Entity-NeRF: Detecting and Removing Moving Entities in Urban Scenes

Supplementary Material

7. MovieMap Dataset Details

To evaluate our approach, we introduce three urban scenes from the MovieMap [38]. The MovieMap Dataset was created by sampling images from 360° videos of varying lengths. Specifically, we extracted 51 images from a 3-second video, 12 images from a 6-second video, and 15 images from a 7-second video. Each image has a resolution of 3840×1920. For each 360° image, we extracted 14 perspective projection images. An example of this extraction process from a single 360° image is illustrated in Fig. 12. A common challenge when capturing 360° images is the inclusion of the photographer in the frame. To address this, we created a mask to exclude the photographer from the images, which is demonstrated in Fig. 13. This masked area was subsequently omitted from both the training and evaluation phases.

8. Additional Implementation Details

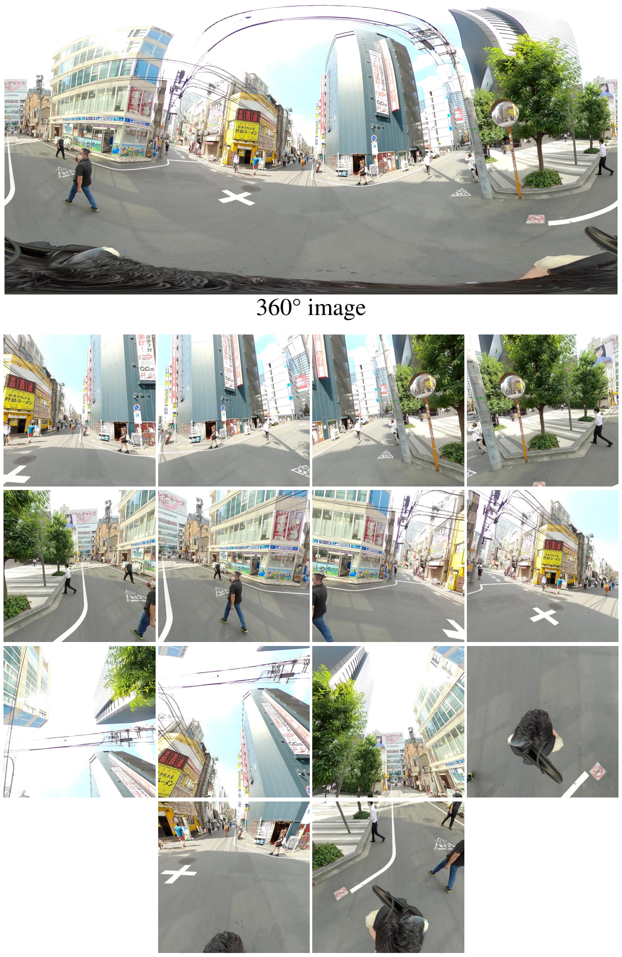
8.1. Rendered Background-only Images of MovieMap Dataset

When training static Neural Radiance Fields (NeRF) by removing masked objects, a significant challenge arises from errors near the edges of segmentation masks. These errors can disrupt the model's ability to accurately render static-only scenes, as they introduce inconsistencies at the boundaries of masked moving objects. To mitigate this, we dilated the mask area of moving objects. We implemented this by applying a convolution operation with a uniformly positive 3×3 kernel. Subsequently, in the output of this convolution, all positive values were converted to 1.

8.2. Robust Approaches for Entity Segmentation Errors

To prevent errors near the edges of entity segmentation, the area where the predicted entity-wise loss weights cover moving objects is increased through dilation. This is achieved by performing convolution with a uniform positive-valued 3×3 kernel and setting any positive values obtained in the result to 1.

Entity segmentation does not assign an entity to every pixel; some pixels are not assigned to any entity. Especially near the edges of objects, there are often pixels that were not assigned to any entity. We choose to include in training all pixels that are not classified as entities. However, we expect that the weight mask dilation process will exclude pixels near the edges of moving objects, which are not assigned to any entities, from the training process.



360° image

Perspective projection images

Figure 12. **Perspective projection images extracted from a single 360° image.**

9. More Results

9.1. Evaluation on RobustNeRF Dataset

Dataset details: Four natural scenes (i.e., Statue, Android, Crab, BabyYoda) from RobustNeRF [33]. Distractor objects are either moved or allowed to move between frames to simulate capture over extended periods. The number of unique distractors varies from 1 (Statue) to 150 (BabyYoda). Additional frames without distractors are provided to enable quantitative evaluation.

Note that we encountered issues with the provided camera parameters for the Statue and Android scenes, and the Crab scene does not provide camera parameters. Consequently, we calibrated the cameras using COLMAP [35]

Loss	Statue PSNR↑	 SSIM↑	 LPIPS↓	Android PSNR↑	 SSIM↑	 LPIPS↓	Crab PSNR↑	 SSIM↑	 LPIPS↓	BabyYoda PSNR↑	 SSIM↑	 LPIPS↓
Mean-squared error (MSE)	18.89	0.70	0.24	18.53	0.63	0.25	24.68	0.80	0.11	22.54	**0.73**	**0.28**
RobustNeRF [33]	21.14	**0.74**	0.19	19.47	0.65	0.21	30.32	0.83	**0.10**	25.16	0.69	0.33
Entity-NeRF (only EARR)	21.10	**0.74**	**0.18**	19.99	**0.69**	**0.20**	30.43	0.83	**0.10**	25.63	0.68	0.33
Entity-NeRF	**21.20**	0.73	0.19	**20.23**	0.67	0.21	**30.65**	**0.84**	0.11	**25.65**	0.68	0.33

Table 2. **Quantitative comparison with RobustNeRF [33] using Mip-NeRF 360 [2] on RobustNeRF Dataset.**

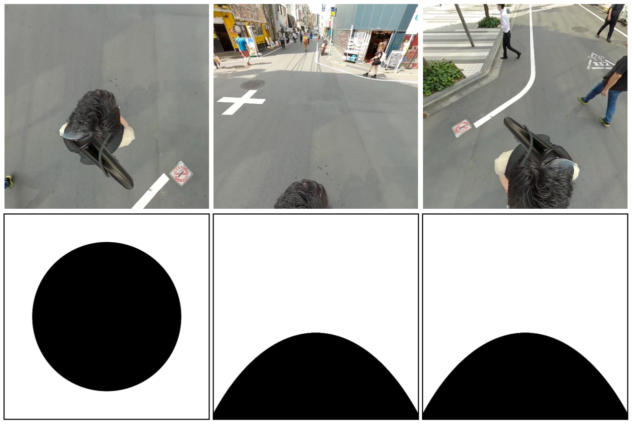


Figure 13. **Masks for photographers.**

for these scenes and used the calibrated parameters for training in the three scenes (Statue, Android, and Crab). The BabyYoda scene was trained using the original camera parameters.

Quantitative comparison: A quantitative evaluation using Mip-NeRF 360 [2] on the RobustNeRF natural scenes (Statue, Android, Crab, and BabyYoda), which were shot with objects centered, is shown in Table 2. Although our proposed method is not intended to improve the performance of scenes shot with the object centered, it showed that the proposed method outperformed RobustNeRF in PSNR, and was equal or better in terms of SSIM and LPIPS. Even when a moving object is photographed at a large size, the same problem as in the urban scene may occur because the object appears at the edge of the patch, and EARR appeared to have solved this problem. In addition, the incorporation of knowledge by the stationary entity classification was also found to be effective in the indoor scenes.

9.2. Qualitative Comparison of Distractiveness using stationary entity classification

As shown in Fig. 14, our thing/stuff segmentation-based stationary entity classification provides more precise $D(\mathbf{r})$ assignments for each entity than EARR in initial learning stages, where predicting accurate diffuse $D(\mathbf{r})$ for all entities is challenging due to large residuals.

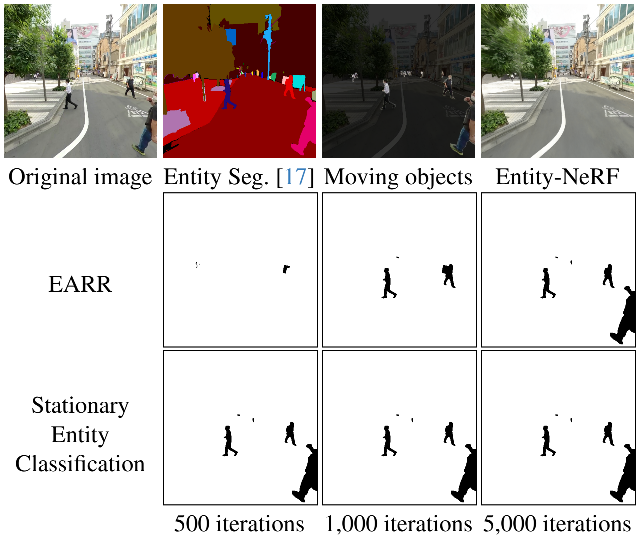


Figure 14. **Visualization of $D(\mathbf{r})$ using stationary entity classification.** Compared to EARR, $D(\mathbf{r})$ in the early stages of training are improved.

	IoU $D(\mathbf{r}) = 1 \uparrow$	IoU $D(\mathbf{r}) = 0 \uparrow$
RobustNeRF [33]	0.84	0.14
Entity-NeRF	**0.98**	**0.59**

Table 3. **Quantitative Comparison of Distractiveness with RobustNeRF [33] on MovieMap Dataset**

9.3. Quantitative Comparison of Distractiveness

In Table 3, the Intersection over Union (IoU) of Distractiveness $D(\mathbf{r})$ in Entity-NeRF at the end of training is compared with the IoU of Distractiveness $D(\mathbf{r})$ in RobustNeRF [33], using masks annotated on moving objects as ground-truth labels. Our proposed method achieves a better IoU for both $D(\mathbf{r}) = 0$ and $D(\mathbf{r}) = 1$, allowing for closer Distractiveness to the annotated mask to be given as a weight in the loss.

9.4. Novel View Synthesis

We conduct a qualitative comparison of our Entity-NeRF's performance in novel view synthesis. The novel view synthesis using the MovieMap Dataset is shown in Fig. 15. We created a circular trajectory around the straight-line path of

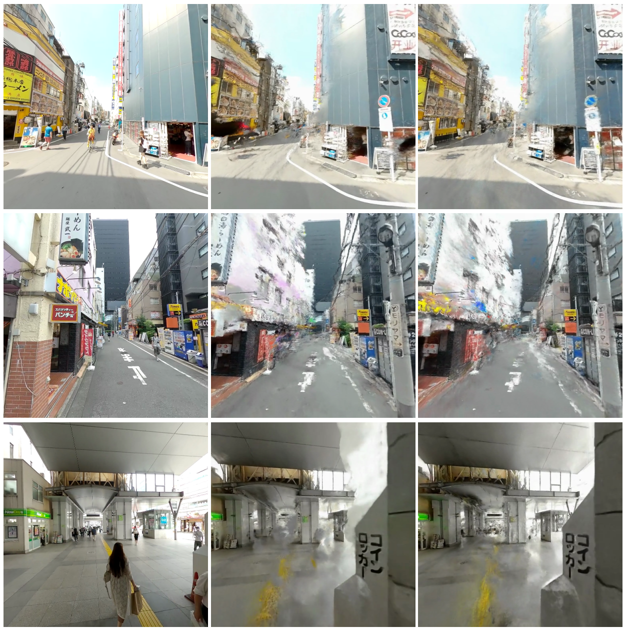


Figure 15. **Novel view synthesis.**

the original video and synthesized new views on the circular path. Entity-NeRF, although suffering from degradation due to the inability to learn correct geometry, shows less deterioration compared to RobustNeRF [33]. This is evident from the comparison of synthesized images from different viewpoints during training, as our approach avoids erroneously including moving objects in the training process and includes more static backgrounds into the training.