SFRF6D: Selective Fewer-Reference Fusion for 6D Pose Estimation

Qinghui Zhang, Chi Zhou, Wei Pan, and Lei Lu

Abstract

Currently, most existing 6D pose estimation methods require object Computer-Aided Design (CAD) models or a large number of reference images. To address this, we propose SFRF6D, a model-free framework that achieves accurate pose estimation using only a few reference images. Our method renders a small set of reference images to generate multi-view information and employs geometry-guided attention with multi-scale feature fusion to improve robustness under occlusion.In addition, inspired by structure-from-motion (SfM) principles, our framework implicitly enforces geometric consistency across views without relying on CAD models. Furthermore, we conduct extensive training and evaluation on the LINEMOD (LM), LINEMOD Occlusion (LMO), and YCB-Video (YCB) datasets, where SFRF6D consistently delivers superior performance than existing few-reference and model-free approaches such as LatentFusion and Any6D by 8-13% in Average Distance of Model Points (Symmetric) (ADD(-S)) accuracy, particularly on the heavily LMO benchmark. These results demonstrate the effectiveness of our method in handling complex real-world scenarios with limited reference images and without CAD model dependency.

1 Introduction

6D pose estimation, which aims to recover an object's 3D translation and 3D rotation, is a long-standing and fundamental problem in computer vision. With the growing integration of automation into real-world applications, accurate pose information has become critical for precision tasks such as robotic grasping and assembly [1].

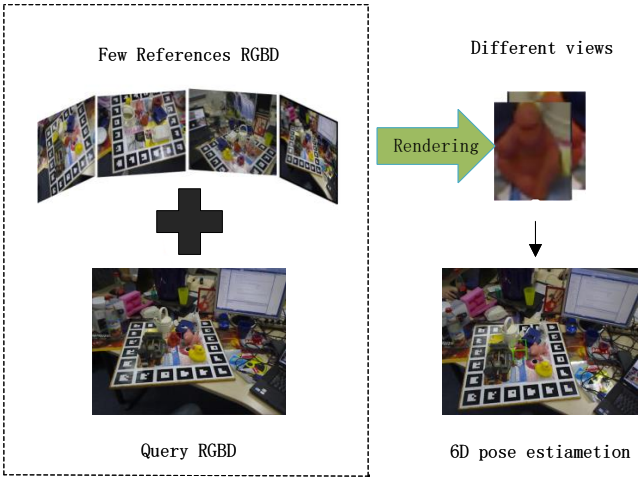

Figure 1: Our SFRF6D is a novel 6D pose estimation method that requires only a small number of RGB-D reference images with ground-truth annotations. Through geometry-guided cross-view fusion, it establishes dense image-point cloud correspondences by rendering and matching the reference point clouds, enabling high accuracy with minimal prior information.

The rapid development of deep learning has replaced traditional hand-crafted feature engineering and further propelled 6D pose estimation, enabling its wide adoption in areas such as augmented reality and autonomous driving [2]. Many existing approaches have achieved high accuracy and efficiency even under challenging conditions such as shadows and occlusions [3], yet they often struggle to balance robustness in complex environments with their dependence on Computer-Aided Design (CAD) models or large collections of reference images. Addressing these limitations remains a major challenge in achieving generalizable and efficient pose estimation.

Currently, mainstream approaches to 6D pose estimation can be broadly divided into two categories: two-stage methods and end-to-end methods. Two-stage pipelines typically begin by segmenting the target re-

gion via a mask to reduce the feature extraction search space and computational cost; they then match 2D image features with 3D object features to establish correspondences. Some methods first predict a coarse pose and refine it through pose-guided rendering to improve correspondence quality, followed by solving for the final 6D pose using Perspective-n-Point (PnP)* and Random Sample Consensus (RANSAC) [4]. This paradigm achieves high accuracy, especially when combining multiple reference views and depth information to crop the point cloud, along with consistency checks [5], symmetry constraints [6], and visibility prediction to further enhance performance. However, these methods accumulate errors from each stage of the pipeline, and compared to end-to-end approaches that optimize a single differentiable objective, they often require substantial manual tuning of stage-specific components - for example, mask thresholds and post-processing heuristics for segmentation, feature matching thresholds, the number of RANSAC iterations and inlier thresholds, depth-based cropping parameters, and settings for subsequent ICP/pose-refinement steps. These hand-designed choices and non-differentiable modules increase sensitivity to dataset characteristics and imaging conditions (illumination, occlusion, sensor noise), which is why significant fine-tuning and optimization are typically needed to reach peak performance. Instead, they rely on surrogate losses based on intermediate representations, which limits their integration with self-supervised learning [7]. Consequently, two-stage approaches are often computationally complex and resource-intensive.

Based on this, we propose SFRF6D, a novel model-free framework for 6D pose estimation that requires only a small number of reference images. The overall workflow is illustrated in Figure 1. Unlike conventional model-free methods, SFRF6D leverages a few RGB-D reference images with known ground-truth poses to render reference-view point clouds from the coarse pose of a query image, constructs a multi-modal, multi-scale feature fusion network, and learns cross-view geometric representations through a self-attention mechanism. Specifically, visibility and directional vectors derived from rendered point clouds guide a set-attention mask, enabling cross-view occlusion awareness and improving pose reliability in highly occluded scenes based on geometric consistency.We conduct experiments on the BOP benchmark datasets, including LineMOD, Occlusion LineMOD, and YCB-V, and the results show that our method performs strongly across diverse environments, particularly achieving high accuracy and robustness in heavily occluded settings such as Occlusion LineMOD [11]. Therefore, SFRF6D effectively enhances performance in complex real-world scenarios, offering a precise and efficient solution for model-free 6D pose estimation using limited reference images.

In summary, our key contributions are as follows:

- We propose SFRF6D, a model-free 6D pose estimation framework that achieves accurate results using only a few RGB-D reference images, reducing dependence on large-scale reference sets and improving real-world deployability.
- A cross-view geometry-guided feature fusion module is introduced, which renders reference point clouds from the query view to establish dense image-point cloud correspondences, enhancing robustness under large viewpoint changes and occlusions.
- During cross-view feature alignment, a multi-scale feature fusion module is introduced to adaptively select the importance of global and local features. Combined with occlusion-aware attention, weights are dynamically adjusted based on visibility and directional vectors, enabling the model to focus on key geometric features of the visible regions.

2 Related Works

2.1 Model-Driven Approaches

Early instance-level pose estimation methods heavily relied on CAD models, solving poses by establishing dense 2D-3D point correspondences and applying PnP. Some studies attempted end-to-end regression, such as Kehl et al., who extended SSD for pose classification [12], or approaches like Manhardt's, which improve robustness in complex environments through multi-hypothesis poses, optimizing the pose via object projection alignment or point-pair matching losses [12]. These methods explore end-to-end learning within the PnP paradigm [13], with recent trends favoring the implicit establishment of dense 2D-3D correspondences to further enhance generalization [14]. Although these approaches achieve higher precision, they still heavily depend on CAD models, limiting their applicability in real-world scenarios. In contrast, our proposed SFRF6D completely eliminates CAD dependency, requiring only a few reference images to render multi-view representations, achieving performance comparable to CAD-dependent methods.

2.2 Model-Free Approaches Based on Extensive References

To mitigate the strong reliance of 6D pose estimation on CAD models, a line of research leverages large collections of reference images or video streams to build stable multi-view representations and perform pose estimation through implicit modeling. Representative works include Gen6D [15], OnePose [16], and OnePose++ [17], which reconstruct point clouds via structured light or SfM and achieve pose estimation through 2D-3D matching. More recent approaches, such as FoundationPose

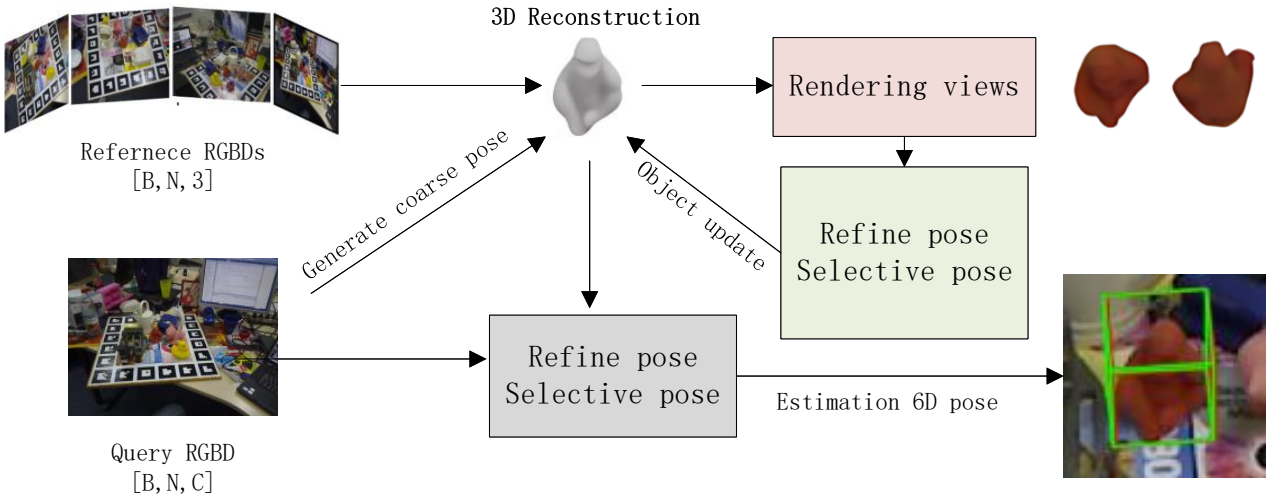

Figure 2: This illustrates the SFRF6D pipeline for addressing the performance degradation in cross-view matching under large viewpoint changes and occlusions. It employs a geometry-guided sparse attention mechanism that explicitly constrains the selection of informative features, reducing the impact of low-matching ambiguity on representative regions. To overcome the non-differentiability of hard attention based on convex optimization while preserving the discreteness of selection, our approach introduces an explicit geometric constraint.

[18] and FS6D [19], incorporate depth information to further enhance accuracy and robustness. However, these methods generally require a substantial number of multi-view reference images as priors, and cross-view representation alignment suffers from inherent limitations-stable and consistent representations become difficult to maintain under severe occlusions or large viewpoint changes, which restricts their practical applicability. In contrast, our method requires only a small set of reference images: we first mitigate inter-view discrepancies through coarse pose estimation and point cloud transformation rendering, and then employ sparse geometry guidance and selective attention to emphasize visible object regions while down-weighting occluded, hard-to-match areas, ultimately achieving robust and stable cross-view feature alignment.

2.3 Model-Free Approaches with Limited Reference

In recent years, there has been notable progress in reducing the reliance on reference data in model-free pose estimation. For instance, LoFTR [20] employs a Transformer attention mechanism to leverage textures from other regions of the reference image, constructing more robust correspondences in weakly textured areas and sparse point clouds. LoFTR enhances features through the Transformer to establish stable matching. Another innovative approach, Oryon [21], uses a language-guided mechanism where user-provided textual prompts specify object regions of interest, combining image and text features to generate masks, thereby expanding applicability in practical scenarios. However, these solutions experience a significant drop in feature matching capability when reference views barely overlap. To address this, GigaPose [22] introduces RGB-based 3D reconstruction to provide geometric priors, but the reconstructed 3D representations are coarse, containing only rough object contours, which limits high-precision pose estimation.

Compared with traditional Structure-from-Motion (SfM) methods that reconstruct dense point clouds or meshes from multiple RGB images, these model-free frameworks do not rely on global multi-view consistency or bundle adjustment. Instead, they perform implicit 3D reasoning by aligning sparse reference views and query observations. While SfM can recover complete geometry, it typically requires dense view coverage and heavy computation, making it unsuitable for few-view scenarios.

SFRF6D further addresses these limitations. Without requiring additional modeling or prior information, it achieves high-precision pose estimation using only a few reference images with known poses. Although SFRF6D does not perform dense 3D reconstruction like SfM, it explicitly reconstructs the object's coarse 3D shape from multi-view reference RGB-D inputs. The reconstructed shape is then rendered under coarse poses to form geometry-aware multi-view representations, providing a reliable geometric basis for cross-view alignment. Subsequently, a geometry-guided visibility-selective attention mechanism emphasizes visible regions while suppressing occluded or non-overlapping areas, achieving

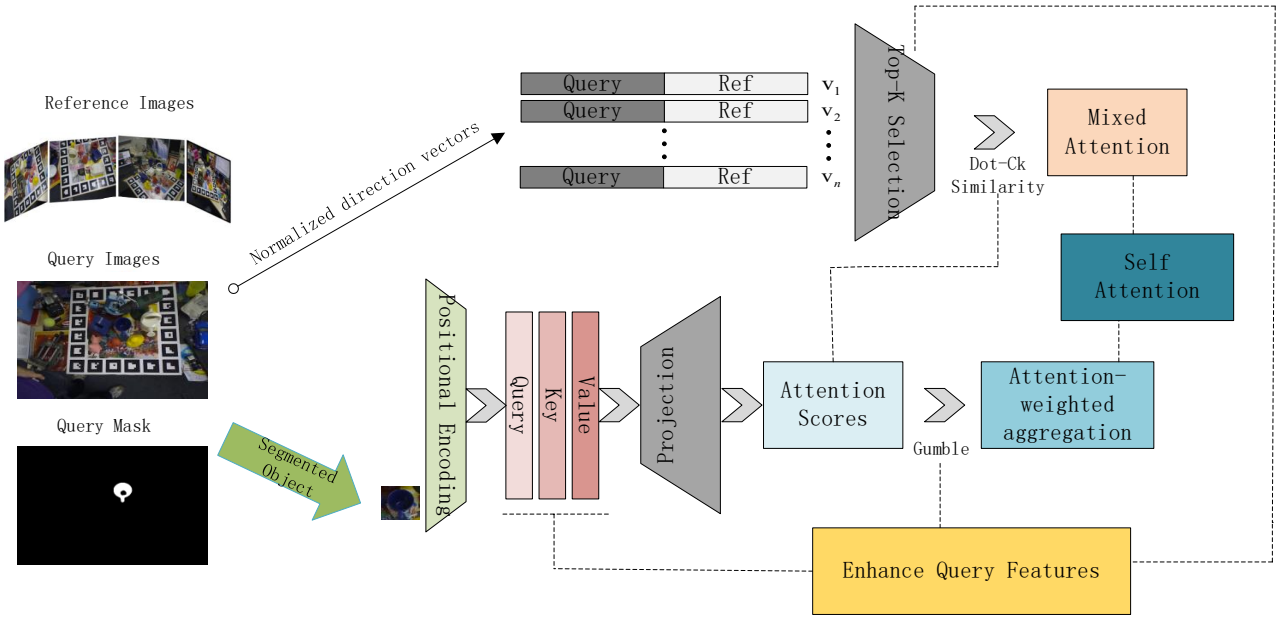

Figure 3: In SFRF6D, sparse geometry-guided attention drives reference rendering, explicitly constraining cross-view correspondences. By leveraging normalized direction vectors and a top-K selection mechanism based on similarity computed after positional encoding, the model's efficiency and lightweight capability are enhanced.

more stable cross-view feature matching. In contrast to SfM-based pipelines, SFRF6D leverages this lightweight 3D shape representation rather than full mesh reconstruction, effectively maintaining robustness and computational efficiency. This strategy alleviates precision degradation caused by occlusion and large viewpoint variations, while maintaining low computational overhead.

3 Method

3.1 Geometry-Guided Sparse Attention Mechanism

As shown in figure 3. We take as input the 2D RGBD features and 3D point cloud representations extracted from a small set of reference images and the query image. The reference images are first processed by a ResNet backbone [23], yielding features represented as [B,M,C],where B denotes the batch size,M the number of point cloud samples, and C the feature dimension; the same representation applies to the query features. The corresponding spatial information for query and reference points, [B,M,3] and [B,N,3], encodes the 3D coordinates of each point in the point cloud, allowing parallel multi-view processing. To establish correspondences between query and reference points, we compute the directional vectors [B,N,M,3] according to the following formulation:

$$dir = ref_{xyz}[:,:,None,:] - query_{xyz}[:,None,:,:] \quad (1)$$

After permutation, the representation becomes [B,M, N,3],which encodes the pairwise correspondence of 3D coordinates between query and reference point clouds. A normalization step is applied to mitigate scale discrepancies and strengthen the attention mechanism's ability to match identical regions, thereby explicitly providing geometric guidance for subsequent attention computation. The position-encoded QKV (query-key-value) mappings [38, 39] are then constructed to facilitate feature aggregation across multiple views and query points. For regions with direct correspondences, the raw 3D coordinates are used, while regions affected by viewpoint changes or occlusion are handled using relative positional encoding as an approximation. The QKV representations are obtained through linear projections, formally defined as follows:

$$Q \in R^{[B,M,C_q]} \quad (2)$$

$$K/V \in R^{[B,N,C_{kv}]} \quad (3)$$

It simultaneously carries both geometric and object-level semantic information, providing richer contextual cues. By retaining the normalized direction vectors, the mechanism strengthens attention toward critical regions and enhances the representational capacity of the subsequent dot-product attention [27]. This design significantly improves the model's geometric awareness, while the sparse attention constraint further boosts robustness under large viewpoint changes and heavy occlusion

[28]. The core idea lies in using the normalized direction vectors to construct a Top-K mask that guides cross-attention. In this module, the Dot-Ck (Dot-Cross) interaction is employed to compute the similarity between query and reference features within the normalized direction vectors. Specifically, the query feature vectors (after linear projection) are multiplied with the transposed key feature vectors to compute the attention logits, producing a tensor of shape [B,M,N]. Regions where matching fails are assigned invalid values, thereby sparsifying the attention map. A softmax operation is then applied to normalize the attention logits, yielding attention scores that represent the correspondence strength between each query and its keys. Finally, the value features are weighted by these attention scores to obtain the enhanced query features:

$$Q' = Softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)v \tag{4}$$

where d_k is the dimension of the key features used for scaling. This operation integrates geometric priors into the attention mechanism, focusing on visible and relevant regions while suppressing mismatched areas. It integrates contextual information from relevant regions into the query, further enhancing its feature representation capacity [29].

In the Top-K selection process, we employ Gumbel Softmax for probabilistic soft sampling, introducing controlled perturbations to the attention logits and using a temperature parameter to improve smoothness. This allows approximate selection of corresponding point sets under challenging viewpoints. Consequently, it effectively enforces explicit constraints on directly matched points while providing implicit constraints on large viewpoint variations, accelerating the network's focus on valid correspondences and its learning ability. Leveraging the self-attention mechanism, the method further establishes internal associations, selecting corresponding regions across multiple views to aggregate cross-view contextual information [38,39]. Through this shared structure, a unified model is constructed to provide robust and stable feature representations for subsequent modules.

3.2 3D Object Reconstruction and Coarse Alignment

To overcome the limitations on pose estimation accuracy caused by the absence of CAD models, our framework reconstructs the 3D scale of the object from a small number of reference images and performs alignment with the query image through projection [30,31,42]. A reference image whose viewpoint is closest to that of the query is selected as the anchor, while the remaining reference views are utilized to supplement geometric information for reconstructing the 3D shape of the query object. Specifically, a convolutional neural network (ResNet18) is employed to extract 2D semantic features of the object from the anchor image [23]. Leveraging the geometric information of the anchor depth map, an initial 3D point cloud of the object is generated [28]. The transformation from the depth map to the object point cloud is defined as follows:

$$P_A = \{(x_i, y_i, z_i)\}_{i=1}^{N} \tag{5}$$

P_A denotes the point cloud data extracted from IA, while x_i, y_i, z_i represents the 3D coordinates of the points. The information from the remaining viewpoints is used as a supplement to further enhance robustness, with the process defined as follows:

$$p_i^{world} = R_i \cdot p_i + t_i \tag{6}$$

Specifically, the point cloud is transformed into the world coordinate system using the camera intrinsics and ground-truth pose information from the reference images, after which all relevant point clouds are uniformly processed [20]. Here, p_i^{world} denotes the point cloud in world coordinates. This establishes a mapping between the query image view and the 3D point cloud in order to reconstruct the object's 3D shape O_N Normalization is then applied to scale the object into the range [-1,1], which facilitates subsequent pose alignment and viewpoint rendering tasks. Next, the query image viewpoint is compared with O_N to obtain an initial estimation of the object's scale and pose. In particular, we employ an oriented bounding box (OBB) to compute the 3D center and rotational components of the object [26,28]. The formula for calculating the object center is given as follows:

$$c_A = \frac{1}{N}\sum_{i=1}^{N} p_i \tag{7}$$

Here, c_A denotes the object center in 3D space, while Pi represents the object point cloud transformed from the i-th reference viewpoint. The rotation angles are determined by first identifying the principal axes of the object and then applying Principal Component Analysis (PCA) to analyze the three main directions of the point cloud. More specifically, given an object point cloud $\{p_i\}_{i=1}^{N}$, the covariance matrix of the point cloud is computed to extract its features [13]. The detailed calculation is defined as follows:

$$C = \frac{1}{N}\sum_{i=1}^{N}(p_i - c_A)(p_i - c_A)^T \tag{8}$$

Here, C denotes the covariance matrix obtained after feature extraction, and its eigenvectors v_1, v_2, v_3 represent the 3D orientation of the point cloud. Thus, the

rotational component is derived as $R = [v_1, v_2, v_3]$ and we employ R to align point clouds across different coordinate systems [28]. Since the alignment between R and the object remains relatively coarse at this stage, we use the Intersection over Union (IoU) as a metric to optimize the object's rotation and scale alignment [9,17]. By maximizing the IoU, the accuracy of object alignment is improved. This process is formulated as follows:

$$IoU(B_A, B_N) = \frac{Areaof Intersection}{Areaof Union} \tag{9}$$

Therefore, we obtain the coarse object alignment output $O_M^{'}$

3.3 Fine Object Alignment and Pose Estimation

After the coarse alignment of the query image with the reconstructed 3D object, the 3D shape of the object is further refined to facilitate precise pose inference. A joint optimization of the object scale s and pose T is performed to improve the accuracy of the final pose estimation. Specifically, the object pose $T_{O_M \to A}$ is estimated from the initially aligned object $O_M^{'}$ [42,44]. The coarse pose is then used to render the object shape in 3D space, and the projection error with respect to the anchor image is measured and taken as the loss. The process is defined as follows:

$$T_{O_M \to A} = arg \min_{T} Loss\left(I_A, Render(O_M, T)\right) \tag{10}$$

R denotes the 2D projection obtained after rendering with T and $O_M^{'}$. By minimizing the error between the projection and the anchor, the pose accuracy is improved. Meanwhile, we sample different object scales and compare them with the reference to optimize the object size s [30]. The process is defined as follows:

$$s = arg \min_{s \in [s_0, s_1]} IoU\left(O_M(s), O_N\right) \tag{11}$$

Here, s_0 and s_1 represent the lower and upper bounds of the sampled object scales. Based on this, the object scale and coarse pose are further refined, thereby improving the accuracy of subsequent pose regression for the query object. Using the jointly optimized object scale and pose, we regress the refined relative pose $T_{A \to Q}$ by combining the transformation matrix between the query image $T_{O_M \to Q}$ and the anchor image $T_{O_M \to A}$.

$$T_{A \to Q} = \left(T_{O_M \to A}\right)^{-1} \cdot T_{O_M \to Q} \tag{12}$$

Based on this, we obtain the refined 6D pose of the query image. For different hypothesized poses, we adopt a two-stage rendering and comparison strategy for selection, where a self-supervised attention mechanism is employed to enhance pose robustness. First, the hypothesized pose $I_{rendered1}$ is rendered into a view and compared with the corresponding rendered view $I_{rendered2}$ of the query image I q. The reprojection error between the two is then computed, producing a hypothesized feature vector e_i to evaluate the alignment accuracy between the hypothesized pose and the query image. The computation of this feature vector is illustrated by the following formula:

$$e_i = Compare(I_{rendered}, I_Q) \tag{13}$$

Subsequently, the feature vectors embedded from these hypothesized poses are weighted through a self-supervised attention mechanism to generate the corresponding selection scores. Benefiting from the strong capability of this mechanism to integrate global context, the model's focus on relevant regions is enhanced, further improving pose accuracy. Specifically, linear concatenation is applied to assemble the feature vectors $e_1, e_2, ..., e_N$ of all hypothesized poses into a single large vector, which serves as the input to the self-attention module. Within this mechanism, the correlation among vectors is evaluated, global information is incorporated to refine different weights, and the weighted output provides the accuracy scores for all hypotheses. The process is formulated as follows:

$$score_i = SelfAttention\left(Concat(e_1, e_2, ..., e_N)\right) \tag{14}$$

Finally, the pose is selected through Top-K filtering. Therefore, our two-stage rendering and self-supervised attention mechanism further enhance robustness under occlusion and large viewpoint variations, while the integration of global contextual information further improves pose accuracy.

4 Experiments

4.1 Datasets

We conduct experiments on publicly available benchmark datasets provided by BOP (Benchmark for 6D Object Pose Estimation) [31], evaluating our method under complex scenarios with occlusion and viewpoint variations, including Occlusion Linemod [11], Linemod [1], and YCB-Video (YCBV) [32].

- Linemod: This dataset is widely used for evaluating industrial object 6D pose estimation, containing RGB-D images of 15 object categories captured from multiple viewpoints. We evaluate our method across all object categories. The variations in texture, occlusion, and illumination conditions among these objects pose significant challenges for pose estimation.
- Occlusion Linemod: Constructed based on the Linemod dataset, this dataset focuses on objects in

highly complex environments with severe occlusions and overlaps. Compared to Linemod, it poses higher demands on pose estimation, as object visibility can be extremely low-sometimes less than 40 percent. Thus, it is particularly suitable for evaluating the generalization ability and robustness of our pose estimation method.

- YCB-Video (YCBV): This dataset includes a larger variety of object categories, greatly expanding the application scenarios. It consists of video sequences of 21 object categories across 12 real-world scenes containing daily objects in cluttered arrangements. With diverse viewpoints, object sizes, and heavy occlusions, YCBV presents significant challenges for testing the robustness and applicability of our method in real-world scenarios.

In summary, the three datasets progressively increase in difficulty-from single-object, multi-view settings to multi-object interactions with severe occlusion-thus providing a solid experimental foundation for comprehensively evaluating the effectiveness, robustness, and practical potential of our method.

4.2 Metrics

We mainly evaluate our method using the Average Distance of Model Points (ADD) [1,25]. Let the model point cloud set be M with size $|M|$, the predicted pose be (R, t), and the corresponding ground-truth pose be (R_{gt}, t_{gt}) ; then the ADD is computed as follows:

$$ADD = \frac{1}{|M|} \sum_{x \in M} \|(Rx + t) - (R_{gt}x + t_{gt})\|. \quad (15)$$

Here, x denotes a 3D point from the object's CAD model. This method measures the accuracy of rigid-body pose estimation by computing the average Euclidean distance between the 3D model points under the predicted pose and those under the ground-truth pose. For objects in the dataset with rotational invariance, ADD for symmetric objects (ADD-S) is used for evaluation. The core idea of this method is to compute the nearest-neighbor distance of each model point to assess pose error while ignoring the effects of rotational or reflectional symmetry. The ADD-S computation process is given as follows:

$$ADD\text{-}S = \frac{1}{|M|} \sum_{x \in M} \min_{y \in M} \|(Rx+t)-(R_{gt}y+t_{gt})\|. \quad (16)$$

In the above equation, y represents the neighboring point of x in the model. In addition, we use AUC (Area Under Curve) and Recall@0.1d to evaluate performance across different datasets. The area under the curve (AUC) is computed to examine the overall error distribution. Specifically, we integrate the recall curve of ADD/ADD-S with respect to varying thresholds $[0, 0.1 \times d]$, where d denotes the maximum distance between any two points in the object CAD model and represents the object's diameter. By adjusting the threshold $[0.1d, 0.05d, 0.01d]$, the robustness of our method under different accuracy levels can be assessed. For Recall@0.1d , we compute the recall at the threshold of $0.1 \times d$; that is, the proportion of predictions satisfying ADD ¡ 0.1d, or for symmetric objects satisfying ADD-S ¡ 0.1d, is considered as true positives. Through recall, we can clearly examine the success rate of our method at different precision levels $[0.1d, 0.05d, 0.01d]$ Our innovation focuses on cross-view feature fusion under sparse geometric constraints with only a small number of reference images, as well as improved robustness to occlusion and overlap. Based on this, we adopt CAD model point-based evaluation metrics (ADD/ADD-S) to more intuitively demonstrate the accuracy of 3D registration and point cloud geometric alignment. AUC reflects the overall error, while Recall@0.1d examines the practicality of our approach under different scenarios and accuracy levels across datasets.

4.3 Running Time

Our experiments were conducted on a workstation equipped with an Intel i5-14600KF CPU and an NVIDIA RTX 4080 Super GPU. For a single object, the total processing time to output the complete pose is approximately 1.1 s, consisting of about 0.12 s for coarse pose estimation and 3-D object reconstruction, 0.91 s for fine pose refinement, and 0.08 s for pose hypothesis selection.

4.4 Comparative Experiments

Figure 4: We conduct a qualitative visualization comparing our model-free, few-reference approach with other state-of-the-art methods. Specifically, we transform 1,000 points sampled from the object CAD model using the predicted pose into the camera coordinate system and project them onto the query image, where they are shown as green points.

We compare our proposed SFRF6D framework with both CAD-model-based methods and reference-image-based methods that do not rely on CAD models. Following the approach in Oryon (Any6D [10]), the predicted relative pose between the query and reference images is transformed into an absolute pose. Specifically, the predicted relative transformation $T_{A \to Q}$ between the query

Table 1: This figure presents recent state-of-the-art methods under few-reference settings (evaluation threshold: 0.1 d), showing that our method achieves higher accuracy with shorter runtime.

Method	CAD model	Modality	Ref-images	Datasets			Mean	Time
				LM	LMO	YCBV		
FS6D + ICP	✗	RGBD	16	88.9	58.6	79.8	75.8	1.2s
LatentFusion	✗	RGBD	16	80.2	53.5	76.6	70.1	2.0s
FoundationPose	✓	RGBD	16	97.8	75.3	93.1	89.7	1.0s
Any6D	✗	RGBD	1	89.3	57.9	82.3	76.5	2.1s
Ours	✗	RGBD	4	**92.1**	**67.3**	**90.2**	**83.2**	**0.9s**

and reference images is multiplied by the ground-truth pose matrix $T_{0\to A}$ of the object in the reference image to obtain the predicted pose of the object in the query image, which can then be directly aligned with the ground truth. The detailed process is expressed as follows:

$$\widehat{T}_{O\to Q} = T_{A\to Q} \cdot T_{O\to A} \tag{17}$$

We compare the proposed SFRF6D with both CAD-model-based and reference-image-based approaches. For CAD-based methods, we adopt FoundationPose as a strong baseline, which achieves higher accuracy but relies on CAD models. For reference-image-based methods, we consider FS6D+ICP, LatentFusion, and Any6D. Among them, Any6D can estimate object poses from a single reference image but its accuracy is limited by the image-to-3D alignment process. FS6D+ICP requires an additional ICP refinement stage on YCB-V and LM datasets to reach competitive RGB-D performance. LatentFusion builds cross-view relationships between reference and query images via voxel field learning; however, the training process is time-consuming, which limits its applicability in real-world scenarios. In contrast, SFRF6D does not require CAD models or per-dataset fine-tuning and outperforms existing RGB-D methods under the same conditions on the YCB-V and LM datasets. It also shows superior robustness in complex scenarios with severe occlusion and large viewpoint changes on the LMO dataset. Quantitative results are reported in Table 1: under identical conditions (no CAD model and a small number of reference images), SFRF6D achieves higher accuracy than the current state-of-the-art methods. FoundationPose attains high accuracy using CAD models but at the cost of longer inference time. Compared to the other non-model methods (FS6D+ICP, LatentFusion, and Any6D), our approach improves accuracy by 7.4 %,13.1%, and 6.7%, respectively, while reducing inference time by 0.3 s, 1.1 s, and 1.2 s. Qualitative comparisons are shown in Figure 4.

Table 2: Ablation Study Comparison Table

Model	ADD(-S)	Time
Origin	67.3	0.94s
No Geometric Attention	64.5	0.91s
No viewselector	66.2	1.32s
8 reference images	68.1	1.03s
2 reference images	63.4	0.93s
1 reference image	57.3	0.90s

4.5 Ablation Study

We validate the effectiveness of each module of SFRF6D on the publicly available LMO (Occlusion Linemod) dataset with severe overlap and occlusion. Table 2 presents the quantitative analysis results under different settings. We first replace the sparse geometry-guided module across reference view images in the model with simple fully connected attention. The runtime decreases by about 30 ms but the model accuracy drops by 2.7%. This demonstrates that, in occlusion environments, our method effectively removes ambiguous matches by means of the normalized direction-vector selection mechanism, thereby improving the generalization performance and robustness of the model. Second, we remove the view selector and use all rendered views for pose estimation. Clearly, the computational overhead thus increases significantly, extending by about 360 ms. Missing this design limits the model's ability to learn the matching between query and reference images, and the estimated pose accuracy decreases by about 1.1In addition, replacing the two-stage coarse-to-fine alignment and rendering process with a single direct regression stage causes a severe performance drop to only 43.2We also conduct an exploratory experiment with a very small number of reference images, setting the application scenarios to 2 and 1 reference image respectively. The experimental results show that when there is only one reference image, the reconstructed 3D shape depends heavily on this single reference image and the viewpoint

difference of the object in the query image. When occlusion or viewpoint difference is large, features cannot be robustly aligned, thus limiting pose accuracy. When the number of reference images increases to two, the accuracy improves by about 6.1%. Compared with increasing from two to four reference images, the accuracy rises by 2.2%. When the number of reference images increases to eight, although the accuracy improves by about 1%, the time overhead increases by about 11%. Therefore, our method achieves the best performance when the number of reference images is set to four, while maintaining a lightweight design.

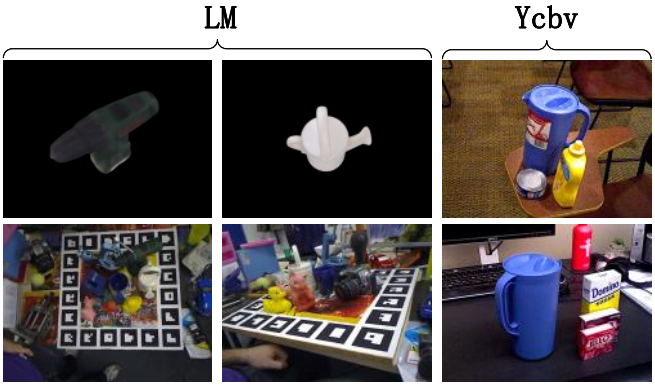

Figure 5: Dataset examples: the first row represents the training data, and the second row represents the testing data.

Therefore, this ablation study validates the complementarity, necessity, and effectiveness of the core modules of SFRF6D. Among them, the geometry-constrained sparse attention mechanism effectively reduces ambiguous feature matching in complex environments with overlap, occlusion, and large viewpoint differences; the view selector effectively lowers computational overhead while ensuring that the model consistently learns the most relevant features, thereby guaranteeing the cross-view fusion matching of reference images; and the two-stage coarse-to-fine 6D pose optimization process ensures the accuracy of the final predicted output. From the experiments with different numbers of reference images, the results show that SFRF6D achieves the optimal balance of performance and accuracy when using four reference images, avoiding the dependence on a large number of reference images in current research methods while still attaining high accuracy. All of the above demonstrates the robust generalization ability and practicality of our method in real-world scenarios.

5 Analysis and Summary

To address the reliance of existing 6D pose estimation methods on object CAD models and large numbers of reference images, we propose SFRF6D, a method that requires no CAD model and only a few reference images, offering a lighter-weight and more robust solution. Leveraging a geometry-guided sparse cross-view attention mechanism, the model exhibits stable performance in complex scenes. Experimental results demonstrate that our method achieves excellent performance on public benchmark datasets Linemod (LM), Occlusion Linemod (LMO), and YCB-Video (YCBV), particularly validating its robustness under severe occlusion in LMO. Compared with CAD-model-based approaches such as FoundationPose, our method shows slightly lower precision but offers higher practicality and reduced computational load; compared to FS6D+ICP, SFRF6D maintains high performance without additional optimization or fine-tuning.

Specifically, SFRF6D provides a new paradigm for 6D pose estimation, eliminating the need for object CAD models, reducing the number of reference images, and ensuring efficient estimation. This makes it particularly suitable for deployment in complex tasks such as robotic grasping. Future work may incorporate pre-trained foundation models, as in FoundationPose, to further improve cross-view generalization. Extensions could include multi-object interaction in dynamic scenes, continued robustness improvements in complex environments, and model lightweighting to meet the requirements of mobile and edge computing platforms.

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 *ECCV*, 2014.

[2] Ordoumpozanis, K. and Papakostas, G. A. Reviewing 6d pose estimation: Model strengths, limitations, and application fields. *Applied Sciences*, 15(6):3284, 2025.

[3] Yisheng He, Haibin Huang, Huihui Pan, Qihao Li, and Jian Sun. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019.

[4] Ryo Araki, Kohei Mano, Tetsuya Hirano, Takumi Hirakawa, et al. Iterative coarse-to-fine 6d-pose estimation using back-propagation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[5] Fabian Manhardt, Wadim Kehl, and Nassir Navab. Deep model-based 6d pose refinement in rgb. In *ECCV*, 2018.

[6] Kartik Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *ECCV*, 2018.

[7] Deng, X. and Xiang, Y. and Mousavian, A. and Fox, D. Self-supervised 6d object pose estimation for robot manipulation. *IEEE Transactions on Robotics*, 2020.

[8] Bugra Tekin, Sudipta N. Sinha, and pascal fua. real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] Chen Song, Jiarui Xu, Xin Yu, and Dacheng Tao. Hybridpose: 6d object pose estimation under hybrid representations. In *CVPR*, 2020.

[10] S. Won, h. lee, and g.h. yang. zero-shot 6d pose tracking for sequential robot manipulation using sam and online segmentation models. In *24th International Conference on Advanced Robotics (ICAR)*, 2024.

[11] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. Introducing a framework for evaluation of 6d object pose estimation under severe occlusion: The occlusion linemod dataset. In *CVPR Workshops*, 2017.

[12] He, W. and Feng, X. and Zhao, Y. and Lv, Y. 6d pose estimation of objects: Recent technologies and challenges. *Applied Sciences*, 2020.

[13] Wang, F. and Manhardt, F. and Tombari, N. and Navab, N. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *CVPR Workshops*, 2021.

[14] Shugurov, S. and Zakharov, S. and Ilic, S. Dpodv2: Dense correspondence-based 6dof pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[15] Hönig, S. and Thalhammer, M. and Vincze, M. Improving 2d-3d dense correspondences with diffusion models for 6d object pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[16] Xingyu Liu, Haoyu Ma, Yiming Zuo, Jianfeng Zhang, and others. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *CVPR*, 2022.

[17] Zhipeng Cai, Zhihao Li, Fei Deng, Tao Yu, and others. Onepose: One-shot object pose estimation without cad models. In *CVPR*, 2022.

[18] Zhipeng Cai, Zhihao Li, Yixuan Wang, Tao Yu, and others. Onepose++: Keypoint-free one-shot object pose estimation without cad models. In *NeurIPS*, 2022.

[19] Wen, Bowen and Yang, Wei and Kautz, Jan and Birchfield, Stan. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024.

[20] Xingyu Liu, Yiming Zuo, Haoyu Ma, Jianfeng Zhang, and others. Fs6d: Few-shot 6d object pose estimation with uncertainty-guided self-training. In *CVPR*, 2023.

[21] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021.

[22] Corsetti, Jaime and Boscaini, Davide and Oh, Changjae and Cavallaro, Andrea and Poiesi, Fabio. Open-vocabulary object 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18071–18080, 2024.

[23] Corsetti, Jaime and Boscaini, Davide and Giuliari, Francesco and Oh, Changjae and Cavallaro, Andrea and Poiesi, Fabio. High-resolution open-vocabulary object 6d pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[24] K. He, X. Zhang, S. Ren, and J. Sun. He, x. zhang, s. ren, and j. sun. deep residual learning for image recognition. In *CVPR*, 2016.

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. Dosovitskiy, l. beyer, a. kolesnikov, et al. an image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[26] A. Vaswani, N. Shazeer, N. Parmar, et al. . Vaswani, n. shazeer, n. parmar, et al. attention is all you need. In *NeurIPS*, 2017.

[27] N. Carion, F. Massa, G. Synnaeve, et al. Carion, f. massa, g. synnaeve, et al. end-to-end object detection with transformers. In *ECCV*, 2020.

[28] Z. Chen, Y. Guo, X. Wang, and W. Wan. Chen, y. guo, x. wang, and w. wan. latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *CVPR*, 2021.

[29] Kendall, A. and Grimes, M. and Cipolla, R. Posenet: A convolutional network for real-time 6-dof camera relocalization. *arXiv preprint arXiv:1505.07427*, 2015.

[30] Sachidanandan, Adnaan Ali. 3d pose estimation and topology reconstruction using foundation models and render and compare.

[31] S. Peng, Y. Liu, Q. Huang, et al. Peng, y. liu, q. huang, et al. pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019.

[32] Kleeberger, Kilian and Landgraf, Christian and Huber, Marco F. Large-scale 6d object pose estimation dataset for industrial bin-picking. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2573–2578. IEEE, 2019.