

Leveraging Global Stereo Consistency for Category-Level Shape and 6D Pose Estimation from Stereo Images

Junning Qiu^{1,2,3*} Minglei Lu^{4*} Fei Wang^{1,2,3} Yu Guo^{1,2,3†} Yonggen Ling^{4†}

¹National Key Laboratory of Human-Machine Hybrid Augmented Intelligence

²National Engineering Research Center for Visual Information and Applications

³Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University ⁴Tencent Robotics X

Abstract

Stereo-based category-level shape and 6D pose estimation methods have the potential to generalize to a wider range of materials than RGBD methods, which often suffer from depth measurement errors. However, without explicit depth from two views, parameters to be estimated can become inherently entangled, negatively impacting performance. To address this, we propose a method that leverages global stereo consistency to constrain optimization directions and mitigate parameter entanglement. We first estimate an intra-category occupancy field to represent a unified shape across views, ensuring consistency and preventing shape ambiguity. Through a divide-and-conquer approach within global shape fitting, we fit this shape to stereo images to obtain the pose, iteratively rendering normalized depth maps and exchanging information across views. This approach improves convergence toward the correct pose and scale. We validated our method on both depth-friendly and depth-challenging materials using our S-RGBD dataset and the TOD benchmark. Our method surpasses RGBD methods on challenging objects and performs comparably on depth-friendly ones. Ablation studies confirm the effectiveness of each component.

1. Introduction

Category-level shape and 6D pose estimation involves estimating 6D pose, scale and shape for unseen objects within a known set of categories. This research area has attracted extensive attention due to its crucial importance in multiple fields such as AR/VR [30, 46], robotics [55], and 3D scene understanding [7]. Different from traditional instance-level 6D pose estimation [12, 15, 18, 24, 35, 50], category-level 6D pose estimation [20, 27, 45, 56] is more challenging as it breaks away from the reliance on a specific

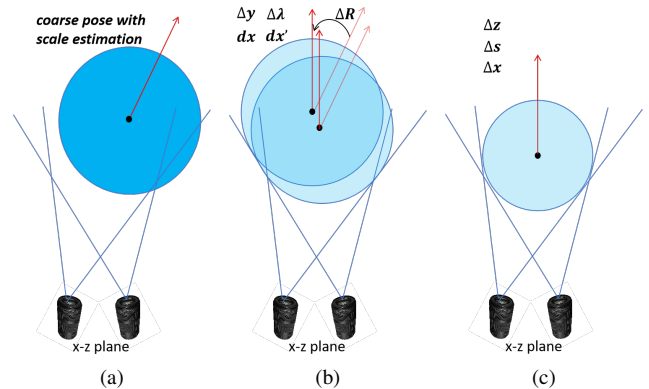


Figure 1. Schematic diagram of how our method optimizes pose and size after the shape has been estimated. (a) Coarse estimate results. (b) The optimization allows fitting shapes to the image in both left and right eye views. Note that $\lambda = \frac{s}{z}$ is related to the 2D image projection. (c) Introduce baseline relationship to remove ambiguity and make global fitting.

3D CAD model for each object. To address the aforementioned challenges, Wang et al. [45] pioneered the use of Normalized Object Coordinate Space (NOCS) to describe intra-category shape, predicting NOCS and utilizing depth information with the Umeyama algorithm [40] to estimate scale and 6D pose. Subsequently, a wave of RGBD methods emerged. Some of these continued to use NOCS or its variants [2, 9, 25, 27, 39, 42], while some directly regressed 6D pose and scale to improve efficiency [4, 6, 26, 28, 58, 59]. Others explored alternative approaches such as [5, 22].

Although RGBD methods have shown promising results, most of them heavily rely on depth sensors [3, 56, 60], which hinders their applicability to a wide range of objects. Consequently, methods that do not require depth input have garnered attention. Challengingly, the lack of depth information not only complicates shape estimation, but also introduces scale ambiguity [60]. To address these issues, most monocular RGB methods [8, 20, 47] use NOCS to describe

*Equal contribution.

†Corresponding author.

category-level shapes and then use estimated metric depth or normals to eliminate scale ambiguity. In contrast, state-of-the-art methods based on stereo image [3, 56] utilize an additional view with a known transformation to establish stereo image matching and perceptually capture 3D information, either explicitly [3] or implicitly [56], for estimating category-level 6D pose.

However, without explicit depth information from the two views, the parameters to be estimated may become entangled, causing the optimization process to get stuck in local optima easily and limiting the direct use of optimization methods like the Umeyama algorithm [40], which are feasible in RGBD approaches. To address this, we propose a novel method that leverages the fundamental constraint of global stereo consistency to mitigate parameter entanglement to estimate category-level shape and 6D pose, as shown in Figure 1. For shape representation, we use a shared intra-category occupancy field, based on the inherent consistency between the left and right views preventing shape ambiguity, with priors from a category embedding. For pose and scale estimation, our Global Shape Fitting Network (GSF-Net) employs a coarse-to-fine strategy, using the high-efficiency Adaptive Occupancy Query to iteratively render a normalized depth map of the current pose as guidance for estimating correction offsets. GSF-Net leverages global stereo consistency and cross-view referencing to estimate each parameter in a divide-and-conquer manner, constraining optimization directions to improve convergence toward accurate pose and scale.

Existing stereo methods mainly outperform RGBD methods in the TOD [29] benchmark, which primarily includes glass objects, but lack a dataset for broader material evaluation. This raises an important question: Which approach, stereo or RGBD methods, is better suited for handling a wide variety of material types that impact depth measurement quality? To investigate, we collected the Stereo-RGBD (S-RGBD) dataset using two parallel-mounted RGBD cameras, covering 10 object categories with depth-friendly and depth-challenging materials across 50 scenes, to evaluate stereo and RGBD methods. Our method achieves state-of-the-art (SOTA) performance among stereo approaches, excelling on depth-challenging objects where RGBD methods struggle, and performing comparably to RGBD on depth-friendly objects. These results underscore our method’s strengths and the versatility of stereo methods across diverse material types. Ablation studies confirm the effectiveness of our components. The main contribution of this paper is as follows:

1. We leverage global stereo consistency to propose a novel method for category-level shape and 6D pose estimation from stereo images, constraining optimization directions to mitigate parameter entanglement.
2. We propose two key designs: Intra-Category Occupancy

Field to achieve consistent shape representation across views, and GSF-Net to obtain accurate 6D pose and scale by iterative global shape fitting.

3. We collected the S-RGBD dataset to more comprehensively evaluate stereo and RGBD category-level 6D pose estimation methods. Experiments on the S-RGBD and TOD benchmarks validate the effectiveness of our approach and demonstrate the potential of stereo methods to generalize across diverse object materials.

2. Related Work

2.1. Deep Learning Based on Stereo Camera

Stereo cameras estimate depth by computing object disparity, akin to human vision, and are used in autonomous driving, robotics, and 3D scene reconstruction [14, 23, 38, 43]. Given that disparity can be derived by matching the left and right views of a calibrated stereo camera, stereo matching has emerged as a significant and enduring research focus. Stereo matching models can be broadly divided into CNN-based and Transformer-based frameworks. Historically, the majority of research has concentrated on CNN-based models, which construct a 3D cost volume [11, 31, 49] to compute the disparity between corresponding pixels in stereo images. Another approach involves concatenating features from both images to create a 4D cost volume [1, 16], followed by cost aggregation to refine the volume. In contrast, Transformer-based models [38, 48] adopt a different strategy. These models leverage cross-attention layers to facilitate information transfer between different views [52], thereby eliminating the need to generate a correlation volume. Different from the above idea of perceiving 3D information by matching, we restore the scale and pose by fitting the estimated intra-category shape to stereo images.

2.2. Category-Level 6D Pose Estimation

Category-level pose estimation aims to determine the pose of previously unseen object instances without relying on their 3D models. Currently, representing intra-category shape variation and resolving scale ambiguity remain significant challenges in this field. The concept of Normalized Object Coordinate Space (NOCS) [45] provides a common coordinate framework for instances within the same category, mitigating the shape variation encountered during pose estimation. Some subsequent methods [2, 9, 25, 27, 39, 42] predict NOCS or its variants to establish 3D-3D correspondences, which can be used to estimate the pose. Alternatively, other approaches [4, 6, 26, 28, 58, 59] aim to enhance efficiency by directly regressing the 9D pose. While these approaches have demonstrated promising outcomes, their applicability is constrained by a heavy reliance on depth observations. Therefore, some efforts have turned to methods that do not require deep data input, such as

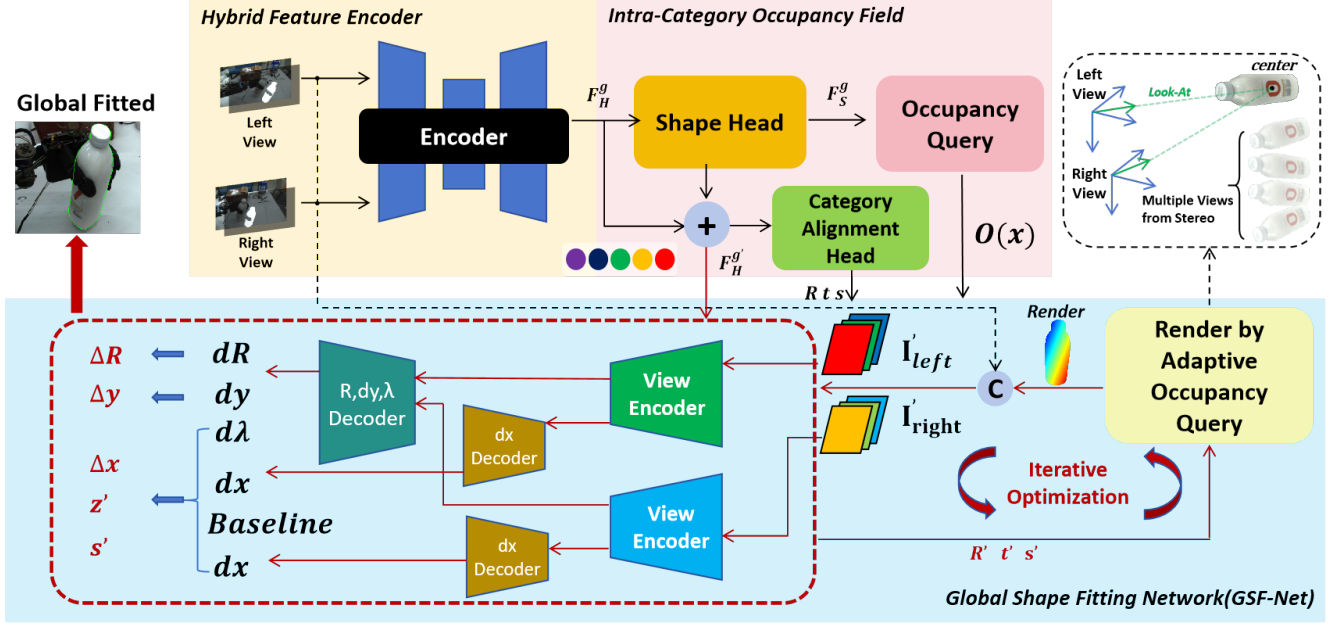


Figure 2. The pipeline of our method. First, Hybrid Feature Encoder is used to obtain features, then Intra-Category Occupancy is predicted by Shape Head to represent the consistent shape across left and right views. Category Alignment Head estimates initial coarse pose and size. Finally, GSF-Net obtains the pose and scale through global shape fitting, and uses global stereo consistency to limit meaningless optimization directions.

monocular RGB. Methods like MSOS[20] and OLDNet [8] estimate poses using the Umeyama algorithm [40], leveraging 3D-3D correspondences derived from NOCS coordinate predictions and metric depth estimation. Additionally, DMSR [47] utilizes object normals and relative depth information obtained from a pretrained DPT model [36] as supplementary inputs, estimating poses through 2D-3D correspondences via the PnP algorithm [21].

Additionally, stereo images are an option where depth is not required. As one of the representative works, Stereopose [3] predicts NOCS maps [45] in the left and right views to represent intra-category shape variation and then resolve scale ambiguity by pixel-level NOCS maps matching. CODERS [56] proposes an Implicit Stereo Matching module that implicitly encodes the 3D depth information for downstream heads. Different from previous methods, we leverage the fundamental constraint of global stereo consistency to achieve accurate category-level shape and 6D pose estimation from stereo images.

3. Methodology

3.1. Overview

Given a pair of stereo images $I_{\text{left}}, I_{\text{right}}$ acquired through a calibrated stereo camera with known baseline length and intrinsic parameters, along with corresponding masks $M_{\text{left}}, M_{\text{right}}$, our goal is to estimate the 3D rotation $\mathbf{R} \in$

$SO(3)$, 3D translation $\mathbf{t} \in \mathbb{R}^3$, and scale of the normalized shape $s \in \mathbb{R}^+$ of the target object, which belongs to known categories. The shape is implicitly represented by an occupancy field.

To achieve our goal, we propose a novel method leveraging global stereo consistency. We first use Hybrid Feature Encoder to extract semantic and geometric features from stereo images and fuse them into a hybrid embedding for subsequent use. We then estimate Intra-Category Occupancy Field shared by the left and right views. Finally, GSF-Net estimates the 6D pose and scale by globally fitting the occupancy field to the stereo images. An overview of our method is shown in Figure 2 and will be explained in detail in the following sections.

3.2. Hybrid Feature Encoder

3.2.1 Semantic Embedding

To capture rich semantic information, we leverage DINOv2 [34], a foundation model proven to extract robust features that support zero-shot semantic correspondences [5, 57]. To effectively utilize DINOv2’s features, we introduce a disparity attention mechanism, where features from left and right views, F_{DINO}^L and F_{DINO}^R , establish attention [41] relationships only within matching pixel rows, following principles similar to stereoscopic disparity mechanisms. Specifically, using a disparity attention mask, attention is restricted to the same object mask regions in

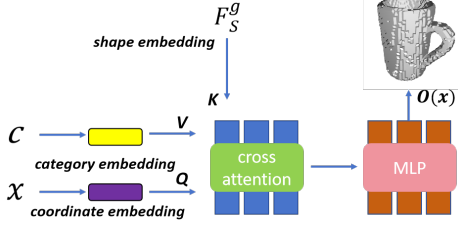


Figure 3. Diagram of the querying in Intra-Category Occupancy Field. Shape, positional, and category embeddings enter a cross-attention mechanism, followed by an MLP to yield the query point’s occupancy value.

the left and right views for semantic features, ensuring that feature interactions occur only within corresponding object regions. For geometric features, attention is constrained to the entire object crop region, maintaining structural consistency. In both cases, attention is limited to the same row of the feature map within their respective regions, preserving the fundamental disparity characteristics of stereoscopic vision. These fused features are then processed through adaptive pooling with varying receptive field sizes to capture both local and global information, producing the final embedding F_{DINO}^g .

3.2.2 Geometry Embedding

To capture geometry information, we use normalized depth predict transformer (NPT) based on DPT [36, 53] to estimate the normalized depth map on the object surface. To maintain the structure of stereo images, a cropping strategy similar to KeyPose [29] is applied, with CNNs extracting left and right view features separately. Similar to the pooling method in Section 3.2.1, we yield the geometry embedding F_{Geo}^g .

At this stage, we have obtained semantic and geometric embeddings from the stereoscopic view. We integrate them using an attention mechanism [41], resulting in the final global hybrid embedding F_H^g that captures comprehensive object information. Specifically, cross-attention is employed, where the semantic embedding serves as the Query and the geometric embedding as the Key and Value to extract relevant geometric information. The roles are then reversed to extract relevant semantic information. Finally, the extracted features from both steps are concatenated to produce global hybrid embedding.

3.3. Intra-Category Occupancy Field

To avoid the optimization difficulties caused by shape ambiguity, we propose the Intra-Category Occupancy Field, which implicitly represents object occupancy [32] as a consistent shape in a category-aligned, normalized coordinate system in the shared space of left and right views. Specifically, we pass the global hybrid feature embedding F_H^g

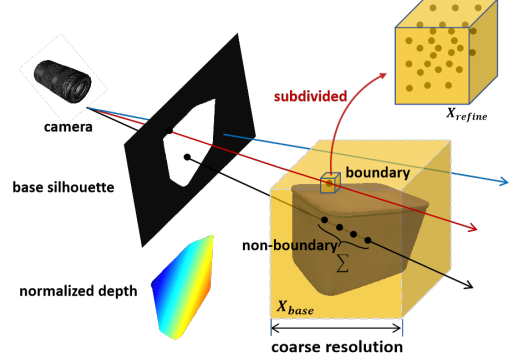


Figure 4. Schematic diagram about Adaptive Occupancy Query. First, confirm the boundary at a coarse resolution, then subdivide the grid to refine query resolution around the boundary, and finally render the depth map.

through the Shape Head, which consists of MLPs, to obtain the global shape embedding:

$$F_S^g = \text{Shape Head}(F_H^g) \quad (1)$$

Let $x = (x_1, x_2, x_3)$ be a vector, where each component x_i is within the range $[-\frac{L_{\text{nl}}}{2}, \frac{L_{\text{nl}}}{2}]$, for $i = 1, 2, 3$.

We apply cross-attention similar in ZeroShape [13] as shown in Figure 3, where the positional embedding of x , denoted as $\text{PosEmb}(x)$, is used as the query, F_S^g as the key, and the category embedding $\text{CatEmb}(c)$ as the value, with c representing the object category:

$$z = \text{CrossAttn}(\text{PosEmb}(x), F_S^g, \text{CatEmb}(c)) \quad (2)$$

The final fused feature z is then passed through a MLP to predict the occupancy value for each x :

$$O(x) = \sigma(\text{MLP}(z)) \quad (3)$$

where σ denotes the sigmoid function, producing occupancy prediction between 0 and 1 for each spatial location.

In order to better preserve the intra-category shape feature, we use a simple addition to supplement the shape features:

$$F_H^{g'} = F_H^g + F_S^g$$

To obtain a coarse category-level pose and scale as initial values, we propose the use of a Category Alignment Head composed of MLPs. Following the 6D rotation representation approach from [44, 56], a 6-dimensional vector $R_{6d} = [r_1 | r_2] \in \mathbb{R}^6$ is predicted and converted into an initial rotation matrix $R^{(0)} \in \mathbb{R}^{3 \times 3}$ through vector normalization.

In this approach, the initial estimates for 6D pose and scale are obtained for a reference view, typically the left eye. The estimations are as follows:

$$R^{(0)}, t^{(0)}, s^{(0)} = \text{Category Alignment Head}(F_H^{g'})$$

Table 1. Experimental Results on the TOD Dataset (higher is better).

Method	Input	Bottle						Mug					
		$3D_{25}$	$3D_{50}$	$3D_{75}$	5° 2cm	10° 5cm	10° 10cm	$3D_{25}$	$3D_{50}$	$3D_{75}$	5° 2cm	10° 5cm	10° 10cm
SPD	RGBD	44.5	7.4	-	-	11.5	17.8	63.6	19.7	-	-	2.3	4.2
SGPA		46.9	9.6	-	-	13.3	22.5	64.6	19.6	-	-	2.8	5.1
KeyPose	Stereo	-	-	-	-	52.7	62.3	-	-	-	-	24.6	25.1
StereoPose		85.4	22.2	-	-	57.8	70.3	97.9	77.4	-	-	34.4	38.2
CODERS		100	99.5	31.5	73.4	99.8	99.8	100	100	92.8	56.2	81.9	81.9
Ours		100	99.8	64.1	84.2	99.9	100	100	100	95.9	79.6	90.8	95.2

Table 2. Evaluation of $3D_{avg}$ performance (higher is better) on the S-RGBD dataset, assessing objects with diverse material types.

Method	Input	Bowl	Box	Can	Cup	Mug	Bottle	Knife	Spoon	Plate	Wine Glass
SPD [39]	RGBD	56.5	41.3	59.2	48.6	54.2	41.5	22.3	16.3	59.1	21.8
SGPA [2]		63.7	57.4	58.6	65.2	57.9	56.3	32.4	26.4	63.4	22.5
Second [5]		86.2	91.2	84.7	83.8	78.8	87.8	49.1	54.3	84.1	24.1
KeyPose [29]	Stereo	45.4	40.3	39.1	34.8	32.9	44.3	47.5	41.2	40.3	40.4
StereoPose [3]		54.2	47.3	59.1	58.5	64.8	63.3	58.1	52.8	41.9	53.4
Ours		82.9	89.7	84.9	85.3	80.2	86.2	67.9	63.5	81.2	74.1

where $R^{(0)} \in \mathbb{R}^{3 \times 3}$ is the initial rotation matrix, and $\mathbf{t}^{(0)} = [t_x^{(0)}, t_y^{(0)}, t_z^{(0)}] \in \mathbb{R}^3$ is the initial translation vector, both estimated for the reference view (left eye). The scale $s^{(0)} \in \mathbb{R}$ is shared across views, representing the object’s initial scale.

3.4. Global Shape Fitting Network(GSF-Net)

GSF-Net renders the occupancy field to represent the current estimation result, iteratively fits the shape globally to correct the pose and scale, and uses global stereo consistency to constrain optimization.

Adaptive Occupancy Query To balance the computational cost of high query resolution and the silhouette blurring caused by low resolution, we employ **adaptive occupancy query** for rendering normalized depth as shown in Figure 4. Define the rendering function Render with inputs as the object’s pose R , scale s , the grid set X , and the corresponding occupancy set $O(X)$, where each grid cell’s center coordinates are queried to obtain $O(X)$. The implementation calculates each pixel’s silhouette value by applying a sigmoid function to a coarsely rendered depth map. First, we render a base depth map

$$D_{\text{base}} = \text{Render}(R, t, s, X_{\text{base}}, O(X_{\text{base}}))$$

over the base grid X_{base} , which is divided at a relatively coarse resolution. By applying a sigmoid function to D_{base} , we obtain the initial silhouette S_{base} . Then, the regions within X_{base} that contribute to the silhouette boundary are

further subdivided to form a refined grid X_{refined} , capturing finer details within those areas. Next, we only need to query $O(X_{\text{refined}})$ for these additional points. By merging these into the original set, we obtain the adaptively refined normalized depth map:

$$D = \text{Render}(R, t, s, X_{\text{base}} \cup X_{\text{refined}}, O(X_{\text{base}}) \cup O(X_{\text{refined}}))$$

View Encoder To enhance the model’s awareness of deviations in the current estimation with all view, we use the ‘lookat’ operation [19, 44] and follow the cropping method in Megapose [19] to crop $\{I_{\text{view}}, D_{\text{view}}, D_{\text{view}}^{\text{lookat}}\}$ where ‘view’ can represent either the left or right view-point. These cropped images and masks are then concatenated to form I'_{view} with dimensions $\mathbb{R}^{H' \times W' \times 6}$. On this basis, CNNs are used to extract features, which are then integrated with the embedding F_H^{gl} through an attention mechanism [41] to obtain f_{left} and f_{right} .

Shared Rotation, Y-Translation, and λ Decoder To reduce unnecessary optimization directions, we share the optimization of rotation, y-axis translation, and λ across the left and right views, as these components should remain consistent. Here, $\lambda = \frac{s}{z}$, which, through simple derivation, can be shown to influence the scale of the 3D-to-2D projection pattern, as similarly discussed in [44]. By inputting feature maps from the left and right views, f_{left} and f_{right} , into the Shared Decoder_{RyL} composed of MLPs, we estimate the offsets for these three components:

$$dR, dy, d\lambda = \text{Decoder}_{\text{RyL}}(f_{\text{left}}, f_{\text{right}}) \quad (4)$$

Table 3. Evaluation of different methods based on depth quality influenced by material types using $3D_{avg}$ metric (higher is better).

Materials	Object	RGBD Methods		Stereo Methods	
		SGPA [2]	Second [5]	StereoPose [3]	Ours
Depth-Friendly	Bowl	70.5	93.3	72.3	87.6
	Cup	68.3	91.3	74.1	86.9
	Spoon	38.4	72.6	51.6	73.7
	Bottle	41.5	89.4	52.3	84.9
Depth-Challenging	Mug	11.5	26.3	62.6	81.2
	Plate	14.7	35.1	37.2	82.6
	Box	15.7	32.1	41.3	67.7
	Knife	13.4	31.6	40.2	66.4

Independent X-Translation Decoder Due to the coupling of $\lambda = \frac{s}{z}$, scale ambiguity is not fully resolved. As a result, when the scale remains incorrect, the x-axis translation cannot be globally accurate. To address this, we estimate independent x-axis offsets for the left and right views as preparation for further refinement. Decoder_x composed of MLPs is employed to estimate these offsets, representing the left and right views, respectively. The inputs are f_{left} and f_{right} , with separate, unshared weights for each decoder:

$$dx_{left} = \text{Decoder}_x(f_{left}), \quad dx_{right} = \text{Decoder}_x(f_{right})$$

These equations define the outputs for each decoder, capturing the independent x-axis offsets for the left and right views.

Merge offsets With the offsets from the shared and independent decoders, we integrate the results to determine scale and depth. Using the stereo disparity $d = x_{left}^{(pixel)} - x_{right}^{(pixel)}$ and the known baseline B , we calculate depth as $z = \frac{f \cdot B}{d}$, where f is the focal length. Given z , the scale s is computed as $s = \lambda \cdot z$, with λ from the shared decoder. The independent x-axis offsets dx_{left} and dx_{right} refine alignment in each view, serving as final adjustments without replacing the stereo disparity for depth. This approach ensures accurate global alignment by consistently determining z and s . The corrected estimations (R' , t' , s') for each view are then iteratively calculated by applying these offsets or update.

If additional correction is needed, the offsets of the current result can be further estimated.

3.5. Loss Function

We employ a three-stage training approach for the model. First, we pre-train the Shape Head; then, we expand training to include the Category Alignment Head. In the final stage, we introduce the GSF-Net to complete the training process.

For shape head loss L_{shape} , we adopt 3D occupancy loss from [13]. The L_{aglin} for category alignment head follows CODERS [56] and MSE loss respectively. The GSF-Net loss L_{GSF} and the training method are similar to Mega-pose [19]. The overall loss function is defined as:

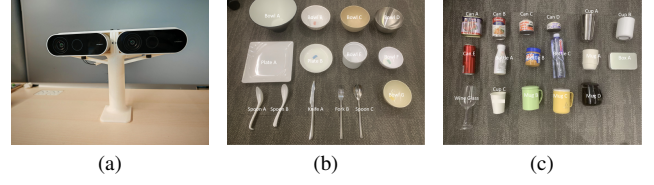


Figure 5. Schematic of S-RGBD dataset collection. (a) Parallel-mounted RGB-D cameras. (b) and (c) are selected objects.

$$\mathcal{L} = \alpha_1 L_{shape} + \alpha_2 L_{aglin} + \alpha_3 L_{GSF}$$

where α_1 , α_2 and α_3 are balancing coefficients that control the influence of each loss term on the total loss $\mathcal{L}$.

4. Stereo-RGBD(S-RGBD) Dataset

The stereo category-level pose estimation benchmark TOD [29] primarily includes transparent objects. However, this dataset alone does not address the question we raised¹. To this end, we collected the Stereo-RGBD (S-RGBD) Dataset using two parallel-mounted RGBD cameras, as shown in Table 5 (a). This dataset encompasses 10 object categories across 50 scenes, with parts of it shown in Table 5 (b) and (c), featuring various material types—such as transparent, reflective, dark, and commonly textured surfaces—that result in a range of depth imaging qualities. This diversity provides a robust foundation for evaluating both stereo and RGBD methods. The target model is acquired through scanning and aligned according to category division. We use scattered AprilTag [33] to determine camera poses, employ the visual foundation model Segment Anything Model 2 (SAM 2) [37] to generate masks, and optimize object poses in the video using PyTorch3D’s differentiable rendering, aligning projected 3D model masks with SAM 2 masks while considering occlusion. Manual adjustments are performed every 20 frames, with differentiable rendering aiding further refinement.

5. Experiment

5.1. Experimental Setup

Datasets We train and test our method on the TOD [29] and our new S-RGBD dataset. The TOD dataset includes 36,000 stereo image pairs of 12 transparent object instances (bottles, mugs, and cups) across 10 backgrounds. Our S-RGBD dataset has been described in detail in Sec 4. The provided stereo images and their corresponding depth maps allow for a fair evaluation of stereo and RGBD methods. All test sets explicitly separate objects from training instances of the same category, covering evaluation, ablation,

¹Which approach, stereo or RGBD methods, is better suited for handling a wide variety of material types that impact depth measurement quality?

Table 4. Ablation studies on different settings using 3DIOU Metrics (higher is better) on the S-RGBD dataset.

Row	Method	$3D_{25}$	$3D_{50}$	$3D_{75}$
AS-1	w/ MonoVersion	45.7	30.5	7.6
	w/o global fitting	42.4	32.4	9.6
	w/o shape constraint	94.1	78.1	49.2
	w/o R constraint	93.4	76.5	49.8
	w/o y constraint	94.1	77.5	51.3
	w/o λ constraint	90.4	71.1	44.2
AS-2	w/ Semantic Only	93.1	74.6	46.4
	w/ Geometric Only	94.4	77.7	49.3
AS-3	w/ Low-resolution	80.2	65.3	23.8
	w/ High-resolution	90.4	74.2	38.6
AS-4	w/ Joint Training	60.4	40.6	5.7
	w/ Separate Training	94.2	75.3	46.3
	Ours (baseline)	95.6	83.7	60.5

and robotics experiments, aiming for fair category-level assessment.

Evaluation Metrics We quantitatively evaluated the performance of category-level object pose estimation using widely adopted metrics from [2, 3, 6, 45, 51, 54, 56]. We measured 3D Intersection over Union (3DIOU) to evaluate the overlap between predicted and ground truth 3D bounding boxes, reporting mean precision at thresholds $3D_{25}$, $3D_{50}$, and $3D_{75}$, corresponding to 25%, 50%, and 75% overlap, respectively. Additionally, $3D_{avg}$ denotes the average precision across these thresholds. To evaluate prediction accuracy, we measure rotation error within thresholds of $\{5^\circ, 10^\circ\}$ and translation error within thresholds of $\{2\text{ cm}, 5\text{ cm}, 10\text{ cm}\}$. A prediction is deemed accurate only if it meets the specified thresholds for both rotation and translation errors.

Implement Details We froze the ViT-B version of DINOv2 [34] and pre-trained DPT [36, 53] throughout the training process, using Adam [17] as the optimizer with an initial learning rate of 1.0×10^{-4} decayed by cosine annealing. The training was divided into three stages. In the first stage, only L_{shape} with $\alpha_1 = 1$ was used. In the second stage, L_{align} with $\alpha_2 = 100$ was added. In the third stage, L_{GSF} with $\alpha_3 = 10$ was introduced. These setups effectively avoid the instability of multitask training and benefit from end-to-end training. All models are trained for 30 epochs on four RTX8000 GPUs with a batch size of 8 and tested on one RTX3090 GPU. All results were obtained after 5 iterations of GSF-Net.

5.2. Performance on Benchmark TOD

Experiments are conducted on the "cup" and "bottle" categories, following [3, 29, 56]. Our method outperforms all other methods on all metrics (see Tab 1), demonstrating its effectiveness. Notably, it outperforms the RGBD methods SPD [39] and SGPA [2] on transparent objects, even though

they use deep refinement [10]. Compared to other stereo methods, we exploit global stereo consistency to make the optimization very effective, achieving good results on objects of any material.

5.3. Performance on S-RGBD Dataset

Experiments on the S-RGBD dataset show that our method achieves state-of-the-art performance on 10 object categories with various materials. The visualization results are shown in Figure 7. We used TransCG [10] for depth refinement in RGBD methods on S-RGBD, as in TOD, since without correction, depth-challenging cases are nearly ineffective—similar to [3]. Unlike RGBD methods that rely on high-quality depth data, our method exhibits higher robustness and outperforms other stereo methods, as shown in Table 2. This highlights its potential to generalize to objects with different depth qualities. In addition, this shows that leveraging global consistency can effectively help stereo images complete the optimization from normalized shape to pose and size.

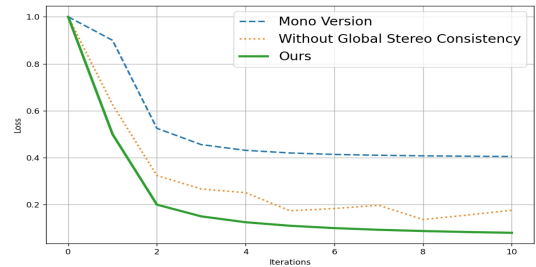


Figure 6. Comparison of convergence speed of different strategies. Reasonable use of global stereo consistency can help the estimated parameters converge quickly globally.

5.4. Ablation Study

[AS-1] Effectiveness of global stereo consistency for constraint To evaluate the effectiveness of global stereo consistency for constraint, we tested six variants. *MonoVersion* refers to the monocular version of our method shown in Table 4 row AS-1. Without global fitting, making it difficult for the left and right views to accurately recover scale through shared information; without shape constraint, the shape is prone to ambiguity; without R and y constraint, a single view can easily fall into local optima. Lack of λ constraint complicates optimization and makes it harder to leverage image information for fitting. Furthermore, we demonstrate that adding constraints can lead to faster iteration convergence, as shown in Figure 6.

[AS-2] Embeddings from Different Sources in The Encoder To verify the necessity of these two embeddings, we trained the proposed model in two variants: (i) with only semantic features and (ii) with only geometric features. Results, shown in Table 4 row AS-2, indicate that removing

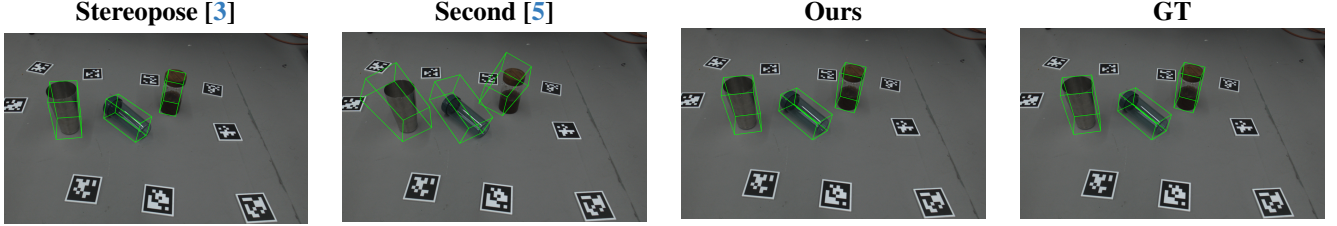


Figure 7. Visual comparison with other methods on the S-RGBD dataset. Note that stereopose [3] is a stereo-based method, which is less affected by the material but not accurate enough. Second [5] is based on RGBD, which is very difficult to deal with depth-challenging objects due to lower-quality depth map. Our method can perform well on a wide range of materials.

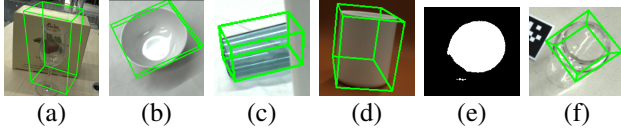


Figure 8. Illustrations of failure cases: (a) to (d) represent common error cases, while (e) and (f) correspond to errors related to masks.

either embedding significantly decreases performance.

[AS-3] Query Approach for Render To validate our Adaptive Occupancy Query, we tested (i) low-resolution and (ii) high-resolution settings. Results in Table 4 row AS-3 show that low-resolution severely hampers pose estimation, while our method surpasses high-resolution rendering by adaptively focusing on critical areas.

[AS-4] Multi-stage Learning Strategy As shown in Table 4 row AS-4, we tested (i) direct joint training and (ii) separate training. Results indicate that direct training struggles with certain objects due to instability in shape estimation, leading to incorrect pose predictions. While separate training converges successfully, the lack of end-to-end learning limits the model’s ability to fully capture the relationship between shape and pose estimation.

[AS-5] Material Robustness Evaluation on Objects of the Same Category We experimented on same-category objects with different materials in the S-RGBD dataset (Table 3) to address footnote 1. When materials support high-quality depth measurements (e.g., opaque plastic), RGBD methods perform well. However, performance drops sharply with low-quality depth measurements (e.g., transparent glass). In contrast, stereo methods show notable robustness to material variations, indicating their potential for generalization across diverse material properties.

5.5. Analysis of Failure Cases

Figure 8 analyzes failure cases: (a) transparent objects reflecting their surroundings, (b) perspective leading to height estimation errors, (c) cross-occlusion resulting in invisibility ambiguity, and (d) self-occlusion causing uncertainty in the shape and pose of objects and their components, e.g., a cup handle’s shape and position having multiple possibilities. Lower-quality masks still yield reliable results, while

Table 5. Robot grasping experiments in real-world. DF: Depth-Friendly, DC: Depth-Challenging.

Material	Success Rate (%)					
	Bowl	Bottle	Mug	Can	Spoon	Cup
DF	100%	100%	97%	93%	88%	99%
DC	97%	95%	95%	90%	84%	95%

only very poor-quality masks(e) cause failure(f).

5.6. Real Robot Experiment

The experiments are conducted on an ABB robot arm equipped with a stereo camera and an INSPIRE-ROBOTS dexterous hand. We perform real-world grasping experiments on objects of different materials using our method. Each scene contains 3-5 randomly placed objects from the S-RGBD dataset, which are sequentially grasped until all are completed. The success rate, defined as successful grasps over attempts, is reported in Table 5, demonstrating our method’s effectiveness across various materials.

6. Conclusion

In this paper, we leverage global stereo consistency to mitigate parameter entanglement by constraining optimization directions for category-level shape and 6D pose estimation. To address shape ambiguity and its impact on optimization, we estimate an intra-category occupancy field as a unified representation. We then iteratively fit this shape to stereo images to obtain pose and scale, predicting each parameter using a divide-and-conquer approach guided by consistency constraints. Additionally, we introduce the S-RGBD dataset, encompassing a variety of object categories and materials, to benchmark stereo-based and RGBD methods. Extensive experiments on the TOD benchmark, the S-RGBD dataset, and in real-world robotic applications demonstrate that our approach achieves state-of-the-art performance on complex objects with diverse materials, highlighting the robust generalization potential of stereo methods.

Acknowledgements

This work was supported in part by the National Major Science and Technology Projects of China under Grant 2009XJTU0016 and the National Key Research and Development Program of China under Grant 2022YFB3303800.

References

- [1] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 2
- [2] Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2773–2782, 2021. 1, 2, 5, 6, 7
- [3] Kai Chen, Stephen James, Congying Sui, Yun-Hui Liu, Pieter Abbeel, and Qi Dou. Stereopose: Category-level 6d transparent object pose estimation from stereo images via back-view nocs. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2855–2861. IEEE, 2023. 1, 2, 3, 5, 6, 7, 8
- [4] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales 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*, pages 1581–1590, 2021. 1, 2
- [5] Yamei Chen, Yan Di, Guangyao Zhai, Fabian Manhardt, Chenyangguang Zhang, Ruida Zhang, Federico Tombari, Nassir Navab, and Benjamin Busam. Secondpose: Se (3)-consistent dual-stream feature fusion for category-level pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9959–9969, 2024. 1, 3, 5, 6, 8
- [6] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6781–6791, 2022. 1, 2, 7
- [7] Yan Di, Chenyangguang Zhang, Ruida Zhang, Fabian Manhardt, Yongzhi Su, Jason Rambach, Didier Stricker, Xiangyang Ji, and Federico Tombari. U-red: Unsupervised 3d shape retrieval and deformation for partial point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8884–8895, 2023. 1
- [8] Zhaoxin Fan, Zhenbo Song, Jian Xu, Zhicheng Wang, Kejian Wu, Hongyan Liu, and Jun He. Object level depth reconstruction for category level 6d object pose estimation from monocular rgb image. In *European Conference on Computer Vision*, pages 220–236. Springer, 2022. 1, 3
- [9] Zhaoxin Fan, Zhenbo Song, Zhicheng Wang, Jian Xu, Kejian Wu, Hongyan Liu, and Jun He. Acr-pose: Adversarial canonical representation reconstruction network for category level 6d object pose estimation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 55–63, 2024. 1, 2
- [10] Hongjie Fang, Hao-Shu Fang, Sheng Xu, and Cewu Lu. Transcg: A large-scale real-world dataset for transparent object depth completion and a grasping baseline. *IEEE Robotics and Automation Letters*, 7(3):7383–7390, 2022. 7
- [11] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020. 2
- [12] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3003–3013, 2021. 1
- [13] Zixuan Huang, Stefan Stojanov, Anh Thai, Varun Jampani, and James M Rehg. Zeroshape: Regression-based zero-shot shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10061–10071, 2024. 4, 6
- [14] Jihong Ju, Ching Wei Tseng, Oleksandr Bailo, Georgi Dikov, and Mohsen Ghafoorian. Dg-recon: Depth-guided neural 3d scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18184–18194, 2023. 2
- [15] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE international conference on computer vision*, pages 1521–1529, 2017. 1
- [16] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE international conference on computer vision*, pages 66–75, 2017. 2
- [17] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7
- [18] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 574–591. Springer, 2020. 1
- [19] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022. 5, 6
- [20] Taeyeop Lee, Byeong-Uk Lee, Myungchul Kim, and In So Kweon. Category-level metric scale object shape and pose estimation. *IEEE Robotics and Automation Letters*, 6(4): 8575–8582, 2021. 1, 3
- [21] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, 81:155–166, 2009. 3
- [22] Guanglin Li, Yifeng Li, Zhichao Ye, Qihang Zhang, Tao Kong, Zhaopeng Cui, and Guofeng Zhang. Generative

- category-level shape and pose estimation with semantic primitives. In *Conference on Robot Learning*, pages 1390–1400. PMLR, 2023. 1
- [23] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7644–7652, 2019. 2
- [24] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018. 1
- [25] Haitao Lin, Zichang Liu, Chilam Cheang, Lingwei Zhang, Yanwei Fu, and Xiangyang Xue. Donet: Learning category-level 6d object pose and size estimation from depth observation. *arXiv preprint arXiv:2106.14193*, 2(4), 2021. 1, 2
- [26] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing 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*, pages 3560–3569, 2021. 1, 2
- [27] Jiehong Lin, Zewei Wei, Changxing Ding, and Kui Jia. Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In *European Conference on Computer Vision*, pages 19–34. Springer, 2022. 1, 2
- [28] Jiehong Lin, Zewei Wei, Yabin Zhang, and Kui Jia. Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14001–14011, 2023. 1, 2
- [29] Xingyu Liu, Rico Jonschkowski, Anelia Angelova, and Kurt Konolige. Keypose: Multi-view 3d labeling and key-point estimation for transparent objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11602–11610, 2020. 2, 4, 5, 6, 7
- [30] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651, 2015. 1
- [31] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 2
- [32] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 4
- [33] Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In *2011 IEEE international conference on robotics and automation*, pages 3400–3407. IEEE, 2011. 6
- [34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafranec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3, 7
- [35] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4561–4570, 2019. 1
- [36] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3, 4, 7
- [37] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 6
- [38] Qing Su and Shihao Ji. Chitransformer: Towards reliable stereo from cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1939–1949, 2022. 2
- [39] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 530–546. Springer, 2020. 1, 2, 5, 7
- [40] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04): 376–380, 1991. 1, 2, 3
- [41] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3, 4, 5
- [42] Boyan Wan, Yifei Shi, and Kai Xu. Socs: Semantically-aware object coordinate space for category-level 6d object pose estimation under large shape variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14065–14074, 2023. 1, 2
- [43] Cheng Wang, Xiaoxian Cui, Shijie Zhao, Konghui Guo, Yang Wang, and Yaqi Song. The application of deep learning in stereo matching and disparity estimation: A bibliometric review. *Expert Systems with Applications*, 238:122006, 2024. 2
- [44] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021. 4, 5
- [45] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas 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*, pages 2642–2651, 2019. 1, 2, 3, 7
- [46] Jiase Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In *2021 IEEE/RSJ International Confer-*

- ence on Intelligent Robots and Systems (IROS), pages 4807–4814. IEEE, 2021. [1](#)
- [47] Jiaxin Wei, Xibin Song, Weizhe Liu, Laurent Kneip, Hongdong Li, and Pan Ji. Rgb-based category-level object pose estimation via decoupled metric scale recovery. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2036–2042. IEEE, 2024. [1](#), [3](#)
- [48] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023. [2](#)
- [49] Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Song Wang, and Lili Ju. Semantic stereo matching with pyramid cost volumes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7484–7493, 2019. [2](#)
- [50] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. [1](#)
- [51] Chi Xu, Jiale Chen, Mengyang Yao, Jun Zhou, Lijun Zhang, and Yi Liu. 6dof pose estimation of transparent object from a single rgb-d image. *Sensors*, 20(23):6790, 2020. [7](#)
- [52] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Reza Tofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. [2](#)
- [53] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiao-gang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. [4](#), [7](#)
- [54] Yang You, Ruoxi Shi, Weiming Wang, and Cewu Lu. Cppf: Towards robust category-level 9d pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6866–6875, 2022. [7](#)
- [55] Guangyao Zhai, Yu Zheng, Ziwei Xu, Xin Kong, Yong Liu, Benjamin Busam, Yi Ren, Nassir Navab, and Zhengyou Zhang. Da² dataset: Toward dexterity-aware dual-arm grasping. *IEEE Robotics and Automation Letters*, 7(4):8941–8948, 2022. [1](#)
- [56] Chuanrui Zhang, Yonggen Ling, Minglei Lu, Minghan Qin, and Haoqian Wang. Category-level object detection, pose estimation and reconstruction from stereo images. *arXiv preprint arXiv:2407.06984*, 2024. [1](#), [2](#), [3](#), [4](#), [6](#), [7](#)
- [57] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. [3](#)
- [58] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation. In *European Conference on Computer Vision*, pages 655–672. Springer, 2022. [1](#), [2](#)
- [59] Ruida Zhang, Yan Di, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Ssp-pose: Symmetry-aware shape prior deformation for direct category-level object pose estimation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7452–7459. IEEE, 2022. [1](#), [2](#)
- [60] Ruida Zhang, Ziqin Huang, Gu Wang, Chenyangguang Zhang, Yan Di, Xingxing Zuo, Jiwen Tang, and Xiangyang Ji. Lapose: Laplacian mixture shape modeling for rgb-based category-level object pose estimation. *arXiv preprint arXiv:2409.15727*, 2024. [1](#)