

SCFlow2: Plug-and-Play Object Pose Refiner with Shape-Constraint Scene Flow

Qingyuan Wang¹, Rui Song¹, Jiaojiao Li¹, Kerui Cheng², David Ferstl³, Yinlin Hu³
¹ State Key Laboratory of ISN, Xidian University ² Taiyuan University of Technology ³ MagicLeap

Abstract

We introduce SCFlow2, a plug-and-play refinement framework for 6D object pose estimation. Most recent 6D object pose methods rely on refinement to get accurate results. However, most existing refinement methods either suffer from noises in establishing correspondences, or rely on retraining for novel objects. SCFlow2 is based on the SCFlow model designed for refinement with shape constraint, but formulates the additional depth as a regularization in the iteration via 3D scene flow for RGBD frames. The key design of SCFlow2 is an introduction of geometry constraints into the training of recurrent matching network, by combining the rigid-motion embeddings in 3D scene flow and 3D shape prior of the target. We train SCFlow2 on a combination of dataset Objaverse, GSO and ShapeNet, and evaluate on BOP datasets with novel objects. After using our method as a post-processing, most state-of-the-art methods produce significantly better results, without any retraining or fine-tuning. The source code is available at <https://scflow2.github.io>.

1. Introduction

6D object pose estimation, which determines an object's 3D rotation and 3D translation relative to a camera, is crucial in fields like robotics, augmented reality, and spatial computing [31, 48].

Most existing pose estimation methods rely on a refinement procedure to get accurate pose results [10, 12, 20, 23, 24, 48, 50], and most of the refinement methods are based on render-and-compare strategies [10, 20, 23, 50]. A group of refinement methods [24, 48] achieves highly accurate results based on retraining or fine-tuning the network for novel objects, which, however, is cumbersome and unfriendly to real applications. Some recent refinement methods [23, 42, 50] have shown promising results without fine-tuning for novel objects. However, they still suffer from several weaknesses.

First, most methods formulate the render-and-compare framework as a general matching problem, but there is a strong prior of the target's 3D shape that is not used, mak-

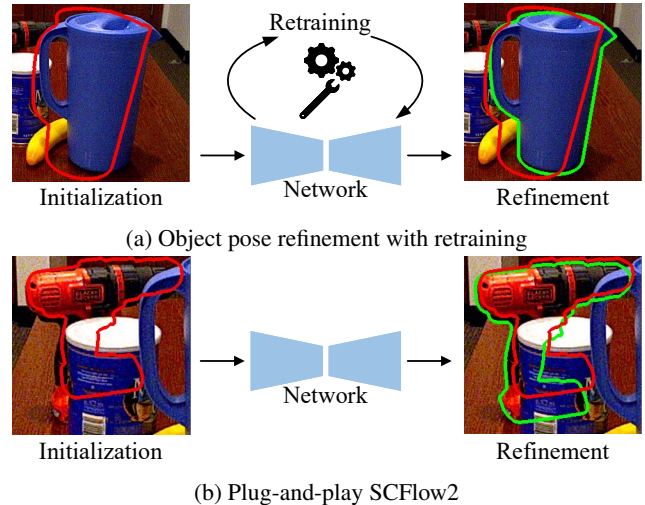


Figure 1. **Object pose refinement** is critical for accurate object pose estimation. (a) Most existing object pose refinement methods, including SCFlow [11], rely on retraining for novel objects to achieve high accuracy. (b) The proposed SCFlow2 achieves even higher accuracy, and more importantly, generalizes well to novel objects without any retraining or fine-tuning.

ing it usually has an unnecessarily large search space for matching and is less effective for 6D object pose estimation [11]. Second, to address the matching issue, most methods attempt to refine the pose starting from multiple pose hypotheses based on the initialization [23, 32, 36]. While this paradigm is effective, it significantly slows down the refinement process. Third, most methods formulate the comparison after the rendering as a 2D matching procedure only, and then use ICP or Kabsch [21] to consume additional depth input as a second stage, which is less effective in capturing 3D motions and local optimal within each stage.

We introduce SCFlow2, a plug-and-play refinement framework for 6D object pose estimation, as shown in Fig. 1. SCFlow2 is based on SCFlow [11] which is a state-of-the-art refinement method embedding 3D shape constraints into the optimization for known objects. We build on the basic shape-constraint structure of SCFlow but extend it in several novel ways.

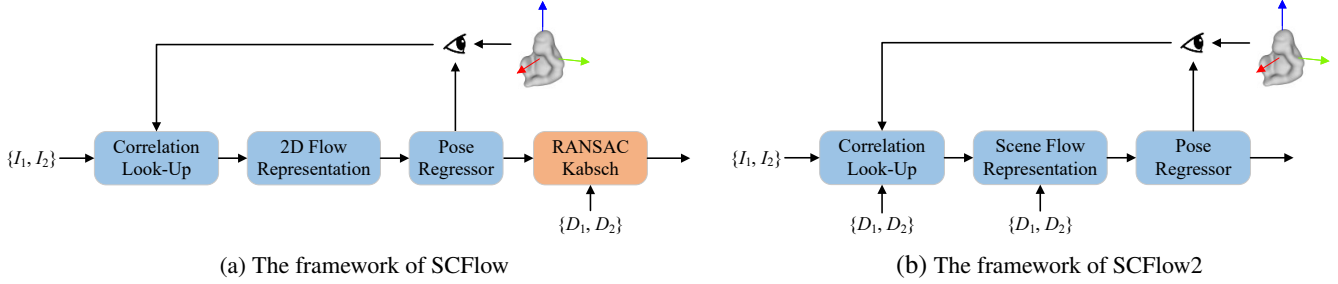


Figure 2. **Design overview of SCFlow and SCFlow2.** Given the object 3D mesh, we render an image I_1 and depth map D_1 based on an initial pose, and then use networks to compare these rendered outputs with the real input I_2 and D_2 to refine the pose. (a) Although SCFlow [11] adds 3D shape constraint into the optimization loop, it formulates the matching process as a pure 2D problem, which is less effective in capturing 3D motions. On the other hand, it cannot work with RGBD images. A common practice is to use RANSAC Kabsch [21] to consume additional depth as a second stage, which however is only local optimal within each stage. (b) SCFlow2 tackles these problems. We introduce an intermediate representation based on 3D scene flow to capture 3D motions in network optimization. Furthermore, we embed depth into the loop by formulating depth as an additional regularization to guide the correlation look-up iteratively, producing an end-to-end trainable system with RGBD images.

Our main design of SCFlow2 is introducing a scene flow based representation into the iterative optimization. Although SCFlow embeds shape constraint into the optimization loop, it formulates the matching process as a pure 2D problem, and suffers in the solution with a second stage of RANSAC Kabsch [21] for RGBD images. On the other hand, SCFlow cannot work with novel objects. To handle these problems, we introduce rigid-motion embeddings in 3D scene flow [47] as our intermediate representation, and formulate depth as an additional regularization to guide the network optimization iteratively, which helps reducing search space and results in an end-to-end system that generalizes to novel objects, as illustrated in Fig. 2.

We train SCFlow2 on a combination of dataset Objaverse [6], GSO [8] and ShapeNet [2], and evaluate on BOP datasets [15] with novel objects. SCFlow2 achieves state-of-the-art accuracy, and can produce accurate results with only one pose hypothesis, outperforming most multiple-hypothesis based refinement methods [23, 32, 36] in both accuracy and efficiency, without any retraining.

2. Related Work

Object pose estimation aims to predict the 3D rotation and 3D translation with respect to the camera, and has shown prominent progress with deep learning techniques [13, 27, 28, 33, 37, 45]. Most existing methods assume the existence of many real images of the target, and train a learning model based on those real images [11, 19, 20, 38, 44], based on either 2D-to-3D matching [16, 26, 53], 3D-to-2D matching [19, 38, 40], or end-to-end training with differentiable perspective-n-point (PnP) [3, 29, 43]. This type of method is accurate on the trained target in general, which however, cannot work with novel objects that were not involved in the training. Retraining or fine-tuning can adapt the net-

work to novel objects, which, however, is cumbersome and limits its applications in real scenarios. Some recent methods show promising results for novel object pose estimation [23, 32, 34, 36, 50], and one of their key components is object pose refinement, which is based on render-and-compare [23, 32, 50] or feature matching [34, 36] strategies. Although these refinement methods work in general, they are less effective at utilizing prior shape information of the target. The recent method SCFlow [11] introduces a shape-constraint recurrent matching framework for object pose estimation, which embeds the target’s 3D shape into the optimization. It produces accurate results for known objects, while has limited generalization ability to novel objects. SCFlow2 is built on top of the basic shape-constraint structure of SCFlow, but extends it by introducing 3D scene flow representation into the network optimization, producing an end-to-end system with RGBD images.

Optical flow estimation is a fundamental computer vision task to estimate pixel-level correspondences between two frames [5, 17, 41, 52]. Most recent methods are based on deep learning models and achieve accurate results in general scenes [5, 41, 46]. The recent method RAFT [46] presents a recurrent framework for flow estimation, and RAFT-3D [47] extends it to estimate 3D scene flow with rigid motion embeddings. Although they show significant advantages over existing flow methods, estimating dense matching with these general flow methods for object pose estimation is less effective, without using the prior 3D shape information of the target. Furthermore, the general flow loss is a surrogate matching loss that does not directly reflect the final 6D object pose [11, 18, 20, 43] in the context of object pose estimation. We combine the shape-constraint formulation [11] and the rigid motion embeddings in 3D scene flow, producing a geometry-guided 6D object pose refiner that is end-to-end trainable and generalizes to novel objects.

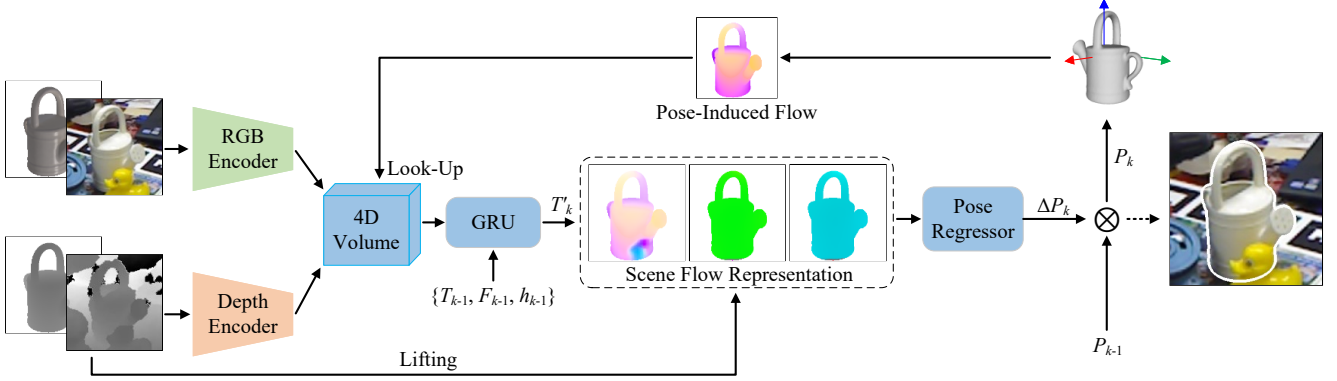


Figure 3. **Overview of SCFlow2.** Given an RGBD image and an initial pose of the target, we first render the target to obtain a synthetic RGBD image as the reference, and use an RGB encoder and depth encoder to extract features from the image pair, which will be used to create a 4D correlation volume. Based on the correlation volume and GRU, We use an intermediate flow regressor to predict 3D scene flow that is represented as a dense 3D transformation field $\mathbf{T}'_k$. We then use a pose regressor to predict a global pose update $\Delta \mathbf{P}_k$ based on implicit voting from the pixel-wise 3D transformation field. Finally, the updated pose $\mathbf{P}_k$ is used to compute the pose-induced flow F_k based on the target mesh to index the correlation volume for the next iteration. Note how the depth and 3D shape of the target are embedded into the framework to guide the optimization iteratively.

3. Method

Given an RGBD image and the 3D mesh of the object, our goal is to estimate the 6D pose of the target object. We assume the camera intrinsic matrix is known. We first introduce the framework at a high level in Section 3.1, and then present the design of the main components in Sections 3.2 and 3.3. We finally discuss the implementation details in Section 3.5.

3.1. Overview

In general, we first render an RGBD image based on an initial pose of the target as a reference, and then predict the pose update based on comparing the rendered image and the real input with a recurrent matching network.

With the RGBD image pair, we first use an RGB encoder and depth encoder to extract features, and then construct a 4D correlation volume based on computing the dot product of the fused RGBD features. Based on the constructed correlation volume, we propose to use an intermediate flow regressor to predict 3D scene flow that is represented as a dense SE3 motion field [47]. We then use a pose regressor to predict a global pose update based on implicit voting from the pixel-wise SE3 motion field. Finally, the updated pose is used to compute a pose-induced flow based on the target mesh to index the correlation volume for the next iteration.

The key of our method is the combination of the dense 3D scene flow representation and the embedding of the target 3D shape prior [11] in a unified recurrent matching framework. Fig. 3 shows the overview of this framework.

3.2. Correlation Volume Construction

We construct a 4D volume pyramid to represent the correlation between any two locations on the two input images. We use an RGB encoder and a depth encoder to extract features from the RGB and the depth image, respectively. For the rendered image I_1 based on the initial pose and the input RGB image I_2 , we use an encoder with shared parameters to obtain their dense feature maps $f_1^i \in \mathbb{R}^{H \times W \times C}$ and $f_2^i \in \mathbb{R}^{H \times W \times C}$ at 1/8 of the original resolution. We use DINOv2 ViT-B [35] as our RGB feature extractor in the encoder, and freeze the pretrained weights during training. For the depths D_1 and D_2 , we first lift the two depth maps to point clouds according to the camera intrinsic matrix, and then extract their 3D structural features with PointNet++ [39].

Similar to DenseFusion [48], we fuse the RGB features and depth features to generate semantic and contextual features. We then compute the 4D correlation volume via computing the dot product of the fused features according to different location pairs.

3.3. Regressing 3D Transformation Field

With a dense matching field as the index, we can extract a 3D correlation map from the 4D correlation volume.

In SCFlow [11], the correlation map is used to predict the optical flow $f(u, v)$, which maps the pixel position $\mathbf{x}_1 = (u, v)$ in I_1 to the corresponding position in I_2 : $\mathbf{x}_2 = (u, v) + f(u, v)$. However, this 2D flow formulation only contains 2D information, which is less effective in capturing 3D motions.

To address this, we propose to predict 3D scene flow [47]

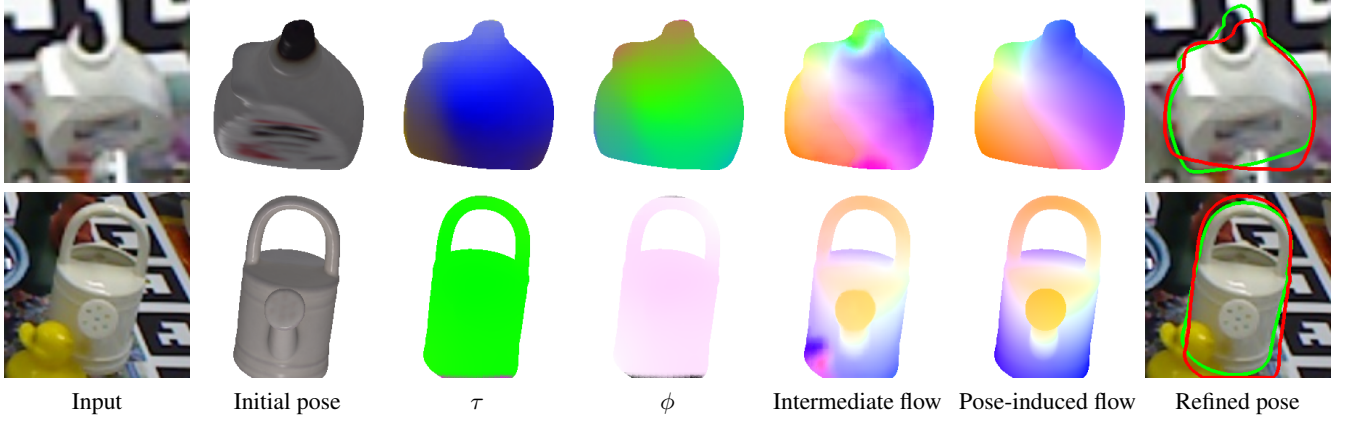


Figure 4. **Visualization of pose-induced flow with 3D scene flow representation.** Given a real input image and the rendered image based on an initial pose, we predict 3D scene flow represented as a dense SE3 motion field. We represent the motion field as a twist field (τ, θ) . In theory, the motion fields should be constant pixels for rigid objects. We use a pose regressor to predict a global object-level pose based on the noisy motion fields. We then use the updated global pose to generate a pose-induced flow based on the target 3D mesh. The pose-induced flow embeds the shape prior of the target and reduces the search space for matching.

as an intermediate representation for the network. Given I_1 and I_2 and their corresponding point clouds C_1 and C_2 , we formulate scene flow as a dense 3D transformation field $\mathbf{T} \in SE(3)^{H \times W}$ to represent the 3D motion between the two point clouds in 3D space.

Specifically, for the point X_i in C_1 with index i and its projected pixel coordinates $\mathbf{x}_i = (u_i, v_i, z_i)$, we can perform the transformation $\mathbf{T}_i \cdot X_i$ to obtain the corresponding point X'_i in C_2 and its projected pixel coordinates $\mathbf{x}'_i = (u'_i, v'_i, z'_i)$. The scene flow vector is defined as the difference $\mathbf{f} = \mathbf{x}'_i - \mathbf{x}_i$, where the first two components represent the standard 2D optical flow and the last component describes the depth difference between the two points.

We use GRU [4] as the recurrent model for the intermediate scene flow predictor. In the k^{th} iteration, the hidden state update can be expressed as:

$$h_k = \text{GRU}(\mathbf{L}_C(F_{k-1}), \mathbf{T}_{k-1}, h_{k-1}; \Theta) \quad (1)$$

where $\mathbf{L}_C(F_{k-1})$ represents the correlation look-up operation guided by the global flow F_{k-1} from the $(k-1)^{th}$ iteration, $\mathbf{T}_{k-1}$ is the dense 3D transformation field from the previous iteration, Θ denotes the GRU network parameters, and h_k is the hidden state feature which will be fed into a dense SE3 layer [47] to update the SE3 motion fields and generate a new transformation field $\mathbf{T}'_k$.

3.4. Regressing Global Object Pose

Our goal is to predict the pose residual of the single object in I_1 and I_2 . Considering the rigidity of objects in 6D pose estimation scenarios, all pixel-level 3D motions in $\mathbf{T}'_k$ theoretically describe the same rigid object and should be the same. However, noise is inevitable in predicting the dense transformation field. Therefore, we designed a small

network to predict the global pose residual using the dense transformation field as input, as shown in Fig. 4. First, we represent the dense transformation field $\mathbf{T}'_k$ as a 4×4 transformation matrix, and then we encode it using a three-layer 2D convolutional network. Finally, we use two fully connected layers to output the global pose residual $\Delta \mathbf{P}_k$, which is represented as a 9-dim vector including 6-min for the rotation [54] and 3-dim for the normalized translation [25].

In the k^{th} iteration, the pose residual $\Delta \mathbf{P}_k$ will be used for the pose update: $\mathbf{P}_k = \mathbf{P}_{k-1} \otimes \Delta \mathbf{P}_k$, and the updated pose $\mathbf{P}_k$ will be used to compute the pose-induced flow based on the object mesh, which imposes shape constraints for the lookup operation in the next iteration [11]. At the same time, we will generate a new dense transformation field $\mathbf{T}_k$ based on the updated pose $\mathbf{P}_k$. Specifically, we calculate the pose residual between the initial pose $\mathbf{P}_0 = [\mathbf{R}_0 | \mathbf{t}_0]$ and $\mathbf{P}_k = [\mathbf{R}_k | \mathbf{t}_k]$ as follows:

$$\Delta \mathbf{R}_k = \mathbf{R}_k \cdot \mathbf{R}_0, \quad \Delta \mathbf{t}_k = \mathbf{t}_k - \Delta \mathbf{R}_k \cdot \mathbf{t}_0 \quad (2)$$

Then, we update the 3D motion field $\mathbf{T}_k \in SE(3)^{H \times W}$, where the motion at every location will be updated as the same new pose residual. This new motion field will be the input for the next iteration.

3.5. Implementation Details

For the input RGB and depth images I_1 and D_2 , we first crop the target out based on the initial pose, and then resize it to a fixed resolution of 256×256 .

Given the ground truth pose, we can compute the ground truth optical flow based on the object mesh. We use the exponentially weighted strategy [11, 46] to calculate the loss

Methods	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V	Avg.
MegaPose [23]	62.0	48.5	84.6	46.2	46.0	72.5	76.4	62.3
MegaPose + Ours	62.4	54.9	88.3	48.4	49.5	76.4	79.0	65.6
FoundPose [36]	61.0	57.0	69.4	47.9	40.7	72.3	69.0	59.6
FoundPose + Ours	69.6	58.6	88.1	54.6	48.4	76.2	84.9	68.6
GenFlow [32]	63.5	52.1	86.2	53.4	55.4	77.9	83.3	67.4
GenFlow + Ours	68.7	53.7	88.0	55.4	56.4	79.2	85.5	69.6
Gigapose [34]	67.8	55.6	81.1	56.3	57.5	79.1	82.5	68.6
Gigapose + Ours	71.0	58.6	82.0	57.7	58.3	80.1	84.4	70.3
SAM6D [27]	69.9	51.5	90.4	58.8	60.2	77.6	84.5	70.4
SAM6D + Ours	72.1	52.3	90.8	60.0	60.5	77.7	85.1	71.2
FoundationPose [50]	75.6	64.6	92.3	50.8	58.0	83.5	88.9	73.4
FoundationPose + Ours	80.1	65.0	95.5	51.7	58.3	84.4	91.4	75.2

Table 1. **Comparison with the state of the art.** All the baselines have their own refinement strategy, and we report their final result published on BOP leaderboard. After using our method as a post-processing (“+ Ours”), the result consistently improves across all baselines on all datasets.

in each iteration for both flow loss and pose loss:

$$\mathcal{L} = \sum_{k=1}^N \gamma^{N-k} (\mathcal{L}_{pose}^k + \alpha \mathcal{L}_{flow}^k) \quad (3)$$

where the weight $\gamma = 0.8$, $N = 8$ and $\alpha = 0.1$. $\mathcal{L}_{flow}$ represents the L1 loss between the first two dimensions of the predicted scene flow and the ground truth optical flow. $\mathcal{L}_{pose}$ refers to pose loss, which is computed as the L1 distance between the 3D point clouds transformed by the predicted pose and the ground truth pose, respectively [11, 49].

4. Experiments

We evaluate SCFlow2 on seven challenging BOP benchmark datasets in this section. Section 4.1 presents our experimental setup, including the datasets used for training and evaluation, training details and the evaluation metrics. In Section 4.2, we compare our approach with current state-of-the-art pose methods. Section 4.3 provides a detailed ablation study of the proposed framework.

4.1. Experimental Settings

Dataset. We train SCFlow2 on a combination of datasets including ShapeNet-Objects [2], Google-Scanned-Objects [8], and Objaverse [6]. All these datasets only contain object meshes, in a total of about 90K objects. We reuse the synthetic images from [23] and [50] which are rendered based on those meshes, with about 3M images in total.

We evaluate on seven datasets from the BOP benchmark [15], including LM-O [1], T-LESS [14], TUD-L [15], IC-BIN [7], ITODD [9], HB [22], and YCB-V [51]. These datasets include 108 objects having different sizes and different texture and symmetry properties. Most of them are

captured in challenging setups, including multiple object instances, significant clutter, and severed occlusions. These objects are not included in the training dataset and are novel to the models.

Training details. We train SCFlow2 using the AdamW optimizer [30] with a batch size of 16, and use a cosine annealing learning rate scheduler. The model is trained for 200k iterations with an initial learning rate of 1e-4. During training, we generate pseudo initial poses by adding random noise to the ground-truth pose. We add rotation jitter sampled from a normal distribution with a standard deviation of 15 degrees in each 3D axis, and add translation jitter sampled with deviations of 15, 15, and 50 millimeters along xyz axes, respectively.

Evaluation metrics We report the result in the common BOP metric, which is the mean Average Recall (AR) of three different metrics, including Visible Surface Discrepancy (VSD), Maximum Symmetry-aware Surface Distance (MSSD), and Maximum Symmetry-aware Projection Distance (MSPD). Readers can refer to [15] for further details about these metrics.

4.2. Comparison with the State of the Art

To explore the plug-and-play nature of our refinement method, we apply it to the final results of MegaPose [23], FoundPose [36], GenFlow [32], GigaPose [34], SAM6D [27], and FoundationPose [50], as shown in Table 1. Although these baselines have a refinement strategy themselves, plugging our refinement method into the end of their framework further improves the results, consistently on all datasets. Most baseline results are downloaded from the BOP leaderboards.

Methods	Depth	LM-O	YCB-V
FoundPose (init)	-	39.6	45.2
FoundPose (final)	✓	61.0	69.0
FoundPose (init) + Ours	✓	<u>61.3</u>	<u>81.8</u>
FoundPose (final) + Ours	✓	69.6	84.9
GenFlow (init)	-	25.0	27.7
GenFlow (final)	✓	63.5	83.3
GenFlow (init) + Ours	✓	<u>64.4</u>	<u>83.6</u>
GenFlow (final) + Ours	✓	68.7	85.5
FoundationPose (init)	-	-	-
FoundationPose (final)	✓	75.6	88.9
FoundationPose (init) + Ours	✓	<u>79.7</u>	<u>90.4</u>
FoundationPose (final) + Ours	✓	80.1	91.4

Table 2. **Effectiveness as a plug-and-play object pose refiner.** All the baselines are based on an initialization-refinement architecture. Our method consistently outperforms their original refinement strategies, and can also improve their final results significantly, in a plug-and-play manner.

	L3	L5	L10	L20	L30	L40	L50
Initialization	92.7	83.8	64.5	35	17.6	7.8	3.5
+ SCFlow2	91.6	91.5	91.5	90.8	88.6	83.2	70.7

Table 3. **Impact of pose initialization.** We add randomly generated rotation and translation noise to GT pose as the initialization, and then use SCFlow2 to refine the pose. L3 to L50 denote different noise levels. SCFlow2 is robust to pose errors in the initialization and works well under noise level 30.

4.3. Ablation Study

Effectiveness as a plug-and-play object pose refiner. We evaluate SCFlow2 in more detail in the context of three different state-of-the-art methods, FoundPose [36], GenFlow [32], and FoundationPose [50]. Table 2 summarizes the results. All of them are based on an initialization-refinement architecture. We denote their initialization results as “init” in the table, and replace their refinement methods with ours in each method (denoted as “init + ours”). Furthermore, we also report the results of applying our refinement to their final results directly, denoted as “final + ours”. Note that, FoundationPose outputs hundreds of pose hypotheses in the initialization stage, and there is no well-defined initialization result for it. In this case, we replace its original refinement method with SCFlow2 in its framework, and keep all other components untouched, including post-processing procedures such as pose scoring and pose selection, which is the reported result of “FoundationPose (init) + Ours” in the table. Overall, our method consistently outperforms their original refinement strategies, and can also improve their final results significantly, in a plug-and-play manner. We show some qualitative results in Fig. 5 and 6.

Datasets	Init.	Seen		Unseen	
		V1	V1+	V1++	V2
LM-O	61.0	62.9	<u>69.3</u>	68.9	69.6
T-LESS	57.0	<u>57.9</u>	54.9	55.3	58.6
TUD-L	69.4	75.9	88.8	86.6	<u>88.1</u>
IC-BIN	47.9	47.9	<u>54.1</u>	53.7	54.6
ITODD	40.7	34.2	49.1	48.1	<u>48.4</u>
HB	72.3	74.5	76.5	76.1	<u>76.2</u>
YCB-V	69.0	80.7	85.5	84.4	<u>84.9</u>
Avg.	59.6	62.0	<u>68.3</u>	67.6	68.6

Table 4. **Advantages over SCFlow.** We use the same initialization based on results from FoundPose (“Init.”), and report results of SCFlow (“V1”) and SCFlow2 (“V2”). SCFlow2 outperforms all “V1” versions without any retraining.

Impact of pose initialization. We add random rotation and translation noise to GT pose as the initialization, and report the results on the YCB-V dataset, as shown in Table 3. L3 to L50 denote different noise levels (rotation error from 3 to 50 degrees, translation error from 3 to 50 mm). SCFlow2 works well under noise level 30.

Evaluation of performance with different iterations. We conduct ablation studies on the trade off between the accuracy/efficiency and the iteration numbers of our method, as shown in Fig. 7. Our method produces meaningful pose refinement results after only 2 iterations and reaches saturation after 6 iterations. In our experiments, we use 8 iterations as the default setting in both training and testing.

Advantages over SCFlow. We compare SCFlow2 and its previous version SCFlow systematically. We summarize the results in Table 4. We use the same initialization based on results from FoundPose, and refine it with SCFlow (“V1”) and SCFlow2 (“V2”), respectively. We evaluate three different versions of SCFlow, where “V1” denotes the original version based on retraining on target objects with RGB images only, “V1+” denotes the version using RANSAC Kabsch to consume additional depth images. Both “V1” and “V1+” need retraining on target objects. We also evaluate a version “V1++” which does not rely on retraining for novel objects, but uses the same large-scale datasets as “V2” for training on RGB images and then refine the pose results using RANSAC Kabsch with additional depths. We can see that, without retraining, “V1++” produces worse results compared with “V1+”. On the other hand, SCFlow2 even outperforms “V1+” without any retraining.

Evaluation with different settings. We use the initialization results of FoundPose as our initialization. Table 5 shows the result, where “Shape” denotes the shape-constraint regularization during the recurrent optimization, and “Scene” is the 3D scene flow representation. We use the intermediate flow for the look-up without relying on the

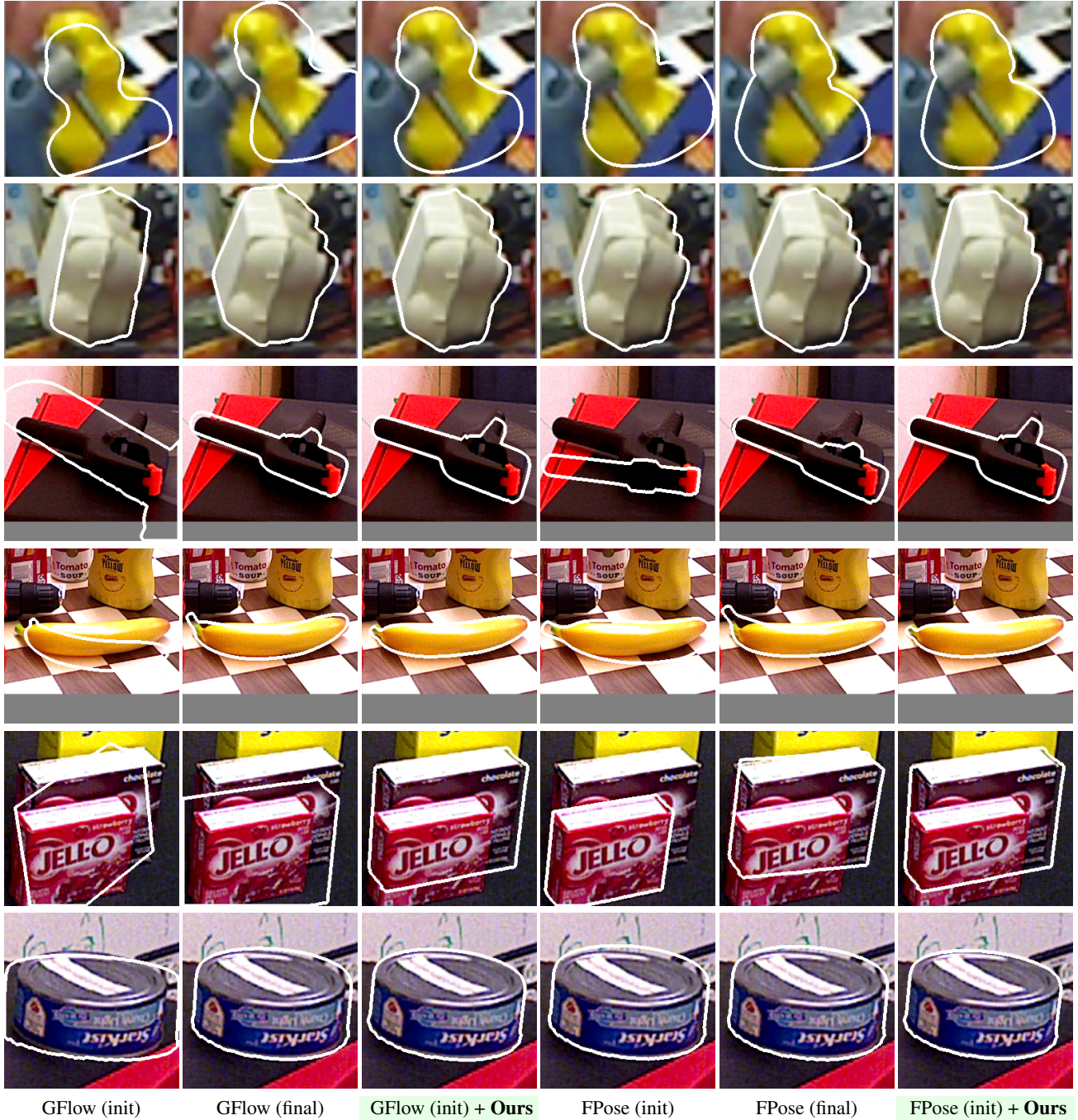


Figure 5. **Qualitative results.** Given the same pose initialization as that in GenFlow (“GFlow”) and FoundPose (“FPose”), denoted as “GFlow (init)” and “FPose (init)” respectively, our refinement method (“+ Ours”) produces considerably more accurate results compared to the refinement approaches in their original methods (note how our reprojected 3D mesh aligns better with the object contours).

target’s 3D mesh when “Shape” is disabled, and we replace the scene flow representation with a plain MLP pose regressor when “Scene” is disabled. The performance becomes worse without the shape prior embedding, and deteriorates significantly without the 3D scene flow formulation.

Timing analysis. The training of SCFlow2 needs about 108 hours on a workstation with 2 NVIDIA RTX-3090 GPU. SCFlow2 considers the pose initialization as the only hypothesis. During inference, SCFlow2 takes about 0.18 seconds to refine a pose on an RTX-3090 GPU with a typical



Initialization + Ours

Figure 6. **More qualitative results.** We show two examples on YCB-V. Although the initialization has large pose error, our refinement method produces accurate results that align well with the target, without retraining on these novel objects.

Shape	Scene	LM-O	YCB-V
✓	-	52.6	67.6
-	✓	<u>60.0</u>	<u>80.4</u>
✓	✓	61.3	81.8

Table 5. **Evaluation with different settings.** “Shape” denotes the shape-constraint regularization during the optimization, and “Scene” is the 3D scene flow formulation. The performance deteriorates significantly without the 3D scene flow formulation.

image from BOP dataset. Taken for comparison, the multi-hypothesis based method MegaPose needs about 1.85 seconds for a refinement on the same workstation, which is 10+ times slower than SCFlow2.

Limitation discussion. In general, our refinement method can improve the accuracy of initial pose significantly. While occasionally, it produces worse results than the initial pose. Fig. 8 shows the histogram of samples along with the relative accuracy improvement after pose refinement on YCB-V and LM-O with the initial pose of FoundPose as our initialization. In the majority of cases, our method improves the initial poses significantly. However, it still happens that the refinement makes the results worse in some extreme cases where the initial poses are already very wrong caused by severe occlusion or extreme light conditions. We will take this as one of our future work.

5. Conclusion

We have introduced SCFlow2, a plug-and-play refinement framework for 6D object pose estimation. We extended

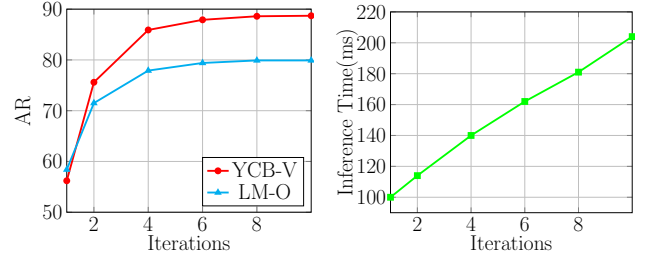


Figure 7. **Evaluation of performance with different iterations.** Given the same pose initialization with noise level 30 as in Table 3, more recurrent iterations produce more accurate results while spending more time in inference. Nevertheless, our approach yields meaningful outcomes after just 2 iterations, and the performance reaches saturation after 6 iterations.

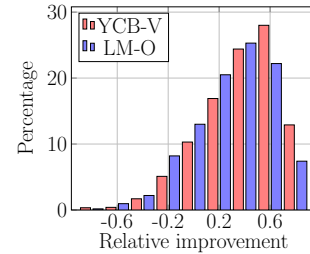


Figure 8. **Limitation discussion.** We show the histogram of samples along with the relative accuracy improvement after pose refinement. In the majority of cases, our method improves the initial poses significantly. However, it still happens that the refinement makes the results worse in some extreme cases.

the RGB-based method SCFlow by formulating additional depth as a regularization in the iteration with an intermediate 3D scene flow representation for RGBD frames, and combined it with the 3D shape prior information of the target, producing a geometry-guided 6D object pose refiner end-to-end trainable and generalizable to novel objects. Experiments on BOP datasets with novel objects show that our method achieves state-of-the-art accuracy, and is one of the most efficient refinement methods.

Acknowledgments. This work was supported in part by the Youth Innovation Team of Shaanxi Universities, the “Scientist + Engineer” team of the Qin Chuang Yuan in Shaanxi Province, the Xi’an Science and Technology Program, the “Leading the Charge” initiative for the industrialization of core technologies in key industrial chains in Shaanxi Province, the National Nature Science Foundation of China under Grant 62371359, the Key Research and Development Program of Shaanxi, the Key Research Program of the Chinese Academy of Sciences.

References

- [1] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6D Object Pose Estimation Using 3D Object Coordinates. In *European Conference on Computer Vision*, 2014. 5
- [2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An Information-Rich 3D Model Repository. *arXiv preprint arXiv:1512.03012*, 2015. 2, 5
- [3] Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, and Tat-Jun Chin. End-to-End Learnable Geometric Vision by Backpropagating PnP Optimization. In *Conference on Computer Vision and Pattern Recognition*, 2020. 2
- [4] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv preprint arXiv:1409.1259*, 2014. 4
- [5] Javier Cuadrado, Ulysse Rançon, Benoit R Cottreau, Francisco Barranco, and Timothée Masquelier. Optical Flow Estimation from Event-based Cameras and Spiking Neural Networks. *Frontiers in Neuroscience*, 17, 2023. 2
- [6] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A Universe of Annotated 3D Objects. In *Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5
- [7] Andreas Dumanoglou, Rigas Kouskouridas, Sotiris Malasiotis, and Tae-Kyun Kim. Recovering 6D Object Pose and Predicting Next-Best-View in the Crowd. In *Conference on Computer Vision and Pattern Recognition*, 2016. 5
- [8] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google Scanned Objects: A High-Quality Dataset of 3D Scanned Household Items. In *International Conference on Robotics and Automation*, 2022. 2, 5
- [9] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. Introducing MVTEC ITODD-A Dataset for 3D Object Recognition in Industry. In *International Conference on Computer Vision Workshops*, 2017. 5
- [10] Yang Hai, Rui Song, Jiaojiao Li, David Ferstl, and Yinlin Hu. Pseudo Flow Consistency for Self-Supervised 6D Object Pose Estimation. In *International Conference on Computer Vision*, 2023. 1
- [11] Yang Hai, Rui Song, Jiaojiao Li, and Yinlin Hu. Shape-Constraint Flow for 6D Object Pose Estimation. In *Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 4, 5
- [12] Rasmus Laurvig Haugaard and Anders Glent Buch. SurfEmb: Dense and Continuous Correspondence Distributions for Object Pose Estimation with Learnt Surface Embeddings. In *Conference on Computer Vision and Pattern Recognition*, 2022. 1
- [13] Yisheng He, Yao Wang, Haoqiang Fan, Jian Sun, and Qifeng Chen. FS6D: Few-Shot 6D Pose Estimation of Novel Objects. In *Conference on Computer Vision and Pattern Recognition*, 2022. 2
- [14] Tomáš Hodaň, Pavel Haluza, Štěpán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-Less Objects. In *Winter Conference on Applications of Computer Vision*, 2017. 5
- [15] Tomáš Hodaň, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. BOP: Benchmark for 6D Object Pose Estimation. In *European Conference on Computer Vision*, 2018. 2, 5
- [16] Tomáš Hodaň, Dániel Baráth, and Jiří Matas. EPOS: Estimating 6D Pose of Objects with Symmetries. In *Conference on Computer Vision and Pattern Recognition*, 2020. 2
- [17] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2
- [18] Yinlin Hu, Pascal Fua, Wei Wang, and Mathieu Salzmann. Single-Stage 6D Object Pose Estimation. In *Conference on Computer Vision and Pattern Recognition*, 2020. 2
- [19] Yinlin Hu, Sebastien Speierer, Wenzel Jakob, Pascal Fua, and Mathieu Salzmann. Wide-Depth-Range 6D Object Pose Estimation in Space. In *Conference on Computer Vision and Pattern Recognition*, 2021. 2
- [20] Yinlin Hu, Pascal Fua, and Mathieu Salzmann. Perspective Flow Aggregation for Data-Limited 6D Object Pose Estimation. In *European Conference on Computer Vision*, 2022. 1, 2
- [21] Wolfgang Kabsch. A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976. 1, 2
- [22] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Homebreweddb: RGB-D Dataset for 6D Pose Estimation of 3D Objects. In *International Conference on Computer Vision Workshops*, 2019. 5
- [23] 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. In *Conference on Robot Learning*. 1, 2, 5
- [24] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. CosyPose: Consistent Multi-View Multi-Object 6d Pose Estimation. In *European Conference on Computer Vision*, 2020. 1
- [25] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIM: Deep Iterative Matching for 6D Pose Estimation. In *European Conference on Computer Vision*, 2018. 4
- [26] Zhigang Li, Gu Wang, and Xiangyang Ji. CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation. In *International Conference on Computer Vision*, 2019. 2
- [27] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. SAM-6D: Segment Anything Model Meets Zero-Shot 6D Object Pose Estimation. In *Conference on Computer Vision and Pattern Recognition*, 2024. 2, 5

- [28] Yongliang Lin, Yongzhi Su, Praveen Nathan, Sandeep Inuganti, Yan Di, Martin Sundermeyer, Fabian Manhardt, Didier Stricker, Jason Rambach, and Yu Zhang. HiPose: Hierarchical Binary Surface Encoding and Correspondence Pruning for RGB-D 6DoF Object Pose Estimation. In *Conference on Computer Vision and Pattern Recognition*, 2024. 2
- [29] Fulin Liu, Yinlin Hu, and Mathieu Salzmann. Linear-Covariance Loss for End-to-End Learning of 6D Pose Estimation. In *International Conference on Computer Vision*, 2023. 2
- [30] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. 5
- [31] 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
- [32] Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. GenFlow: Generalizable Recurrent Flow for 6D Pose Refinement of Novel Objects. In *Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2, 5, 6
- [33] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatin, Yinlin Hu, Renaud Marlet, Mathieu Salzmann, and Vincent Lepetit. NOPE: Novel Object Pose Estimation from a Single Image. In *Conference on Computer Vision and Pattern Recognition*, 2024. 2
- [34] Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. GigaPose: Fast and Robust Novel Object Pose Estimation via One Correspondence. In *Conference on Computer Vision and Pattern Recognition*, 2024. 2, 5
- [35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, 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
- [36] Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomáš Hodaň. FoundPose: Unseen Object Pose Estimation with Foundation Features. *European Conference on Computer Vision*, 2024. 1, 2, 5, 6
- [37] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. LatentFusion: End-to-End Differentiable Reconstruction and Rendering for Unseen Object Pose Estimation. In *Conference on Computer Vision and Pattern Recognition*, 2020. 2
- [38] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. PVNet: Pixel-Wise Voting Network for 6DoF Pose Estimation. In *Conference on Computer Vision and Pattern Recognition*, 2019. 2
- [39] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *arXiv preprint arXiv:1706.02413*, 2017. 3
- [40] Mahdi Rad and Vincent Lepetit. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth. In *International Conference on Computer Vision*, 2017. 2
- [41] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The Surprising Effectiveness of Diffusion Models for Optical Flow and Monocular Depth Estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [42] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. OSOP: A Multi-Stage One Shot Object Pose Estimation Framework. In *Conference on Computer Vision and Pattern Recognition*, 2022. 1
- [43] Chen Song, Jiaru Song, and Qixing Huang. HybridPose: 6D Object Pose Estimation under Hybrid Representations. In *Conference on Computer Vision and Pattern Recognition*, 2020. 2
- [44] Yongzhi Su, Mahdi Saleh, Torben Fetzner, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. ZebraPose: Coarse to Fine Surface Encoding for 6DoF Object Pose Estimation. *arXiv preprint arXiv:2203.09418*, 2022. 2
- [45] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. OnePose: One-Shot Object Pose Estimation without CAD Models. In *Conference on Computer Vision and Pattern Recognition*, 2022. 2
- [46] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *European Conference on Computer Vision*, 2020. 2, 4
- [47] Zachary Teed and Jia Deng. RAFT-3D: Scene Flow using Rigid-Motion Embeddings. In *Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3, 4
- [48] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. In *Conference on Computer Vision and Pattern Recognition*, 2019. 1, 3
- [49] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation. In *Conference on Computer Vision and Pattern Recognition*, 2021. 5
- [50] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. In *Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2, 5, 6
- [51] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In *Robotics: Science and Systems*, 2018. 5
- [52] Zhaolin Xiao, Zhan Su, Fengyuan Zuo, and Haiyan Jin. Low-Light Optical Flow Estimation with Hidden Feature Supervision Using a Siamese Network. *Journal of Image and Graphics*, 29:0231–0242, 2024. 2
- [53] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. DPOD: 6D Pose Object Detector and Refiner. In *International Conference on Computer Vision*, 2019. 2
- [54] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Conference on Computer Vision and Pattern Recognition*, 2019. 4