CylinderDepth: Cylindrical Spatial Attention for Multi-View Consistent Self-Supervised Surround Depth Estimation

Supplementary Material

1. Qualitative Results

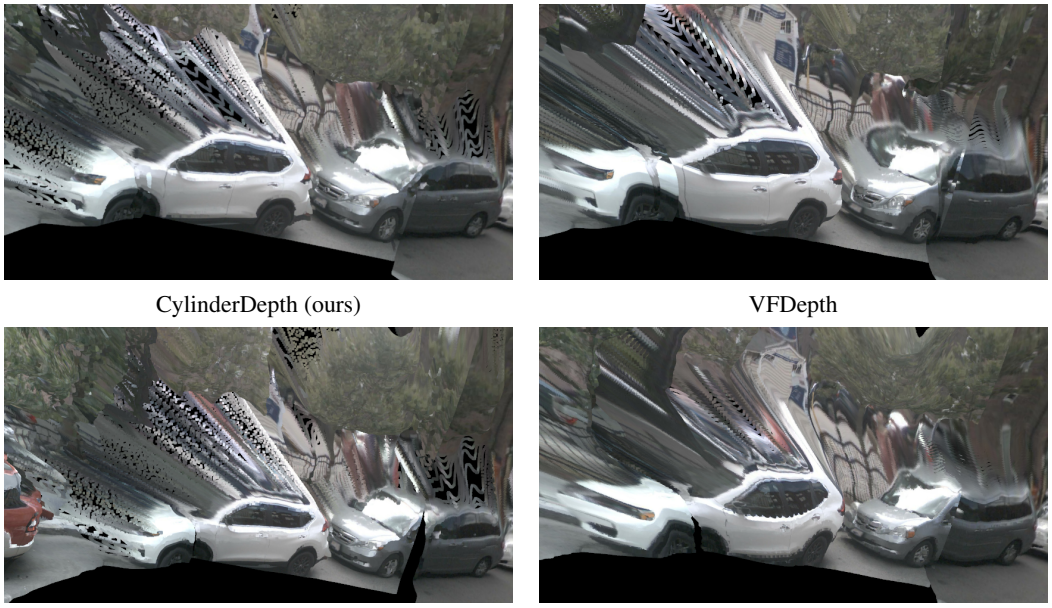


Figure 1. Exemplary 3D reconstructions, comparing our method to the state-of-the-art on nuScenes. It shows the reconstruction of the overlap regions of the front-right, back-right and the back camera.

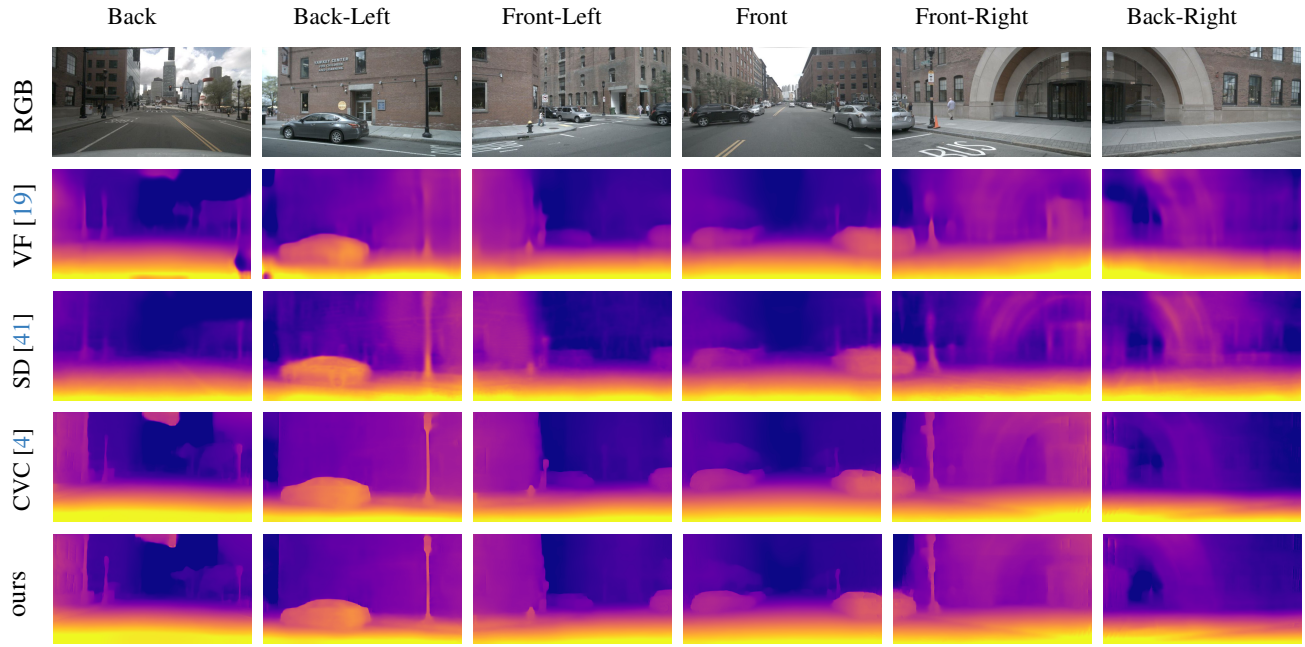
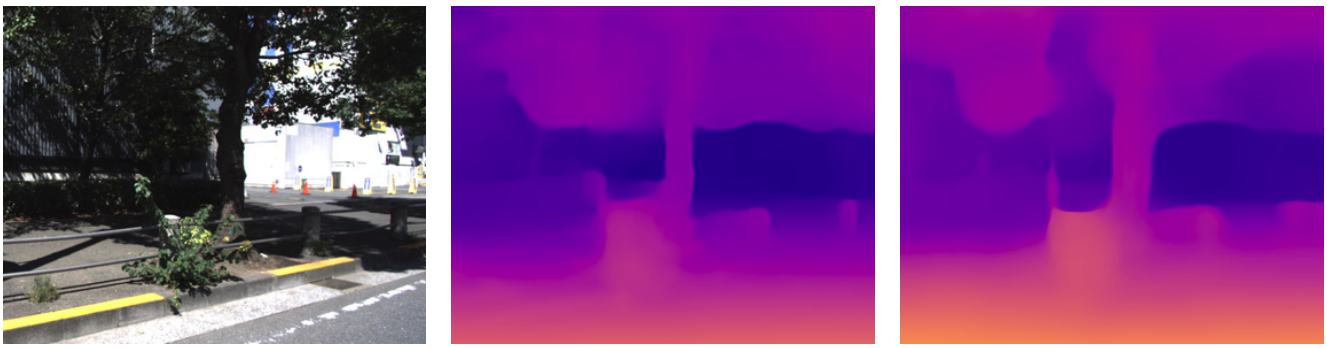


Figure 2. Comparison of depth maps predicted by our method and by state-of-the-art methods on nuScenes. Depth is shown from close in yellow to distant in blue.

2. Low-Resolution Spatial Attention



(a) RGB image (b) Only low resolution (c) All resolutions

Figure 3. Depth maps when applying attention only at the lowest resolution (b) versus at all resolutions after them being downsampled (c). Finer details are preserved when restricting attention only to low-resolution feature maps.

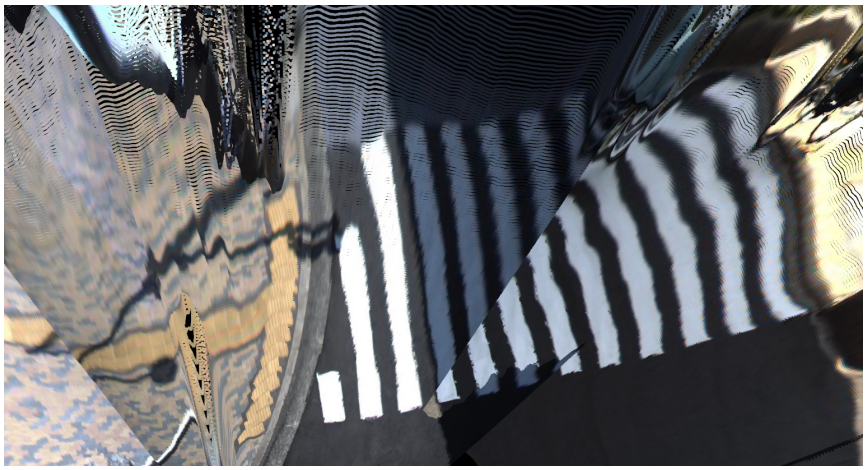


Figure 4. An exemplary limitation of our method: the approach struggles to enforce multi-view consistency at high resolution, as our attention mechanism is only applied on feature maps at low resolution.

3. Depth Consistency Evaluation Metric

For a pixel $\mathbf{p}_{\mathbf{I}_{t,i}}$ in depth map $\hat{\mathbf{D}}_{\mathbf{I}_{t,i}}$, we define its Euclidean distance map from a common reference coordinate system as:

$$\hat{\mathbf{D}}^{\text{Euclid}}_{\mathbf{I}_{t,i}}(\mathbf{p}_{\mathbf{I}_{t,i}}) = \left\| {}^{\text{ref}}\mathbf{T}_{\mathbf{I}_{t,i}} \left(\hat{\mathbf{D}}_{\mathbf{I}_{t,i}}(\mathbf{p}_{\mathbf{I}_{t,i}}) \, \mathbf{K}^{-1}_{\mathbf{I}_{t,i}} \mathbf{p}_{\mathbf{I}_{t,i}} \right) \right\|_2, \tag{1}$$

where $\mathbf{K}_{\mathbf{I}_{t,i}}$ is the intrinsic matrix, and ${}^{\text{ref}}\mathbf{T}_{\mathbf{I}_{t,i}}$ denoting the pose relative to the reference coordinate system.

The depth consistency between the depth predictions of images $\mathbf{I}_{t,i}$ and $\mathbf{I}_{t,j}$ is defined as the Root Mean Square Error (RMSE) between their predicted depth maps $\hat{\mathbf{D}}_{\mathbf{I}_{t,i}}$ and $\hat{\mathbf{D}}_{\mathbf{I}_{t,j}}$, expressed as the Euclidean distance from a common reference coordinate system, and is given as:

$$\text{DepthCons}(\hat{\mathbf{D}}_{\mathbf{I}_{t,i}}, \hat{\mathbf{D}}_{\mathbf{I}_{t,j}}) = \sqrt{\frac{1}{|\Omega_{\mathbf{I}_{t,i},\mathbf{I}_{t,j}}|} \sum_{\mathbf{p}_{\mathbf{I}_{t,i}} \in \Omega_{\mathbf{I}_{t,i},\mathbf{I}_{t,j}}} \left(\hat{\mathbf{D}}^{\text{Euclid}}_{\mathbf{I}_{t,i}}(\mathbf{p}_{\mathbf{I}_{t,i}}) - \hat{\mathbf{D}}^{\text{Euclid}}_{\mathbf{I}_{t,j}}(\pi_{\mathbf{I}_{t,i} \to \mathbf{I}_{t,j}}(\mathbf{p}_{\mathbf{I}_{t,i}})) \right)^2}, \tag{2}$$

where $\Omega_{\mathbf{I}_{t,i},\mathbf{I}_{t,j}}$ denotes the set of corresponding pixels in the overlapping region, and $\pi_{\mathbf{I}_{t,i} \to \mathbf{I}_{t,j}}(\mathbf{p}_{\mathbf{I}_{t,i}})$ is the corresponding pixel in image $\mathbf{I}_{t,j}$ for pixel $\mathbf{p}_{\mathbf{I}_{t,i}}$ in image $\mathbf{I}_{t,i}$. Correspondences are established using the ground-truth depth together with the camera intrinsics and the pose between the cameras.