Completion as Enhancement: A Degradation-Aware Selective Image Guided Network for Depth Completion

Supplementary Material

6. Setup

6.1. Dataset

NYUv2 [35] is the most commonly used dataset for depth completion, consisting of 1,449 sets from 464 different indoor scenes using Microsoft Kinect. AGG-Net divides this dataset into 420 images for training and 1,029 for testing. The initial resolution of the RGB-D pairs is 640 × 480, which are randomly cropped and resized to 324 × 288.

DIML [8] is a newly introduced dataset featuring a series of RGB-D frames captured by the Kinect V2 for indoor scenes and the ZED stereo camera for outdoor scenes. In addition to the usual invalid patterns, it includes numerous edge shadows and irregular holes, making it ideal for evaluating the adaptability of depth completion models to various invalid patterns. AGG-Net focuses solely on the indoor portion of the dataset, which comprises 1,609 RGB-D pairs for training and 503 pairs for testing. The resolution is randomly cropped and resized from 512 × 288 to 320 × 192.

SUN RGB-D [36] is a comprehensive dataset containing 10,335 refined RGB-D pairs, captured using four different sensors across 19 major scene categories. According to the official scheme, the training set includes 4,845 pairs, while the testing set comprises 4,659 pairs. The resolution is randomly cropped and resized from 730 × 530 to 384 × 288.

TOFDC [50] is collected using the time-of-flight (TOF) depth sensor and RGB camera of a Huawei P30 Pro, encompassing various scenes such as textures, flowers, bodies, and toys under different lighting conditions and in open spaces. It includes 10,000 RGB-D pairs of resolution 512 × 384 for training and 560 pairs for evaluation. The ground truth depth maps were captured by a Helios TOF camera.

6.2. Metric

Following prior works [2, 29, 31, 37, 50], we evaluate our model using the Root Mean Squared Error (RMSE), Absolute Relative Error (REL), and Accuracy under the threshold $\delta_{1.25^i}$, where $i = 1, 2, 3$. These metrics provide a comprehensive assessment of the model's performance in terms of depth estimation accuracy and error tolerance. The specific definitions and calculations are provided in Tab. 5.

6.3. Implementation Detail

Our SigNet is implemented in PyTorch and trained on a single NVIDIA RTX 3090 GPU. We train the model for 20 epochs using the Adam optimizer [13] with momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, an initial learning rate of 1×10^{-4}, and a weight decay of 1×10^{-6}. The balance coefficient γ in the loss function is set to 0.1. To further improve model performance and generalization, we apply data augmentation techniques, including random cropping, flipping, and normalization during training, which helps the model better handle variations in input data.

For one pixel p in the valid pixel set $\mathbb{P}$ of $\mathbf{y}$:	
– RMSE	$\sqrt{\frac{1}{\lvert\mathbb{P}\rvert}\sum\left(\hat{\mathbf{y}}_p - \mathbf{y}_p\right)^2}$
– REL	$\frac{1}{\lvert\mathbb{P}\rvert}\sum\left\lvert\hat{\mathbf{y}}_p - \mathbf{y}_p\right\rvert/\hat{\mathbf{y}}_p$
– $\delta_{1.25^i}$	$\frac{\lvert\mathbb{S}\rvert}{\lvert\mathbb{P}\rvert}$, $\mathbb{S} : \max\left(\hat{\mathbf{y}}_p/\mathbf{y}_p, \mathbf{y}_p/\hat{\mathbf{y}}_p\right) < 1.25^i$

Table 5. Metric definition. $\mathbf{y}$: prediction, $\hat{\mathbf{y}}$: ground truth.

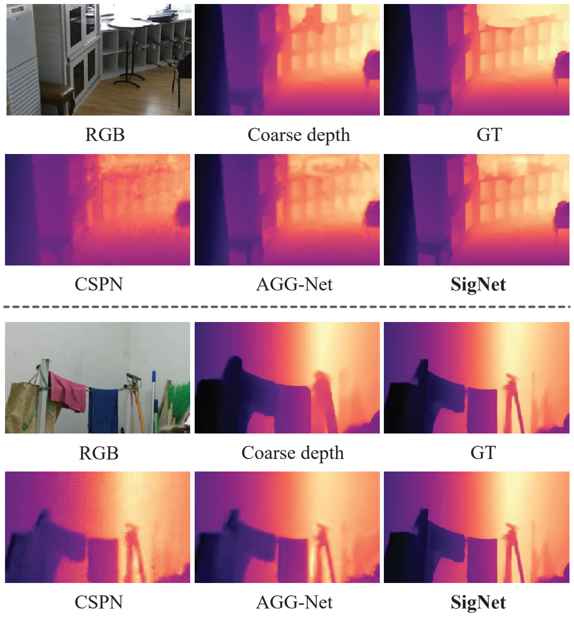

Figure 10. Visual comparisons with SOTAs on DIML.

7. More Visualizations

Fig. 10 shows additional visual results on the DIML dataset. As evident, our method reconstructs depth structures with better accuracy and more detailed information. For example, in the first row, the cabinet predicted by our SigNet appears clearer and sharper compared to the other methods, demonstrating the superior performance of our approach in preserving intricate depth details.