Supplementary Material of 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`

1 Network Architecture

The network architectures of SG-FFA and SG-AECR are shown in Fig. 1 and Fig. 2, respectively.

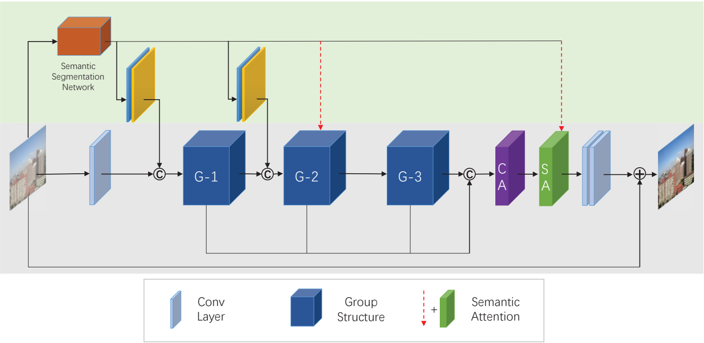


Fig. 1. SG-FFA network architecture.

2 Implementation Details

Hyperparameters

- SG-AOD: Total epoch is 10, training batch is 8, initial learning rate is 0.001 and it decays by 0.1 at the epoch of $2,\ 5,\ 8$, $\lambda_{\text{sem}} = 0.003$.
- SG-GCA: Training batch is 8, initial learning rate is 0.001. For ITS, total epoch is 100 and learning rate decays by 0.1 at the epoch of $40,\ 80$; for OTS, total epoch is 30 and learning rate decays by 0.1 at the epoch of $15,\ 25$.

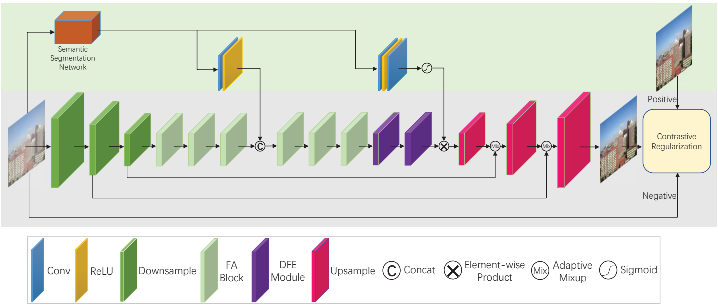


Fig. 2. SG-AECR network architecture.

$\lambda_{\text{sem}} = 5$. Note that for the input image, its pixel ranges within $0 \sim 255$ and subtracts 128 first (behaving better than standardized to $0 \sim 1$), and its channel is 4 with 1 channel from the calculated edge.

- SG-FFA: Training batch is 2, initial learning rate is 0.0001 with cosine annealing. ITS with total iteration 500000 and OTS with total epoch 1000000. $\lambda_{\text{sem}} = 0.05$. The training dataset is augmented with randomly rotated by $90, 180, 270$ degrees, horizontally flipped, and cropped by 240×240.
- SG-AECR: Training batch is 16, initial learning rate is 0.0002 with cosine annealing. ITS with total iteration 80000 and the augmentation methods are the same as SG-FFA. For Dense-Haze and NH-HAZE, we only change total iteration to 8000 and cropped size to 240×320. $\lambda_{\text{sem}} = 0.05$, and $\lambda_{\text{con}} = 0.1$ is taken from the default value.

Baseline Network Modification We have made some modifications on AOD-Net and GCA-Net to generate new baselines, so that a better dehazing benchmark will be reached (all reported results in this script correspond to the new baselines). Compared to the original AOD-Net [2], our baseline AOD-Net has some modifications: adding CA modules which are borrowed from FFA-Net, residual learning, and extending ResBlock from 3 to 7, *etc.* Since the gated fusion module of the original GCA-Net [1] is a kind of CA, so we just add a PA after the CA as our baseline GCA-Net. But we also want to emphasize that these changes do not interfere in any way with the verification of the validation of SG mechanisms. In other words, if we remove these changes to restore the original networks, applying SG mechanisms still gets significant outperformance.

3 Visualization Analysis

Besides the visual comparisons between AOD-Net and SG-AOD, we also provide the results of GCA-Net and SG-GCA here, concentrating on the effectiveness of attention mechanism. As Fig. 3 shows, after a relatively deep stage of feature propagation, the PA only focuses on the local edges of objects, while our SA still

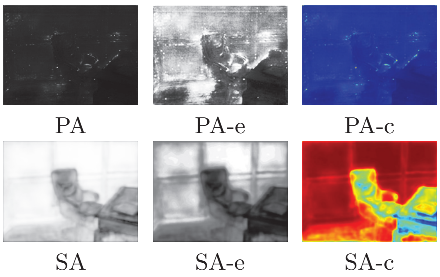


Fig. 3. Visual comparisons on attention mechanism between the PA of GCA-Net (top) and the SA of our SG-GCA (bottom). For a clearer observation, e and c mean *histogram equalization* and *colormap*, respectively.

has an accurate grasp of the global contour, which is a more effective guidance for image dehazing.

4 Segmentation Model

As described in the main text, unless otherwise specified, we mainly adopt the RF-LW-ResNet-50 version of RefineNet as the semantic segmentation model, which no longer participates in the training process of dehazing. Abbreviating the segmentation model trained on NYUv2 dataset as NYU-Res50, and trained on PASCAL_VOC as VOC-Res50. The output channel of the trained segmentation model is 40 for NYUv2 and 21 for VOC (*i.e.* the class number of the corresponding dataset), we convert it to 16 in the SF mechanism, then concatenate the semantic feature maps S_F with the middle feature maps F whose channel is usually 64 (SG-AECR with 256).

Intuitively, the stronger the segmentation model is, the better the dehazing effect is. As Table 1 shows, we have tried different segmentation models: different network architectures or trained on different datasets. On ITS, from NYU-Res50 to the relatively stronger NYU-Res152, there is not much difference in the dehazing metrics, probably due to the close function of the two networks. Though the images of OTS do not seem to be very consistent with NYUv2-trained segmentation model, their segmentation results 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 (but still very slightly), because PASCAL_VOC is more consistent with outdoor images. In a word, the performances under different segmentation models are not much different. 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.

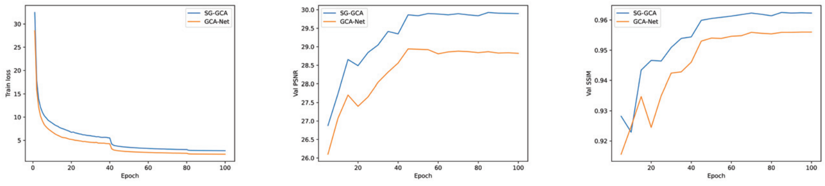
As for the output of the segmentation model, *i.e.* $\mathcal{S}(I)$, we take the soft logits of the last network layer (size $C \times H \times W$) instead of the final hard segmentation results (size $1 \times H \times W$). The former soft logits are more proper to couple with neural networks, which have more channels than the original hard outputs (e.g. 40 vs. 1) to convey semantic guidance, especially for the SF mechanism. The hard results are taken from an *argmax* operation on the soft logits, which has already lost much information. Then we first have to convert the hard outputs with one-hot encoding before feeding them to the SG modules. From another perspective, the soft logits can be thought of as a form of regularization, somewhat similar to smoothing labels. Once we convert the hard results to one-hot encodings to replace the soft logits, the dehazed metrics drop a lot and can not even exceed the baseline. Thus replacing the neural network segmentation model with traditional segmentation model does not seem to work, since the latter always outputs hard segmentation results. Otherwise, the efficiency of SG-Net would be promoted.

Table 1. Quantitative comparisons of SG-AOD on SOTS for different semantic segmentation models.

Seg Model		NYU-Res50	NYU-Res152
ITS	PSNR	23.24	23.31
	SSIM	0.8468	0.8453
Seg Model		NYU-Res50	VOC-Res50
OTS	PSNR	26.18	26.38
	SSIM	0.9362	0.9372

5 Quantitative and Qualitative Evaluation

We display more detailed experiment results here. Taking the training course of SG-GCA on ITS as a representation, its superiority over the baseline GCA-Net is shown in Fig. 4, including validation PSNR and validation SSIM. Note that the training loss of SG-GCA is higher than GCA-Net since the former contains an extra term of semantic perception loss.



(a) Train loss (b) Val PSNR (c) Val SSIM

Fig. 4. Comparisons of training courses on ITS between GCA-Net and our SG-GCA.

We first concentrate on dehazing for regular hazy images, so the mainly compared works are all only for the benchmark RESIDE dataset. To further verify the effectiveness of our work, we transfer SG mechanisms to the more challenging Dense-Haze and NH-HAZE datasets. Table 2 shows the quantitative comparisons on several datasets between the baseline AECR-Net [3] and our SG-AECR. AECR-Net adopts a module named Contrastive Regularization (CR) built upon contrastive learning to exploit both the information of hazy images and clear images as negative and positive samples, respectively. They have done an ablation study about the different positive and negative sample rates on CR, and found that the evaluation metrics with the rate of 1 : 10 exceed 1 : 1 a little. As our training batch is set to 16, so we do three groups of experiments with the CR rate of 1 : 1, 1 : 10, and 1 : 16, respectively. Fix any CR rate group of Table 2 to compare, we can see that our SG-AECR performs almost better than AECR-Net. And the CR rate of 1 : 10 is a relatively optimal value.

Table 2. More detailed quantitative comparisons on indoor SOTS, Dense-Haze and NH-HAZE between AECR-Net and our SG-AECR.

Methods	CR Rate	Indoor SOTS		Dense-Haze		NH-HAZE	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
AECR-Net	1:1	33.34	0.9824	14.36	0.4229	18.43	0.6432
	1:10	32.37	0.9799	14.43	0.4450	18.50	0.6562
	1:16	32.92	0.9812	14.56	0.4444	17.39	0.6335
SG-AECR	1:1	**33.67**	**0.9832**	14.81	0.4351	18.60	0.6531
	1:10	32.54	0.9806	**14.91**	**0.4641**	**18.68**	**0.6609**
	1:16	32.84	0.9805	14.54	0.4530	17.84	0.6521

As for the data division of Dense-Haze and NH-HAZE, we have mentioned in the text that we adopt the same division method as AECR-Net, *i.e.*, the size of training set and test set are 40 and 5 for Dense-Haze, while 45 and 5 for NH-HAZE. These two datasets both contain a total of 55 images. AECR-Net adopts this kind of division method because the labels of 5 validation images or 5 test images were not released before. Now they are all released, thus we also have tried to divide them into: 50 for training and 5 for test. Note that all the division methods are arranged in order of image number. Due to the change of test images and increase of training images, with the CR rate of 1 : 10, our SG-AECR gets PSNR of 16.62 and SSIM of 0.5501 on Dense-Haze, while 18.57 and 0.6357 on NH-HAZE.

Furthermore, we present more visual results on indoor SOTS, Dense-Haze, and NH-HAZE, as shown in Fig. 5, 6, and 7, respectively. Dehazing task is relatively easier on Indoor SOTS, but much more challenging on Dense-Haze and NH-HAZE. Therefore the visual effects in Fig. 5 are pretty good. For dense haze and non-homogeneous haze, there are still some gaps between the dehazed images and the ground truths. For severely degraded hazy images, they are

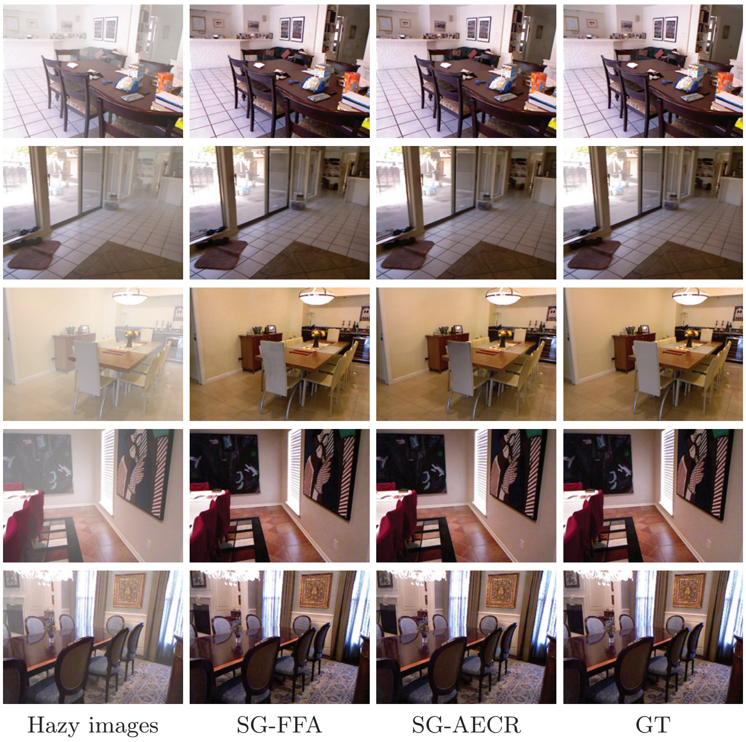
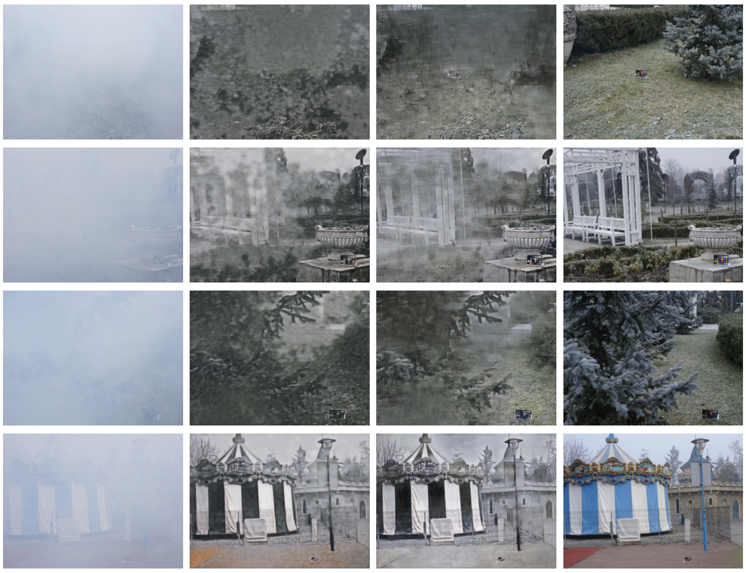


Fig. 5. More visual display on indoor SOTS for our SG-FFA and SG-AECR.

worthy of further exploration. Focusing on Fig. 6 and 7, we can observe the superiority of our SG-AECR over AECR-Net, such as the scratches around the street lamps in the 2nd row of NH-HAZE. In general, SG-AECR achieves clearer object contours and fewer unrealistic halo artifacts. We can zoom in for more details.

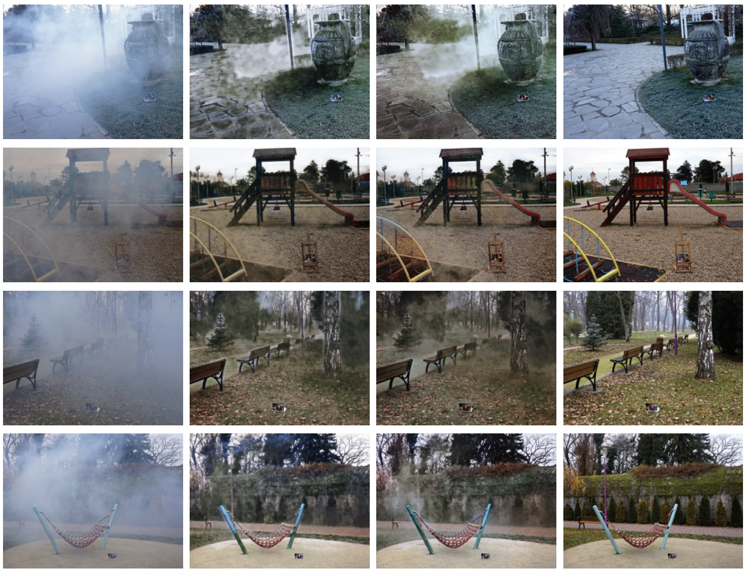
References

1. 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)
2. 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)
3. 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)



Hazy images AECR-Net SG-AECR GT

Fig. 6. More visual display on Dense-Haze for AECR-Net and our SG-AECR.



Hazy images AECR-Net SG-AECR GT

Fig. 7. More visual display on NH-HAZE for AECR-Net and our SG-AECR.