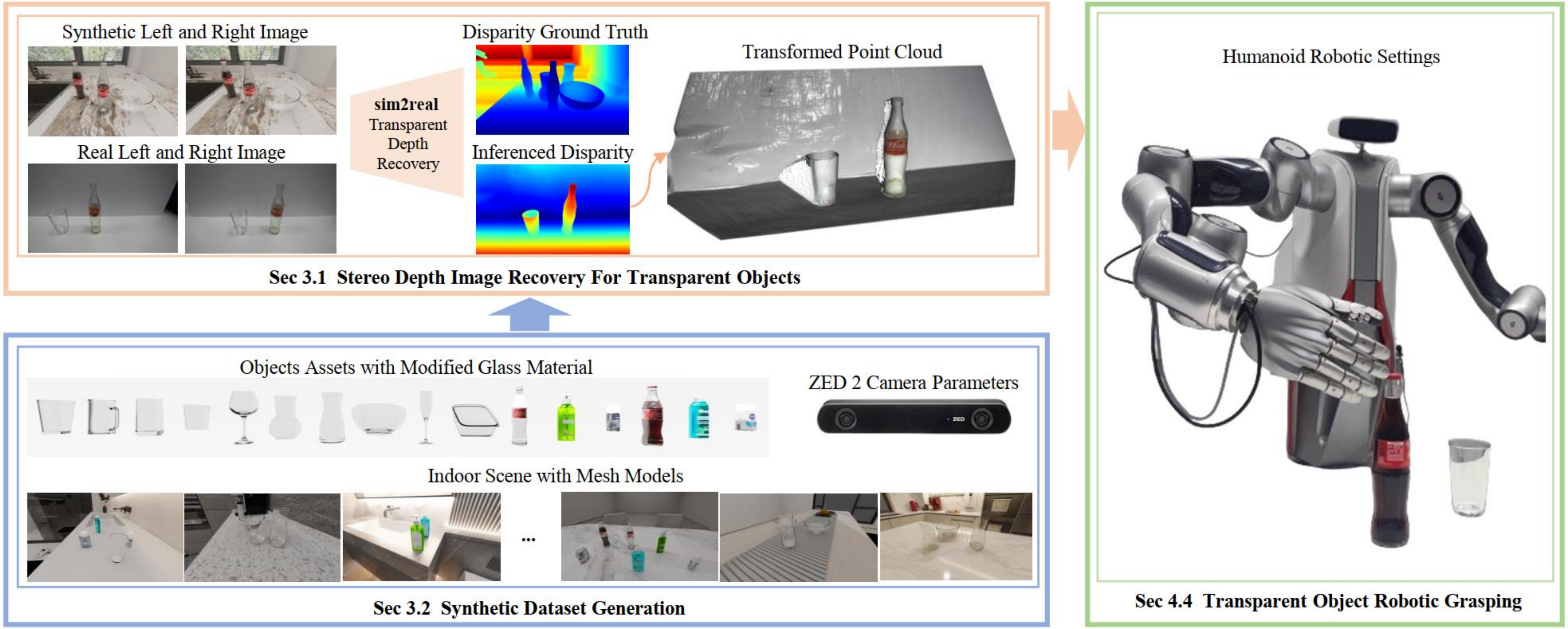
ClearDepth: Enhanced Stereo Perception of Transparent Objects for Robotic Vision

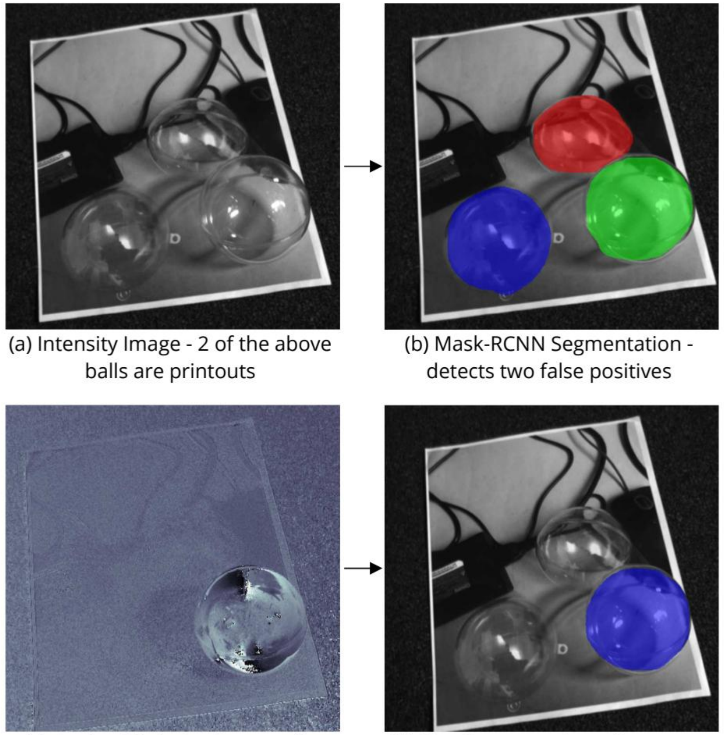
Kaixin Bai

23.04.2024

Introduction


Synthetic Left and Right Image
Real Left and Right Image
sim2real
Transparent
Depth
Recovery
Disparity Ground Truth
Inferenced Disparity
Transformed Point Cloud
Sec 3.1 Stereo Depth Image Recovery For Transparent Objects
Objects Assets with Modified Glass Material
ZED 2 Camera Parameters
Indoor Scene with Mesh Models
...
Sec 3.2 Synthetic Dataset Generation
Humanoid Robotic Settings
Sec 4.4 Transparent Object Robotic Grasping

Related Work – Perception of Transparent Objects



(a) Intensity Image - 2 of the above balls are printouts

(b) Mask-RCNN Segmentation - detects two false positives

(c) Angle of Polarization - easily seperates real ball from printout

(d) Our Segmentation - detects no false positives

[2020 CVPR] Deep polarization cues for transparent object segmentation

Figure 1: **Our reconstruction results and their transparent renderings**

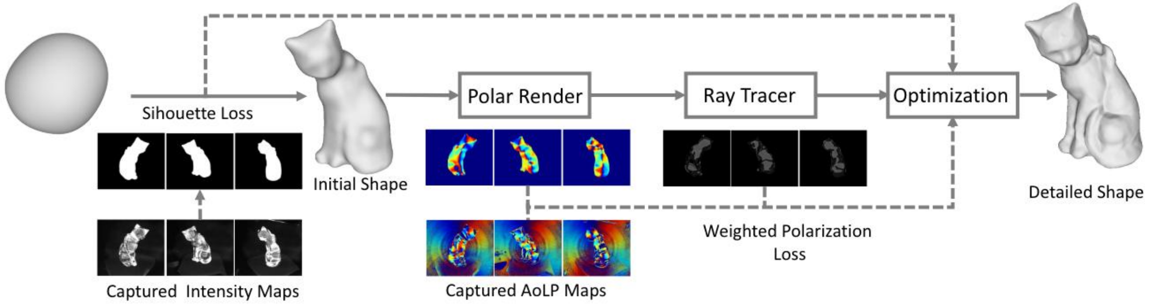


Figure 2: **Overview of our method.** The silhouette loss is used as supervision of initial shape reconstructing, then the polarimetric render and ray tracer calculate the weighted polarization loss for detailed shape optimization

[2022] Polarimetric Inverse Rendering for Transparent Shapes Reconstruction

Related Work – Perception of Transparent Objects

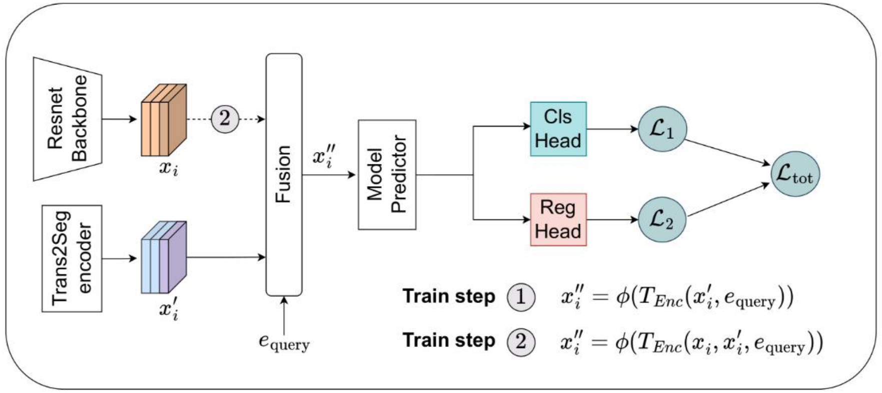

Resnet Backbone
x_i
2
Trans2Seg encoder
x_i'
Fusion
e_{query}
x_i''
Model Predictor
Cls Head
Reg Head
$\mathcal{L}_1$
$\mathcal{L}_2$
$\mathcal{L}_{\text{tot}}$
Train step 1 $x_i'' = \phi(T_{Enc}(x_i', e_{\text{query}}))$
Train step 2 $x_i'' = \phi(T_{Enc}(x_i, x_i', e_{\text{query}}))$

Fig. 3: The two step process of training the fusion module

[2023] Transparent Object Tracking with Enhanced Fusion Module

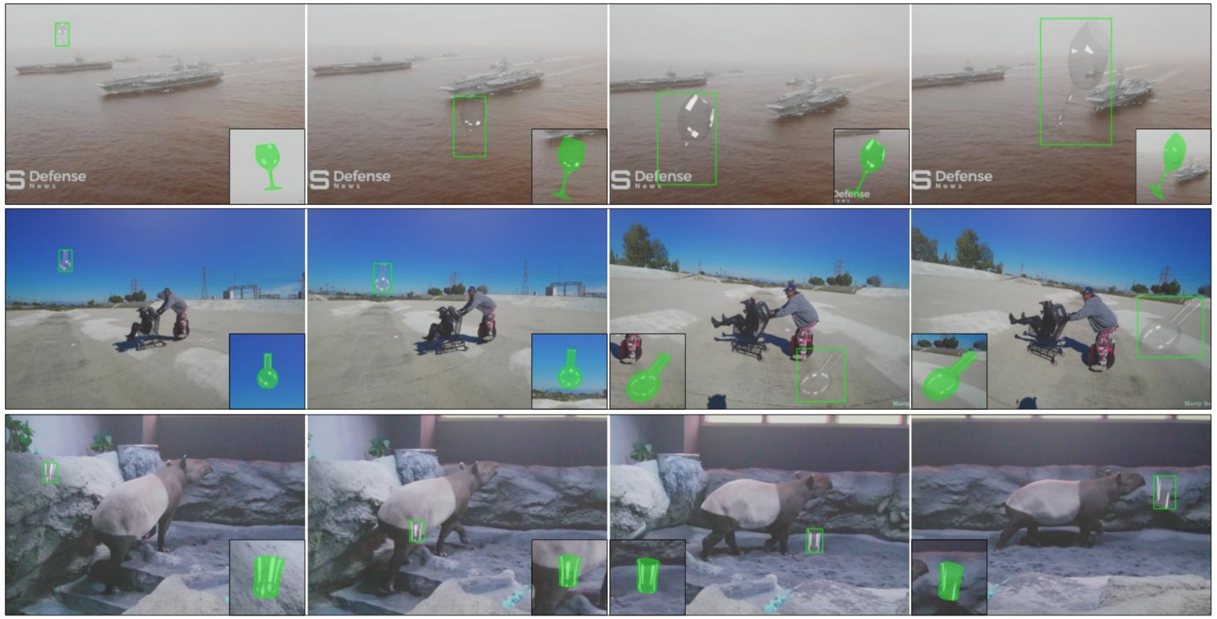


Figure 7: Frames from three sequences of the proposed Trans2k dataset. Objects are labeled by bounding boxes and segmentation masks (cropped in square).

[2022 BMVA "best paper"] Trans2k: Unlocking the Power of Deep Models for Transparent Object Tracking

Related Work – Perception of Transparent Objects

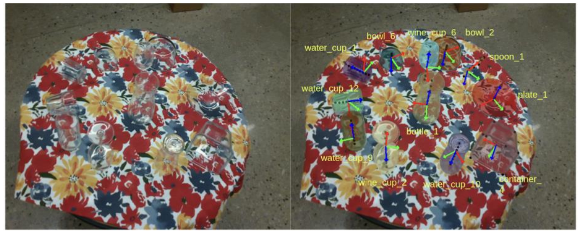


Fig. 4. Object names and labeled keypoints in set 3 of 5.

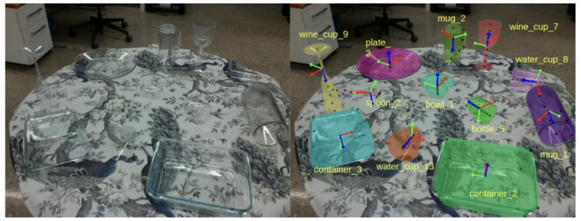


Fig. 5. Object names and labeled keypoints in set 4 of 5.

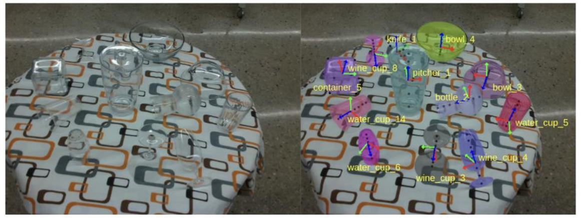


Fig. 6. Object names and labeled keypoints in set 5 of 5.

[2022 ECCV] ClearPose: Large-scale Transparent Object Dataset and Benchmark

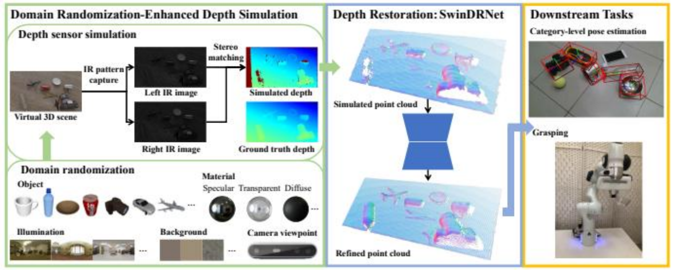


Fig. 1. Framework overview. From the left to right: we leverage domain randomization-enhanced depth simulation to generate paired data, on which we can train our depth restoration network SwinDRNet, and the restored depths will be fed to downstream tasks and improve estimating category-level pose and grasping for specular and transparent objects.

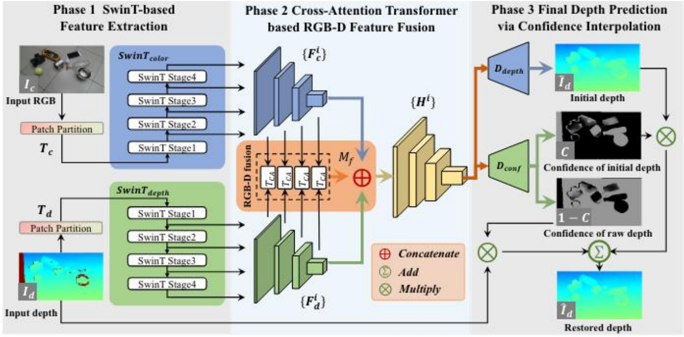


Fig. 4. Overview of our proposed depth restoration network SwinDRNet. We first extract the multi-scale features of RGB and depth image in phase 1, respectively. Next, in phase 2, our network fuse features of different modalities. Finally, we generate the initial depth map and confidence maps via two decoders, respectively, and fuse the raw depth and initial depth using the predicted confidence map.

[2022 ECCV] Domain Randomization-Enhanced Depth Simulation and Restoration for Perceiving and Grasping Specular and Transparent Objects

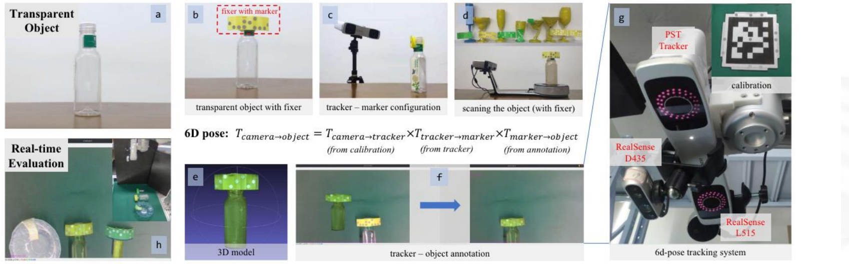

Transparent Object
a
b
fixer with marker
transparent object with fixer
c
tracker – marker configuration
d
scaning the object (with fixer)
g
PST Tracker
calibration
RealSense D435
RealSense L515
6d-pose tracking system
Real-time Evaluation
h
6D pose: $T_{camera\to object} = T_{camera\to tracker} \times T_{tracker\to marker} \times T_{marker\to object}$
(from calibration)
(from tracker)
(from annotation)
e
3D model
f
tracker – object annotation

Fig. 3. System setup process. Given a transparent object (a), we attach a fixer with IR markers to it (b) and record its pattern with an optical tracker (c), which can enable tracking afterwards. Then we scan the object (d) and get its 3D model (e). After that, we manually perform tracker-object annotation to get the transformation matrix from marker to object (f), where an GUI is developed for real-time evaluation (h). The whole annotation and tracking process is assisted by our 6D pose tracking system (g), which consists of a PST tracker, an Intel RealSense D435 camera and an Intel RealSense L515 camera.

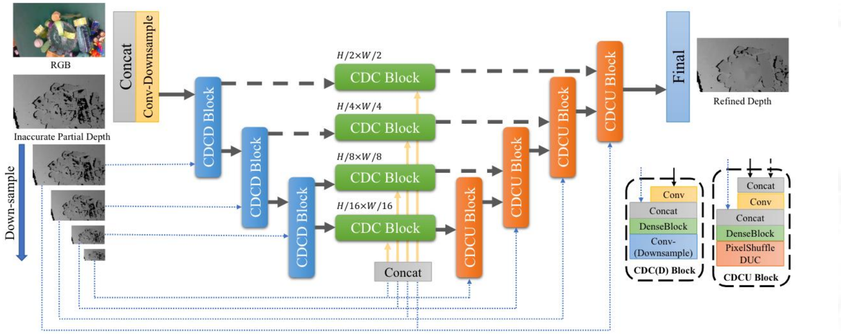


Fig. 4. The architecture of our proposed end-to-end depth completion network DFNet. Our network utilizes a U-Net architecture with CDCD blocks, CDC blocks and CDCU blocks. These blocks are mainly composed of dense blocks [12], with DUC [36] replacing deconvolution layer in up-sampling of CDCU block. All convolution layers except the last one are followed by batch normalizations [14] and ReLU activations, and have 3 × 3 kernels.

[2022 RAL] TransCG: A Large-Scale Real-World Dataset for Transparent Object Depth Completion and a Grasping Baseline

Related Work – Perception of Transparent Objects

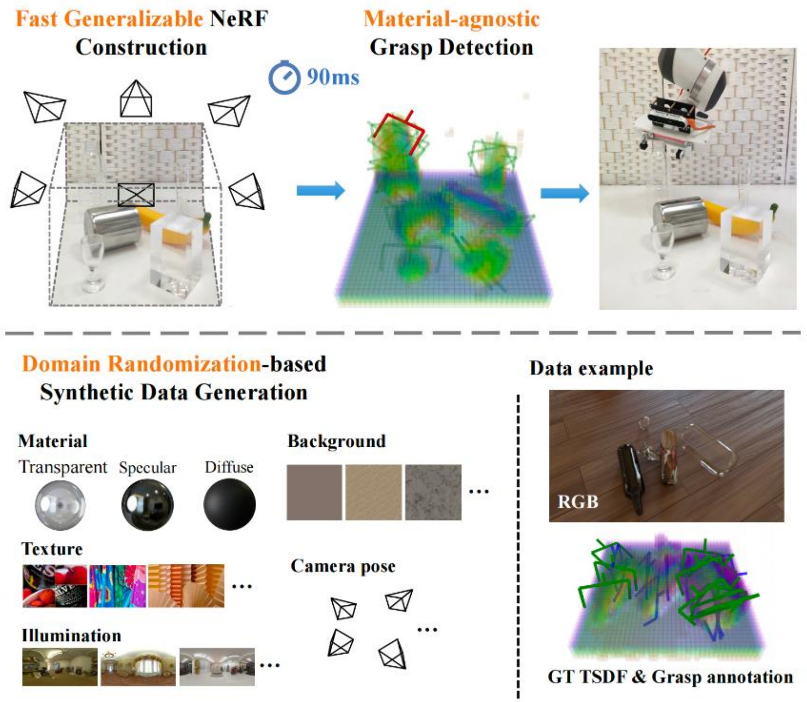

Fast Generalizable NeRF Construction
90ms
Material-agnostic Grasp Detection
Domain Randomization-based Synthetic Data Generation
Material
Transparent
Specular
Diffuse
Background
Texture
Camera pose
Illumination
Data example
RGB
GT TSDF & Grasp annotation

Figure 1. Overview of the proposed GraspNeRF and the dataset. Our method takes sparse multiview RGB images as input, constructs a neural radiance field, and executes material-agnostic grasp detection within 90ms. We train the model on the proposed large-scale synthetic multiview grasping dataset generated by photorealistic rendering and domain randomization.

[2023 ICRA] GraspNeRF: Multiview-based 6-DoF Grasp Detection for Transparent and Specular Objects Using Generalizable NeRF

Related Work – Deep Learning-based Stereo Depth Recovery

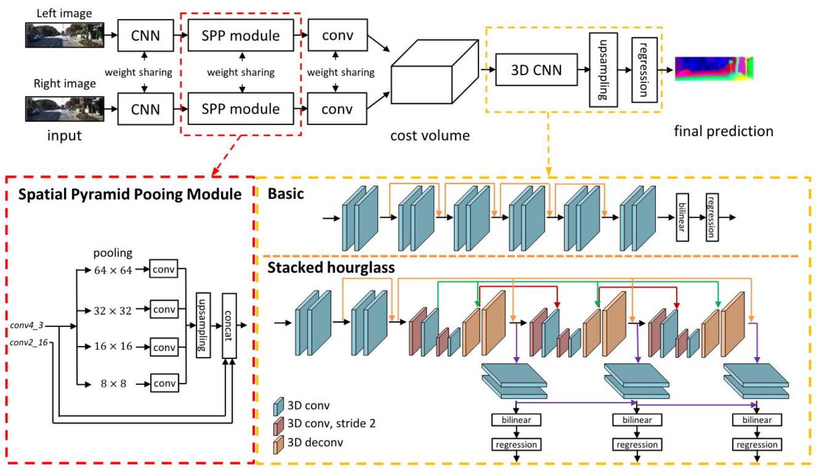
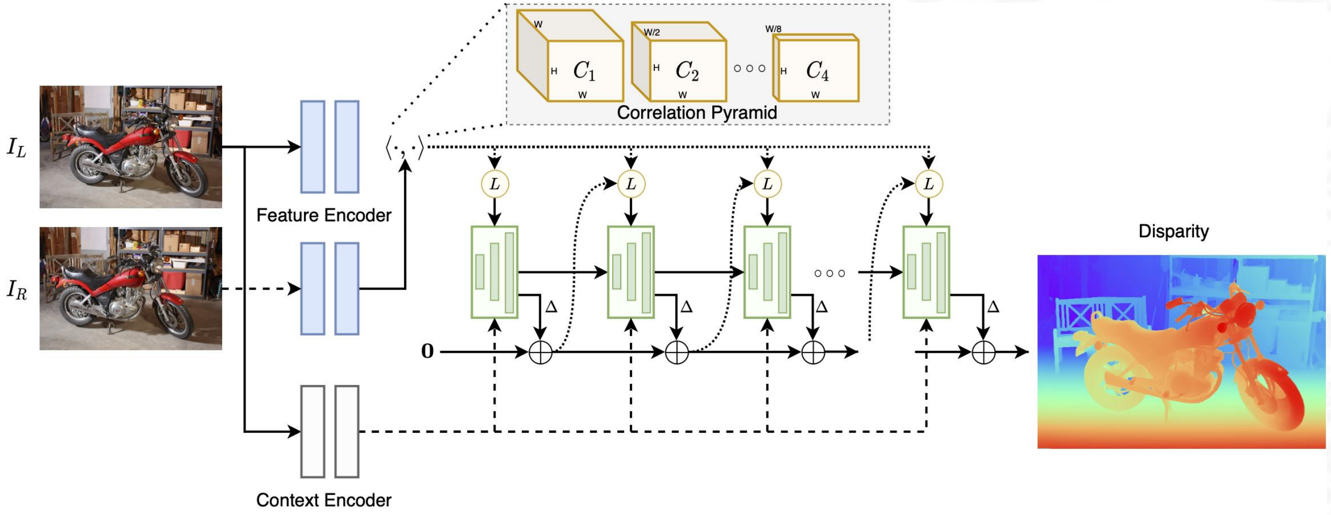


Figure 1. Architecture overview of proposed PSMNet. The left and right input stereo images are fed to two weight-sharing pipelines consisting of a CNN for feature maps calculation, an SPP module for feature harvesting by concatenating representations from sub-regions with different sizes, and a convolution layer for feature fusion. The left and right image features are then used to form a 4D cost volume, which is fed into a 3D CNN for cost volume regularization and disparity regression.

[2018 CVPR] Pyramid Stereo Matching Network



[2021 3DV] RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching

Related Work – Transparent Object Datasets

3 Toronto Transparent Objects Depth Dataset (TODD)

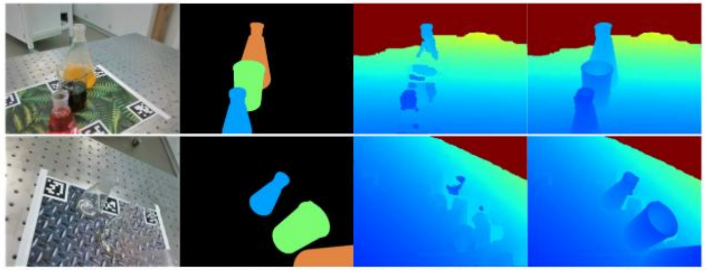
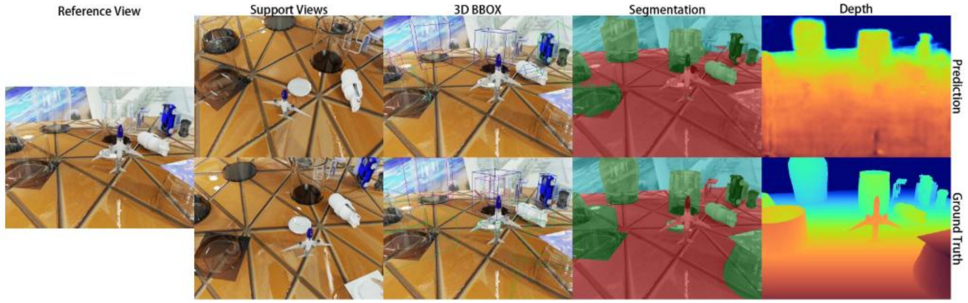


Figure 4: Samples from the proposed dataset. (a) RBG image (b) Instance Segmentation (c) Raw depth from RGB-D sensor (d) Ground truth depth obtained through automatic depth annotation. The dataset also includes 6DoF object pose information, which is not depicted in the figure.

TODD



MVTrans:Syn-TODD

Real robot demo

Check out our source codes!

Many objects

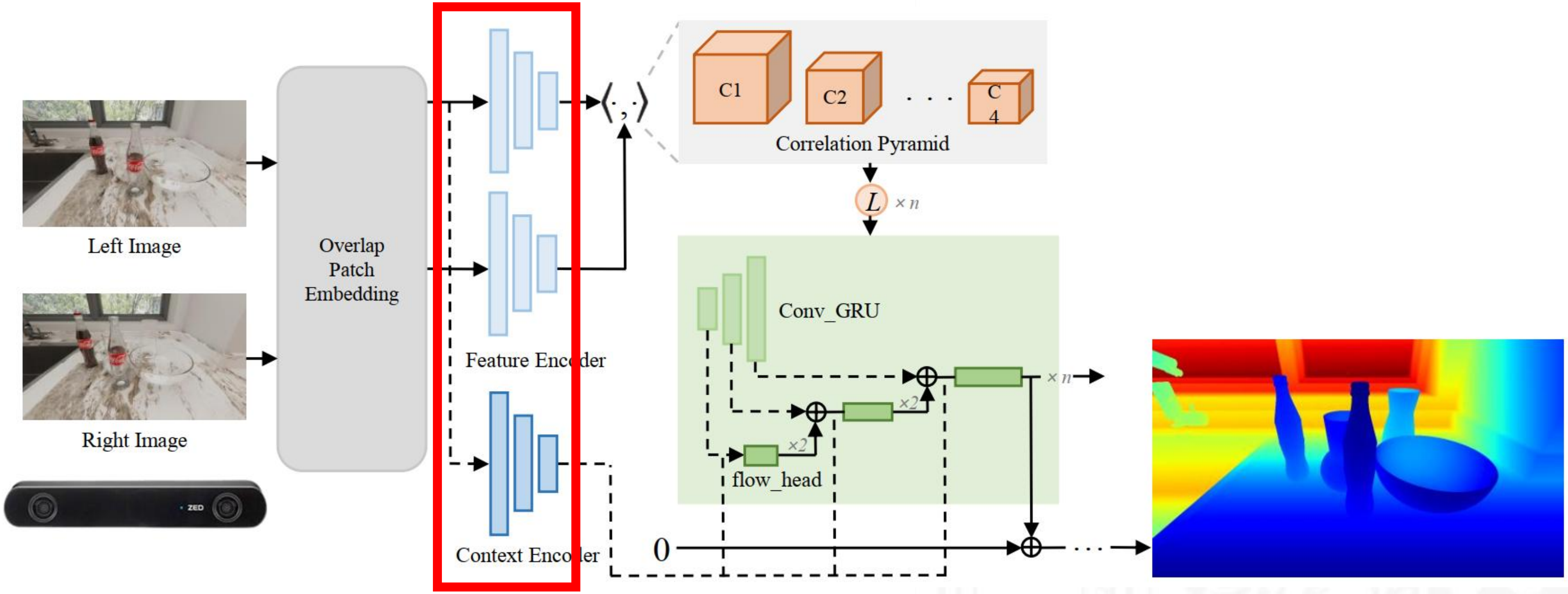
Robust to large variance in scenes

Various scenes

Closer to real-world environments

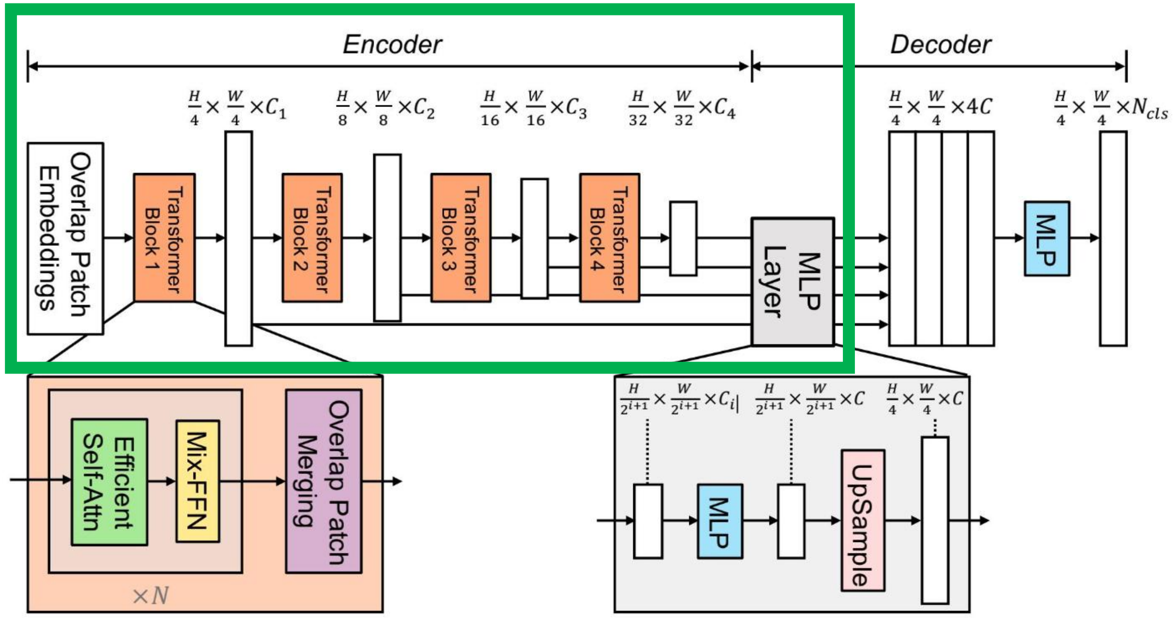
TransCG

Method – Network Overview

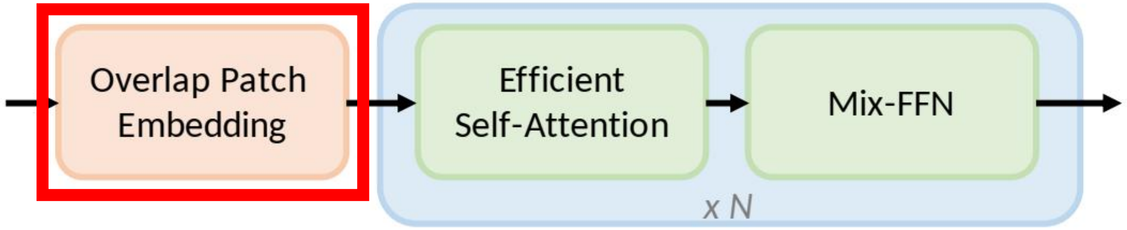


shared weights

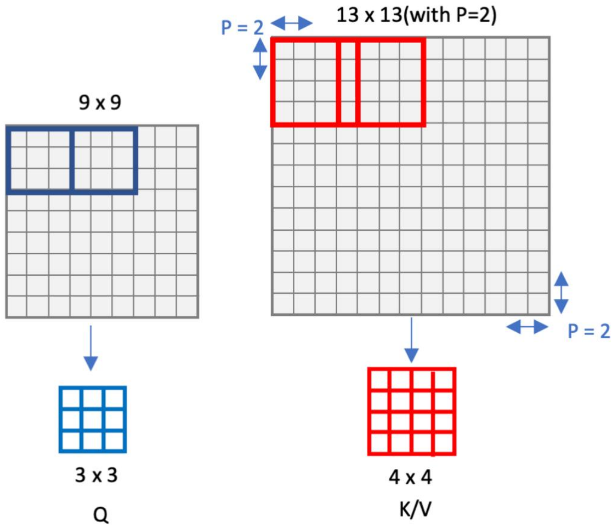
Method – Network Overview


Encoder
Decoder
$\frac{H}{4}\times\frac{W}{4}\times C_1$
$\frac{H}{8}\times\frac{W}{8}\times C_2$
$\frac{H}{16}\times\frac{W}{16}\times C_3$
$\frac{H}{32}\times\frac{W}{32}\times C_4$
$\frac{H}{4}\times\frac{W}{4}\times 4C$
$\frac{H}{4}\times\frac{W}{4}\times N_{cls}$
Overlap Patch Embeddings
Transformer Block 1
Transformer Block 2
Transformer Block 3
Transformer Block 4
MLP Layer
MLP
Efficient Self-Attn
Mix-FFN
Overlap Patch Merging
×N
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i$
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C$
$\frac{H}{4}\times\frac{W}{4}\times C$
MLP
UpSample

SegFormer with mixvisiontransformer as encoder

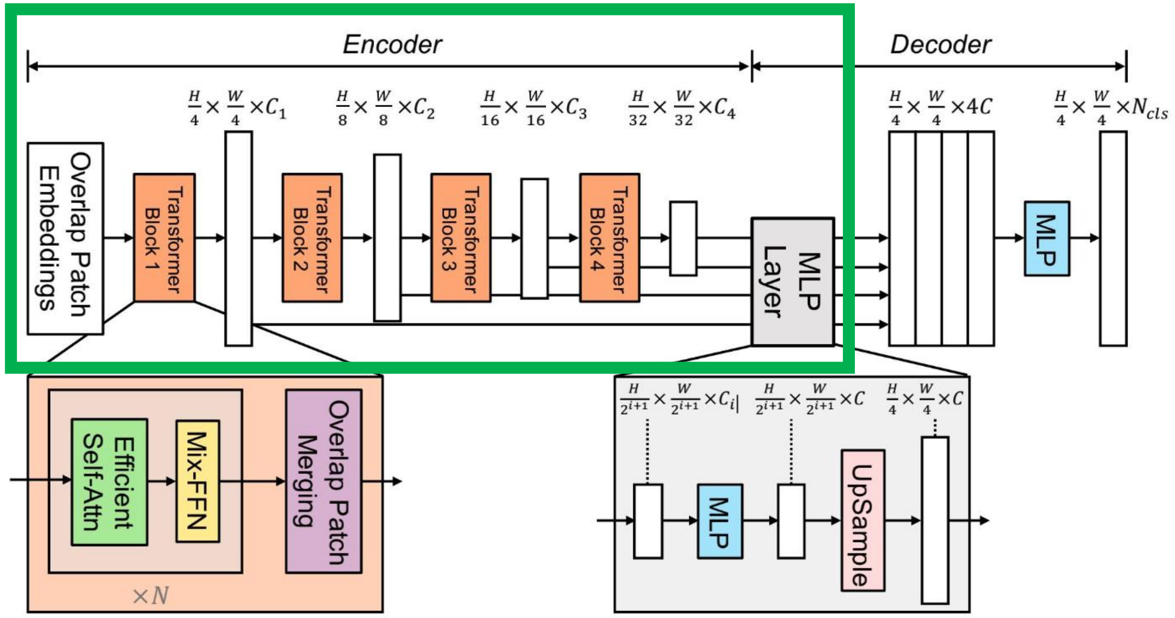


mixvisiontransformer encoder

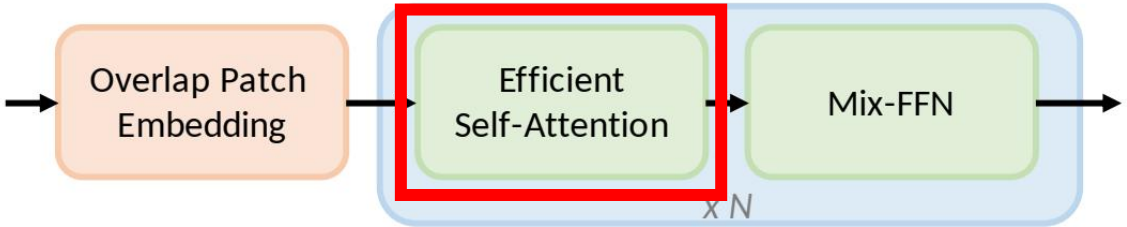


overlap patch embedding

Method – Network Overview


Encoder
Decoder
$\frac{H}{4}\times\frac{W}{4}\times C_1$
$\frac{H}{8}\times\frac{W}{8}\times C_2$
$\frac{H}{16}\times\frac{W}{16}\times C_3$
$\frac{H}{32}\times\frac{W}{32}\times C_4$
$\frac{H}{4}\times\frac{W}{4}\times 4C$
$\frac{H}{4}\times\frac{W}{4}\times N_{cls}$
Overlap Patch Embeddings
Transformer Block 1
Transformer Block 2
Transformer Block 3
Transformer Block 4
MLP Layer
MLP
Efficient Self-Attn
Mix-FFN
Overlap Patch Merging
×N
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i$
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C$
$\frac{H}{4}\times\frac{W}{4}\times C$
MLP
UpSample

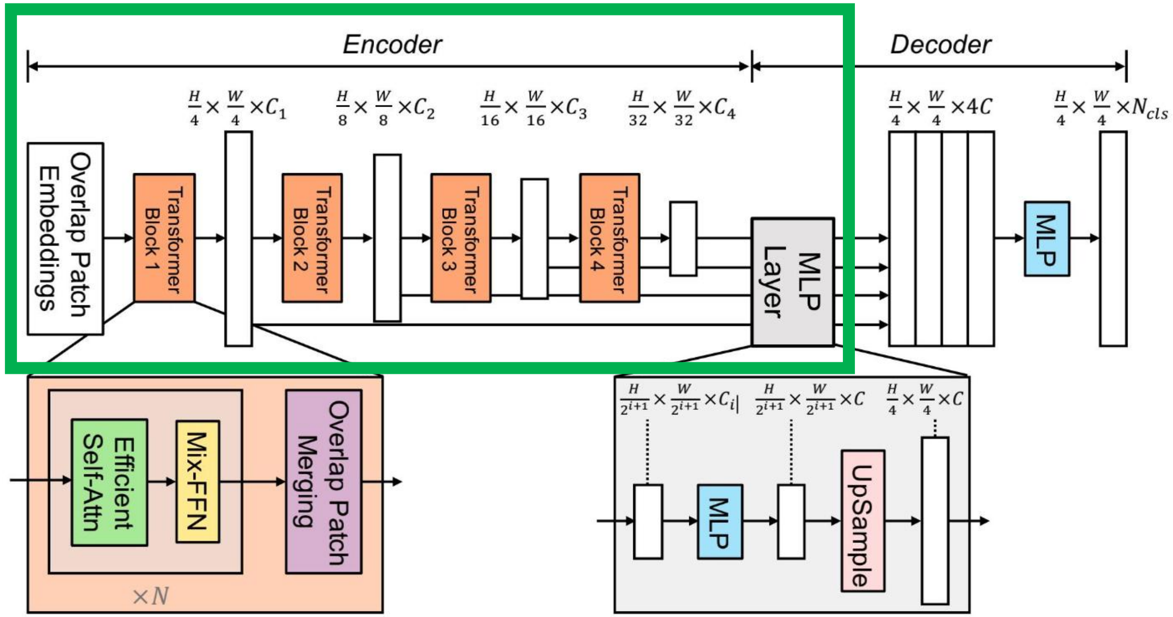
SegFormer with mixvisiontransformer as encoder



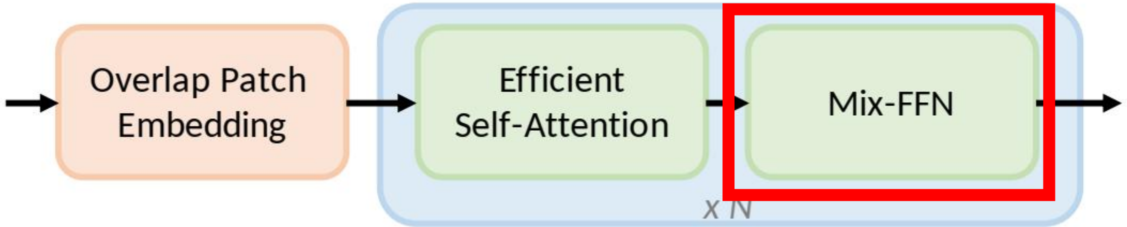
$$\hat{K} = Reshape(\frac{N}{R}, C \cdot R)(K)$$

$$K = Linear(C \cdot R, C)(\hat{K})$$

Method – Network Overview

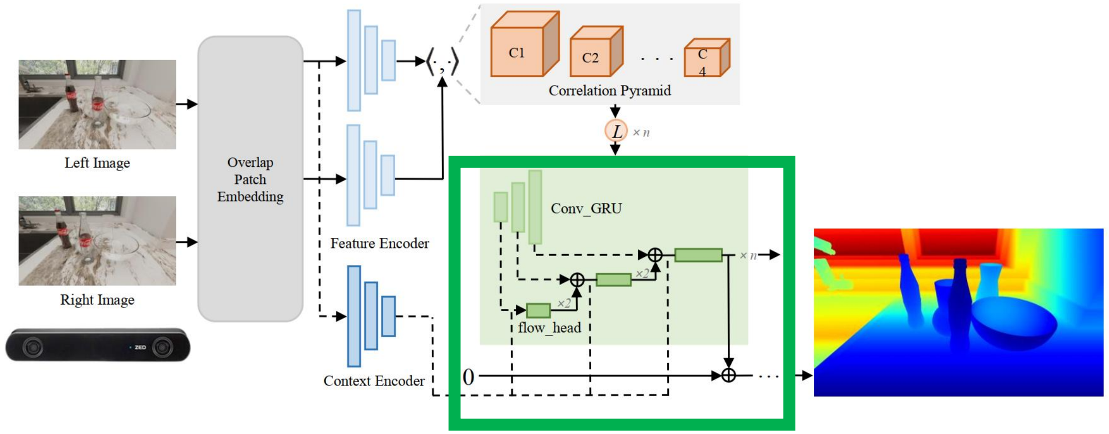

Encoder
Decoder
$\frac{H}{4}\times\frac{W}{4}\times C_1$
$\frac{H}{8}\times\frac{W}{8}\times C_2$
$\frac{H}{16}\times\frac{W}{16}\times C_3$
$\frac{H}{32}\times\frac{W}{32}\times C_4$
$\frac{H}{4}\times\frac{W}{4}\times 4C$
$\frac{H}{4}\times\frac{W}{4}\times N_{cls}$
Overlap Patch Embeddings
Transformer Block 1
Transformer Block 2
Transformer Block 3
Transformer Block 4
MLP Layer
MLP
Efficient Self-Attn
Mix-FFN
Overlap Patch Merging
×N
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i$
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C$
$\frac{H}{4}\times\frac{W}{4}\times C$
MLP
UpSample

SegFormer with mixvisiontransformer as encoder



$$x_{out} = MLP(GELU(Conv3 \times 3(MLP(x_{in})))) + x_{in},$$

Method – Network Overview



$$z_k = \sigma(\mathrm{Conv}([h_{k-1}, x_k], W_z) + c_k), \tag{3}$$

$$r_k = \sigma(\mathrm{Conv}([h_{k-1}, x_k], W_r) + c_r), \tag{4}$$

$$\tilde{h}_k = \tanh(\mathrm{Conv}([r_k \odot h_{k-1}, x_k], W_h) + c_h), \tag{5}$$

$$h_k = (1 - z_k) \odot h_{k-1} + z_k \odot \tilde{h}_k, \tag{6}$$

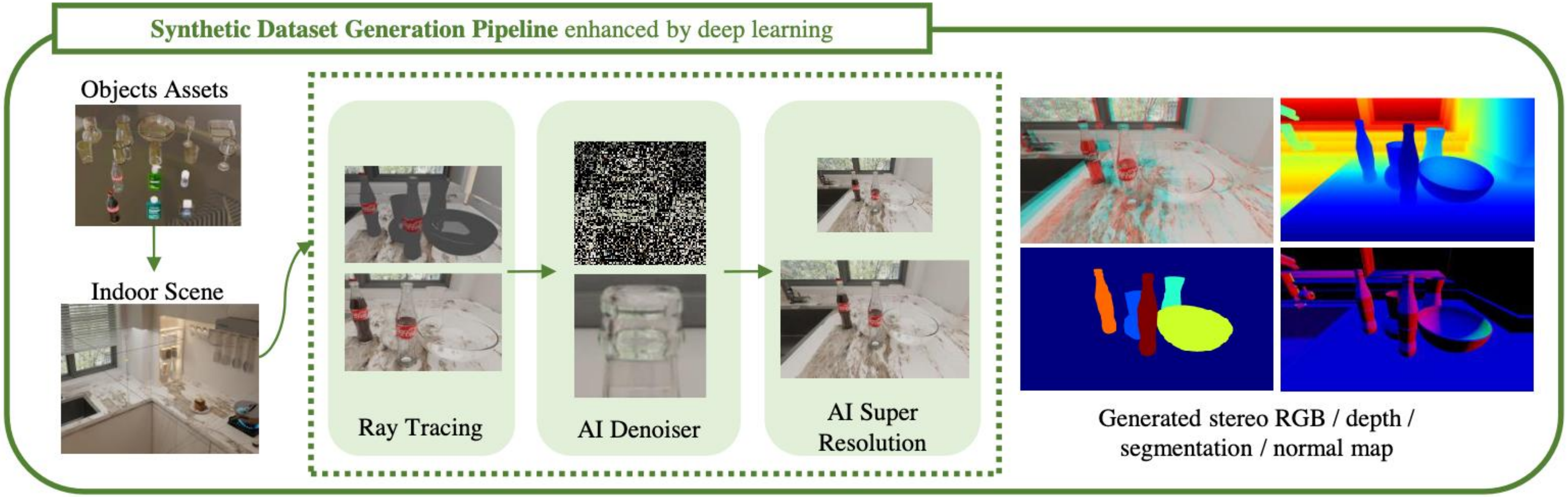
$$\Delta\mathbf{d}_{k,\frac{1}{32}} = \mathrm{Decoder}(h_{k,\frac{1}{32}}), \tag{7}$$

$$\Delta\mathbf{d}_{k,\frac{1}{16}} = \mathrm{Decoder}(h_{k,\frac{1}{16}} + \mathrm{Interp}(\Delta\mathbf{d}_{k,\frac{1}{32}})), \tag{8}$$

$$\Delta\mathbf{d}_{k,\frac{1}{8}} = \mathrm{Decoder}(h_{k,\frac{1}{8}} + \mathrm{Interp}(\Delta\mathbf{d}_{k,\frac{1}{16}})), \tag{9}$$

$$\mathbf{d}_{k+1} = \mathbf{d}_k + \Delta\mathbf{d}_k \tag{10}$$

Method – Synthetic Dataset Generation



Synthetic Dataset Generation Pipeline enhanced by deep learning

Objects Assets

Indoor Scene

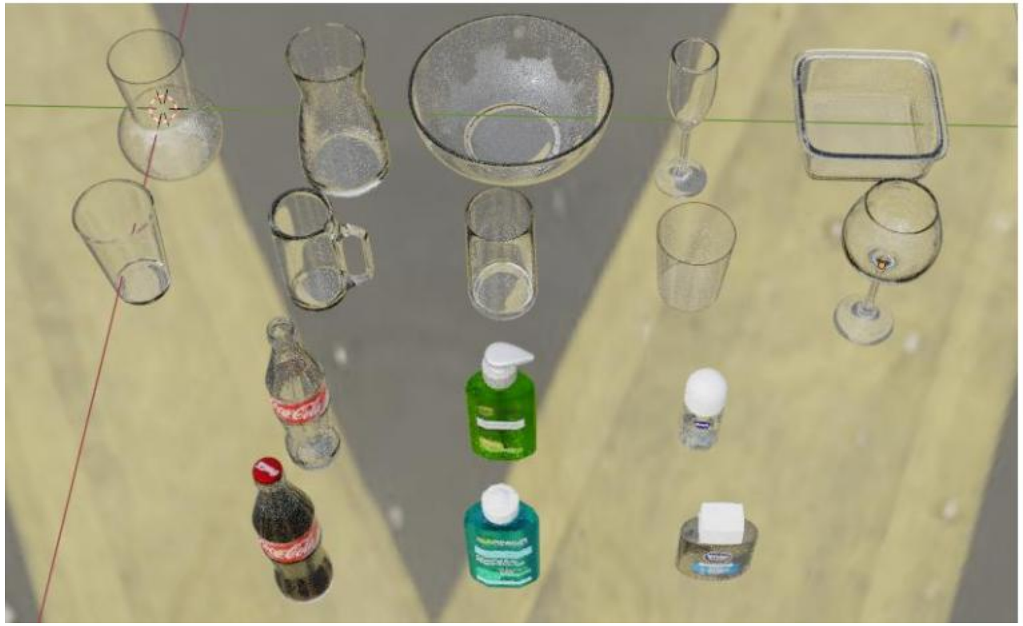
Ray Tracing

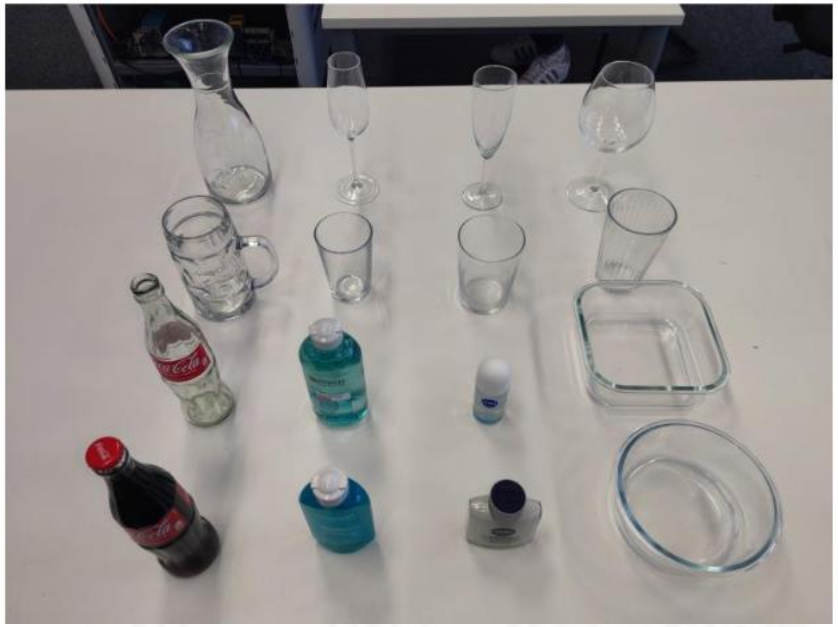
AI Denoiser

AI Super Resolution

Generated stereo RGB / depth / segmentation / normal map

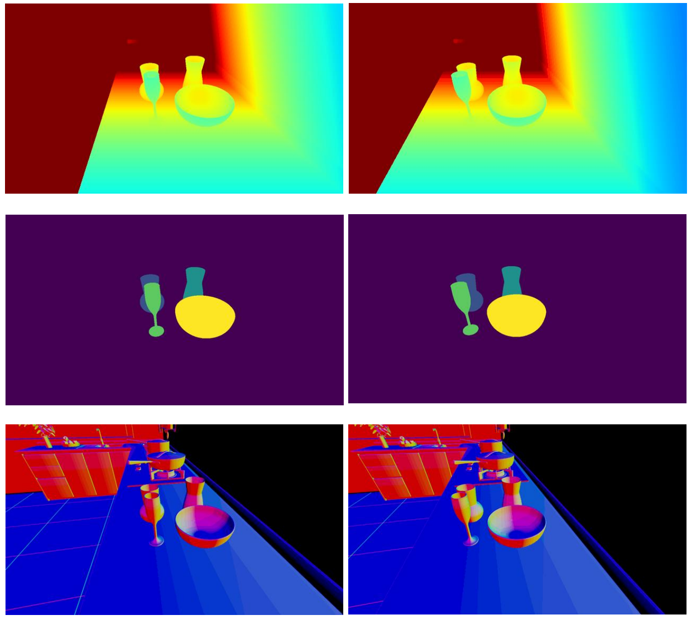
Method – Synthetic Dataset Generation





Method – Synthetic Dataset Generation

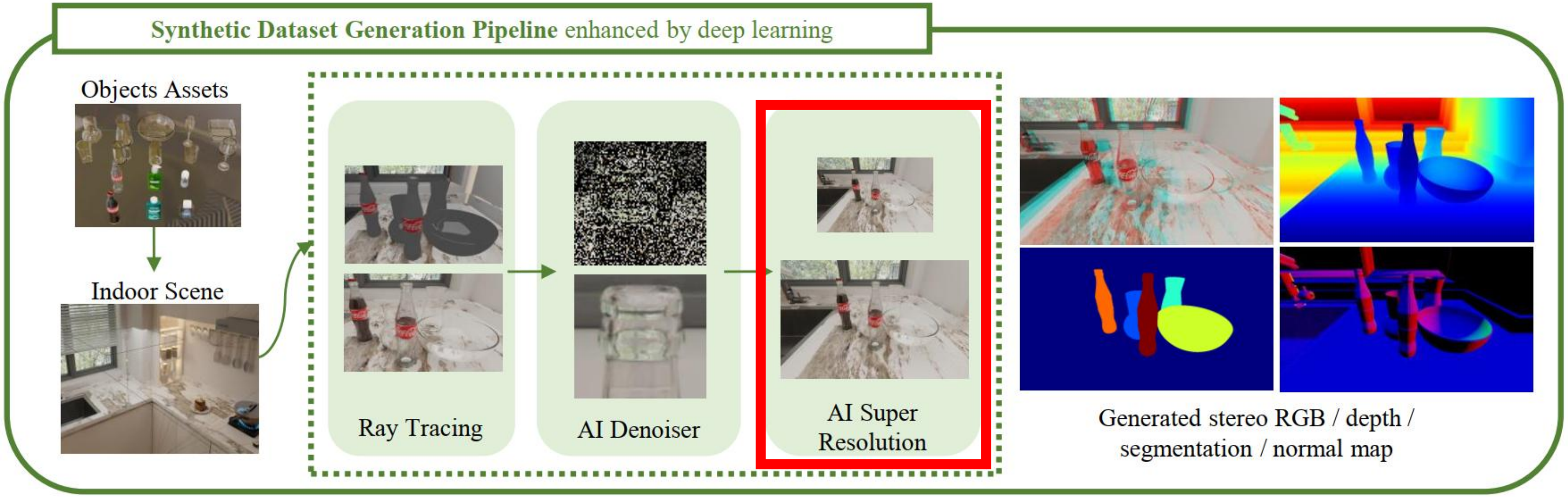
Left/Right RGB Image



Left/Right Depth Image,
Segmentation Map,
normal Map
(Objects' Poses)

Sample images from our
SynClearDepth dataset

Method – Synthetic Dataset Generation



Synthetic Dataset Generation Pipeline enhanced by deep learning

Objects Assets

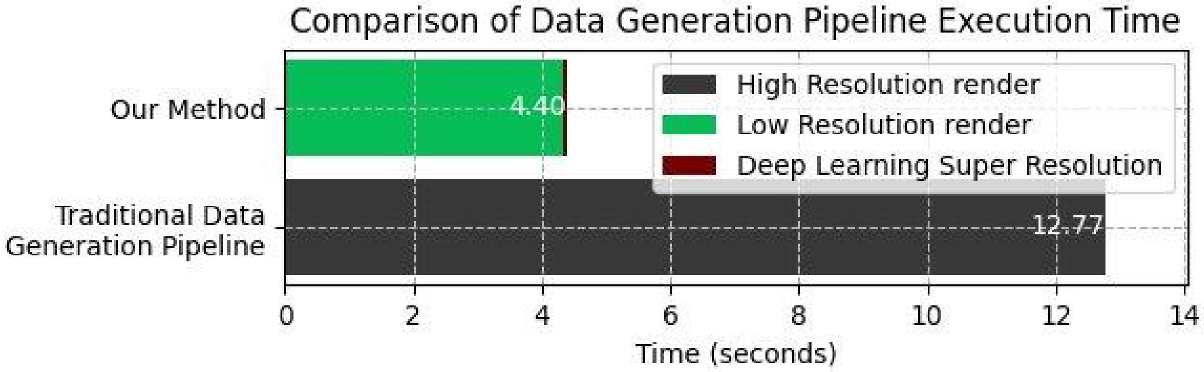
Indoor Scene

Ray Tracing

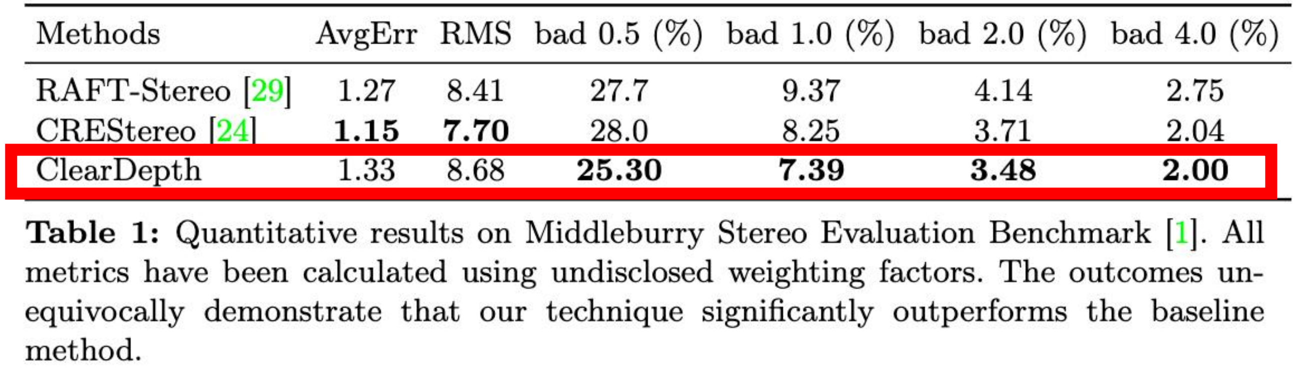
AI Denoiser

AI Super Resolution

Generated stereo RGB / depth / segmentation / normal map

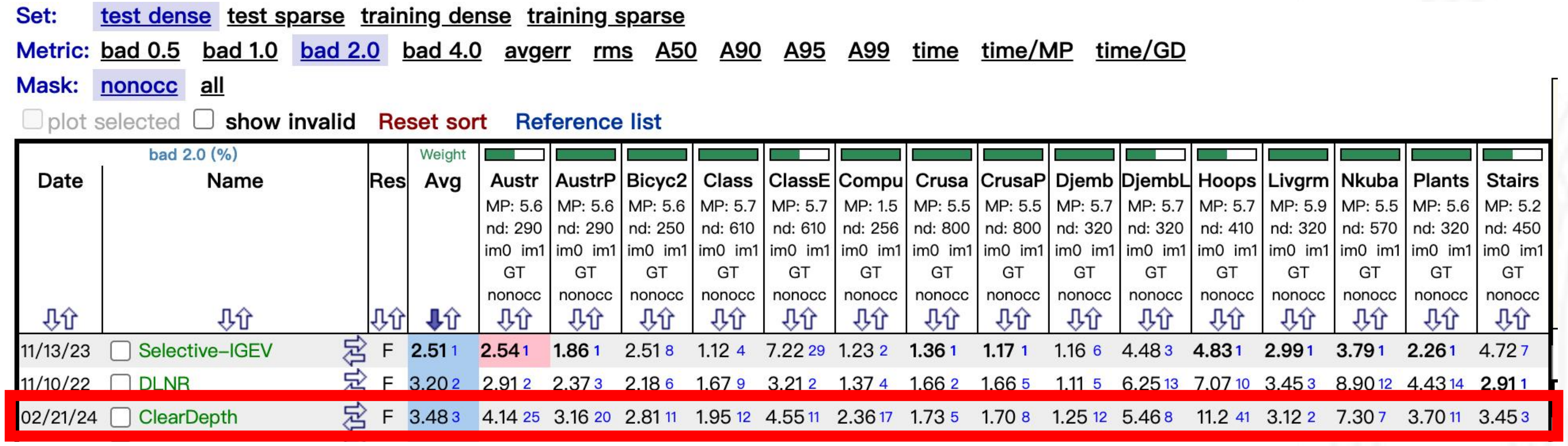


Experiments - Quantitative Analysis on Middlebury Dataset



Methods	AvgErr	RMS	bad 0.5 (%)	bad 1.0 (%)	bad 2.0 (%)	bad 4.0 (%)
RAFT-Stereo [29]	1.27	8.41	27.7	9.37	4.14	2.75
CREStereo [24]	**1.15**	**7.70**	28.0	8.25	3.71	2.04
ClearDepth	1.33	8.68	**25.30**	**7.39**	**3.48**	**2.00**

Table 1: Quantitative results on Middlebury Stereo Evaluation Benchmark [1]. All metrics have been calculated using undisclosed weighting factors. The outcomes unequivocally demonstrate that our technique significantly outperforms the baseline method.



Set: test dense test sparse training dense training sparse

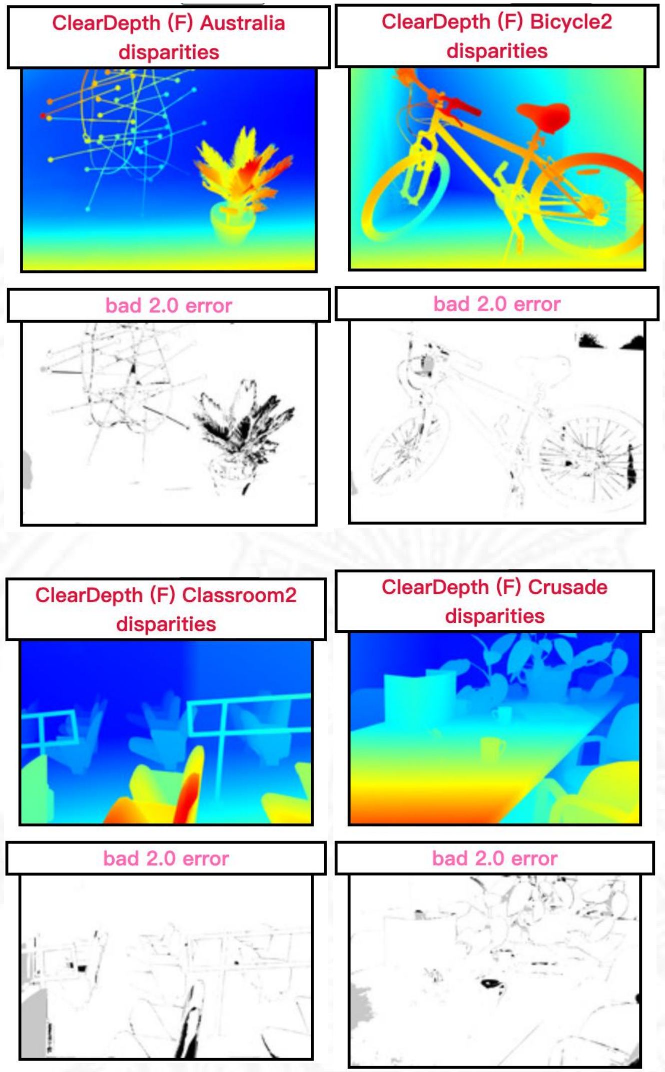
Metric: bad 0.5 bad 1.0 bad 2.0 bad 4.0 avgerr rms A50 A90 A95 A99 time time/MP time/GD

Mask: nonocc all

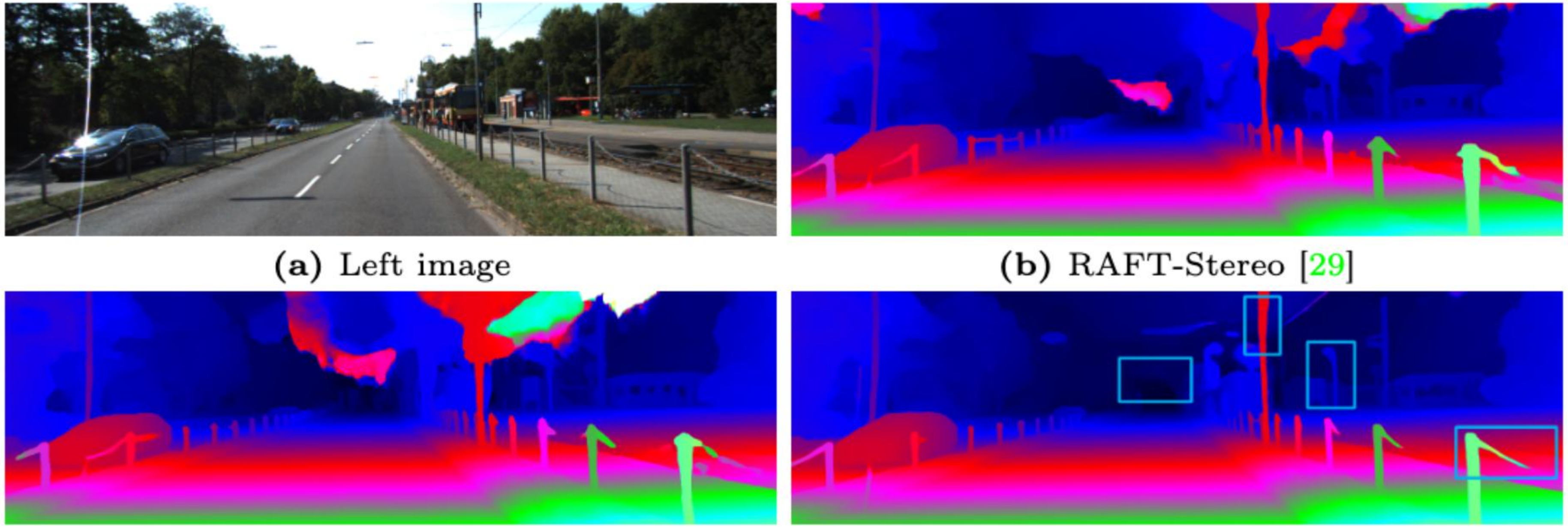
□ plot selected □ show invalid Reset sort Reference list

Date ⇩⇧	bad 2.0 (%) Name ⇩⇧		Res ⇩⇧	Weight Avg ⇩⇧	Austr MP: 5.6 nd: 290 im0 im1 GT nonocc ⇩⇧	AustrP MP: 5.6 nd: 290 im0 im1 GT nonocc ⇩⇧	Bicyc2 MP: 5.6 nd: 250 im0 im1 GT nonocc ⇩⇧	Class MP: 5.7 nd: 610 im0 im1 GT nonocc ⇩⇧	ClassE MP: 5.7 nd: 610 im0 im1 GT nonocc ⇩⇧	Compu MP: 1.5 nd: 256 im0 im1 GT nonocc ⇩⇧	Crusa MP: 5.5 nd: 800 im0 im1 GT nonocc ⇩⇧	CrusaP MP: 5.5 nd: 800 im0 im1 GT nonocc ⇩⇧	Djemb MP: 5.7 nd: 320 im0 im1 GT nonocc ⇩⇧	DjembL MP: 5.7 nd: 320 im0 im1 GT nonocc ⇩⇧	Hoops MP: 5.7 nd: 410 im0 im1 GT nonocc ⇩⇧	Livgrm MP: 5.9 nd: 320 im0 im1 GT nonocc ⇩⇧	Nkuba MP: 5.5 nd: 570 im0 im1 GT nonocc ⇩⇧	Plants MP: 5.6 nd: 320 im0 im1 GT nonocc ⇩⇧	Stairs MP: 5.2 nd: 450 im0 im1 GT nonocc ⇩⇧
11/13/23	□ Selective–IGEV	⇄	F	**2.51** 1	**2.54** 1	**1.86** 1	2.51 8	1.12 4	7.22 29	1.23 2	**1.36** 1	**1.17** 1	1.16 6	4.48 3	**4.83** 1	**2.99** 1	**3.79** 1	**2.26** 1	4.72 7
11/10/22	□ DLNR	⇄	F	3.20 2	2.91 2	2.37 3	2.18 6	1.67 9	3.21 2	1.37 4	1.66 2	1.66 5	1.11 5	6.25 13	7.07 10	3.45 3	8.90 12	4.43 14	**2.91** 1
02/21/24	□ ClearDepth	⇄	F	3.48 3	4.14 25	3.16 20	2.81 11	1.95 12	4.55 11	2.36 17	1.73 5	1.70 8	1.25 12	5.46 8	11.2 41	3.12 2	7.30 7	3.70 11	3.45 3

https://vision.middlebury.edu/stereo/eval3/


ClearDepth (F) Australia disparities
ClearDepth (F) Bicycle2 disparities
bad 2.0 error
bad 2.0 error
ClearDepth (F) Classroom2 disparities
ClearDepth (F) Crusade disparities
bad 2.0 error
bad 2.0 error

Experiments - Quantitative Analysis on KITTI Dataset



(a) Left image

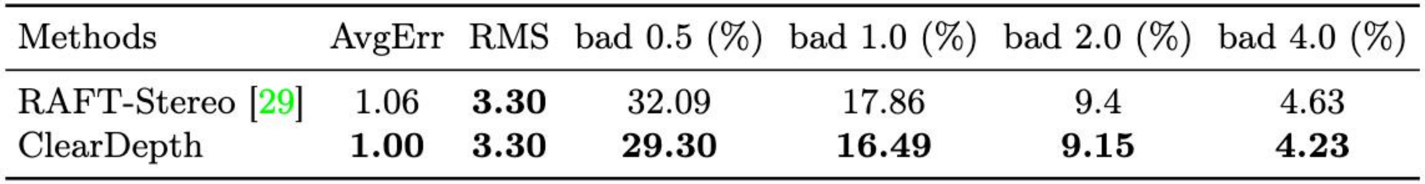
(b) RAFT-Stereo [29]

(c) StereoBase [17]

(d) Ours

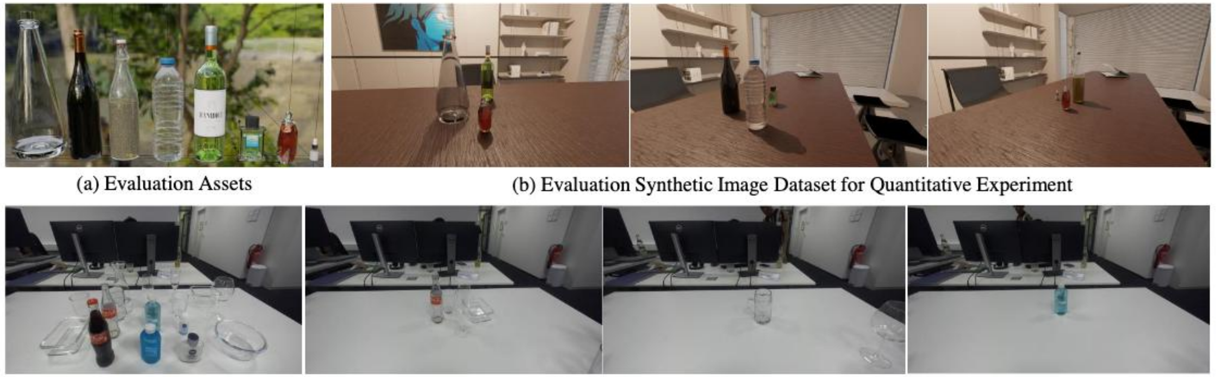
Fig. 7: Visual comparisons on KITTI 2015 test set. Our method is more robust to overall scene details.

Experiments - Evaluation on Our Transparent Object Dataset



Methods	AvgErr	RMS	bad 0.5 (%)	bad 1.0 (%)	bad 2.0 (%)	bad 4.0 (%)
RAFT-Stereo [29]	1.06	**3.30**	32.09	17.86	9.4	4.63
ClearDepth	**1.00**	**3.30**	**29.30**	**16.49**	**9.15**	**4.23**

Table 2: Quantitative results on our synthetic transparent object dataset as shown in Figure 8. The results clearly show that our method comprehensively surpasses the baseline approach.



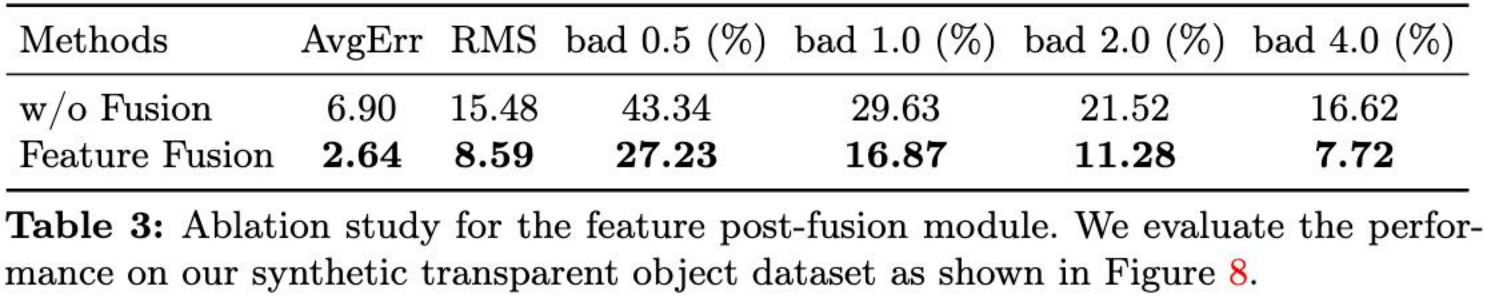
(a) Evaluation Assets

(b) Evaluation Synthetic Image Dataset for Quantitative Experiment

(c) Evaluation Real Image Dataset for Qualitative Experiment

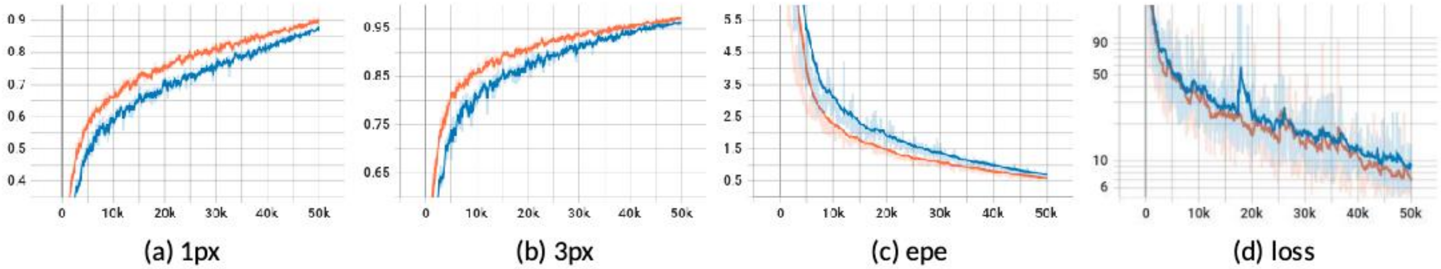
Fig. 8: The validation experiments for stereo depth recovery of transparent objects are divided into quantitative and qualitative experiments. The quantitative experiments utilize a synthetic image dataset with objects different from those in the SynClearDepth dataset, while the qualitative experiments employ a real-world image dataset of transparent objects collected for this purpose.

Experiments - Ablation Study of Feature Post-Fusion Module



Methods	AvgErr	RMS	bad 0.5 (%)	bad 1.0 (%)	bad 2.0 (%)	bad 4.0 (%)
w/o Fusion	6.90	15.48	43.34	29.63	21.52	16.62
Feature Fusion	**2.64**	**8.59**	**27.23**	**16.87**	**11.28**	**7.72**

Table 3: Ablation study for the feature post-fusion module. We evaluate the performance on our synthetic transparent object dataset as shown in Figure 8.



(a) 1px

(b) 3px

(c) epe

(d) loss

Fig. 9: Ablation study for the feature post-fusion module. The figure illustrates the curves of metrics and loss during the training process. The orange curve represents our model with the feature post-fusion module, while the blue curve corresponds to the model without this module.

Experiments - Robotic Grasping Experiments

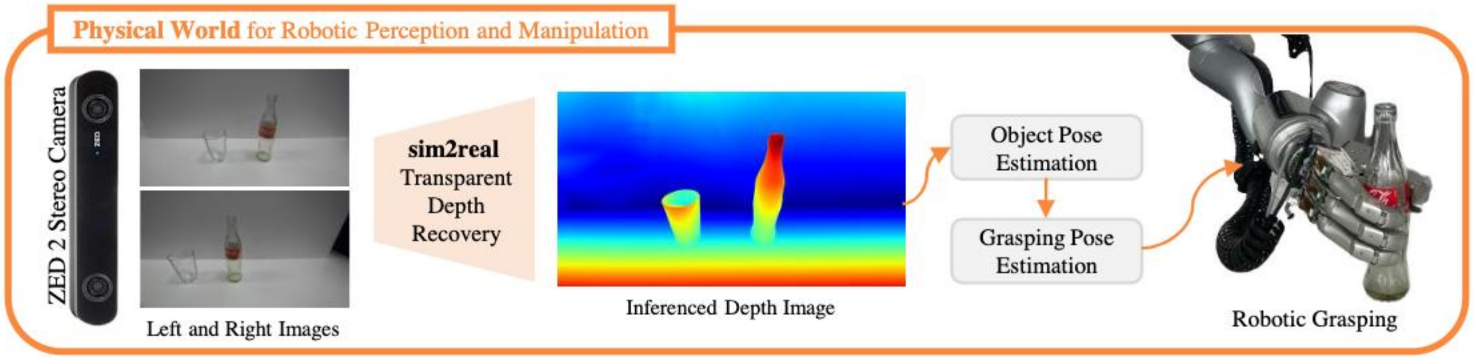
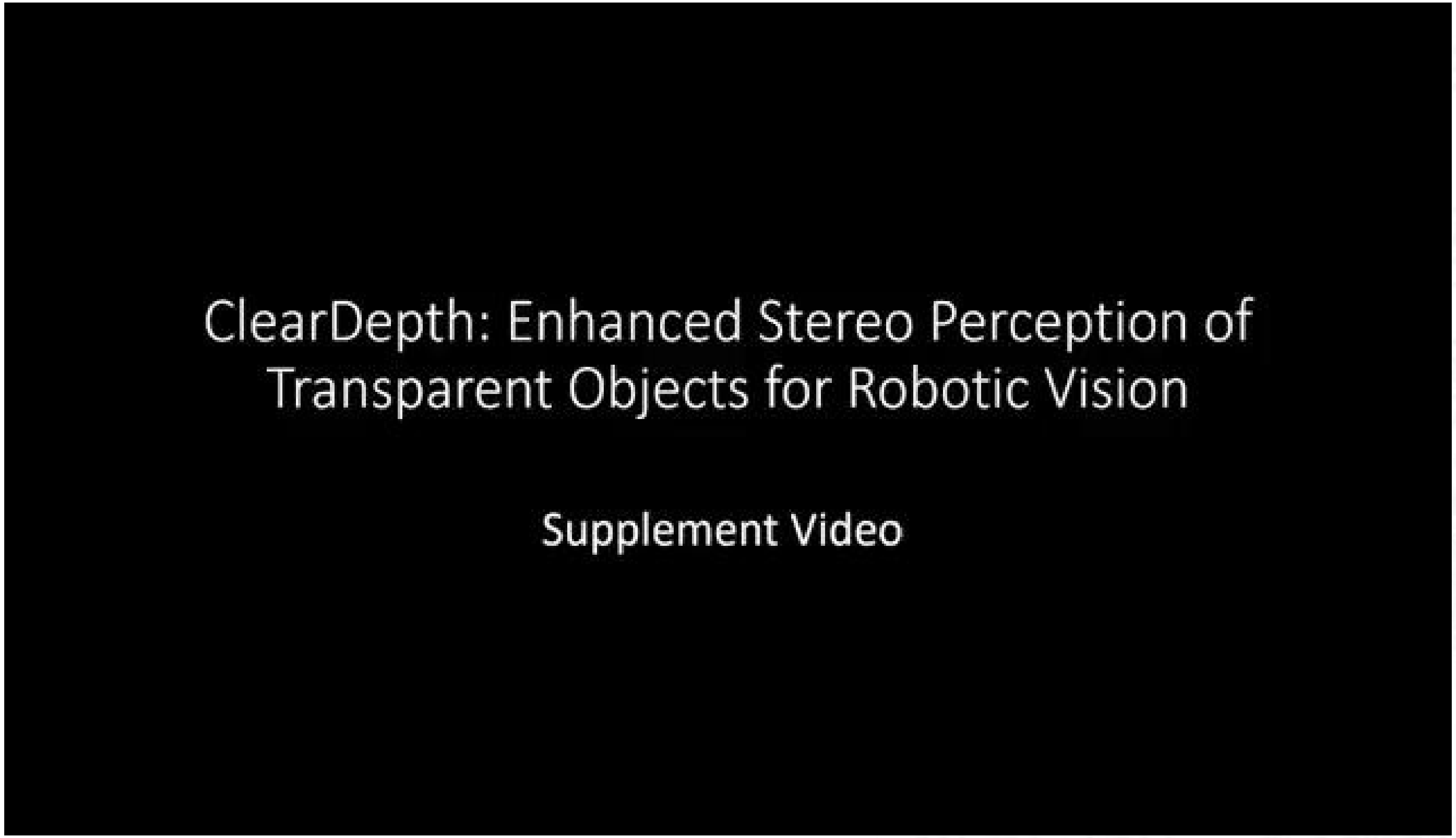


Fig. 10: Our proposed model processes left and right RGB images from the ZED 2 stereo camera to generate depth maps of transparent objects. Object poses are estimated via point clouds, enabling precise, pre-programmed grasping and manipulation with a five-fingered dexterous hand.

Experiments - Robotic Grasping Experiments



ClearDepth: Enhanced Stereo Perception of Transparent Objects for Robotic Vision

Supplement Video

Thank You!