Complementary Feature Enhanced Network with Vision Transformer for Image Dehazing

Dong Zhao, Jia Li, Senior Member, IEEE, Hongyu Li, Long Xu

Abstract— Conventional CNNs-based dehazing models suffer from two essential issues: the dehazing framework (limited in interpretability) and the convolution layers (content-independent and ineffective to learn long-range dependency information). In this paper, firstly, we propose a new complementary feature enhanced framework, in which the complementary features are learned by several complementary subtasks and then together serve to boost the performance of the primary task. One of the prominent advantages of the new framework is that the purposively chosen complementary tasks can focus on learning weakly dependent complementary features, avoiding repetitive and ineffective learning of the networks. We design a new dehazing network based on such a framework. Specifically, we select the intrinsic image decomposition as the complementary tasks, where the reflectance and shading prediction subtasks are used to extract the color-wise and texture-wise complementary features. To effectively aggregate these complementary features, we propose a complementary features selection module (CFSM) to select the more useful features for image dehazing. Furthermore, we introduce a new version of vision transformer block, named Hybrid Local-Global Vision Transformer (HyLoG-ViT), and incorporate it within our dehazing networks. The HyLoG-ViT block consists of the local and the global vision transformer paths used to capture local and global dependencies. As a result, the HyLoG-ViT introduces locality in the networks and captures the global and long-range dependencies. Extensive experiments on homogeneous, non-homogeneous, and nighttime tasks, where the reflectance and shading prediction subtasks are used to extract the color-wise and texture-wise complementary features. To effectively aggregate these complementary features, we propose a complementary features selection module (CFSM) to select the more useful features for image dehazing. Furthermore, we introduce a new version of vision transformer block, named Hybrid Local-Global Vision Transformer (HyLoG-ViT), and incorporate it within our dehazing networks. The HyLoG-ViT block consists of the local and the global vision transformer paths used to capture local and global dependencies. As a result, the HyLoG-ViT introduces locality in the networks and captures the global and long-range dependencies. Extensive experiments on homogeneous, non-homogeneous, and nighttime

dehazing network (EPDN) [15], GridDehazeNet (GDNet) [16], multi-scale boosted dehazing network (MSBDN) [17] and feature fusion attention network (FFA) [18]. Trained in an end-to-end fashion, these models directly recover the haze-free image. However, the ASM cannot completely represent the complex hazy imaging process since it ignores the emission and absorption of atmospheric particles, leading to ineffective dehazing results [13]. Another category is the fully learning-based [11, 14] (as illustrated in Fig. 1 (b)), including enhanced Pix2pixHD dehazing network (EPDN) [15], GridDehazeNet (GDNet) [16], multi-scale boosted dehazing network (MSBDN) [17] and feature fusion attention network (FFA) [18]. Trained in an end-to-end fashion, these models directly recover the haze-free image. However, they have limitations in interpretability [11] and usually appear ineffective in dense haze removal. As to the second aspect, on the one hand, convolution is content-independent as the same convolution kernel is shared to the entire image, ignoring the distinct nature between different image regions [19]. On the other hand, due to inductive biases such as locality, the convolutions are ineffective to learn long-range dependency information [20, 21]. The above discussions inspire us to provide a new framework and a more powerful mechanism to replace some of the convolution layers.
Atmospheric Scattering Model

Hybrid Local-Global Vision Transformer (HyLoG-ViT)

Complementary Feature Enhanced Framework

II. RELATED WORK

A. Single Image Dehazing

Single image dehazing is an ill-posed problem. Without any auxiliary information, previous dehazing algorithms require the introduction of specific prior knowledge. Through research in the last decade, many priors have been proposed based on the cues of color (such as the DCP [3], CAP [4], and HLP [5]) and texture (such as the CoDP [6] and GCP [7]). For example, Berman et al. [5] find that for a haze-free image, its pixels that belong to the same color cluster form to a point while forming a line in a hazy image, namely the haze-line. The GCP uses the image gradient to estimate the depth information and the atmospheric light, preserving texture details more efficiently.
Recently, CNNs-based dehazing networks have been extensively studied. One kind of CNNs-based network restores the clear image by estimating the intermediate variables in the atmospheric scattering model [32] and then calculates the clear image. For example, the multi-scale CNNs dehazing model (MSCNN) [9] utilizes a coarse-scale network to estimate the complete transmission map and use a fine-scale network to refine the results. More recent CNNs-based networks tend to learn hazy-to-clear image translation directly [33]. For example, the MSBDN [17] removes haze in an end-to-end manner by incorporating boosting and error feedback principles into a U-Net [34] architecture with a dense feature fusion module.

Unlike the previous dehazing frameworks, our method relies neither on the atmospheric scattering model nor on a completely black box system. The subnetworks of intrinsic image decomposition provide color- and texture-wise complementary features, enabling the dehazing subnetwork to yield haze-free images with natural color and fine details.

Non-homogeneous and nighttime dehazing. For different complicated haze patterns, the dehazing method should be specifically designed, such as the TBNN [27], DW-GAN [28] and EDN [29] for non-homogeneous dehazing, and MRP [30] and OSFD [31] methods for nighttime dehazing. Unlike these models, training on different datasets, our network exhibits good generalization performances on homogeneous, non-homogeneous, and nighttime dehazing tasks without any changes in the network architecture.

B. Vision Transformer

Very recently, vision transformers [24] have received increasing research interest in image and video vision tasks, including object detection [35], image classification [24] and semantic segmentation [20]. Many new versions of vision transformers [36, 37, 38, 39] have been proposed to relieve the high computational cost problems. For example, the Swin Transformer [37] uses a locally-grouped self-attention [26], where the input features are separated into a grid of non-overlapped windows and the vision transformer is operated only within each window. Many other methods have been proposed to bring inductive biases into the vision transformer [26, 40, 41, 42], such as the LocalViT [41], which brings a locality mechanism to ViT by employing the depth-wise convolution into the feed-forward network. CvT [40] proposes a convolutional token embedding to model local spatial contexts and a convolutional projection layer to provide efficiency benefits.

New versions of vision transformers are also developed and used in low-level vision tasks. For example, the Uformer [43] designs a general U-shaped Transformer-based structure for image restoration. It also proposes a locally-enhanced window Transformer block to reduce the computation cost. The SwinIR [21] utilizes residual Swin Transformer blocks to extract deep features for image restoration.

In our HyLoG-ViT, the local vision transformer path is similar to the LeWin transformer proposed in Uformer. However, the local version used in Uformer alone fails to effectively model global dependencies and preserve the local continuity around those regions. By contrast, our HyLoG-ViT block can capture both local and global dependencies simultaneously.

III. METHOD

A. Complementary Features Enhanced Framework

To achieve high performance of haze removal, we propose a novel framework from a new perspective, namely the complementary feature enhanced framework. In this framework, each group of complementary features are learned by corresponding complementary tasks, and together they serve the primary task (i.e. the dehazing task in this paper). Unlike previous CNNs-based approaches that learn haze-relevant features implicitly, complementary features here can be purposively chosen and learned by training under specific subnetworks and datasets. Problems that arise here are: which complementary features should be considered for image dehazing, and which complementary tasks are used to learn such features effectively?
Before answering these problems, we here analyse the essential criterion the complementary features should follow. First, each kind of complementary feature represents one interference factor closely related to the prime task. Second, the different complementary features should be weakly correlated with each other so that each complement subtask can focus on doing one thing more effectively. Otherwise, multiple subnetworks learn similar representations, making the network very inefficient. That is, ‘Each performs one’s own functions; together serve their common mission’.

**Complementary Features:** Intuitively, the images captured in hazy or foggy scenes suffer from some interference factors, and the most conspicuous ones are the variation of color and blurriness of edges. Many previous works have witnessed that the information of color [3, 4, 5] and texture [6, 7] are the crucial priors for estimating the haze distribution. Inspired by these observations, we endeavour to build the corresponding subnetworks to learn the color-wise and texture-wise complementary features.

**Complementary Tasks:** In further investigations, we fund that the intrinsic image decomposition (which decomposes an image into reflectance and shading, as shown in Fig.2 (a)) is a good candidate as the complementary task; as the reflectance component contains the actual color of the scenes, and the shading component contains the structure and texture information [45]. To demonstrate these, we analyse the characters of the reflectance and shading of the hazy and clear images on the homogeneous, non-homogeneous and nighttime haze datasets. Fig.2 (b) shows the $\Delta E$ maps and CIEDE2000 metrics of the hazy image $I_H$, hazy reflectance $I_{HR}$ and clear reflectance $I_{CR}$ compared with the clear image $I_C$. The investigations demonstrate that the clear reflectance $I_{CR}$ maintain the color information of clear image $I_C$ with very low $\Delta E$ and CIEDE2000. Similarly, we visualise the visible gradients maps and calculate the Rate of Newly Visible Edges $r_{nve}$ [47] of hazy shading $I_{HS}$, clear shading $I_{CS}$ and clear image $I_C$ compared with the hazy image $I_C$, as illustrated in Fig.2 (c). We find that the $r_{nve}$ of $I_{CS}$ is very close to the one of $I_C$ on the tree datasets, indicating that the shading $I_{CS}$ preserves rich edge and texture information.

Finally, we choose the Unsupervised Learning for Intrinsic Image Decomposition (USI3D) [44] method to generate clear reflectance and shading samples in our training datasets. Most importantly, the USI3D assumes reflectance-shading independence, coinciding with the second criterion of complementary features as aforementioned. The reflectance predilection subnetwork in our model can focus on color-wise complementary features learning, avoiding the interferences of shading components; and vice versa.

**B. Dehazing Network**

In our dehazing model, the intrinsic image decomposition and dehazing are considered a joint model by exploring the former’s complementary features that the latter requires. Specifically, our network is a multi-task learning model with a decoder-focused architecture [48].

1) **Overall Architecture**

The overview of our dehazing network consists of a shared encoder and three decoders, as illustrated in Fig.3. Denote the hazy image is $I_H$, the shared encoder $E$ is used to extract the shallow and deep features:

$$e^0 = F^0_E(I_H), \quad e^z = F^z_E(e^{z-1}),$$

where $F^0_E$ denotes the feature extraction layer used to extract shallow feature $e^0$; $F^z_E$ denotes $z$-th stage of encoder $E$, $e^z$ refers to the deep feature at stage $z$, $z \in [1, \cdots, Z]$, and $Z$ is the total number of stages. Three parallel decoders follow the encoder. Decoders $D_R$ and $D_S$ are used to predict the reflectance and shading of the haze-free image, respectively, and their intermediate features are served as complementary features to the decoder $D_D$ to generate the high-quality haze-free image. The decoders are described as:

$$d^0_R = F^0_D(d^0_R), \quad d^z_R = F^z_D(d^{z-1}_R, e^z),$$

$$d^0_S = F^0_D(d^0_S), \quad d^z_S = F^z_D(d^{z-1}_S, e^z),$$

$$d^0_D = F^0_D(e^0, d^0_R, d^0_S), \quad d^z_D = F^z_D(d^{z-1}_D, d^z_R, d^z_S),$$
where \( F_{DB}^0 \), \( F_{DS}^0 \) and \( F_{DD}^0 \) denote the \( z \)-th stage of decoder \( D_R \), \( D_S \) and \( D_D \), respectively, \( z \in [1, \ldots, Z] \), \( d_R^z \), \( d_S^z \) and \( d_D^z \) are the intermediate features of the decoder \( D_R \), \( D_S \) and \( D_D \) at stage \( z \), respectively. The final reflectance \( I_R \), shading \( I_S \), and haze-free image \( I_D \) are generated through three image reconstruction layers, i.e., \( F_{DB}^0 \), \( F_{DS}^0 \) and \( F_{DD}^0 \), respectively:

\[
I_R = F_{DB}^0(d^1_R, c^0), \quad I_S = F_{DS}^0(d^1_S, c^0), \quad I_D = F_{DD}^0(d^1_D, d^1_R, d^1_S, c^0). \quad (6)
\]

2) Details of the Encoder and Decoder: The encoder \( E \) comprises of a feature extraction layer and a series of HyLoG-ViT blocks (see section III-D). The extraction layer is built by a \( 5 \times 5 \) convolution and a basic ResNet block [49] to extract shallow features. Each HyLoG-ViT block is followed by a \( 4 \times 4 \) convolution to downscale the spatial resolution with stride 2 and double the channel number. The convolutions used here can also bring the inductive bias into this ‘transformer-based’ encoder. In the decoder \( D_R \) and \( D_S \), except the bottom stage \( Z \), feature \( d_{RS}^{z-1} \) from the previous decoder block is first upscaled by a \( 4 \times 4 \) deconvolution to expand the spatial resolution with stride 2 and halve the channel number. Then, the output is concatenated with the feature \( c^z \) from the same stage of encoder \( E \). Therefore, the subnetworks \( E-D_R \) and \( E-D_S \) are formed into two U-shaped structures, which alleviates the issue of spatial information loss caused by downscaling. Different from decoder \( D_R \) and \( D_S \), for \( z \)-th stage in \( D_D \), there have three inputs: the feature \( d_{RD}^{z-1} \) from the previous stage; the complementary features \( d^z_R \) and \( d^z_S \) from the same stage of \( D_R \) and \( D_S \), respectively. These input features are fed into a complementary features selection module to select the most useful channels from \( d^z_R \) and \( d^z_S \) dynamically.

C. Complementary Features Selection Module (CFSM)

One can simply aggregate the complementary features \( d^z_R \) and \( d^z_S \) via element-wise summation or concatenation operation, which, however, is inefficient. Therefore, we propose the CFSM to fuse the complementary features in a nonlinear fashion. The architecture of the CFSM is illustrated in Fig.4.

Given the complementary features \( d^z_R \) and \( d^z_S \) and the feature \( d^z_{RD} \) \( (\in \mathbb{R}^{H \times W \times c}) \), they are first combined via element-wise summation. Then, the outputs are transformed to two channel-wise statistics \( s_{ave} \) and \( s_{max} \) \((\in \mathbb{R}^{1 \times 1 \times c}) \) by a global average pooling and a global max pooling, respectively. Each statistic is separated into two streams: one stream for \( d^z_R \) feature selection and another for \( d^z_S \). Take \( d^z_R \) stream as an example, the \( s_{ave} \) and \( s_{max} \) are followed by channel-downscaling \( 1 \times 1 \) convolution layers to calculate the compact feature vectors, \( v_{ave-R} \) and \( v_{max-R} \) \((\in \mathbb{R}^{1 \times 1 \times c/r}) \), respectively. The feature vectors are fed into two parallel channel-upsampling \( 1 \times 1 \) convolution layers and provide feature descriptors \( t_{ave-R} \) and \( t_{max-R} \) \((\in \mathbb{R}^{1 \times 1 \times c}) \). The final feature descriptor is defined as \( t_R = t_{ave-R} + t_{max-R} \). The \( t_R \) passes through the Sigmoid function to generate attention score \( a_R \) \((\in \mathbb{R}^{1 \times 1 \times c}) \) for \( d^z_R \). Similar, we can get the attention score \( a_S \) for \( d^z_S \). The overall process of feature recalibration and aggregation is defined as:

\[
\hat{d}^{-1}_{RD} = d^{-1}_{RD} + a_R \cdot d^z_R + a_S \cdot d^z_S. \quad (7)
\]

D. Hybrid Local-Global Vision Transformer (HyLoG-ViT)

We propose a HyLoG-ViT model to address the challenges of high computation cost and lack of locality. As shown in Fig.3, a HyLoG-ViT block consists of two paths. In the local vision transformer path, the input feature map is grouped into grids of non-overlapped regions, and the transformer block is applied only within each region. Given the feature maps \( X \in \mathbb{R}^{H \times W \times C} \) where the \( H \), \( W \) and \( C \) are the height, width and channel number of the maps, the computation of the local vision transformer path can be expressed as:

\[
X = \{X^1, X^2, \ldots, X^{N_l \times N_l}\};
\]

\[X^i_l = \text{Transformer}(X^i), \quad i \in [1, 2, \ldots, N_l \times N_l];\quad (8)
\]

\[X_l = \{X^1_l, X^2_l, \ldots, X^{N_l \times N_l}_l\},
\]

where \( X^i_l \) is the output of the \( i \)-th region, \( X_l \) is the combined output of the local vision transformer path; \( \text{Transformer}(\cdot) \) represents the standard transformer layer, as illustrated in Fig.5 (b). \( N_l \) is the region number per column/row. This design enables the vision transformer to focus on capturing region-level attention and exploring local context information. However, the local vision transformer path cannot model global dependencies and lose local continuity around those regions. Therefore, we also introduce the global vision transformer path, where the input feature is down-sampled by the average pooling operation to reduce the spatial resolution. The output is fed into the standard transformer layer. The computation of the global vision transformer path is:

\[
X_g = \text{Upsamp}_{g_{N_g}}(\text{Transformer}(\text{Avepool}_{g_{N_g}}(X))); \quad (9)
\]

where \( X_g \) is the output of the global vision transformer path. \( \text{Avepool}_{g_{N_g}}(\cdot) \) is the average pooling with reduction ratio \( N_g \); \( \text{Upsamp}_{g_{N_g}}(\cdot) \) represents the upsampling operation with upscale ratio \( N_g \). The global vision transformer path improves efficiency while still maintaining the capability of aggregating global information. The hybrid outputs of the two paths are concatenated and transformed to the original dimension by a \( 3 \times 3 \) convolution layer:

\[
X_h = \text{Conv}_{3 \times 3}(\text{Concat}(X_l, X_g)); \quad (10)
\]
where $X_h$ is the final output of HyLoG-ViT block; $\text{Concat}(\cdot)$ is the channel-wise concatenation; $\text{Conv}_{3 \times 3}(\cdot)$ is a $3 \times 3$ convolution layer. This fuse operation maintains the merits of the local and global vision transformer and introduces locality into the network.

**Positional Encoding:** The positional encoding used in transformers aims to retain positional information. However, for low-level vision tasks, the input images with varying sizes are commonly different from the training ones. As a result, it may decrease performance [51, 52] and break the translation invariance [26]. Note that our model can achieve good results even without additional positional encoding modules. We argue that the $3 \times 3$ convolution used in (10) is sufficient to provide positional information for Transformers.

**Complexity Analysis:** For the local vision transformer path, denote the input feature $X \in \mathbb{R}^{H \times W \times C}$ is grouped into $N_l \times N_l$ regions and the spatial-resolution of each region is $\frac{H}{N_l} \times \frac{W}{N_l} \times C$. Then, the complexity of the self-attention in the local vision transformer path is reduced from $O((HW)^2C)$ to $O\left((\frac{HW}{N_l})^2C\right)$ compared with the standard self-attention. For the global vision transformer path, denote the spatial dimension of the input feature is downscaled by a reduction ratio $N_g$, and the complexity is $O\left((\frac{HW}{N_l})^2C\right)$. Therefore, the total self-attention computational complexity of HyLoG-ViT block is $O\left((\frac{1}{N_l} + \frac{1}{N_g})(HW)^2C\right)$.

### E. Loss Function

In our training, the ground truths of reflectance $I_{CR}$ and shading $I_{CS}$ are generated by operating the pre-trained intrinsic image decomposition model USID [44] on the ground truth of clear image $I_C$. For reflectance and haze-free image estimations, we train the networks using the L2 reconstruction loss, conditional adversarial loss [53, 54], and SSIM loss (as used in [55]). For the shading estimation, we use the L2 reconstruction loss and an edge preserve loss which is defined as:

$$L_e = E_I \| \nabla_x I_S(p) - \nabla_x I_{CS}(p) \|_2 + \| \nabla_y I_S(p) - \nabla_y I_{CS}(p) \|_2,$$

(11)

where $E$ is the mean operation on a batch of samples; $\nabla_x$ and $\nabla_y$ are the spatial derivatives at $x$ and $y$ directions, respectively. $p$ refers to the spatial location. Overall, the hybrid loss function contains three parts, that is:

$$L = \lambda_R L_R + \lambda_S L_S + \lambda_D L_D,$$

(12)

where $L_R$, $L_S$, and $L_D$ are the loss functions for the three vision tasks, respectively.

### IV. Experiment

#### A. Experiment Setup

1) **Dataset:** Our dehazing model is trained on outdoor datasets, including RESIDE dataset [56], NH-HAZE dataset [57] and NHR dataset [31] for homogeneous haze, non-homogeneous haze and nighttime haze removal, respectively. For RESIDE dataset, we randomly select 41240 samples from OTS [56] (outdoor training subset in RESIDE) for training and 500 samples from SOTS [56] (synthetic objective testing subset in RESIDE) for testing. For the NH-HAZE dataset, we synthetic 9800 samples augmented from 50 original high-resolution samples by randomly cropping, flipping, and rotating; and the 41~45-th samples are used for qualitative evaluations. The NHR dataset contains 17900 samples; we select the last 475 for testing and others for training. We also collect 372 and 132 real-world daytime and nighttime hazy images to evaluate the performance of our model.

2) **Implemental Details:** Our dehazing model is implemented with PyTorch library and trained on one NVIDIA GeForce RTX 3090 GPU. Our model is trained for 32 epochs on the RESIDE dataset, 20 epochs on the NH-HAZE dataset and 5 epochs on the NHR dataset. We adopt Adam [58] optimizer with the initial learning rate is $10^{-4}$ for both CNN and vision transformer backbones. We use the layer normalization [59] in the vision transformer blocks and Activation Normalization [60] in CNN layers. The parameters of the loss functions are set as $(\lambda_R, \lambda_S, \lambda_D) = (1, 1, 1.5)$. For the HyLoG-ViT block, we set $N_l = 8$ and the patch size as $2 \times 2$ in the local vision transformer path; and in the global vision transformer path, we set $N_g = 4$ and the patch size as $4 \times 4$. The code of our model is available at [https://github.com/phoenixtreesky/CFEN-ViT-Dehazing](https://github.com/phoenixtreesky/CFEN-ViT-Dehazing).

3) **Metrics:** We use the peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) [61], learned perceptual image patch similarity (LPIPS) [62] and CIEDE2000 [46] to evaluate the dehazing methods on the synthetic datasets. LPIPS measures the ‘perceptual distance’ of the two compared images using deep features, which coincides well with human perceptual similarity judgments [63]. We use the CIEDE2000 to measure the color difference between the dehazed image and its corresponding clear image. We set the parameters $[K_L, K_1, K_2]$ in CIEDE2000 are $[2, 0.045, 0.015]$ in the experiments. Another metric we used is the bind contrast enhancement assessment [47] which contains three indicators, that are: 1) the rate of newly visible edges ($r_{nve}$) to evaluate the ability of edge restoration; 2) the geometric mean of the normalized gradient ($m_{ng}$) to measure the quality of the image contrast; 3) the rate of saturated pixels ($r_{sp}$) to measure the degree of over-saturation [47]. These three indicators are used to evaluate dehazing models on real-world datasets.
B. Comparisons with State-of-the-art Methods

1) Homogeneous Dehazing: To demonstrate the effectiveness of our dehazing model on homogeneous image dehazing, we firstly evaluate it on the synthetic datasets, including SOTS (outdoor) [56], O-HAZE [64], and HazeRD [65]. We compare our method with the prior-based and learning-based models, including DCP [3], HLP [5], MSCNN [9], GDNet [16], EPDN [15], MSBDN [17], and FFA [18].

It can be found from Fig.6 (i), over-saturation phenomenon emerges at the sky regions in the prior-based methods of DCP and HLP, indicating that those priors may be invalid under such conditions. The deep learning-based approaches can provide visually pleasant dehazed results on the SOTS dataset [56]. The metrics of PSNR, SSIM and LPIPS also demonstrate that GDNet and MSBDN obtain the top two scores on the SOTS dataset. However, those methods exhibit relatively poor performances on the other two challenging datasets, either leading to color distortion results (such as EPDN) or leaving haze at distant scenes (such as GDNet, MSBDN and FFA), as illustrated in Fig.6 (ii) and (iii). Our model obtains the best PSNR, SSIM, LPIPS and CIEDE2000 on the O-HAZE and HazeRD datasets.

We further evaluate our method on real-world datasets. The visual comparisons are shown in Fig.7. The prior-based methods, i.e. DCP and HLP, tend to yield over-enhanced visual artifacts, especially in the sky regions. Although the CNNs-based methods mitigate the over-estimation artifacts successfully, some of them would produce insufficient dehazing in distant scenes. For example, the dehazed images of MSBDN and FFA appeared almost indistinguishable from the original hazy ones. The reason might lie in the overfitting on the training datasets. As obviously shown in Fig.7 (b), our dehazing model produces plausible visual dehazing results with vivid color and fine details. The clear superiority of our model on the real-world dataset is attributed to the complementary feature enhanced framework. The experimental results also demonstrate that our model learns the color- and texture-wise complementary features effectively.

2) Non-Homogeneous Dehazing: We then evaluate our method on the non-homogeneous dehazing dataset NH-
TABLE I
HOMOGENEOUS IMAGE DEHAZING RESULTS ON THE RESIDE, O-HAZE, HazeRD, AND REAL-WORLD DATASETS. [8]

| Dataset | Metric | DCP | HLP | MSCNN | GDNet | EPDN | MSBDN | FFA | Ours |
|---------|--------|-----|-----|-------|-------|------|-------|-----|------|
| SOTS    | PSNR↑  | 14.57 | 16.44 | 19.15 | 29.78 | 27.42 | 33.79 | 29.92 | 32.17 |
|         | SSIM↑  | 0.749 | 0.784 | 0.855 | 0.978 | 0.946 | 0.985 | 0.975 | 0.970 |
|         | LPIPS↓ | 0.157 | 0.140 | 0.112 | 0.036 | 0.056 | 0.038 | 0.043 | 0.041 |
|         | CIEDE200↓ | 33.01 | 28.65 | 21.65 | 10.86 | 25.55 | 10.36 | 11.24 | 14.32 |
| O-HAZE  | PSNR↑  | 15.09 | 15.28 | 17.22 | 16.67 | 17.17 | 16.76 | 16.09 | 29.87 |
|         | SSIM↑  | 0.449 | 0.466 | 0.521 | 0.480 | 0.600 | 0.460 | 0.465 | 0.758 |
|         | LPIPS↓ | 0.318 | 0.306 | 0.285 | 0.296 | 0.292 | 0.318 | 0.323 | 0.155 |
|         | CIEDE200↓ | 45.72 | 46.81 | 37.11 | 36.39 | 33.79 | 36.67 | 35.15 | 28.80 |
| HazeRD  | PSNR↑  | 15.92 | 15.03 | 15.64 | 15.32 | 15.76 | 15.67 | 16.09 | 17.39 |
|         | SSIM↑  | 0.643 | 0.601 | 0.624 | 0.673 | 0.608 | 0.679 | 0.667 | 0.706 |
|         | LPIPS↓ | 0.228 | 0.284 | 0.235 | 0.231 | 0.243 | 0.232 | 0.232 | 0.224 |
|         | CIEDE200↓ | 31.21 | 27.83 | 26.88 | 26.08 | 30.61 | 26.25 | 26.46 | 25.75 |
| Real World | r_{ave} | 1.993 | 1.418 | 0.822 | 0.704 | 1.664 | 1.429 | 0.600 | 1.767 |
|         | m_{avg} | 1.570 | 1.699 | 1.437 | 1.116 | 1.587 | 1.303 | 1.123 | 2.165 |
|         | r_{sp} | 0.0003 | 0.0195 | 0.0267 | 0.0010 | 0.0001 | 0.0023 | 0.0010 | 0.0000 |

The bold results indicate the best performances, and the second-best are underlined. ↑: The larger the better. ↓: The smaller the better.

Fig. 8. Dehazing results of the Non-Homogeneous hazy images (on the NH-HAZE dataset). (a): the hazy images. (b)-(f): the dehazing results of EDN-AT [29], EDN-3J[29], TWNN [27], DWGAN [28] and Ours, respectively. (g): the ground truth.

HAZE. Our model is compared with several latest non-homogeneous dehazing models, including EDN-AT [29], EDN-3J[29], TWNN [27] and DWGAN [28]. Fig.8 reveals the visual comparisons. As we can see, all of the methods remove the non-homogeneous haze successfully. However, the EDN-3J fails to remove haze from the stone steps; the other compared methods seem to fail to restore accurate color at the bushes and lawns (see the second row in Fig.8). While our model successfully avoids these issues due to the reflectance subnetwork learning accurate color information, achieving more accurate dehazing results on the NH-HAZE dataset than other methods. The quantitative results are shown in Table II also reveal that our model obtains the best results, surpassing the second best methods by a considerable margin, surpassing the second best methods 2.13 dB on PSNR, 0.005 on SSIM, 0.025 on LPIPS, and 0.48 on CIEDE2000, respectively. The reason may lie in the proposed HyLoG-ViT, which has a stronger ability in modelling long-range context information than the vanilla CNN.

3) Nighttime Dehazing: We further evaluate our dehazing model on the nighttime dehazing, comparing with a low-light image enhancement and several latest nighttime dehazing models, i.e. the LIME [66], LCD [67], MRP [30] and OSFD [31]. The visual results tested on the synthetic NHR dataset [31] and real-world dataset are illustrated in Fig.9 and Fig.10, respectively. The low-light image enhancement method LIME only improves the luminance of the nighttime image while failing to remove the color shift and haze. Our method and the other nighttime dehazing models can improve the luminance, correct color shift and remove haze simultaneously. However, all of LCD, MRP and OSFD yield unnatural results at the sky regions in Fig.9 and bright regions around the street lamps in Fig.10 with over-saturation artifacts. While our method produces pleasingly smooth results in such regions, which looks more natural. The qualitative results as shown in Table III also demonstrate that our model can perform surprisingly well on nighttime dehazing, outperforming the state-of-the-art MRP and OSFD methods with respect to PSNR, SSIM, LPIPS and CIEDE2000. Especially, our CIEDE2000 is significantly lower than the MRP with 14.54, indicating that the color-wise complementary features learned from the reflectance subnetwork play a beneficial role in preserving accurate color.
Fig. 9. Dehazing results of the Synthetic Nighttime hazy images (on NHR dataset). (a): the hazy images. (b)-(f): the dehazing results of LIME [66], LCD [67], MRP [30], OSFD [31] and Ours, respectively. (g): the ground truth.

TABLE III
Nighttime dehazing results on the NHR and Real-World datasets.

| Dataset     | Metric | LIME | LCD | MRP | OSFD | Ours |
|-------------|--------|------|-----|-----|------|------|
|             | PSNR ↑ | 13.37 | 13.31 | 15.07 | 19.88 | 22.94 |
|             | SSIM ↑ | 0.767 | 0.608 | 0.694 | 0.700 | 0.814 |
|             | LPIPS ↓ | 0.470 | 0.439 | 0.387 | 0.371 | 0.207 |
|             | CIEDE2000 ↓ | 43.36 | 35.02 | 32.79 | 33.35 | 18.25 |

|             | r_{ave} ↑ | 0.144 | 0.190 | 0.169 | 0.149 | 0.219 |
|             | r_{min} ↑ | 5.049 | 5.589 | 5.406 | 4.906 | 5.846 |
|             | r_{sp} ↓ | 0.0104 | 0.0136 | 0.0190 | 0.0163 | 0.0000 |

C. Discussion

Due to the stronger ability in modelling long-range context information, the network structure with transformer blocks is naturally good at learning the spatially variant features [43]. Our experiments verify this conclusion. The samples in the SOTS are synthetic by setting the atmospheric light and scattering coefficient as global constants. While in the O-HAZE and NH-HAZE datasets, hazy images are captured in natural scenes where the haze is generated by professional haze machines [64]. Under foggy and hazy conditions, the atmospheric light and scattering coefficient will no longer be globally constant in natural scenes. Therefore, as shown in Table I and II, our method can surpass the SOTAs by a considerable margin on the O-HAZE and NH-HAZE, performing relatively better than on the homogeneous haze SOTS dataset.

D. Ablation Study

For quick experiments, all of the models are trained for 10 epochs and evaluated on 300 samples randomly selected from the SOTS [56] dataset in our ablation studies.

1) Evaluations on the Framework: We conduct the ablation studies to demonstrate the contribution of the joint model. 1) w/o-RS: our dehazing network without any complementary features, removing the decoders \( D_R \) and \( D_S \); 2) w/o-S: our dehazing model removing the decoder \( D_S \); 3) w/o-R: our dehazing model removing the decoder \( D_R \). Comparing models of w/o-RS, w/o-S, and w/o-R, the latter two models obtain higher PSNR and SSIM than the former, indicating that the joint learning mechanism with complementary features of reflectance or shading indeed boosts the network’s dehazing performance. Furthermore, when both complementary features are leveraged in our model, it achieves the best performance, demonstrating the effectiveness of the proposed framework.

2) Evaluations on the CFSM: For checking the contribution of the CFSM, we conduct the experiment where the model w/o-CFSM refers to the dehazing model that with-out CFSM yet merge the complementary features \( d_{BR} \), \( d_{BS} \) and features \( d_{DS} \) by summation. The dehazing results are shown in the second row in Table V. As we can find, without the CFSM, the dehazing results reduce 0.671 dB on PSNR and 0.0029 on SSIM, indicating the importance and benefits of using the CFSM in the dehazing network.

3) Evaluations on the HyLoG-ViT: We further evaluate the effectiveness of the proposed HyLoG-ViT model for image dehazing by involving the following different configurations,
where the HyLoG-ViT block is replaced by: 1) two basic ResNet [49] blocks (CNN); 2) the basic ViT block (ViT); 3) the local ViT path (L-ViT); 4) the global ViT path (G-ViT); 5) the sequential stacked local and global ViT paths (LoG-ViT). The model CNN achieves better dehazing performance than the models of ViT and G-ViT. One of the reasons is that Transformers-based models lack some of the inductive biases inherent to model CNN, such as translation equivariance and locality, and therefore do not generalize well when trained on insufficient data. Models of LoG-ViT and HyLoG-ViT obtain the top two PSNR and SSIM, indicating that combining local and global interactions are more effective than other models. Comparing them, we can find that our parallel hybrid scheme is better than the sequentially stacked one.

4) Position Encoding: In this experiment, we will show that, without the position encoding/embedding, our model can also achieve high performance on image dehazing. We also verify that the $3 \times 3$ convolution layer that fuses the hybrid features from local and global ViT paths is useful. Hence, we compared the following models: 1) PE: our HyLoG adds the learnable position encodings as used in [24, 68]; 2) ADD: the hybrid features are fused by element-wise summation. The quantitative results are shown in Table VI. As we can find, there are no significant performance gains to the SSIM when adding the position encoding. The PE model gets even lower PSNR than Our model. However, the ADD model performs worse than the other two models with respect to the PSNR and SSIM, demonstrating that the $3 \times 3$ convolution layer for hybrid features’ fusion is vital in the HyLoG-ViT.

### V. Conclusion

This paper aims to mitigate the issues of CNNs-based dehazing networks by introducing a new complementary feature enhanced framework and a hybrid local and global vision transformer. To these ends, we first propose a new dehazing framework, which jointly learns the intrinsic image decomposition and dehazing. The reflectance and shading prediction tasks encourage the networks to learn more useful complementary features for image dehazing task. We propose a complementary features selection module to effectively aggregate those complementary features to enhance the useful complementary features while weakening the irrelevant ones. Then, we introduce vision transformers into the dehazing task. We design a new hybrid local-global vision transformer (HyLoG-ViT) which is more computationally efficient than the standard ViT, as it can capture both local and global dependencies. We conduct extensive experiments on homogeneous, non-homogeneous, and nighttime dehazing tasks to evaluate our method. Qualitative and quantitative results reveal that our method achieves comparable or even better performance than CNNs-based dehazing methods, demonstrating the effectiveness of the proposed framework and the HyLoG-ViT. We strongly believe that the complementary feature enhanced framework can be deeply explored and achieve outstanding performances on other image restoration tasks, such as image deraining, low-light image enhancement and old photo restoration.

### References

[1] S. Nayar and S. Narasimhan, “Vision in bad weather,” in Proceedings of the IEEE International Conference on Computer Vision, vol. 2, 1999, pp. 820–827.
[2] J. Dong and J. Pan, “Physics-based feature dehazing networks,” in Proceedings of the European Conference on Computer Vision. Springer, 2020, pp. 188–204.
[3] K. He, J. Sun, and X. Tang, “Single image haze removal using dark channel prior,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 12, pp. 2341–2353, 2011.
[4] Q. Zhu, J. Mai, and L. Shao, “A fast single image haze removal algorithm using color attenuation prior,” IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 3522–3533, 2015.
[5] D. Berman, T. Treibitz, and S. Avidan, “Single image dehazing using haze-lines,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 3, pp. 720–734, 2018.
[6] J. Li, H. Zhang, D. Yuan, and M. Sun, “Single image dehazing using the change of detail prior,” Neurocomputing, vol. 156, pp. 1–11, 2015.
[7] M. Kaur, D. Singh, V. Kumar, and K. Sun, “Color image dehazing using gradient channel prior and guided l 0 filter,” Information Sciences, vol. 521, no. C, p. 326–342, 2020.
[8] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, “Dehazenet: An end-to-end system for single image haze removal,” IEEE Transactions on Image Processing, vol. 25, no. 11, pp. 5187–5198, 2016.
[9] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, “Single image dehazing via multi-scale convolutional neural networks,” in Proceedings of the European Conference on Computer Vision, 2016, pp. 154–169.
[10] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, “Aod-net: All-in-one dehazing network,” in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4770–4778.
[11] P. Li, J. Tian, Y. Tang, G. Wang, and C. Wu, “Deep retinex network for single image dehazing,” IEEE Transactions on Image Processing, vol. 30, pp. 1100–1115, 2020.
[12] R. Liu, X. Fan, M. Hou, Z. Jiang, Z. Luo, and L. Zhang, “Learning aggregated transmission propagation networks for haze removal and beyond,” IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 10, pp. 2973–2986, 2019.
[13] H. Li, J. Li, D. Zhao, and L. Xu, “Dehazeflow: Multi-scale conditional flow network for single image dehazing,” in Proceedings of the ACM international conference on Multimedia, 2021, pp. 2577–2585.
M. Fu, H. Liu, Y. Yu, J. Chen, and K. Wang, “Dw-gan: A discrete wavelet transform gan for nonhomogeneous dehazing,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8160–8168.

X. Liu, Y. Ma, Z. Shi, and J. Chen, “Griddehazenet: Attention-based multi-scale network for image dehazing,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7314–7323.

H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M.-H. Yang, “Multi-scale boosted dehazing network with dense feature fusion,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 2157–2167.

X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, “Ffa-net: Feature fusion attention network for single image dehazing,” in Proceedings of the Association for the Advancement of Artificial Intelligence, 2020, pp. 11908–11915.

M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, “Intriguing properties of vision transformers,” arXiv preprint arXiv:2105.10497, 2021.

S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr et al., “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” in Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 6881–6890.

J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “Swinir: Image restoration using swin transformers,” in Proceedings of the IEEE International Conference on Computer Vision Workshop, 2021.

D. Zhou, “Human texture perception,” https://www.cs.auckland.ac.nz/~georgy/research/texture/thesis-html/node6.html, 2006, accessed: 2006-2-22.

B. Julesz, “Visual pattern discrimination,” IRE transactions on Information Theory, vol. 8, no. 2, pp. 84–92, 1962.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 1973–1983.

M. Fu, H. Liu, Y. Yu, J. Chen, and K. Wang, “Dw-gan: A discrete wavelet transform gan for nonhomogeneous dehazing,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021, pp. 203–212.

M. Yu, Y. Cherukuri, T. Guo, and V. Monga, “Ensemble dehazing networks for non-homogeneous haze,” in Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop, 2020.

J. Zhang, Y. Cao, S. Fang, Y. Kang, and C. Wen Chen, “Fast haze removal for nighttime image using maximum reflectance prior,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7418–7426.
mans, D. Dai, and L. Van Gool, “Multi-task learning for dense prediction tasks: A survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

K. He, X. Zhang, S. Ren, and S. Jian, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, “Cbam: Convolutional block attention module,” in Proceedings of the European Conference on Computer Vision, 2018, pp. 3–19.

C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, “Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images,” in Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop, 2020, pp. 444–445.

D. P. Kingma and P. Dhariwal, “Glow: Generative flow with invertible 1x1 convolutions,” arXiv preprint arXiv:1803.02872, 2021.

J. Cai, S. Gu, and L. Zhang, “Learning a deep single image contrast enhancer from multi-exposure images,” IEEE Transactions on Image Processing, vol. 27, no. 4, pp. 2049–2062, 2018.

J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” arXiv preprint arXiv:1511.06434, 2015.

D. P. Kingma and P. Dhariwal, “Glow: Generative flow with invertible 1x1 convolutions,” arXiv preprint arXiv:1807.03097, 2018.

D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proceedings of the International Conference on Learning Representations, 2015.

Y. Zhang, L. Ding, and G. Sharma, “Hazed: an outdoor scene dataset and benchmark for single image dehazing,” in Proceedings of the IEEE International Conference on Image Processing, 2017, pp. 3205–3209.

Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli et al., “Image quality assessment: from error visibility to structural similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, 2004.

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

ACKNOWLEDGMENT

This work is supported by grants from National Natural Science Foundation of China (No.61922006, No.62132002) and CAAI-Huawei MindSpore Open Fund.

Hongyu Li is currently pursuing the master degree with the State Key Laboratory of Virtual Reality Technology and System, School of Computer Science and Engineering, Beihang University, Beijing, China, in 2020. His research interests include computer vision and image processing.

Long Xu (M’12-SM’15) received the M.S. degree from Tsinghua University in 2005 and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2011. He is currently a Full Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. Before he joined Beihang University in Jun. 2014, he used to conduct research in Nanyang Technological University, Peking University and Shanda Innovations. He is the author or coauthor of over 70 technical articles in refereed journals and conferences such as TPAMI, IJCV, TIP, CVPR and ICCV. His research interests include computer vision and multimedia big data, especially the understanding and generation of visual contents. He is supported by the Research Funds for Excellent Young Researchers from National Nature Science Foundation of China since 2019. He was also selected into the Beijing Nova Program (2017) and ever received the Second-grade Science Award of Chinese Institute of Electronics (2018), two Excellent Doctoral Thesis Award from Chinese Academy of Sciences (2012) and the Beijing Municipal Education Commission (2012), and the First-Grade Science-Technology Progress Award from Ministry of Education, China (2010). He is a senior member of IEEE, CIE and CCF. More information can be found at http://cvteam.net.

Jia Li (M’12) received his M.S. degree in applied mathematics from Xidian University, Xi’an, China, in 2002, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He was a Postdoc with the Department of Computer Science, City University of Hong Kong, the Department of Electronic Engineering, Beihang University, Beijing, China. He used to conduct research in Nanyang Technological University, Singapore. Currently, he is with the Key Laboratory of Solar Activity, National Astronomical Observatories, Chinese Academy of Sciences. His research interests include image/video processing, solar radio astronomy, wavelet, machine learning, and computer vision. He was selected into the 100-Talents Plan, Chinese Academy of Sciences, 2014.