ReFlow6D: Refraction-Guided Transparent Object 6D Pose Estimation via Intermediate Representation Learning

Hrishikesh Gupta, Stefan Thalhammer, *Member, IEEE*, Jean-Baptiste Weibel, *Member, IEEE*, Alexander Haberl, and Markus Vincze, *Member, IEEE*

***Abstract*—Transparent objects are ubiquitous in daily life, making their perception and robotics manipulation important. However, they present a major challenge due to their distinct refractive and reflective properties when it comes to accurately estimating the 6D pose. To solve this, we present *ReFlow6D*, a novel method for transparent object 6D pose estimation that harnesses the *refractive-intermediate representation*. Unlike conventional approaches, our method leverages a feature space impervious to changes in RGB image space and independent of depth information. Drawing inspiration from image matting, we model the deformation of the light path through transparent objects, yielding a unique object-specific intermediate representation guided by light refraction that is independent of the environment in which objects are observed. By integrating these intermediate features into the pose estimation network, we show that *ReFlow6D* achieves precise 6D pose estimation of transparent objects, using only RGB images as input. Our method further introduces a novel transparent object compositing loss, fostering the generation of superior *refractive-intermediate* features. Empirical evaluations show that our approach significantly outperforms state-of-the-art methods on *TOD* and *Trans32K-6D* datasets. Robot grasping experiments further demonstrate that *ReFlow6D's* pose estimation accuracy effectively translates to real-world robotics task.**

***Index Terms*—Deep learning for visual perception, perception for grasping and manipulation, visual learning.**

I. Introduction

ROBOT object manipulation using 6D pose estimation is a well-studied problem in robotics, e.g., [1][2]. It uses the estimation of the 6D pose of objects, i.e., 3D rotation and 3D translation, for solving a wide variety of real-world tasks such as object grasping, scene understanding, and complex robotic object manipulation. Researchers have extensively studied this problem for opaque objects [3] such as texture-less objects [4], [5], symmetrical objects [6][7] and also occluded objects [8]. However, even though transparent objects are as pervasive in daily life and household settings as opaque objects, their 6D pose estimation has been approached only in comparably few works [9], [10].

Methods developed for 6D pose estimation of opaque and Lambertian objects cannot be directly applied to transparent objects without degradation of performance. This is mainly because transparent objects pose two prominent challenges to visual perception system. Firstly, transparent objects do not exhibit consistent RGB color and texture features across varying scenes. A transparent object's appearance depends on the scene's lighting, background, and setup. Hence, the visual features drastically differ between scenes thereby confounding RGB feature-based learning methods. Secondly, the non-Lambertian nature of transparent objects poses challenges for commercial depth sensors and leads to inaccurate depth measurements [11], [12]. This limitation significantly impacts state-of-the-art pose estimation approaches reliant on precise depth data. Research on transparent objects can be divided into the following primary directions. One approach focuses on completing missing or erroneous depth information for transparent objects [11], [13] followed by utilizing a secondary network for pose estimation [12]. The second approach focuses on estimating poses from implicit depth cues, such as stereo images [14]. Both approaches leverage depth cues, yet recent advancements such as [15] have demonstrated the effectiveness of estimating poses using the RGB image space only, while incorporating the relevant geometric features for opaque objects [16]. Furthermore, [17] has extended this methodology to transparent objects by showing the effectiveness of intermediate geometric and edge representations.

In this paper, we present *ReFlow6D*, a novel approach for transparent object 6D pose estimation that leverages *refractive-intermediate representation* (see Fig. 1), a more reliable intermediate feature space for transparent objects. This representation captures the deformation of the light path induced by the given transparent object. This deformation is consistent and

Received 8 March 2024; accepted 2 August 2024. Date of publication 6 September 2024; date of current version 23 September 2024. This article was recommended for publication by Associate Editor Domenico G. Sorrenti and Editor Abhinav Valada upon evaluation of the reviewers' comments. This work was supported by the EU-program EC Horizon 2020 for Research and Innovation under Grant 101017089, through Project TraceBot. *(Corresponding author: Hrishikesh Gupta.)*

Hrishikesh Gupta, Jean-Baptiste Weibel, Alexander Haberl, and Markus Vincze are with the Vision for Robotics Laboratory, Automation and Control Institute, 1040 TU Wien, Austria (e-mail: Gupta@acin.tuwien.ac.at; weibel@acin.tuwien.ac.at; vincze@acin.tuwien.ac.at; haberl@acin.tuwien.ac.at).

Stefan Thalhammer is with the Industrial Engineering Department, UAS Technikum Vienna, 1200 Vienna, Austria (e-mail: stefan.thalhammer@technikum-wien.at).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2024.3455897, provided by the authors.

Digital Object Identifier 10.1109/LRA.2024.3455897

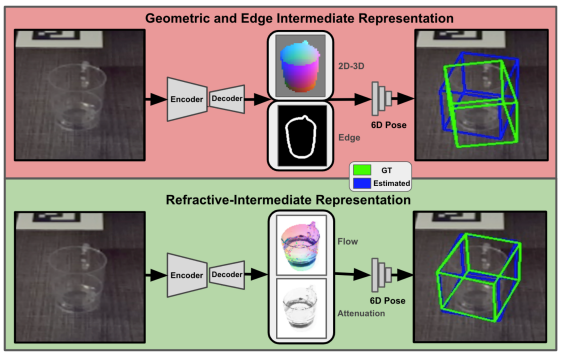

Fig. 1. **Intermediate representation for pose estimation**: The figure shows the effectiveness of the refractive-intermediate representation vs Geometric and edge intermediate layers applied to 6D pose estimation. The green 3D Bbox shows groundtruth, while the blue shows the estimation.

independent of the environment and accounts for symmetries of the object. As a result, making *ReFlow6D* robust to changes in RGB image space and operate independently of the depth information.

In detail, for a given pose, the deformation of light induced by a transparent object can be expressed as a refractive flow and an attenuation. The refractive flow is the offset between a foreground pixel and its refracted counterpart on the background, while the attenuation captures light intensity change at each pixel. These properties were utilized by Tom-Net [18] for transparent object matting and compositing. Image matting is the process of extracting a foreground object from a background image, yielding an opacity value (and refraction in the case of a transparent object) for each pixel of the extracted object referred to as *matte*. Inversely, image compositing combines the foreground object with a new background, guided by the extracted *matte* to reveal or conceal specific regions [19]. The refractive flow and attenuation as detailed before are unique optical properties of a transparent object regardless of the environment where it is placed. We combine these optical properties mentioned in [18] with the object binary mask and surface region attention maps (Surface Regions), forming our *refractive-intermediate representation* for transparent objects. Using a learned Patch-PnP [15] we directly regress the 6D object pose from these learned *refractive-intermediate representation*.

To summarize, the contributions towards solving transparent object 6D pose estimation for robot object handling are:

1) Incorporation of the *refractive-intermediate representation* in the pose estimation architecture as intermediate features that model the deformation of light paths through transparent objects, which is a unique matte for a transparent object and invariant to environment changes, enabling more robust and accurate 6D pose estimation.
2) Evaluations against the state of the art showing the improvement of estimation of 6D poses of transparent objects using the refractive over the geometry intermediate representation.
3) Robot transparent object manipulation experiments demonstrating the real applicability of *ReFlow6D*.

The paper is organized as follows: Section II covers related work. Section III explains ReFlow6D and the network architecture. Section IV presents the experimental evaluation and results against state-of-art methods. Section V showcases robot grasping experiments in a real-world environment using the pose estimates of ReFlow6D. Finally, we conclude this paper in Section VI.

II. Related Works

In this section, we briefly review representative and recent works on the perception of the transparent objects, their manipulation and transparent object pose estimation and image matting.

A. Transparent Object Perception and Manipulation

Perception of transparent objects is a necessary precursor to robotic manipulation. In order to address the challenges associated with transparent objects, CNN models for the detection of transparent objects were introduced by [20] using the RGB image space only. Further, [21] proposed a deep segmentation model that achieved state-of-the-art accuracy in segmentation tasks. Specialized sensors such as in [22] training Mask R-CNN using polarization images to outperform baseline models have also been explored.

A solution to accurately measure the depth information of transparent objects is to estimate the missing depth through depth completion. In the context of pose estimation and robotic grasping, ClearGrasp [11] utilized depth completion techniques, training DeepLabv3+ models for image segmentation, surface normal estimation, and boundary segmentation before robot grasping. Further advancements in in-depth completion include methods involving implicit functions [13] and NeRF features [23].

B. Transparent Object 6D Pose Estimation and Image Matting

Solving pose estimation problems using implicit depth cues, Keypose [14] was introduced for regressing 2D keypoints using stereo images. It outperformed DenseFusion [24], even with groundtruth depth. [9] proposed a transparent object grasping approach estimating the object 6D poses from the proposed model-free pose estimation approach using multiview-geometry.

In the realm of monocular RGB-based methods, Transnet [25] directly regresses transparent object pose from images using depth completion and surface normal estimation. GDR-Net [15] unifies direct and geometry-based indirect methods, providing direct pose information output for opaque objects. Building upon CDPN [26] and EPOS [6] GDR-Net uses geometric features such as surface region attention and dense correspondence maps as intermediate layers, directly regressing 6D poses using their proposed Patch-PnP method. Further, building upon GDR-Net, TGF-Net [17] regresses direct 6D poses with the Patch-PnP method using their proposed edge intermediate representations, demonstrating improvement over the proposed geometric representations presented in [15] for transparent objects.

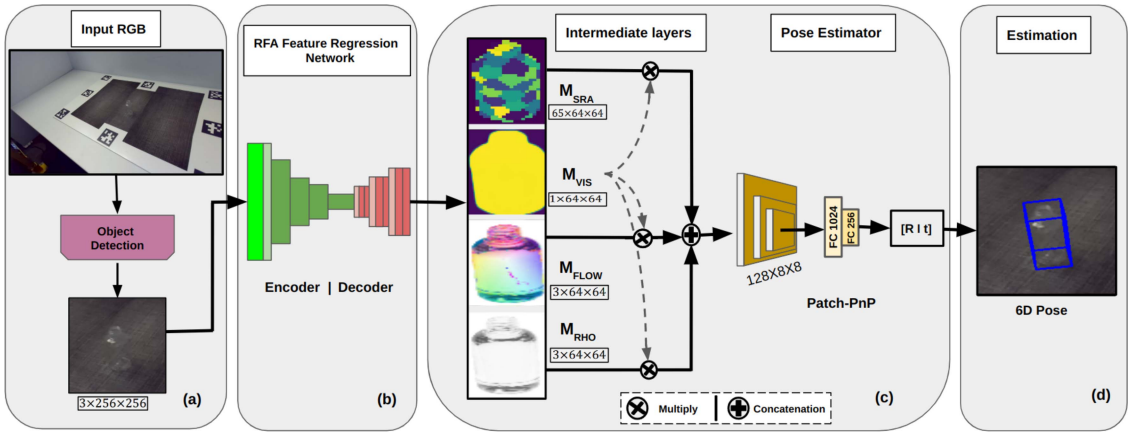

Fig. 2. **Framework of ReFlow6D**: (a) Given an RGB image I we use off-the-shelf object detector for detecting transparent objects. (b) The RFA feature regression network takes then zoomed-in RoI as input and predicts several refractive-intermediate representation. (c) These intermediate features are then concatenated and provided as input to the Patch-PnP. (d) The Patch-PnP directly regresses the 6D object pose of the transparent object.

Object matting and compositing were detailed early by Zongker et al. [19] and [27]. Building on this TOM-Net [18] frame the transparent object image matting problem as a refractive flow and attenuation estimation problem. They show that this matte can be effectively estimated and learned by CNN-based methods for colorless transparent objects. Further, by compositing the transparent object on varying backgrounds using the learned matte, they also show that the extracted matte (RFA) is a unique property of the transparent object, independent of its environment.

We aim to leverage this distinct matte to tackle transparent object pose estimation, over the geometric [15] and edge intermediate representation [17].

III. ReFlow6D

We now present our method to solve the transparent object 6D pose estimation problem, given the input RGB image I and a set of N transparent objects $O = \{ O_i \mid \text{i} = 1..\text{N} \}$ together with their corresponding 3D CAD models $S = \{ S_i \mid \text{i} = 1..\text{N} \}$. First, all transparent objects of interest from O in the image I are detected using a standard object detector like [28], [29]. We then aim to learn the *refractive-intermediate representation* for the object O_i in the image I.

Using these learned intermediate representations we then perform the 6D pose regression $P_i = [R_i|t_i]$ w.r.t. the camera for each object O_i present in I, where R_i describes the 3D rotation and t_i denotes the 3D translation of the detected object. Fig. 2 provides a schematic overview of the proposed ReFlow6D. We utilize the Patch-PnP method proposed by GDR-Net [15], enabling us to perform a direct pose regression from the concatenation of these learned intermediate features. This stands in contrast to earlier methods that employed indirect pose estimation approaches [26].

Following, Section III-A illustrates the proposed *refractive-intermediate representation* and how it is estimated. Further in Section III-B, we detail the losses we employ to learn this intermediate representation and the 6D pose regression from the Patch-PnP method. Additionally, Section III-B will provide details on the object compositing loss that we use for further supervision to refine the estimated *refractive-intermediate representation*.

A. Refractive-Intermediate Representation

In this subsection, we introduce in detail the intermediate representation we use for estimating a direct 6D pose regression using Patch-PnP. For learning the *refractive-intermediate representation*, we employ an encoder-decoder architecture, specifically utilizing the Geometric feature regression network proposed in [15]. This choice enables us to learn the intermediate representations with minimal adjustments to their network and directly regress poses using their proposed Patch-PnP. Specifically, we retain the layers for regressing the surface-region map and object mask while adding channels necessary for refractive flow and attenuation regression. We refer to this modified network as *RFA regression network* (as seen in Fig. 2).

Refractive flow and attenuation (RFA): The refractive flow M_{FLOW} represents the deformation of the light ray traversing through the transparent object from its background environment to the foreground object's surface. Where, each pixel on the refractive flow image is a 2D vector $(\delta x, \delta y)$, which indicates the offset between the observed foreground pixel and its corresponding background pixel after the refraction of the ray [18]. The attenuation or the attenuation index M_{RHO}, as mentioned in TOM-Net, represents the intensity of the light-ray traversing through the object on its surface. The attenuation represents the magnitude or light intensity at a pixel. The RFA of an object $O_i \in O$ can be described as a function of index of refraction(IOR) of the O_i and S_i. The parameter S_i plays a crucial role in deriving the shape and geometry of the object. Since IOR, the shape, and geometry of the object O_i are constant attributes, their RFA properties (refractive flow and attenuation) remain constant as well. These RFA properties of an object $O_i \in O$ remain unchanged, regardless of the environment it is placed in. Forming a set of physically plausible intermediate features independent to changes in the environment of the transparent object.

Handling Object Symmetries with Surface Region Attention maps: We implement Surface Regions M_{SRA}, akin to prior works [15], [17], [30]. These fragments encapsulate rich geometry information, while providing additional ambiguity-aware supervision for the 6D pose estimation. We predict M_{SRA} similarly to [15], treating it as an intermediate layer. Inspired by [6], M_{SRA} serves as a symmetry-aware feature map guiding Patch-PnP feature learning. M_{SRA} derived from Dense Correspondence Maps by employing farthest points sampling. [15], classifying each pixel in the dense correspondence maps into the corresponding regions. Thus the classification probabilities predicted for each pixel in M_{SRA} implicitly represent the symmetry of the object.

In addition to the above mentioned intermediate representations, we also predict the mask of the object. This implicitly encodes the geometric information of the object in addition to M_{SRA}. Specifically, we predict the object visibility mask M_{VIS} for each detected object in image I.

B. Loss Functions

The final loss L function of our optimization scheme comprises two distinct loss functions:

$$L = L_{\text{inter}} + L_{\text{pose}} \tag{1}$$

Where L_{inter} denotes the loss function for our intermediate features and L_{pose} represents loss for our output patch-PnP [15] pose. L_{inter} function can individually be denoted as:

$$L_{\text{inter}} = L_{\text{FLOW}} + L_{\text{RHO}} + L_{\text{M}_{\text{VIS}}} \tag{2}$$

Where,

$$\begin{cases} L_{\text{FLOW}} = ||\hat{M}_{\text{VIS}} \cdot (\tilde{M}_{\text{FLOW}} - \hat{M}_{\text{FLOW}})||_1 \\ L_{\text{RHO}} = ||\hat{M}_{\text{VIS}} \cdot (\tilde{M}_{\text{RHO}} - \hat{M}_{\text{RHO}})||_1 \\ L_{\text{M}_{\text{VIS}}} = ||\tilde{M}_{\text{VIS}} - \hat{M}_{\text{VIS}}||_1 \end{cases} \tag{3}$$

Where, $\hat{.}$ and $\tilde{.}$ denote groundtruth and estimation, respectively. L_{pose} function can individually be denoted as:

$$L_{\text{pose}} = L_{\text{R}} + L_{\text{center}} + L_{\text{Z}} \tag{4}$$

Where following [15] we employ a disentangled 6D pose loss via individually supervising the rotation R, the scale-invariant 2D object center (δ_x, δ_y), and the distance δ_z. Where the components of (4) are:

$$\begin{cases} L_{\text{R}} = avg_{x \in S} ||\hat{R}x - \tilde{R}x||_1 \\ L_{\text{center}} = ||\hat{\delta_x} - \tilde{\delta_x}, \hat{\delta_y} - \tilde{\delta_y}||_1 \\ L_{\text{Z}} = ||\hat{\delta_z} - \tilde{\delta_z}||_1 \end{cases} \tag{5}$$

Object compositing as additional supervision: The quality of predicted Refractive Flow and Attenuation (RFA) features as intermediate layers strongly influence the accuracy of the final 6D pose estimation for transparent objects, as demonstrated in subsequent ablation studies. Observations in [18] show that higher quality RFA features improve the quality of image compositing of transparent objects on various backgrounds. As [18] introduced RFA for transparent object compositing, we propose to use this task as additional supervision. We introduce L_{comp}, as an additional supervisory loss, to further refine the estimated RFA intermediate features by the RFA feature regression network, thus improving the 6D pose estimation. We use the estimated matte (i.e., RFA), for compositing the transparent object on a random background image for both groundtruth and predictions.

Fig. 3. **Transparent object Compositing:** Examples of transparent object compositing from the TOD and Trans32K-6D datasets on random COCO backgrounds. Used for additional supervision loss for refining the estimated RFA.

Drawing inspiration from [18] and [19] for transparent object compositing, we apply the compositing equation to each pixel [18]:

$$C = (1 - M_{\text{VIS}}).B + M_{\text{VIS}}.M_{\text{RHO}}.f(\mathbf{T}, M_{\text{FLOW}}) \tag{6}$$

Here, C is the final computed compositing for a given pixel, B is the background color, and $\mathbf{T}$ are the set of calibration images [18], and f is the matting function. The variable M_{VIS} in (6) distinguishes between background and foreground. Equation 6 is then estimated by using the groundtruth and estimated M_{VIS}, M_{RHO}, and M_{FLOW} for each pixel of the transparent object. Fig. 3 shows examples of the final compositing of the transparent objects on random COCO backgrounds [31].

Thus we add L_{comp} to the loss function 1, redefining our final loss function L as follows:

$$L = L_{\text{inter}} + L_{\text{pose}} + L_{\text{comp}} \tag{7}$$

$$L_{\text{comp}} = ||\hat{M}_{\text{VIS}} \cdot \hat{C} - \tilde{M}_{\text{VIS}}.\tilde{C}|| \tag{8}$$

$\hat{C}$ is computed using groundtruth RFA and $\tilde{C}$ is calculated using estimated RFA intermediate features during the training procedure using (6).

IV. Experiments

In this section, we first introduce our experimental setup. We will detail the implementation of our method and the datasets being used, followed by the evaluation metrics. We then proceed to present the evaluation results for commonly employed benchmark transparent pose estimation methods against our method.

A. Experimental Setup

Implementation Details: Our experiments are implemented using PyTorch [32] and GDRNPP, the version of [15] presented in the BOP challenge 2022 [33]. We train our networks on Nvidia-GTX 3090 and Nvidia A6000. Our method is trained end-to-end using the Ranger optimizer [34] [35], with a batch size of 8 and a base learning rate of 1e-4, which we anneal at 72% of the training phase using a cosine schedule [36]. For transparent object compositing detailed in section 3. (b), we use COCO images [31] as the random background.

RFA matte rendering: The training dataset is generated using BlenderProc [37]. Refractive flow, attenuation, and binary mask for each object under each viewpoint are obtained following the gray code-based calibration method used in [18]. We keep the index-of-refraction fixed at 1.5 for all objects. *Datasets:* We conduct our experiments on two datasets: *TOD* (Transparent object dataset) which is introduced by the Keypose method [14], and the Trans6D-32K introduced by TGF-Net [17]. The *TOD* dataset provides 2D and 3D groundtruth keypoints and contains 15 unique transparent objects with varying shapes, scales, and symmetries. The dataset contains approximately 2700 training samples and 350 test samples per object, with each image sample of resolution 720X 1080. For obtaining the groundtruth 6D pose from the provided groundtruth keypoints, we use the Orthogonal Procrustes algorithm [38], as done in [14]. *Trans6D-32K* contains 10 unique common types of household transparent objects, of which 5 are symmetric and 5 are non-symmetric objects. The dataset contains 400 synthetic training images and 2800 synthetic test images per object, so the entire dataset contains 32000 images.

Evaluation Metrics: We use two standard metrics for 6D object pose evaluation, i.e. *ADD(-S)* [3] [39] and *Average Recall (AR)* [16]. The ADD metric assesses whether the average deviation of transformed model points is within 10% of the object's diameter (0.1d). For symmetric objects, ADD-S measures error as the average distance to the closest model point. Average Recall (AR) is computed as the mean of three metrics: Maximum Symmetry-Aware Projection Distance (MSPD), Maximum Symmetry-Aware Surface Distance (MSSD) and Visible Surface Discrepancy (VSD). For detailed explanations, please refer to [40].

The Keypose method [14] does not directly predict the 6D pose of the object, unlike our method, but instead estimates the 2D keypoints of the object. To ensure a fair comparison, we evaluate our method using Mean Absolute Error (MAE) scores for predicted keypoints. We leverage our trained 6D pose estimation model, and transform using the groundtruth keypoints provided by *TOD*. Then we project it on the 2D image space of image I using the camera intrinsics.

B. Evaluation

We evaluate our method against the Keypose [14], GDR-Net [15] and TGF-Net [17] method. For comparision with the Keypose method we train our method and GDR-Net with the $TOD\ dataset$ and use the inference output per object, as provided by the trained Keypose models. For comparison with the TGF-Net, we train our method using the $Trans6D-32K$ dataset and use the evaluation results as provided in [17]. The results of the comparison experiment trained with the $TOD\ dataset$ are illustrated in Table I, and with the $Trans6D-32K$ dataset is illustrated in Table II.

TABLE I
AVERAGE RECALL RESULTS ON THE *TOD* DATASET OBJECTS

Objects	KeyPose	Ours	Keypose	GDR-Net	Ours
Metric	MAE↓		AR↑		
Bottle0	**4.6**	4.9	**92.4**	85.9	88.8
Bottle1	5.1	**1.8**	18.6	61.3	**68.9**
Cup0	6.8	**1.3**	99.6	99.9	**100.0**
Cup1	7.1	**1.4**	87.1	83.9	**89.1**
Mug0	**8.8**	11.2	83.2	84.8	**99.2**
Mug1	21.9	**2.1**	88.1	97.7	**98.0**
Mug2	10.1	**1.6**	80.1	**89.2**	88.7
Mug3	11.3	**1.3**	60.0	88.8	**89.8**
Mug4	12.1	**2.1**	77.6	**91.7**	88.8
Mug5	9.0	**1.9**	89.2	**94.4**	92.3
Mug6	9.7	**2.0**	95.3	**98.6**	98.2
Tree0	15.6	**6.1**	**91.7**	88.8	89.3
heart0	12.8	**5.8**	38.10	72.6	**84.6**
Mean	10.4	**3.4**	77.0	87.5	**90.4**

The bold entities show quantitative values of the best-performing models in that category.

TABLE II
ADD(-S) ON *TRANS6D-32K* DATASET OBJECTS

Objects	GDR-Net	TGF-Net	Ours
#01	73.5	83.4	**89.7**
#02	76.3	83.5	**95.2**
#03	67.4	67.7	**82.8**
#04	82.9	85.4	**91.0**
#05	78.5	89.6	**91.8**
#06	91.5	89.6	**94.7**
#07	90.6	92.6	**94.8**
#08	96.6	97.3	**99.3**
#09	96.0	97.5	**98.6**
#10	92.4	**94.9**	93.8
Mean	84.6	88.2	**93.2**

The bold entities show quantitative values of the best-performing models in that category.

C. Comparison With the Benchmark

Results on TOD: Quantitative results are mentioned in Table I. Our method achieves the best average recall of **90.4**% against 87.5% and 77.0% for GDR-Net and Keypose method respectively. ReFlow6D outperforms the other two state-of-the-art methods for highly symmetric and textureless objects such as $Bottle1$, for which we achieve an AR score of **68.9**%. For highly asymmetric objects such as $Tree0$, our method and GDR-Net method perform a bit worse than the Keypose method. This is because both the GDR-Net and our method rely on Surface Regions, which are difficult to predict for a complex geometry such as $Tree0$. Fig. 4(a) shows the qualitative results on TOD. For keypoint prediction, our method achieves the best *MAE* score of **3.4** compared to Keypose's score of 10.4. $Ball0$ was excluded due to the absence of a groundtruth binary mask, and $Bottle2$ was excluded due to inaccurate test-set groundtruth 6D poses provided by TOD . Thus only evaluating on 13 TOD dataset objects. In addition to comparing MAE scores with the [14]

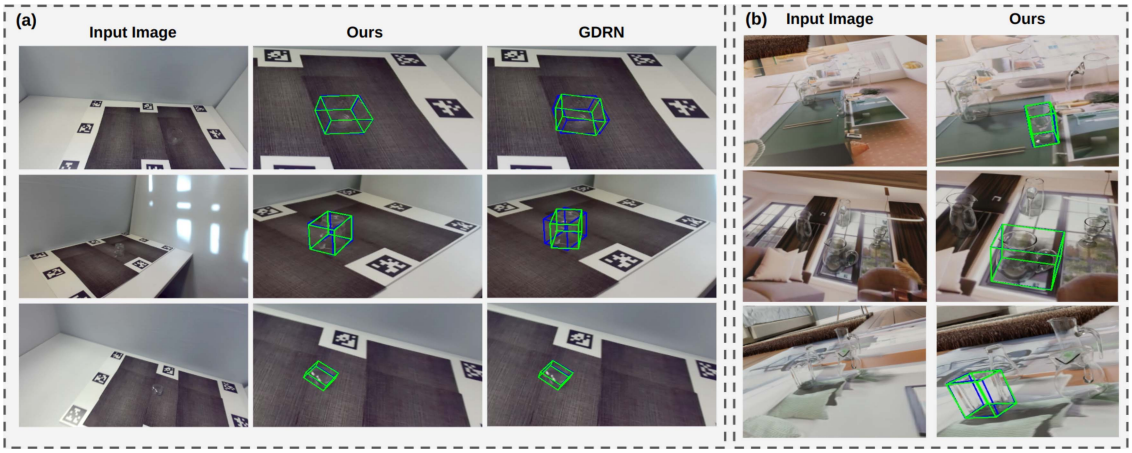

Fig. 4. **Qualitative Results of ReFlow6D:** (a) Qualitative results on the TOD dataset. (b) Qualitative results on the Trans32K-6D dataset. Estimates are shown in cropped images for visibility. No estimates are shown for the TGF-Net method as the authors did not publish their code.

TABLE III
ABLATION STUDY OF RFA FEATURES IN ADD(-S) ON *TRANS6D-32K*

Features	L_{comp}	Flow	Rho	M_{SRA}	ADD(-S)
w/o Compositing loss	×	✓	✓	✓	90.5
w/o Flow	×	×	✓	✓	80.3
w/o Rho	×	✓	×	✓	93.0
Full model	✓	✓	✓	✓	**93.2**

The bold entities show quantitative values of the best-performing models in that category.

method, we also compare against MVTrans [41], a multi-view approach for transparent object perception and pose estimation. Our method achieves the best *MAE* score of **3.4**, averaged over 13 TOD objects, compared to MVTrans's score of **7.4**.

Results on Trans6D-32K: We take the evaluation criteria i.e, the ADD(-s) and results report by the TGF-Net in their original paper [17]. Our method outperforms TGF-Net and GDR-Net on the Trans6D-32K dataset, as indicated in Table II. Notably, our method achieves the best ADD(-S) score of **93.2**, outperforming TGF-Net (88.2) and GDR-Net (84.6). Even complex non-symmetric objects like #03 *ReFlow6D* perform much better with minimal degradation compared to the other methods. Qualitative results on TOD are depicted in Fig. 4(b).

D. Ablation Studies

In order to verify the effectiveness of the RFA intermediate layers, we conducted ablation experiments. All ablation experiments use the same network initialization and training scheme. We consider the original GDR-Network [15] without Dense correspondence maps as our base network since we only add the RFA intermediate layers to this.

Effect without compositing Loss: We exclude the L_{comp} detailed in (9), while keeping the other losses and predicting all the RFA, M_{SRA} and M_{VIS}. We can see from Table III that there is a decline in performance to 90.5 from 93.2 (as indicated in Table II). Proving that compositing loss as an additional supervision loss contributes to the performance of our method. This also substantiates that improved RFA intermediate features contribute to improved accurate 6D pose estimation of transparent objects. *Effect without Flow:* We remove the prediction of Flow as intermediate layers. While keeping the prediction of other RFA, M_{SRA}, M_{VIS} and losses. We can see in Table III the big performance drop to 80.3 from 93.2in Table I. This notable decrease in performance is primarily attributed to the crucial role played by refractive flow features in influencing the final 6D pose estimation, surpassing the contribution of other RFA features.

Effect without Rho: We remove the prediction of attenuation M_{RHO} as intermediate layers while keeping the prediction of other RFA components, M_{SRA}, M_{VIS} and losses. From Table III, we observe a marginal drop in performance to 93.0 compared to the full model. This marginal drop is attributed to the fact that Flow along with Surface Regions encodes all the essential information required for predicting the 6D pose of the transparent object. This proves, that learning M_{RHO} for pose estimation does not contribute greatly towards 6D pose estimation of the transparent objects, explaining the increase in performance when compared to the ablation studies of removing L_{comp}. Indicating that its absence facilitates easier learning for the Patch-PnP method.

We have opted not to present the ablation results for ReFlow6D trained exclusively with M_{SRA}. This is because the Patch-PnP network fails to converge when trained solely with M_{SRA} .

V. REAL-WORLD ROBOT EXPERIMENTS

To showcase that our proposed method ReFlow6D can be directly applied to real-world scenarios, we conduct robot grasping experiments.

Dataset and Scene setup: Due to the unavailability of the real physical object instances from the TOD and $Trans32K-6D$ datasets for real robot grasping experiments, we use items from the TraceBot project .[1] We construct a physically based synthetic dataset of the available objects we term *TraceBot dataset*.

[1]H2020 program under grant agreement No 101017089

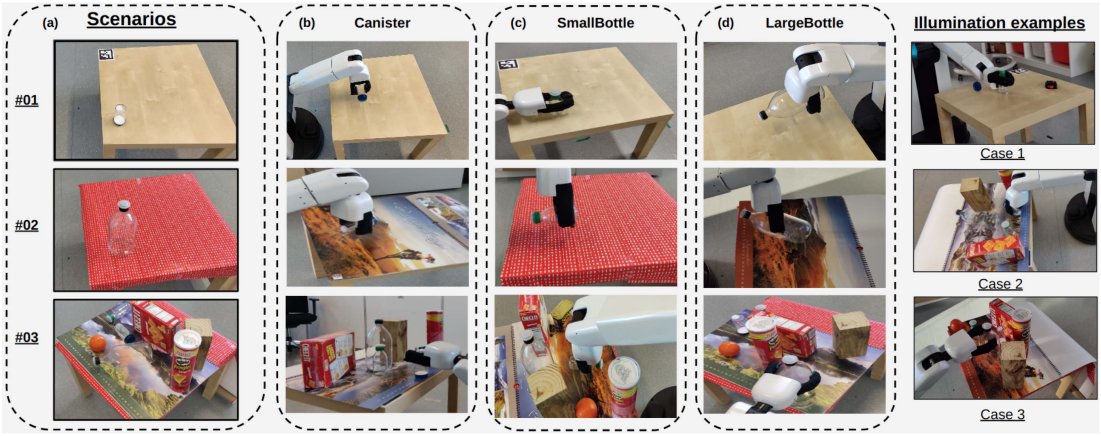

Fig. 5. **Grasping qualitative results:** On the left we show (a) Examples of all three different scenarios. (b) Grasping example of the object "Canister" for all 3 scenarios. (c) Grasping example of the object "SmallBottle" for all 3 scenarios. (d) Grasping of the object "LargeBottle" for all 3 scenarios. On the right we show four cases of illumination (a) Light sifting through the semi-permeable blind covering the only window of the room. (b) Artificial ambient light. (c) Natural light. While the fourth case of illumination i.e, superposition of artificial ambient and natural light is shown on the left side of images with the grasping senarios.

The *TraceBot dataset* includes a mixture of three transparent and six non-transparent objects as distractors for real-world grasping experiments. The three transparent objects are, the "$Canister$", "$SmallBottle$" and "$LargeBottle$". The object CAD models were constructed by precisely measuring each of the three objects by a caliper and constructing the model using SolidWorks. To render this dataset we used Blender [42] with varied background and illumination for each image in the dataset, where all the nine objects were placed in random poses in the setup. For each of the three transparent objects, 5000 images were rendered for training. The scene setup for the real-world grasping was set up in front of the robot on a table, as can be seen in Fig. 5. In order to create challenging and varying scene illumination the light is varied for each grasp. In particular we have consider four cases of illumination detailed in 5.

In total, we created and test three scenarios with considerable domain shift to the training data. The three scenarios can be described as follows:

- *Scenario 1:* includes only the raw table plane unseen in the training data, with the single transparent object placed randomly on it. Fig. 5(a)(01) illustrates scenario 1.
- *Scenario 2:* includes a random textured background unseen in the training data, placed on the table for each grasp. The single transparent object is also placed randomly on it for each real-robot grasp experiment. Fig. 5(a)(02)illustrates scenario 2.
- *Scenario 3:* is a cluttered scene consisting of the two other transparent objects with 5 other randomly chosen unseen opaque objects, together with unseen random background texture for each grasp. Both the background texture and the opaque objects were not part of the training data. All the objects are randomly placed on the table with random backgrounds. Fig. 5(a)(03) illustrates scenario 3.

Hardware and Implementation: We employ the Toyota HSR robot [43] which comes with the RGB-D camera Xtion PRO LIVE mounted, for all our grasping experiments. We train ReFlow6D network following the implementation details mentioned in Section 4. Fixed grasp points were manually selected and annotated based on the CAD model of the objects and the robot's gripper in advance. We first train a YOLOv3 [29] detection network to detect the 2D bounding box of the object, after which we trained the ReFlow6D method for 6D pose estimation per object. Finally, we used the robot to grasp transparent objects to demonstrate the applicability of our method in real-world scenarios. We used the same grasp annotation method and grasping pipeline as described in [44]. We conducted a total of 30 grasps per object, i.e., 10 grasps for each of the mentioned scenarios.

TABLE IV
GRASPING EXPERIMENTS RESULTS

	Canister	SmallBottle	LargeBottle	**Mean**
Scenario 01	90%	100%	80%	90%
Scenario 02	90%	90%	80%	86.6%
Scenario 03	80%	70%	60%	70%
Total Success	86.6%	86.6%	73.3%	82.2%

Results: In the qualitative experimental results fig. 5(b)(c)(d), we showcase performance for each object and scenario. Quantitative results in Table IV detail 10 grasps per object per scenario, yielding a mean success rate of **82.2%**. $LargeBottle$ exhibits the lowest success rate **73.3%** success rate for 30 grasps, attributed to its size similarity with our robot's gripper. This highlights the need for precise grasp-planning, object detection, and 6D pose estimation, leaving a smaller margin of error. These results also highlight how ReFlow6D's accurate pose estimation translates effectively to real-world tasks.

VI. CONCLUSION

In this letter, we propose ReFlow6D, a monocular instance-level 6D pose estimation approach tailored specifically for transparent objects. Our method proposes a novel set of *refractive-intermediate representations*, guided by refractive principles and enabling robust transparent object pose estimation. We demonstrated that integrating these refractive feature attributes (RFAs) alongside surface-region attention as intermediate features better guide the network toward more precise 6D pose estimations for transparent objects. Through comprehensive

empirical evaluations, we demonstrate the effectiveness of Re-Flow6D in real-world scenarios compared to existing state-of-the-art methods. These results underscore the efficacy of the *refractive-intermediate representation* over geometric and edge-based representations. Future work will investigate pose estimation of objects with more complex transparent objects with diverse thickness, geometry, and indices of refraction.

REFERENCES

[1] S. Thalhammer, M. Leitner, T. Patten, and M. Vincze, "Pyrapose: Feature pyramids for fast and accurate object pose estimation under domain shift," in *Proc. 2021 IEEE Int. Conf. Robot. Automat.*, 2021, pp. 13909–13915.

[2] S. Stevšić, S. Christen, and O. Hilliges, "Learning to assemble: Estimating 6D poses for robotic object-object manipulation," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 1159–1166, Feb. 2020.

[3] T. Hodaň, J. Matas, and Š. Obdržálek, "On evaluation of 6D object pose estimation," in *Proc. 2016 Comput. Vis.–ECCV Workshops*, Amsterdam, The Netherlands, Oct. 8-10 and 15-16, 2016, Proceedings, Part III 14. Springer, pp. 606–619.

[4] T. Hodan et al., "T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects," in *Proc. 2017 IEEE Winter Conf. Appl. Comput. Vis.*, 2017, pp. 880–888.

[5] C. Wu, L. Chen, Z. He, and J. Jiang, "Pseudo-Siamese graph matching network for textureless objects'6-D pose estimation," *IEEE Trans. Ind. Electron.*, vol. 69, no. 3, pp. 2718–2727, Mar. 2022.

[6] T. Hodan, D. Barath, and J. Matas, "Epos: Estimating 6D pose of objects with symmetries," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2020, pp. 11703–11712.

[7] J. Richter-Klug and U. Frese, "Handling object symmetries in CNN-based pose estimation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 13850–13856.

[8] S. Peng et al., "Pvnet: Pixel-wise voting network for 6DoF pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4561–4570.

[9] J. Chang et al., "Ghostpose: Multi-view pose estimation of transparent objects for robot hand grasping," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 5749–5755.

[10] I. Lysenkov and V. Rabaud, "Pose estimation of rigid transparent objects in transparent clutter," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2013, pp. 162–169.

[11] S. Sajjan, M. Moore, and e. a. Pan, "Clear grasp: 3D shape estimation of transparent objects for manipulation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 3634–3642.

[12] X. Chen, H. Zhang, Z. Yu, A. Opipari, and O. Chadwicke Jenkins, "Clearpose: Large-scale transparent object dataset and benchmark," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 381–396.

[13] L. Zhu et al., "RGB-D local implicit function for depth completion of transparent objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4649–4658.

[14] X. Liu, R. Jonschkowski, A. Angelova, and K. Konolige, "Keypose: Multi-view 3D labeling and keypoint estimation for transparent objects," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2020, pp. 11602–11610.

[15] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6D object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16611–16621.

[16] T. Hodan et al., "Bop: Benchmark for 6D object pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 19–34.

[17] H. Yu, S. Li, H. Liu, C. Xia, W. Ding, and B. Liang, "TGF-Net: Sim2real transparent object 6D pose estimation based on geometric fusion," *IEEE Robot. Automat. Lett.*, vol. 8, no. 6, pp. 3868–3875, Jun. 2023.

[18] G. Chen, K. Han, and K.-Y. K. Wong, "TOM-Net: Learning transparent object matting from a single image," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2018, pp. 9233–9241.

[19] D. E. Zongker, D. M. Werner, B. Curless, and D. H. Salesin, "Environment matting and compositing," in *Proc. Seminal Graph. Papers: Pushing Boundaries*, vol. 2, 2023, pp. 537–546.

[20] P.-J. Lai and C.-S. Fuh, "Transparent object detection using regions with convolutional neural network," in *Proc. IPPR Conf. Comput. vision, graphics, image Process.*, vol. 2, 2015.

[21] E. Xie, W. Wang, W. Wang, M. Ding, C. Shen, and P. Luo, "Segmenting transparent objects in the wild," in *Proc. Comput. Vis.–ECCV 2016 16th Euro. Conf.*, Glasgow, U.K., Aug. 23–28, 2020, Proceedings, Part XIII 16. Springer, 2020, pp. 696–711.

[22] A. Kalra et al., "Deep polarization cues for transparent object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8602–8611.

[23] J. Ichnowski, Y. Avigal, J. Kerr, and K. Goldberg, "Dex-NeRF: Using a neural radiance field to grasp transparent objects," 2021, *arXiv:2110.14217*.

[24] C. Wang et al., "Densefusion: 6 D object pose estimation by iterative dense fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2019, pp. 3343–3352.

[25] H. Zhang et al., "Transnet: Category-level transparent object pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 148–164.

[26] Z. Li, G. Wang, and X. Ji, "Cdpn: Coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7678–7687.

[27] Y.-Y. Chuang et al., "Environment matting extensions: Towards higher accuracy and real-time capture," in *Proc. 27th Annu. Conf. Comput. Graph. interactive Techn.*, 2000, pp. 121–130.

[28] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

[29] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.

[30] Y. Su et al., "Zebrapose: Coarse to fine surface encoding for 6DoF object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6738–6748.

[31] T. Lin et al., "Microsoft COCO: Common objects in context," CoRR, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[32] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," *Adv. neural Inf. Process. Syst.*, vol. 32, 2019.

[33] M. Sundermeyer et al., "BOP challenge 2022 on detection, segmentation and pose estimation of specific rigid objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2785–2794.

[34] L. Liu et al., "On the variance of the adaptive learning rate and beyond," 2019, *arXiv:1908.03265*.

[35] M. Zhang, J. Lucas, J. Ba, and G. E. Hinton, "Lookahead optimizer: K steps forward, 1 step back," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[36] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

[37] M. Denninger et al., "Blenderproc2: A procedural pipeline for photorealistic rendering," *J. Open Source Softw.*, vol. 8, no. 82, 2023, Art. no. 4901, doi: 10.21105/joss.04901.

[38] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, pp. 33–51, 1975.

[39] S. Hinterstoisser et al., "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes," in *Proc. Comput. Vis.–ACCV 2012 11th Asian Conf. Comput. Vis.*, Daejeon, Korea, Nov. 5-9, 2012, Revised Selected Papers, Part I 11. Springer, 2013, pp. 548–562.

[40] T. Hodaň et al., "BOP challenge 2020 on 6 D object localization," in Glasgow, U.K., Aug. 23–28, 2020, Proceedings, Part II 16. Springer, 2020, pp. 577–594.

[41] Y. R. Wang et al., "MVTrans: Multi-view perception of transparent objects," in *Proc.2023 IEEE Int. Conf. Robot. Automat.*, 2023, pp. 3771–3778.

[42] B. O. Community, "Blender - a 3D modelling and rendering package, blender foundation, stichting blender foundation," Amsterdam, 2018. [Online]. Available: http://www.blender.org

[43] T. Yamamoto et al., "Development of the research platform of a domestic mobile manipulator utilized for international competition and field test," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 7675–7682.

[44] H. Gupta, S. Thalhammer, M. Leitner, and M. Vincze, "Grasping the inconspicuous," 2022, *arXiv:2211.08182*.