Tokenize Image Patches: Global Context Fusion for Effective Haze Removal in Large Images

Supplementary Material

6. More Details of Experiments

In our experiments, we utilize 3 datasets: *8KDehaze*, *4KID*, and *O-HAZE*. The *4KID* dataset includes 3 subsets: Daytime, Night, and Realtime. For consistency with the other datasets, we select the Daytime subset for training and testing. Since both the *4KID* and *O-HAZE* datasets do not provide pre-split training and test sets, we randomly select 500 samples from the *4KID* dataset and 5 samples from the O-HAZE dataset as the test set. The remaining images from these datasets are used for training. To further enhance the generalization capability of the model, we apply random rotations to the input images during training, ensuring greater variability and robustness in the model's performance.

7. More Details of the *8KDehaze* dataset

To the best of our knowledge, the proposed *8KDehaze* dataset is the first ultra-high-resolution dataset in the field of image dehazing. It consists of 9,000 training pairs and 1,000 test pairs, each with a resolution of 8192×8192 pixels. This dataset offers a valuable resource for advancing large image inference in the dehazing domain. In this section, we provide a detailed description of the dataset's key features and discuss its potential impact on the development of dehazing algorithms.

Geographical Diversity. The *8KDehaze* dataset encompasses a wide range of geographical environments, ensuring a comprehensive representation of real-world scenarios. As shown in Figure 9, the images are sourced from diverse regions, including desert, inland waters, urban areas, rural landscapes, forest, coastlines, and mountainous terrains. This geographical diversity enables the dataset to capture the various ways in which haze manifests across different topographies and ecosystems. The distribution of images across these categories is uniform, ensuring that models trained on the *8KDehaze* dataset can generalize well to different types of scenes in practical applications.

Seasonal Variation. Another key aspect of the *8KDehaze* dataset is its seasonal variation. The dataset includes images captured in all four seasons: spring, summer, autumn, and winter. Figure 10 presents some representative samples from each season. These seasonal differences present unique challenges for dehazing algorithms. For example, models may struggle to differentiate between snow-covered ground and haze, as both can appear similarly gray or white. Additionally, the varying vegetation and atmospheric conditions across seasons require models to adapt to diverse environmental conditions. The inclusion of seasonal variation in the *8KDehaze* dataset is essential for developing robust dehazing algorithms capable of generalizing across different times of the year.

Haze Distribution. The *8KDehaze* dataset contains images with diverse haze characteristics. Figure 11 presents the haze distribution across the dataset, including varying coverage areas and haze intensities. Unlike conventional images captured from a ground-level perspective, aerial images often exhibit irregular and random haze patterns. This non-uniform haze distribution poses significant challenges for dehazing algorithms, as it requires models to effectively utilize global image information and adapt to a wide range of unpredictable haze patterns. Given that most existing dehazing methods primarily address uniform haze in small image patches, the *8KDehaze* dataset's large-sized, non-uniform haze distributions offer new opportunities for advancing the field of image dehazing.

Dataset Availability. We provide two versions of the dataset: a full version for training and evaluation, and a mini version for debugging. Both the full dataset and the mini version are publicly available at https://github.com/CastleChen339/DehazeXL.

8. Performance of CNN Backbone Models

As stated in Section Methodology, the proposed DehazeXL uses Swin Transformer [29] as the backbone for both the Encoder and Decoder. Given the advancements made by CNN-based backbones in the dehazing domain, we conduct additional experiments in this section using several CNN-based backbones for both the Encoder and Decoder. We select the following CNN-based architectures for comparison: VGG-19 [39], ResNet-50 [19], DenseNet-201 [21], and EfficientNet-B4 [44]. The models were trained and evaluated on the *8KDehaze*, *4KID*, and *O-HAZE* datasets under the same experimental settings as DehazeXL. The quantitative results are summarized in Table 4 to 6.

The results indicate that models using CNN-based architectures as both Encoder and Decoder achieve competitive performance in terms of PSNR and SSIM, demonstrating the general applicability of the proposed framework. However, models based on Swin Transformer as the backbone outperform these CNN-based models. This superior performance underscores the advantages of Transformer-based Encoder-Decoder architectures, which are more effective at capturing long-range dependencies compared to traditional

Figure 9. Geographic Diversity of the *8KDehaze* Dataset. The samples in the *8KDehaze* dataset cover seven distinct terrain types: desert, inland waters, urban areas, rural areas, forest, coast, and mountain areas.

Figure 10. Seasonal Variation in the *8KDehaze* Dataset. Samples can be categorized into 4 seasons: spring, summer, autumn, and winter.

Table 4. Performance comparison of different backbones on the *8KDehaze* dataset in terms of PSNR and SSIM

Model Backbone	PSNR	SSIM
VGG-19	27.15	0.9389
ResNet-50	28.18	0.9583
DenseNet-201	29.49	0.9620
EfficientNet-B4	30.15	0.9746
Swin-T (DehazeXL)	**32.35**	**0.9863**

Table 5. Performance comparison of different backbones on the *4KID* dataset in terms of PSNR and SSIM

Model Backbone	PSNR	SSIM
VGG-19	23.17	0.8782
ResNet-50	24.20	0.8903
DenseNet-201	24.81	0.8892
EfficientNet-B4	25.17	0.8946
Swin-T (DehazeXL)	**26.62**	**0.9073**

CNN architectures. These capabilities make Transformer-based models better suited for image dehazing tasks. Therefore, we choose Swin Transformer as the default encoder and decoder backbone for DehazeXL. Notably, our method achieves state-of-the-art performance without any modifications to the original Swin Transformer, further validating the potential of the proposed framework.

9. Additional Visual Results

In this section, we present additional visual results to further highlight the effectiveness of the proposed DehazeXL model for large image dehazing. Figure 12 illustrates the comparative results of all methods on the *8KDehaze*, *4KID*, and *O-HAZE* datasets. Figure 13 showcases the attribution maps for DehazeXL's dehazed results using the proposed

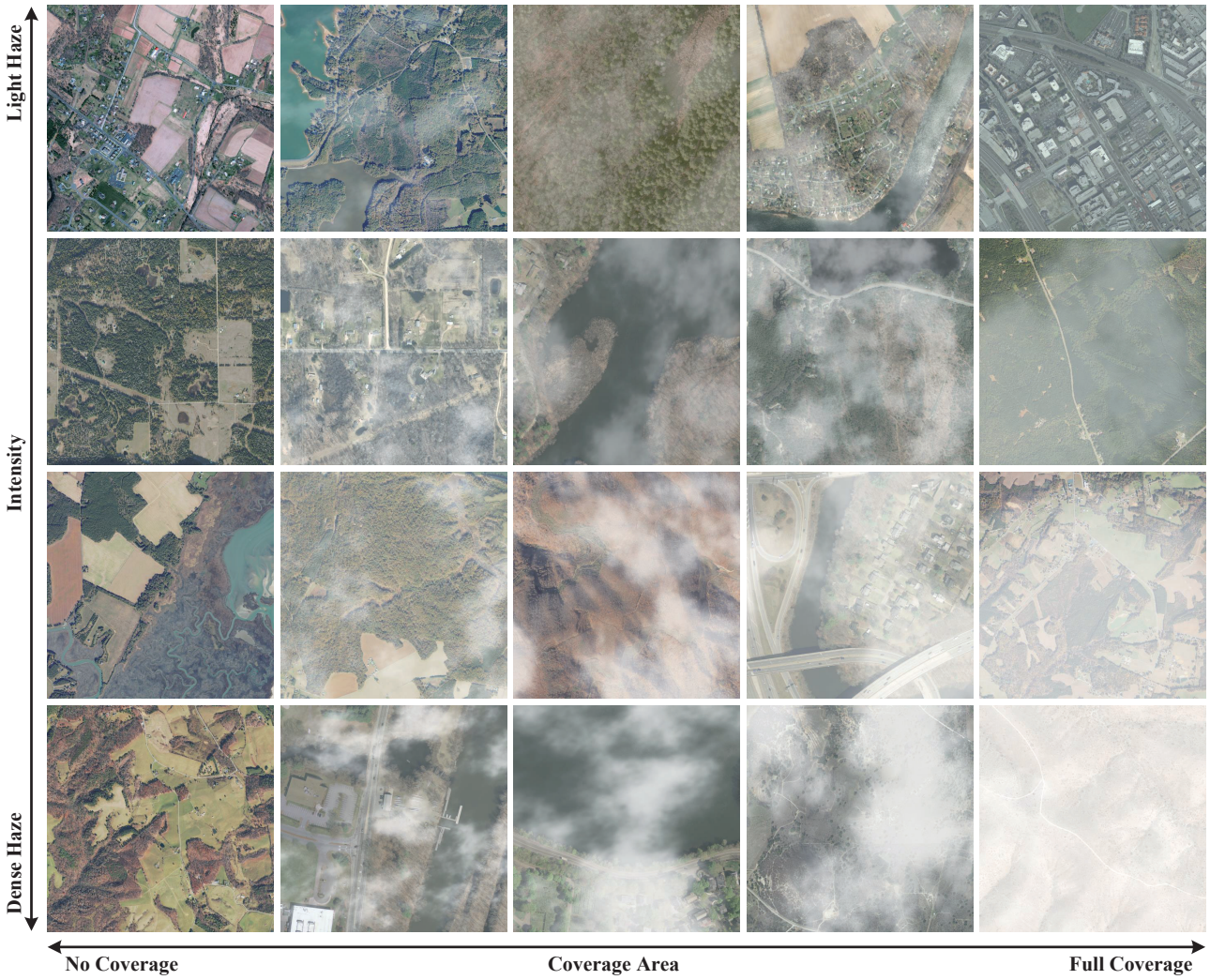


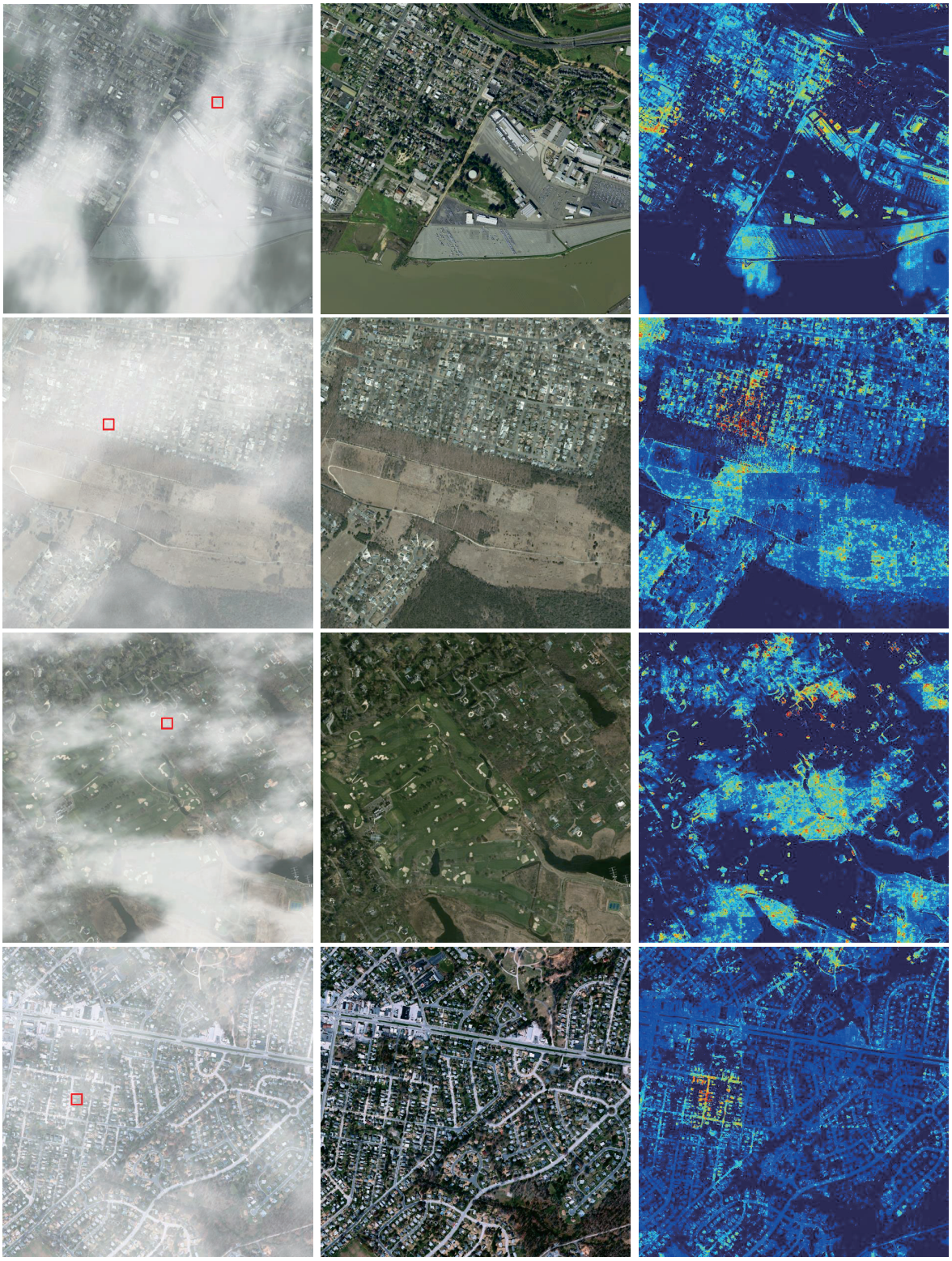
Figure 11. Haze Distribution in the *8KDehaze* Dataset. The x-axis represents the haze coverage area, ranging from no coverage to full coverage, while the y-axis indicates the haze intensity, spanning from low to high.

Table 6. Performance comparison of different backbones on the O-HAZE dataset in terms of PSNR and SSIM

Model Backbone	PSNR	SSIM
VGG-19	19.95	0.7056
ResNet-50	20.62	0.7184
DenseNet-201	20.88	0.7215
EfficientNet-B4	21.06	0.7330
Swin-T (DehazeXL)	**21.49**	**0.7348**

DAM. It offers insight into the specific contributions of each pixel to the dehazed results in the specified region. The source code of DAM is available at https://github.com/fengyanzi/DehazingAttributionMap.

Figure 12. Comparisons of dehazed results on the *8KDehaze*, *4KID*, and *O-HAZE* datasets.



(a) Inputs (b) Dehazed Results (c) Attribution Maps

Figure 13. Attribution maps for DehazeXL's dehazed results using the proposed DAM method. The red box on (a) indicate the regions of interest for attribution. In the attribution maps, the color intensity corresponds to the degree of influence on the dehazed results, with warmer colors (e.g., red) indicating higher influence and cooler colors (e.g., blue) indicating lower influence.