Supplementary Material for "ActiveZero: Mixed Domain Learning for Active Stereovision with Zero Annotation"

1. Additional Ablation Study

1.1. Effect of Simulation Ground-truth

In this section, we study the effect of using the supervised simulation disparity loss $\mathcal{L}_{disp}$ during training. To do so, we conduct experiments with and without $\mathcal{L}_{disp}$ added to the final loss term and observe their convergence rate as well as final converged solution. Figure 1 shows that adding simulation disparity loss (blue) helps the network converge faster to the global optima.

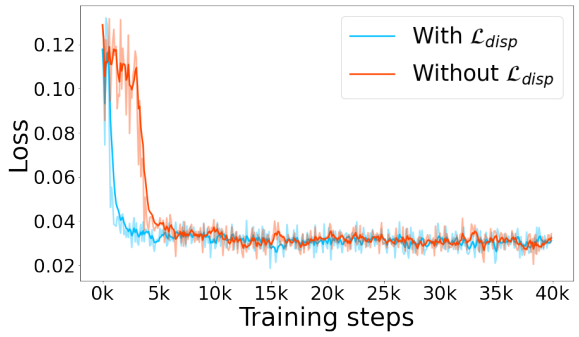


Figure 1. Loss curve of training with and without simulation disparity ground-truth

Method	Abs depth err (mm) ↓	> 4mm ↓
w/o sim disp loss	4.729	0.367
sim disp loss	**4.377**	**0.335**

Table 1. Performance with and without auxiliary simulation supervision

1.2. Patch Size of Reprojection Loss

In this section, we conduct an ablation study on the patch size of the patch-wise reprojection loss. In the main paper, we chose a patch size of 11. For this study, we change patch size to 7, 15 and 21, train each one with only the real reprojection loss term, and evaluate them on the same testing dataset. Table 2 suggests patch size 15 has the best result on the absolute depth error (*abs depth err*) metric while patch size 21 has the lowest percentage of depth outliers with absolute depth error larger than 4mm (*>4mm*). However, the loss curve in Fig. 2 indicates that patch size 11 converges faster than the other patch sizes. Considering patch size 11 also occupies less GPU memory during training, we choose patch size 11 in our main experiments.

Patch size	Abs depth err (mm) ↓	> 4mm ↓
7	5.507	0.466
11	5.115	0.393
15	**5.114**	0.386
21	5.402	**0.385**

Table 2. Performance of different patch size

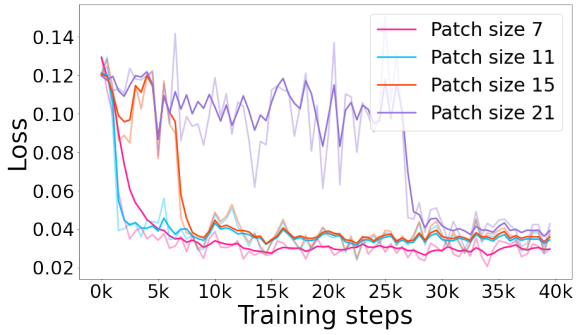


Figure 2. Loss curve of training using different patch sizes

1.3. Loss Ratio between Simulation and Real Domain

In this section, we conduct an ablation study on the loss weight λ_s and λ_r described in Sec. 3.3 of the main paper. In our main experiment, we use $\lambda_s = 0.01$ and $\lambda_r = 2$. We change λ_s and λ_t to different values and test the trained models on the testing dataset. The results in Tab. 3 indicate that when $\lambda_s = 0.01$ and $\lambda_r = 2$, the network achieves the best result, which is consistent with our experiment setting.

λ_s	λ_r	Abs depth err (mm) ↓	$>$ 4mm ↓
1	0.5	7.578	0.548
1	1	6.064	0.455
1	2	5.672	0.446
0.05	2	5.543	0.433
0.01	2	**4.377**	**0.335**
0.002	2	4.683	0.368

Table 3. Performance of different loss weight

1.4. Choice of backbones

In this section, we study different backbones in our proposed pipeline, which includes DispNet (2016), RAFT (2020), as well as PSMNet (2018). The results are shown in Tab. 4. Compared to their original performance, our framework greatly improved each backbone's depth estimation performance.

Backbone	Direct transfer	LCN re-proj	Temporal IR re-proj
DispNet	82.906	43.781	**18.069**
PSMnet	16.854	10.598	**4.377**
Raft	6.521	5.890	**4.738**

Table 4. Absolute Depth Error (mm) of different backbones

2. Time Budget and Inference Time

Temporal IR reprojection is only used during training on real images. Collection is done offline and takes ∼0.7s to capture one frame. During testing, the temporal IR image sequences are not required. We measure the inference time of our proposed pipeline in Tab. 5. Our method has an average inference time of 0.25 seconds per image pair with a resolution of 960×540. Compared to StereoGAN with PSMNet backbone, our method achieves faster inference times while also having better performance. We will continue to reduce our inference time in future studies.

Method	Inference Time(s) ↓
StereoGAN+PSM	0.303
Our Method	0.256

Table 5. Inference time of StereoGAN+PSM and our method

3. Pointcloud visualization of the estimated depth

We provide visualizations from *novel views* for RealSense measurements and our depth prediction on the same test scene below. As shown, our prediction contains less noise and is more complete in the highlighted transparent area.

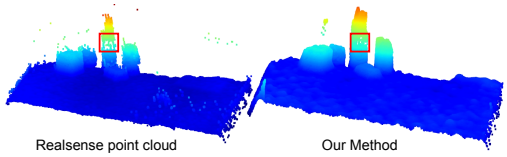


Figure 3. Depth estimation (point cloud) at a novel view

4. Performance On Different Dataset

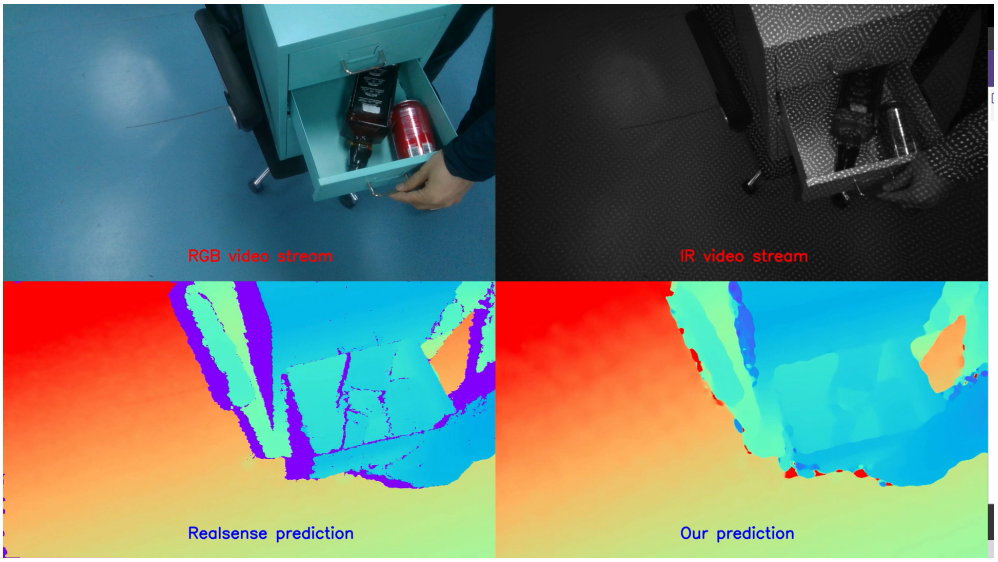
In this section, we test our trained model on scene structures and moving objects different from our training data. Our model can make accurate and complete depth predictions on *moving objects* and *people*, demonstrating the generalizablity of the network from our method. Results are shown in Fig. 4

5. More Details of Datasets

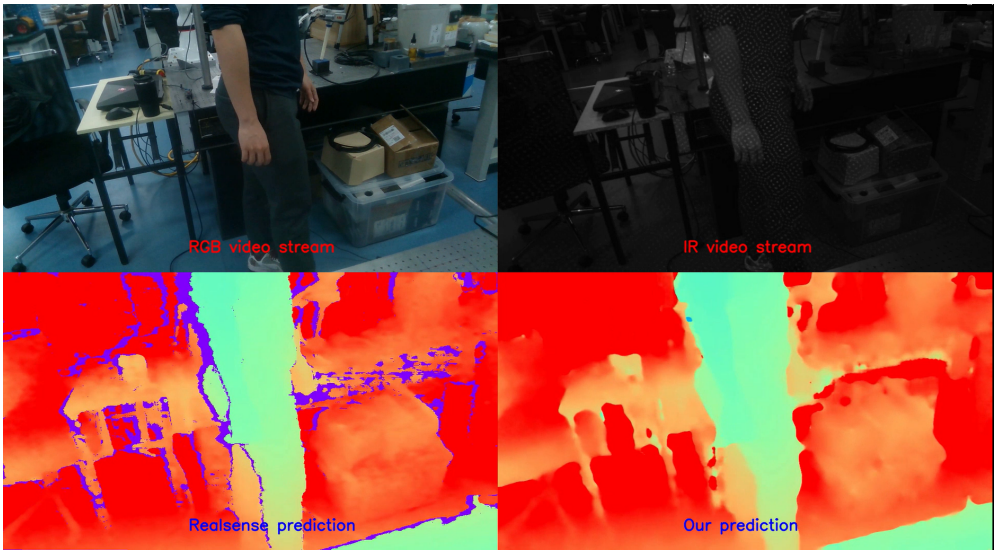
The training simulation dataset has 18000 image pairs with random camera extrinsics, shape primitives, textures and poses. As in Fig. 5 (a), in order to make the scene more complicated, the primitives can overlap with each other and are not strictly attached to the table. Therefore, they can either overlap with the table or float above the table. In Fig. 5 (a), the textures are randomly selected to improve generalizability. For IR images in Fig. 5 (a), the simulated IR pattern is projected onto each scene of the simulation dataset.

Samples of the training real dataset are shown in Fig. 5 (b). The objects in the training dataset are not present in the testing dataset and the ground truth depths are not required for this dataset. To preserve its generalizability, the optical properties of the objects are diversely selected. In Fig. 5 (b), there exists objects that are transparent (glass bottle), specular (the cover of the glass bottle) and diffused (black paper box). These objects have different abilities to reflect IR pattern as seen in Fig. 5 (b). Temporal IR images are collected by adjusting the power of the pattern emitter. There are 6 images with increasing IR power in each scene.

The testing dataset contains objects that are never used in training to best represent the generalizability of our method. As shown in Fig. 5 (c) and (d), the object properties are also diversely selected. For example, this dataset contains specular objects (metal ball), transparent objects (bottled water) and diffused objects (printed cell phone). The IR pattern is collected by adjusting the IR emitter to the max power used in the training dataset. To obtain accurate ground truth, we align the scene using the same object poses and camera parameters in simulation, as shown in Fig. 5 (c) and (d).



(a) Results and comparison of our method and realsense on dynamic scenes



(b) Results and comparison of our method and realsense scenes with moving objects and people

Figure 4

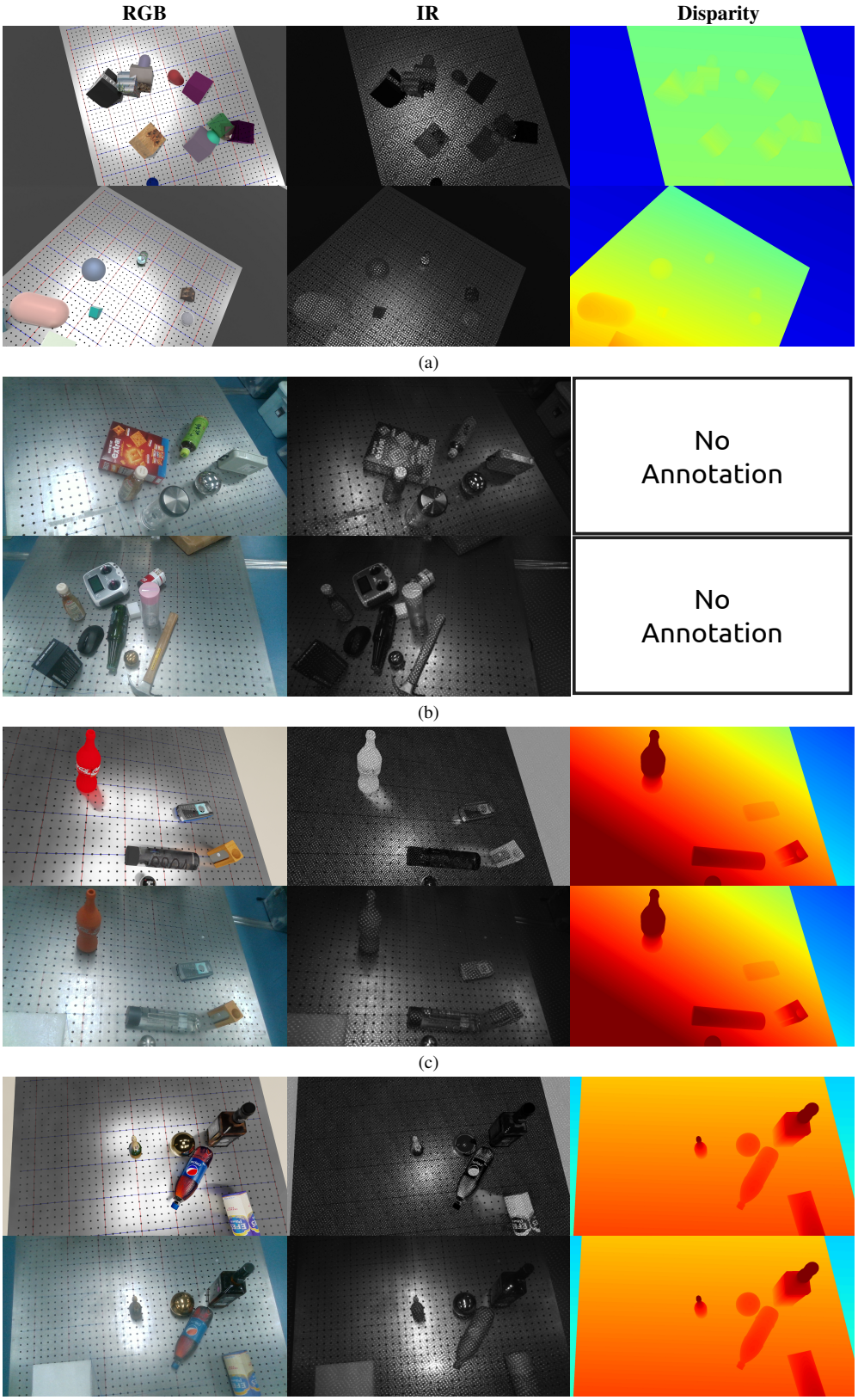


Figure 5. More examples of our datasets. (a) is training simulation dataset, (b) is training real dataset, (c) and (d) are testing simulation and real pixel-wise aligned pairs.