IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 March 2024, accepted 2 May 2024, date of publication 6 May 2024, date of current version 15 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3397715

RESEARCH ARTICLE

Object Pose Estimation Using Color Images and Predicted Depth Maps

DINH-CUONG HOANG[1], PHAN XUAN TAN[2], (Member, IEEE), ANH-NHAT NGUYEN[3], DUY-QUANG VU[3], VAN-DUC VU[3], THU-UYEN NGUYEN[3], QUANG-TRI DUONG[1], VAN-THIEP NGUYEN[3], NGOC-ANH HOANG[3], KHANH-TOAN PHAN[3], DUC-THANH TRAN[3], NGOC-TRUNG HO[3], CONG-TRINH TRAN[3], VAN-HIEP DUONG[3], AND PHUC-QUAN NGO[1]

[1]Greenwich Vietnam, FPT University, Hanoi 10000, Vietnam
[2]College of Engineering, Shibaura Institute of Technology, Tokyo 135-8548, Japan
[3]IT Department, FPT University, Hanoi 10000, Vietnam

Corresponding author: Dinh-Cuong Hoang (cuonghd12@fe.edu.vn)

ABSTRACT The task of object pose estimation in computer vision heavily relies on both color (RGB) and depth (D) images to provide crucial appearance and geometric information, assisting algorithms in understanding occlusions and object geometry, thereby enhancing accuracy. However, the dependency on specialized sensors capable of capturing depth poses challenges in terms of cost and availability. Consequently, researchers are exploring methods to estimate object poses solely from RGB images. Nevertheless, this approach encounters difficulties in handling occlusions, discerning object geometry, and resolving ambiguities arising from similar color or texture patterns. This paper introduces a novel geometry-aware method for object pose estimation utilizing RGB images as input to determine the poses of multiple object instances. Our approach leverages both depth and color images during training but only relies on color images during inference. Departing from traditional depth sensors, our method computes predicted point clouds directly from estimated depth images derived from RGB inputs. A key innovation lies in the formulation of a multi-scale fusion module adept at seamlessly integrating features extracted from RGB images with those inferred from the predicted point clouds. This fusion process significantly fortifies the pose estimation pipeline by harnessing the strengths of both modalities, resulting in notably improved object poses. Extensive experimentation demonstrates that our approach markedly outperforms state-of-the-art RGB-based methods on Occluded-LINEMOD and YCB-Video datasets. Moreover, our method achieves competitive results compared to RGB-D approaches that necessitate both RGB and depth data from physical sensors.

INDEX TERMS Pose estimation, robot vision systems, intelligent systems, deep learning, supervised learning, machine vision.

I. INTRODUCTION

Object pose estimation, known as determining an object's precise position and orientation in 3D space from visual data, holds immense significance across various fields [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. This capability is vital for machines to interact intelligently with the physical world. In robotics, it facilitates tasks like object manipulation and assembly [11], [12], [13], [14]. Autonomous vehicles rely on accurate pose estimation to navigate safely and interact with their surroundings [5], [6], [11], [15], [16]. In augmented reality, it enhances user experience by realistically overlaying virtual objects onto the real world [7], [8], [17]. However, achieving accurate object poses faces challenges due to lighting variations, sensor noise, occlusion, and object truncation.

Traditional methods relied on manual feature engineering to establish image-object correspondences but struggled with

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

adapting to different lighting conditions and heavily occluded scenes [9], [18]. The emergence of machine learning and Deep Neural Networks (DNNs), coupled with RGB-D sensors [19], has transformed pose estimation. These advancements show promise in overcoming challenges, offering substantial improvements in accuracy and robustness [20], [21], [22]. Yet, RGBD-based methods often depend on specialized depth cameras, limiting their applicability where such sensors aren't available. In contrast, using color (RGB) images is more cost-effective and accessible. Recent studies leverage convolutional neural networks (CNNs) that operate solely on RGB images [23]. However, these RGB-based approaches face challenges due to the absence of geometry information, impacting performance in various scenarios like low-contrast scenes, textureless objects, lighting variations, sensor noise, and occlusions. Unlike RGBD-based methods that provide depth cues, RGB-based methods generate less discriminative features, affecting the quality of feature representations. Depth information is crucial for understanding object spatial relationships and scene geometry, significantly enhancing the discriminative power of features.

In this study, our aim is to bridge the gap between RGB-only and RGB-D approaches by utilizing the potential of monocular depth estimation techniques. Recent advancements in deep learning have showcased the potential of models to estimate depth from RGB images, offering approximate depth maps for each pixel. These depth maps contribute additional depth information, enriching the features derived from RGB images and facilitating the generation of geometry-informed features. This augmentation captures geometric cues, enhancing the discriminative capacity of representations and minimizing the disparity between RGB-only and RGB-D approaches. While utilizing estimated depth information can enhance performance, it can also be susceptible to inaccuracies inherent in the extracted depth maps. We introduce two multi-scale fusion modules designed for the integration of multi-modal features. The initial module, employing a transformer-based fusion approach, combines visual features with geometric counterparts, producing enhanced geometric features. Subsequently, the second module, utilizing an attention-based fusion technique, integrates geometric features into the visual domain, resulting in enriched visual features. These improved features from both branches are subsequently fused to generate geometry-aware features crucial for the final stage pose estimation. Our pose estimation module builds upon our prior work [13], employing deep Hough voting mechanisms, to ensure accurate and robust estimation. Through this approach, we aim to achieve a more cohesive integration of depth-derived insights with visual features, advancing the precision of object pose estimation in complex scenarios.

The main contributions of this work are as follows:

- We introduce a novel deep learning network tailored for monocular object pose estimation, aiming to leverage estimated depth maps as a replacement for measured depth images. These estimated depth maps serve as supplementary data to RGB images.
- Our proposal includes a transformer-based fusion module designed to generate context-aware geometric features.
- Moreover, we introduce an attention-based fusion module aimed at selectively integrating geometric features into visual representations.
- We meticulously design a voting-based 6D object pose estimation framework relying on enhanced features, resilient to noise and occlusion, adept at handling multiple objects within cluttered scenes.
- Through extensive experimentation, our proposed approach showcases superior performance compared to state-of-the-art monocular 6D object pose methods, validated across the YCB-Video [24] and Occlusion LineMOD [25] datasets.

The structure of this article is as follows. In Section II, we present related work, specifically addressing 6D object pose estimation in Section II-A, RGB-D Fusion in Section II-B and monocular depth estimation in Section II-C. Section III outlines our proposed methodology, breaking down the process into distinct components such as depth estimation to 3D point cloud (Section III-A) and feature extraction and fusion (Section III-B), which encompasses backbones, attention-based geometric-to-visual fusion, and transformer-based visual-to-geometric fusion. The voting-based object pose estimation is detailed in Section III-C. Moving on to Section IV, we cover the evaluation process, including datasets (Section IV-A), implementation and training details (Section IV-B), evaluation metrics (Section IV-C), and the results (Section IV-D). Ablation study is presented in Section IV-E, while Section IV-F explores the runtime efficiency of our model and Section IV-G provides a discussion. Finally, in Section V, we draw conclusions based on experimental results.

II. RELATED WORK

A. 6D OBJECT POSE ESTIMATION

Early methods in 6D pose estimation predominantly relied on handcrafted feature engineering and geometric matching techniques [26], [27], [28]. This classical approach involved extracting local features from input images and aligning them with model features, followed by the application of Perspective-n-Point (PnP) algorithms on resultant 3D-to-2D correspondences. This technique, while foundational, encountered challenges in robustness against transformations and varying viewpoints. Considerable research efforts have focused on refining local feature descriptors to achieve invariance to diverse transformations, aiming for more robust matching capabilities [29], [30], [31]. Simultaneously, advancements in PnP algorithms [32], [33], [34] have emerged, enhancing the robustness of pose estimation against noise and mismatches. These improvements have significantly bolstered the speed and reliability of

feature-based pose estimation, particularly for well-textured objects. However, a major drawback of these methods lies in the manual design of features by experts. This pre-defined set of characteristics might overlook certain scenarios or object properties, leading to suboptimal performance in new or unexpected environments. Moreover, the reliance on pre-processing or post-processing steps adds computational overhead, making these techniques unsuitable for real-time applications.

Recent advancements in 6D pose estimation have seen a surge in learning-based paradigms, particularly leveraging CNNs. These approaches manifest in two main trends: direct regression to 6D pose [24], [35], [36], [37], or predicting 2D keypoint locations for subsequent pose derivation via PnP. Direct methods aim to predict the 6DoF pose of objects through regression or classification directly from input images. PoseCNN [24], a pioneering CNN architecture, performs 6D object pose regression from a single RGB image by breaking down the object pose estimation task into distinct components. The method estimates the 3D translation of an object by localizing its center in the image and predicting the distance between the object and the camera. The estimation of 3D rotation is achieved through regression to a quaternion representation. However, direct estimation of 3D rotation encounters challenges due to the nonlinearity of the rotation space, potentially limiting the generalizability of CNNs. To overcome the complexity of directly estimating 3D rotation, several approaches by Kehl et al. [35] and Sundermeyer et al. [36] discretize the rotation space, transforming the estimation into a classification task. This discretization involves dividing the rotation space into discrete bins, simplifying the estimation process. However, due to this discretization, these methods often produce coarse results, necessitating a post-refinement step to enhance accuracy. In a recent advancement by Trabelsi et al. [37], an innovative approach integrates object classification, initial pose estimation, and iterative refinement into an end-to-end framework. This method utilizes both appearance features and flow vectors to enhance the accuracy of 6D pose estimation. Despite these advancements, achieving generalization in natural scene settings remains challenging due to the inherently ill-posed nature of the problem.

In contrast to direct methods, a prevalent strategy involves establishing 2D-3D correspondences and deducing pose-related parameters [38], [39], [40], [41]. Among these correspondence-based techniques, keypoint-based methods [38], [39] have demonstrated promising results in accurate 6DoF pose prediction without extensive post-processing. These methods typically employ deep networks to detect the 2D projections of 3D keypoints, followed by leveraging a Perspective-n-Point (PnP) solver for pose estimation [33]. PVNet [40] introduces a voting-based keypoint localization strategy adept at performing well even under occlusion or truncation. It utilizes a CNN to regress pixel-wise vectors representing keypoints, enabling voting for keypoint locations based on visible parts. This vector-field representation ensures robust recovery of occluded or truncated keypoints. Several works, such as those by Song et al. [41] and Yu et al. [42], have also adopted pixel-wise voting schemes to enhance keypoint localization performance. These keypoint-based methods establish correspondences independently, with post-detection consistency enforced through the PnP algorithm, which lies outside the deep network. In an endeavor to streamline this process, Hu et al. [43] implement the PnP step as a deep network component, enabling an end-to-end trainable 6D pose estimation framework. However, akin to two-stage keypoint-based methods, the accuracy of pose estimation is heavily reliant on the quality of keypoint detection. The quality of extracted features plays a pivotal role in keypoint detection performance. Robust and representative learned features yield accurate detection, whereas features lacking discriminative power or essential characteristics may compromise detection accuracy. Thus, enhancing the quality of extracted features remains critical for accurate keypoint detection and, consequently, reliable pose estimation results.

In recent advancements aimed at enhancing feature extraction, methods have increasingly integrated depth data as a complementary source alongside RGB images [20], [21], [44], [45]. By leveraging this fusion of RGB and depth information, these approaches have achieved state-of-the-art results, notably on standard datasets such as Occluded-LINEMOD [25] or YCB-Video [24]. The integration of depth data addresses crucial limitations present in RGB-only methods, offering substantial improvements in accuracy and robustness in 6D object pose estimation tasks. However, one inherent drawback of RGBD-based methodologies is their reliance on specialized depth sensors. This dependence restricts the applicability of these approaches in scenarios where dedicated depth cameras are unavailable or impractical. In this context, our work deviates from traditional RGB-D sensor reliance. Instead, we propose an innovative approach that utilizes monocular depth estimation frameworks to generate depth maps from RGB images.

B. RGB-D FUSION

RGB images typically contain rich semantic information, while depth images or point clouds offer precise geometric descriptions. Therefore, fusing these two modalities is a promising direction as they provide complementary information. With the advancement of inexpensive RGB-D sensors, numerous works have explored how to fully leverage this complementary information across various tasks such as detection [46], [47], [48], [49], segmentation [50], [51], and pose estimation [20], [52]. Fusion strategies are often categorized based on the flow of information. Undirected fusion [20], [21], [50] directly concatenates or adds features from both modalities. Unidirectional fusion [48], [49], [51] involves one modality guiding the fusion process of the other. However, these methods may

not fully exploit the interconnection between modalities. Bidirectional fusion [52], [53] has emerged as a recent approach, demonstrating the benefits of joint optimization across different modalities.

In RGB-D fusion-based pose estimation, traditional approaches [24], [35], [54] often adopt a coarse-to-fine strategy, initially computing poses from RGB images and refining them using depth maps as compensatory cues. Others [54], [55] treated depth information as an extra channel of RGB images or converted it into a bird-eye-view (BEV) image for input to CNN-based networks. However, these methods may be time-consuming due to expensive pose-processing steps and may not fully exploit spatial geometric structure information. In contrast, approaches like [20], [44], [45] used two separate branch networks to extract appearance and geometric features, followed by fusion for pose computation. While effective, these methods are sensitive to modality data loss. Recently, FFB6D [52] introduced an early fusion module to enhance communication between feature extraction branches, albeit with sensitivity to local noises. To address this, we propose two multi-scale fusion modules designed for integrating multi-modal features. The initial module employs a transformer-based fusion approach to enhance geometric features by combining them with visual features. Subsequently, the second module utilizes attention-based fusion to integrate geometric features into the visual domain, resulting in enriched visual features. These improved features from both branches are then fused to generate geometry-aware features crucial for the final stage of pose estimation.

C. MONOCULAR DEPTH ESTIMATION

Monocular depth estimation (MDE) refers to the process of inferring the depth information of a scene or an image using a single camera or a single image [56]. Depth estimation involves determining the distance of objects or points within the scene from the camera. The aim is to predict a dense depth map, where each pixel in a 2D image is associated with a corresponding depth value, indicating how far away the object or scene point is from the camera. Traditionally, depth estimation required specialized hardware or multiple images captured from different viewpoints (stereo vision) to triangulate depth information [57]. However, with the advancements in deep learning and computer vision, MDE has seen significant progress using neural network architectures that can learn to estimate depth directly from single images [58], [59], [60], [61], [62]. These neural networks are trained on large datasets where both RGB images and corresponding depth maps are available, allowing the network to learn the relationship between visual features and depth information.

Eigen et al. [58] pioneered deep learning-based MDE. Their approach utilized two CNNs to leverage global and local information, employing a scale-invariant error as a loss function to optimize training. Liu et al. [59] formulated depth estimation as a deep continuous conditional random fields (CRF) learning problem, employing superpixel-based segmentation and network modules to predict depth values. However, their reliance on fully connected (FC) layers resulted in computational overhead. Laina et al. [60] introduced the Fully Convolutional Residual Network (FCRN) for depth estimation. By modifying the encoder from ResNet-50 [61], removing FC layers, and guiding upscaling via a decoder, they significantly reduced the number of learnable parameters. Their experiments showed that increasing network depth improved accuracy by capturing more context information, inspiring the application of deeper CNNs like ResNet-101/152 [61], DenseNet-169 [63], or SENet-154 [62] to MDE. Mancini et al. [64] incorporated optical flow information to estimate depth, fusing current RGB images with generated optical flow maps in an encoder-decoder architecture. This approach aimed to enhance depth estimation from monocular images by leveraging additional information. Alhashim and Wonka [65] innovated a densely connected encoder-decoder architecture based on DenseNet-169 [63], employing a simplified decoder. Their approach is enriched by augmented training and deeper network design. Lee et al. [66] introduced a Local Planar Guidance (LPG) layer to facilitate precise mapping of features for depth estimation. Yin et al. [67] incorporated geometric constraints, enhancing depth map accuracy and 3D point cloud quality. Hu et al. [68] combined an encoder-decoder framework with Multi-Scale Feature Fusion (MSFF) and Refinement (RF) modules, introducing a hybrid loss function encompassing depth, gradients, and surface normals. Chen et al. [69] developed a Structure-Aware Residual Pyramid Network (SARPN), exploiting scene structures across multiple scales for accurate depth estimation. Fu et al. [70] and Bhat et al. [71] respectively addressed depth discretization issues using ordinal regression and adaptive bin division strategies, both contributing to more accurate depth maps.

Relative to depth sensors, the accuracy and precision of predicted depth maps from monocular estimation networks can sometimes fall short. This discrepancy often stems from their reliance on training data quality and the intricacies of neural network architectures. Consequently, to harness the potential of estimated depths as valuable additions to RGB images, a meticulously crafted fusion strategy becomes essential for effectively merging these disparate data sources.

III. METHODOLOGY

Given an RGB image, the objective is to predict a transformation matrix that converts the object's coordinate system to the camera coordinate system. We assume that the 3D model of the object is available and the object coordinate system is defined in the 3D space of the model. The 6D pose, denoted as ξ, is represented as a homogeneous transformation matrix $p \in SE(3)$, where R represents the rotation matrix belonging to $SO(3)$, and $\mathbf{t}$ represents the translation in $\mathbb{R}^3$, such that $\xi = [R|\mathbf{t}]$. Since the estimation is performed from camera images,

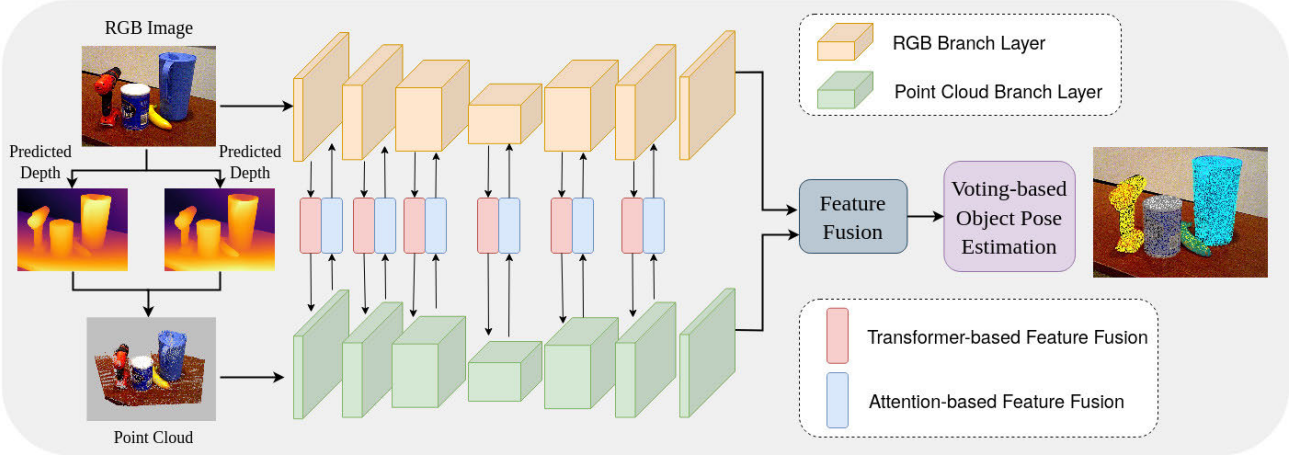


FIGURE 1. Overview of our network architecture. The diagram illustrates the key components of our network, including depth estimation, transformer-based adaptive fusion for context-aware geometric feature learning, attention-based adaptive fusion for geometry-aware visual feature learning, and voting-based object pose estimation.

the poses are defined relative to the camera coordinate frame. Our objective is to estimate the 6D pose of a set of known objects in an RGB image of a cluttered scene, where each pose is defined with respect to the camera coordinate frame. During the inference stage, only RGB images are available, while depth images are accessible during training.

A. OVERVIEW

The proposed approach takes an input RGB image $\mathcal{I}$ and outputs a transformation matrix $\xi = [R|\mathbf{t}]$ that transforms the object from its coordinate system to the camera coordinate system. Such transformation consists of a rotation matrix $R \in SO(3)$ and a translation matrix $\mathbf{t} \in \mathbb{R}^3$. Initially, off-the-shelf monocular depth estimation frameworks are employed to generate depth maps. Areas with uncertain depth information in these maps are identified and excluded, isolating reliable regions to form a 3D point cloud denoted as $\mathcal{P}$. Subsequently, $\mathcal{I}$ and $\mathcal{P}$ are inputted into two distinct branches of heterogeneous networks designed for feature extraction. Visual and geometric features are separately extracted using dedicated networks, and they are integrated fully and bidirectionally across all stages through adaptive fusion modules. These fused features are then utilized to estimate the pose of objects within the scene, building upon our prior work as described in [13].

Figure 1 presents an overview of our approach, comprised of integral components: depth estimation, transformer-based adaptive fusion for context-aware geometric feature learning, attention-based adaptive fusion for geometry-aware visual feature learning, and voting-based object pose estimation. Given an input RGB image $\mathcal{I}$, we first employ off-the-self monocular depth estimation frameworks to produce depth maps. These components synergistically enhance feature discriminability and robustness, culminating in improved accuracy and efficiency in object pose estimation. In the following sections, we delve into detailed explanations of each component's role in facilitating the success of our system.

B. DEPTH ESTIMATION TO 3D POINT CLOUD

Let $\mathcal{I}$ represent a single RGB image with dimensions $w \times h$, and $\mathcal{D}$ denotes the corresponding depth map of the same size as $\mathcal{I}$. The task of Monocular Depth Estimation (MDE) aims to establish a non-linear mapping $\Psi : \mathcal{I} \rightarrow \mathcal{D}$. Existing monocular depth estimation methods primarily cater to expansive outdoor scenes, posing challenges when applied to relatively smaller objects, such as those encountered in the Benchmark for 6D Object Pose Estimation (BOP) [72]. Furthermore, a notable limitation arises from the degraded quality around object edges and surfaces, directly influencing the precision of 3D perception, thereby impacting pose estimation. Predicted depth maps often exhibit blurriness around object boundaries due to inherent characteristics of 2D convolutions and bilinear upsampling. The convolutional kernel's aggregation across object boundaries often leads to undesired interpolations between foreground and background regions. Consequently, the associated 3D point clouds fail to accurately reflect the object's true 3D structure.

To overcome these constraints, our focus lies in enhancing depth map quality, specifically tailored for such objects. We leverage two distinct depth estimation networks: DPT [73] and iDisc [74], to derive individual depth images denoted as $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. By computing the disparity between these images, regions exhibiting substantial differences beyond a predefined threshold are identified as uncertain areas. Our methodology involves excluding these uncertain regions from the depth images and substituting the depth values within other areas with their mean values. This refinement process aims to enhance depth information, resulting in an improved and more precise depth image $\mathcal{D}$. The depth map $\mathcal{D}$ consists of depth pixels $d : \Omega \rightarrow \mathbb{R}$. We define the 3D back projection of a point

$\mathbf{u} \in \Omega$ given a depth map $\mathcal{D}$ as $\mathbf{p}(\mathbf{u}, \mathcal{D}) = \mathbf{K}^{-1}\tilde{u}d(\mathbf{u})$, where $\mathbf{K}$ represents the camera intrinsics matrix and $\tilde{u}$ denotes the homogeneous form of $\mathbf{u}$. The perspective projection of a 3D point $\mathbf{p} = [x, y, z]^\top$ is defined as $\mathbf{u} = \pi(\mathbf{Kp})$, where $\pi(\mathbf{p}) = [x/z, y/z]^\top$. Ultimately, this process yields a predicted point cloud $\mathcal{P}$.

C. FEATURE EXTRACTION AND FUSION

1) BACKBONES

Given the domain gap between RGB images and predicted 3D point clouds, our approach leverages two distinct backbone network branches to process these modalities independently. Illustrated in Figure 1, the visual branch handles RGB images, while the geometric branch manages point clouds derived from predicted depth images. To accomplish this, we employ ResNet34 [61] for extracting visual features $\mathbf{F}_v$ from RGB images and PointNet++ [75] to extract geometric features $\mathbf{F}_g$ from the resulting point cloud $\mathcal{P}$. These networks facilitate bidirectional information flow through multi-scale adaptive fusion modules. This approach enables both branches to harness local and global information interchangeably, enriching the representation learning process.

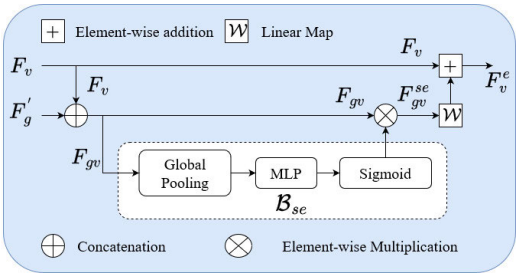


FIGURE 2. Architecture of attention-based geometric-to-visual fusion module $\mathcal{M}_{gv}$.

2) ATTENTION-BASED GEOMETRIC-TO-VISUAL FUSION

Depth images offer valuable geometric insights, yet they might contain noise or inaccuracies. Hence, when merging geometric features from the depth branch into the visual features of the RGB branch, it's crucial to design a fusion module that selectively integrates this information. The proposed attention-based geometric-to-visual fusion module (Figure 2), denoted as $\mathcal{M}_{gv}$, selectively incorporates geometric cues from $\mathbf{F}_g^i$ into the visual features $\mathbf{F}_v^i$ at each i-th stage.

Rather than employing a straightforward concatenation operation, this module emphasizes the channel relationships and leverages global information to adaptively highlight informative features while suppressing less useful ones. Initially, we identify the k_1 nearest 3D points for each pixel from the point cloud. In cases where fewer than k_1 points exist in the corresponding region, null features are padded. These collected geometric features are integrated using max pooling and processed through Multi-Layer Perceptrons (MLPs), resulting in $\mathbf{F}'_g$. Next, $\mathbf{F}_v^i$ and $\mathbf{F}'_g$ are concatenated to form $\mathbf{F}_{gv}$ and are subjected to a squeeze-and-excitation block $\mathcal{B}_{se}$ [62], also known as a channel attention block. This block employs global average pooling to condense each feature map within $\mathbf{F}_{gv}$ to a single pixel, generating a 1D vector of length C. This vector undergoes an MLP network with a hidden layer and sigmoid activation, followed by element-wise multiplication with $\mathbf{F}_{gv}$. This process recalibrates the feature responses, amplifying significant channels while suppressing less relevant ones. The output of $\mathcal{B}_{se}$, denoted as $\mathbf{F}_{gv}^{se}$, is described as follows:

$$\mathbf{F}_{gv} = \mathbf{F}'_g \oplus \mathbf{F}_v \tag{1}$$

$$\mathcal{B}_{se}(\mathbf{F}_{gv}) = \sigma(MLP(AvgPool(\mathbf{F}_{gv}))) \tag{2}$$

$$\mathbf{F}_{gv}^{se} = \mathcal{B}_{se}(\mathbf{F}_{gv}) \otimes \mathbf{F}_{gv} \tag{3}$$

Here, $\oplus$ denotes the concatenation operation and $\otimes$ represents element-wise multiplication. σ represents the sigmoid function. Subsequently, the fused features $\mathbf{F}_{gv}^{se}$ are treated as a residue and added to the original visual features $\mathbf{F}_v$ to produce enhanced features $\mathbf{F}_v^e$:

$$\mathbf{F}_v^e = \mathbf{F}_v + \mathcal{W}\mathbf{F}_{gv}^{se} \tag{4}$$

where $\mathcal{W}$ denotes a linear layer that maps $\mathbf{F}_{gv}^{se}$ to a fused feature of the same dimension as $\mathbf{F}_v$.

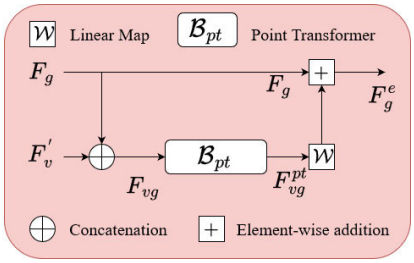


FIGURE 3. Architecture of transformer-based visual-to-geometric fusion module $\mathcal{M}_{vg}$.

3) TRANSFORMER-BASED VISUAL-TO-GEOMETRIC FUSION

In order to enhance features extracted in the point cloud branch, integrating visual representations $\mathbf{F}_v$ from the RGB branch into geometric features $\mathbf{F}_g$ can create context-aware geometric representations $\mathbf{F}_g^e$. However, unlike structured image data, point clouds lack a fixed grid arrangement. They exhibit variable point densities and irregular arrangements, lacking specific patterns in their data structures. This variability leads to regions with differing point densities; some areas might densely capture intricate details while others remain sparser. Given these irregularities, conventional attention-based fusion modules, commonly used in RGB data, may not suit point cloud processing. Inspired by Point Transformer [76], we propose a transformer-based fusion module $\mathcal{M}_{vg}$ (Figure 3). This module is tailored to efficiently integrate visual features into the point cloud branch, enabling the production of context-aware geometric features $\mathbf{F}_g^e$.

For each geometric feature paired with its respective 3D point coordinate, we retrieve visual features, denoted as $\mathbf{F}_v$, by projecting its local neighborhood (with a radius r) onto the image. Subsequently, we extract visual features by sampling the k_2 nearest neighboring pixels within this region, collecting their respective visual representations. In instances where fewer than k_2 pixels are present in the specified region, we pad null features. These gathered visual features undergo integration via max pooling and subsequent processing through Multi-Layer Perceptrons (MLPs) to align their channel size with the corresponding point cloud feature. This step yields modified visual features, denoted as $\mathbf{F}'_v$. Following this, we concatenate the integrated visual features $\mathbf{F}'_v$ with the geometric features $\mathbf{F}_g$ and apply a shared MLP to obtain the fused geometric feature $\mathbf{F}_{vg}$. Consequently, the network enriches a set of N 3D points with high-dimensional features, represented as $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{N}$ and $\mathbf{F}'_{vg} = \{\mathbf{f}_i\}_{i=1}^{N}$, where $\mathbf{p}_i = [\mathbf{x}_i; \mathbf{f}_i]$. Here, $\mathbf{x}_i \in \mathbb{R}^3$ signifies the point's spatial coordinates, while $\mathbf{f}_i$ represents the associated feature vector. These enriched points, denoted as $\{\mathbf{p}_i\}_{i=1}^{N}$ and now embedded with the fused features, serve as input to the Point Transformer block $\mathcal{B}_{pt}$ [76] to generate augmented features $\mathbf{F}_{vg}^{pt}$. This process enhances the representations of the 3D points through the integration of visual and geometric features. Subsequently, the fused features $\mathbf{F}_{vg}^{pt}$ are treated as a residue and added to the original visual features $\mathbf{F}_g$ to produce enhanced features $\mathbf{F}_g^e$:

$$\mathbf{F}_g^e = \mathbf{F}_g + \mathcal{W}\mathbf{F}_{vg}^{pt} \tag{5}$$

where $\mathcal{W}$ denotes a linear layer that maps $\mathbf{F}_{vg}^{pt}$ to a fused feature of the same dimension as $\mathbf{F}_g$.

4) FEATURE FUSION

After bidirectional fusion within both the $\mathcal{M}_{gv}$ and $\mathcal{M}_{vg}$ modules, distinct features are extracted from the visual and geometric branches. In order to enhance correspondence reliability and derive more distinctive features, a straightforward undirected fusion occurs in the final stage. This fusion involves projecting each point onto the image plane using the camera intrinsic matrix, establishing correspondences between visual and geometric features. These pairs are then concatenated to create the dense fused feature representation $\mathbf{F}$, which subsequently serves as input for the voting-based object pose estimation module in the following step.

D. VOTING-BASED OBJECT POSE ESTIMATION

Utilizing the fused feature map $\mathbf{F}$, our approach predicts object poses through a voting-based module, as previously detailed in our work [13]. We initialize M seeds as $\{\mathbf{s}_i\}_{i=1}^{M}$, where each seed $\mathbf{s}_i = [\mathbf{x}_i; \mathbf{f}_i]$ encompasses the 3D spatial location $\mathbf{x}_i \in \mathbb{R}^3$ and its corresponding feature vector $\mathbf{f}_i \in \mathbb{R}^C$. Subsequently, a voting module $\mathcal{M}_v$ generates votes from each seed independently. Specifically, this module employs a multi-layer perceptron (MLP) network with fully connected layers, ReLU activation, and batch normalization. The MLP takes the seed feature $\mathbf{f}_i$ and outputs the Euclidean space offset $\Delta\mathbf{x}_i \in \mathbb{R}^3$ and a feature offset $\Delta\mathbf{f}_i \in \mathbb{R}^C$. Thus, the vote $\mathbf{v}_i = [\mathbf{y}_i; \mathbf{g}_i]$ generated from the seed $\mathbf{s}_i$ has $\mathbf{y}_i = \mathbf{x}_i + \Delta\mathbf{x}_i$ and $\mathbf{g}_i = \mathbf{f}_i + \Delta\mathbf{f}_i$.

FIGURE 4. Example of generated votes projected on a 2D image. Green points indicate precise predictions closely aligned with ground-truth object centers, while red points represent predictions deviating further from the ground-truth.

Supervising the learning of the 3D offset $\Delta\mathbf{x}_i$ involves employing a regression loss function:

$$L_{vote} = \frac{1}{M_o}\sum_i \|\Delta\mathbf{x}_i - \Delta\mathbf{x}_i^*\|_H \cdot \mathbb{1}(\mathbf{s}_i) \tag{6}$$

Here, M_o represents the count of seeds on the object surface, $\|\cdot\|_H$ denotes the Huber norm, and $\mathbb{1}(\cdot)$ serves as a binary function indicating whether a seed point s_i belongs to an object. $\Delta\mathbf{x}_i^*$ is the ground truth displacement from the seed position $\mathbf{x}_i$ to the object center.

Votes, represented similarly to seeds in tensor form, are no longer anchored on object surfaces. Their significant difference lies in their positions. Votes generated from seeds on the same object are now closer to each other than the seeds, facilitating the combination of cues from different parts of the object. Figure 4 shows an example of generated votes projected on a 2D image. This semantic-aware locality is then leveraged to aggregate vote features for object proposals. After clustering the votes, we aggregate their features to generate object proposals and classify them. While numerous clustering methods exist, we adopt a straightforward strategy of uniform sampling and grouping based on spatial proximity. Specifically, from a set of votes $\{\mathbf{v}_i = [\mathbf{y}_i; \mathbf{g}_i] \in \mathbb{R}^{3+C}\}_{i=1}^{M}$, we sample a subset of J votes using farthest point sampling based on $\{\mathbf{y}_i\}$ in 3D Euclidean space. We then create J clusters by identifying neighboring votes. Although simple, this clustering technique integrates seamlessly into an end-to-end pipeline and performs effectively in practice. The votes within the J clusters are aggregated to produce J feature proposals $\{\mathbf{h}_i \in \mathbb{R}^C\}_{i=1}^{J}$ using a PointNet-like module, as described in [13]. At this stage, we obtain J proposals

comprising 3D positions $\mathbf{y}_i = \mathbf{x}_i + \Delta\mathbf{x}_i$ centered around object positions, proposal features $\mathbf{h}_i \in \mathbb{R}^C$ characterizing local geometry, and associated sets of seed points $\mathbf{s}_i$.

Pose Estimation With Multi-Task Loss: To jointly supervise the learning of modules, we utilize a multi-task loss:

$$L = \lambda_1 L_{vote} + \lambda_2 L_{pose} \tag{7}$$

where λ_1 and λ_2 denote the task weights. This loss comprises a voting loss L_{vote} and a pose loss L_{pose}, decomposed as:

$$L_{pose} = L_t + \alpha L_{rot} + \beta L_{obj} + \gamma L_{sem} \tag{8}$$

Here, α, β, and γ scale the losses proportionately. The pose loss includes a translation loss L_t (as regression), an L2 loss measuring the disparity between output and ground-truth rotation matrices, an objectness loss L_{obj}, and a semantic classification loss L_{sem}. The objectness loss employs cross-entropy for binary classification (object or not), while the semantic classification loss utilizes cross-entropy for semantic classes.

The pose loss comprises a translation loss along with an L2 loss between the predicted and ground truth rotation matrices. However, this rotation loss is primarily suitable for asymmetric objects. It fails to handle symmetric objects effectively due to the inherent ambiguity in their orientations. Symmetric objects can exhibit multiple correct 3D rotations, making the use of such a loss function inappropriate as it penalizes the network for regressing to any of these alternative rotations [20]. Consequently, employing this loss function on symmetric objects could result in inconsistent training signals. In essence, symmetric objects pose a challenge due to the existence of multiple, and sometimes infinite, canonical frames, leading to ambiguous learning objectives. Therefore, for symmetric objects, an alternative approach is necessary. Instead of directly minimizing the L2 distance between predicted and ground truth rotation matrices, a more suitable strategy involves minimizing the distance between each point on the estimated model orientation and the closest corresponding point on the ground truth model [20]. This approach addresses the ambiguity by focusing on point-to-point correspondence, enabling the network to learn effectively despite the symmetric nature of the objects. Therefore, we compute L_{rot} as:

$$L_{rot} = \frac{1}{m} \sum_{x_1 \in M} \parallel \min_{x_2 \in M} (\bar{R}x + \bar{T} - \hat{R}x + \hat{T}) \parallel \tag{9}$$

Here, M denotes the set of 3D model points, and m signifies the number of points. The loss is determined as the average distance from the vertices of the object model in the ground-truth pose to the closest vertices of the model in the estimated pose. This approach minimizes the loss when both 3D models are aligned.

IV. EVALUATION

In this section, we present a comprehensive evaluation of our proposed approach through a series of experiments conducted on 6D object pose estimation benchmark datasets. Our aim is to rigorously assess the efficacy and performance of the proposed framework in addressing the targeted challenges. The experiments are conducted using a high-performance workstation equipped with an NVIDIA RTX-3090 GPU and an Intel Xeon CPU 12 cores 2.1GHz. Through this empirical analysis, we provide compelling evidence of the capabilities and effectiveness of our proposed method.

A. DATASETS

1) YCB-VIDEO DATASET

Representing a comprehensive benchmark for 6D object pose estimation, the YCB-Video dataset [24] encompasses a diverse collection of 21 objects, varying in sizes and textures. It offers approximately 130,000 real images sourced from 92 video sequences, coupled with an additional 80,000 synthetically generated images primarily focusing on foreground objects. Each object within this dataset is meticulously annotated with accurate pose details and corresponding segmentation masks, facilitating precise evaluation of pose estimation methods. Notably, the test subset of the YCB-Video dataset exhibits a wide spectrum of challenges, including diverse illumination conditions, varying levels of noise, and significant occlusions. This assortment of challenges significantly intensifies the difficulty level for accurate pose estimation. We split the dataset into 75 videos for training, while the remaining 17 videos, constituting 4000 keyframes, are reserved for evaluation during testing.

2) OCCLUDED-LINEMOD DATASET

A subset derived from the LINEMOD dataset by Hinterstoisser et al. [77], the Occluded-LINEMOD dataset [25] distinguishes itself by annotating multiple objects within individual images, making it an ideal choice for assessing multi-object pose estimation methods. This dataset introduces several challenges beyond those in the LINEMOD dataset, including cluttered backgrounds, textureless objects, varying lighting conditions, and notably, severe occlusions between multiple object instances. These added complexities significantly elevate the task's difficulty in accurately estimating object poses. Despite its utility, the Occluded-LINEMOD dataset comprises only 1214 testing images and lacks explicit training data. Consequently, we adopt methodologies proposed in prior works [40], [43] to generate a synthetic training dataset consisting of 50,000 images, aiding in the training process.

B. IMPLEMENT AND TRAINING DETAILS

All images within the datasets are uniformly cropped and resized to 256×256 pixels. Our implementation utilizes a pre-trained ResNet34 model, trained on the ImageNet dataset, serving as the encoder for RGB images. This encoder is coupled with a PSPNet decoder [78], producing an output appearance feature comprising 128 channels. For point cloud feature extraction, we adopt a PointNet++-based feature

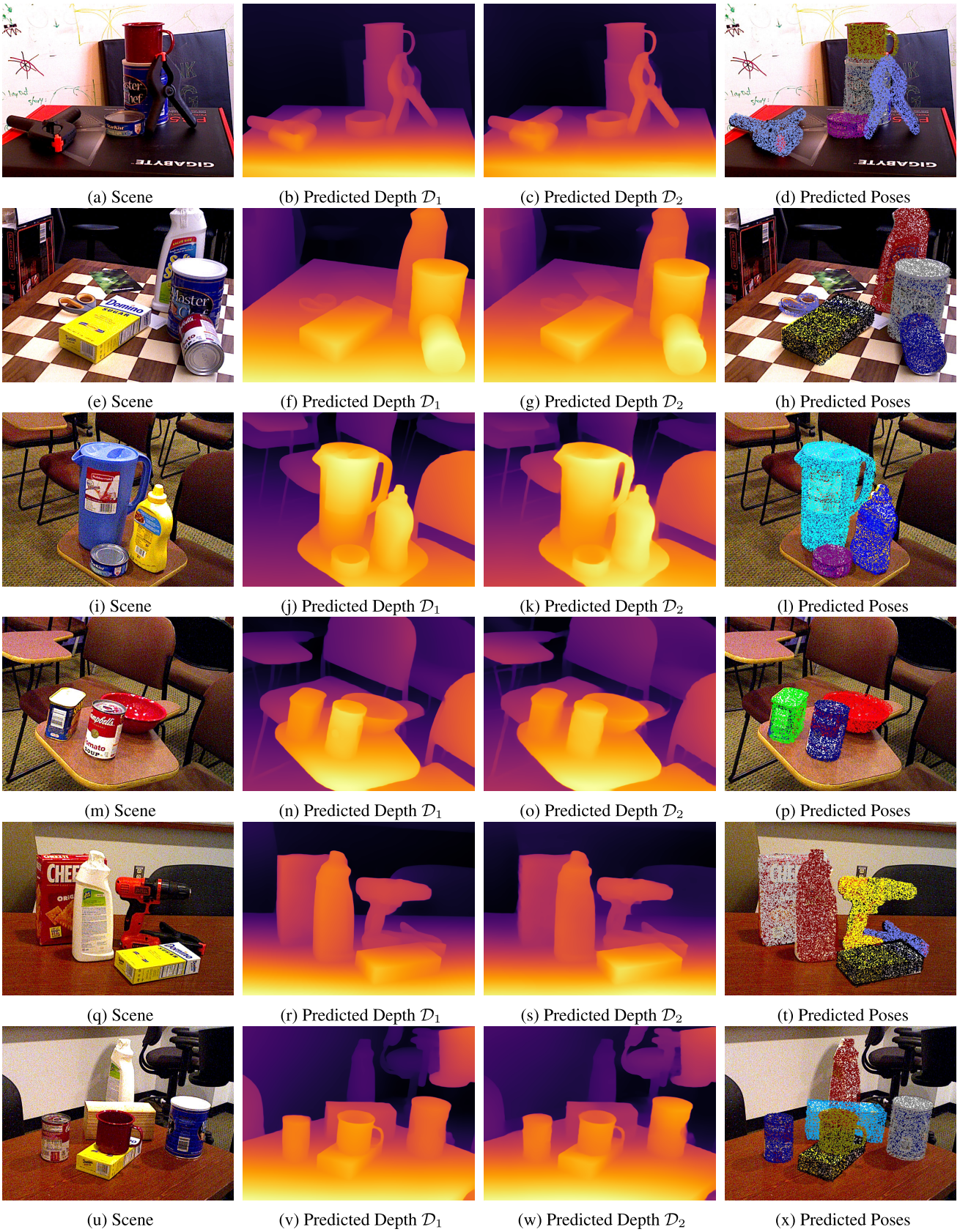


FIGURE 5. Qualitative results on YCB-Video dataset. The point cloud of each object, distinguished by various colors, undergoes a transformation based on the predicted pose and is subsequently projected onto the 2D image frame.

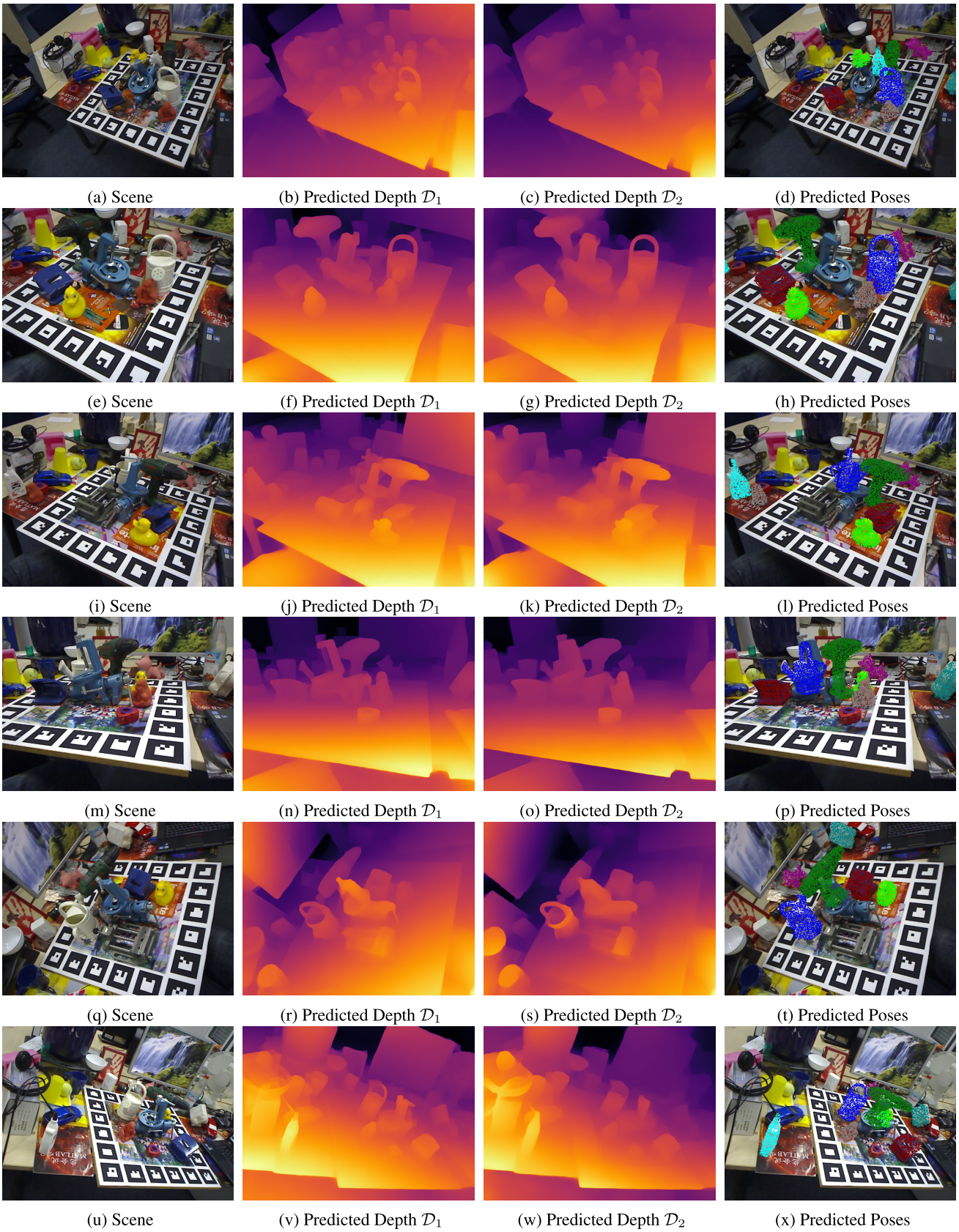


FIGURE 6. Qualitative results on Occluded-LINEMOD dataset. The point cloud of each object, distinguished by various colors, undergoes a transformation based on the predicted pose and is subsequently projected onto the 2D image frame.

learning network [75], randomly sampling 12288 points from depth images. This process yields a 128-channel output. In Eq. 7 and 8, we set $\lambda_1 = \lambda_2$ and $\alpha = \beta = \gamma$. Our implementations are realized using PyTorch and Python platforms, executed on an NVIDIA RTX-3090 GPU paired with an Intel Xeon CPU (12 cores, 2.1GHz), leveraging CUDA and the Linux operating system. The monocular depth estimation configurations align with those detailed in the original papers [73], [74].

For the YCB-Video dataset, we employ the Adam optimizer, initiating with a learning rate of 0.01. To facilitate learning, we schedule learning rate decay at epochs 100, 140, 180, reducing rates respectively to 0.1, 0.1, 0.1. The training utilizes a batch size of 8 and incorporates standard data augmentation techniques, spanning 220 epochs, taking approximately 16 hours to converge.

In the case of the Occluded-LINEMOD dataset, we employ the Adam optimizer [79], a batch size of 8, and common data augmentation techniques. We initiate training with a learning rate of 0.001, conducting training for 200 epochs. Learning rate decay occurs at epochs 100, 140, and 200, reducing rates to 0.1 at each step. The training process, until convergence, is completed in approximately 10 hours.

C. EVALUATION METRICS

In evaluating 6D object pose estimation, two widely adopted metrics are employed: ADD(-S) [24], [77] and 2D reprojection error (REP) [80].

The ADD(-S) Metric calculates the mean distance between two transformed model points: one using the estimated pose and the other using the ground-truth pose. A pose estimation is considered accurate if the distance is less than $\delta = 5\%$ of the model's diameter. For symmetric objects, the ADD-S variant [24] measures the mean distance based on the closest point distance. Moreover, to gauge performance comprehensively, the Area Under Curve (AUC) [24] of the ADD(-S) metric is computed by varying the distance threshold, capped at a maximum threshold of 5 cm.

The 2D Reprojection Error Metric proposed by Brachmann et al. [80] quantifies the mean distance between the projections of 3D model points utilizing the estimated pose against those using the ground-truth pose. An accurate pose estimation is attained when the distance is less than $\eta = 5$ pixels, effectively evaluating the precision of projecting 3D model points onto the 2D image plane. This metric serves as a vital indicator of accuracy in the alignment of 3D models with their corresponding 2D representations.

D. RESULTS

1) COMPARISON WITH RGB-BASED METHODS

For a fair and comprehensive comparison, we re-trained the baseline RGB-based pose estimation methods on the same datasets with settings as described in their original papers, utilizing the code provided by the respective authors. The evaluated methods include PoseCNN [24], PVNet [40], Single-Stage [43], and GDR-Net [81], alongside the recent RADet [82]. Each method represents a state-of-the-art approach in RGB-based object pose estimation, offering a benchmark for assessing the performance of our proposed method.

The evaluation results on the YCB-Video (Figure 5) and Occluded-LINEMOD (Figure 6) datasets, as depicted in Tables 1 and 2, underscore the exceptional performance of our pose estimation method. Employing the widely recognized ADD(-S) metric and 2D reprojection error (REP), our approach consistently outperforms state-of-the-art methods across a spectrum of object categories. On the YCB-Video dataset, our method showcases an average AUC of 84.8%, surpassing the AUC averages of PoseCNN (72.0%), PVNet (73.5%), Single-Stage (78.3%), and GDR-Net (82.5%). The superiority extends to 2D REP, with our method achieving an average score of 81.0, outperforming PoseCNN (33.0%), PVNet (49.0%), Single-Stage (67.2%), and GDR-Net (70.1%). The standout performance is evident in specific objects, such as "003_cracker_box," where our AUC of 89.8% and REP of 83.5% outshine all benchmarked methods.

Moving to the Occluded-LINEMOD dataset, crafted to assess robustness in the face of occlusion, our method demonstrates an average AUC of 81.2%, surpassing the averages of PoseCNN (57.3%), PVNet (67.4%), Single-Stage (71.7%), and GDR-Net (75.8%). The superiority is further corroborated by 2D REP, where our method attains an average score of 66.2, outperforming PoseCNN (17.6%), PVNet (62.3%), Single-Stage (63.6%), and GDR-Net (58.7%). Notably, for objects like "Can," our AUC of 79.4% and REP of 85.9% demonstrate the method's remarkable aptitude in accurately estimating poses amid challenging occlusion scenarios.

2) COMPARISON WITH DEPTH-BASED METHODS

The table demonstrates the performance of our proposed method compared to existing depth-based object pose estimation techniques on the YCB-Video and Occluded-LINEMOD (LMO) datasets. While other methods utilize real depth data, our approach relies on predicted depth maps derived from RGB images. Despite this reliance on predicted depth, our method achieves notable results. Specifically, in the YCB-Video dataset, our method achieves an AUC of 84.8 and a REP of 81.0, outperforming competing methods such as Drost et al. [83] (AUC: 74.1, REP: 79.8), Li et al. [84] (AUC: 77.8, REP: 68.4), Chen et al. [85] (AUC: 80.1, REP: 70.3), Gao et al. [86] (AUC: 74.4, REP: 76.3), and Zhuang et al. [87] (AUC: 76.3, REP: 77.2). Similarly, on the Occluded-LMO dataset, our method achieves an AUC of 81.2 and a REP of 66.2, surpassing the performance of other methods. The success of our method, even without the use of real depth data, can be attributed to several factors. Firstly, by integrating both RGB and predicted depth information, our method captures richer spatial relationships

TABLE 1. Quantitative comparison on YCB-Video dataset with state-of-the-art RGB methods.

Method	PoseCNN [24]		PVNet [40]		Single-Stage [43]		GDR-Net [81]		RADet [82]		Ours	
	AUC	REP	AUC	REP	AUC	REP	AUC	REP	AUC	REP	AUC	REP
002_master_chef_can	75.4	33.6	76.1	45.9	80.0	65.3	84.5	70.8	79.2	70.1	**83.5**	**83.5**
003_cracker_box	77.2	35.5	77.4	47.6	82.1	68.1	85.3	71.4	81.2	78.5	**89.8**	**83.5**
004_sugar_box	70.1	30.1	70.2	41.8	**79.2**	64.1	74.4	50.2	76.7	70.4	78.7	**80.5**
005_tomato_soup_can	76.1	34.2	77.0	46.1	81.2	66.7	84.9	71.3	79.9	73.6	**88.1**	**83.7**
006_mustard_bottle	78.5	47.2	79.9	55.3	84.2	75.1	86.7	76.1	84.6	80.3	**89.1**	**85.1**
007_tuna_fish_can	72.1	37.1	73.3	44.7	78.6	69.1	82.0	72.9	77.3	74.4	**84.8**	**81.6**
008_pudding_box	79.3	53.1	80.2	65.9	84.3	75.7	86.7	80.2	84.5	82.9	**90.5**	**85.0**
009_gelatin_box	78.8	41.2	78.3	57.8	77.0	75.1	85.0	74.9	70.1	64.6	**85.8**	**80.0**
010_potted_meat_can	75.1	34.2	77.3	46.4	81.2	66.2	**84.9**	71.2	78.2	73.0	84.4	**86.1**
011_banana	77.3	35.3	77.2	48.2	82.1	68.0	85.2	73.2	80.0	81.4	**86.1**	**83.1**
019_pitcher_base	75.2	43.2	73.0	45.4	73.0	65.2	82.5	70.4	78.5	73.2	**84.7**	**81.5**
021_bleach_cleanser	65.4	23.3	67.3	40.1	70.8	64.1	75.2	60.3	70.4	67.8	**79.1**	**73.5**
024_bowl	74.2	32.3	72.3	44.5	81.2	64.2	83.2	68.2	77.3	76.1	**81.2**	**80.1**
025_mug	78.4	42.1	77.2	46.2	81.2	64.2	87.1	72.3	79.0	73.3	**82.3**	**84.6**
035_power_drill	67.2	23.3	70.2	49.3	83.1	62.3	**85.3**	71.9	70.2	62.1	84.2	**73.0**
036_wood_block	73.1	36.2	79.0	58.4	75.2	62.6	83.8	72.1	71.2	60.2	**87.1**	**82.0**
037_scissors	66.1	13.1	56.0	43.6	70.2	61.2	74.3	62.4	53.1	19.1	**85.7**	**73.1**
040_large_marker	55.1	13.4	62.0	44.3	62.1	60.2	72.0	65.3	57.3	20.5	**85.1**	**79.2**
051_large_clamp	59.2	13.3	70.2	49.3	79.2	63.1	**83.0**	**71.3**	55.3	41.5	80.1	77.1
052_extra_large_clamp	62.1	18.1	70.2	42.3	75.1	**82.4**	**82.0**	72.2	78.8	73.2	82.0	80.1
061_foam_brick	76.2	53.2	79.0	65.2	82.3	68.1	84.1	73.4	80.3	76.1	**89.1**	**85.2**
Average	72.0	33.0	73.5	49.0	78.3	67.2	82.5	70.1	74.4	66.3	**84.8**	**81.0**

TABLE 2. Quantitative comparison on Occluded-LINEMOD dataset with state-of-the-art RGB methods.

Method	PoseCNN [24]		PVNet [40]		Single-Stage [43]		GDR-Net [81]		RADet [82]		Ours	
	AUC	REP	AUC	REP	AUC	REP	AUC	REP	AUC	REP	AUC	REP
Ape	50.6	34.6	65.4	69.1	69.6	70.3	**80.6**	71.0	75.3	70.1	80.5	**72.5**
Can	55.2	15.1	66.8	**86.1**	75.1	85.2	75.1	73.0	79.2	81.0	**79.4**	85.9
Cat	48.9	10.4	66.7	65.1	68.9	67.2	80.4	65.8	76.4	65.1	**87.4**	**72.1**
Driller	71.4	10.4	75.7	73.1	79.0	71.8	81.7	62.5	80.5	68.3	**87.8**	**75.6**
Duck	69.6	31.8	75.2	61.4	75.3	63.6	82.2	59.5	84.2	63.5	**88.1**	**67.1**
Eggbox	32.0	1.9	50.2	18.4	62.0	22.7	60.1	20.3	67.8	42.0	**69.2**	**26.4**
Glue	68.5	13.8	69.6	55.4	67.8	56.5	68.3	57.4	66.4	57.3	**73.3**	**59.1**
Holepun	62.1	23.1	69.7	69.8	75.6	**71.0**	77.6	60.1	80.2	70.1	**83.5**	70.6
Average	57.3	17.6	67.4	62.3	71.7	63.6	75.8	58.7	76.3	64.7	**81.2**	**66.2**

TABLE 3. Quantitative comparison on YCB-Video and Occluded-LINEMOD (LMO) datasets with depth-based object pose estimation methods.

Method	YCB-Video [24]		Occluded-LMO [77]	
	AUC	REP	AUC	REP
Drost et al. [83]	74.1	79.8	73.3	60.2
Li et al. [84]	77.8	68.4	72.5	59.1
Chen [85]	80.1	70.3	74.9	60.8
Gao et al. [86]	74.4	76.3	71.4	56.3
Zhuang et al. [87]	76.3	77.2	76.5	58.1
Ours	**84.8**	**81.0**	**81.2**	**66.2**

TABLE 4. Quantitative comparison on YCB-Video and Occluded-LINEMOD (LMO) datasets with state-of-the-art RGB-D methods. All methods use RGB images and predicted depth maps as input.

Method	YCB-Video [24]		Occluded-LMO [77]	
	AUC	REP	AUC	REP
DenseFusion [20]	78.9	74.2	78.6	64.5
PVN3D [21]	81.1	72.2	77.4	63.2
FFB6D [52]	82.6	75.4	79.8	64.6
Kumar et al. [22]	77.2	78.5	75.7	59.7
Wu et. al. [88]	79.1	80.4	79.3	62.5
Ours	**84.8**	**81.0**	**81.2**	**66.2**

TABLE 5. Quantitative comparison on YCB-Video and Occluded-LINEMOD (LMO) datasets with state-of-the-art RGB-D methods. All methods use RGB images and real depth maps as input.

Method	YCB-Video [24]		Occluded-LMO [77]	
	AUC	REP	AUC	REP
DenseFusion [20]	82.6	78.5	80.3	66.1
PVN3D [21]	84.2	75.8	79.0	65.1
FFB6D [52]	85.4	78.1	81.7	67.0
Kumar et al. [22]	83.1	78.5	77.3	60.1
Wu et. al. [88]	85.2	80.4	81.6	64.1
Ours	**85.9**	**82.3**	**82.8**	**67.6**

within the scene, enabling more precise pose estimations. While predicted depth maps may be less accurate than real depth data, they exhibit lower noise levels, which can be advantageous in certain scenarios. Additionally, our approach incorporates geometry information derived from predicted depth maps, which, when combined with our attention-based fusion network and voting module, enhances the accuracy of pose estimations.

3) COMPARISON WITH RGBD-BASED METHODS

The tables 4 and 5 present a quantitative comparison of the proposed method with several state-of-the-art RGB-D methods on the YCB-Video and Occluded-LINEMOD datasets. In Table 4, which utilizes RGB images and predicted depth maps, the proposed method outperforms existing approaches in terms of both AUC and REP on both datasets.

Specifically, on the YCB-Video dataset, our method achieves an AUC of 84.8 and a REP of 81.0, surpassing DenseFusion [20], PVN3D [21], FFB6D [52], Kumar et al. [22], and Wu et al. [88] Similarly, on the Occluded-LMO dataset, our method achieves the highest AUC of 81.2 and REP of 66.2, demonstrating superior performance compared to the other methods. In Table 4, where real depth maps are utilized along with RGB images, the proposed method continues to exhibit superior performance. On the YCB-Video dataset, our method achieves the highest AUC of 85.9 and REP of 82.3, surpassing all other methods, including DenseFusion, PVN3D, FFB6D, Kumar et al., and Wu et al. Similarly, on the Occluded-LMO dataset, our method achieves the highest AUC of 82.8 and REP of 67.6, once again outperforming the comparative methods.

Upon closer examination of the tables, it's evident that the proposed method achieves the best results with a significant margin when utilizing predicted depth maps, outperforming all other methods on both datasets. When real depth maps are used, while our method still achieves the best results, the improvement over other methods is relatively smaller compared to when predicted depth maps are employed. the superior performance of our method, especially when utilizing predicted depth maps, can be attributed to the innovative attention-based fusion module and voting mechanism integrated into our approach. The attention mechanism enables the model to dynamically focus on relevant regions within the input data, effectively capturing intricate spatial relationships and semantic details crucial for accurate pose estimation. By selectively attending to salient features, our model can effectively integrate information from both RGB images and predicted depth maps, thereby enhancing its ability to discern object poses with greater precision. Furthermore, the incorporation of voting mechanism facilitates robust aggregation of pose estimates from multiple viewpoints, leveraging the diversity of information available in the input data. This mechanism enables our model to mitigate ambiguities and uncertainties inherent in the estimation process by aggregating predictions across multiple spatial locations and orientations.

E. ABLATION STUDY

To gain insights into the efficacy of different components in our proposed framework, we conducted an ablation study by training and evaluating our model under various configurations. The components under investigation include enhanced depth information, $\mathcal{M}_{gv}$ (geometry-to-visual attention), $\mathcal{M}_{vg}$ (visual-to-geometry attention), or a combination of both $\mathcal{M}_{gv}$ and $\mathcal{M}_{vg}$. The ''Full'' entry represents our complete model with all components enabled. Table 6 presents the results of this ablation study on the YCB-Video [24] and the Occluded-LMO [77] datasets. The results underscore the critical role of each component in enhancing the performance of our model. When the model is trained without enhanced depth information, we observe a noticeable decline in performance, with an AUC of 81.1% and REP of 79.4% on the YCB-Video dataset, and an AUC of 78.2% and REP of 65.5% on the Occluded-LMO dataset. This highlights the importance of enhanced depth information in refining the model's accuracy, even though RGB images with raw predicted depth maps still contribute significantly, showcasing the significance of the fusion network. Furthermore, removing either $\mathcal{M}_{gv}$ or $\mathcal{M}_{vg}$ results in a substantial drop in accuracy. Without $\mathcal{M}_{gv}$, the AUC decreases to 75.3% (YCB-Video) and 70.2% (Occluded-LMO), and the REP drops to 70.7% and 56.2%, respectively. Similarly, without $\mathcal{M}_{vg}$, the AUC decreases to 74.6% (YCB-Video) and 71.3% (Occluded-LMO), and the REP drops to 69.9% and 58.3%, respectively. These results emphasize the complementary nature of the proposed attention mechanisms, each playing a crucial role in enhancing the model's accuracy. The combination of both $\mathcal{M}_{gv}$ and $\mathcal{M}_{vg}$ in the ''Full'' model provides the most comprehensive and effective framework. The model achieves an impressive AUC of 84.8% (YCB-Video) and 81.2% (Occluded-LMO), along with a REP of 81.0% and 66.2%, respectively. These results highlight the synergistic contributions of enhanced depth information and the proposed attention mechanisms, affirming their crucial role in elevating the accuracy of our RGB-based pose estimation method.

TABLE 6. Ablation study: Evaluating the impact of different components in the proposed framework by training and assessing our model under various configurations, including without (w/o) enhanced depth, M_{gv} (geometry-to-visual attention), M_{vg} (visual-to-geometry attention), or both M_{gv} and M_{vg}. The "w/o Enhanced Depth" configuration indicates the use of only the depth estimation framework DPT [73], excluding the combination of depths from DPT [73] and iDisc [74] as described in section III-B. The "Full" entry represents our model with all components enabled. Results are reported for the YCB-Video dataset [24] and the Occluded-LMO dataset [77].

Method	YCB-Video [24]		Occluded-LMO [77]	
	AUC	REP	AUC	REP
w/o Enhanced Depth	81.1	79.4	78.2	65.5
w/o $\mathcal{M}_{gv}$	75.3	70.2	70.7	56.2
w/o $\mathcal{M}_{vg}$	74.6	71.3	69.9	58.3
w/o ($\mathcal{M}_{gv} + \mathcal{M}_{vg}$)	65.2	32.9	60.3	29.5
Full	**84.8**	**81.0**	**81.2**	**66.2**

F. RUNTIME

Figure 7 illustrates the comparison of running speed (in Hz) and ADD(-S) on the YCB-Video and Occluded-LINEMOD datasets for various methods. Our proposed model operates at 30 Hz. This is particularly noteworthy when compared to the processing speeds of RGB-D methods, where DenseFusion operates at 10 Hz, PVN3D at 22 Hz, and FFB6D at 27 Hz. The trade-off between accuracy and processing speed is evident, with RGB-D methods offering higher precision at a slightly reduced speed. For other RGB methods, PoseCNN operates at 4 Hz, PVNet at 25 Hz, Single-Stage at 40 Hz, GDR-Net at 38 Hz, and RADet at 33 Hz.

G. DISCUSSION

While the above experimental results demonstrate the effectiveness of the proposed model for indoor scenarios,

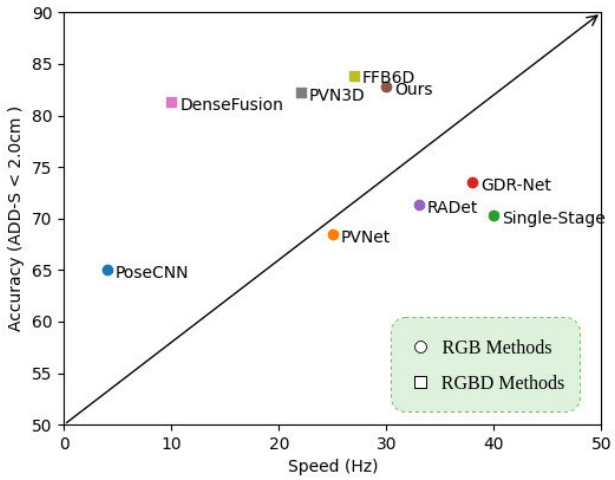


FIGURE 7. Comparison of running speed (Hz) and ADD(-S). The arrow indicates the direction where the method performs better, achieving higher accuracy in less inference time.

TABLE 7. Performance comparison of the proposed method on various datasets. No. Obj represents the number of object categories, and Avg. Size denotes the average size of objects in the dataset (in centimeters).

Dataset	No. Obj	Avg. Size	AUC	REP
YCB-Video [24]	21	10x10x6	84.8	81.0
Occluded-LMO [77]	8	8x10x5	81.2	66.2
Warehouse [11]	11	50x60x30	83.6	80.1
Fraunhofer IPA [89]	10	30x20x10	84.5	79.3

it is important to investigate its generalization capabilities, particularly when applied to outdoor, larger-scale environments. However, to the best of our knowledge, there is no outdoor dataset available specifically for the 6D object pose estimation task. Outdoor datasets often contain labels for tasks such as object detection or scene understanding, whereas datasets for 6D object pose estimation typically consist of full 3D models of objects captured in indoor environments, commonly associated with applications like robot manipulation, indoor navigation, and augmented reality within buildings. For example, the BOP dataset, one of the most popular benchmarks for 6D object pose estimation, includes 12 datasets, all of which are indoor environments. YCB-Video and Occluded-LMO, also part of the BOP Benchmark, are among the most widely used datasets in previous works, but they are also indoor datasets.

While we were unable to find an outdoor dataset to evaluate the generalization capabilities of our proposed model, we conducted additional experiments on the Warehouse dataset [11] and the Fraunhofer IPA Bin-Picking Dataset [89]. Although the data in these two datasets were not captured in open outdoor environments, they feature a wide range of diverse objects with diverse sets of materials. The Warehouse environment [11] presents complex challenges, including low illumination inside shelves, low-texture and symmetric objects, clutter, and occlusions. The dataset contains 11 objects and is focused on challenges in detecting warehouse object poses. The Fraunhofer IPA Bin-Picking Dataset [89] consists of 10 categories of objects and offers large-scale data designed for complex industrial scenarios and multiple parts for industrial grasping.

Table 7 presents the performance of our method on the four datasets, demonstrating its generalization capabilities across various scenarios. In future work, we plan to explore the adaptation of the proposed approach to other tasks in vast and open areas, including landscapes, streets, and natural surroundings. This will involve experiencing variations in natural lighting conditions, including changes in sunlight intensity, shadows, and weather-related factors, as well as accounting for natural elements, vehicles, people, and architectural structures.

V. CONCLUSION

In this work, we presented a novel RGB-based 6D object pose estimation method that leverages enhanced depth information and attention mechanisms to achieve remarkable accuracy and efficiency. Through extensive evaluations on the YCB-Video and Occluded-LINEMOD datasets, our proposed model consistently outperformed state-of-the-art RGB methods, showcasing its robustness and effectiveness. Notably, the ablation study highlighted the crucial contributions of enhanced depth information and the proposed attention mechanisms, affirming their roles in elevating the accuracy of our pose estimation approach. Moreover, in a comparative analysis with RGB-D methods, our model demonstrated competitive performance, underlining its potential utility in scenarios where depth sensors may be impractical or unavailable. The ablation study further emphasized the importance of each component in our framework, with enhanced depth information, geometry-to-visual attention ($\mathcal{M}gv$), and visual-to-geometry attention ($\mathcal{M}vg$) playing distinct yet complementary roles. In terms of runtime efficiency, our proposed model exhibited a processing speed of 30 Hz, surpassing several RGB methods and holding its ground against RGB-D counterparts. This trade-off between accuracy and processing speed positions our method as a versatile solution for real-world applications. The comprehensive evaluation, ablation study, and runtime analysis collectively demonstrate the effectiveness and versatility of our proposed RGB-based 6D object pose estimation method. As we continue to explore avenues for improvement and extension, this work contributes to advancing the field of computer vision and holds promise for applications in robotics, augmented reality, and beyond.

REFERENCES

[1] G. Du, K. Wang, S. Lian, and K. Zhao, "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: A review," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1677–1734, Mar. 2021.

[2] D.-C. Hoang, T. Stoyanov, and A. J. Lilienthal, "Object-RPE: Dense 3D reconstruction and pose estimation with convolutional neural networks for warehouse robots," in *Proc. Eur. Conf. Mobile Robots (ECMR)*, Sep. 2019, pp. 1–6.

[3] D.-C. Hoang, A. J. Lilienthal, and T. Stoyanov, "Panoptic 3D mapping and object pose estimation using adaptively weighted semantic information," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1962–1969, Apr. 2020.

[4] L. F. Rocha, M. Ferreira, V. Santos, and A. Paulo Moreira, "Object recognition and pose estimation for industrial applications: A cascade system," *Robot. Comput.-Integr. Manuf.*, vol. 30, no. 6, pp. 605–621, Dec. 2014.

[5] S. Hoque, S. Xu, A. Maiti, Y. Wei, and M. Y. Arafat, "Deep learning for 6D pose estimation of objects—A case study for autonomous driving," *Expert Syst. Appl.*, vol. 223, Aug. 2023, Art. no. 119838.

[6] S. Hoque, Md. Y. Arafat, S. Xu, A. Maiti, and Y. Wei, "A comprehensive review on 3D object detection and 6D pose estimation with deep learning," *IEEE Access*, vol. 9, pp. 143746–143770, 2021.

[7] N. Haouchine, P. Juvekar, M. Nercessian, W. M. Wells, A. Golby, and S. Frisken, "Pose estimation and non-rigid registration for augmented reality during neurosurgery," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 4, pp. 1310–1317, Apr. 2022.

[8] E. Marchand, H. Uchiyama, and F. Spindler, "Pose estimation for augmented reality: A hands-on survey," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 12, pp. 2633–2651, Dec. 2016.

[9] D.-C. Hoang, L.-C. Chen, and T.-H. Nguyen, "Sub-OBB based object recognition and localization algorithm using range images," *Meas. Sci. Technol.*, vol. 28, no. 2, Feb. 2017, Art. no. 025401.

[10] D.-C. Hoang, J. A. Stork, and T. Stoyanov, "Context-aware grasp generation in cluttered scenes," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 1492–1498.

[11] D.-C. Hoang, A. J. Lilienthal, and T. Stoyanov, "Object-RPE: Dense 3D reconstruction and pose estimation with convolutional neural networks," *Robot. Auto. Syst.*, vol. 133, Nov. 2020, Art. no. 103632.

[12] D.-C. Hoang, A.-N. Nguyen, V.-D. Vu, D.-Q. Vu, V.-T. Nguyen, T.-U. Nguyen, C.-T. Tran, K.-T. Phan, and N.-T. Ho, "Grasp configuration synthesis from 3D point clouds with attention mechanism," *J. Intell. Robotic Syst.*, vol. 109, no. 3, p. 71, Nov. 2023.

[13] D.-C. Hoang, J. A. Stork, and T. Stoyanov, "Voting and attention-based pose relation learning for object pose estimation from 3D point clouds," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 8980–8987, Oct. 2022.

[14] D.-C. Hoang, A.-N. Nguyen, V.-D. Vu, T.-U. Nguyen, D.-Q. Vu, P.-Q. Ngo, N.-A. Hoang, K.-T. Phan, D.-T. Tran, V.-T. Nguyen, Q.-T. Duong, N.-T. Ho, C.-T. Tran, V.-H. Duong, and A.-T. Mai, "Graspability-aware object pose estimation in cluttered scenes," *IEEE Robot. Autom. Lett.*, vol. 9, no. 4, pp. 3124–3130, Apr. 2024.

[15] A. Palleschi, M. Gugliotta, C. Gabellieri, D.-C. Hoang, T. Stoyanov, M. Garabini, and L. Pallottino, "Fully autonomous picking with a dual-arm platform for intralogistics," in *Proc. I-RIM Conf. I-RIM*, 2020, pp. 109–111.

[16] V.-D. Vu, D.-D. Hoang, P. Xuan Tan, V.-T. Nguyen, T.-U. Nguyen, N.-A. Hoang, K.-T. Phan, D.-T. Tran, D.-Q. Vu, P.-Q. Ngo, Q.-T. Duong, A.-N. Nguyen, and D.-C. Hoang, "Occlusion-robust pallet pose estimation for warehouse automation," *IEEE Access*, vol. 12, pp. 1927–1942, 2024.

[17] L.-C. Chen, D.-C. Hoang, H.-I. Lin, and T.-H. Nguyen, "Innovative methodology for multi-view point cloud registration in robotic 3D object scanning and reconstruction," *Appl. Sci.*, vol. 6, no. 5, p. 132, May 2016.

[18] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit, "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 858–865.

[19] M. G. Diaz, F. Tombari, P. Rodriguez-Gonzalvez, and D. Gonzalez-Aguilera, "Analysis and evaluation between the first and the second generation of RGB-D sensors," *IEEE Sensors J.*, vol. 15, no. 11, pp. 6507–6516, Nov. 2015.

[20] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "DenseFusion: 6D object pose estimation by iterative dense fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3338–3347.

[21] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun, "PVN3D: A deep point-wise 3D keypoints voting network for 6DoF pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11629–11638.

[22] A. Kumar, P. Shukla, V. Kushwaha, and G. C. Nandi, "Context-aware 6D pose estimation of known objects using RGB-D data," 2022, *arXiv:2212.05560*.

[23] Z. Fan, Y. Zhu, Y. He, Q. Sun, H. Liu, and J. He, "Deep learning on monocular object pose detection and tracking: A comprehensive overview," *ACM Comput. Surv.*, vol. 55, no. 4, pp. 1–40, Apr. 2023.

[24] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," 2017, *arXiv:1711.00199*.

[25] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6D object pose estimation using 3D object coordinates," in *Computer Vision—(ECCV)*. Cham, Switzerland: Springer, 2014, pp. 536–551.

[26] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, May 1999, pp. 1150–1157.

[27] A. Collet, M. Martinez, and S. S. Srinivasa, "The MOPED framework: Object recognition and pose estimation for manipulation," *Int. J. Robot. Res.*, vol. 30, no. 10, pp. 1284–1306, Sep. 2011.

[28] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce, "3D object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints," *Int. J. Comput. Vis.*, vol. 66, no. 3, pp. 231–259, Mar. 2006.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[30] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.

[31] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, and K. Daniilidis, "6-DoF object pose from semantic keypoints," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 2011–2018.

[32] Y. Zheng, Y. Kuang, S. Sugimoto, K. Åström, and M. Okutomi, "Revisiting the PnP problem: A fast, general and optimal solution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2344–2351.

[33] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *Int. J. Comput. Vis.*, vol. 81, pp. 155–166, Jul. 2009.

[34] S. Li, C. Xu, and M. Xie, "A robust O(n) solution to the Perspective-n-Point problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1444–1450, Jul. 2012.

[35] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1530–1538.

[36] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel, "Implicit 3D orientation learning for 6D object detection from RGB images," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 699–715.

[37] A. Trabelsi, M. Chaabane, N. Blanchard, and R. Beveridge, "A pose proposal and refinement network for better 6D object pose estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2381–2390.

[38] M. Rad and V. Lepetit, "BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3848–3856.

[39] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6D object pose prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 292–301.

[40] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "PVNet: Pixel-wise voting network for 6DoF pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4556–4565.

[41] C. Song, J. Song, and Q. Huang, "HybridPose: 6D object pose estimation under hybrid representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 428–437.

[42] X. Yu, Z. Zhuang, P. Koniusz, and H. Li, "6D of object pose estimation via differentiable proxy voting loss," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2020, pp. 1–17.

[43] Y. Hu, P. Fua, W. Wang, and M. Salzmann, "Single-stage 6D object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2927–2936.

[44] K. Wada, E. Sucar, S. James, D. Lenton, and A. J. Davison, "MoreFusion: Multi-object reasoning for 6D pose estimation from volumetric fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14528–14537.

[45] G. Zhou, Y. Yan, D. Wang, and Q. Chen, "A novel depth and color feature fusion framework for 6D object pose estimation," *IEEE Trans. Multimedia*, vol. 23, pp. 1630–1639, 2021.

[46] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4603–4611.

[47] C. Wang, C. Ma, M. Zhu, and X. Yang, "PointAugmenting: Cross-modal augmentation for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11789–11798.

[48] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zho, Q. V. Le, A. Yuille, and M. Tan, "DeepFusion: LiDAR-camera deep fusion for multi-modal 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1717–17161.

[49] Z. Wang and K. Jia, "Frustum ConvNet: Sliding frustums to aggregate local point-wise features for amodal 3D object detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 1742–1749.

[50] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Computer Vision—(ECCV)*. Cham, Switzerland: Springer, 2014, pp. 345–360.

[51] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3D graph neural networks for RGBD semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5209–5218.

[52] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "FFB6D: A full flow bidirectional fusion network for 6D pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3002–3012.

[53] W. Hu, H. Zhao, L. Jiang, J. Jia, and T.-T. Wong, "Bidirectional projection network for cross dimension scene understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14368–14377.

[54] C. Li, J. Bai, and G. D. Hager "A unified framework for multi-view multi-class object pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 254–269.

[55] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3D object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 641–656.

[56] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: A review," *Neurocomputing*, vol. 438, pp. 14–33, May 2021.

[57] H. Laga, L. V. Jospin, F. Boussaid, and M. Bennamoun, "A survey on deep learning techniques for stereo-based depth estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1738–1764, Apr. 2022.

[58] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–12.

[59] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.

[60] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 239–248.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[62] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[63] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[64] M. Mancini, G. Costante, P. Valigi, and T. A. Ciarfuglia, "Fast robust monocular depth estimation for obstacle detection with fully convolutional networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 4296–4303.

[65] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," 2018, *arXiv:1812.11941*.

[66] J. Han Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," 2019, *arXiv:1907.10326*.

[67] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5683–5692.

[68] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1043–1051.

[69] X. Chen, X. Chen, and Z.-J. Zha, "Structure-aware residual pyramid network for monocular depth estimation," 2019, *arXiv:1907.06023*.

[70] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.

[71] S. Farooq Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4008–4017.

[72] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. G. Buch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, C. Sahin, F. Manhardt, F. Tombari, T. K. Kim, J. Matas, and C. Rother, "BOP: Benchmark for 6D object pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 19–34.

[73] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12159–12168.

[74] L. Piccinelli, C. Sakaridis, and F. Yu, "IDisc: Internal discretization for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21477–21487.

[75] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–15.

[76] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16239–16248.

[77] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes," in *Computer Vision—(ACCV)*. Cham, Switzerland: Springer, 2012, pp. 548–562.

[78] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[79] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–18.

[80] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother, "Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3364–3372.

[81] C. Liu, W. Sun, K. Zhang, J. Liu, X. Zhang, and S. Fan, "Prior geometry guided direct regression network for monocular 6D object pose estimation," in *Proc. 41st Chin. Control Conf. (CCC)*, Jul. 2022, pp. 6241–6246.

[82] Y. Hai, R. Song, J. Li, M. Salzmann, and Y. Hu, "Rigidity-aware detection for 6D object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8927–8936.

[83] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3D object recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 998–1005.

[84] D. Li, H. Wang, N. Liu, X. Wang, and J. Xu, "3D object recognition and pose estimation from point cloud using stably observed point pair feature," *IEEE Access*, vol. 8, pp. 44335–44345, 2020.

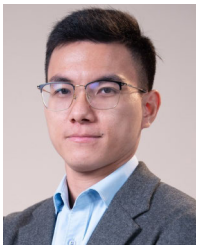
[85] W. Chen, X. Jia, H. J. Chang, J. Duan, and A. Leonardis, "G2L-Net: Global to local network for real-time 6D pose estimation with embedding vector features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4232–4241.

[86] G. Gao, M. Lauri, Y. Wang, X. Hu, J. Zhang, and S. Frintrop, "6D object pose regression via supervised learning on point clouds," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 3643–3649.

[87] C. Zhuang, S. Li, and H. Ding, "Instance segmentation based 6D pose estimation of industrial objects using point clouds for robotic bin-picking," *Robot. Comput.-Integr. Manuf.*, vol. 82, Aug. 2023, Art. no. 102541.

[88] C. Wu, L. Chen, S. Wang, H. Yang, and J. Jiang, "Geometric-aware dense matching network for 6D pose estimation of objects from RGB-D images," *Pattern Recognit.*, vol. 137, May 2023, Art. no. 109293.

[89] K. Kleeberger, C. Landgraf, and M. F. Huber, "Large-scale 6D object pose estimation dataset for industrial bin-picking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 2573–2578.



DINH-CUONG HOANG received the Ph.D. degree in computer science from Orebro University, Sweden, in 2021. He is currently a Lecturer with FPT University, Greenwich Vietnam. His research interests include the intersection of computer vision, robotics, machine learning, and autonomy for robots, with a focus on perception algorithms.

PHAN XUAN TAN (Member, IEEE) received the B.E. degree in electrical-electronic engineering from the Military Technical Academy, Vietnam, the M.E. degree in computer and communication engineering from Hanoi University of Science and Technology, Vietnam, and the Ph.D. degree in functional control systems from the Shibaura Institute of Technology, Japan. He is currently an Associate Professor with the Shibaura Institute of Technology. His current research interests include deep learning for visual computing, image and video processing, computational light field, 3D view synthesis, multimedia quality of experience, and multimedia networking.

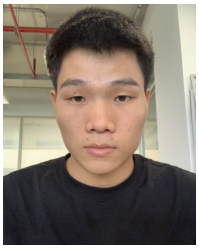
ANH-NHAT NGUYEN received the B.S. degree in computer science from Duy Tan University, Da Nang, Vietnam, in 2012, and the M.S. degree from Huazhong University of Science and Technology (HUST), China, in 2018. He is currently a Lecturer with FPT University, Vietnam. His research interests include image processing, information security, physical layer secrecy, radio-frequency energy harvesting, and wireless sensor networks.

DUY-QUANG VU is currently pursuing the B.S. degree in artificial intelligence with FPT University, Hanoi, Vietnam, with a primary focus on research in the field of computer vision.

VAN-DUC VU is currently pursuing the B.S. degree in artificial intelligence with FPT University, Hanoi, Vietnam, with a primary focus on research in the field of computer vision.

THU-UYEN NGUYEN is currently pursuing the B.S. degree in artificial intelligence with FPT University, Hanoi, Vietnam, with a primary focus on research in the field of computer vision.

QUANG-TRI DUONG is currently pursuing the B.S. degree in computing with FPT University, Greenwich Vietnam, Hanoi, Vietnam, with a primary focus on research in the field of computer vision and the Internet of Things (IoT).



VAN-THIEP NGUYEN is currently pursuing the B.S. degree in artificial intelligence with FPT University, Hanoi, Vietnam. His current interests include the manipulation of objects by robots and object recognition.

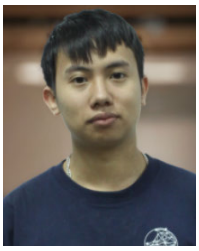
NGOC-ANH HOANG is currently pursuing the B.S. degree in artificial intelligence with FPT University, Hanoi, Vietnam, with a primary focus on research in the field of computer vision.

KHANH-TOAN PHAN is currently pursuing the B.S. degree in artificial intelligence with FPT University, Hanoi, Vietnam, with a primary focus on research in the field of computer vision.

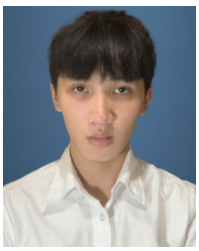
DUC-THANH TRAN is currently pursuing the B.S. degree in artificial intelligence with FPT University, Hanoi, Vietnam, with a primary focus on research in the field of computer vision.

NGOC-TRUNG HO is currently pursuing the B.S. degree in artificial intelligence with FPT University, Hanoi, Vietnam, with a primary focus on research in the field of computer vision and robotics.

CONG-TRINH TRAN is currently pursuing the B.S. degree in artificial intelligence with FPT University, Hanoi, Vietnam, with a primary focus on research in the field of computer vision.



VAN-HIEP DUONG is currently pursuing the B.S. degree in artificial intelligence with FPT University, Hanoi, Vietnam, with a primary focus on research in the field of computer vision and robotics.



PHUC-QUAN NGO is currently pursuing the B.S. degree in computing with FPT University, Greenwich Vietnam, Hanoi, Vietnam, with a primary focus on research in the field of computer vision and the Internet of Things (IoT).