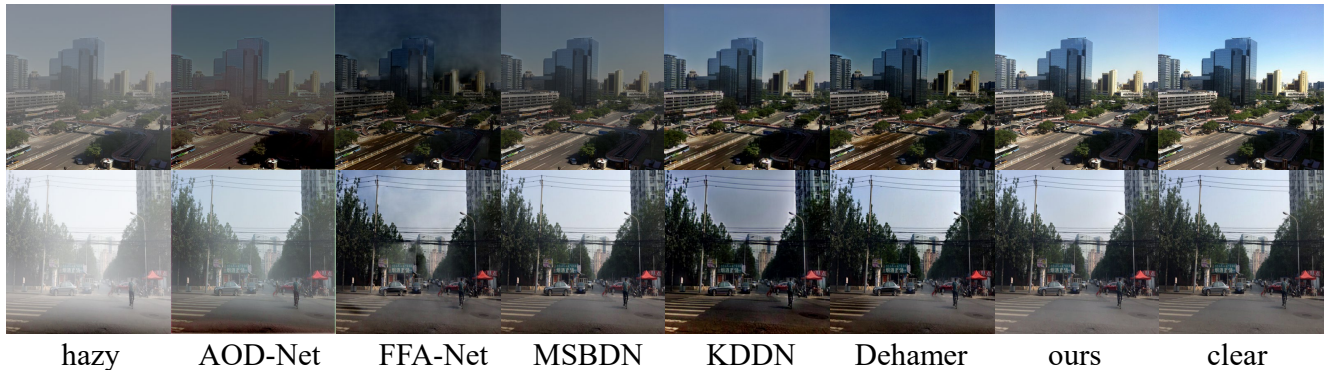| hazy | AOD-Net | FFA-Net | MSBDN | KDDN | Dehamer | ours | clear |

Fig. 4 Visual results comparisons on Haze4K dataset [27]. Zoom in for best view.

Table 2 Ablation study of our Ksformer on the Haze4k Dataset [27].

Model	PSNR (dB)	SSIM
Base(U-Net)	25.46	0.92
Base+MKRA	32.25	0.94
Base+LFPM	28.52	0.92
Base+MKRA+LFPM	33.23	0.95
Base+MKRAM (Full)	**33.74**	**0.98**

5. Experiments

5.1 Implementation Details

During our experiments, we employ PyTorch version 1.11.0 and utilize the capabilities of four NVIDIA RTX 4090 GPUs to perform all tests. In the training phase, images are randomly cropped into 320×320 pixel patches. For assessing the model's computational complexity, we adopt a size of 128×128 pixels. The Adam optimizer is engaged for optimization, with decay rates set at 0.9 for β_1 and 0.999 for β_2. The initial learning rate is configured at 0.00015, and we apply a cosine annealing strategy for its scheduling. The batch size is maintained at 64. Through empirical determination, we set the penalty parameter λ to 0.2 and γ to 0.25, and we proceed with training for 80,000 iterations.

5.2 Quantitative and Qualitative Experiments

Visual Comparison. To thoroughly assess our method, we tested it on both the synthetic Haze4K [27] dataset and the real-world RTTS [34] dataset. As shown in Fig.3 and Fig.4, it's clear that our method outperforms others in terms of edge sharpness, color fidelity, clarity of texture details, and handling of sky areas, whether on synthetic or real datasets. Quantitative Comparison. We quantitatively compared Ksformer with the current state-of-the-art methods on the SOTS indoor [34] and Haze4K [27] datasets. As shown in Table.1, for the SOTS indoor [34] dataset, Ksformer achieved a PSNR of 39.40 and an SSIM of 0.994, which is a 0.09 PSNR improvement over the second-best method, and it did so with only 30% of the parameter volume. For the Haze4K [27] dataset, Ksformer reached a PSNR of 33.74 and an SSIM of 0.98. The quantitative comparison fully demonstrates that Ksformer outperforms other state-of-the-art methods in terms of performance.

5.3 Ablation Study

To prove the effectiveness of our method, we conducted an ablation study. We first built a U-Net as the base network, and then gradually added modules to the baseline. As shown in Table.2, both PSNR and SSIM improved with the step-by-step addition of modules, and the metrics reached their best values after effectively combining the modules we proposed.

6. Conclusion

This paper introduces Ksformer, which combines a top-k operator with multi-scale windows, giving the network the char-

acteristics of content awareness and low complexity. At the same time, it obtains spectral features with ultra-lightweight parameters, narrowing the spectral gap between clean and foggy images. On the SOTS indoor [34] dataset, it achieved a PSNR of 39.4 and an SSIM of 0.994 with only 5.8M.

Although the Ksformer has a relatively small parameter count of just 5.8 million, it's unfortunate that it can't be implemented on embedded systems due to its high GFLOPs. We plan to further explore the balance between performance and computational complexity. By appropriately reducing the number of channels and modules, we aim to make the Ksformer suitable for embedded systems, allowing it to play a significant role in a broader range of fields.

Acknowledgments

References

[1] S. Hao and Y. Zhou and Y. Guo, *A brief survey on semantic segmentation with deep learning*, *Neurocomputing*, volume 406, pages 302–321, 2020.

[2] Z. Zou and K. Chen and Z. Shi and Y. Guo and J. Ye, *Object detection in 20 years: A survey*, *Proc. IEEE*, volume 111, number 3, pages 257–276, 2023.

[3] B. Li and X. Peng and Z. Wang and J. Xu and D. Feng, *Aod-net: All-in-one dehazing network*, in *Proc. IEEE Int. Conf. Comput. Vis.*, pages 4770–4778, 2017.

[4] X. Qin and Z. Wang and Y. Bai and X. Xie and H. Jia, *FFA-Net: Feature fusion attention network for single image dehazing*, in *Proc. AAAI Conf. Artif. Intell.*, volume 34, number 07, pages 11908–11915, 2020.

[5] H. Wu and Y. Qu and S. Lin and J. Zhou and R. Qiao and Z. Zhang and Y. Xie and L. Ma, *Contrastive learning for compact single image dehazing*, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 10551–10560, 2021.

[6] K. He, J. Sun, and X. Tang, *Single image haze removal using dark channel prior*, *IEEE Trans. Pattern Anal. Mach. Intell.*, volume 33, number 12, pages 2341–2353, 2010.

[7] X. Liu and Y. Ma and Z. Shi and J. Chen, *Griddehazenet: Attention-based multi-scale network for image dehazing*, in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 7314–7323, 2019.

[8] D. Berman and S. Avidan et al., *Non-local image dehazing*, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1674–1682, 2016.

[9] H. Dong and J. Pan and L. Xiang and Z. Hu and X. Zhang and F. Wang and M.-H. Yang, *Multi-scale boosted dehazing network with dense feature fusion*, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 2157–2167, 2020.

[10] T. Ye and M. Jiang and Y. Zhang and L. Chen and E. Chen and P. Chen and Z. Lu, *Perceiving and modeling density is all you need for image dehazing*, arXiv preprint arXiv:2111.09733, 2021.

[11] B. Li and W. Ren and D. Fu and D. Tao and D. Feng and W. Zeng and Z. Wang, *Benchmarking single-image dehazing and beyond*, *IEEE Trans. Image Process.*, volume 28, number 1, pages 492–505, 2021.

[12] Y. Liu and L. Zhu and S. Pei and H. Fu and J. Qin and Q. Zhang and L. Wan and W. Feng, *From synthetic to real: Image dehazing collaborating with unlabeled real data*, in *Proc. 29th ACM Int. Conf. Multimedia*, pages 50–58, 2021.

[13] M. Hong and Y. Xie and C. Li and Y. Qu, *Distilling image dehazing with heterogeneous task imitation*, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 3462–3471, 2020.

[14] C. Guo and Q. Yan and S. Anwar and R. Cong and W. Ren and C. Li, *Image dehazing transformer with transmission-aware 3D position embedding*, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 5812–5820, 2022.

[15] K. Zhang and W. Ren and W. Luo and W.-S. Lai and B. Stenger and M.-H. Yang and H. Li, *Deep image deblurring: A survey*, *Int. J. Comput. Vis.*, volume 130, number 9, pages 2103–2130, 2022.

[16] Y. Cui and Y. Tao and Z. Bing and W. Ren and X. Gao and X. Cao and K. Huang and A. Knoll, *Selective frequency network for image restoration*, in *Proc. 11th Int. Conf. Learn. Represent.*, 2022.

[17] N. Parmar and A. Vaswani and J. Uszkoreit and L. Kaiser and N. Shazeer and A. Ku and D. Tran, *Image transformer*, in *Proc. Int. Conf. Mach. Learn.*, PMLR, pages 4055–4064, 2018.

[18] H. Chen and Y. Wang and T. Guo and C. Xu and Y. Deng and Z. Liu and S. Ma and C. Xu and C. Xu and W. Gao, *Pre-trained image processing transformer*, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 12299–12310, 2021.

[19] J. Liang and J. Cao and G. Sun and K. Zhang and L. Van Gool and R. Timofte, *Swinir: Image restoration using swin transformer*, in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 1833–1844, 2021.

[20] J. Ke and Q. Wang and Y. Wang and P. Milanfar and F. Yang, *Musiq: Multi-scale image quality transformer*, in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 5148–5157, 2021.

[21] Z. Liu and Y. Lin and Y. Cao and H. Hu and Y. Wei and Z. Zhang and S. Lin and B. Guo, *Swin transformer: Hierarchical vision transformer using shifted windows*, in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 10012–10022, 2021.

[22] Z. Wang and X. Cun and J. Bao and W. Zhou and J. Liu and H. Li, *Uformer: A general u-shaped transformer for image restoration*, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 17683–17693, 2022.

[23] S. W. Zamir and A. Arora and S. Khan and M. Hayat and F. S. Khan and M.-H. Yang, *Restormer: Efficient transformer for high-resolution image restoration*, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 5728–5739, 2022.

[24] Y. Qiu and K. Zhang and C. Wang and W. Luo and H. Li and Z. Jin, *MB-TaylorFormer: Multi-branch efficient transformer expanded by Taylor formula for image dehazing*, in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 12802–12813, 2023.

[25] X. Mao and Y. Liu and W. Shen and Q. Li and Y. Wang, *Deep residual Fourier transformation for single image deblurring*, arXiv preprint arXiv:2111.11745, volume 2, number 3, pages 5, 2021.

[26] Z. Tu and H. Talebi and H. Zhang and F. Yang and P. Milanfar and A. Bovik and Y. Li, *Maxim: Multi-axis mlp for image processing*, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 5769–5780, 2022.

[27] Y. Liu and L. Zhu and S. Pei and H. Fu and J. Qin and Q. Zhang and L. Wan and W. Feng, *From synthetic to real: Image dehazing collaborating with unlabeled real data*, in *Proc. 29th ACM Int. Conf. Multimedia*, pages 50–58, 2021.

[28] Q. Zhu and J. Mai and L. Shao, *Single image dehazing using color attenuation prior*, in *Proc. BMVC*, pages 1–10, 2014, organization Citeseer.

[29] I. W. Selesnick and R. G. Baraniuk and N. C. Kingsbury, *The dual-tree complex wavelet transform*, *IEEE Signal Process. Mag.*, volume 22, number 6, pages 123–151, 2005.

[30] H. Yang and Y. Fu, *Wavelet u-net and the chromatic adaptation transform for single image dehazing*, in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pages 2736–2740, 2019.

[31] W. Chen and H. Fang and C. Hsieh and C. Tsai and I. Chen and J. Ding and S. Kuo, *All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss*, in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 4196–4205, 2021.

[32] W. Zou and M. Jiang and Y. Zhang and L. Chen and Z. Lu and Y. Wu, *Sdwnet: A straight dilated network with wavelet transformation for image deblurring*, in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 1895–1904, 2021.

[33] H. Yu and N. Zheng and M. Zhou and J. Huang and Z. Xiao and F. Zhao, *Frequency and spatial dual guidance for image dehazing*, in *Proc. Eur. Conf. Comput. Vis.*, pages 181–198, 2022.

[34] B. Li and W. Ren and D. Fu and D. Tao and D. Feng and W. Zeng and Z. Wang, *Benchmarking single-image dehazing and beyond*, *IEEE Trans. Image Process.*, volume 28, number 1, pages 492–505, 2021.