Pattern Recognition 158 (2025) 111051

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Image dehazing via self-supervised depth guidance

Yudong Liang [a,b], Shaoji Li [b], De Cheng [c,*], Wenjian Wang [a,b], Deyu Li [a,b], Jiye Liang [a,b]

[a] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, PR China
[b] The School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, Shanxi, PR China
[c] The State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an, 710071, Shaanxi, PR China

ARTICLE INFO

Keywords:
Image dehazing
Self-supervised
Depth guidance
Transformer
Hybrid attention

ABSTRACT

Self-supervised learning methods have demonstrated promising benefits to feature representation learning for image dehazing tasks, especially for avoiding the laborious work of collecting hazy-clean image pairs, while also enabling better generalization abilities of the model. Despite the long-standing interests in depth estimation for image dehazing tasks, few works have fully explored the interactions between depth and dehazing tasks in an unsupervised manner. In this paper, a self-supervised image dehazing framework under the guidance of self-supervised depth estimation has been proposed, to fully exploit the interactions between depth and hazes for image dehazing. Specifically, the hazy image and the corresponding depth estimation are generated and optimized from the clear image in a dual-network self-supervised manner. The correlations between depth and hazy images are exploited in depth-guided hybrid attention Transformer blocks, which adaptively leverage both the cross-attention and self-attention to effectively model hazy densities via cross-modality fusion and capture global context information for better feature representations. In addition, the depth estimations of hazy images are further explored for the detection tasks on hazy images. Extensive experiments demonstrate that the depth estimation not only enhances the model generalization ability across different dehazing datasets, leading to state-of-the-art self-supervised dehazing performance, but also benefits downstream detection tasks on hazy images. Our code is available at https://github.com/DongLiangSXU/Depth-Guidance-dehazing.git.

1. Introduction

Image dehazing [1–5] plays a crucial role in enhancing visual quality and facilitating high-level vision tasks in hazy weather conditions. Existing image dehazing methods [6–8] have enjoyed the merits of supervised learning [9], which relies on the laborious work of collecting large sets of hazy-clean image pairs. Besides, such supervised methods can easily overfit to the training data, which sacrifices the model generalization ability. As a result, these methods cannot be well applied to the real-world scenarios, and the model performance will degrade heavily when applying to other datasets.

Therefore, to improve the generalization ability of the learned model remains a highly challenging issue to be explored. Recently, self-supervised learning methods have exhibited impressive performances in feature representation learning, occasionally even outperforming their supervised learning counterparts in a couple of high-level tasks. Benefiting from better feature representations, self-supervised techniques have also been applied in low-level vision tasks including image dehazing. Li et al. [1] propose to predict the transmission map, atmospheric light and dehazed images without ground-truth, from which the hazy inputs are reconstructed in a self-supervised manner. Although this layer disentanglement mechanism achieves much better performances compared with other unsupervised learning-based image dehazing methods, some assumptions or priors to construct the loss functions may be violated in real application. Liang et al. [2] generate hazy images from collected clear images with depth estimations. Self-supervised learning was then employed to restore the clear images and further adapt restorations from the real hazy images. It does not depend on the expensive collection of hazy-clean image pairs and profits from extra clear images to improve the performance. It demonstrates strong generalization abilities when applied to different test data. The rapid progress in self-supervised learning gives promising directions for further developments in image dehazing areas.

On the other hand, the interactions between the image depth and dehazing tasks have attracted increasing attention. In hazy weather conditions, visibility decreases as distance increases. As Fig. 1, the depth information exhibits a strong correlation with the hazy densities, which can be explained in the atmospheric scattering model. Recently, Guo et al. [10] apply Dark Channel Prior [11] to model the depth information and exploit 3D position embedding for the proposed

* Corresponding author.
E-mail address: dcheng@xidian.edu.cn (D. Cheng).

https://doi.org/10.1016/j.patcog.2024.111051
Received 26 January 2024; Received in revised form 2 July 2024; Accepted 23 September 2024
Available online 30 September 2024

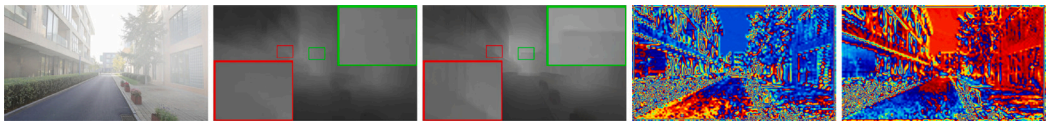

Fig. 1. Visual representations of a hazy image, the corresponding depth map, our preliminary fusion results, one typical channel of attention maps by self-attention mechanism without depth guidance, and by our hybrid attention mechanism. Note that our preliminary fusion brings more details compared with depth map and our hybrid attention map is highly correlated to the semantic structures and hazy densities. Redder colors in attention maps mean larger attention values.

dehazing Transformer. Yang et al. [12] employ the scattering coefficients and depth estimations to formulate dehazing and rehazing cycles with unpaired data to improve the generalization ability of the model. The depth information proves to be important heuristic information for the image dehazing task. How to explore depth guidance for image dehazing is an interesting and promising issue for this community.

In this paper, we exploit depth estimations for hazy images to guide image dehazing tasks in a self-supervised manner. Following [2], we generate hazy images from clear images with the aid of depth estimation from the pre-trained depth estimation models. In addition, self-supervision is adopted for the depth predictions from the generated hazy images. We design depth-guided hybrid attention Transformer blocks to adaptively leverage both the cross-attention and self-attention. It enjoys the merits of cross-attention which enables cross-modality fusion of depth and hazy images to effectively model hazy densities, and takes advantages of self-attention which could model long-range correlations and capture global context information for better feature representations. As Fig. 1, the hybrid attention map has successfully captured the variations of hazy density and semantic information which are crucial for image dehazing.

The depth provides additional positional information in the 3D world. Furthermore, the depth estimations for hazy images are explored for the detection tasks on hazy images. The state-of-the-art dehazing performances are obtained, and extensive experimental results demonstrate that the depth estimations not only improve the generalization ability of the model across different dehazing datasets, but also benefit the downstream detection tasks on the hazy images. In summary, we make the following main contributions,

- A self-supervised image dehazing framework with self-supervised depth guidance has been built, which sequentially generates hazy inputs, estimates the depth for hazy images with the aid of the depth estimations from clear images, and exploits the interactions between depth and hazes for image dehazing.
- We design depth-guided hybrid attention Transformer blocks to exploit the correlations between the image depth and hazy densities in the images, which adaptively leverage both the cross-attention and self-attention to effectively model hazy densities via cross-modality fusion and capture global context information for better feature representations.
- The state-of-the-art dehazing performances are obtained compared with the unsupervised or self-supervised dehazing methods. The self-supervised depth estimations not only improve the model generalization ability across different dehazing datasets, but also benefit the downstream detection tasks on hazy images.

The remainder of this paper is organized as follows. Section 2 briefly reviews the existing related literature of Transformer-based dehazing methods, self-supervised learning for image dehazing and object detection methods on hazy images. Section 3 describes our approach in detail, covering the hazy image generation, the architectures for the depth-guided image dehazing, and the depth guidance for detection on hazy images. Section 4 reports the implementation details, the performance comparisons of dehazing and detection tasks for hazy images, as well as the ablation study. Finally, in Section 5, we conclude this paper.

2. Related work

Image dehazing tasks have been widely concerned for decades. The existing image dehazing methods can be broadly categorized into traditional handcrafted image dehazing methods [11] and learning-based image dehazing methods [2,13]. Among the learning-based image dehazing methods, transform-based and self-supervised learning-based dehazing methods are reviewed to better understand our approach.

2.1. Transformer-based dehazing methods

Recently, Vision Transformers have been successfully applied in lots of computer vision tasks [14–16] which enjoy the merits of modeling long-range correlations and paralleling the computations for attention. For dehazing tasks, vision Transformers are also explored with different architectural designs [10,16]. Valanarasu et al. [16] develop learnable weather query embedding for decoders of Transformers to enable enhancement for multiple kinds of distortions. Guo et al. [10] attempt to combine CNN and Transformers for dehazing which modulates the CNN features via global context information. Cross-attention mechanisms have demonstrated great potential for cross-modality fusion problems [15,17]. Wei et al. [17] design the Multi-Modality Cross Attention Network for matching image and sentence, which exploits the inter-modality relationship between the sentence words and image regions. Wang et al. [15] apply cross-attentions to fulfill the multi-modal token fusion method with Transformers and obtains the state-of-the-art performances for several fusion tasks for different modalities. To better fuse the information from depth and hazy images, cross-attention mechanisms implemented by Transformers are explored in this paper.

The performances of supervised Transformer-based dehazing methods are limited by the size of dataset as the hazy-clean image pairs are inherent unavailable or too expensive to collect. The performances deteriorate when the supervised models are applied to other datasets. Improving the generalization ability of models is a crucial issue for dehazing methods.

2.2. Self-supervised learning for image dehazing

To improve the generalization ability of model and the feature representation abilities, unsupervised or self-supervised learning is developed to train models without extra labels. Depth-Aware Unpaired Video Dehazing [18] employs depth to simulate ego-motion and models haze variations with unpair data. Visual-quality-driven unsupervised image dehazing [19] develops interactive fusion modules and iterative optimization modules to refine dehazed results. Generative Adversarial and Self-Supervised Dehazing Network [20] restores hazy image in a self-supervised learning and adversarial learning manner.

Self-supervised learning could enable large-scale training from the input itself by pretext learning or contrastive learning, obtaining more discriminative feature representations which could further benefit the downstream tasks. There are some self-supervised learning based works [2,21,22] which similarly integrate pre-trained deep models to extract features or obtain priors without data annotations. Liang et al. [2] perform self-supervised learning and adaptation to restore clear images

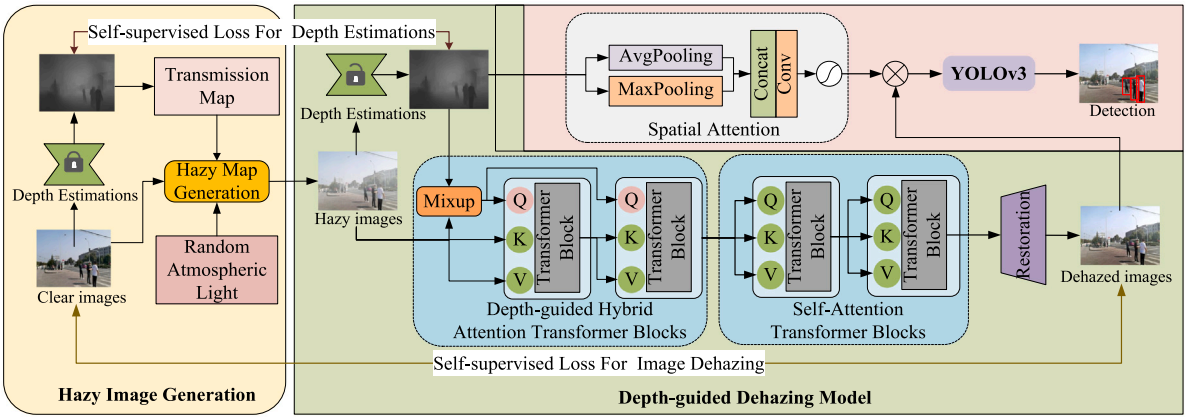

Fig. 2. The framework of the proposed depth-guided self-supervised dehazing method. It consists of the self-supervised hazy image generation module, the self-supervised depth estimation block, the depth-guided hybrid attention Transformer blocks, the self-attention Transformer blocks, and the image restoration module.

with aid of the pre-trained depth estimation model. In this paper, we further develop depth estimation model for hazy images via self-supervised learning and the estimated depth information would benefit the following dehazing and the detection tasks for hazy images.

Although these pre-trained deep models are obtained with the aid of human information, this human information does not overlap with the current tasks. Utilizing pre-trained deep models in our approach serves to extract superior features than hand-designed features or introduces effective priors, thereby circumventing the typically arduous and labor-intensive manual annotations or collection for the current task.

3. Our approach

As shown in Fig. 2, we propose a self-supervised image dehazing framework to sequentially generate hazy inputs, estimate the depth for hazy images in a self-supervised dual-network manner from clear images, and exploit the interactions between depth and hazes for image dehazing. To demonstrate the benefits of depth estimations, depth estimations and dehazed results are further explored for the detection task on hazy images in a manner of spatial attention. Our approach integrates hazy image generation and depth-guided image dehazing process into an end-to-end model. Only depth-guided image dehazing process is applied in the inference phase. The details are depicted in the following section.

3.1. Hazy image generation

We generate hazy images from clear images as [2], which avoids expensive or unprocurable collections of hazy-clean image pairs. According to the atmospheric scattering model, hazy images can be generated with global atmospheric light A and the transmission map $t(x)$ as illustrated in Eq. (1), which is related to the depth information.

$$I(x) = J(x)t(x) + A(1 - t(x)), \tag{1}$$

$$t(x) = e^{-\beta d(x)}. \tag{2}$$

The transmission map $t(x)$ attenuates exponentially with the scene depth $d(x)$, where β is the scattering coefficient of the atmosphere and x indicates the position in the scene. This formulation aligns with the observed phenomenon where hazy density increases with distance from the observer.

Benefiting from the significant process in the area of single image depth estimation, a pre-trained single depth estimation model [23] or [24] has been explored to provide depth information from clear images. Global atmospheric light A and scattering coefficient β are randomly produced for hazy image generations using the depth estimation from clear images. The pre-trained depth estimation model for hazy image generation is frozen during the training. The hazy image generation is quite efficient and is abandoned during inference.

3.2. The architectures for the depth-guided image dehazing

The architecture for image dehazing consists of the self-supervised depth estimation block, the depth-guided hybrid attention Transformer blocks, the self-attention Transformer blocks, and the restoration block. The correlations between depth and hazy images are exploited in the proposed depth-guided hybrid attention Transformer blocks which adaptively integrate cross-attentions and self-attentions to effectively model hazy densities and capture global context information for better feature representations.

For the generated hazy images, self-supervised depth estimations are performed firstly using the same deep architectures as the depth estimation model for clear images such as [23] or [24]. There is a lock in Fig. 2 to indicate that the pre-trained model is fixed. With the supervision of depth estimations from clear images as illustrated in Eq. (3), the estimations can be successfully adapted to hazy images during the learning of dehazing process. L_1 loss is applied in the Eq. (3) as follows,

$$L_{depth}(\hat{d}(x), \tilde{d}(x)) = |\hat{d}(x) - \tilde{d}(x)|, \tag{3}$$

where $\tilde{d}(x)$ and $\hat{d}(x)$ are the depth estimations of hazy-free and hazy image pairs by depth estimation models pre-trained and trained by self-supervised learning respectively.

Afterwards, as shown in Figs. 2 and 3, the estimated depth is then fed into the depth-guided hybrid attention Transformer blocks to explore depth guidance for dehazing process. The depth-guided hybrid attention Transformer blocks firstly fuse the depth and hazy images with an adaptive mixup operation [25] as Eq. (4),

$$Mix(\hat{d}(x), I(x)) = \sigma(\theta) * \hat{d}(x) + (1 - \sigma(\theta)) * I(x), \tag{4}$$

where the learnable parameter θ controls the preliminary fusions of depth estimations $\hat{d}(x)$ of hazy images and the hazy image $I(x)$.

Then the fused depth-hazy modality is fed into the hybrid attention Transformer blocks to acquire the correlations between depth and hazy image. In specific, the attentions are calculated as:

$$Attn(Q, K, V) = softmax(\frac{QK^T}{\sqrt{n}})V, \tag{5}$$

where Q, K, and V are the query, keys, and values as in the original self-attention calculations, and n is the dimensionality of heads. In this hybrid attention calculation, Q is the head for the preliminarily fused depth-hazy modality, K and V are the heads for hazy images. As patch embedding applies convolution and normalization, the patch embedding for the preliminary fusion results is simplified to be separable. Integrating Eq. (4) into Eq. (5), it arrives,

$$\begin{aligned} Attn(Q, K, V) \approx & \sigma(\theta) * Attn(Q_{\hat{d}}, K, V) \\ & + (1 - \sigma(\theta)) * Attn(Q_I, K, V), \end{aligned} \tag{6}$$

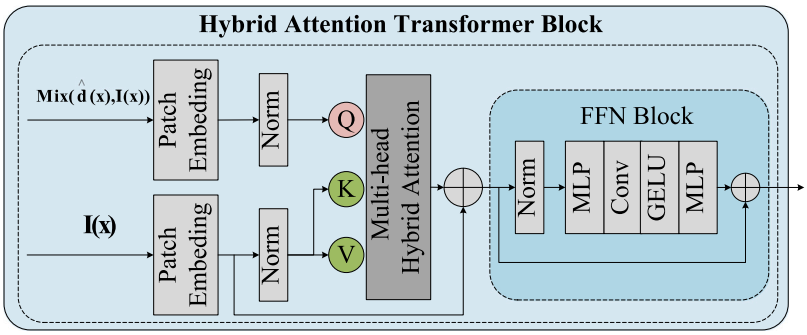

Fig. 3. The architecture of the applied Hybrid Attention Transformer block.

where $Q_{\hat{d}}$ and Q_I could be considered as the corresponding heads for depth estimation $\hat{d}(x)$ of hazy images and the hazy image $I(x)$. Then the former term of Eq. (6) denotes the cross-attention calculations for cross-modality fusions that exploits the correlations between depth estimations and hazy input, and the latter term indicates the self-attention calculations which model the long-range correlations and capture the global context for the dehazing task. As θ is a learnable parameter, the integration of cross-attention and self-attention calculations is adaptive, which could easily enjoy the merits of both cross-attention and self-attention. In fact, the attention calculates the similarities between query and key. The preliminary adaptive mixup fusion has reduced the discrepancy between depth estimations and hazy images which makes capturing the correlations between depth and hazy images easier for the following attention calculations. As shown in Fig. 1, the preliminary fusion results of depth and hazy images have provided more texture details than the depth alone. The hybrid attention map has successfully described the variations of hazy density and semantic information which are essential for good dehazing performances. In Fig. 1, the attention values are larger in the area of sky than in the road area, as sky is more distant and with a larger hazy densities. On the other hand, some structures like buildings are brighter which indicates the hybrid attentions have successfully captured the semantic information.

The preliminary fusions of depth and hazy images by adaptive mixup operations play an important role in our proposed depth-guided hybrid attention Transformer blocks. Without the preliminary fusions of depth and hazy images, the hybrid attention mechanism degrades to the conventional cross-attention mechanism. As demonstrated in Table 2 of the ablation study part in the experiments, although depth-guidance via the conventional cross-attention mechanisms still benefits the dehazing results, performances degrade compared with our proposed depth-guided hybrid attention mechanisms. Adaptive integration of cross-attention and self-attention has facilitated exploiting the depth-guidance for dehazing tasks. Attention mechanisms are important for the cross-modality fusions. Only applying preliminary fusions without the following Transformer blocks but with a plain stacking convolutional neural network fails to converge the training. As shown in Fig. 3, we apply a conventional calculation of the transformers as:

$$y_i = MA(Fea_i^Q, Fea_i^K, Fea_i^V) + Fea_i^V, \tag{7}$$

where Fea_i^Q, Fea_i^K, Fea_i^V are the input feature maps of query, key and value the i_{th} Transformer blocks for the multi-head attention operation MA, of which the attention is calculated as Eq. (5). y_i is the output features for the multi-head attention calculation, which is passed to the feed-forward neural network for the calculation F_{FFN} similar to [16] as

$$F_{FFN}(y_i) = F_{M1}(GELU(conv_d(F_{M2}(Norm(y_i))))) + y_i. \tag{8}$$

F_{M1} and F_{M2} are different mapping realized by Multi-Layer Perceptron network, $conv_d$ is the depth-wise convolution, $GELU$ is Gaussian error linear units, $Norm$ is the Layer Normalization for Add&Norm operations.

In the following self-attention calculations, Q, K and V are the heads of calculated feature embedding. The self-attention Transformer blocks have the same architectures as the previous hybrid attentive Transformer blocks. The self-attention calculations could further model the long-range correlations and capture the global context for the dehazing task, which could also benefit modeling the uneven density of the hazy images.

Finally, several up-sample convolution layers and nonlinear activations are applied to restore the dehazed results. The whole architecture for the depth-guided image dehazing performs an end-to-end learning with aid of the hazy image generation part. L_1 loss, perceptual loss L_p and contrastive loss L_{CR} are utilized during training as Eq. (9),

$$L_{dehaze} = \lambda_r L_1(\hat{J}, J) + \lambda_p L_p(\hat{J}, J) + \lambda_{cr} L_{CR}(\hat{J}, J, I) \tag{9}$$

where $\hat{J}$, J, I and λ are dehazed restorations, clear images, generated hazy images and multiplier parameters respectively. In the implementations, λ_r, λ_p and λ_{cr} take the values 1, 0.04 and 1 respectively. The hyperparameters in Eq. (9) are chosen empirically. Specifically, the contrastive loss L_{CR} aims to make the restorations away from the generated hazy images and towards the clear images as Eq. (10),

$$L_{CR}(\hat{J}, J, I) = \sum_k \omega_k \cdot \frac{D(f_k(J), f_k(\hat{J}))}{D(f_k(I), f_k(\hat{J}))}. \tag{10}$$

Following [2], we apply the contrastive loss to the intermediate features from the fixed pre-trained model *e.g.*, VGG features, where f_k denotes the kth layer feature obtained from the pre-trained model, D denotes the metric to quantify the differences, and ω_k corresponding to the weight.

3.3. The depth guidance for detection on hazy images

Image dehazing methods not only aims to enhance visual quality but also play a crucial role in improving the performance of high-level tasks under hazy conditions, such as object detection, which is one of the most common high-level task in real applications. The depth provides positional information in 3D world which may benefit the following detection task on hazy images. As shown in Fig. 2, a simple spatial attention is applied for the fusions of the depth and dehazed images by our method for detection. Dehazed images are multiplied by the spatial attention weights calculated from depth estimations as Eq. (11), which are further handled by the YOLOv3 models.

$$F_s(\hat{d}, \hat{J}) = \sigma(conv(Concat(Maxpool(\hat{d}), Avgpool(\hat{d})))) \cdot \hat{J} \tag{11}$$

In Eq. (11), the depth estimations after average-pooling $Avgpool(\hat{d})$ and max-pooling $Maxpool(\hat{d})$ are concatenated and fused with a convolution operation, then sigmoid activation σ is applied to reweight the dehazed restoration $\hat{J}$ to generate depth-guided feature representations $F_s(\hat{d}, \hat{J})$. Then, this depth-guided feature representations are fed into a very basic detection model YOLOv3 [26]. Benefiting from exploiting the inner correlations of image dehazing and depth, the depth estimations from hazy images could not only improve the dehazing process

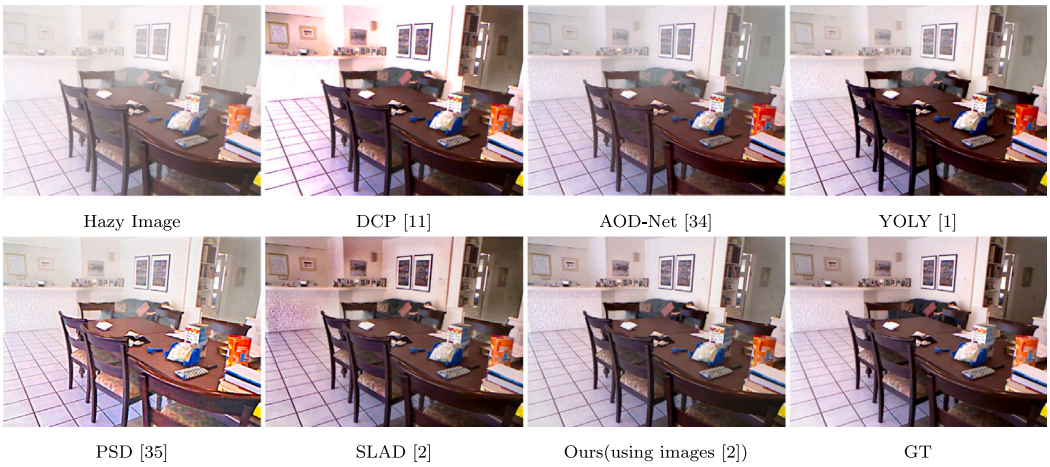

Fig. 4. Visual comparisons on the SOTS-indoor dataset.

but also facilitate the downstream detection task on hazy images. This highlights the practical benefits of our image dehazing method beyond visual enhancement, directly impacting the performance of high-level computer vision applications like object detection. Note that in the implementation, the pre-trained YOLOv3 detection model by MS COCO dataset [27] is imported.

4. Experiments

Our proposed network is implemented by PyTorch 1.7.0, and experiments can be conducted on two NVIDIA TITAN XP GPUs of $12GB$ of memory. The models are trained using an Adam optimizer with exponential decay rates β_1 and β_2 of 0.9 and 0.999, respectively. The initial learning rates for training the depth estimation model and dehazing model are 0.00001 and 0.0005 respectively. Batch size are set to 8. The cosine annealing strategy is applied to adjust the learning rate. We train the models on the training dataset used by [2] and the training sets of RESIDE-6K [28] separately to investigate the influences of the training data for our method. The same 2000 pictures with diverse content utilized by [2] are adopted as the clear images for training. RESIDE-6K is the training set used in [28], which contains 3000 indoor image pairs from ITS and 3000 outdoor image pairs from OTS. ITS and OTS are subsets of RESIDE dataset [29]. We only use the clear images from the training data of RESIDE-6K for the training. The same image pre-processing operations as [2] are applied, which randomly crop the input to 256×256 and normalize the value to $0 \sim 1$. We train the model using clear images from [2] for 200 epochs denoted as Ours(using images [2]), while we train the model using the clear images from RESIDE-6K for 300 epochs denoted as Ours(RESIDE-6K). The applied losses are the weighted combinations of $L1$ loss, perceptual loss and contrastive loss, which are commonly used in various image enhancement tasks.

4.1. Depth-guided image dehazing

To demonstrate the generalization abilities of our model to different datasets, five datasets, *i.e.*, Synthetic Objective Testing Set (SOTS) [29], Hybrid Subjective Testing Set (HSTS) [29], 4KID [30], Haze4K [31], I-HAZE [32], and O-HAZE [33] are investigated.

Following [1], the SOTS and parts of HSTS are evaluated, which are the subsets of the RESIDE v0 version. The SOTS consists of the "SOTS-indoor" subset and the "SOTS-outdoor" subset, having 500 indoor hazy images and 500 outdoor hazy images for testing respectively. 10 outdoor hazy-clean image pairs of HSTS are also tested. Haze4K is a large-scale dehazing dataset that contains 4000 hazy-clean image pairs [31]. There are some very similar images between Haze4K and RESIDE datasets.

To further evaluate the generalization ability of some supervised dehazing methods, we choose I-HAZE, O-HAZE, 4KID and real hazy images from RTTS of RESIDE dataset and Internet for comparisons. Since I-HAZE is not used for training, we use all the data in it for testing. 4KID is established by [30], which focuses on large-size 4K (*i.e.*, 3840×2160) synthetic hazy images. According to [30], 200 hazy images in the Haze4K dataset are randomly selected and tested. To reduce the heavy computational burden in attention mechanisms, we first resize the original image to quarter size (*i.e.*, 1024×576) by bilinear interpolation, then dehaze the image, and finally bilinearly upsample the restorations to the original size. It is interesting the performances of our model and the compared models could be improved by this downsample and upsample operations, partly due to the distortions that have been suppressed during the downsample process. All the compared methods have applied the same operations for fair comparisons.

As our method is the self-supervised method, the following methods are compared quantitatively: traditional prior-based algorithms, DCP [11], NLD [34], the SOTA zero-shot unsupervised dehazing network YOLY [1] that requires iterative optimizations for input images, the SOTA unsupervised dehazing network PSD [35] that performs end-to-end inferences for hazy inputs, the self-supervised dehazing method SSDN [36] and ZID [37]. In addition, recent unsupervised or self-supervised methods are compared, including the SOTA self-supervised dehazing method SLAD [2], recent dehazing method SLP [38] that proposes Saturation Line Prior, the SOTA unpair dehazing method CDD-GAN [39], Depth-Aware Unpaired Video Dehazing (DUVD) [18], Visual-quality-driven unsupervised image dehazing (VQD) [19] and SZDNet [40], a recent self-supervised learning based image dehazing method. Earlier fully supervised deep network AOD [41] is also provided as a reference.

As the code and models of some methods above are not released, some scores are unavailable in their public paper and absent in Table 1. Our method consistently demonstrates either superior or comparable performance to these state-of-the-art methods, validating the efficacy of our proposed self-supervised image dehazing framework with self-supervised depth guidance.

Estimating depth information in a self-supervised learning mode from hazy images is barely explored in the existing works, which further provides depth guidance for the self-supervised image dehazing process. The experiments demonstrate the great benefits of our self-supervised depth guidance for image dehazing process.

Our framework is adaptable and not restricted to a particular pre-trained depth model. We validate this by incorporating two different pre-trained depth estimation models, *i.e.*, densely connected depth estimation model [23] and light-weight pre-trained depth estimation

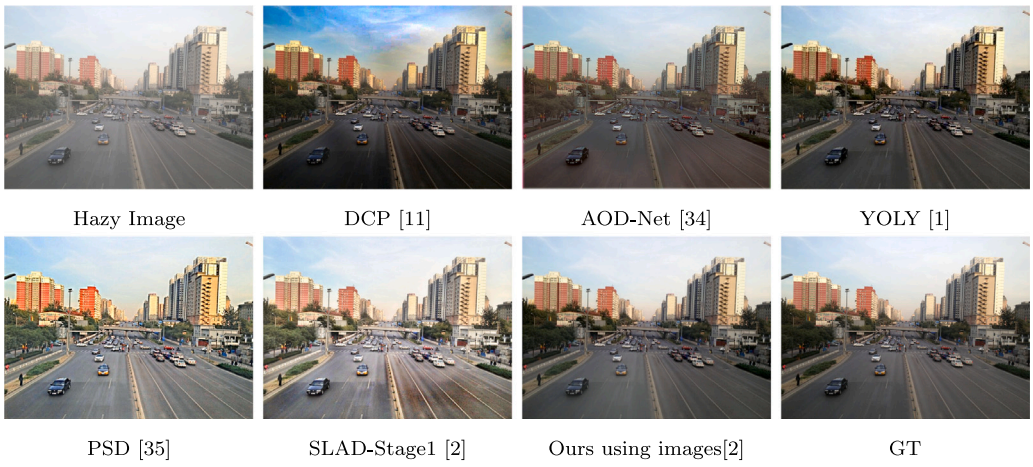

Fig. 5. Visual comparisons on the SOTS-outdoor dataset.

Table 1
Performance Comparisons of different methods on the SOTS-indoor, SOTS-outdoor, HSTS, Haze4K, I-HAZE, O-HAZE, 4KID dataset.

	SOTS-indoor		SOTS-outdoor		HSTS		Haze4k		I-HAZE		O-HAZE		4KID		Average value	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DCP [11]	16.62	0.8197	17.16	0.8514	17.01	0.8030	14.01	0.7600	14.43	0.7520	16.78	0.6500	17.50	**0.8756**	16.21	0.7873
AOD [41]	19.06	0.8504	19.58	0.8711	19.68	0.8350	17.15	0.8300	15.34	0.7909	16.68	0.7578	14.02	0.8270	17.35	0.8231
NLD [34]	17.29	0.7489	13.06	0.7449	18.92	0.7411	15.27	0.6700	14.12	0.6540	15.58	0.5850	16.47	0.8018	15.81	0.7065
YOLY [1]	19.41	0.8327	22.41	0.9023	21.02	0.9050	15.75	0.5913	15.75	0.7853	16.35	0.7298	13.97	0.7786	17.80	0.7892
PSD [35]	12.45	0.6812	15.42	0.7624	19.37	0.8240	14.77	0.7114	13.78	0.7904	12.01	0.7799	12.78	0.7614	14.36	0.7586
SLAD [2]	20.49	0.8578	24.33	0.9323	25.05	0.9280	21.63	0.9164	16.40	0.8659	14.87	0.8284	18.48	0.8404	20.17	**0.8813**
SLP [38]	20.16	0.8579	19.68	0.8864	19.27	0.8308	18.95	0.8713	13.88	0.7703	16.53	0.8192	17.29	0.8658	17.98	0.8431
CDD-GAN [39]	**24.61**	0.9180	–	–	22.16	0.9110	–	–	–	–	–	–	–	–	–	–
VQD [19]	–	–	22.53	0.8750	22.25	0.8470	–	–	–	–	16.75	0.6770	–	–	–	–
SSDN [36]	19.56	0.8330	19.51	0.8270	19.84	0.8510	–	–	–	–	–	–	–	–	–	–
SZDNet [40]	–	–	–	–	–	–	–	–	16.52	0.7186	15.83	0.6459	–	–	–	–
ZID [37]	19.32	0.8225	20.27	0.8777	22.65	0.9011	18.96	0.8192	16.08	0.7787	17.33	0.7753	18.46	0.8601	19.01	0.8335
DUVD [18]	15.84	0.7829	21.71	0.8891	19.20	0.8516	20.14	0.8718	17.60	0.7657	19.72	0.7401	12.98	0.7874	18.17	0.8126
Ours(DenseDepth [23]) (using images [2])	22.91	0.8810	24.94	0.9356	**25.30**	0.9368	**22.90**	**0.8975**	**16.90**	**0.8365**	**17.39**	**0.8264**	18.91	0.8853	**21.32**	0.8855
Ours(LiteMonoDepth [24]) (using images [2])	20.62	0.8594	24.12	0.9324	24.36	0.9264	22.31	0.8958	16.19	0.7407	16.62	0.7794	17.80	0.8568	21.52	0.8794
Ours(Densedepth [23]) (RESIDE-6K)	25.39	**0.8993**	26.09	0.9431	24.83	0.9253	23.16	0.8856	16.54	0.8248	16.88	0.8209	**18.51**	0.8691	20.39	0.8576
Ours(LiteMonoDepth [24]) (RESIDE-6K)	22.00	0.8616	**25.81**	**0.9383**	25.62	**0.9301**	23.18	0.8968	16.24	0.7329	17.73	0.7781	18.16	0.8572	21.25	0.8564

model [24]. This results in two variations of models, denoted as Ours(DenseDepth [23]) and Ours(LiteMonoDepth [24]) respectively, which obtains comparable performances as Table 1. We anticipate further improvements in performance with the adoption of larger and more sophisticated depth estimation models. Note that our framework is also adaptable to different training images. Specifically, we denote our model trained using clear images from [2] as "Ours (using images [2])", and the model trained using clear images from RESIDE-6K as "Ours (RESIDE-6K)". Besides Table 1, all the results are obtained with the densely connected pre-trained depth estimation model [23].

From Table 1, Figs. 4, 5 and 7, our method compares favorably against the state-of-the-art unsupervised or self-supervised methods, even surpassing an early proposed fully supervised method AOD [41] on SOTS-indoor and SOTS-outdoor datasets. In Fig. 4, some unsupervised methods such as YOLY [1] and PSD [35] leave the haze in the left corner remained, and the self-supervised method SLAD [2] produces color shift distortions in the left corner. Our method successfully dehazes the images without obvious distortions. For the outdoor images in Fig. 5, our method restores the color tunes more correctly compared with other methods. There are some distortions in the road by SLAD [2].

The results of PSD [35] seem unreal especially for the building parts. There are color shifts in the sky area in the dehazing results of YOLY [1].

For the 4KID dataset, state-of-the-art supervised methods are additionally evaluated for the generalization ability of the model in Table Table 1. On 4KID, our method achieves the best quantitative scores and significantly outperforms the state-of-the-art self-supervised image dehazing method SLAD [2]. Compared with other methods in Fig. 6, our method restores more natural results with fewer artifacts or color shifts, especially in the distant building and sky areas. Some fully supervised methods [41–43] struggle to effectively dehaze images or may generate severe artifacts when applied to test data sets that do not perfectly align with their training datasets.

For the real hazy images in Fig. 7, although results of all the dehazing methods are far from perfect, our results generate fewer artifacts and color shifts, especially for the distant area like the sky, benefitting from the depth guidance exploited by our model.

4.2. Ablation study

In this section, we perform ablation studies on SOTS test set to investigate the benefits of depth guidance, depth-guidance via the hybrid

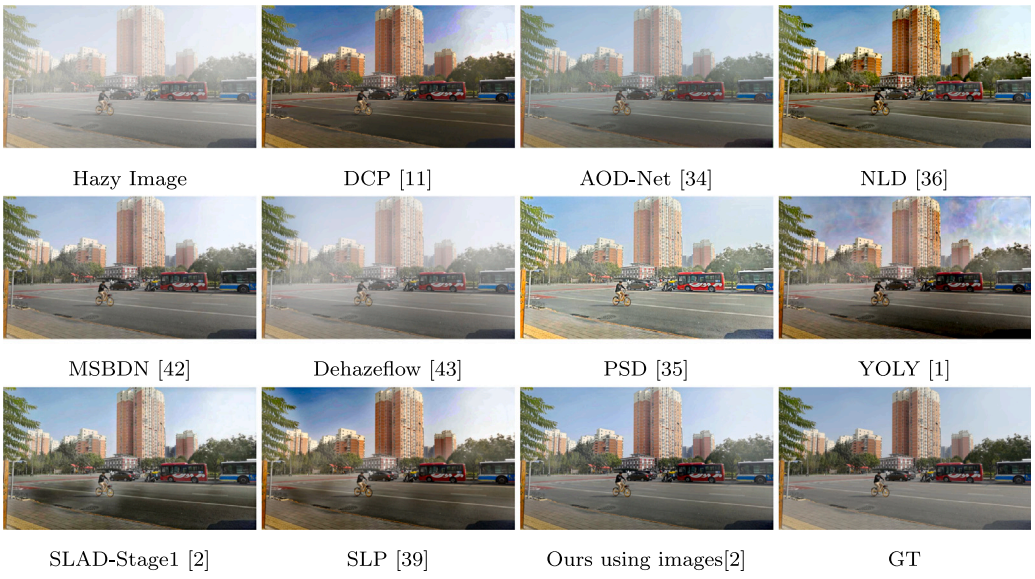

Fig. 6. Visual comparisons on the 4KID dataset.

Table 2
PSNR Comparing the different manners of providing depth-guidance.

	w/o Depth	Cross attention	Hybrid attention	CNN fusion
SOTS-indoor	18.96/0.8458	21.42/0.8722	**22.91/0.8810**	8.87/0.3442
SOTS-outdoor	24.45/0.9291	24.77/0.9355	**24.94/0.9356**	7.58/0.3181
4KID	16.54/0.8420	17.41/0.8701	**18.91/0.8853**	8.42/0.4211

attention Transformer blocks and contrastive loss to the dehazing task. The models are trained using the training sets from [2].

In Tables 2 and 3, the ablation study for providing depth guidance is performed using our proposed model. When no depth is provided, the depth-guided hybrid attention Transformer blocks are converted to the traditional self-attention Transformer blocks. Without the preliminary fusions of depth and hazy images, the hybrid attention mechanisms degrade to the conventional cross-attention mechanisms. As shown in Table 2, the performances gradually improve for the models with no depth guidance, with depth-guided cross-attention, and with our hybrid attention, which demonstrates the importance of providing depth guidance and our hybrid attention mechanisms. The performances are largely impaired for all the datasets without the depth guidance. This observation underscores the significance of incorporating depth guidance and employing our hybrid attention mechanisms.

The preliminary fusions of depth and hazy images by adaptive mixup operations play an important role in our proposed depth-guided hybrid attention Transformer blocks. Although depth-guidance provided in this cross-attention mechanism brings improvements (+2.46 dB on SOTS-indoor) for dehazing, the performances get worse (−1.49 dB) compared with our proposed depth-guided hybrid attention mechanisms. Adaptive integrations of cross-attentions and self-attentions have facilitated exploiting the correlations between depth and dehazing tasks. The attention-based cross-modality fusion appears to be essential as only applying preliminary fusion without the subsequent Transformer blocks but with a plain stacking convolutional neural network fails to get converged during training.

In Figs. 1 and 8, visual representations of a hazy image, our preliminary fusion results, one typical channel of attention maps by our hybrid attention mechanism, by cross attention without the preliminary fusion, by self-attention mechanism without the depth guidance are represented. These typical attention channels are selected with the largest summations of absolute values in each channel of attention maps. Redder colors in attention maps mean larger attention values.

Table 3
The performance comparisons of our self-supervised learning framework with and without depth guidance or Contrastive loss (denoted as CR loss).

Dataset	depth		w/o depth	
	CR	w/o CR	CR	w/o CR
SOTS-outdoor	**24.94/0.9356**	22.25/0.9096	24.45/0.9291	21.69/0.9065
SOTS-indoor	**22.91/0.8810**	19.30/0.8260	18.96/0.8458	16.81/0.7955
4KID	**18.91/0.8853**	17.42/0.8903	16.54/0.8420	16.07/0.8785

Table 4
mAP comparisons of detection for different settings.

Method	Train dataset	Test dataset	
		VOC_Foggy_test	RTTS
(a) YOLOv3 (clear images)	VOC_norm	42.32	31.48
(b) YOLOv3 (hazy images)	VOC_Foggy	57.63	31.79
(c) Dehaze+YOLOv3	VOC_Foggy	57.09	31.23
(d) Dehaze+depth+YOLOv3	VOC_Foggy	**58.04**	**33.03**

The typical channels of attention maps by our hybrid attention mechanism have larger values in distant area such as in the sky and capture richer information for semantic structures, which brings better visual restorations. Self-attention mechanism seems to have fewer correlations to the haze densities. Cross attention without the preliminary fusion reveals fewer details of structures and textures as depth images often lacks of details.

In Table 3, the benefits of contrastive loss are investigated, revealing its significant impact on performance improvement, irrespective of the presence of depth guidance. The restorations are pulled towards clear images and pushed away from the hazy input, which assists feature learning process and improves the results.

4.3. Depth-guided detection on hazy images

For detection tasks on hazy images, synthetic foggy images with detection labels are generated from the detection dataset PSCAL VOC [44], *i.e.*, VOC_Foggy and a real-world foggy dataset RTTS [29] are applied to demonstrate the benefits of depth guidance. We generate the hazes with the applied hazy generation methods for the synthetic VOC_Foggy dataset and trained the detection model, then the trained model is directly applied on RTTS to investigate the real-world performances.

The benefits of introducing depth for detection tasks on hazy images have been investigated in Table 4 compared with mAP (mean

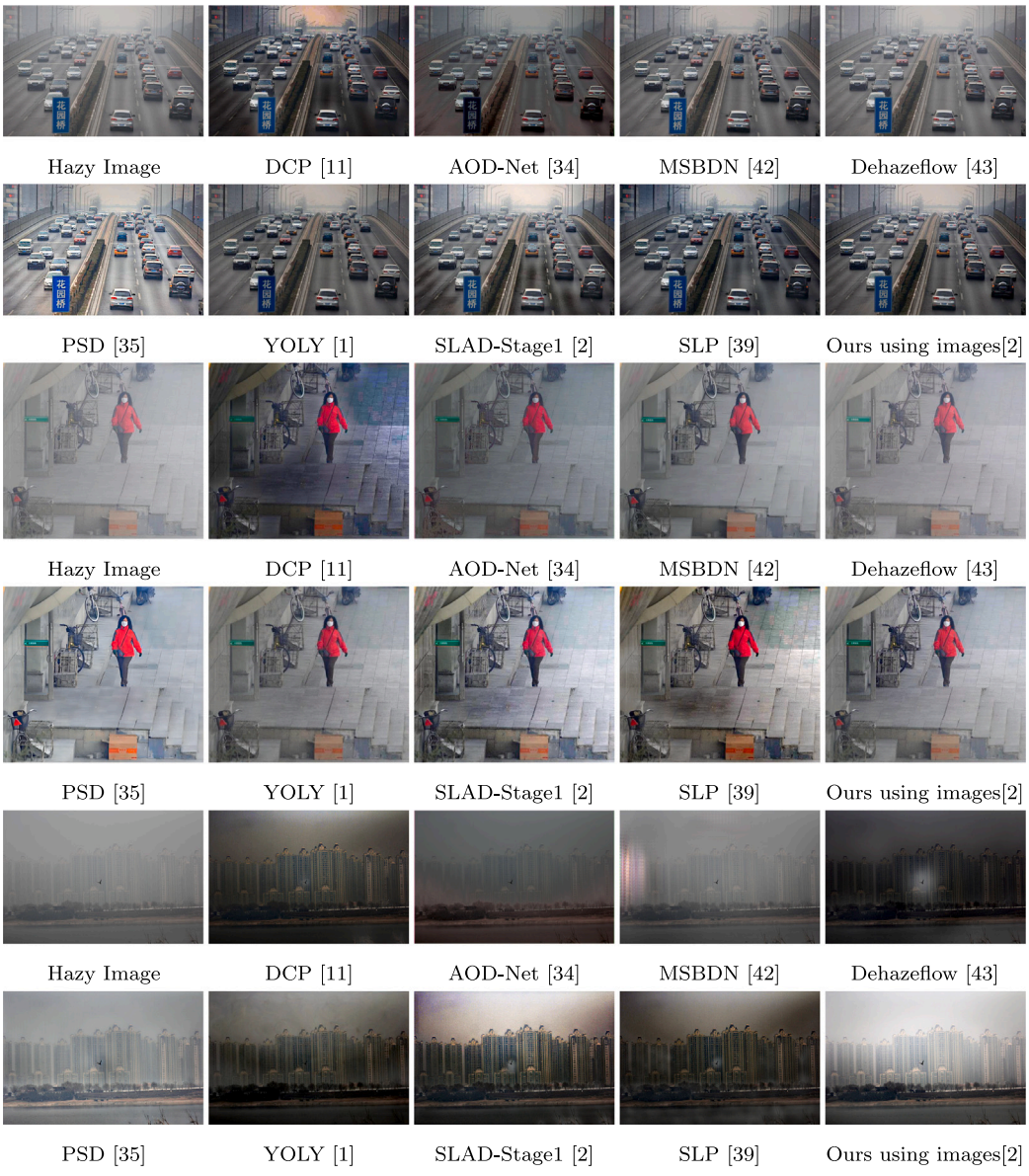

Fig. 7. Dehazing result on the real-world hazy images.

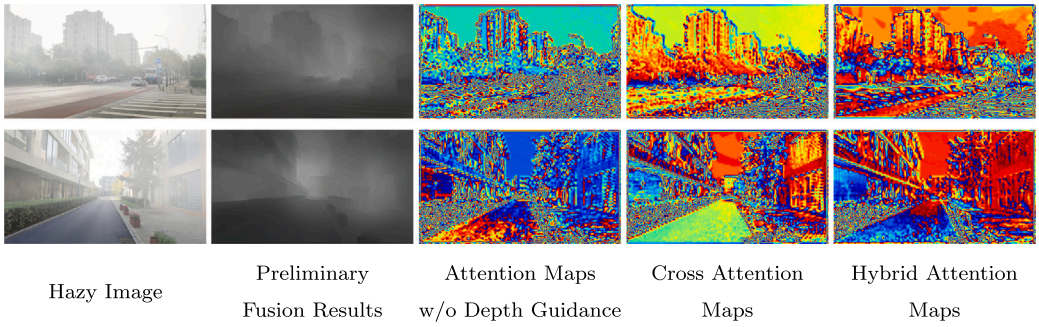

Fig. 8. Visual representations of a hazy image, our preliminary fusion results, one typical channel of attention maps by self-attention mechanism without the depth guidance, by cross attention without the preliminary fusion, and by our hybrid attention mechanism are represented. These typical attention channels are selected with the largest summations of absolute values in each channel of attention maps. Redder colors in attention maps mean larger attention values.

Average Precision). Pre-trained YOLOv3 detection models from the MS COCO dataset [27] are trained on the PASCAL VOC dataset [44] with several settings: (a) ground truth clear images, (b) synthetic hazy input images, (c) dehazed images, (d) dehazed images and depth estimations, collaborating in the proposed attentive multiplication manner. In Fig. 9, image dehazing and detection results for real-world hazy images from the RTTS dataset are represented for comparisons. Hazy images, YOLOv3 detection results by the model trained on dehazed images without

Fig. 9. Image dehazing and YOLOv3 detection results for real-world hazy images from the RTTS dataset. From left to right: hazy images, YOLOv3 detection results on the hazy images, YOLOv3 detection results by models trained on dehazed images without depth, and by models trained on the dehazed images and estimated depth by our method.

depth and the proposed model trained on dehazed images plus the estimated depth by our method. The images are arranged from left to right in Fig. 9. For the real-world hazy images from RTTS, our method could improve the visibility of the hazy images although some failures exist. Even for the failure cases such as the second row in Fig. 9, the detection process still benefits from the depth estimations. Consequently, the model manages to detect cars in distant regions, even when severe occlusions are present.

As shown in Table 4 and Fig. 9, the depth guidance has largely improved the detection accuracies on both synthetic and real hazy image. From Table 4, simply dehazing the images cannot guarantee the detection accuracies to be improved. When the detection model, trained on clear images, is applied to hazy images, it suffers a considerable degradation due to substantial domain gaps. Comparatively, detection on clear images achieves an mAP of 74.47, highlighting ample room for improvement when dealing with hazy images. Nevertheless, the depth guidance holds promise for detection. We investigate the performances of the detection model trained and tested on the clear images, *i.e.*, the detection model is trained on VOC_norm dataset [44] and evaluated on the VOC_norm_test dataset [44]. Comparatively, the detection model trained and tested using clear images achieves an mAP of 74.47, highlighting ample room for improvement when dealing with hazy images.

5. Conclusions

We have proposed a self-supervised image dehazing framework with self-supervised depth guidance to exploit the interactions between depth and hazes for image dehazing tasks. The depth estimation from clear image enables an effective hazy generation and self-supervised depth estimations for hazy inputs. The proposed depth-guided hybrid attentive Transformer blocks effectively explore the depth-guidance to model hazy densities and capture global context information for better feature representations. Our method compares favorable against the state-of-the-art unsupervised or self-supervised dehazing methods. The self-supervised depth estimations not only improve the generalization ability of the model to different dehazing datasets, but also benefit the downstream detection tasks on the hazy images.

Our model applies Transformer blocks, of which computational complexities grow quadratically with the sizes of the input images. Larger sizes of input images for the Transformer blocks bring larger receptive fields to better capture the global correlations. However, the computational burden may prevent further expansion of the receptive field, damaging the promotion of the performances. In the future, efficient architecture design of Transformers could be explored to alleviate the quadratical growth computational problem of Transformers.

CRediT authorship contribution statement

Yudong Liang: Writing – original draft, Software, Project administration, Methodology, Formal analysis, Conceptualization. **Shaoji Li:** Validation, Software. **De Cheng:** Writing – review & editing, Project administration, Conceptualization. **Wenjian Wang:** Writing – review & editing, Supervision, Investigation, Funding acquisition. **Deyu Li:** Writing – review & editing, Supervision, Conceptualization. **Jiye Liang:** Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (No. 62476157, 61802237, 62176198, U21A20513), Fundamental Research Program of Shanxi Province (Nos. 202203021221002, 202203021211291), Science and Technology Major Project of Shanxi Province (Nos. 202101020101019), Key R&D Program of Shanxi Provine (Nos. 202102070301019), Natural Science Foundation of Shanxi Province (Nos. 201901D211176, 202103021223464), Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (Nos. 2019L0066), and the Special Fund for Science and Technology Innovation Teams of Shanxi (Nos. 202204051001015).

References

[1] B. Li, Y. Gou, S. Gu, J.Z. Liu, J.T. Zhou, X. Peng, You only look yourself: Unsupervised and untrained single image dehazing neural network, Int. J. Comput. Vis. 129 (5) (2021) 1754–1767.

[2] Y. Liang, B. Wang, W. Zuo, J. Liu, W. Ren, Self-supervised learning and adaptation for single image dehazing, in: Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI-22, 2022, pp. 1137–1143.

[3] N. Jiang, K. Hu, T. Zhang, W. Chen, Y. Xu, T. Zhao, Deep hybrid model for single image dehazing and detail refinement, Pattern Recognit. 136 (2023) 109227.

[4] S.K. Yadav, K. Sarawadekar, Robust multi-scale weighting-based edge-smoothing filter for single image dehazing, Pattern Recognit. (2023) 110137.

[5] T. Wang, G. Tao, W. Lu, K. Zhang, W. Luo, X. Zhang, T. Lu, Restoring vision in hazy weather with hierarchical contrastive learning, Pattern Recognit. 145 (2024) 109956.

[6] S. Yin, Y. Wang, Y.-H. Yang, A novel image-dehazing network with a parallel attention block, Pattern Recognit. 102 (2020) 107255.

[7] S. Yin, X. Yang, Y. Wang, Y.-H. Yang, Visual attention dehazing network with multi-level features refinement and fusion, Pattern Recognit. 118 (2021) 108021.

[8] Y. Liu, X. Hou, Local multi-scale feature aggregation network for real-time image dehazing, Pattern Recognit. 141 (2023) 109599.

[9] Z. Su, J. Zhang, L. Wang, H. Zhang, Z. Liu, M. Pietikäinen, L. Liu, Lightweight pixel difference networks for efficient visual representation learning, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[10] C.-L. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, C. Li, Image dehazing transformer with transmission-aware 3D position embedding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5812–5820.

[11] K. He, J. Sun, X. Tang, Single image haze removal using dark channel prior, IEEE Trans. Pattern Anal. Mach. Intell. 33 (12) (2010) 2341–2353.

[12] Y. Yang, C. Wang, R. Liu, L. Zhang, X. Guo, D. Tao, Self-augmented unpaired image dehazing via density and depth decomposition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2037–2046.

[13] B. Cai, X. Xu, K. Jia, C. Qing, D. Tao, Dehazenet: An end-to-end system for single image haze removal, IEEE Trans. Image Process. 25 (11) (2016) 5187–5198.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[15] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, Y. Wang, Multimodal token fusion for vision transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12186–12195.

[16] J.M.J. Valanarasu, R. Yasarla, V.M. Patel, Transweather: Transformer-based restoration of images degraded by adverse weather conditions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2353–2363.

[17] X. Wei, T. Zhang, Y. Li, Y. Zhang, F. Wu, Multi-modality cross attention network for image and sentence matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10941–10950.

[18] Y. Yang, C.-L. Guo, X. Guo, Depth-aware unpaired video dehazing, IEEE Trans. Image Process. 33 (2024) 2388–2403.

[19] A. Yang, Y. Liu, J. Wang, X. Li, J. Cao, Z. Ji, Y. Pang, Visual-quality-driven unsupervised image dehazing, Neural Netw. 167 (2023) 1–9.

[20] S. Zhang, X. Zhang, S. Wan, W. Ren, L. Zhao, L. Shen, Generative adversarial and self-supervised dehazing network, IEEE Trans. Ind. Inform. (2023).

[21] V. Sharma, M. Tapaswi, M.S. Sarfraz, R. Stiefelhagen, Self-supervised learning of face representations for video face clustering, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2019, IEEE, 2019, pp. 1–8.

[22] Y. Zheng, C. Zhong, P. Li, H.-a. Gao, Y. Zheng, B. Jin, L. Wang, H. Zhao, G. Zhou, Q. Zhang, et al., Steps: Joint self-supervised nighttime image enhancement and depth estimation, in: 2023 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2023, pp. 4916–4923.

[23] I. Alhashim, P. Wonka, High quality monocular depth estimation via transfer learning, 2018, arXiv e-prints arXiv:1812.11941.

[24] N. Zhang, F. Nex, G. Vosselman, N. Kerle, Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18537–18546.

[25] H. Wu, Y. Qu, S. Lin, J. Zhou, R. Qiao, Z. Zhang, Y. Xie, L. Ma, Contrastive learning for compact single image dehazing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10551–10560.

[26] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, 2018, arXiv preprint arXiv:1804.02767.

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[28] Y. Shao, L. Li, W. Ren, C. Gao, N. Sang, Domain adaptation for image dehazing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2808–2817.

[29] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, Z. Wang, Benchmarking single-image dehazing and beyond, IEEE Trans. Image Process. 28 (1) (2018) 492–505.

[30] Z. Zheng, W. Ren, X. Cao, X. Hu, T. Wang, F. Song, X. Jia, Ultra-high-definition image dehazing via multi-guided bilateral learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16185–16194.

[31] Y. Liu, L. Zhu, S. Pei, H. Fu, J. Qin, Q. Zhang, L. Wan, W. Feng, From synthetic to real: Image dehazing collaborating with unlabeled real data, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 50–58.

[32] C.O. Ancuti, C. Ancuti, R. Timofte, C.D. Vleeschouwer, I-HAZE: a dehazing benchmark with real hazy and haze-free indoor images, 2018, arXiv:1804.05091v1.

[33] C.O. Ancuti, C. Ancuti, R. Timofte, C. De Vleeschouwer, O-haze: a dehazing benchmark with real hazy and haze-free outdoor images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 754–762.

[34] D. Berman, S. Avidan, et al., Non-local image dehazing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1674–1682.

[35] Z. Chen, Y. Wang, Y. Yang, D. Liu, PSD: Principled synthetic-to-real dehazing guided by physical priors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7180–7189.

[36] G. Ju, Y. Choi, D. Lee, J.H. Paik, G. Hwang, S. Lee, Self-supervised dehazing network using physical priors, in: Asian Conference on Computer Vision, Springer, 2022, pp. 290–305.

[37] B. Li, Y. Gou, J.Z. Liu, H. Zhu, J.T. Zhou, X. Peng, Zero-shot image dehazing, IEEE Trans. Image Process. 29 (2020) 8457–8466.

[38] P. Ling, H. Chen, X. Tan, Y. Jin, E. Chen, Single image dehazing using saturation line prior, IEEE Trans. Image Process. (2023).

[39] X. Chen, Z. Fan, P. Li, L. Dai, C. Kong, Z. Zheng, Y. Huang, Y. Li, Unpaired deep image dehazing using contrastive disentanglement learning, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII, Springer, 2022, pp. 632–648.

[40] X. Xiao, Y. Ren, Z. Li, N. Zhang, W. Zhou, Self-supervised zero-shot dehazing network based on dark channel prior, Front. Optoelectron. 16 (1) (2023) 7.

[41] B. Li, X. Peng, Z. Wang, J. Xu, D. Feng, Aod-net: All-in-one dehazing network, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4770–4778.

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

[43] H. Li, J. Li, D. Zhao, L. Xu, DehazeFlow: Multi-scale conditional flow network for single image dehazing, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2577–2585.

[44] W. Liu, G. Ren, R. Yu, S. Guo, J. Zhu, L. Zhang, Image-adaptive YOLO for object detection in adverse weather conditions, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 2, 2022, pp. 1792–1800.

Yudong Liang received the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 2017. He is currently an associate professor in the School of Computer and Information Technology, Shanxi University, Shanxi, China. His research interests include computer vision, pattern recognition and deep learning, with a focus on image enhancement tasks, such as image super-resolution, image dehazing, low-light image enhancement, image quality assessment.

Shaoji Li received the master degree from Shanxi University, in 2024. Her current research interests include image dehazing and low-light enhancement.

De Cheng is an associate professor with School of Telecommunications Engineering, Xidian University, China. He received the B.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2011 and 2017, respectively. From 2015 to 2017, he was a visiting scholar in Carnegie Mellon University, Pittsburgh, USA. His research interests include pattern recognition, machine learning, and multimedia analysis.

Wenjian Wang received the Ph.D. degree from Xi'an Jiaotong University. She is a professor and the Ph.D. supervisor with the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University. Her research interests include machine learning, data mining, and artificial intelligence. She has published more than 260 articles in her research fields, including JMLR, IEEE TKDE, PR, NN, and IEEE TSC.

Deyu Li received the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 2002. He is currently a Professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, School of Computer and Information Technology, Shanxi University, Taiyuan, China. He has published more than 80 articles in international journals. His current research interests include artificial intelligence, granular computing, data mining and machine learning.

Jiye Liang is currently a Professor with the School of Computer and Information Technology, Shanxi University. His current research interests include data mining and machine learning. He has published more than 200 papers in his research fields, including AI, IEEE TPAMI, IEEE TKDE, IEEE TNNLS, PR, NeurIPS, ICML, and AAAI.