2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

MVTrans: Multi-View Perception of Transparent Objects

Yi Ru Wang[†♯*], Yuchi Zhao[‡*], Haoping Xu[†*], Sagi Eppel[†],
Alàn Aspuru-Guzik[†], Florian Shkurti[†], Animesh Garg[†∓]

***Abstract*— Transparent object perception is a crucial skill for applications such as robot manipulation in household and laboratory settings. Existing methods utilize RGB-D or stereo inputs to handle a subset of perception tasks including depth and pose estimation. However transparent object perception remains to be an open problem. In this paper, we forgo the unreliable depth map from RGB-D sensors and extend the stereo based method. Our proposed method, MVTrans, is an end-to-end multi-view architecture with multiple perception capabilities, including depth estimation, segmentation, and pose estimation. Additionally, we establish a novel procedural photo-realistic dataset generation pipeline and create a large-scale transparent object detection dataset, Syn-TODD, which is suitable for training networks with all three modalities, RGB-D, stereo and multi-view RGB. https://ac-rad.github.io/MVTrans/**

I. INTRODUCTION

Transparent objects are prevalent in our daily lives, and their use spans household, laboratory, and industrial settings. However, the unique specular properties of transparent objects cause perception challenges, particularly in areas of depth estimation, segmentation, and pose estimation. Specifically, transparent objects differ from common objects in their ability to inherit visual properties of the background, as well as distort light rays and hence the depth modality of commodity RGB-D sensors, which operate on the assumption that objects have opaque lambertian surfaces [1].

Existing methods have addressed transparent object perception challenges in two ways. RGB-D based depth completion methods [2–4] recover estimated transparent object depth from raw sensor depth and RGB features, and use the predicted depth for pose estimation tasks. Another stream of works [1, 5] skip the unreliable sensor depth and directly work on transparent object pose estimation using stereoscopic imagery. Bypassing the depth completion problem brings advantages over RGB-D based methods. Namely, it unifies models from multi-step pipelines into a single end-to-end model. These stereo-based models have higher capacity than build-in depth sensor algorithms to handle non-lambertian surface objects.

Multi-view estimation extends stereo vision by providing richer information for a given scene especially in complex settings with occlusion. Multi-view methods [8–10] demonstrate superior performance in 3D vision tasks and have the potential to surpass stereo vision methods when handling transparency by enabling fusion of diverse viewing angles. This is especially convenient for autonomous settings, where an eye-in-hand camera can capture varying views to perceive and manipulate transparent objects [1, 3], shown in Figure 1.

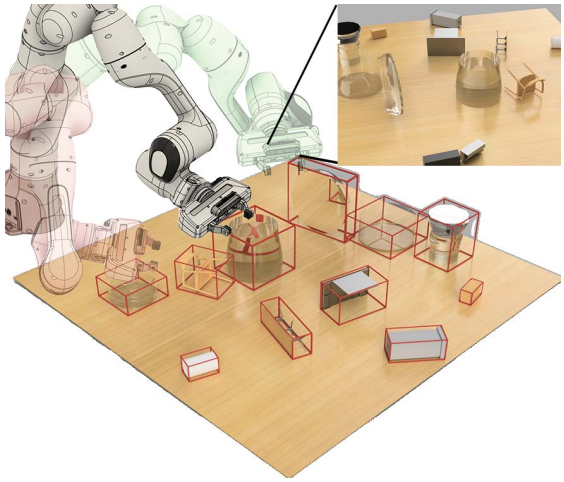

Fig. 1: Given a set of multi-view images of the scene, the proposed method (MVTrans) predicts segmentation mask, depth map, pose, and 3D bounding box for both opaque and transparent objects using an end-to-end multi-task perception network.

To train such multi-view networks, large-scale transparent object datasets are needed, yet existing ones are not suitable. Some works [2, 7] focus on single view tasks and lack multi-view annotations. Other datasets collected using robots or SLAM [1, 3, 6] provide the multi-view images necessary for training. However, these real-world datasets are limited in object diversity and scene complexity.

In this work, we introduce an end-to-end multi-view architecture for perception of transparent objects, including depth estimation, segmentation, and scene understanding (pose and 3D bounding box prediction). Our method outperforms state-of-the-art RGB-D and stereo models on both real-world and synthetic datasets for all transparent object perception tasks. To ensure model generalizability, we present a novel pipeline for procedural photo-realistic dataset generation, and a large-scale transparent object dataset (Syn-TODD). The dataset generation pipeline enables procedural generation of transparent objects, paired with domain randomized scene setup. In total, our dataset includes 1996 photo-realistic tabletop scenes with transparent and opaque objects, and 57 fully annotated views for each scene. In summary, our contributions are twofold:

1) A novel end-to-end multi-view architecture, MVTrans, for multi-task perception of transparent objects. It can perform depth estimation, segmentation, and scene understanding including pose and 3D bound box prediction for every

[†]University of Toronto & Vector Institute,
[‡]University of Waterloo, [∓]Nvidia
[♯]University of Washington, yiruwang@cs.washington.edu
[*] Authors contributed equally

Fig. 2: Sample images of transparent object datasets. (a) Top row: existing methods including ClearGrasp [2], TODD [3], ClearPose [6], KeyPose [1] and Trans10K [7]. (b) Bottom row: Ours (Syn-TODD). Syn-TODD's scene complexity and object diversity are superior when compared with existing datasets.

object in a given scene. Its performance exceeds current RGB-D and stereo-based methods.

2) A large-scale transparent object dataset, Syn-TODD. It has wide compatibility with RGB, RGB-D, stereo and multi-view based methods and superior scene complexity, object diversity and annotation richness. To create it, a rendering pipeline with domain randomization and procedural generation are employed.

II. Related Work

A. Transparent Object Perception

Transparent objects pose additional problems for both 2D and 3D perception tasks, mainly due to its non-opaque, non-lambertian surface which breaks traditional methods' assumption of opaqueness [1]. Similar to opaque object perception, transparent object perception tasks include segmentation, depth estimation, 3D bounding box and pose prediction.

For **segmentation**, [11–15] aim to improve the performance of general object recognition and segmentation on particular transparent objects. [13, 15–17] segment transparent and reflective surfaces for visual navigation and scene understanding. Several monocular RGB methods leverage the unique difference in appearance and texture along the edge of transparent vessels by incorporating boundary cues [13–15].

Depth completion is particularly important for RGB-D based networks when handling transparent objects, whose depth appears distorted when captured raw. Some works demonstrate the effectiveness of a global optimization approach, which leverages the combination of predicted surface normal, occlusion boundary and original depth for depth estimation guidance [2, 18]. Other methods use encoder-decoder or Generative Adversarial Networks (GANs) to generate the completed depth map by regression [19, 20].

3D BBox and pose prediction methods concern either axis-aligned bounding box or oriented bounding box. Some existing works use CNNs to train from RGB-D input [21–23], where the depth is generated from aforementioned depth completion modules. Stereo vision methods avoid the distorted depth and directly take stereo image pairs as input. KeyPose [1] is a stereo RGB and keypoint based method, and SimNet [5] is a stereo and oriented 3D BBox based method. Our proposed MVTrans architecture is capable of preforming all three perception tasks: segmentation, depth completion and scene understanding which includes pose and 3D BBox for every object in the scene.

B. Multi-view Perception

Single-view perception refers to estimation of scene parameters and properties using a single monocular image input. Multi-view perception refers to the broad category of using more than one view to infer 3D information from the captured scene. Some works impose the epipolar constraint, which leverages a pair of stereo images to learn the associated disparity and depth [1, 5]. Other works use supervisory signals for depth estimation guidance [9], or enforce constraints regarding the spatio-temporal consistency between consecutive frames [8]. Recent works also demonstrate the advantage of plane sweeping volume algorithm for in multi-view 3D feature fusion [5, 10]. Our work is within the broad multi-view perception category, in which we leverage multiple overlapping views for multi-task learning.

C. Transparent Object Dataset

Transparent objects lack an ideal benchmark dataset for 3D perception tasks. Existing transparent datasets come with different limitations. Trans10K [7] is a 2D segmentation dataset that consists of real images of transparent objects created by manual annotation. ClearGrasp [2] proposed a synthetic 3D dataset with 9 unique objects. KeyPose [1] collects a 3D real-world dataset of 15 unique objects using an eye-in-hand robot with only single object scenes. TODD [3] automates the collection and annotation process using eye-in-hand camera and AprilTags, its dataset includes 8 unique objects and complex scenes with cluttered and filled glassware. Recently, ClearPose [6] proposed to use SLAM and manual CAD model alignment for large-scale real world transparent object dataset creation. To improve these limitations, we create Syn-TODD, a large-scale transparent dataset with superior scene complexity, object diversity, and annotation richness which supports model training with RGB-D, stereo and multi-view modalities, a sample of which is shown in Figure 2.

III. MVTrans: Multi-Task Model Architecture

Our proposed method, MVTrans, as shown in Figure 3, is an end-to-end multi-view architecture with multiple perception capabilities, namely depth estimation, segmentation, and scene understanding (pose and 3D bounding box prediction for every opaque and transparent object in a given scene). MVTrans takes a set of multi-view images as input, whose

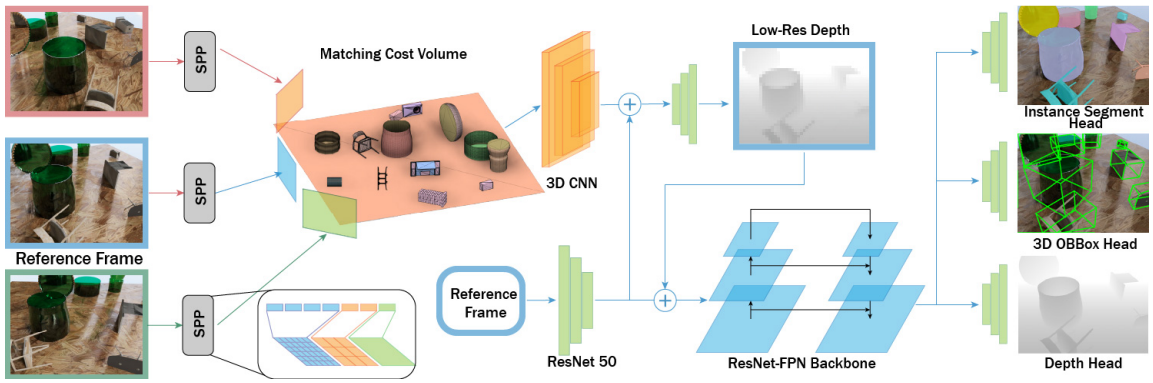

Fig. 3: MVTrans Architecture. Images from different viewing angles are used as input, in which one frame is selected as the reference frame for predictions. For each image, a shared spatial pyramid pooling (SPP) module extracts its features. Subsequent plane-sweeping warps are applied to non-reference frames to build the matching volume, which is then regularized by a 3D CNN. In parallel, the reference frame's 2D context features is extracted by ResNet-50. Concatenated 3D matching volume and 2D context features are used to generate a low-res depth prediction. Both RGB and depth features are fed to a ResNet-FPN backbone as well as downstream output heads, which predicts instance segmentation, 3D OBB and full resolution depth map.

2D features are back-projected into a 3D matching volume. The matching volume and reference image's 2D features are used to predict a low resolution depth map. Both the RGB features and low-res depth feature then passed through a backbone feature extractor before reaching the multi-task prediction heads for segmentation, 3D bounding box and pose estimation, and depth prediction.

A. Local and Global Context Fusion for Multi-view Input

Given a collection of RGB input $\{I_1, ..., I_N\}$, where $N \geq 2$ is the number of multi-view images, we refer to the first image as the reference image, and the others as support images whose captured view partially overlaps with the reference image scene. Each image has dimensions $H \times W \times 3$. Similar to the feature extraction process in [24], expert networks are used to produce a Matching Volume (3D) and a Context Volume, which are jointly fused into a cost volume that encapsulates the local and global features.

Matching volume. Each multi-view image is passed through a spatial pyramid pooling (SPP) module [25] to encapsulate and aggregate context across different scales and locations to form feature maps. The feature maps of the support images are back-projected into the coordinate system of the reference image, at a stack of parallel planes with depths sampled based on ranges observed in the dataset. The goal of the back-projection is to capture and incorporate the photo-consistency of the warped images on the pixel level. At each sampled depth z_d, where $d \in [1, D]$ is the plane index, a planar homography transformation is applied to obtain the coordinate mapping from the support to the reference image:

$$u'_i \sim \mathbf{K}[R_r|t_r] \begin{bmatrix} (\mathbf{K}^{-1}u)z_i \\ 1 \end{bmatrix} \quad (1)$$

where u denotes the coordinate of a pixel in the support image frame, $\mathbf{K}$ denotes the camera intrinsic matrix, $\{R, t\}$ denotes the rotation and translation of the support image coordinates to the coordinates of the reference image.

The transformed coordinate mapping is used to construct the warped feature volume, with dimensions of $C \times D \times H' \times W'$, where C denotes the number of channels, and H' and W' denotes the scaled down height and width. When the reference and support feature volumes are concatenated together, we obtain a raw matching volume of $2C \times D \times H' \times W'$ dimension. For $N-1$ support images, we will have $N-1$ raw matching volumes. 3D convolution layers process the raw matching volumes and reduce the dimension to $C \times D \times H' \times W'$. Information across the $N-1$ raw matching volumes are then aggregated through a view average pooling operation. We then apply a series of 3D convolutions to further regularize the aggregated volume to produce the regularized matching volume, which encapsulates the local features for matching.

Context volume. To learn the global 2D context, we apply a ResNet-50 [26] architecture and obtain a feature volume with dimensions of $D \times H' \times W'$. Notice the number of channels is equivalent to the number of sampled depths D in matching volume. This 2D context volume is then fused with the 3D matching volume through expanding the dimensions to $1 \times D \times H' \times W'$. After concatenation of the regularized matching volume and the 2D context volume, we obtain a final cost volume of size $(C+1) \times D \times H' \times W'$ which encapsulates the local and global context.

Rough depth map. The cost volume is fed through consecutive convolution layers to obtain volumes with size $D \times H' \times W'$. A softmax operator is applied on the D dimension to create a probability volume P, from which the expected depth is extracted using the soft argmax operation.

B. Perception Predictions

Extraction of high-level perception predictions begins with the concatenation between the low resolution depth map and extracted features of the reference image $I_{reference}$, which enable learning with combined depth, colour, and textual cues, similar to [5, 27]. The output passes through a ResNet-18-FPN feature backbone before reaching the perception heads.

Pose and 3D bounding box estimation. This involves predicting the oriented bounding box of rigid objects, including the translation $t \in \mathbb{R}^{1\times3}$, rotation $R \in \mathbb{R}^{3\times3}$, and size of each object instance along three axis $S \in \mathbb{R}^3$. The procedure used for estimating object pose involves several components,

including differentiating object instances, deriving object size, and predicting object translation and rotation.

To differentiate object instances, we model each object as a bivariate normal distribution around its center and conduct peak detection. The heatmap can be computed as:

$$\text{HeatMap}(\mathbf{p}) = \max_{i \in \mathcal{O}} (\mathcal{N}(\mathbf{p}; \mu_i, \sigma_i)) \tag{2}$$

where $\mathcal{I}_{reference}$ has pixels $\{\mathbf{p}\}$, $\mathcal{O}$ is the set of object instances within image $\mathcal{I}_{reference}$, and μ_i and σ_i denotes the centroid and covariance of object i, respectively.

Determining object size requires prediction of the displacement field, which encapsulates information about the distance between each pixel and the eight vertices of the associated object. The vertex offset can be calculated for an image-plane projected vertex $\mathbf{v}$, and a pixel $\mathbf{p}$ as:

$$\text{VertexOffset}(\mathbf{p}) = \mathbf{v} - \mathbf{p} \tag{3}$$

We operate at the $H/8 \times W/8$ resolution to reduce computation requirements, which results in a displacement field with dimensions of $H/8 \times W/8 \times 16$. To combine displacement fields for all objects within the scene, we consider probabilistic values for each pixel from the heatmap, and merge based on the object with the highest probability.

To recover the translation, the objects' centroid distances from the camera are regressed as a $H/8 \times W/8$ tensor, which can be used in combination with the camera pose to derive the object translations in world space. Note that the centroid distance field contains information for all objects in the image, the partitioning of which is based on the heatmaps and the object with the highest probability at each pixel, similar to the combined displacement field. Rotation estimation is based on covariance matrix prediction. Computation of the ground-truth covariance of the object begins with sampling points on the object mesh surface in simulation, in the object's local space. The points are then converted to camera space and used to compute the covariance:

$$\text{Covariance}(\mathbf{c}_C) = \text{Covariance}(R_W^C \cdot R_L^W \cdot \mathbf{c}_L) \tag{4}$$

where C denotes camera frame, W denotes world frame, and L denotes local frame. $\mathbf{c}$ refers to the coordinates of points on the object surface. We then compute covariance on $\mathbf{c}_C$. The ground truth covariance matrix of each object is combined in a similar manner as the displacement field and the centroid distance field, and is used as a supervision signal to enable covariance prediction, which is regressed as a tensor of $H/8 \times W/8 \times 6$, consisting of the elements of the upper triangular matrix of the standard covariance matrix for 3D point clouds. Rotation is then recovered through the Singular Value Decomposition (SVD) of the covariance matrix.

Segmentation. Instance segmentation differentiates the table surface, background, and objects on the table. Training is done using the up-scaling branch approach introduced in [28] with multiple up-sampling layers, where each layer consists of 3×3 convolution, group norm, ReLU and $2\times$ bilinear up-sampling. We use cross entropy loss for training.

Depth. To predict the full resolution depth map, we apply the up-scaling branch similar to our segmentation head, which enables aggregation of several features across different scales. Training is achieved using the Huber loss to minimize:

$$L_{depth} = \text{Huber}(f_{depth}(I_{1:N}), D_{reference}) \tag{5}$$

where $D_{reference}$ denotes the ground truth depth map of the reference frame.

TABLE I: Transparent Dataset Comparison. Comparison of Syn-TODD (ours) with ClearGrasp [2], Trans10K [7], Keypose [1], TODD [3], ClearPose [6]. Our 3D transparent object dataset has significant advantage over others in terms of sample size, scene complexity, object diversity and annotation richness.

	Trans10K	ClearGrasp	TODD	ClearPose	KeyPose	**Syn-TODD (Ours)**
Samples	10K	50K	15K	360K	15K	113K
Objects	10K	10	8	63	10	16K
Scenes	10K	33	22	63	10	1996
Objs/scene	1-20	1-5	1-3	1-25	1	3-15
RGB	mono	mono	mono	mono	stereo	multi-view
Segment	semantic	semantic	instance	instance	instance	instance
Depth	✘	✔	✔	✔	✔	✔
Pose	✘	✘	✔	✔	✔	✔
3D Bbox	✘	✘	✘	✘	✘	✔
Normal	✘	✔	✘	✘	✘	✔
Keypoints	✘	✘	✘	✘	✔	✔

IV. Synthetic Transparent Object Dataset

Despite recent improvements of real-world data collection, synthetic datasets still lead in terms of throughput, annotation accuracy, object diversity, and scene complexity [29–35]. Given the promising performance of stereo vision models [1] and sim-to-real training [2, 4] for transparent object detection, a photo-realistic and large-scale transparent object dataset is needed. We present Syn-TODD, which has wide compatibility with RGB, RGB-D, stereo, and multi-view based methods. For a given scene, we render stereo image pairs from a grid of viewing angles. Additionally, procedural generation of objects and domain randomization of scenes enhance the dataset's complexity and aids model generalization.

Scene Setup. We use Blender [36] for high-fidelity, photo-realistic synthetic data generation [37]. Each scene consists of three parts: background, tabletop, and objects. To diversify scene appearances, we apply domain randomization to select the background from 1000+ High Dynamic Range Image (HDRI) for environments and illumination variances, and the tabletop surface from 1400+ Physics Based Rendering (PBR) materials with varying textures and visual appearances. Multiple light sources are also introduced at random locations.

Procedural Generation. Transparent objects are procedurally generated. We employ a method that creates the vessel curvatures using 2D function combinations of linear, polynomial, and sinusoidal functions, with coefficients and parameters differing across vessel. For each generated vessel, we apply a transparent material with randomized properties, including color, index of refraction, transparency, reflection, and roughness, among others. Each scene contains up to seven random transparent objects, with possible occlusion.

Object Selection. As shown in Figure 2, up to seven generated transparent objects are placed in each scene. Additionally, we randomly place up to eight different objects from a subset of ShapeNet [38] with 13000+ models to

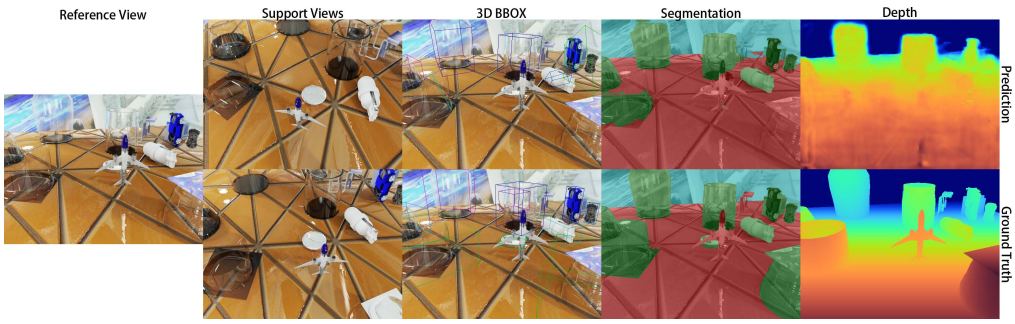

Fig. 4: Samples and prediction results from Syn-TODD. (a) Multi-view RGB image (reference view and support views) (b) 3D OBB (GT and prediction) (c) Instance Segmentation (GT and prediction) (d) Depth (GT and prediction)

TABLE II: Pose estimation results on the KeyPose [1]. We assess the performance of DenseFusion, Keypose, SimNet and MVTrans with the KeyPose real world dataset for RGB and RGB-D based 6-DoF pose estimation. All models are trained on the KeyPose dataset. DenseFusion is trained twice using raw sensor depth and ground truth depth respectively. Both 2 view and 5 view MVTrans significantly outperform all baseline methods.

	modality	AUC (↑)	$< 2cm$ (↑)	MAE (↓)
DenseFusion [39]	RGB-D (truth depth)	71.9	37.5	35.1
DenseFusion [39]	RGB-D (raw depth)	63.8	18.9	37.2
SimNet[5]	stereo RGB	87.9	83.1	12.6
KeyPose [1]	stereo RGB	90.0	90.1	9.9
MVTrans (Ours)	2-view RGB	**92.7**	**93.6**	**7.4**
MVTrans (Ours)	5-view RGB	**92.9**	**94.0**	**7.2**

simulate occlusion and diversity. The objects are further diversified by randomized scale and orientation. To further mimic a real-life setting, we provide samples in which vessels are filled with varying colors and transparencies of liquid.

Annotations. The dataset contains a diverse set of annotations and saved scene files. We provide annotations for 57 sets of viewing angles for each scene, where the views are spaced in a grid of an upper-half sphere with random radius. Each view consists of a stereo image pair, and the following annotations are provided for the left image: 2D & 3D bounding box, object pose, furthest point sampled keypoints, instance segmentation, depth, surface normal, and object centroid heatmap.

Dataset Statistics. As shown in Table I, Syn-TODD consists of 113,772 stereo image pairs of 1996 different scenes containing a combination of 9012 unique opaque objects from ShapeNet [38] and 7010 unique procedural generated transparent objects. Syn-TODD is split into training & validation sets, with 1575 and 421 scenes each.

V. Results

In this section, we analyze the performance of the proposed multi-task model on tasks of depth prediction, segmentation, 3D object detection, and pose estimation. We evaluate and show the robustness of the method against strong baselines using our proposed dataset and a real world dataset [1].

A. Implementation Details

MVTrans and SimNet are trained on four Nvidia A100 GPUs, with a batch size of 8 to 24 based on view count for 70 epochs on both KeyPose and Syn-TODD datasets. We use the Adam optimizer with $\alpha = 0.0006$, $\beta_1 = 0.9$, and $\beta_2 = 0.99$, and weight decay of 1e-4.

B. Metrics

We evaluate MVTrans for all three of its prediction heads: depth estimation, pose and 3D bounding box estimation and instance segmentation.

For **depth prediction**, the standard metrics as described in [2] are followed. The prediction and ground truth arrays are first resized to 144×256 resolution prior to evaluation. Errors are computed using the following metrics, root mean squared error (RMSE), absolute relative difference (REL), and mean absolute error (MAE).

For **6-DoF pose**, Area Under the Curve (AUC), percentage of 3D keypoint errors $< 2cm$ and Mean Absolute Error (MAE). AUC percentage is calculated based on an X-axis range from 0 to 10 cm, where the curve shows the cumulative percentage of errors under that metric value.

For **3D bounding box**, 3D intersection over union (3D IoU) is used to measure box fit, and 3D mean average precision (3D mAP) is calculated using 3D IoU > 0.25 as criteria, which is correlated with grasp success rate as shown in [5].

For **instance segmentation**, intersection over union (IoU) and mean average precision (mAP) are used to evaluate the predicted mask. For mAP, IoU > 0.5 is used as the threshold.

C. Experiment 1: Multi-view and RGB-D Comparisons

For transparent objects, the raw depth captured by commodity RGB-D sensors is incomplete and distorted, which naturally form the depth completion task when using RGB-D based methods. However, stereo and multi-view based models do not rely on depth information, thus is advantageous for transparent object related 3D tasks. To demonstrate the claimed advantage and our MVTrans's pose prediction capability, as shown in Table II, we compare the pose predicted by RGB-D based DenseFusion [39] with MVTrans trained on KeyPose [1] dataset. MVTrans's two and five view versions both outperform DenseFusion, regardless of whether it is trained on raw distorted depth, or ground truth depth.

D. Experiment 2: Multi-view and Stereo Vision Comparisons

We conduct experiments to test MVTrans against and other stereo RGB based methods. First, we focus on the pose estimation task, where MVTrans is trained on the KeyPose [1] dataset and compared with current state-of-the-art networks. The results are listed in Table II. For the baselines, KeyPose [1] and SimNet [5] are both stereo image based models, where

TABLE III: KeyPose Multi-task results. We train SimNet and MVTrans (2/3/5 views) on KeyPose and evaluate their performances in 3D bounding box, 6 DoF pose and segmentation predictions. MVTrans has better performance compared to SimNet on KeyPose for all tasks, for both settings.

	3D Bbox		Pose			Segmentation	
	3D mAP (↑)	3D IoU (↑)	AUC (↑)	$< 2cm$ (↑)	MAE (mm) (↓)	mAP (↑)	IoU (↑)
Real Dataset: Training and Evaluation on KeyPose Dataset							
SimNet[5]	89.40	49.80	87.89	83.14	12.55	99.10	92.40
MVTrans (2 images)	91.20	**61.40**	92.68	93.59	7.37	99.30	92.20
MVTrans (3 images)	90.40	58.30	92.14	92.76	8.00	99.00	90.90
MVTrans (5 images)	**92.20**	60.90	**92.89**	**93.98**	**7.15**	**99.80**	**92.80**

TABLE IV: Syn-TODD Multi-task results. We train SimNet and MVTrans (2/3/5 views) on our Syn-TODD dataset, and evaluate depth, 3D bounding box and segmentation prediction performances. MVTrans has better performance compared to SimNet on Syn-TODD dataset for all tasks.

	Depth			3D Bbox		Segmentation	
	RMSE (↓)	MAE (↓)	REL (↓)	3D mAP (↑)	3D IoU (↑)	mAP (↑)	IoU (↑)
Synthetic Dataset: Training and Evaluation on Syn-TODD Dataset							
SimNet[5]	1.229	1.020	0.975	4.65	34.92	48.21	50.52
MVTrans (2 images)	0.134	0.089	0.135	40.79	45.95	84.94	79.52
MVTrans (3 images)	0.125	0.083	0.125	42.53	46.17	**87.75**	**81.89**
MVTrans (5 images)	**0.124**	**0.080**	**0.117**	**46.99**	**48.44**	87.24	81.30

KeyPose predicts object pose by keypoints, and SimNet is a multitask model with a pose estimation branch. MVTrans takes two views of the scene as input to match the amount of information received by RGB-D and stereo methods. MVTrans demonstrates a significant advantage in all three pose estimation metrics when compared to all baselines.

E. Experiment 3: Evaluating Multi-task Performance

We conduct experiments to study MVTrans's multi-task capabilities when compared against the previous SOTA method [5] across depth estimation, 3D orientated bounding box or pose estimation, and segmentation tasks. Our evaluations are done on two datasets, namely KeyPose [1], and Syn-TODD. For each experiment, all models are trained and evaluated on the same corresponding dataset. For simple scenes, including single object scenes from KeyPose, as shown in Table III, MVTrans outperforms SimNet [5] with all experimented view counts, including two views. This shows the advantage of multi-view over stereo vision without increasing view count. For complex settings, including multiple objects from Syn-TODD, as shown in Table IV, MVTrans has better performance compared to SimNet by a significant margin. However, 3D bounding box and pose estimation for complex scenes and novel objects presented in Syn-TODD remains challenging given lower results for all methods, in comparison to experimental results on the KeyPose dataset.

F. Ablation Study

To quantitatively evaluate the performance gain from increasing image views, MVTrans with view counts of two, three, and five are trained. We present the evaluation of their performances on both Keypose and Syn-TODD datasets, in Table III and Table IV, respectively. For both datasets, MVTrans has superior performance over the baseline stereo SimNet method, and increasing view count generally yields improved results, especially for harder tasks, for example, the 3D bounding box estimation result for the complex Syn-TODD dataset. For simpler tasks like segmentation, we see that increasing the view count leads to marginal improvement, the reason for this is because segmentation is a 2D task, and hence will not benefit as much from the richer 3D information that additional views provide. Overall, we see that there is still room for improvement for all metrics, which reveals the challenging nature of the dataset.

VI. Conclusion

In this work, we proposed a large-scale photo-realistic multiview dataset, Syn-TODD, for pre-training multiview networks, in addition to a novel end-to-end multiview-based method for multi-task learning, MVTrans. We evaluate the performance of MVTrans on both synthetic and real datasets, including Syn-TODD (Ours) and KeyPose Dataset [1], and observe that we outperform previous baselines by a large margin in depth estimation, segmentation, and scene understanding (3D bounding box and pose estimation). Future directions worth exploring include sim-to-real transfer using the large-scale photo-realistic dataset, and leveraging the perception predictions for scene-graph and/or grasp generation for downstream manipulation and planning tasks. We hope this work can help accelerate future research in household manipulation and laboratory automation.

ACKNOWLEDGMENT

AG is a CIFAR AI Chair. AAG is a CIFAR AI Chair and Lebovic Fellow. AG and FS are also supported in part through the NSERC Discovery Grants Program. The authors would like to acknowledge Vector Institute and Compute Canada for computing services. AAG and HX thank the Canada 150 Research Chair funding from NSERC, Canada. AAG is thankful for the generous support of Dr. Anders G. Frøseth. The authors would like to thank Kourosh Darvish for constructive feedback and discussions on the manuscript.

REFERENCES

[1] X. Liu, R. Jonschkowski, A. Angelova, and K. Konolige, "Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects," *CoRR*, vol. abs/1912.02805, 2019. [Online]. Available: http://arxiv.org/abs/1912.02805

[2] S. S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, and S. Song, "Cleargrasp: 3d shape estimation of transparent objects for manipulation," *CoRR*, vol. abs/1910.02550, 2019. [Online]. Available: http://arxiv.org/abs/1910.02550

[3] H. Xu, Y. R. Wang, S. Eppel, A. Aspuru-Guzik, F. Shkurti, and A. Garg, "Seeing glass: Joint point cloud and depth completion for transparent objects," 2021. [Online]. Available: https://arxiv.org/abs/2110.00087

[4] L. Zhu, A. Mousavian, Y. Xiang, H. Mazhar, J. van Eenbergen, S. Debnath, and D. Fox, "Rgb-d local implicit function for depth completion of transparent objects," 2021.

[5] T. Kollar, M. Laskey, K. Stone, B. Thananjeyan, and M. Tjersland, "Simnet: Enabling robust unknown object manipulation from pure synthetic data via stereo," in *Proceedings of the 5th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 08–11 Nov 2022, pp. 938–948. [Online]. Available: https://proceedings.mlr.press/v164/kollar22a.html

[6] X. Chen, H. Zhang, Z. Yu, A. Opipari, and O. C. Jenkins, "Clearpose: Large-scale transparent object dataset and benchmark," 2022. [Online]. Available: https://arxiv.org/abs/2203.03890

[7] E. Xie, W. Wang, W. Wang, M. Ding, C. Shen, and P. Luo, "Segmenting transparent objects in the wild," 2020. [Online]. Available: https://arxiv.org/abs/2003.13948

[8] X. Long, L. Liu, W. Li, C. Theobalt, and W. Wang, "Multi-view depth estimation using epipolar spatio-temporal network," *CoRR*, vol. abs/2011.13118, 2020. [Online]. Available: https://arxiv.org/abs/2011.13118

[9] H. Afzal, D. Aouada, D. Font, B. Mirbach, and B. Ottersten, "Rgb-d multi-view system calibration for full 3d scene reconstruction," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 2459–2464.

[10] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *ECCV*, 2018.

[11] K. McHenry, J. Ponce, and D. Forsyth, "Finding glass," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2, 2005, pp. 973–979 vol. 2.

[12] K. McHenry and J. Ponce, "A geodesic active contour framework for finding glass," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, 2006, pp. 1038–1044.

[13] E. Xie, W. Wang, W. Wang, M. Ding, C. Shen, and P. Luo, "Segmenting transparent objects in the wild," 2020. [Online]. Available: https://arxiv.org/abs/2003.13948

[14] Y. Cao, Z. Zhang, E. Xie, Q. Hou, K. Zhao, X. Luo, and J. Tuo, "Fakemix augmentation improves transparent object detection," 2021. [Online]. Available: https://arxiv.org/abs/2103.13279

[15] H. He, X. Li, G. Cheng, J. Shi, Y. Tong, G. Meng, V. Prinet, and L. Weng, "Enhanced boundary learning for glass-like object segmentation," *CoRR*, vol. abs/2103.15734, 2021. [Online]. Available: https://arxiv.org/abs/2103.15734

[16] J. Zhang, K. Yang, A. Constantinescu, K. Peng, K. Müller, and R. Stiefelhagen, "Trans4trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world," *CoRR*, vol. abs/2107.03172, 2021. [Online]. Available: https://arxiv.org/abs/2107.03172

[17] E. Xie, W. Wang, W. Wang, P. Sun, H. Xu, D. Liang, and P. Luo, "Trans2seg: Transparent object segmentation with transformer," *CoRR*, vol. abs/2101.08461, 2021. [Online]. Available: https://arxiv.org/abs/2101.08461

[18] Y. Zhang and T. Funkhouser, "Deep depth completion of a single rgb-d image," 2018. [Online]. Available: https://arxiv.org/abs/1803.09326

[19] D. Senushkin, M. Romanov, I. Belikov, A. Konushin, and N. Patakin, "Decoder modulation for indoor depth completion," 2021.

[20] A. Noguchi and T. Harada, "RGBD-GAN: unsupervised 3d representation learning from natural image datasets via RGBD image synthesis," *CoRR*, vol. abs/1909.12573, 2019. [Online]. Available: http://arxiv.org/abs/1909.12573

[21] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3338–3347, 2019.

[22] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun, "Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 629–11 638, 2020.

[23] K. Wada, E. Sucar, S. James, D. Lenton, and A. J. Davison, "Morefusion: Multi-object reasoning for 6d pose estimation from volumetric fusion," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14 528–14 537, 2020.

[24] X. Long, L. Liu, W. Li, C. Theobalt, and W. Wang, "Multi-view depth estimation using epipolar spatio-temporal networks," 2021.

[25] J.-R. Chang and Y. Chen, "Pyramid stereo matching network," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5410–5418, 2018.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[27] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "The best of both modes: Separately leveraging rgb and depth for unseen object instance segmentation," in *CoRL*, 2019.

[28] A. Kirillov, R. B. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6392–6401, 2019.

[29] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A survey of embodied ai: From simulators to research tasks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 2, pp. 230–244, 2022.

[30] A. Dasgupta, J. Duan, M. H. Ang Jr, and C. Tan, "Avoe: a synthetic 3d dataset on understanding violation of expectation for artificial cognition," *arXiv preprint arXiv:2110.05836*, 2021.

[31] C. Eppner, A. Mousavian, and D. Fox, "Acronym: A large-scale grasp dataset based on simulation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 6222–6227.

[32] J. Tremblay, T. To, and S. Birchfield, "Falling things: A synthetic dataset for 3d object detection and pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2038–2041.

[33] J. Duan, S. Yu, and C. Tan, "Space: A simulator for physical interactions and causal learning in 3d environments," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2058–2063.

[34] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2901–2910.

[35] F. Baradel, N. Neverova, J. Mille, G. Mori, and C. Wolf, "Cophy: Counterfactual learning of physical dynamics," *arXiv preprint arXiv:1909.12000*, 2019.

[36] B. O. Community, "Blender - a 3d modelling and rendering package," Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: http://www.blender.org

[37] S. Eppel, H. Xu, Y. R. Wang, and A. Aspuru-Guzik, "Predicting 3d shapes, masks, and properties of materials, liquids, and objects inside transparent containers, using the transproteus cgi dataset," 2021. [Online]. Available: https://arxiv.org/abs/2109.07577

[38] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "Shapenet: An information-rich 3d model repository," 2015. [Online]. Available: https://arxiv.org/abs/1512.03012

[39] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," *CoRR*, vol. abs/1901.04780, 2019. [Online]. Available: http://arxiv.org/abs/1901.04780