Supplementary Materials for AGG-Net: Attention Guided Gated-convolutional Network for Depth Image Completion

Dongyue Chen[1,2,3,*], Tingxuan Huang[1,*], Zhimin Song[1], Shizhuo Deng[1,2], Tong Jia[1,3]
[1] College of Information Science and Engineering,
Northeastern University Shenyang 110819, Liaoning, China
[2] National Frontiers Science Center for Industrial Intelligence and Systems Optimization,
Northeastern University, Shenyang 110819, Liaoning, China
[3] Foshan Graduate School of Innovation, Northeastern University, Foshan 528311, Guangdong, China
chendongyue@ise.neu.edu.cn, cangshuhuang@gmail.com
sg1997k@gmail.com, dengshizhuo@ise.neu.edu.cn, jiatong@ise.neu.edu.cn

1. More Details of Data Preprocessing

In the main paper, we conduct several experiments on three popular datasets: NYU-Depth V2, DIML, and SUN RGB-D. In this section, we provide more details about the data pre-processing of the training and testing sets, including raw depth maps, ground truth, and color maps. As mentioned in section 4.1 of the main paper, images of the three datasets are in different sizes. Further details of the procedures are described as follows.

- **Data Augmentation of raw depth images.** In training, firstly, the raw depth image is randomly rotated within a range of -5 to 5 degrees, randomly center cropped within a range of 50% to 80%, and then resized to a fixed shape. Secondly, a randomly chosen area of the depth map is erased and replaced by void pixels. Existing works indicate that this operation can improve the robustness and generalization ability of CNN models [4]. Thirdly, the final depth image is directly mapped to the range of $[0, 1]$. In testing, we only crop the center patch in half size of the raw depth image, and then resize it to a specific size with the same setting above.

- **Data Augmentation of ground truth.** In training, we implement the same augmentation operations including random rotating, cropping and resizing with the same parameters as we did to the corresponding raw depth map, and finally we mapped the augmented ground truth image to the range of $[0, 1]$ directly. In testing, the process is the same as the raw depth map.

- **Data Augmentation of color maps.** In training, firstly, the color maps are randomly rotated, cropped, and resized with the same parameters as the corresponding raw depth map. Secondly, we adopt the ColorJitter to change their brightness, contrast and saturation randomly to simulate various shooting conditions such as different illuminations and exposure levels. At last, the color maps are normalized using the mean value and std of the ImageNet [1]. In testing, we only center-crop and resize the color map as the raw depth map, and normalize it as above.

2. Contribution of the Pre-filling module

In the main paper, we proposed a pre-filling network to restore all the missing areas coarsely. It is worth noting that the pre-filling module keeps all the valid depth values unchanged while replacing the missing parts only. Due to the space limitation, we didn't show the complete ablation study results on the pre-filling network in the main paper. In this section, we provide more results about the contribution of the pre-filling module to demonstrate its effectiveness.

Benchmark	Pre-filling	RMSE↓	Rel↓	$\delta_{1.10}$↑
(a) NYU-Depth V2.	–	0.105	0.016	97.4
	✓	0.092	0.014	98.3
(b) DIML	–	0.093	0.016	97.7
	✓	0.078	0.011	98.5
(c) SUN RGB-D	–	0.147	0.040	95.9
	✓	0.128	0.035	97.1

Table 1. Ablation study of the Pre-filling module on three datasets.

As shown in Tab 1, the ablative experimental results on the three datasets prove that the pre-filling module does contribute to the task of depth completion in a positive way.

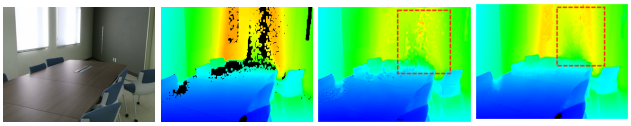

Figure 1. Visualized comparisons on a typical sample between with/without pre-filling.

Fig. 1 shows the visualized comparison between the models with or without the pre-filling module. in which plaques and artifacts closely related to the original missing area can be seen at the marked region when the pre-filling module is absent. In contrast, the adverse impacts of these invalid pixels are largely reduced when the pre-filling module is applied.

3. Comparison between AG-SC and AGs

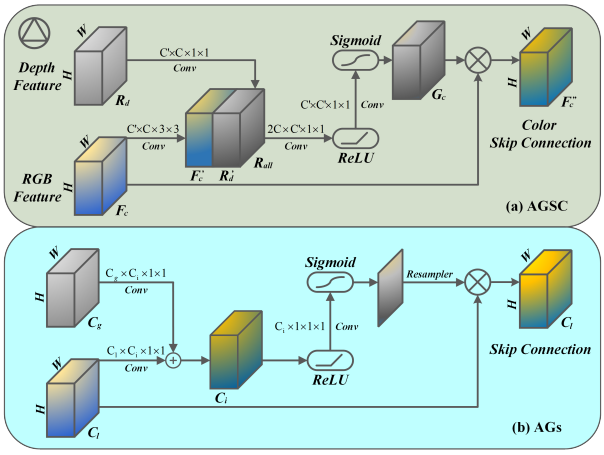

Figure 2. Detailed structures of the (a) AG-SC and (b) AGs. [2].

Our proposed AG-SC module is used to refine the color skip connection from the encoder to the decoder in different scales, of which the effectiveness has been proved in the main paper. However, it is easy to confuse the proposed AG-SC with a similar work named Attention Gates (AGs for short) [2] which is also presented to improve the skip connection of the U-Net [3]. The difference between our proposed AG-SC and the AGs can be clarified as shown in Fig. 2. The AGs. was originally designed for combining features of the same modality to filter the skip connections. In the architecture of the AGs., the channel-wise summation is applied to combine the color features and the depth features, which cannot provide a good accommodation for both of the two modalities because the simple addition of the two modalities is not only meaningless but also causing the loss of information, for example, $x = 5$ and $y = 3$ are apparently more informative than $x + y = 8$. In contrast, our proposed AG-SC module concatenates the two feature tensors together along the channel axis and then employs a VConv unit to conduct their fusion, which preserves more information of color and depth features and provides more learning space for cross-modal attention. To validate our contribution, we carry out the comparison experiment by replacing the traditional skip connection with our proposed AG-SC and the AGs respectively. The corresponding results are shown in Tab 2.

AGs[2]	AG-SC	RMSE↓	Rel↓	$\delta_{1.10}$↑
✓		0.128	0.016	97.2
	✓	0.092	0.014	98.3

Table 2. Comparison results of the AGs and AG-SC module on NYU-Depth V2.

It can be seen from the results clearly that the proposed AG-SC module is more precise compared to the AGs. It is because the correlation between color and depth features is too complex to build on the simple addition of the two modalities provided by the AGs. In comparison, our proposed AG-SC can learn the complicated joint distribution of the two modalities through the convolution on the concatenation of the two modalities of feature tensors, which provides a better way to filter the color skip connections and then suppress the interference comes with the depth-irrelevant color features.

4. More Details of Completion Results

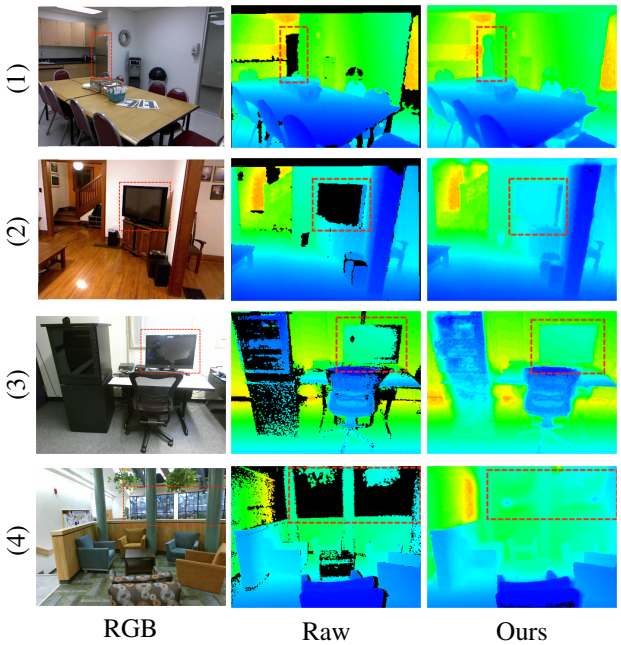

Figure 3. Visualizations of the color and raw depth images, and the corresponding results of our method on the two datasets, rows (1) and (2) are from the NYU-Depth V2, rows (3) and (4) are selected from SUN RGB-D. Most significantly improved regions are marked with boxes.

Though the quantitative results of our model were given in Tab. 3 of the main paper, limited by the paper length, only the visualization results on some typical cases of the DIML dataset were displayed in Fig. 6 of the main paper. In this section, we provide more results of our model on the datasets NYU-Depth V2 and SUN RGB-D as a supplement, as shown in Fig 3.

In Fig. 3 (1), the depth values of the refrigerator are entirely missing because of the specular reflection, while our model can effectively restore these missing areas with clear boundaries. It demonstrates that the proposed AG-GConv module can improve the prediction of invalid depth values by introducing color features into the dual-branch framework. In Fig. 3 (2), (3), despite the mirror images of the environment reflected from the screens or the scenery outside the windows, the missing depth values of these areas are still recovered well by our model, regardless of the disturbance of color information. It proves that the proposed AG-SC module and the edge persistence loss improve the resistance of our model to the depth-irrelevant color features. Especially in Fig. 3 (4), the huge missing holes of the windows with rich colored textures can still be filled with robust depth values. It is because that the proposed CA mechanism makes the AG-GConv module predict the missing depth values well based on a wider range of the background pixels.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[2] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 2

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, 2015. 2

[4] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020. 1