

TransNet: Transparent Object Manipulation Through Category-Level Pose Estimation

Huijie Zhang¹ Anthony Opipari¹ Xiaotong Chen¹ Jiyue Zhu¹ Zeren Yu² Odest Chadwicke Jenkins¹

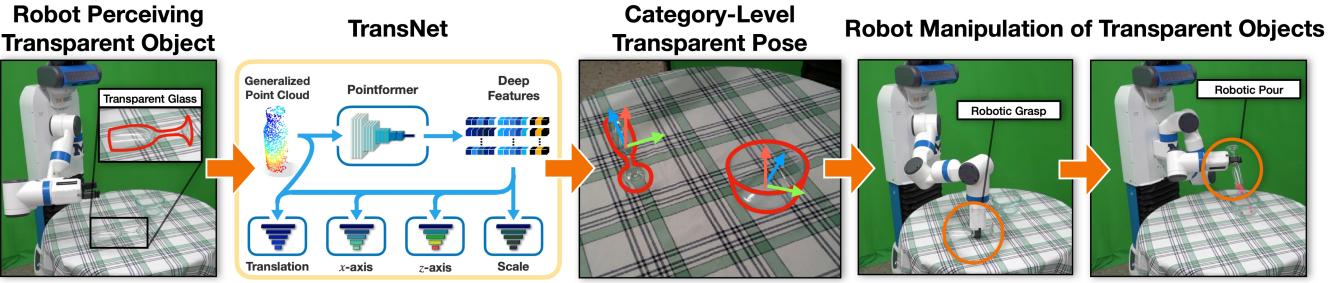


Fig. 1. TransNet is a pipeline for category-level transparent object pose estimation and manipulation. Given an RGB-D image of a scene containing transparent objects, TransNet estimates the 6D pose and 3D scale of each object in the scene for use in downstream manipulation tasks such as grasping and pouring.

Abstract—Transparent objects present multiple distinct challenges to visual perception systems. First, their lack of distinguishing visual features makes transparent objects harder to detect and localize than opaque objects. Even humans find certain transparent surfaces with little specular reflection or refraction, like glass doors, difficult to perceive. A second challenge is that depth sensors typically used for opaque object perception cannot obtain accurate depth measurements on transparent surfaces due to their unique reflective properties. Stemming from these challenges, we observe that transparent object instances within the same category, such as cups, look more similar to each other than to ordinary opaque objects of that same category. Given this observation, the present paper explores the possibility of category-level transparent object pose estimation rather than instance-level pose estimation. We propose *TransNet*, a two-stage pipeline that estimates category-level transparent object pose using **localized depth completion and surface normal estimation**. TransNet is evaluated in terms of pose estimation accuracy on a large-scale transparent object dataset and compared to a state-of-the-art category-level pose estimation approach. Results from this comparison demonstrate that TransNet achieves improved pose estimation accuracy on transparent objects. Moreover, we use TransNet to build an autonomous transparent object manipulation system for robotic pick-and-place and pouring tasks. The project webpage is available at <https://progress.eecs.umich.edu/projects/transnet/>.

I. INTRODUCTION

From glass doors and windows to kitchenware and all kinds of containers, objects with transparent materials are prevalent throughout daily life. Thus, perceiving the pose (position and orientation) of transparent objects is a crucial capability for autonomous perception systems seeking to interact with their environment. However, transparent objects present unique perception challenges both in the RGB and depth domains.

For RGB, the color appearance of transparent objects is highly dependent on the background, viewing angle, material, and lighting condition, due to light reflection and refraction effects. For depth, common commercially available depth sensors record mostly invalid or inaccurate depth values within the transparent region. Such visual challenges, especially missing detection in the depth domain, pose severe challenges for autonomous object manipulation and obstacle avoidance tasks. This paper addresses these challenges by studying how category-level transparent object pose estimation may be achieved using end-to-end learning.

Recent works have shown promising results on grasping transparent objects by completing the missing depth values followed by the use of a geometry-based grasp engine [1], [2], [3], transfer learning from RGB-based grasping neural networks [4], light-field feature learning [5], or domain-randomized depth noise simulation [6]. For more advanced manipulation tasks such as rigid body pick-and-place or liquid pouring, geometry-based estimations, such as symmetrical axes, edges [7] or object poses [8], [9], [6], are required to model the manipulation trajectories. Instance-level transparent object poses could be estimated from keypoints on stereo RGB images [10], [11], a light-field camera [12], [13], or directly from a single RGB-D image [9] with support plane assumptions. Recently emerged large-scale transparent object datasets [1], [14], [11], [3], [15] pave the way for addressing the problem using deep learning.

In this paper, we set out to extend the frontier of 3D transparent object perception by building upon recent work. We introduce *TransNet*, a category-level pose estimation pipeline for transparent objects that outperforms a state-of-the-art baseline. We further explore the effect of input modalities, feature embedding methods, and depth-normal consistency through the learning process to provide insights for future research. Moreover, we build an autonomous transparent object manipulation system using TransNet and

¹ H. Zhang, A. Opipari, X. Chen, J. Zhu, and O. Jenkins are with the Electrical Engineering and Computer Science, Mechanical Engineering, and Robotics departments, University of Michigan, Ann Arbor, MI, USA.

² Z. Yu is with Department of Computer Science, University of Southern California, Los Angeles, CA, USA.

demonstrate its efficacy in pick-place and pouring tasks.

II. RELATED WORK

A. Transparent Object Visual Perception for Manipulation

Transparent objects must be perceived before they can be acted on. Lai *et al.* [16] and Khaing *et al.* [17] developed CNN models to detect transparent objects from RGB images. Xie *et al.* [18] proposed a deep segmentation model that achieved state-of-the-art accuracy. ClearGrasp [1] employed depth completion for use with pose estimation on robotic grasping tasks, where they trained three DeepLabv3+ [19] models to learn transparency mask, surface normal, and boundary, respectively. Follow-on studies developed different approaches for depth completion, including implicit functions [20], NeRF reconstruction [2], combined point cloud and depth features [14], adversarial learning [21], multi-view geometry [22], RGB image completion [3], and sim2real transfer [6]. Weng *et al.* [4] used transfer learning from the RGB to the depth sensor domain without completing raw depth. For instance-level pose estimation, Xu *et al.* [9] utilized segmentation, surface normal, and image coordinate UV-map as input to a network similar to [23] that can estimate 6 DOF object pose. Keypose [10] proposed to estimate 2D keypoints and regress object poses from stereo images using triangulation. For other special sensors, Xu *et al.* [24] used light-field images to perform segmentation using a graph-cut-based approach. Kalra *et al.* [25] trained Mask R-CNN [26] using polarization images as input to outperform the baseline that was trained on only RGB images by a large margin. Zhou *et al.* [12], [5], [13] employed light-field images to learn features for robotic grasping and object pose estimation. Along with the proposed methods, massive datasets, across different sensors and both synthetic and real-world domains, have been collected and made public for various related tasks [18], [1], [10], [13], [25], [11], [20], [14], [3], [15]. Compared with these previous works, and to the best of the authors' knowledge, we propose the first category-level pose (6D pose + 3D scale) estimation approach developed specifically for transparent objects.

B. Opaque Object Category-level Pose Estimation

Category-level object pose estimation aims to estimate unseen objects' 6D pose within seen categories and their scale or canonical shape. Wang *et al.* [27] introduced the Normalized Object Coordinate Space (NOCS) for dense 3D correspondence learning and solved for object pose and scale using the learned NOCS map and masked depth [28]. More recent works explored different aspects to improve pose estimation accuracy. A category-level shape prior is found to be beneficial for pose estimation accuracy in [29] and further improved in [30], [31], [32]. DualPoseNet [33], 6D-ViT [34], ACR-Pose [35], and CPPF [36] proposed to incorporate rotation-invariant embedding, Transformer networks, Generative Adversarial Networks, and deep point-pair-feature, respectively. However, these techniques require high-quality depth input provided by opaque objects with Lambertian light reflectance. One recent work by Dai *et*

al. [6] proposed a data generation system that simulates the noise on non-Lambertian surfaces of active stereo depth cameras and demonstrated its usage in category-level pose estimation and robotic grasping.

Compared with these techniques, TransNet takes advantage of recent advances in transparent depth completion approaches and directly learns category-level pose estimation from imperfect depth predictions, carefully considering multi-modal input, feature embedding architecture, and depth-normal consistency.

III. TRANSNET

Given an input **RGB-D pair** $(\mathcal{I}, \mathcal{D})$ viewing a scene that includes transparent objects, our goal is to predict the objects' 6D rigid body transformations $[\mathbf{R}|\mathbf{t}]$, 3D scales \mathbf{s} in the camera coordinate frame and object classes within a predefined category set, where $\mathbf{R} \in SO(3), \mathbf{t} \in \mathbb{R}^3$ and $\mathbf{s} \in \mathbb{R}_+^3$. In this problem, inaccurate or invalid depth readings exist within the image region of transparent objects (represented as a binary mask). To solve the category-level pose estimation problem, we propose a novel two-stage deep neural network pipeline, *TransNet*.

A. Architecture Overview

TransNet first applies a fine-tuned instance segmentation module to extract individual object patches from the overall image. As shown in Figure 2, in the first stage, TransNet takes the patches as input and extracts multi-modal features corresponding to each transparent object instance. Depth completion and surface normal estimation are applied on RGB-D patches to obtain depth-normal pairs through cross-task consistency learning. The estimated depth-normal pairs, RGB, and ray direction patches are concatenated to feature patches, followed by a random sampling strategy within the instance masks to generate features referred to as the **generalized point cloud**. In the second stage, the generalized point cloud is processed through Pointformer [34] to produce embedded feature vectors. The pose is then separately estimated in four decoder modules for object translation, x -axis, z -axis, and scale, respectively.

B. Object Instance Detection and Segmentation

Similar to other category-level pose estimation work [32], we fine-tune a Mask R-CNN [26] model to obtain the object's **bounding box** \mathcal{B} , **segmentation mask** \mathcal{M} and **category label** \mathcal{P}_c . Patches of ray direction $\mathcal{R}_{\mathcal{B}}$, RGB $\mathcal{I}_{\mathcal{B}}$ and raw depth $\mathcal{D}_{\mathcal{B}}$ are extracted according to the bounding box \mathcal{B} and serve as input to the first stage of TransNet. The UV mapping itself is an important cue when estimating poses from patches [37], as it provides information about the relative position and size of the patches within the overall image. We use ray direction instead of UV mapping because it also contains camera intrinsic information. Here the ray direction \mathcal{R} represents the direction from camera origin to each pixel in the camera

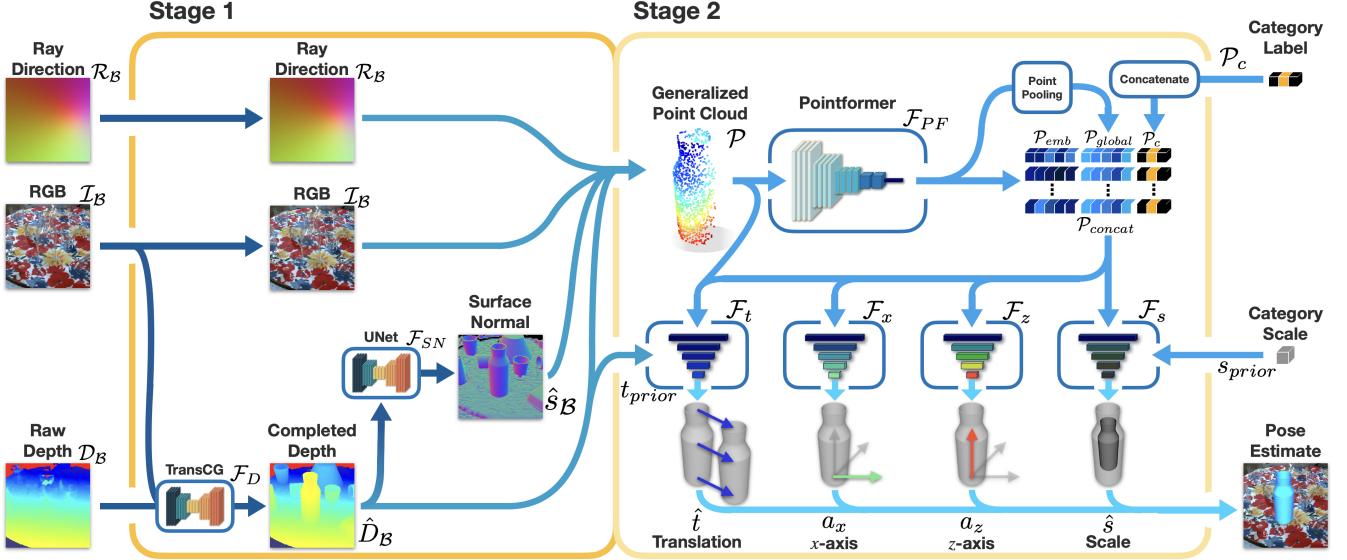


Fig. 2. The two-stage architecture of TransNet. Image patches of RGB, ray direction, surface normal, and completed depth are concatenated and embedded to learn the translation, x-axis, z-axis, and scale separately of the estimated category-level pose.

plane under the camera frame. For each pixel (u, v) :

$$p = [u \ v \ 1]^T$$

$$\mathcal{R} = \frac{K^{-1}p}{\|K^{-1}p\|^2} \quad (1)$$

where p is the homogeneous UV coordinate in the image plane and K is the camera intrinsic matrix.

C. Cross-task Consistency for Depth-Normal Pairs

We apply a recent state-of-the-art approach, TransCG [3] for depth completion (\mathcal{F}_D) and U-Net [38] for surface normal estimation (\mathcal{F}_{SN}).

$$\hat{\mathcal{D}}_{\mathcal{B}} = \mathcal{F}_D(\mathcal{I}_{\mathcal{B}}, \mathcal{D}_{\mathcal{B}})$$

$$\hat{\mathcal{S}}_{\mathcal{B}} = \mathcal{F}_{SN}(\hat{\mathcal{D}}_{\mathcal{B}}) \quad (2)$$

where the depth estimation $\hat{\mathcal{D}}_{\mathcal{B}}$ is used as input to the surface normal estimation module. The two models are first trained separately with depth and surface normal losses, \mathcal{L}_d and \mathcal{L}_s , and then trained together using cross-task consistency. Following [39], we design a consistency loss \mathcal{L}_{con} to train both networks. Both \mathcal{L}_d and \mathcal{L}_s are implemented using L_2 loss for depth completion and surface normal estimation within transparent mask \mathcal{M} .

$$\mathcal{L}_d = \frac{1}{N_p} \sum_{p \in \mathcal{M}} \left\| \hat{\mathcal{D}}_p - \mathcal{D}_p^* \right\|^2$$

$$\mathcal{L}_s = \frac{1}{N_p} \sum_{p \in \mathcal{M}} \left\| \hat{\mathcal{S}}_p - \mathcal{S}_p^* \right\|^2$$

$$\mathcal{L}_{con} = \frac{1}{N_p} \sum_{p \in \mathcal{M}} \left\| \hat{\mathcal{S}}_p - \mathcal{F}_{SN}(\mathcal{D}_{\mathcal{B}}^*)_p \right\|^2 \quad (3)$$

where \mathcal{D}^* and \mathcal{S}^* are the ground truth depth and surface normal images, and N_p refers to the number of masked pixels, respectively. During training, these losses are used according

to the perceptual loss training strategy described by Zamir et al. [39]. Following cross-task consistency training, the depth completion and surface normal estimation networks are frozen and used to generate input for the second stage of TransNet.

D. Generalized Point Cloud

The generalized point cloud is implemented as a concatenation of multi-modal input features: $\mathcal{I}_{\mathcal{B}}$, $\mathcal{R}_{\mathcal{B}}$, $\hat{\mathcal{D}}_{\mathcal{B}}$, and $\hat{\mathcal{S}}_{\mathcal{B}}$. Specifically, we randomly sample N pixels within the transparent mask \mathcal{M} of the feature patch to obtain a generalized point cloud $\mathcal{P} \in \mathbb{R}^{N \times 10}$. The experiment described in Section IV-B explores the best choice of features for use in the generalized point cloud.

E. Transformer Feature embedding

Given a generalized point cloud \mathcal{P} corresponding to a detected object, TransNet employs Pointformer [34], a multi-stage transformer-based point cloud embedding method to encode object features $\mathcal{P}_{emb} \in \mathbb{R}^{N \times d_{emb}}$.

$$\mathcal{P}_{emb} = \mathcal{F}_{PF}(\mathcal{P}) \quad (4)$$

Experiments described in Section IV-B evaluate the effectiveness of using alternative point cloud embedding methods, such as 3D-GCN [40], for use in the transparent object context with noisy depth data.

Next, a Point Pooling layer (a multilayer perceptron (MLP) with max-pooling) is applied to extract global features \mathcal{P}_{global} . The global features are then concatenated with local features \mathcal{P}_{emb} and a one-hot category \mathcal{P}_c label from the instance segmentation network. The concatenations result in a feature vector to be used as input to the decoder:

$$\mathcal{P}_{global} = \text{MaxPool}(\text{MLP}(\mathcal{P}_{emb}))$$

$$\mathcal{P}_{concat} = [\mathcal{P}_{emb}, \mathcal{P}_{global}, \mathcal{P}_c] \quad (5)$$

F. Pose and Scale Estimation

After the feature embeddings are extracted based on multi-modal input, four separate decoders are applied to estimate the object’s translation, x -axis, z -axis, and scale. Similar to [31].

Translation Residual Estimation The translation decoder \mathcal{F}_t learns a 3D translation residual from the object translation prior t_{prior} calculated as the average of predicted 3D coordinate over the sampled pixels in \mathcal{P} .

$$t_{prior} = \frac{1}{N_p} \sum_{p \in N} K^{-1} [u_p \ v_p \ 1]^T \hat{\mathcal{D}}_p \quad (6)$$

$$\hat{t} = t_{prior} + \mathcal{F}_t([\mathcal{P}_{concat}, \mathcal{P}])$$

where u_p, v_p are the 2D pixel coordinate for the selected pixel. We use the L_1 loss between the ground truth and estimated position:

$$\mathcal{L}_t = |\hat{t} - t^*| \quad (7)$$

Rotation Estimation We decouple the 3D rotation matrix R into two orthogonal axes, x -axis a_x and z -axis a_z , and estimate them separately. The network learns confidence values to deal with the problem that the regressed two axes are not orthogonal:

$$\begin{aligned} [\hat{a}_i, c_i] &= \mathcal{F}_i(\mathcal{P}_{concat}), \ i \in \{x, z\} \\ \theta_z &= \frac{c_x}{c_x + c_z} \left(\theta - \frac{\pi}{2} \right) \\ \theta_x &= \frac{c_z}{c_x + c_z} \left(\theta - \frac{\pi}{2} \right) \end{aligned} \quad (8)$$

where c_x, c_z denotes the confidence for the learned axes. θ represents the angle between a_x and a_z . θ_x, θ_z are obtained by solving an optimization problem and then used to rotate a_x and a_z within their common plane. For the training loss, we use L_1 plus cosine similarity loss $\mathcal{L}_{r_x}, \mathcal{L}_{r_z}$ for each individual axis. Angular loss, \mathcal{L}_a , is also used to constrain a perpendicular relationship between the two axes and confidence loss $\mathcal{L}_{con_x}, \mathcal{L}_{con_z}$ to learn axis confidence:

$$\begin{aligned} \mathcal{L}_{r_i} &= |\hat{a}_i - a_i^*| + 1 - \langle \hat{a}_i, a_i^* \rangle, \ i \in \{x, z\} \\ \mathcal{L}_a &= \langle \hat{a}_x, \hat{a}_z \rangle \end{aligned} \quad (9)$$

$$\mathcal{L}_{con_i} = |c_i - \exp(\alpha \|\hat{a}_i - a_i^*\|_2)|, \ i \in \{x, z\}$$

where α is a constant to scale the distance. Thus the overall loss for the second stage is:

$$\begin{aligned} \mathcal{L} &= \lambda_s \mathcal{L}_s + \lambda_t \mathcal{L}_t + \lambda_{r_x} \mathcal{L}_{r_x} + \lambda_{r_z} \mathcal{L}_{r_z} + \\ &\quad \lambda_{r_a} \mathcal{L}_a + \lambda_{con_x} \mathcal{L}_{con_x} + \lambda_{con_z} \mathcal{L}_{con_z} \end{aligned} \quad (10)$$

where $\lambda_{r_x}, \lambda_{r_z}, \lambda_{r_a}, \lambda_t, \lambda_s, \lambda_{con_x}, \lambda_{con_z}$ are weights for each loss.

Scale Residual Estimation Similar to the translation decoder, we define the scale prior s_{prior} as the average of scales of all object 3D CAD models within each category. Then the scale of a given instance is calculated as follows:

$$\hat{s} = s_{prior} + \mathcal{F}_s(\mathcal{P}_{concat}) \quad (11)$$

The loss function is defined as the L_1 loss between the ground truth scale and estimated scale:

$$\mathcal{L}_s = |\hat{s} - s^*| \quad (12)$$

IV. EXPERIMENTS

Dataset We evaluated TransNet and baseline models on the Clearpose Dataset [15] for depth completion, surface normal estimation, and categorical transparent object pose estimation tasks. As shown in Figure 3, 33 object instances from 3 ClearPose categories (*bowl*, *water_cup*, and *wine_cup*) were selected for evaluating the relevant models on each task. 25 of the objects are used in the training set, totaling 190K RGB-D images. Hence, 8 objects not seen during training are used for testing, with 60K test images available. In practice, we train the models using the full 190K image training set and uniformly sample images of the test object to form a representative 5K image test set. For the robot manipulation experiments, we used 1 *bowl* object and 1 *wine_cup* object from the test set along with another 1 new *wine_cup* and 2 new *water_cups* purchased from the market that fit the robot parallel gripper.



Fig. 3. Transparent objects used in experiments. The image on the left show the 25 objects in the training set. The image on the right shows 8 novel objects in the test set for vision evaluation and robotic experiments.

Implementation Details Our model was trained in several stages. For experiments included in the ablation study and baseline comparison, ground truth instance segmentations are used to generate input patches, while for the robot experiments, a Mask R-CNN [26] model fine-tuned on the Clearpose dataset is used for generating input patches. The image patches were generated from object bounding boxes and re-scaled to a fixed shape of 256×256 pixels. Cross-task consistency training is performed using the AdamW optimizer [41] with learning rate of $1e^{-3}$. The pretrained surface normal estimation model provided by Zamir *et al.* [39] is used for initial network parameters. Cross-task training hyperparameters follow the perceptual loss strategy described in [39] until convergence using a batch size of 8. For TransCG, the AdamW optimizer [41] is used for training with $\lambda_{smooth} = 0.001$ and the overall learning rate is 0.001 to train the model until converge. For U-Net, we used the Adam optimizer [42] with a learning rate of $1e^{-4}$ to train the model until convergence. For both surface normal estimation and depth completion, the batch size was set to 24. For the second stage, the training hyperparameters of Pointformer and pose and scale estimation were selected following [34], [32]. The learning rate for all loss terms were kept the same during training, $\{\lambda_{r_x}, \lambda_{r_z}, \lambda_{r_a}, \lambda_t, \lambda_s, \lambda_{con_x}, \lambda_{con_z}\} = \{8, 8, 4, 8, 8, 1, 1\} e^{-4}$. The Ranger optimizer [43], [44], [45] was used with a linear warm-up for the first 1000 iterations,

then a cosine annealing method at the 0.72 anneal point was used. All the experiments for pose estimation were trained on a 16G RTX3080 GPU until loss convergence.

Evaluation metrics For category-level pose estimation, this study follows [32], [31] in using 3D intersection over union (IoU) between the ground truth and estimated 3D bounding box at 25%, 50% and 75% thresholds. Additionally, $5^{\circ}5cm$, $10^{\circ}5cm$, and $10^{\circ}10cm$ are used as metrics. The reported measurements for each of these three metrics represent the percentage of a model’s pose estimates with error less than the specified metric’s degree and distance thresholds.

For depth completion evaluation, the root of mean squared error (RMSE), absolute relative error (REL) and mean absolute error (MAE) are used as metrics along with $\delta_{1.05}$, $\delta_{1.10}$, $\delta_{1.25}$, where δ_n is calculated as:

$$\delta_n = \frac{1}{N_p} \sum_p \mathbf{I} \left(\max \left(\frac{\hat{D}_p}{D_p^*}, \frac{D_p^*}{\hat{D}_p} \right) < n \right) \quad (13)$$

where $\mathbf{I}(\cdot)$ represents the indicator function. \hat{D}_p and D_p^* represent estimated and ground truth depth for each pixel p .

For surface normal estimation, RMSE and MAE errors of normalized vectors are used as metrics with thresholds of 11.25° , 22.5° , and 30° . Here, measurements reported for the 11.25° metric represent the percentage of estimates with an angular distance less than 11.25° from the ground truth surface normals. The mean angular error in degrees (MEAN) is also reported.

A. Comparison with Baseline

We chose one state-of-the-art categorical opaque object pose estimation model (GPV-Pose [32]) as a baseline, which was trained with estimated depth from TransCG [3] for a fair comparison. From Table I, TransNet outperformed the baseline in most of the metrics on the Clearpose dataset. The $3D_{25}$ metric is relatively easy to achieve strong performance, so there is no substantial difference between the two models’ corresponding $3D_{25}$ performance. When compared to the baseline on the remaining metrics, TransNet achieved around a $3 \times$ improvement on $3D_{75}$, $3.5 \times$ on $10^{\circ}5cm$, $10^{\circ}10cm$ and $8 \times$ on $5^{\circ}5cm$. Qualitative results are shown in Figure 4 for TransNet.

TABLE I

COMPARISON WITH THE BASELINE ON THE CLEARPOSE DATASET.

Method	$ 3D_{25} $	$ 3D_{50} $	$ 3D_{75} $	$ 5^{\circ}5cm $	$ 10^{\circ}5cm $	$ 10^{\circ}10cm $
GPV-Pose	95.4	65.1	13.2	2.7	12.5	15.5
TransNet	97.3	80.3	39.9	22.7	45.4	50.6

B. Ablation Study

In Table II, we compared the performance of TransNet to that of modified versions of the network architecture. The results of this study informed Trial 6 as the final architecture of TransNet and the one that was used in robot experiments. Results from the ablation study may also be used to guide future work in the area of transparent object pose estimation.



Fig. 4. Qualitative results of category-level pose estimates from TransNet. White bounding boxes depict ground truth pose annotations, while colored boxes depict TransNet’s estimates. Different colors represent different categories. The top two rows show accurate estimates, while the bottom row highlights inaccurate estimates when TransNet is confronted with heavily cluttered scenes. Objects with no bounding box correspond to categories not used during training and evaluation. For axial symmetric objects, we use the ground truth x-axis to produce a 3D bounding box.

Embedding Method Between trial 1 and trial 2, we compared the effect of the embedding method, 3D-GCN [40] and Pointformer [34], on TransNet accuracy. The use of Pointformer resulted in higher accuracy for TransNet on most metrics compared to when 3D-GCN was used. More details are shown in Table III. Crucially, when the ground truth depth was changed to estimated depth (modeling the change from opaque to transparent setting), Pointformer retained substantially more accuracy than 3D-GCN. A possible explanation for this result is based on how 3D-GCN propagates information between nearest neighbors. While this is an efficient method for propagating information locally, it suffered when given the noisy completed depth as input. Pointformer, on the other hand, shares information across the whole point cloud, possibly contributing to increased robustness when input data is noisy. Therefore, given depth with large uncertainty, the transformer-based embedding method might be more powerful than embedding methods using nearest neighbors.

Modalities for Generalized Point Cloud The comparison of trials 2–4 was used to identify the optimal combination of feature inputs for the generalized point cloud in TransNet. For trials 2 and 4, we compared the effect of adding the estimated surface normal to the generalized point cloud. Results from this comparison imply that surface normals serve as a good complement when depth estimation is not accurate, as in the transparent setting. For trials 3 and 4,

TABLE II
ABLATION STUDY FOR TRANSNET

Trial	Embedding	Generalized Point Cloud			Consistency	Separate Category	$3D_{25} \uparrow$	$3D_{50} \uparrow$	$3D_{75} \uparrow$	$5^{\circ}5cm \uparrow$	$10^{\circ}5cm \uparrow$	$10^{\circ}10cm \uparrow$
	PF 3D-GCN	RGB	Depth	Normal Ray-direction								
1	✓	✓	✓	✓			98.3	72.7	17.6	6.6	22.5	28.0
2	✓	✓	✓	✓			97.9	76.9	28.4	9.4	8.0	29.6
3	✓	✓	✓	✓			95.4	61.6	11.6	3.6	15.3	31.3
4	✓	✓	✓	✓			98.4	81.6	34.4	13.0	38.6	45.1
5	✓	✓	✓	✓	✓		98.1	81.8	39.2	15.8	41.4	46.1
6	✓	✓	✓	✓	✓	✓	97.3	80.3	39.9	22.7	45.4	50.6

TABLE III
COMPARISON BETWEEN DIFFERENT EMBEDDING METHODS

Depth type	Embedding	$3D_{25} \uparrow$	$3D_{50} \uparrow$	$3D_{75} \uparrow$	$5^{\circ}5cm \uparrow$	$10^{\circ}5cm \uparrow$	$10^{\circ}10cm \uparrow$
Ground truth	3D-GCN	99.8	96.6	55.4	70.9	86.0	90.3
	Pointformer	99.7	96.6	74.5	61.9	83.7	86.5
Estimation	3D-GCN	98.3	72.7	17.6	6.6	22.5	28.0
	Pointformer	97.9	76.9	28.4	9.4	29.6	35.3

TABLE IV
ACCURACY FOR DEPTH-NORMAL PAIR ESTIMATION

Depth Metrics	RMSE \downarrow	REL \downarrow	MAE \downarrow	$\delta_{1.05} \uparrow$	$\delta_{1.10} \uparrow$	$\delta_{1.25} \uparrow$
w/o consistency	0.056	0.044	0.041	68.61	89.35	98.89
w consistency	0.057	0.044	0.041	70.49	89.47	98.57
Surface Normal Metrics	RMSE \downarrow	MAE \downarrow	MEAN \downarrow	$11.25^{\circ} \uparrow$	$22.5^{\circ} \uparrow$	$30^{\circ} \uparrow$
w/o consistency	0.19	0.13	11.43	56.75	88.45	96.64
w consistency	0.12	0.08	8.96	73.62	92.56	96.58

we compared the potential benefit of including ray-direction. Notably, the ray-direction input contains both camera intrinsic and pixel positional information, which we hypothesize is critical for the network to be able to model accurate 3D spatial relationships. All the metrics showed that including both surface normal and ray-direction results in improved pose estimation accuracy.

Consistent Learning For trials 4 and 5 in Table II, TransNet is trained without and with cross-task consistency, respectively. For trial 4, both surface normal estimation and depth completion networks in TransNet’s first stage are trained separately and without the \mathcal{L}_{con} loss. Results from this comparison show that adding cross-task consistency training improved TransNet performance on high accuracy metrics: $5^{\circ}5cm$, $3D_{75}$. From Table IV, adding a cross-task consistency loss resulted in improved accuracy for depth-normal pair estimation, particularly for the surface normal estimation network. Therefore, cross-task consistent training improves the performance of category-level transparent object pose estimation and is included in TransNet.

Separate Category As shown in Table II when comparing results in trial 5 and trial 6, it was found that TransNet’s accuracy improves when a single instantiation of the model is trained per category as opposed to being trained on multiple categories. Notably, this result is most apparent for the high accuracy metric $5^{\circ}5cm$.

Why TransNet Outperforms GPV-Pose Comparing GPV-Pose with the trial 1 model leads to an observation that both models use 3D-GCN and a similar network architecture. The only difference is that instead of using the generalized point cloud as input, GPV-Pose directly uses a point cloud (the multiplication of depth and ray direction) as input. Yet, trial 1 outperforms GPV-Pose on pose estimation tasks. We

hypothesize that the camera intrinsic and pixel positional information contained in ray-direction degrades in accuracy after being multiplied directly with noisy depth data to form the point cloud input to GPV-Pose. In contrast, TransNet retains ray-direction and depth information as distinct feature channels in the generalized point cloud with learned transformations to combine each feature modality. These learned transformations potentially allow TransNet to be more robust in the face of noisy depth data.

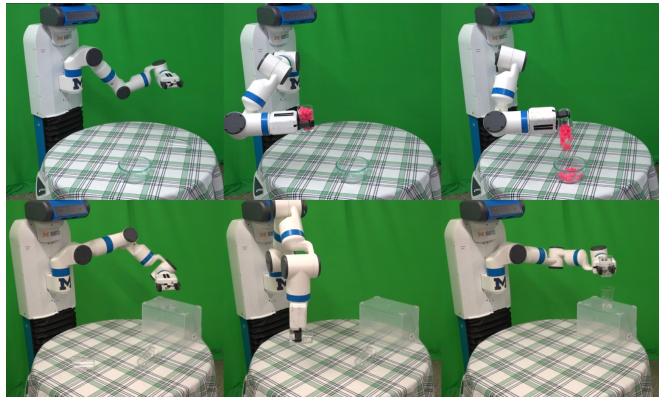


Fig. 5. Robotic experiments. The top row is the demo of pouring using a water cup and a bowl. The bottom row is the demo of pick-and-place using two water cups. The left column is the initial state, the middle column is picking, and the right column is pouring or placing.

C. Robot Experiment

We used a Fetch robot for manipulation and an Intel RealSense L515 camera to take RGB-D images as used in the ClearPose dataset. The robot-camera extrinsic calibration is done using AprilTags [46]. Inspired from [47], we designed two robot manipulation tasks: pouring and pick-and-place to evaluate the efficacy of TransNet to support robot manipulation of transparent objects using 3 *water_cups*, 2 *wine_cups*, and 1 *bowl* not appearing in the training set. For grasping the object, we calculate the approaching direction as orthogonal to the object’s symmetrical axis and the grasp location as the object’s center point.

Pick-and-place This task is composed of a pick action and a place action. For the pick action, the robot is to grasp a cup (either *water_cup* or *wine_cup*) on the table-top with its parallel gripper based on its pose estimate from TransNet. For the place action, the robot is to place the cup upright in a pre-defined location. The target poses for the robot’s gripper are solved by aligning the object’s estimated bottom surface parallel to the target location surface with a 2 cm offset.

TABLE V
ROBOT EXPERIMENT SUCCESS RATE

Task	Objects	#Pick	Success Rate	#Place(Pour)	Success Rate
Pick & Place	water.cup	16/20	80%	12/16	75%
	wine.cup	13/20	65%	12/13	92.3%
Pouring	water.cup & bowl	15/20	75%	10/15	66.7%
	wine.cup & bowl	14/20	70%	11/14	78.6%

Pouring This task is composed of a pick action and a pouring action. The pick action follows the same protocol as *pick-and-place*. For the pouring action, the robot moved the cup above the bowl based on the bowl’s estimated location while also keeping the cup’s z axis upright to prevent spilling. Then the content in the cup is poured into the bowl by tilting the cup’s z axis.

Result The poses and scales used by the robot are output directly from TransNet without additional post-processing. Quantitative results measuring the robot’s success rate are included in Table V. TransNet enabled the robot to perform pour and pick-place actions with promising success rates. Visualization examples are included in Figure 5.

V. CONCLUSIONS

In this paper, we introduced *TransNet*, a two-stage pipeline for category-level transparent object pose estimation that outperformed the baselines. Ablation studies were performed to understand the benefits of using multi-modal input, feature embedding modules, and cross-task consistency. The efficacy of TransNet was demonstrated by enabling a robot system to perform transparent object manipulation tasks with high success rates. In the future, we could further improve the pose estimation accuracy as there is still a large gap between TransNet and state-of-the-art opaque object datasets. One direction is to consider objects’ material information, as we observed a performance gap between perceiving glass and plastic objects.

REFERENCES

- [1] S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, and S. Song, “Clear grasp: 3d shape estimation of transparent objects for manipulation,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3634–3642.
- [2] J. Ichnowski, Y. Avigal, J. Kerr, and K. Goldberg, “Dex-nerf: Using a neural radiance field to grasp transparent objects,” *arXiv preprint arXiv:2110.14217*, 2021.
- [3] H. Fang, H.-S. Fang, S. Xu, and C. Lu, “Transcg: A large-scale real-world dataset for transparent object depth completion and grasping,” *arXiv preprint arXiv:2202.08471*, 2022.
- [4] T. Weng, A. Pallanckize, Y. Tang, O. Kroemer, and D. Held, “Multi-modal transfer learning for grasping transparent and specular objects,” *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3791–3798, 2020.
- [5] Z. Zhou, T. Pan, S. Wu, H. Chang, and O. C. Jenkins, “Glass-loc: Plenoptic grasp pose detection in transparent clutter,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4776–4783.
- [6] Q. Dai, J. Zhang, Q. Li, T. Wu, H. Dong, Z. Liu, P. Tan, and H. Wang, “Domain randomization-enhanced depth simulation and restoration for perceiving and grasping specular and transparent objects,” *arXiv preprint arXiv:2208.03792*, 2022.
- [7] C. J. Phillips, M. Lecce, and K. Daniilidis, “Seeing glassware: from edge detection to pose estimation and shape recovery,” in *Robotics: Science and Systems*, vol. 3, 2016, p. 3.
- [8] I. Lysenkov, V. Eruhimov, and G. Bradski, “Recognition and pose estimation of rigid transparent objects with a kinect sensor,” *Robotics*, vol. 273, no. 273–280, p. 2, 2013.
- [9] C. Xu, J. Chen, M. Yao, J. Zhou, L. Zhang, and Y. Liu, “6dof pose estimation of transparent object from a single rgb-d image,” *Sensors*, vol. 20, no. 23, p. 6790, 2020.
- [10] X. Liu, R. Jonschkowski, A. Angelova, and K. 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*, 2020, pp. 11602–11610.
- [11] X. Liu, S. Iwase, and K. M. Kitani, “Stereobj-1m: Large-scale stereo image dataset for 6d object pose estimation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10870–10879.
- [12] Z. Zhou, Z. Sui, and O. C. Jenkins, “Plenoptic monte carlo object localization for robot grasping under layered translucency,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.
- [13] Z. Zhou, X. Chen, and O. C. Jenkins, “Lit: Light-field inference of transparency for refractive object localization,” *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4548–4555, 2020.
- [14] 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,” *arXiv preprint arXiv:2110.00087*, 2021.
- [15] X. Chen, H. Zhang, Z. Yu, A. Oripipari, and O. C. Jenkins, “Clearpose: Large-scale transparent object dataset and benchmark,” *arXiv preprint arXiv:2203.03890*, 2022.
- [16] P.-J. Lai and C.-S. Fuh, “Transparent object detection using regions with convolutional neural network,” in *IPPR Conference on Computer Vision, Graphics, and Image Processing*, vol. 2, 2015.
- [17] M. P. Khaing and M. Masayuki, “Transparent object detection using convolutional neural network,” in *International Conference on Big Data Analysis and Deep Learning Applications*. Springer, 2018, pp. 86–93.
- [18] E. Xie, W. Wang, W. Wang, M. Ding, C. Shen, and P. Luo, “Segmenting transparent objects in the wild,” in *European Conference on Computer Vision*. Springer, 2020, pp. 696–711.
- [19] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.
- [20] L. Zhu, A. Mousavian, Y. Xiang, H. Mazhar, J. van Eenbergen, S. Debnath, and D. Fox, “Rbg-d local implicit function for depth completion of transparent objects,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4649–4658.
- [21] Y. Tang, J. Chen, Z. Yang, Z. Lin, Q. Li, and W. Liu, “Depthgrasp: Depth completion of transparent objects using self-attentive adversarial network with spectral residual for grasping,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 5710–5716.
- [22] J. Chang, M. Kim, S. Kang, H. Han, S. Hong, K. Jang, and S. Kang, “Ghostpose*: Multi-view pose estimation of transparent objects for robot hand grasping,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 5749–5755.
- [23] M. Tian, L. Pan, M. H. Ang, and G. H. Lee, “Robust 6d object pose estimation by learning rgb-d features,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6218–6224.
- [24] Y. Xu, H. Nagahara, A. Shimada, and R.-i. Taniguchi, “Transcut: Transparent object segmentation from a light-field image,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3442–3450.
- [25] A. Kalra, V. Taamazyan, S. K. Rao, K. Venkataraman, R. Raskar, and A. Kadambi, “Deep polarization cues for transparent object segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8602–8611.
- [26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.
- [27] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, “Normalized object coordinate space for category-level 6d object pose and size estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2642–2651.

- [28] S. Umeyama, “Least-squares estimation of transformation parameters between two point patterns,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 13, no. 04, pp. 376–380, 1991.
- [29] M. Tian, M. H. Ang, and G. H. Lee, “Shape prior deformation for categorical 6d object pose and size estimation,” in *European Conference on Computer Vision*. Springer, 2020, pp. 530–546.
- [30] K. Chen and Q. Dou, “Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2773–2782.
- [31] W. Chen, X. Jia, H. J. Chang, J. Duan, L. Shen, and A. Leonardis, “Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1581–1590.
- [32] Y. Di, R. Zhang, Z. Lou, F. Manhardt, X. Ji, N. Navab, and F. Tombari, “Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting,” *arXiv preprint arXiv:2203.07918*, 2022.
- [33] J. Lin, Z. Wei, Z. Li, S. Xu, K. Jia, and Y. Li, “Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3560–3569.
- [34] L. Zou, Z. Huang, N. Gu, and G. Wang, “6d-vit: Category-level 6d object pose estimation via transformer-based instance representation learning,” *arXiv preprint arXiv:2110.04792*, 2021.
- [35] Z. Fan, Z. Song, J. Xu, Z. Wang, K. Wu, H. Liu, and J. He, “Acr-pose: Adversarial canonical representation reconstruction network for category level 6d object pose estimation,” *arXiv preprint arXiv:2111.10524*, 2021.
- [36] Y. You, R. Shi, W. Wang, and C. Lu, “Cppf: Towards robust category-level 9d pose estimation in the wild,” *arXiv preprint arXiv:2203.03089*, 2022.
- [37] X. Jiang, D. Li, H. Chen, Y. Zheng, R. Zhao, and L. Wu, “Uni6d: A unified cnn framework without projection breakdown for 6d pose estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11174–11184.
- [38] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015, pp. 234–241.
- [39] A. R. Zamir, A. Sax, N. Cheerla, R. Suri, Z. Cao, J. Malik, and L. J. Guibas, “Robust learning through cross-task consistency,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11197–11206.
- [40] Z.-H. Lin, S.-Y. Huang, and Y.-C. F. Wang, “Convolution in the cloud: Learning deformable kernels in 3d graph convolution networks for point cloud analysis,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [41] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [42] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [43] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, “On the variance of the adaptive learning rate and beyond,” *arXiv preprint arXiv:1908.03265*, 2019.
- [44] H. Yong, J. Huang, X. Hua, and L. Zhang, “Gradient centralization: A new optimization technique for deep neural networks,” in *European Conference on Computer Vision*. Springer, 2020, pp. 635–652.
- [45] M. Zhang, J. Lucas, J. Ba, and G. E. Hinton, “Lookahead optimizer: k steps forward, 1 step back,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [46] M. Krogius, A. Haggenmiller, and E. Olson, “Flexible layouts for fiducial tags,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2019.
- [47] X. Chen, K. Zheng, Z. Zeng, S. Basu, J. Cooney, J. Pavlasek, and O. C. Jenkins, “Manipulation-oriented object perception in clutter through affordance coordinate frames,” *arXiv preprint arXiv:2010.08202*, 2020.