Open-Vocabulary Object 6D Pose Estimation via Modulated Textual Semantics

Zixuan Sun[1,2], **Hui Shuai**[1,2] and **Qingshan Liu**[1,2]
[1]Interdisciplinary Center for Future Information, Nanjing University of Posts and Telecommunications
[2]National Key Laboratory of Tibetan Language Intelligence

Abstract

Estimating the 6D pose of novel objects without CAD models or video sequences remains a challenging problem. Recent works explore text-driven approaches to address this challenge in an open-vocabulary manner. However, these methods typically treat text embeddings as static priors, which lack the flexibility to adapt to specific visual scenes, thereby limiting reliable cross-view correspondences. In this paper, we propose a framework that modulates static textual embeddings into adaptable semantic guidance for RGB-D open-vocabulary 6D pose estimation. Specifically, we introduce a Text Feature Modulation (TFM) module that uses learnable queries to transform fixed text embeddings into adaptive semantic representations. These representations are explicitly calibrated by a Foreground-Grounded Semantic Alignment (FGSA) strategy, which suppresses background distractions via contrastive learning. Furthermore, we design a Semantic-Anchored Cross-view Fusion (SACF) module that aggregates image features from the reference and query views using modulated semantics as an invariant anchor. Finally, we predict object segmentation and establish semantic-aware correspondences, leveraging modulated textual semantics to filter mismatches for robust 6D pose estimation. Extensive experiments on standard benchmarks demonstrate that our method achieves superior performance.

1 Introduction

6-degree-of-freedom (6D) pose estimation aims to recover the 3D rotation and 3D translation of an object relative to the camera coordinate system, serving as a fundamental component in robotic manipulation [Liu *et al.*, 2023] and augmented reality [Tang *et al.*, 2019]. Although existing instance-level and category-level methods [Peng *et al.*, 2019; Labbé *et al.*, 2020; Wang *et al.*, 2019; Ji *et al.*, 2025] achieve high-precision pose estimation, they rely heavily on large-scale training data with object-specific or category-normalized priors, limiting their applicability in open-world scenarios.

To overcome these limitations, recent research has shifted toward novel object pose estimation, aiming to generalize to objects unseen during training. However, most existing approaches [Liu *et al.*, 2022; Wen *et al.*, 2024; Deng *et al.*, 2025; Huang *et al.*, 2025] typically require high-quality 3D models or multi-view reference images of the target object during inference. While effective, preparing such data for every novel object incurs significant preparation and computational overheads, hindering their deployment in real-time applications. To reduce this dependence, a more practical paradigm is to infer the 6D pose using only a single reference view [Corsetti *et al.*, 2024; Lee *et al.*, 2025; Liu *et al.*, 2025]. Given the limited visual information in this setting, augmenting the sparse observation with auxiliary representations has emerged as a critical solution to ensure robustness. Recent methods [Corsetti *et al.*, 2024; Corsetti *et al.*, 2025] demonstrate that textual prompts can be used to facilitate pose estimation of novel objects. These methods exhibit remarkable adaptability, performing effectively even when reference and query images are captured from completely different environments. However, treating text as a rigid prior limits its robustness in complex scenarios. Without adapting to the visible content, such static semantics fail to capture reliable cross-view feature correspondences essential for precise pose estimation.

In this work, we propose a unified framework that modulates static textual embeddings into adaptable semantic guidance for RGB-D open-vocabulary 6D pose estimation. Unlike prior works that directly use generic text priors, we introduce a TFM module. TFM employs a set of learnable latent queries to interact with frozen text embeddings, optimized to transform static textual descriptions into task-adaptive representations. To ensure these representations are robust to clutter, we propose a FGSA strategy. FGSA employs a metric learning objective [Oord *et al.*, 2018] to strictly align the modulated textual representations with the visual foreground, while explicitly suppressing interference from visually similar background regions. Furthermore, we design the SACF module to bridge the visual gap caused by distinct environments in the reference and query views. By fusing the visual features of both the reference and query views with modulated semantics as an invariant anchor, SACF enables consistent feature aggregation across different views. Finally, the fused features are utilized to predict object segmentation and

establish semantic-aware matching. Unlike methods [Corsetti *et al.*, 2024; Corsetti *et al.*, 2025] that rely solely on visual similarity for matching, we introduce a textual semantic consistency score to dynamically suppress mismatches.

In summary, our main contributions are as follows:

- We propose a framework that modulates static textual embeddings into adaptable semantic guidance for open-vocabulary object 6D pose estimation.
- We introduce the TFM module for dynamic textual semantics modulation, the FGSA strategy for discriminative foreground grounding, and the SACF module for semantically-guided cross-view feature fusion.
- The experimental results on RGB-D open-vocabulary object pose estimation benchmarks demonstrate that our method achieves superior performance.

2 Related Work

Recent research in 6D pose estimation has primarily focused on handling novel objects that were not seen during training. Depending on the reference information available at inference time, existing approaches are generally divided into three main groups: model-based methods, multi-reference view methods and single-reference view methods. Below, we provide a brief overview of all three approaches.

Model-based novel object pose estimation. CAD model-based approaches typically require high-quality 3D models of the target objects during inference to establish correspondences with the input image. OVE6D [Cai *et al.*, 2022] establishes a universal framework for pose estimation from depth images by encoding object viewpoints relative to CAD models. MegaPose [Labbé *et al.*, 2022] adopts a coarse-to-fine strategy, first classifying a coarse pose from a massive set of synthetic templates and then refining it via a render-and-compare network. To improve efficiency, GigaPose [Nguyen *et al.*, 2024b] decouples pose prediction and renders only a small number of templates using CAD models, thereby significantly accelerating the inference process. Although these methods achieve high accuracy, render-and-compare is computationally intensive and makes it difficult to obtain CAD models for novel objects in real-world scenarios.

Multi-reference view novel object pose estimation. To eliminate the dependency on CAD models, several methods utilize video sequences or multiple reference images. Early learning-based approaches like LatentFusion [Park *et al.*, 2020] construct a latent 3D representation from a small set of reference views and regress the pose via a render-and-compare strategy. To improve robustness against occlusion and clutter, FS6D [He *et al.*, 2022b] employs a prototypical network to aggregate features from multiple support views for pose estimation. OnePose [Sun *et al.*, 2022] and OnePose++ [He *et al.*, 2022a] reconstruct a sparse 3D point cloud from a video sequence using Structure-from-Motion and match 2D keypoints in the query image to this 3D representation. FoundationPose [Wen *et al.*, 2024] establishes a unified framework for both model-based and model-free settings. In its model-free configuration, it leverages a neural field representation synthesized from a small number of reference images to perform robust tracking and estimation. Although effective, collecting complete video scans or calibrated multi-view image sets for each new object still incurs significant preparation overhead.

Single-reference view novel object pose estimation. This category targets the most challenging setting where only a single reference view is available. Due to the limited visual information available, directly recovering the 6D pose is a highly ill-posed problem. To overcome this, recent approaches attempt to augment the sparse observation with *auxiliary representations*. For instance, NOPE [Nguyen *et al.*, 2024a] proposes to learn a canonical representation directly from a single image to infer the object pose. However, due to the inherent scale ambiguity in monocular estimation, it is primarily limited to estimating 3D rotation rather than the full 6D pose. Reconstruction-based methods like Any6D [Lee *et al.*, 2025] attempt to lift the single reference image into a pseudo-3D model to facilitate 3D-2D matching. In addition, some methods explore text prompts as an additional source of semantic information. Oryon [Corsetti *et al.*, 2024] and Horyon [Corsetti *et al.*, 2025] leverage Vision-Language Models (VLMs) to guide the pose estimation process using text descriptions. These methods exhibit remarkable adaptability to environmental changes. However, they typically treat text embeddings as static priors, failing to adapt semantics to the specific visual context of the target object. In contrast, our framework modulates these static embeddings into task-adaptive semantic cues, ensuring robust cross-modal alignment and reliable feature correspondences.

3 Method

We follow Oryon [Corsetti *et al.*, 2024] to address the problem of open-vocabulary 6D pose estimation, where the target object is defined solely by a user-provided textual prompt and is unseen during training.

Formally, given a textual description T and two RGB-D images $\{(I_r, D_r), (I_q, D_q)\}$ captured from different scenes, our framework aims to jointly predict object segmentation masks, establish cross-view correspondences, and recover the relative pose. An overview of the proposed pipeline is shown in Fig. 1. First, the textual description T is encoded by a frozen text encoder [Radford *et al.*, 2021] to obtain textual embeddings. These embeddings are then refined by the TFM module, which employs learnable queries to transform generic semantics into task-adaptive representations. The input images I_r and I_q are processed by a frozen image encoder [Radford *et al.*, 2021] to extract dense visual feature maps. To ensure the modulated semantics are discriminative, we introduce the FGSA strategy. FGSA leverages foreground supervision to guide the calibrated textual semantics toward object-related regions while explicitly pushing away background distractors via a contrastive objective. This strategy operates only during training, introducing no additional computational overhead at inference time. Then, the SACF module utilizes the calibrated semantics as an invariant anchor to aggregate visual features across the reference and query views. Based on the fused representations, the framework predicts segmentation masks of object regions and computes

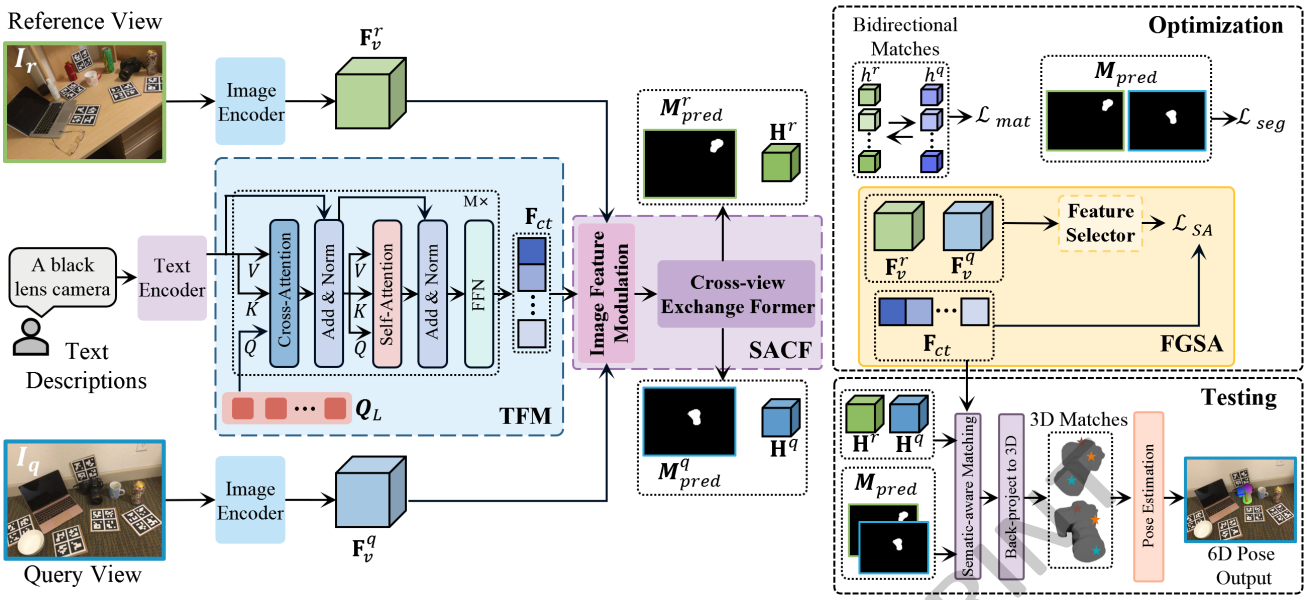

Figure 1: Overview of the proposed framework. Taking a reference image I_r, a query image I_q, and a text description T as input, our method first uses the TFM module to transform static text embeddings into task-adaptive semantic representations $\mathbf{F}_{ct}$. During training, the FGSA strategy explicitly calibrates $\mathbf{F}_{ct}$ to align with visual foreground features $(\mathbf{F}_v^r, \mathbf{F}_v^q)$ via the semantic alignment loss $\mathcal{L}_{SA}$. Conditioned on the calibrated semantics, the SACF module aggregates features across views to produce the fused representations $\mathbf{H}^r, \mathbf{H}^q$. Finally, the network predicts segmentation masks $\mathbf{M}_{pred}$ and dense correspondences, which are lifted to 3D using depth maps D to estimate the 6D pose.

semantic-aware cross-view correspondences. Finally, following [Corsetti *et al.*, 2024], the predicted matches are projected into 3D space using the depth maps D_r, D_q, and camera parameters, and the relative 6D pose is then solved via a point cloud registration algorithm [Bai *et al.*, 2021].

3.1 Text Feature Modulation Module

CLIP [Radford *et al.*, 2021] text embeddings align with global image representations and work well for image-level retrieval, but these frozen embeddings often lack the fine-grained detail and task-specific adaptation required for feature matching and 6D pose estimation. To address this, we propose the TFM module, which transforms static user-provided descriptions into adaptable, task-aware semantic representations without fine-tuning the heavy text encoder.

Let $\mathbf{F}_t$ denote the initial text embeddings extracted from the description using a frozen text encoder [Radford *et al.*, 2021]. Rather than using these fixed features directly, we initialize a set of learnable latent queries $\mathbf{Q}_L$ from a Gaussian distribution. These queries act as a flexible medium to extract and reorganize semantic information from the fixed text embeddings. The TFM module is structured as a stack of M transformer blocks. As illustrated in Fig. 1, unlike standard decoder architectures, we design our block to prioritize the extraction of semantic cues. Formally, let $\mathbf{Q}^{(m-1)}$ denote the input queries to the m-th block. The modulation starts with a multi-head cross-attention (MCA) layer, where $\mathbf{Q}^{(m-1)}$ serves as the query, and the text embeddings $\mathbf{F}_t$ serve as both key and value. The features are then processed sequentially by a multi-head self-attention (MSA) layer and a feed-forward network (FFN). Each sub-layer adopts a post-normalization design, with a residual connection followed by layer normalization (LN):

$$\mathbf{Z}_c^{(m)} = \mathrm{LN}\left(\mathbf{Q}^{(m-1)} + \mathrm{MCA}\left(\mathbf{Q}^{(m-1)}, \mathbf{F}_t, \mathbf{F}_t\right)\right), \quad (1)$$

$$\mathbf{Z}_s^{(m)} = \mathrm{LN}\left(\mathbf{Z}_c^{(m)} + \mathrm{MSA}\left(\mathbf{Z}_c^{(m)}\right)\right), \quad (2)$$

$$\mathbf{Q}^{(m)} = \mathrm{LN}\left(\mathbf{Z}_s^{(m)} + \mathrm{FFN}\left(\mathbf{Z}_s^{(m)}\right)\right). \quad (3)$$

The output of the final block is denoted as the calibrated text embeddings $\mathbf{F}_{ct}$. Driven by the end-to-end training objective, these learnable queries are optimized to selectively aggregate semantic cues relevant to the target object, yielding a compact representation that is better suited for the subsequent visual matching task than the original generic embeddings.

3.2 Semantic-Anchored Cross-View Fusion

Reliable 6D pose estimation requires accurate correspondences between reference and query views. To achieve this, we propose the SACF module, which aggregates multi-view features guided by the shared textual calibrated semantics. As shown in Fig. 1, SACF consists of two stages: *Image Feature Modulation* and the *Cross-view Exchange Former*.

First, we inject the calibrated textual semantics $\mathbf{F}_{ct}$ from the TFM module into the visual features $\mathbf{F}_v^i$ of both views $(i \in \{r, q\})$. We adopt the cross-attention architecture following [Liu *et al.*, 2021; Cho *et al.*, 2024] to perform image feature modulation and obtain semantically enhanced feature maps $\hat{\mathbf{F}}_v^i$, ensuring that subsequent fusion is grounded in the object's specific semantic attributes described by the user. After image feature modulation, the next step is to establish interaction between the two views. Specifically, the features $\hat{\mathbf{F}}_v^i$

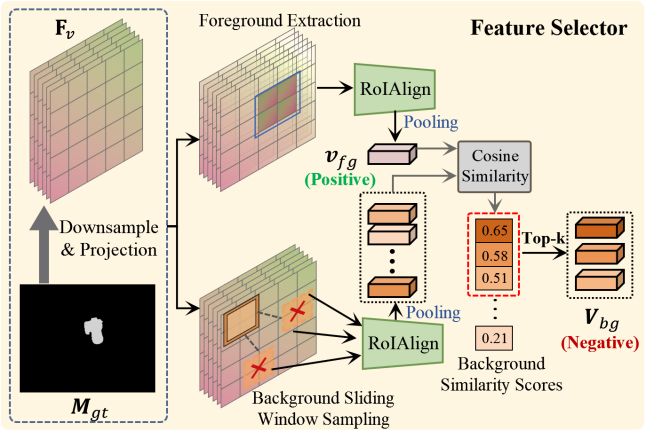

Figure 2: Illustration of the feature selector module. Given the image feature map and ground-truth mask, we first extract the foreground feature vector (positive sample) via RoIAlign and pooling. Then, we employ a sliding window strategy to sample background regions and compute their cosine similarity with the foreground representation. The top-k background regions with the highest similarity scores are selected as hard negative samples to supervise the semantic alignment.

are first augmented with learnable 2D positional embeddings $\mathbf{P}$ and processed by depth-wise convolutions to incorporate local geometric context:

$$\mathbf{F}_g^i = \mathrm{DWConv}(\hat{\mathbf{F}}_v^i + \mathbf{P}). \tag{4}$$

Then, inspired by [Cui *et al.*, 2022], we project $\mathbf{F}_g^i$ into queries, keys, and values, and design a bidirectional exchange mechanism that swaps queries between the reference and query views to enhance shared semantic regions across the two views. We formulate the exchange former as:

$$\mathbf{H}_{attn}^r = \mathrm{MultiHeadAttn}(\mathbf{Q}(\mathbf{F}_g^r), \mathbf{K}(\mathbf{F}_g^q), \mathbf{V}(\mathbf{F}_g^q)), \tag{5}$$

$$\mathbf{H}_{attn}^q = \mathrm{MultiHeadAttn}(\mathbf{Q}(\mathbf{F}_g^q), \mathbf{K}(\mathbf{F}_g^r), \mathbf{V}(\mathbf{F}_g^r)). \tag{6}$$

Subsequently, we upsample features $\mathbf{H}_{attn}^r$ and $\mathbf{H}_{attn}^q$ to obtain the fused features $\mathbf{H}^r$ and $\mathbf{H}^q$, which are then used to predict the object segmentation mask and compute cross-view feature correspondences.

3.3 Foreground-Grounded Semantic Alignment

The TFM module implicitly calibrates textual semantic embeddings for pose estimation, but without explicit constraints, it may overfit to generic scene contexts rather than the target object. To this end, we propose the FGSA strategy. As illustrated in Fig. 1, FGSA operates only during training and consists of two components: a feature selector and a semantic alignment objective.

Feature Selector. The feature selector is designed to identify the visual foreground and mine hard background negatives, enabling explicit contrastive supervision. Given an image feature map $\mathbf{F}_v$ and the binary object mask $\mathbf{M}_{gt}$, we first extract the foreground representation. Specifically, we apply RoIAlign [He *et al.*, 2017] on the object bounding box derived from $\mathbf{M}_{gt}$, followed by global average pooling (GAP) to produce a compact foreground vector $\mathbf{v}_{fg}$. To identify visually confusing background regions, we employ a sliding window strategy over $\mathbf{F}_v$. The window size is dynamically adapted to the object's scale to ensure consistency. This generates a set of candidate background windows $\mathcal{W}_{bg} = \{w_i\}$, excluding those overlapping with the foreground. For each window w_i, we extract its feature representation using the same RoIAlign and GAP operations as the foreground to ensure feature space alignment. Let $\mathbf{v}_i$ denote the extracted vector for window w_i. We then compute the visual similarity between each background candidate and the foreground vector:

$$s_i = \mathrm{sim}(\mathbf{v}_i, \mathbf{v}_{fg}), \tag{7}$$

where $\mathrm{sim}(\cdot)$ is the cosine similarity. We then select the top-k background windows with the highest scores s_i as the hard negative samples, denoted as $\mathbf{V}_{bg} = \{\mathbf{v}_{bg}^j\}_{j=1}^k$. These regions share similar visual statistics with the object and are most likely to confuse the model.

Semantic Alignment Objective. The goal of FGSA is to align the calibrated text embeddings $\mathbf{F}_{ct}$ with the visual foreground while pushing away hard negatives. We formulate this using a contrastive InfoNCE loss [Oord *et al.*, 2018]. Here, we treat $\mathbf{F}_{ct}$ as anchors, $\mathbf{v}_{fg}$ as positive keys, and the hard background features $\mathbf{V}_{bg}$ as negative keys. The semantic alignment objective is formulated as:

$$\mathcal{L}_{\mathrm{SA}} = -\sum_{n=1}^{N} \log \frac{s(\mathbf{f}_{ct}^{(n)}, \mathbf{v}_{fg})}{s(\mathbf{f}_{ct}^{(n)}, \mathbf{v}_{fg}) + \sum_{j=1}^{k} s(\mathbf{f}_{ct}^{(n)}, \mathbf{v}_{bg}^j)}, \tag{8}$$

where $\mathbf{f}_{ct}^{(n)}$ denotes the n-th calibrated query vector in $\mathbf{F}_{ct}$, and $s(\mathbf{t}, \mathbf{v}) = \exp(\langle \mathbf{t}, \mathbf{v} \rangle / \tau)$ defines the exponentiated cosine similarity with temperature τ.

By integrating $\mathcal{L}_{\mathrm{SA}}$ during training, the TFM module is guided to learn calibrated text semantics that are strictly grounded in object regions, reducing background interference and better supporting robust 6D pose estimation.

3.4 Semantic-Aware Matching

During inference, we estimate the 6D pose by establishing correspondences between reference and query views. Unlike prior works [Corsetti *et al.*, 2024] that rely solely on visual similarity, which can be ambiguous in cluttered scenes, we leverage calibrated textual semantics to verify match reliability. Specifically, we propose a semantic-aware matching score that combines visual and semantic similarity.

Given the fused feature maps $\mathbf{H}^r, \mathbf{H}^q$ and predicted masks $\mathbf{M}_{pred}^r, \mathbf{M}_{pred}^q$, we first extract the valid features $\mathcal{V}^r$ and $\mathcal{V}^q$. For a pair of feature points $\mathbf{h}_i \in \mathcal{V}^r$ and $\mathbf{h}_j \in \mathcal{V}^q$, the visual similarity is computed as $S_{vis}(i,j) = \langle \mathbf{h}_i, \mathbf{h}_j \rangle$. To incorporate semantic guidance, we compute a consistency score by measuring the alignment of each visual feature with the textual representation. We derive a holistic semantic prototype $\bar{\mathbf{t}}_{ct} = \mathrm{Mean}(\mathbf{F}_{ct})$ to serve as a global reference. In practice, since the fused features $\mathbf{H}^r$ and $\mathbf{H}^q$ have different dimensions from the modulated textual features, we perform this verification by retrieving feature vectors from the feature map $\mathbf{F}_v$ at the spatial locations corresponding to $\mathbf{h}_i$ and $\mathbf{h}_j$. The semantic consistency score is defined as:

$$S_{sem}(i,j) = \sigma(\langle \mathbf{f}_v(\mathbf{p}_i), \bar{\mathbf{t}}_{ct} \rangle) \cdot \sigma(\langle \mathbf{f}_v(\mathbf{p}_j), \bar{\mathbf{t}}_{ct} \rangle), \tag{9}$$

where $\mathbf{f}_v(\mathbf{p}_i)$ denotes the visual feature vector sampled from $\mathbf{F}_v$ at the position $\mathbf{p}_i$ of point i, and $\sigma(\cdot)$ is the sigmoid function. This acts as a soft gate that suppresses matches where either point fails to align with the user-provided description. The final matching score is a weighted combination:

$$S_{final}(i,j) = (1-\lambda_{sem})S_{vis}(i,j) + \lambda_{sem}S_{sem}(i,j), \quad (10)$$

where λ_{sem} balances the contribution of visual and semantic cues. We retain matches with $S_{\text{final}}(i,j) > \theta_{\text{conf}}$ using a confidence threshold θ_{conf}. The filtered matches are back-projected to 3D using depth maps and camera intrinsics, and the pose is solved via the PointDSC [Bai *et al.*, 2021] algorithm.

3.5 Loss Functions

Our framework is trained in an end-to-end manner using a multi-task objective. The total loss function $\mathcal{L}_{total}$ is composed of matching loss $\mathcal{L}_{match}$, segmentation loss $\mathcal{L}_{seg}$ and semantic alignment loss $\mathcal{L}_{\text{SA}}$.

Matching Loss ($\mathcal{L}_{match}$). We establish the set of ground-truth correspondences $\mathcal{P}_{gt}$ by projecting pixels from the reference view to the query view using the ground-truth depth and relative camera poses. To avoid ambiguous many-to-one matches, we employ a symmetric dual-softmax loss [Cheng *et al.*, 2021; Sun *et al.*, 2021], which enforces cycle consistency, requiring the matched features to be mutual nearest neighbors in both views. Let $\mathbf{S}$ be the similarity matrix where $\mathbf{S}_{ij} = \langle \mathbf{h}_i^r, \mathbf{h}_j^q \rangle / \tau$. We compute the matching probability from reference to query ($P_{r\to q}$) and from query to reference ($P_{q\to r}$) using the softmax function:

$$P_{r\to q}(j|i) = \frac{\exp(\mathbf{S}_{ij})}{\sum_k \exp(\mathbf{S}_{ik})}, P_{q\to r}(i|j) = \frac{\exp(\mathbf{S}_{ij})}{\sum_k \exp(\mathbf{S}_{kj})}. \quad (11)$$

The matching loss is defined as the negative log-likelihood of the joint probability for all valid ground-truth pairs:

$$\mathcal{L}_{match} = -\sum_{(i,j)\in\mathcal{P}_{gt}} \log\left(P_{r\to q}(j|i)\cdot P_{q\to r}(i|j)\right). \quad (12)$$

By maximizing the product of bidirectional probabilities, this loss strictly penalizes ambiguous matches and encourages a one-to-one correspondence between $\mathbf{H}^r$ and $\mathbf{H}^q$.

Segmentation Loss ($\mathcal{L}_{seg}$). We utilize the Dice Loss for segmentation supervision:

$$\mathcal{L}_{seg} = 1 - \frac{2\sum_i \mathbf{M}_{pred}^{(i)}\mathbf{M}_{gt}^{(i)}}{\sum_i \mathbf{M}_{pred}^{(i)} + \sum_i \mathbf{M}_{gt}^{(i)} + \epsilon}, \quad (13)$$

where i iterates over the reference and query views, and $\mathbf{M}_{pred}^i, \mathbf{M}_{gt}^i$ denote the predicted probability map and the binary ground-truth mask for view i, respectively.

Semantic Alignment Loss ($\mathcal{L}_{\text{SA}}$). As described in Sec. 3.3, this loss explicitly aligns the calibrated textual embeddings with the visual foreground features while pushing away mined hard background negatives.

Finally, the total loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{match} + \lambda_{seg}\mathcal{L}_{seg} + \lambda_{align}\mathcal{L}_{\text{SA}}, \quad (14)$$

where λ_{seg} and λ_{align} are balancing weights.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. Under the standard novel-object 6D pose estimation setting, we train on synthetic data and evaluate on real-world datasets without fine-tuning on test objects. Our model is trained and evaluated on benchmarks established by Oryon [Corsetti *et al.*, 2024] based on the ShapeNet6D [He *et al.*, 2022b], REAL275 [Wang *et al.*, 2019], and Toyota-Light [Hodan *et al.*, 2018] datasets. We term them Shape6D-T, REAL-T, and TL-T, respectively, which focus on determining the relative pose of a text-prompted object captured in two distinct scenes. Shape6D-T is a large-scale synthetic dataset containing approximately 800K images, where the training text descriptions are constructed from object names and synonyms available in the metadata. We evaluate our model on REAL-T and TL-T, each of which contains 2K image pairs selected from real-world scenes with mild occlusions and challenging lighting variations. The textual descriptions for REAL-T and TL-T are constructed by combining specific visual attributes with object categories.

Evaluation Metrics. Following the standard BOP protocol [Hodan *et al.*, 2018], we report Visual Surface Discrepancy (VSD), Maximum Symmetry-Aware Surface Distance (MSSD), and Maximum Symmetry-Aware Projection Distance (MSPD), and calculate their mean as the Average Recall (AR). We also use the ADD(S)-0.1d [Xiang *et al.*, 2017] metric to evaluate pose recall with a 10% diameter threshold, which implicitly handles object symmetry. For segmentation, we employ the mean Intersection-over-Union (mIoU).

4.2 Implementation Details

For the visual and text encoders, we utilize the CLIP [Radford *et al.*, 2021] backbone, which is frozen during training. The TFM module consists of $M = 4$ transformer layers. We train our model in an end-to-end manner on 4 NVIDIA RTX 4090 GPUs. Optimization is performed using the AdamW optimizer [Loshchilov and Hutter, 2017] with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. We employ a cosine annealing scheduler with a warm-up period of 1,000 iterations. The batch size is set to 16, and the model is trained for 25 epochs. The loss weights are empirically set to $\lambda_{seg} = 1.0$, $\lambda_{align} = 0.5$ and the number of hard negative samples in FGSA is set to k=3. For the semantic-aware matching, the balancing weight is set to $\lambda_{sem} = 0.2$, and we filter matches with a confidence threshold $\theta_{conf} = 0.3$. To prevent overfitting, we apply data augmentations including random color jittering, Gaussian blur and flipping.

4.3 Comparison with State-of-the-Art

We primarily compare our method with state-of-the-art single reference-view approaches, including the point cloud registration method ObjectMatch [Gümeli *et al.*, 2023], the cross-view matching method Oryon [Corsetti *et al.*, 2024], and the 3D model reconstruction method Any6D [Lee *et al.*, 2025].

Quantitative results. Table 1 shows a comprehensive comparison on the REAL-T and TL-T datasets. In the practical setting where segmentation masks must be predicted, our

Figure 3: Qualitative comparison with state-of-the-art methods using predicted segmentation masks on the REAL-T and TL-T datasets. The top row visualizes the estimated 6D poses, while the bottom row displays the corresponding predicted segmentation masks. We visualize the estimated 6D poses by projecting the coordinate axes onto the query images.

Method	Mask	REAL-T						TL-T					
		VSD	MSSD	MSPD	AR	ADD	mIoU	VSD	MSSD	MSPD	AR	ADD	mIoU
Obj. Mat.	Oryon	14.1	27.9	25.2	22.4	13.2	66.5	2.2	10.5	12.1	8.3	3.8	68.1
	Ours	18.4	32.0	31.8	27.4	12.9	74.2	2.7	12.6	15.1	10.1	5.2	72.3
Oryon	Oryon	23.6	36.6	36.4	32.2	24.3	66.5	12.1	37.5	41.4	30.3	20.9	68.1
	Ours	26.7	39.5	40.3	35.5	30.1	74.2	11.6	40.5	43.3	31.8	21.7	72.3
Ours	Oryon	25.6	39.8	40.7	35.4	32.3	66.5	13.4	41.5	46.1	33.7	23.2	68.1
	Ours	31.8	44.6	45.6	40.7	38.7	**74.2**	14.1	47.8	50.8	37.6	28.4	**72.3**
Obj.Mat.	GT	15.5	31.7	30.8	26.0	13.4	100.0	2.4	13.0	14.0	9.8	5.4	100.0
Oryon	GT	32.1	50.9	56.7	46.5	34.9	100.0	13.9	42.9	45.5	34.1	22.9	100.0
Any6D	GT	31.1	56.5	65.3	51.0	**53.5**	100.0	15.8	**55.8**	58.4	**43.3**	**32.2**	100.0
Ours	GT	**35.4**	**56.9**	**66.5**	**52.9**	51.4	100.0	**16.3**	53.1	**59.6**	43.0	31.2	100.0

Table 1: Quantitative comparison on the REAL-T and TL-T datasets. We report the performance using both predicted segmentation masks and ground-truth (GT) masks. **Bold** indicates the best result, and underlined indicates the second best.

method achieves significant improvements. Specifically, on the REAL-T dataset, our method achieves an AR of 40.7, surpassing the previous SOTA Oryon by 8.5 points. On the challenging TL-T dataset, which features severe lighting variations, our method achieves an AR of 37.6 compared to 30.3 for Oryon. This demonstrates that our calibrated textual semantics offer more targeted and robust guidance for pose estimation in complex environments. Our framework also achieves higher mIoU scores of 74.2 and 72.3 on the REAL-T and TL-T datasets, respectively. Notably, as shown in Table 1, when the baseline methods are equipped with masks predicted by our framework, their performance improves consistently. For instance, Oryon's AR on REAL-T rises from 32.2 to 35.5, and Obj. Mat.'s AR increases from 22.4 to 27.4.

In the setting using ground-truth masks, our method continues to outperform competitors on most metrics. On REAL-T, we achieve an AR of 52.9, outperforming the reconstruction-based method Any6D (51.0) and matching-based method Oryon (46.5). This indicates that beyond better segmentation, our proposed method establishes more reliable feature correspondences for pose estimation. The slightly inferior performance compared with Any6D on some metrics is attributed to the fact that Any6D adopts an explicit 3D shape reconstruction strategy, which enforces strong geometric consistency. However, such generated geometric structures often suffer from visual artifacts, causing alignment deviations on the image plane. In contrast, our method achieves superior performance on projection-aware metrics.

Qualitative results. Figure 3 shows the results on the REAL-T and TL-T datasets using predicted masks. Compared to Oryon, our framework demonstrates superior robustness. For example, Oryon generates a noisy mask for the "brown mug", failing to filter out semantically similar background distractors. In contrast, our method produces a clean mask and an accurate pose, validating the effectiveness of the proposed framework. Figure 4 presents the results using ground-truth masks. Any6D is sensitive to sparse view information and struggles with complex textures. In contrast, our method can effectively aggregate multi-view features under the guidance of calibrated textual semantics, thereby estimating more accurate 6D poses.

Robustness to segmentation errors. To evaluate the robustness to segmentation errors, we quantified the errors in the masks generated by our framework using the IoU metric relative to the ground truth. Specifically, we evaluate the AR using masks under varying IoU thresholds. As shown

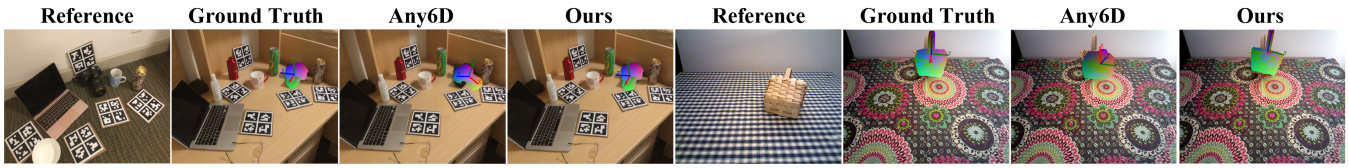

Figure 4: Qualitative comparison with state-of-the-art methods using ground-truth segmentation masks on the REAL-T and TL-T datasets. We visualize the estimated 6D poses by projecting the coordinate axes onto the query images.

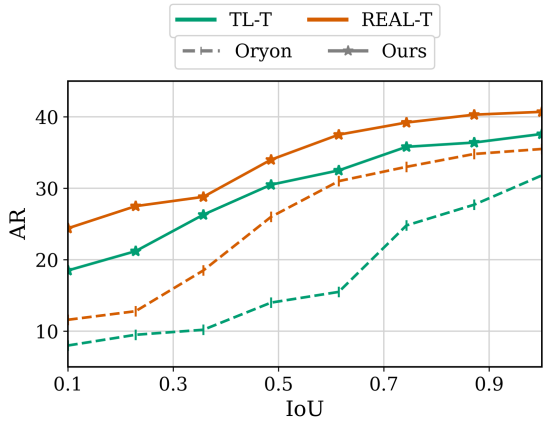

Figure 5: Robustness to segmentation errors. We evaluate the pose estimation performance (AR) under varying levels of segmentation quality (IoU) on the REAL-T and TL-T datasets.

Components			REAL-T		
TFM	FGSA	SACF	AR	ADD	mIoU
			30.1	23.8	64.6
		✓	34.4	29.2	68.9
✓		✓	<u>36.8</u>	33.7	71.7
	✓	✓	36.0	<u>35.1</u>	<u>74.0</u>
✓	✓	✓	**40.7**	**38.7**	**74.2**

Table 2: Ablation study of different components on REAL-T dataset with our predicted masks.

in Fig. 5, our method consistently outperforms Oryon across all IoU thresholds on both the REAL-T and TL-T datasets. Moreover, the AR scores of the proposed framework exhibit greater stability compared to those of Oryon.

4.4 Ablation Study

Effectiveness of individual components. Table 2 validates the contribution of each module. The baseline method adopts fixed text embeddings and performs fusion only via image feature modulation to predict segmentation and matching results, yet this configuration yields inferior performance across all metrics compared to the full implementation of SACF. Incorporating TFM boosts AR to 36.8 by transforming fixed global text priors into task-adaptive feature representations, thereby providing more flexible semantic guidance for subsequent feature fusion. Applying FGSA with SACF significantly boosts mIoU to 74.0 by effectively suppressing background distractions. The full model achieves the best performance with an AR of 40.7, validating the effectiveness of modulated textual semantics for 6D pose estimation.

λ_{sem}	REAL-T		TL-T	
	AR	ADD	AR	ADD
0	36.5	33.5	33.2	24.1
0.1	<u>38.2</u>	<u>35.1</u>	34.1	25.0
0.2	**40.7**	**38.7**	**37.6**	**28.4**
0.5	37.4	34.4	<u>34.8</u>	<u>25.5</u>
0.8	33.1	31.0	30.7	21.2

Table 3: Ablation study on the balancing weight λ_{sem} for semantic-aware matching on REAL-T and TL-T datasets.

Impact of Semantic-Aware Matching Weight. We verify the contribution of semantic consistency by varying the balancing weight λ_{sem} in Eq. (10). As shown in Table 3, introducing semantic guidance ($\lambda_{sem} > 0$) initially boosts performance, peaking at $\lambda_{sem} = 0.2$ with AR scores of 40.7 and 37.6 on REAL-T and TL-T, respectively. This confirms that calibrated textual semantics serve as a critical constraint to reject ambiguous visual matches. However, performance degrades significantly when λ_{sem} exceeds 0.2. This indicates that while semantic cues are essential for verification tasks, overemphasizing semantics degrades the accuracy of 6D pose estimation, implying that semantic consistency should act as a soft constraint rather than a dominant factor.

5 Conclusion

In this paper, we presented a framework for RGB-D open-vocabulary 6D pose estimation, guided by modulated textual semantics. To overcome the limitations of using static text embeddings for visual matching, we introduced the Text Feature Modulation (TFM) module, which transforms frozen descriptions into task-adaptive semantic representations. Then, we proposed the Foreground-Grounded Semantic Alignment (FGSA) strategy to explicitly ground these semantics in object regions while suppressing background distractions. Furthermore, the Semantic-Anchored Cross-View Fusion (SACF) module was designed to leverage these calibrated semantics as an invariant anchor, enabling robust feature aggregation across distinct views. Extensive experiments on standard benchmarks demonstrate that our method achieves superior performance in both segmentation quality and pose estimation accuracy. Future work will explore integrating image caption and depth estimation to extend our framework to an RGB-only setting, enabling broader applications and greater practicality.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant U24B20155 and Grant 62506175, in part by the Basic Research Program of Jiangsu under Grant BK20250666, and in part by the Jiangsu Province Higher Education Institution Basic Science (Natural Science) Research General Program under Grant 5KJB520038.

References

[Bai *et al.*, 2021] Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. Pointdsc: Robust point cloud registration using deep spatial consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15859–15869, 2021.

[Cai *et al.*, 2022] Dingding Cai, Janne Heikkilä, and Esa Rahtu. Ove6d: Object viewpoint encoding for depth-based 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6803–6813, 2022.

[Cheng *et al.*, 2021] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021.

[Cho *et al.*, 2024] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024.

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

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

[Cui *et al.*, 2022] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13608–13618, 2022.

[Deng *et al.*, 2025] Weijian Deng, Dylan Campbell, Chunyi Sun, Jiahao Zhang, Shubham Kanitkar, Matt E Shaffer, and Stephen Gould. Pos3r: 6d pose estimation for unseen objects made easy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16818–16828, 2025.

[Gümeli *et al.*, 2023] Can Gümeli, Angela Dai, and Matthias Nießner. Objectmatch: Robust registration using canonical object correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13082–13091, 2023.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[He *et al.*, 2022a] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *Advances in Neural Information Processing Systems*, 35:35103–35115, 2022.

[He *et al.*, 2022b] Yisheng He, Yao Wang, Haoqiang Fan, Jian Sun, and Qifeng Chen. Fs6d: Few-shot 6d pose estimation of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6814–6824, 2022.

[Hodan *et al.*, 2018] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.

[Huang *et al.*, 2025] Junwen Huang, Shishir Reddy Vutukur, Peter KT Yu, Nassir Navab, Slobodan Ilic, and Benjamin Busam. Raypose: Ray bundling diffusion for template views in unseen 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9102–9112, 2025.

[Ji *et al.*, 2025] Yuzhu Ji, Mingshan Sun, Jianyang Shi, Xiaoke Jiang, Yiqun Zhang, and Haijun Zhang. Seqpose: an end-to-end framework to unify single-frame and video-based rgb category-level pose estimation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 1224–1232, 2025.

[Labbé *et al.*, 2020] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *European Conference on Computer Vision*, pages 574–591. Springer, 2020.

[Labbé *et al.*, 2022] 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.

[Lee *et al.*, 2025] Taeyeop Lee, Bowen Wen, Minjun Kang, Gyuree Kang, In So Kweon, and Kuk-Jin Yoon. Any6d: Model-free 6d pose estimation of novel objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11633–11643, 2025.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Liu *et al.*, 2022] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *European Conference on Computer Vision*, pages 298–315. Springer, 2022.

[Liu *et al.*, 2023] Jian Liu, Wei Sun, Chongpei Liu, Xing Zhang, and Qiang Fu. Robotic continuous grasping system by shape transformer-guided multiobject category-level 6-d pose estimation. *IEEE Transactions on Industrial Informatics*, 19(11):11171–11181, 2023.

[Liu *et al.*, 2025] Mengya Liu, Siyuan Li, Ajad Chhatkuli, Prune Truong, Luc Van Gool, and Federico Tombari. One2any: One-reference 6d pose estimation for any object. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6457–6467, 2025.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Nguyen *et al.*, 2024a] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Yinlin Hu, Renaud Marlet, Mathieu Salzmann, and Vincent Lepetit. Nope: Novel object pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17923–17932, 2024.

[Nguyen *et al.*, 2024b] Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9903–9913, 2024.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Park *et al.*, 2020] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10710–10719, 2020.

[Peng *et al.*, 2019] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: 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.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[Sun *et al.*, 2021] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[Sun *et al.*, 2022] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022.

[Tang *et al.*, 2019] Fulin Tang, Yihong Wu, Xiaohui Hou, and Haibin Ling. 3d mapping and 6d pose computation for real time augmented reality on cylindrical objects. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2887–2899, 2019.

[Wang *et al.*, 2019] 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.

[Wen *et al.*, 2024] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. 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.

[Xiang *et al.*, 2017] 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.