SG-Net: Semantic Guided Network for Image Dehazing

Tao Hong[0000−0002−8054−503X], Xiangyang Guo[0000−0002−6426−5804], Zeren Zhang[0000−0003−0573−0339], and Jinwen Ma✉[0000−0002−7388−4295]

Department of Information and Computational Sciences, School of Mathematical Sciences and LMAM, Peking University, Beijing 100871, China
{paul.ht, guoxy}@pku.edu.cn, Eric_Zhang@stu.pku.edu.cn, jwma@math.pku.edu.cn

Abstract. From traditional handcrafted priors to learning-based neural networks, image dehazing technique has gone through great development. In this paper, we propose an end-to-end Semantic Guided Network (SG-Net[1]) for directly restoring the haze-free images. Inspired by the high similarity (mapping relationship) between the transmission maps and the segmentation results of hazy images, we found that the semantic information of the scene provides a strong natural prior for image restoration. To guide the dehazing more effectively and systematically, we utilize the information of semantic segmentation with three easily portable modes: Semantic Fusion (SF), Semantic Attention (SA), and Semantic Loss (SL), which compose our Semantic Guided (SG) mechanisms. By embedding these SG mechanisms into existing dehazing networks, we construct the SG-Net series: SG-AOD, SG-GCA, SG-FFA, and SG-AECR. The outperformance on image dehazing of these SG networks is demonstrated by the experiments in terms of both quantity and quality. It is worth mentioning that SG-FFA achieves the state-of-the-art performance.

Keywords: Image dehazing · Semantic attention · Perception loss.

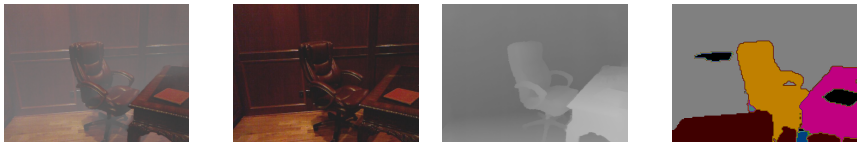
1 Introduction

As a representative task with lots of application value in low-level computer vision, image dehazing has attracted the interest of many researchers in recent years. Like other similar tasks such as image denoising, image deraining, *etc.*, image dehazing can be summarized as an image restoration problem. The atmosphere scattering model [17,19] is formulated as:

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

where $\boldsymbol{I}(x)$ and $\boldsymbol{J}(x)$ are the degraded hazy image and the target haze-free image respectively. $\boldsymbol{A}$ is the global atmosphere light, and $t(x)$ is the medium transmission map. Moreover, we have $t(x) = e^{-\beta d(x)}$ with β and $d(x)$ being the

[1] Codebase page: https://github.com/PaulTHong/Dehaze-SG-Net



(a) Hazy image (b) GT (c) Transmission (d) Segmentation

Fig. 1. Visualization of a sample from NYUv2. We can observe the high similarity (mapping relationship) between the transmission map and the segmentation result, which inspires our exploration on semantic guidance for image dehazing.

atmosphere scattering parameter and the scene depth, respectively. Since the transmission map $t(x)$ and the atmosphere light $\boldsymbol{A}$ are often unknown in real scenarios, image dehazing is an ill-posed problem. Therefore the core challenge is to estimate $t(x)$ and $\boldsymbol{A}$ properly, then we can restore the haze-free image as:

$$\boldsymbol{J}(x) = \frac{\boldsymbol{I}(x) - \boldsymbol{A}}{t(x)} + \boldsymbol{A} \tag{2}$$

We can divide dehazing methods into two classes, traditional prior-based methods and learning-based methods. Classic traditional methods include DCP [11], BCCR [18], CAP [35], *etc.* With the rise of deep learning, many neural network methods are successively proposed, such as DehazeNet [3], MSCNN [21], AOD-Net [13], GCA-Net [4], -FFANet [20], AECR-Net [31]. Driven by the supervised data, these networks are well designed to fulfill the dehazing task, each with pros and cons.

Fig. 1 displays one sample from NYUv2 [25]. As we can see, the transmission map and the semantic segmentation [10] result of this hazy image are highly related. From formula (1) we know that the transmission map $t(x)$ is dependent on the scene depth $d(x)$, while one object in the segmentation result often has similar $d(x)$. The ideal estimated maps of haze-free images shall be smooth in the regions of the same object and only discontinuous across the boundaries of different objects. The semantic segmentation information of images seems to provide a relatively accurate prior for this requirement, by building a mapping relationship with the transmission map. Cheng *et al.* firstly clarified to use semantic information to resolve image dehazing, but calling ordinary convolutional feature maps as semantic module and color module [9] is far-fetched. Song *et al.* proposed a multi-task network for semantic segmentation and dehazing [26], while it is not very targeted and requires a higher resource consumption. Ren *et al.* proposed to incorporate global semantic priors as input to regularize the transmission maps for video dehazing [22], whose effectiveness is partly due to the coherent similarity between video frames. This work enlightens us to migrate to image dehazing and fulfill semantic guidance more abundantly. Zhang *et al.* proposed a semantic-aware dehazing network with adaptive feature fusion [34], but this approach requires the ground-truth semantic label.

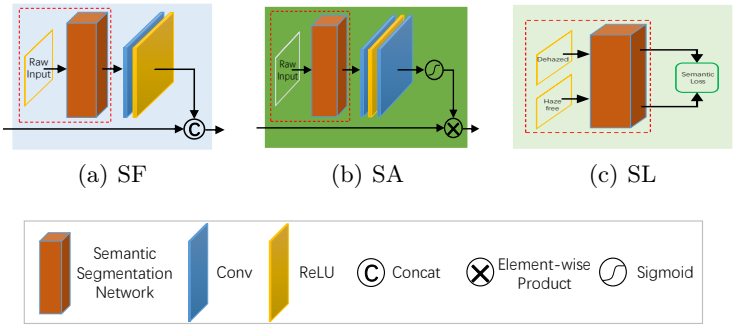


Fig. 2. Semantic Fusion (left), Semantic Attention (middle) and Semantic Loss (right). Note that the module in the red dotted rectangle is shared in one network.

Therefore the key challenge is how to fit the mapping relationship from semantic segmentation to haze transmission with accurately guided tools, rather than negative interference. Apart from feature fusion, attention is considered to be an effective mechanism for neural network learning to be in line with human learning. The attention mechanism [32,28,29] has been researched in detail in the design of neural networks, and is widely applied in Natural Language Processing and Computer Vision. The FFA-Net proposed a feature attention module, which combines the Channel Attention (CA, the gated fusion module of GCA-Net is just a kind of CA) and Pixel Attention (PA) in channel-wise and pixel-wise features, respectively. The excellent performance of FFA-Net inspires us the great potential of attention for integrating semantic information. Furthermore, based on the reconstruction loss between dehazed and haze-free images, we propose a new kind of semantic perception loss to regularize their feature maps through semantic segmentation, resulting in finer dehazed results.

Elegantly combining semantic information with attention mechanism, perception loss, *etc.*, we propose a comprehensive Semantic Guided Network, *i.e.* SG-Net, from two perspectives: network structure and learning course. We mainly adopt three operation modes to fulfill semantic guidance, whose schematics are shown in Fig. 2. SF plus SA improves the network structure, while SL facilitates the optimization process.

- **Semantic Fusion (SF):** Incorporate feature maps of semantic segmentation as new branches into current dehazing networks. Fusion makes shallow features propagation more directional and effective.
- **Semantic Attention (SA):** Directly transfer the feature maps of semantic segmentation as attention. Refine high-level features more specifically at the pixel level.
- **Semantic Loss (SL):** Impose constraints on the perception loss between the semantic segmentation feature maps of dehazed and haze-free images. Optimize towards a more semantic-aware direction.

In the experiments, we choose four representative networks as baselines, *i.e.* AOD-Net, GCA-Net, FFA-Net and AECR-Net, respectively. For the first three networks, the dehazing effect gradually increases in order, while the inference time consumption also gradually increases. After adding our SG mechanisms, we get a stronger SG-Net series, named SG-AOD, SG-GCA, SG-FFA and SG-AECR, respectively.

The main contributions of our work are as follows:

- We propose a novel end-to-end network to restore haze-free images, outperforming existing methods both in quantity and quality, of which SG-FFA gets the state-of-the-art performance.
- We elaborately design SF, SA, and SL to give full play to the guidance of semantic information. With the detailed exploration of cooperative strategies, these operation modes aggregate dehazing effects from different scales and levels.
- Our simple but efficient SG mechanisms can be embedded into the existing network series at will, improving accuracy while only adding a little extra time consumption.

2 Related Work

2.1 Image Dehazing

As introduced in the previous section, image dehazing has evolved from traditional prior-based methods to learning-based methods. The dark channel prior (DCP) [11] is a brilliant discovery. Moreover, the boundary constraint and contextual regularization (BCCR) [18] and color attenuation prior (CAP) [35] are successively proposed.

As for neural network methods, they usually adopt an encoder-decoder structure to learn restoration. AOD-Net [13] directly generates the clean image through a lightweight CNN, named All-in-One Dehazing Network. GCA-Net [4] means Gated Context Aggregation Network, which adopts the smoothed dilation convolution [30] to help remove the gridding artifacts, and leverages a gated subnetwork to fuse the features from different levels. As for FFA-Net [20], *i.e.* Feature Fusion Attention Network, it combines Channel Attention with Pixel Attention mechanism. AECR-Net [31] proposes a contrastive regularization built upon contrastive learning to exploit both the information of hazy images and clear images as negative and positive samples, respectively. And Chen *et al.* proposed a Principled Synthetic-to-real Dehazing (PSD) framework [8], *i.e.* a synthetic data pre-trained backbone, followed by unsupervised fine-tuning with real hazy images. In addition to the synthetic hazy image pairs, Yang *et al* proposed a disentangled dehazing network to generate realistic haze-free images only using unpaired supervision [33], which leads to a new challenge. In this paper, we focus on dehazing with supervised mode.

2.2 Semantic Segmentation

Semantic segmentation aims to cluster image pixels of the same object class with assigned labels. The general semantic segmentation architecture can be considered as an encoder-decoder network. The encoder is usually a pre-trained classification network, like ResNet [12]. And the task of the decoder is to semantically project the discriminable features (lower resolution) learned by the encoder into the pixel space (higher resolution) to obtain a dense classification.

The classic development path of semantic segmentation networks includes FCN [16], U-Net [23], DeepLab series: v1 [5], v2 [6], v3 [7], RefineNet [15], and MTI-Net [27] *etc.* We adopt RefineNet as the semantic segmentation branch unless otherwise specified.

3 Proposed Method

For dehazing task, the hazy image and the haze-free image are usually denoted as I and J. Denoting the whole dehazing network as $\mathcal{D}$, then in general it is optimized towards

$$\min \ \mathscr{L}\left(\mathcal{D}(I), J\right) \tag{3}$$

where $\mathscr{L}$ is the defined restoration loss function.

Our semantic guidance works on network $\mathcal{D}$ in the form of SF and SA, and works on loss $\mathscr{L}$ in the form of SL, respectively. Combining the power of designing network constructures and loss functions, our SG-Net series takes advantage of semantic information to perform well in image dehazing.

3.1 Semantic Fusion and Semantic Attention

To illustrate SF and SA in detail, we can refer to Fig. 2. Feed the raw hazy input to the pretrained semantic segmentation network (denoted as $\mathcal{S}$), then it exports the semantic feature maps ($\mathcal{S}(\cdot)$ stands for the output logits of the last layer). Note that in one whole network, all the SG branches share the same $\mathcal{S}$, therefore we just need to generate the semantic feature maps once and then impose different operations on them.

Denote the convolution layer, ReLU activation function, Sigmoid activation function as Conv, δ, σ, and denote the operation of concatenation, element-wise sum, element-wise product as $\cup$, $\oplus$, $\otimes$, respectively. Besides, denote the middle feature maps from the baseline network branch as F (size $C \times H \times W$), then we can fulfill SF and SA function by operation $\cup$ and $\otimes$, as shown below, where S_F and S_A are generated feature maps from the corresponding SG mechanisms.

Semantic Fusion

$$S_F = \delta(\mathrm{Conv}(\mathcal{S}(I))) \tag{4}$$

$$\check{F} = F \cup S_F \tag{5}$$

Let the size of F and S_F be $C_1 \times H \times W$ and $C_2 \times H \times W$ respectively, then the size of $\check{F}$ is $(C_1 + C_2) \times H \times W$. If the size $H \times W$ of F and S_F are different, we only need to add an upsampling or downsampling in the SF branch, so as the SA branch. Removing all other elements and only leaving the concatenation will degenerate SF into an ordinary skip-layer connection.

Semantic Attention

$$S_A = \sigma(\text{Conv}(\delta(\text{Conv}(\mathcal{S}(I))))) \tag{6}$$

$$\tilde{F} = F \otimes S_A \tag{7}$$

The size of S_A is $1 \times H \times W$. Note that the size of Channel Attention (CA) and Pixel Attention (PA) in FFA-Net [20] are $C \times 1 \times 1$ and $1 \times H \times W$, respectively. And the formula of PA can be expressed as:

$$P_A = \sigma(\text{Conv}(\delta(\text{Conv}(F)))) \tag{8}$$

$$\tilde{F} = F \otimes P_A \tag{9}$$

The key difference between SA and PA is the source of attention. Generating attention through the pretrained semantic segmentation network fully excavates the semantic prior information, leading the dehazing networks to learn the transmission map more specifically. It can be seen from the visualization analysis in Section 4.5 that after a relatively deep stage of feature propagation, the PA only focuses on the local edges of objects, while our SA still has an accurate grasp of the global contour.

As for the general usage strategy of SG mechanisms, we recommend **adopting SF for shallow feature maps and SA for deep feature maps**. SF and SA could be considered as guiding from channel-wise (coarse) and pixel-wise (fine) levels, respectively. Since the $\mathcal{S}(I)$ is more matched with shallow F, so concatenating them with high-level deep F at the back layer is not very appropriate. In that case, the element-wise product provides accurate and efficient semantic guidance at pixel-level to make up for deep F. In addition, we recommend **adopting SF at a relatively low-resolution scale** if there is a downsampling operation in the networks. We infer that a low-resolution scale could alleviate the mismatches between S_F and F, especially for the object edges. More specifically quantified, we recommend placing one or two SF in the front and center (denoting the number of whole layers as n, then we may consider the layer of $[\frac{n}{2}]$, where [] is a rounding operation), and placing one SA in the back (e.g. $n - 3$). A more detailed exploration on the design of SG mechanisms can refer to Section 4.4, including multi-branch strategy, fusion position, *etc.*

3.2 Semantic Loss

For the common reconstruction loss of image dehazing, we adopt residual learning rather than directly learning the haze-free images since the former learning

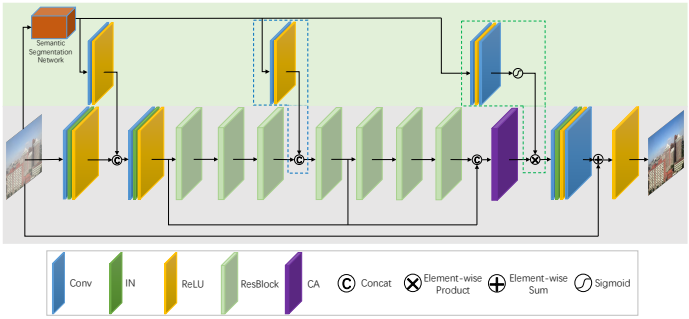


Fig. 3. SG-AOD network architecture.

method is more effective. Hence, the restoration loss function is calculated between the network output $\mathcal{D}(I)$ and residual $J - I$.

$$\begin{aligned} r &= J - I \\ \hat{r} &= \mathcal{D}(I) \\ \mathscr{L}_{\text{rec}} &= ||\hat{r} - r||_1 \end{aligned} \tag{10}$$

where $||\cdot||_1$ is the L1 norm. Through the experiments, we discover that L1 loss performs better than L2 loss, especially when comparing SSIM (Structural Similarity) metrics.

What's more, we propose a new kind of semantic perception loss. To strengthen the semantic relationship between haze-free and dehazed images, we apply a regularization on their feature maps through the pretrained semantic segmentation network, *i.e.* $\mathcal{S}(\mathcal{D}(I))$ and $\mathcal{S}(J)$:

$$\mathscr{L}_{\text{sem}} = ||\mathcal{S}(\mathcal{D}(I)) - \mathcal{S}(J)||_1 \tag{11}$$

where $\mathcal{S}(\cdot)$ could be considered to be substituted by features $\mathcal{S}_i(\cdot)$ from different stages i.

Then, we can combine the reconstruction loss and semantic loss to get our final loss function as

$$\mathscr{L} = \mathscr{L}_{\text{rec}} + \lambda_{\text{sem}}\mathscr{L}_{\text{sem}} \tag{12}$$

where λ_{sem} is an adjustable positive weight. Note that the AECR-Net has also adopted an extra contrastive loss $\mathscr{L}_{\text{con}}$ [31], thus our SG-AECR loss composes of three parts:

$$\mathscr{L} = \mathscr{L}_{\text{rec}} + \lambda_{\text{con}}\mathscr{L}_{\text{con}} + \lambda_{\text{sem}}\mathscr{L}_{\text{sem}} \tag{13}$$

where λ_{con} is also a positive weight. The experiments indicate that our semantic loss brings a significant promotion, especially on SSIM metrics. For more detailed results, please refer to Section 4.4.

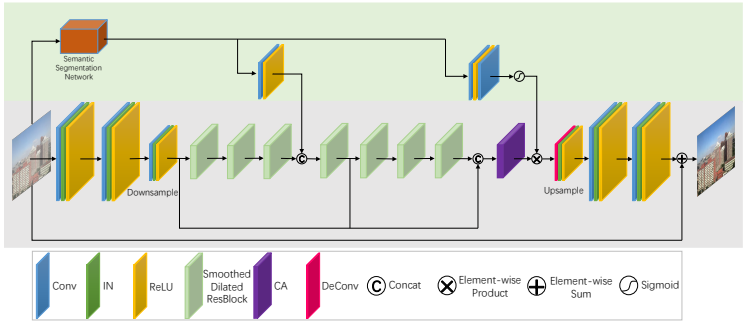


Fig. 4. SG-GCA network architecture.

3.3 SG-Net Series

Firstly, we take AOD-Net as an example to give a detailed introduction of SG-AOD, as shown in Fig. 3. Then we can easily master the other members of SG-Net series: SG-GCA, SG-FFA and SG-AECR. They are shown in Fig. 4 and Appendix, respectively. Each of them can be considered as a combination of the baseline network and the SG module.

SG-AOD Embedding our proposed SG module into the baseline AOD-Net, then the SG-AOD network architecture can be seen in Fig. 3 (SL is not presented, the same as below). Dividing this network into two parts, the lower part is the baseline AOD-Net (grey background), and the upper part is the SG module (green background). Focusing on the SG module, the red cube represents the semantic segmentation network. The SG module has two different kinds of branches, which exactly correspond to the two different semantic guided modes: 2 SFs and 1 SA, as shown in the blue and green dotted rectangle, respectively.

SG-GCA Speaking of the SG-GCA, the baseline GCA-Net adopts the smoothed dilated convolution and a downsampling-upsampling framework. On the basis of it, we add an SF branch and an SA branch at the downsampling scale as the SG module, then SG-GCA is constructed.

SG-FFA As for the SG-FFA, the baseline FFA-Net fully adopts the CA and PA modules. In addition to the final CA and PA, every block in the group structure contains a pair of CA and PA (3 groups contain 19 blocks). We add 2 SF branches in front of the G-1 and G-2 modules. Besides, we update the last PA and the PA of G-3 to SA, as indicated by the red dotted line.

SG-AECR The baseline AECR-Net consists of autoencoder-like downsampling-upsampling framework and contrastive regularization, and the former includes 6 FA blocks, 1 DFE module, and 2 adaptive mixup operations. We add an SF branch after the 3rd FA block and add an SA branch after the DFE module.

4 Experiments

4.1 Dataset and Evaluation Metrics

The image dehazing benchmark universally adopted nowadays is RESIDE [14], which contains synthetic hazy images in both indoor and outdoor scenarios. We adopt ITS (Indoor Training Set, generated from NYUv2 [25] which contains the scene depth $d(x)$) and OTS (Outdoor Training Set) for training respectively, and SOTS (Synthetic Objective Test Set) for test. ITS contains 1399 clean images and 13990 (1399×10) synthetic hazy images, and OTS contains 2061 clean images and 72135 (2061×35) synthetic hazy images. SOTS contains 500 indoor images and 500 outdoor images. The synthesis method is setting different atmosphere light $\boldsymbol{A}$ and scattering parameter β within a certain range.

To further evaluate the robustness of dehazing models in the real-world scene, we also adopt two challenging real-world datasets: Dense-Haze [1] and NH-HAZE (Non-Homogeneous HAZE) [2]. The haze of Dense-Haze is very heavy and the haze of NH-HAZE is not uniformly distributed. These two datasets both contain 55 1200×1600 size images, consisting of 45 training images, 5 validation images and 5 test images. Following the division of AECR-Net, the size of training set and test set are 40 and 5 for Dense-Haze, while 45 and 5 for NH-HAZE.

As for the evaluation metrics, we adopt the common PSNR (Peak Signal to Noise Ratio) and SSIM (Structural Similarity).

4.2 Implementation Details

We finish the experiments on NVIDIA GPU (Tesla V100) by PyTorch framework. The configuration of our SG-Net series that does not appear in the detailed description (please refer to the Appendix) is just the same as the baseline networks. All the SG-Nets adopt Adam as optimizer with momentum $\beta_1 = 0.9$, $\beta_2 = 0.999$. Unless otherwise specified, our utilized semantic segmentation model is RefineNet: RF-LW-ResNet-50. As for the loss weight λ_{sem}, we adjust it so that the reconstruction loss and weighted semantic loss are about at the same level.

4.3 Quantitative and Qualitative Evaluation

As Table 1 shows, we choose the classic DCP method and four representative networks, *i.e.* AOD, GCA, FFA, AECR, to make comparisons. For a certain type of network, from the baseline network to our SG-Net, PSNR and SSIM both gradually get varying degrees of improvement, which strongly demonstrates the effectiveness of our proposed SG mechanisms. During our research, the investigated state-of-the-art methods are FFA-Net on RESIDE, and AECE-Net on Dense-Haze and NH-HAZE, respectively. Our SG-FFA still gets some breakthroughs to reach a new state-of-the-art performance. After increasing the training batch from 2 to 5 with the same iteration, FFA gets further promotion on (PSNR, SSIM) for indoor SOTS: $(38.61, 0.9913)$ of FFA-Net and $(39.18, 0.9932)$ of SG-FFA. It should be noted that the performance on dataset ITS is almost

Table 1. Quantitative comparisons on different datasets for different dehazing methods.

Methods	Indoor SOTS		Outdoor SOTS		Dense-Haze		NH-HAZE	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DCP	16.62	0.8179	19.13	0.8148	-	-	-	-
AOD-Net	21.30	0.8251	25.78	0.9293	-	-	-	-
SG-AOD	**23.33**	**0.8707**	**26.18**	**0.9362**	-	-	-	-
GCA-Net	27.79	0.9452	28.39	0.9500	-	-	-	-
SG-GCA	**29.78**	**0.9621**	**29.15**	**0.9593**	-	-	-	-
FFA-Net	36.39	0.9886	31.69	**0.9800**	-	-	-	-
SG-FFA	**37.56**	**0.9915**	**32.11**	0.9791	-	-	-	-
AECR-Net	33.34	0.9824	-	-	14.43	0.4450	18.50	0.6562
SG-AECR	**33.67**	**0.9832**	-	-	**14.91**	**0.4641**	**18.68**	**0.6609**

the same as the reported performance in the original paper of FFA-Net, but the performance on OTS drops a lot. So we have to take the run result as the real baseline and then adopt SG modules on it for a fair comparison. Similarly, with the open-source code from the authors of AECR-Net [31], we are still not able to reach the best level reported in the paper. So we need to reproduce the code as a baseline to make a comparison, and would not claim the outperformance as new state-of-the-art performance. Our SG-AECR beats the AECR-Net on all three datasets, which once again reveals the effectiveness of semantic guidance for image dehazing. Dense-Haze and NH-HAZE are far more challenging than RESIDE, thus common networks such as AOD-Net behave not well on them. Thanks to the power of contrastive regularization, AECR-Net could get relatively better results.

Furthermore, we display the qualitative comparisons for different dehazing methods here. In Fig. 5, the top 2 rows correspond to the ITS-trained models, while the bottom 4 rows correspond to the OTS-trained models. We can observe that DCP suffers from severe color distortion because of their underlying prior assumptions. AOD-Net is often unable to entirely remove the haze and tends to output low-brightness images. GCA-Net is unsatisfactory at processing high-frequency detail such as textures and edges. Compared to the baseline series, our SG-Net series is superior in detail maintenance and color fidelity, such as the sky region. Concentrating on the last pumpkin image, our SG-FFA has the most obvious dehazing effect, especially on the ground surface. And zooming in on the red wall area of the first image, the white haze of SG-FFA is the weakest, close to nothing. Some visualization results of SG-AECR are shown in Fig. 6. We can observe the superiority of our SG-AECR over AECR-Net, for example, the string 'OUTDOOR' on the ping pong table in the 3th column of images. More quantitative and qualitative results (including training curve, *etc.*) are demonstrated in the supplementary material.

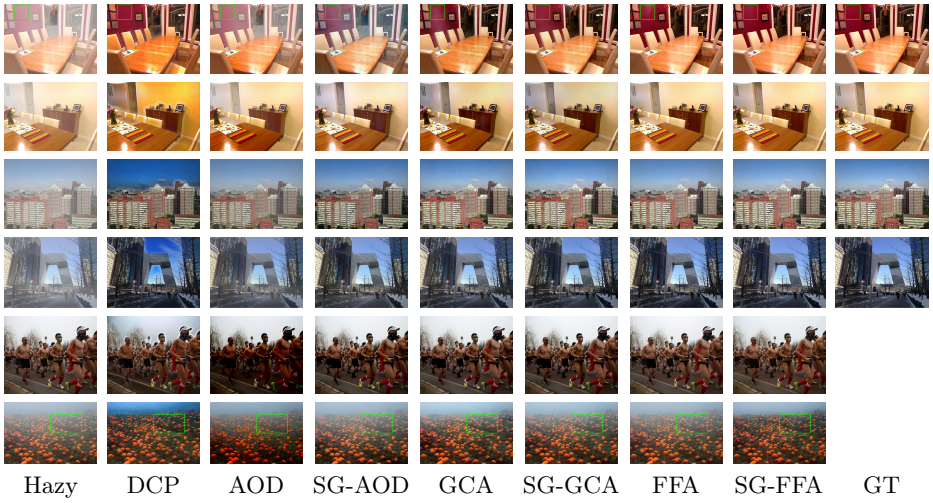


Fig. 5. Qualitative comparisons on SOTS (the top 2 rows for indoor and the middle 2 rows for outdoor) and real-world hazy images (the bottom 2 rows, without corresponding ground truth images) for different dehazing methods. Zoom in on the green rectangle area for more details.

Fig. 6. Qualitative comparisons on indoor SOTS, Dense-Haze and NH-HAZE (corresponding in column order, two columns each) for our SG-AECR. Zoom in on the green rectangle area for more details.

4.4 Ablation Study

To further analyze the function of SG mechanisms, we make a comprehensive ablation study as shown in Table 2. Without any mechanism corresponds to the baseline models, and including SA module means replacing PA with SA. Taking SG-AOD and SG-GCA on ITS as examples, SA and SF both bring promotion and the appropriate combination of them with SL achieves better performance. From PA to SA, the more specific guidance of semantic feature

Table 2. Ablation analysis with different SG mechanisms on indoor SOTS.

	SA		✓		✓	✓
	SF			✓	✓	✓
	SL					✓
SG-AOD	PSNR	21.30	22.07	23.14	23.24	**23.33**
	SSIM	0.8251	0.8360	0.8432	0.8468	**0.8707**
SG-GCA	PSNR	27.79	28.28	28.98	29.06	**29.78**
	SSIM	0.9452	0.9475	0.9493	0.9531	**0.9621**

Table 3. Ablation analysis of different SF positions on indoor SOTS for SG-AOD.

SF Position	none	c1+r1	c1+r3	cr+r3+r7	c2+r3+r7	r3+r7	r1+r3	r1+r3+r5
PSNR	21.30	**23.23**	23.14	22.93	22.60	22.40	23.02	22.27
SSIM	0.8251	**0.8585**	0.8432	0.8322	0.8403	0.8260	0.8312	0.8271

maps and their superiority as attention are fully embodied. Moreover, the flexible transplantation of SG mechanisms is worth mentioning.

And we briefly introduce the design details of SG mechanisms here. For SA, we increase the convolutional layer from 1 to 2 (with a 0.81 increase of PNSR in a set of AOD-ITS comparative experiments, abbreviated as PSNR $\uparrow$ 0.81, the same as below). For SF, if a multi-branch strategy is adopted, we add independent SF branches instead of sharing the same SF parameters (PSNR $\uparrow$ 0.26). These modifications bring positive effects because the fitting capacity of the network has been further strengthened. Besides, the concatenation position of SF is carefully explored. For SG-AOD, adding another SF branch after the 7th ResBlock brings a negative effect (PSNR $\downarrow$ 0.21). For SG-GCA, embedding SF branch before downsampling does not perform better than embedding after downsampling (PSNR $\downarrow$ 0.39). These phenomena reflect that the semantic fusion is more suitable for shallow feature maps and relatively low-resolution scales, which is consistent with our inference. Denoting the layers of SG-AOD as c1-c2 (convolution layer), r1-r7 (residual layer) in order, a more detailed exploration is shown in Table 3 (without SA and SL). We can see the superiority of front c1 over back r7, etc.

On the other hand, we have mentioned that we set the SL weight λ_{sem} to satisfy that the reconstruction loss and weighted semantic loss are about at the same order of magnitude. Therefore, the network will take into account the guidance of these loss functions with nearly equal importance when training. Following this simple and effective selection principle, we have already achieved good results. As for different semantic weights λ_{sem}, we make a simple study on ITS for SG-AOD. From Table 4 we can see that our semantic loss has a significant promotion on dehazing effect, especially on SSIM metrics.

Table 4. Ablation analysis of different SL weights on indoor SOTS for SG-AOD.

SL Weight	0	0.0005	0.001	0.003	0.005
PSNR	23.24	22.82	22.81	**23.33**	23.20
SSIM	0.8468	0.8540	0.8639	0.8707	**0.8747**

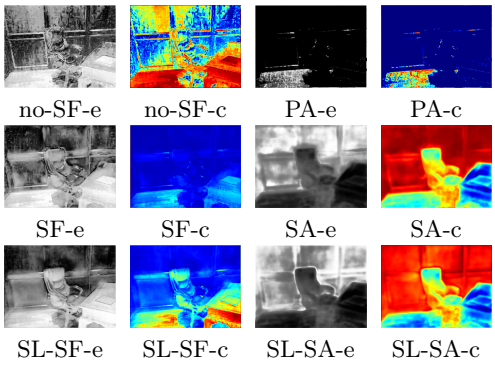


Fig. 7. Visual comparisons on middle feature maps between AOD-Net (top) and our SG-AOD (middle without SL and bottom with SL). The feature map of SF is from the first convolution module after SF module, and PA is equivalent to no-SA. For a clearer observation, *e* and *c* mean *histogram equalization* and *colormap*, respectively.

4.5 Visualization Analysis

As Fig. 7 shows, still taking the demo image in Fig. 1 as an example, we display the middle feature maps from the same position for comparison, with or without semantic guidance (results of SG-GCA are in the Appendix). The feature maps with SG mechanisms contain more details and fit the contour of objects better, thus generating more smooth and accurate dehazed results. We have also tried to implement a visual explanation with Grad-CAM [24], which uses the Gradient-weighted Class Activation Mapping to produce a coarse localization map to highlight the important regions. Yet the results is not as ideal as in Fig. 7. We infer that the dehazing task is not like a classification task which only focuses on partial saliency regions.

4.6 Segmentation Model

We mainly adopt the RF-LW-ResNet-50 trained on NYUv2 dataset as the semantic segmentation model. On ITS, from NYU-Res50 to the relatively stronger NYU-Res152, there is not much difference in the dehazing metrics. We infer that the improved dehazing effect is mainly due to our proposed SG mechanisms, that is, how to better impose semantic guidance, while the impact of segmentation models is relatively slight. Though the images of OTS do not seem to be very consistent with NYUv2-trained segmentation model, their segmentation results

Table 5. Time (training and inference) and parameter analysis.

Methods	AOD		GCA		FFA		AECR	
	Base	SG	Base	SG	Base	SG	Base	SG
Train (h)	3.2	4.1	22.6	25.3	193.4	181.2	40.1	47.2
Infer (s)	0.16	0.19	0.30	0.32	0.47	0.49	0.39	0.46
#Params (M)	0.62	0.66	0.71	0.72	4.46	4.49	2.61	3.93

still play a good role in the SG-Nets. Moreover, substituting PASCAL_VOC-trained model (21 classes) for NYUv2-trained model partially improves the dehazing metrics on OTS, because PASCAL_VOC is more consistent with outdoor images. For more details on the exploration of segmentation models such as the superiority of soft logits over hard outputs, please refer to the Appendix.

4.7 Efficiency Analysis

Finally, we give simple comparisons of training time (on ITS), inference time (per image on average, on SOTS), and parameters between our SG-Nets and the baseline series, as shown in Table 5. Note that the time consumption corresponds to 1 GPU, and the parameters of pretrained semantic segmentation model are not counted. We can see that the efficient SG mechanism does not bring a lot of extra time and space consumption, which are mainly dominated by the segmentation model. Thus pre-storing the semantic segmentation feature maps of the training data can save the training time if needed. And it is worth noting that the training time of SG-FFA is less than FFA-Net. This is due to that we replace many SAs with PAs, while the input channel numbers of SA's 2 layers are less than PA's, $[40, 5]$ *versus* $[64, 8]$.

5 Conclusion

In this paper, we have proposed an end-to-end Semantic Guided Network for image dehazing. Semantic guidance is fulfilled with three simple yet effective designs: Semantic Fusion, Semantic Attention and Semantic Loss. The outperformance over existing methods is demonstrated both in quantity and quality. And our SG mechanisms could be flexibly embedded into a certain network so that a better tradeoff between accuracy and speed would be sought. In future work, it is worth studying to further explore the explanation of semantic mechanism (similar function of field depth or edge contour) and extend it to other low-level vision tasks.

Acknowledgements This work was supported by the Natural Science Foundation of China under grant 62071171.

References

1. Ancuti, C.O., Ancuti, C., Sbert, M., Timofte, R.: Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In: 2019 IEEE international conference on image processing (ICIP). pp. 1014–1018. IEEE (2019)
2. Ancuti, C.O., Ancuti, C., Timofte, R.: Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 444–445 (2020)
3. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D.: Dehazenet: An end-to-end system for single image haze removal. IEEE Transactions on Image Processing **25**(11), 5187–5198 (2016)
4. Chen, D., He, M., Fan, Q., Liao, J., Zhang, L., Hou, D., Yuan, L., Hua, G.: Gated context aggregation network for image dehazing and deraining. In: 2019 IEEE winter conference on applications of computer vision (WACV). pp. 1375–1383. IEEE (2019)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
7. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
8. Chen, Z., Wang, Y., Yang, Y., Liu, D.: Psd: Principled synthetic-to-real dehazing guided by physical priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7180–7189 (2021)
9. Cheng, Z., You, S., Ila, V., Li, H.: Semantic single-image dehazing. arXiv preprint arXiv:1804.05624 (2018)
10. Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Garcia-Rodriguez, J.: A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857 (2017)
11. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE transactions on pattern analysis and machine intelligence **33**(12), 2341–2353 (2010)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: Aod-net: All-in-one dehazing network. In: Proceedings of the IEEE international conference on computer vision. pp. 4770–4778 (2017)
14. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Reside: A benchmark for single image dehazing. arXiv preprint arXiv:1712.04143 **1** (2017)
15. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1925–1934 (2017)
16. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)

17. McCartney, E.J.: Optics of the atmosphere: scattering by molecules and particles. New York (1976)
18. Meng, G., Wang, Y., Duan, J., Xiang, S., Pan, C.: Efficient image dehazing with boundary constraint and contextual regularization. In: Proceedings of the IEEE international conference on computer vision. pp. 617–624 (2013)
19. Narasimhan, S.G., Nayar, S.K.: Vision and the atmosphere. International journal of computer vision **48**(3), 233–254 (2002)
20. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: Ffa-net: Feature fusion attention network for single image dehazing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11908–11915 (2020)
21. Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., Yang, M.H.: Single image dehazing via multi-scale convolutional neural networks. In: European conference on computer vision. pp. 154–169. Springer (2016)
22. Ren, W., Zhang, J., Xu, X., Ma, L., Cao, X., Meng, G., Liu, W.: Deep video dehazing with semantic segmentation. IEEE Transactions on Image Processing **28**(4), 1895–1908 (2018)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
24. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
25. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012)
26. Song, T., Jang, H., Ha, N., Yeon, Y., Kwon, K., Sohn, K.: Deep multi-task network for simultaneous hazy image semantic segmentation and dehazing. Journal of Korea Multimedia Society **22**(9), 1000–1010 (2019)
27. Vandenhende, S., Georgoulis, S., Van Gool, L.: Mti-net: Multi-scale task interaction networks for multi-task learning. In: European Conference on Computer Vision. pp. 527–543. Springer (2020)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
29. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
30. Wang, Z., Ji, S.: Smoothed dilated convolutions for improved dense prediction. Data Mining and Knowledge Discovery pp. 1–27 (2021)
31. Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., Ma, L.: Contrastive learning for compact single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10551–10560 (2021)
32. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
33. Yang, X., Xu, Z., Luo, J.: Towards perceptual image dehazing by physics-based disentanglement and adversarial training. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)

34. Zhang, S., Ren, W., Tan, X., Wang, Z.J., Liu, Y., Zhang, J., Zhang, X., Cao, X.: Semantic-aware dehazing network with adaptive feature fusion. IEEE Transactions on Cybernetics (2021)
35. Zhu, Q., Mai, J., Shao, L.: A fast single image haze removal algorithm using color attenuation prior. IEEE transactions on image processing **24**(11), 3522–3533 (2015)