Supplementary Material

Category-level Object Detection, Pose Estimation and Reconstruction from Stereo Images

1. Experiment Details

1.1. Network Details

In this section, we will provide a detailed introduction to the specific aspects of our network. As shown in Table 1, our method utilizes ConvNext-B [5] as image backbone. We output the last two layers of the backbone and adopt FPN [2] to further aggregate multi-dimensional information. Then our Implicit Stereo Matching produces 3D stereo embeddings with the same dimension as stereo features. We simply sum the 3D stereo features with 3D stereo embeddings as discussed in our paper. In the transformer decoder stage, we use 150 object queries to generate object embeddings. We use a 64-dimensional vector to represent the shape of objects.

Table 1. **Network Details of CODERS.**

Layers	Dimensions
Stereo Images	$2 \times 3 \times 608 \times 960$
Backbone (ConvNext-B)	$2 \times 512 \times 38 \times 60$ $2 \times 1024 \times 19 \times 30$
Neck (FPN)	$2 \times 256 \times 38 \times 60$
3D Stereo Embeddings	$2 \times 256 \times 38 \times 60$
Stereo-aware Features	$2 \times 256 \times 38 \times 60$
Object Queries	150×256
Object Embedding	256
Shape Embedding	64

Our transformer decoder contains six decoder layers. In each layer, we process object queries with the order of self-attention, norm, FFN, cross-attention, norm, FFN. The details of the decoder layer are shown in Table 2. Inspired by DETR [1], We use multi-head attention for all attention operations.

1.2. Experiment Settings

The TOD [4] dataset provides stereo images with a resolution of 720×1280 pixels. To ensure consistency with our SS3D dataset, we randomly resize and crop the input images to 600×960 pixels.

To maintain uniformity, we train cups and mugs in the same category, labeled as "cup" during the training process.

The Origin TOD dataset only offers key point annotations. To determine the object scale, we measure it using CAD models. Based on these measurements, we generate 6D pose annotations utilizing the provided key points.

Our real-world data is captured at a resolution of 1200×1920 pixels. However, for our purpose, we resize the input images to 600×960 pixels.

Table 2. **Network Details of Decoder Layer.**

Layer	Q	KV
Self-attention	Queries 150×256	Queries 150×256
Cross-attention	Queries 150×256	Features $2 \times 38 \times 60 \times 256$

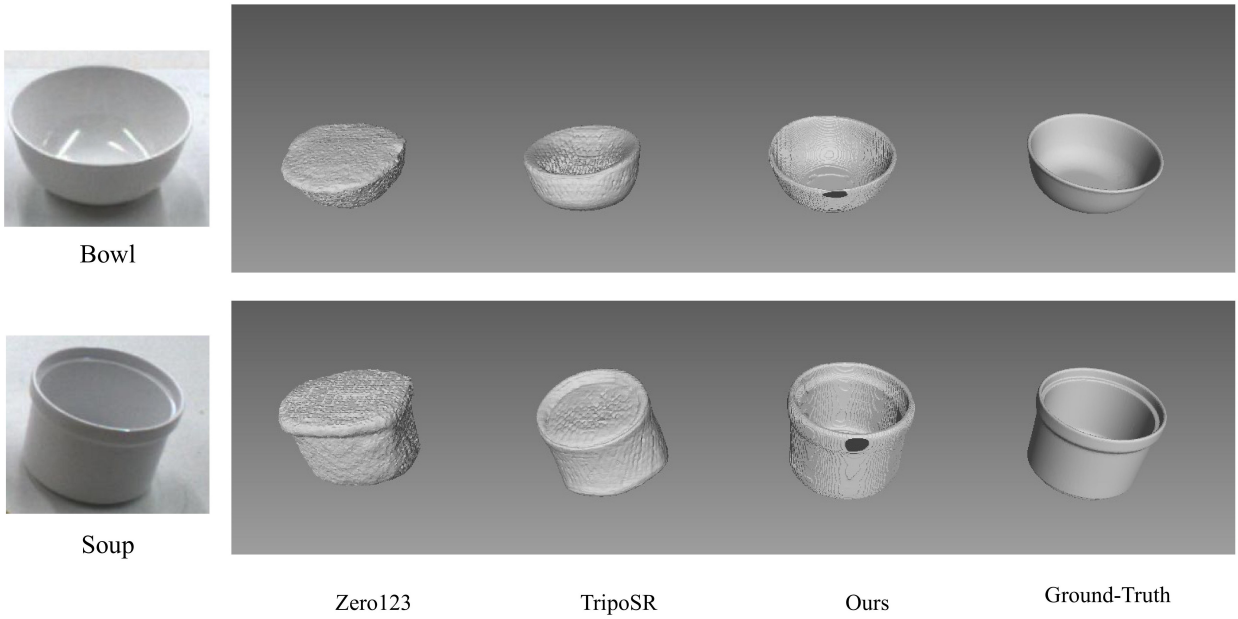

Figure 1. **Qualitative Results of Reconstruction** Our approach generates meshes with high quality and can reconstruct blade shapes.

For the purpose of reconstruction comparison, we selected zero123 and TripoSR. To meet their requirements, we preprocess our input images to be object-centric and free from background interference.

In the ablation study, we utilize the TOD dataset and follow the same settings as discussed above.

2. Qualitative Results of Reconstruction

We do more comparison for CODERS with Zero123 [3] and TripoSR [6] on real-world data. To ensure a fair comparison, we adopt the same setting as Zero123 and TripoSR, which require object-centric images without background, while CODERS utilizes the entire image. As depicted in Figure 1, CODERS is capable of generating high-quality meshes with only single-view stereo images as input.

3. Qualitative Results on SS3D Test Dataset

We generate a large-scale stereo category-level object dataset called **SS3D**. To build the SS3D dataset, we select 427 objects from OmniObject3D [7] and reserve 64 objects from 16 categories for testing. The results are shown in Figure 2. Our method demonstrates excellent generalization ability for scenes and objects.

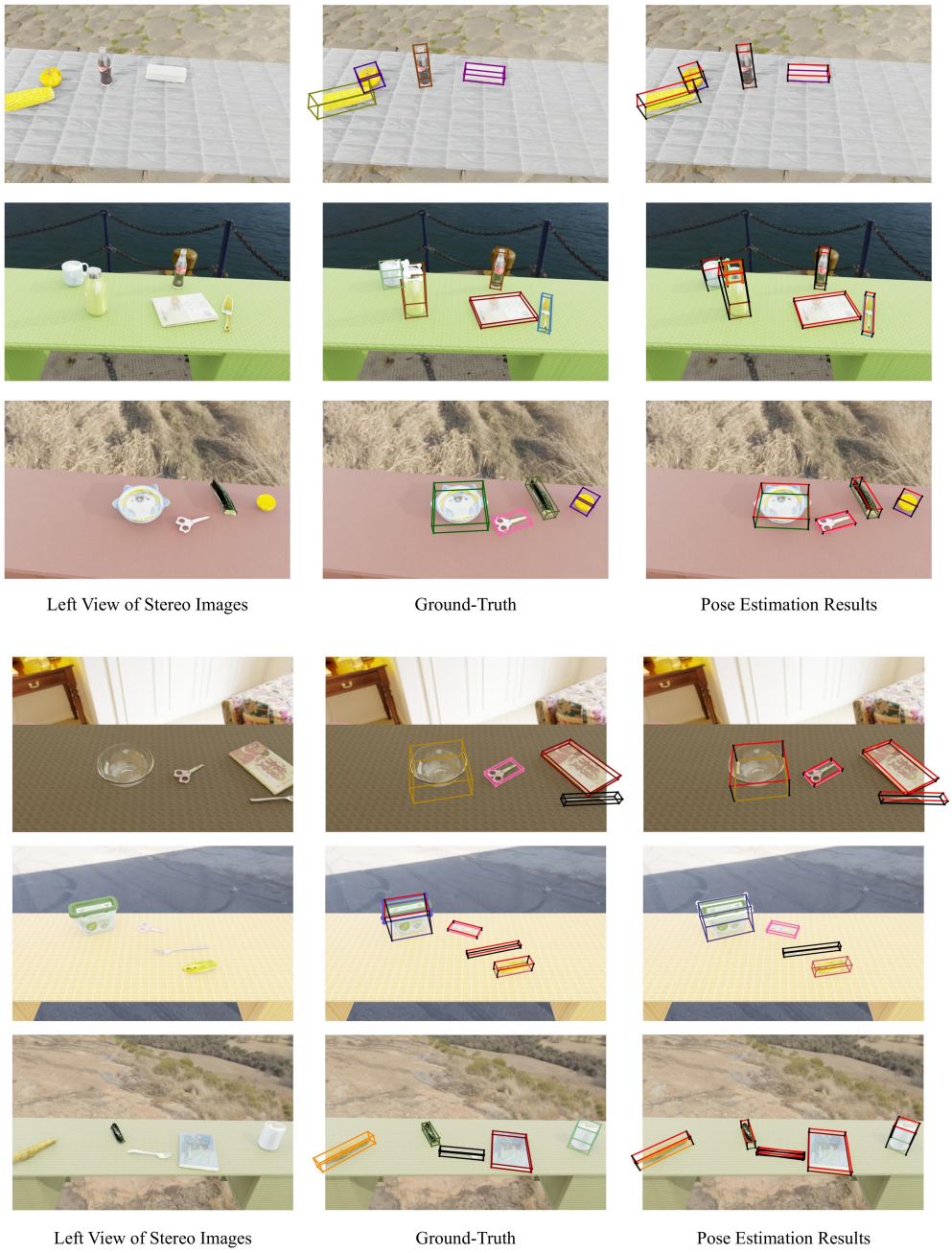

Figure 2. **Qualitative Results on SS3D Test Dataset** The bottom color of the 3D bounding boxes represents the object category. Our method can handle all 16 object categories using a single model. Importantly, these objects have varying surface properties including specular, transparent, and diffuse.

4. Failure Cases

In this section, we show some failure cases of CODERS on unseen objects see Figure 3. Objects with outlier scales, strange materials, and occlusions remain significant issues. These issues are common problems in the field of computer vision, which require further exploration.

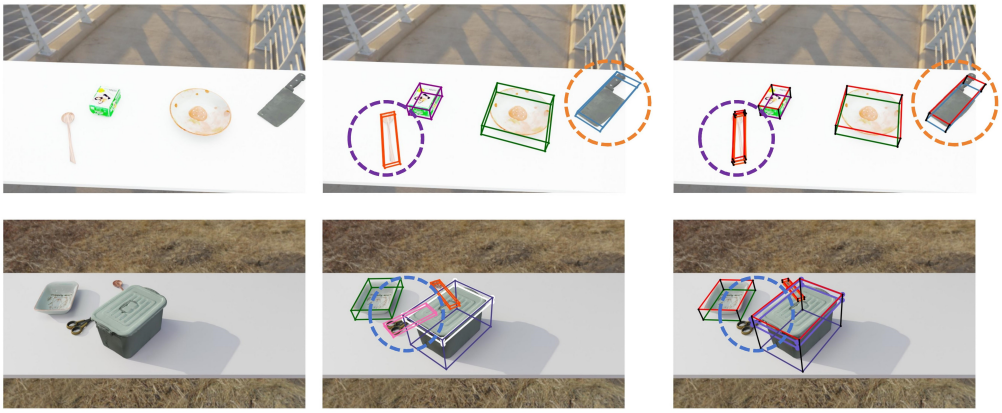

Figure 3. **Failure Cases** The purple circle indicates a wooden spoon with low confidence. The orange circle represents a knife with an outlier scale. The blue circle denotes a pair of scissors that is occluded. These issues are common problems in the field of computer vision, which require further exploration.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[2] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[3] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 2

[4] Xingyu Liu, Rico Jonschkowski, Anelia Angelova, and Kurt Konolige. Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11602–11610, 2020. 1

[5] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 1

[6] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 2

[7] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 2