2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 23-27, 2022, Kyoto, Japan

Towards Two-view 6D Object Pose Estimation: A Comparative Study on Fusion Strategy

Jun Wu[1], Lilu Liu[1], Yue Wang[1] and Rong Xiong[1] .

***Abstract*— Current RGB-based 6D object pose estimation methods have achieved noticeable performance on datasets and real world applications. However, predicting 6D pose from single 2D image features is susceptible to disturbance from changing of environment and textureless or resemblant object surfaces. Hence, RGB-based methods generally achieve less competitive results than RGBD-based methods, which deploy both image features and 3D structure features. To narrow down this performance gap, this paper proposes a framework for 6D object pose estimation that learns implicit 3D information from 2 RGB images. Combining the learned 3D information and 2D image features, we establish more stable correspondence between the scene and the object models. To seek for the methods best utilizing 3D information from RGB inputs, we conduct an investigation on three different approaches, including Early-Fusion, Mid-Fusion, and Late-Fusion. We ascertain the Mid-Fusion approach is the best approach to restore the most precise 3D keypoints useful for object pose estimation. The experiments show that our method outperforms state-of-the-art RGB-based methods, and achieves comparable results with RGBD-based methods.**

I. Introduction

6D object pose estimation aspires to estimate the rotation and translation of interested objects with regard to certain canonical coordinates. Accurate object pose estimation is the key to many real-world applications, such as robotic manipulation, augmented reality, and human-robot interactions. This is a challenging problem due to the variety of objects appearance, occlusions between objects, and clutter in the scene.

Based on the sensors they adopt, current model-based object pose estimation methods can be roughly categorized into two classes: RGB-based methods [1] [2] [3] and RGBD-based methods [4] [5] [6]. Though the scale is a known quantity in model-based object pose estimation problem, previous researches have shown that the performance of RGB-based methods are generally less competitive than the RGBD-based methods [7], mostly due to the lacking of 3D structure information. Since an image is the projection of the object under certain lighting condition and observation angle, predicting object pose with only RGB inputs could be affected by low resolutions, ill observation pose, changing of environment, and textureless or resemblant surface, thus involves more ambiguities. On the other hand, RGBD-based methods usually integrate 3D structures with 2D image features to estimate object poses [8] [9]. The factors impacting one kind of input modality could barely impact the other, thus this integration of two modalities of input features creates more robust prediction results.

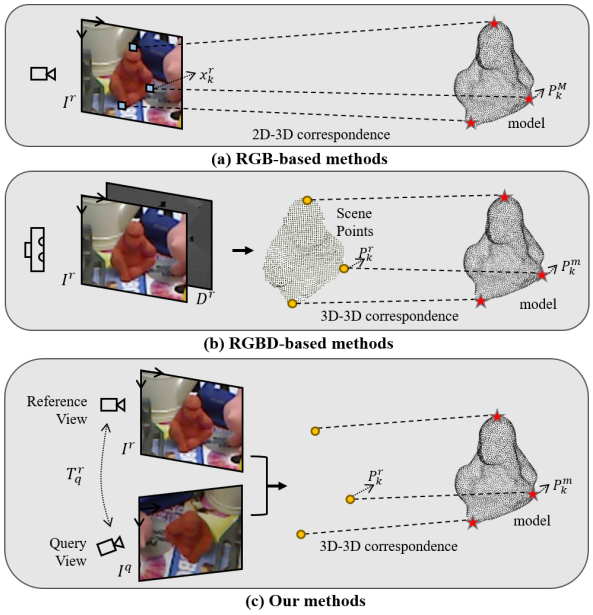

Fig. 1. **Illustration of our ideas.** We show the general pipelines of RGB-based, RGBD-based methods and our methods. RGB-based methods take RGB image as input, establish correspondence between image features and object models, then solve for the pose. While RGBD-based methods deploy image features and point cloud features to build the correspondence, which enhances the robustness of estimation. Our motivation is to narrow down the performance gap between RGB-based methods and RGBD-based methods, by proposing a framework combining image features from 2 input images to learn the implicit 3D information.

Therefore, a reasonable approach to enhance the pose estimation accuracy of RGB-based methods would be learning 3D geometric information from the RGB inputs. Though 2D images lack awareness of 3D geometric information in nature, two or more frames of such images combined together implicitly bring out the depth information, as has been verified in many stereo or multi-view stereo tasks [10] [11] [12]. Thus, we opt to learn the implicit 3D information from two or more RGB images to enhance the accuracy in estimating object poses. Luckily, making two or more observations of the scene with a certain relative transformation is commonly accessible for robots, rendering the applicability of this pipeline. Besides, when the clutter grows in the environment, objects are more likely to be occluded,

This work was supported in part by the National Key R&D Program of China under Grant 2021ZD0114500 and in in part by the Zhejiang Provincial Natural Science Foundation of China (LD22E050007).

[1]Jun Wu, Lilu Liu, Yue Wang, and Rong Xiong are with the State Key Laboratory of Industrial Control Technology and Institue of Cyber-Systems and Control, Zhejiang University, Zhejiang, China.

Corresponding author, wangyue@iipc.zju.edu.cn, Co-corresponding author, rxiong@zju.edu.cn

2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-7927-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IROS47612.2022.9981630

leading to worse estimation results, even for RGBD-based methods. With more observation perspectives, occlusion is easier to tackle.

In this paper, we investigate 3 different approaches to learn the 3D information from 2 input images, including fusing images to restore depth map (Early-Fusion), fusing 2D keypoints predicted from each image to restore 3D keypoints (Late-Fusion), and fusing 2D feature maps to restore characterized 3D space points (Mid-Fusion). We analyze all approaches on real word datasets and discover that the Mid-Fusion approach is able to restore the most precise 3D information useful for object pose estimation task. And by comparing to other state-of-the-art methods, we prove that combining 2D image features with learned 3D information effectively enhance the pose estimation accuracy.

To this end, the major contributions of this paper are as follows:

- To narrow down the performance gap between RGB-based methods and RGBD-based methods, we propose a framework for 6D object pose estimation, which learns the implicit 3D information from two RGB images to solve the pose.
- To seek for the method best utilizing 3D information useful for object pose estimation task, we conduct an investigation on 3 different approaches, including Early-Fusion, Mid-Fusion, and Late-Fusion, and show that the Mid-Fusion approach yields more precise 3D information.
- We compare our method to state-of-the-art methods on benchmarks. Our results show that integrating 3D information with 2D features effectively enhance the object pose estimation accuracy.

II. Related Works

RGB-based Object Pose Estimation. Many object pose estimation methods take RGB images as input, and tackle this problem by learning the correspondence between 2D image features and 3D CAD models [2] [13] [3] [14] [15]. [16] and [17] employs CNN to predict the 3D bounding box corners of the object in the image, and associate the corners with those in 3D CAD models to solve the pose. Since the corners are artificial, the estimation results are not satisfactory. To use more reliable correspondence, PVNet [18] selects keypoints from the object's model, and train a CNN to predict the vertex from every pixel to those keypoints then vote with confidence. Besides keypoints, HybridPose [19] employs edge vectors and symmetry correspondence to enrich the feature space for better estimation. Beyond establishing correspondence between 2D features and 3D models, some methods seek to directly learn the rationale between 2D features and 6D poses with neural networks [1] [20]. Furthermore, [21] and [22] explicitly learn the dense 2D-3D correspondence between images and canonical models, then regress poses by a Patch-PnP network.

RGBD-based Object Pose Estimation. To further improve the estimation accuracy, another pipeline, RGBD-based methods additionally deploy depth information [4] [5] [23] [24]. Early research [25] [26] compose contour vectors from RGB image and surface normal vectors from depth image, and estimate the pose by template matching. PVN3D [9] uses a neural network to separately extract image and point features, then encode them pointwisely to vote for 3D keypoints, and solve the pose with 3D-3D correspondence. REDE [6] and L6DNet [27] both encode multimodal features, and apply a 3D registration algorithm to solve the pose. Because of the extra depth information, RGBD-based methods generally achieve better results than RGB-based methods.

Multi-view Stereo Object Pose Estimation. Recently, some methods deploy multiframe observations to further refine pose estimation results in a optimization framework [28] [29] [30] [31] [7]. They formulate a global optimization function to refine the object poses predicted from each observation by minimizing the reprojection loss or average distance loss. This backend framework is generally adaptive to work with other frontend pose estimation methods, including our proposed one. [32] is the method most similar to ours, which proposes a pose estimation framework for transparent objects. They fuse the stereo RGB inputs to restore dense depth map, assigning depth value to the predicted 2D keypoints, then solve the 3D-3D registration problem. Our proposed Early-Fusion approach shares the same idea with them, and will be discussed later in the experiments.

Multi-view Stereo Scene Reconstruction. Restoring scene geometry from multiple overlapping images is a problem widely studied. Early Methods decouples the complex multi-view problem into relatively small problems with only one reference and a few source images at a time [33] [34]. Based on plane-sweep stereo, many recent learning-based methods [35] [36] [37] build cost volumes with warped 2D image features from multiple observations, regularize them with 3D CNNs and regress the depth. Cas-MVSNet [38] proposes a cascade cost volume mechanism based on feature pyramids and estimates the depth in a coarse-to-fine manner to reduce the computation resources. We follow them to build our Early-Fusion pipeline to restore the depth and evaluate its capability in restoring useful 3D keypoints for pose estimation.

III. Method

In this section, we introduce our overall pipeline to combine two images for object pose estimation, as illustrated in Fig. 2. Taking the input reference image I^r and query image I^q, we extract 2D image feature F^r and F^q with a shared feature extractor network. Then, 3 different approaches are developed to predict 3D keypoints P^r, including Early-Fusion, Mid-Fusion, and Late-Fusion. The Early-Fusion approach fuse the two input images to obtain depth map D^r, thus the predicted 2D keypoints x^r are projected to P^r. The Mid-Fusion approach fuse the two feature maps to restore a characterized 3D geometry volume V^r, then predicts P^r with 3D CNNs. And the Late-Fusion approach fuses x^r and x^q directly to get P^r. Notice that for a fair comparison, the 2D keypoints x^r are predicted by an additional head of the

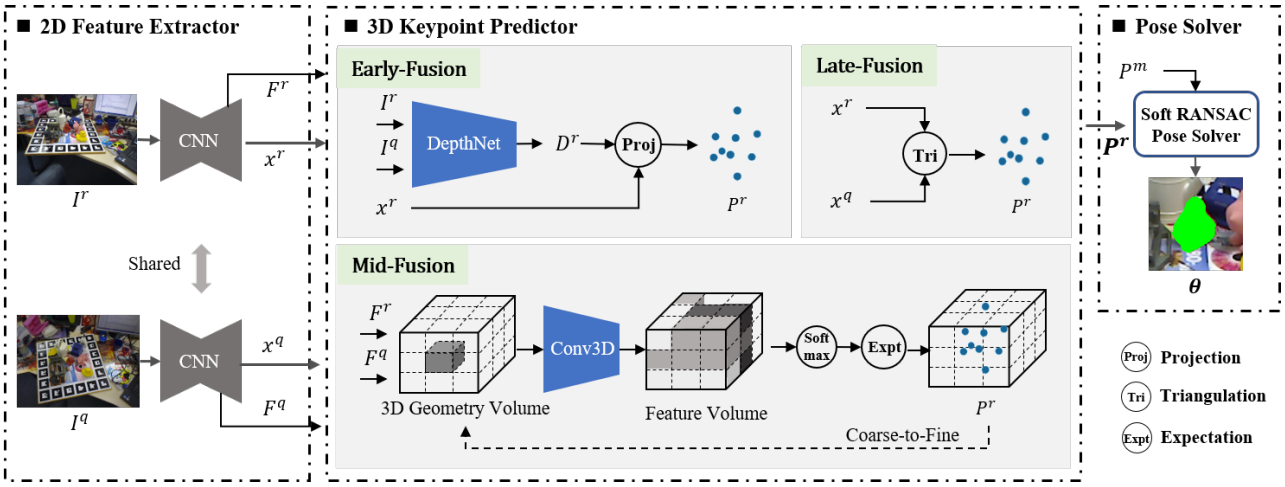

Fig. 2. **Methods Overview.** We show the overall pipeline of our proposed methods. Taking the input reference image I^r and query image I^q, we extract 2D image feature F^r and F^q with a shared multi-scale feature extractor network. Then, 3 different approaches are developed to predict 3D keypoints P^r, including Early-Fusion, Mid-Fusion, and Late-Fusion. The Early-Fusion approach fuses the two input images to predict depth map D^r via a DepthNet, thus the predicted 2D keypoints x^r are projected to P^r. The Late-Fusion approach fuses x^r and x^q by directly triangulating them to get P^r. The Mid-Fusion approach fuse the two feature maps to restore a characterized 3D geometry volume V_r, then predicts P^r with 3D CNNs. Notice that for a fair comparison, the 2D keypoints x^r are predicted by an additional head of the shared feature extractor. Last, P^r are used to solve the pose in a closed form with model keypoints P^m.

shared feature extractor. Last, P^r are used to solve the pose in a closed form with model keypoints P^m.

A. 2D Feature Extraction

For a fair comparison, we adopt a shared feature extraction network for all our approaches. Since feature extraction is not the focus of this paper, we follow [18] to build a multi-scale convolutional neural network, to extract features at multiple resolutions, and predicts 2D keypoints $\{x^r, x^q\}$ for Early-Fusion pipeline and Late-Fusion pipeline.

B. 3D Keypoints Prediction

With the goal of learning useful 3D information from two input RGB images, we explore 3 different approaches to predict the 3D keypoints, including Early-Fusion, Mid-Fusion, and Late-Fusion. Here we present how we develop these approaches separately.

Early-Fusion Approach: The most intuitive way to learn 3D information from input images is to directly recover depth from them. Without loss of generality, we follow [38] to build a cascade of 3D geometry volumes from warped 2D image features, regularize them with independent 3D CNN blocks and regress the depth map. Then, for a fair comparison with the other two approaches, we predict 2D keypoints from the input reference image following [18]. Last, the 3D keypoints are obtained by assigning depth values to the predicted 2D keypoints.

Late-Fusion Approach: Another nature approach to learn 3D information is to fuse the predicted 2D keypoints from both inputs into 3D keypoints. Likewise, we follow [18] to predict 2D keypoint for equitable contrast. Then, we triangulate these 2D keypoints to 3D with their relative camera poses by OpenCV library algorithm.

Mid-Fusion Approach: Despite the two straightforward approaches, we take a deeper look at the essence of keypoint prediction process. Predicting keypoints is to seek the location where the features best match the target keypoint previously chosen in the object models. Hence, the distinctiveness of the features is significant. Considering this, the Early-Fusion and the Late-Fusion approaches both rely solely on the 2D image features to predict, and the 3D information is only used to project the 2D keypoints to 3D. Therefore, we consider this Mid-Fusion approach to combine 2D image features and 3D information together to predict the keypoints.

As shown in Fig. 2, we build a 3D geometry volume from the two image feature maps, regularize it with 3D CNNs, then reduce the divergence between the feature field and the local keypoint heatmaps to predict 3D keypoints.

To build the 3D geometry volume, we divide the interested 3D space into regular 3D grids of size (H_v, W_v, D_v), and the size of each grid is (h_v, w_v, d_v). The axes of the grid coordinates are centralized in an initial guess and parallel to the reference camera coordinates. We create a many-to-one projection from 3D grid space to the image space by known camera intrinsic parameters. The grid space is also projected to the query image space by the relative camera pose between the two views. Then we assign the features extracted before of the 2D point to its related 3D point. Points projected outside the image are assigned to initial feature values. At each 3D point, we simply concatenate the two feature vectors from two inputs to keep as much information as possible.

Then, to predict 3D keypoints, we regularize this 3D volume with 3D CNNs to obtain a field representing the distribution of target keypoint locations. This distribution is learned by reducing the divergence between the 3D feature field and the local target keypoint heatmaps. The local heatmap of the ith keypoint is defined as

$$Q_i^k(p|P_i^m, \theta) \sim \mathcal{N}(\theta \cdot P_i^m, \sigma_i) \tag{1}$$

where $\{P_i^m\}_{i=1}^N$ refers to the N model keypoints, θ is

the target pose, and $\{\sigma_i\}_{i=1}^N$ are the hyperparameters. This distribution represents the local heatmap we expect to be highlighted as target keypoints locations. And the feature field of the ith keypoint is defined as

$$Q_i^v(P|V(\cdot)) = \sum_{j\in\Omega} w_j V([p_j]) \tag{2}$$

where $\{w_j\}_{j\in\Omega}$ are the trilinear interpolation coefficients, $[p_j]$ are the 8 neighbour grids of the conditioned position p, and $V(\cdot)$ is the value operation in the feature field. This distribution describes how likely a position in the field is to be the keypoint. To minimize the divergence of these two distributions, we adopt a Kullback-Leibler divergence loss

$$Loss_{KL} = D_{KL}(Q^v \| Q^k) = -\sum_{i=1}^{N} Q_i^v log(\frac{Q_i^k}{Q_i^v}) \tag{3}$$

Last, we predict the 3D keypoint by calculating the expectation of the distribution in the 3D volume. Although this is a super-resolution operation, the quality of the distribution still depends on the volume resolution. Hence, we employ a simple coarse-to-fine strategy to further enhance the prediction accuracy.

C. Soft RANSAC Solver

Given the predicted 3D keypoints $\{\hat{P}_i^r\}_{i=1}^N$ and the model keypoints $\{P_i^m\}_{i=1}^N$, we then solve the object pose by minimizing the distance between predicted points and model points. The optimization problem is

$$\hat{\theta} = \arg\min_{\theta} \|\theta \cdot P_i^m - \hat{P}_i^r\|_2 \tag{4}$$

where θ represents the 6D object pose.

Though this optimization problem can be solved by SVD in closed form, it is easy to be disturbed by outliers introduced by sensor noises or incorrect predictions. Therefore, we propose a soft RANSAC solver to softly count the inliers to solve a more robust pose. Taking the predicted keypoints $\{\hat{P}_i^r\}_{i=1}^N$, we calculate all possible poses with every 3 points by SVD, which brings us a set of pose hypothesis $\{\theta_k\}_{k=1}^{K=C_N^3}$. For each hypothesis, we evaluate the distance for every predicted and model keypoints under the hypothesis pose

$$d_{k,i} = \|\theta_k \cdot P_i^m - \hat{P}_i^r\|_2 \tag{5}$$

Then we sum up the soft inlier numbers for each hypothesis to evaluate the pose, and softly aggregate all the hypothesis into a final pose with a regularized scores

$$\hat{\theta} = \sum_{k}^{K} \theta_k \cdot \sum_{i}^{N} sigmoid(\gamma_1(-d_{k,i} + \gamma_2)) \tag{6}$$

This solver is shared among all the 3 proposed approaches.

For the differentiability in the entire process in Mid-Fusion approach, the pose error could be end-to-end back-propagated to train the 3D networks. The loss to evaluate the predicted pose is

$$Loss_{pose} = \|\hat{t} - t\|_2 + \alpha\|\hat{R}R^T - I\|_F \tag{7}$$

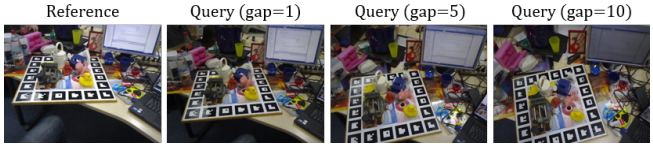

Fig. 3. **Example of query images under different pairing gaps.** The reference image and query image are paired up by a fixed gap, in which $gap = 1$ means pairing two images with a gap of one frame, and so on. Generally larger gaps lead to farther views.

where R is the rotation matrix and t is the translation vector. In total, the Mid-Fusion approach is trained with joint losses

$$Loss = \sum_{j} \beta_1 Loss_{pose_j} + \beta_2 Loss_{kpt_j} + \beta_3 Loss_{KL_j} \tag{8}$$

where $j = \{0, 1\}$ refers to the coarse and fine levels, and $\{\beta_1, \beta_2, \beta_3\}$ are hyperparameters.

IV. Experiments

In this section, we evaluate our method by experiments to answer two questions: (1) Which of the proposed fusing approaches can recover the most precise and useful 3D information from 2 RGB images for object pose estimation? (2) Combined with 2D image features and 3D information, can our method enhance the pose estimation accuracy? To answer (1), we assess and compare the precision of the predicted 3D keypoints of the proposed approaches. To answer (2), we compare the estimation results between our method and other state-of-the-art methods on public benchmark LineMOD dataset [26] and Occlusion LineMOD dataset [39].

TABLE I
EVALUATION ON 3D KEYPOINTS PREDICTION.
$(3D\text{-}Keypoint\text{-}Distance/m)$

	Early-Fusion	Mid-Fusion	Late-Fusion	REDE [6]
$gap = 1$	0.044	**0.024**	0.079	
$gap = 5$	0.031	**0.009**	0.014	
$gap = 10$	0.031	**0.007**	0.011	
mean	0.035	**0.013**	0.035	0.024

A. Dataset

LineMOD dataset [26] is a widely used benchmark for object 6D pose estimation tasks. It contains 13 objects with varieties of textures and structures. For a fair comparison, we follow [18] [6] to split it into train, valid and test data. For each object, about 180 images are for training and more than 1000 images are for testing. Also, we follow [18] to further create 10000 images by "Cut and Paste" strategy to train all the objects.

Occlusion LineMOD [39] dataset shares the same objects with LineMOD, but additionally annotates a subset of images with serious occlusion and clutters. It contains 8 different objects, and 1214 images. All images are used for testing with models trained on LineMOD dataset.

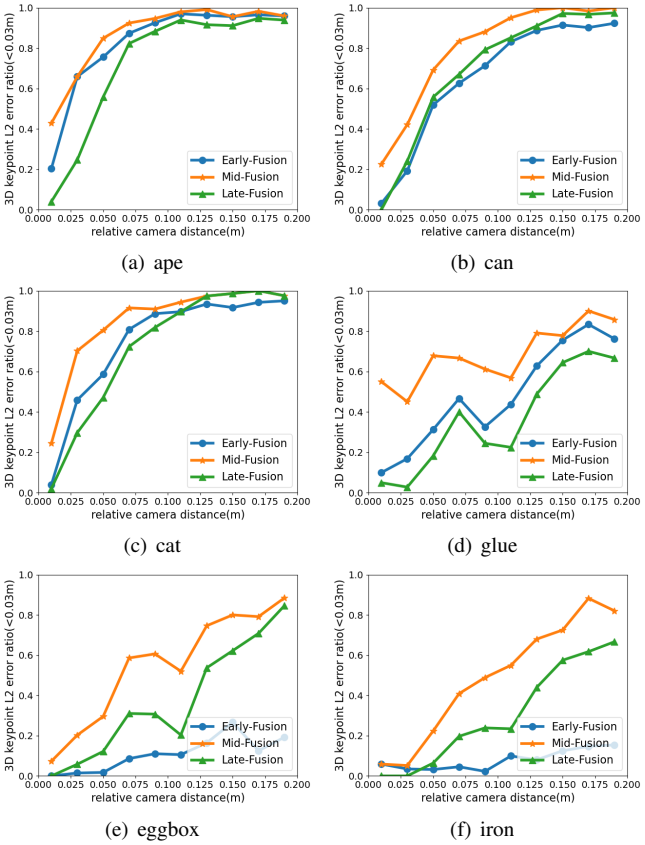

Fig. 4. **3D Keypoint Recall ($< 0.03m$) with respect to relative camera distances.** Each point represents a ratio of the numbers of keypoints with an error less than $0.03m$ to the number of all keypoints under certain relative camera distances. The relative camera distances are computed as the distances between two focal points between the two cameras. The Mid-Fusion approach achieves better results in most of the camera distances, especially in small parallax.

B. Metrics

We follow convention to use ADD [25] and ADD-S [1] as evaluation metrics. Given an object model of M points, ADD metric evaluates the average distance between the model points transformed with predicted and ground truth pose respectively

$$ADD = \frac{1}{M}\sum_{i=1}^{M}\|\theta p_i - \hat{\theta} p_i\|_2 \tag{9}$$

While ADD-S metric calculates the average distance between the closest points, which is used for evaluating symmetric object

$$ADD\text{-}S = \frac{1}{M}\sum_{i=1}^{M}\min_{j\in M}\|\theta p_j - \hat{\theta} p_i\|_2 \tag{10}$$

An estimation is regarded successful if the ADD(-S) is less than 10% of the object's diameter.

Besides, to evaluate the quality of restored 3D information, we apply 3D keypoint Euclidean distance between predicted 3D keypoints and ground truth 3D keypoints.

C. Implementation Details

In implementation, we follow [18] to select 9 keypoints for every object, and perform the same data augmentation.

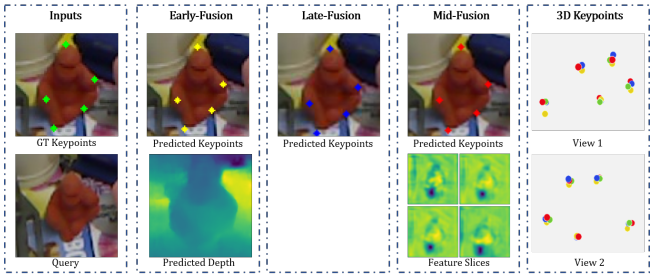

Fig. 5. **Visualization of keypoint prediction results from different fusion approach.** From left to right: input image pairs with ground truth 3D keypoints (green), predicted keypoints from Early-Fusion (yellow) and the predicted depth map, predicted keypoints from Late-Fusion (blue), predicted keypoints from Mid-Fusion (red) and four slices of the feature volume, and all the ground truth and predicted 3D keypoints in two different views with respective colors.

Every two adjacent frames are paired up for training, and every two frames with a certain gap are paired up for testing. The relative camera poses are obtained from the dataset.

In the Early-Fusion approach, we finetune the pretrained models provided by [38] on our datasets with an initial learning rate of 0.0001 and a stepwise decay strategy. The depth range is set according to the depth range of the dataset, and the depth resolution is set to $3.5mm$.

In the Mid-Fusion approach, the keypoint prediction network contains three 3D convolutional layers all with one 3D BatchNorm layer and 1 ReLU layer respectively, 1 output 3D convolutional layer, and finally 1 LogSoftmax layer. The 3D geometric volume is set in range $[-0.3, 0.3] \times [-0.3, 0.3] \times [-0.3, 0.3]$(meters) with grid size of $0.01m$ in the coarse level. While the range of the fine volume is dependent on the diameter of each object and its grid size is set to $0.005m$.

We run all our experiments on a machine equipped with an Intel(R) Xeon(R) Silver 4216 CPU at 2.10GHz, and an NVIDIA GeForce RTX 3090 GPU.

D. Evaluation on 3D Keypoint Precision

To validate the ability of our method to restore 3D information from the input RGB images, we train all the networks in the 3 proposed approaches on LineMOD dataset, then analyze and compare the quality of the predicted 3D object keypoints in terms of average 3D Keypoint Distance metric. During testing, we adopt a fixed gapping strategy to pair up images as inputs for our proposed method. Since the dataset is collected with handheld cameras, the relative camera poses between each frame are diverse even with fixed gaps. But generally a larger gap still leads to two farther views, as the example in Fig. 3 shows. Thus, to analyze the effect of pairing gaps, the evaluation is conducted under different gaps ($\{1, 5, 10\}$), in which $gap = 1$ means pairing two frames with a gap of one frame, and so on.

As can be seen in Table. I, with the increase of gap frames, the 3D keypoints prediction precision of the proposed methods are all enhanced. Among all the 3 approaches, the Mid-Fusion approach achieves the best prediction results, with an average keypoint error of $0.007m$ in $gap = 10$. The Late-Fusion approach is the second-best in $gap = 5$ and $gap = 10$ occasions. While the Early-Fusion approach

TABLE II

PERFORMANCE COMPARISON ON LINEMOD (ADD(-S)$< 0.1d$).

	RGB								RGBD	
	Pix2Pose [2]	PVNet [18]	DPOD [3]	CDPN [40]	HybridPose [19]	SO-Pose [22]	GDR-Net [21]	**Ours**	DenseFusion [4]	REDE [6]
ape	58.1	43.6	53.3	64.4	63.1			**94.9**	92.3	96.3
benchvise	91.0	99.9	95.3	97.8	**99.9**			**99.9**	93.2	98.9
cam	60.9	86.9	90.4	91.7	90.4			**93.7**	94.4	99.8
can	84.4	95.5	94.1	95.9	98.5			**99.0**	93.1	99.4
cat	64.0	79.3	60.4	83.8	89.4			**98.3**	96.5	99.2
driller	76.3	96.4	97.7	96.2	98.5			**99.1**	87.0	99.3
duck	43.8	52.6	66.0	66.8	65.0			**96.0**	92.3	96.2
eggbox*	96.8	99.1	99.7	99.7	**100.0**			99.3	99.8	100.0
glue*	79.4	95.7	93.8	**99.6**	98.8			99.0	100.0	100.0
holepuncher	74.8	81.9	65.8	85.8	89.7			**94.5**	86.9	99.2
iron	83.4	98.9	99.8	97.9	**100.0**			98.8	97.0	99.9
lamp	82.0	99.3	88.1	97.9	**99.5**			98.4	95.3	99.5
phone	45.0	92.4	74.2	90.1	94.9			**96.5**	92.8	98.9
mean	72.4	86.3	83.0	89.9	91.3	96.0	94.1	**97.5**	94.3	99.0

*denotes symmetric objects.

obtains the worst results in all pairing settings. We also compare the keypoints predicted by our method and those predicted by a SOTA RGBD-based method [6]. Table. I shows that, compared to [6] ($0.024m$), the Early-Fusion approach achieves slightly worse results, and the Late-Fusion approach gets better results with $gap = 5$ ($0.014m$) and $gap = 10$ ($0.011m$). What's more, the Mid-Fusion approach achieves the same results even under $gap = 1$, and better results with larger gaps.

To further verify the performance of the 3 proposed approaches under different parallax, we draw an accuracy ratio curve under different relative camera distances in 4 different objects (ape, can, cat, and glue), as shown in Fig. 4. Each point in the figure represents a ratio of the numbers of keypoints with an error less than $0.03m$ to the number of all keypoints under certain relative camera distances. The relative camera distances are computed as the distances between two focal points between the two cameras. The Mid-Fusion approach achieves the best results in most of the camera distances, especially in small ones. It exhibits the robustness and superiority of the Mid-Fusion approach in different situations.

Fig. 5 shows a visualized result comparison among the 3 proposed approaches. The 3D predicted keypoints are projected to the reference image by camera parameters, as well as shown in 2 different views in 3D space. Also, the depth map predicted by the Early-Fusion approach and the feature slices learned by the Mid-Fusion approach are exhibited. As can be seen, though some predictions show good performance when projected to 2D image plane, their errors are more clear when observed in the 3D space, such as the top keypoint predicted by the Early-Fusion approach.

To this end, we argue that among all the 3 proposed methods, the Mid-Fusion approach achieves the best prediction results under different situations, which shows its ability to restore the most useful 3D information for object pose estimation from two input RGB images. Therefore, we adopt the Mid-Fusion approach as the keypoint predictor to conduct experiments compared to other state-of-the-art methods.

E. Evaluation on Object Pose Estimation

To validate the enhancement of our method after combining 2D image features and learned 3D information, we evaluate the performance in LineMOD dataset and Occlusion LineMOD dataset and compete with state-of-the-art RGB-based methods and RGBD-based methods. In this section, all our methods are evaluated with $gap = 10$ pairing strategy.

Table. II displays the performance on LineMOD dataset. Best ADD(-S) recalls for each object are in bold, excluding RGBD-based methods. Our method excels all other RGB-based methods with a mean ADD recall of 97.5, and most of the methods in 9 object classes. Also, our method performs slightly better than the RGBD-based method [4] in 10 object classes and in the mean recall metric. Both indicate the advantage of combining 2D image features and 3D information together for pose estimation. Nonetheless, though predicts more accurate 3D keypoints than [6], our method performs worse than [6] in pose estimation. We consider the reason lies in the pose solver they adopt. In [6], after predicting sparse 3D keypoints, they deploy a mechanism called minimal solver bank to use dense real scene points recovered from the depth sensors to softly aggregate all the possible poses calculated from 3D registration, which is indeed an exhaustive RANSAC algorithm. The solver is robust enough to remedy the uncertainty brought by less accurate keypoints. Yet out Mid-Fusion approach is unable to deploy such an algorithm due to the lack of dense 3D points, while the Early-Fusion approach restores dense points with certain errors which will bring more disturbance to the solver.

Pose estimation under serious occlusion situations is a difficult task, especially for RGB-based methods. Table. III shows the performance on Occlusion LineMOD dataset. Best ADD(-S) recalls for each object is in bold, excluding RGBD-based methods. In this dataset with more occlusion and clutter, the mean ADD(-S) recall of our method beats all other RGB-based and RGBD-based methods. If excluding RGBD-based, our method beats RGB-based methods in 5 object classes. Note that we also beat [6] in this dataset.

TABLE III

PERFORMANCE COMPARISON ON OCCLUSION LINEMOD (ADD(-S)$< 0.1d$).

	RGB							RGBD		
	PoseCNN [1]	PVNet [18]	DPOD [3]	HybridPose [19]	SO-Pose [22]	GDR-Net [21]	**Ours**	PVN3D [9]	REDE [6]	FFB6D [24]
ape	9.6	15.8		20.9	48.4	46.8	**49.2**	33.9	55.9	47.2
can	45.2	63.3		75.3	85.8	**90.8**	89.3	88.6	87.9	85.2
cat	0.93	16.7		24.9	32.7	**40.5**	37.7	39.1	37.5	45.7
driller	41.4	65.7		70.2	77.4	82.6	**90.1**	78.4	75.3	81.4
duck	19.6	25.2		27.9	48.9	46.9	**57.8**	41.9	48.5	53.9
eggbox*	22.0	50.2		52.4	52.4	54.2	**57.3**	80.9	73.3	70.2
glue*	38.5	49.6		53.8	**78.3**	75.8	71.6	68.1	74.8	60.1
holepuncher	22.1	39.7		54.2	75.3	60.1	**80.0**	74.7	61.9	85.9
mean	24.9	40.8	47.3	47.5	62.3	62.2	**66.6**	63.2	64.2	66.2

*denotes symmetric objects.

Under the heavily occluded situation, our method not only take advantage of both 2D image features and learned 3D information, but is also benefited from extra observation views. Hence, we have more opportunities to acquire useful information for keypoint prediction and pose estimation.

Some visualization results are shown in Fig. 6. We project the object CAD model transformed by the estimated pose and draw the points on the reference view. All the images are cropped for better visualization. Compared with RGB-based method PVNet [18], our method achieves more accurate results in textureless surfaces, bizarre viewing angle, and some normal situations. Also, our method beats both [18] and RGBD-based method [6] in occluded occasions.

V. CONCLUSIONS

In this paper, we propose a framework to learn implicit 3D information from 2 input RGB images for 6D object pose estimation. In this framework, we conduct an investigation on 3 different 3D information learning approaches. Based on the experiments, we reveal a Mid-Fusion mechanism that fuses 2D image features geometrically to 3D space produces the most precise keypoints. We further show that with this Mid-Fusion mechanism, our framework is capable of enhancing the pose estimation performance of RGB-based methods, and achieving comparable results with RGBD-based methods. However, our performance in less occluded environments are still inferior to RGBD-based methods, mostly because of the lack of dense reliable scene points. Thus, we plan to tackle the challenge by employing two or more views of RGBD inputs to solve the pose in the future.

REFERENCES

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

[2] K. Park, T. Patten, and M. Vincze, "Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7668–7677, 2019.

[3] S. Zakharov, I. Shugurov, and S. Ilic, "Dpod: 6d pose object detector and refiner," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1941–1950, 2019.

[4] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3343–3352, 2019.

[5] K. Wada, E. Sucar, S. James, D. Lenton, and A. J. Davison, "Morefusion: Multi-object reasoning for 6d pose estimation from volumetric fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14540–14549, 2020.

[6] W. Hua, Z. Zhou, J. Wu, H. Huang, Y. Wang, and R. Xiong, "Rede: End-to-end object 6d pose robust estimation using differentiable outliers elimination," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2886–2893, 2021.

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

[8] Y. Shi, J. Huang, X. Xu, Y. Zhang, and K. Xu, "Stablepose: Learning 6d object poses from geometrically stable patches," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15222–15231, 2021.

[9] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11632–11641, 2020.

[10] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5418, 2018.

[11] P. Li, X. Chen, and S. Shen, "Stereo r-cnn based 3d object detection for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7644–7652, 2019.

[12] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 767–783, 2018.

[13] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3828–3836, 2017.

[14] T. Hodan, D. Barath, and J. Matas, "Epos: Estimating 6d pose of objects with symmetries," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11703–11712, 2020.

[15] F. Manhardt, D. M. Arroyo, C. Rupprecht, B. Busam, T. Birdal, N. Navab, and F. Tombari, "Explaining the ambiguity of object detection and 6d pose from visual data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6841–6850, 2019.

[16] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6d object pose prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 292–301, 2018.

[17] Q. Luo, H. Ma, L. Tang, Y. Wang, and R. Xiong, "3d-ssd: Learning hierarchical features from rgb-d images for amodal 3d object detection," *Neurocomputing*, vol. 378, pp. 364–374, 2020.

[18] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4561–4570, 2019.

[19] C. Song, J. Song, and Q. Huang, "Hybridpose: 6d object pose estimation under hybrid representations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 431–440, 2020.

[20] Y. Hu, P. Fua, W. Wang, and M. Salzmann, "Single-stage 6d object pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2930–2939, 2020.

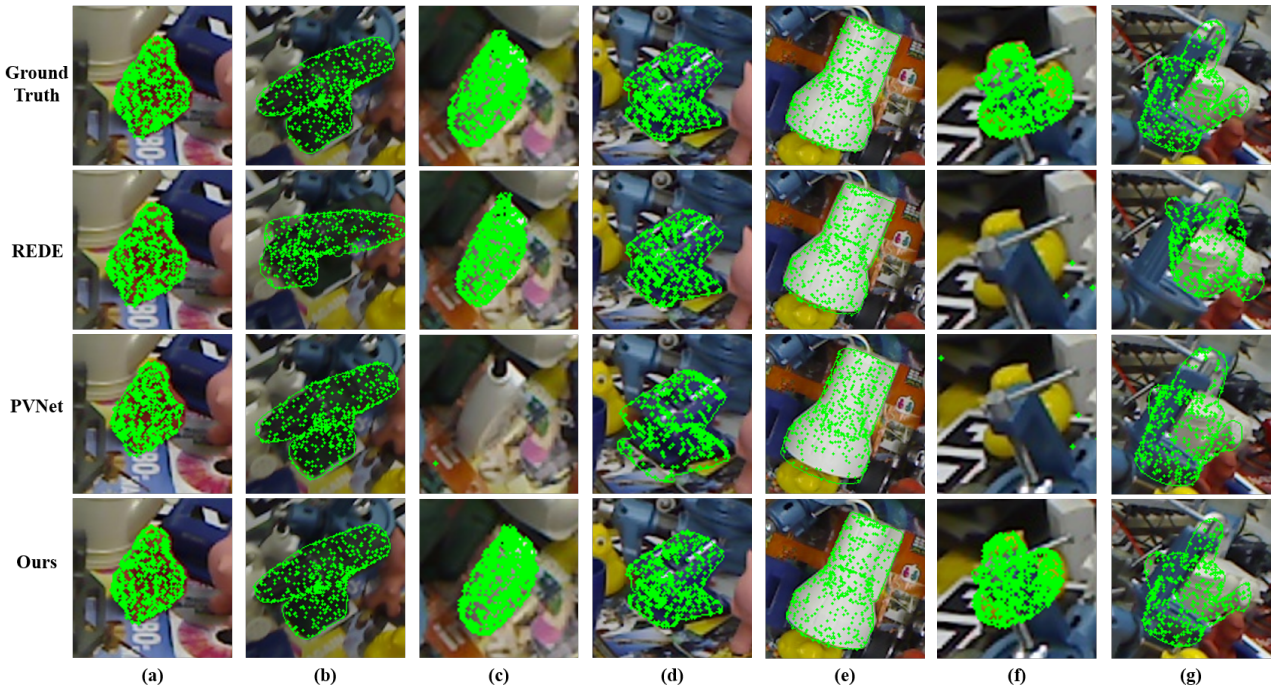

Fig. 6. **Visualization Results Samples.** Some visualization results on LindMOD and Occlusion LineMOD dataset. The pose results are shown by projecting the model points via the predicted pose from each method. We also show the contour of the model projection mask. All images are cropped for better visualization. The cases with few dislocated projected points represent a predicted pose with large errors. The last row shows our result with Mid-Fusion approach. Compared with RGB-based method PVNet [18], our method achieves more accurate results in textureless surfaces (e), bizarre viewing angle (c), and some normal situations (a)(b)(d). Also, our methods beats [18] and RGBD-based method [6] in occluded occasions (f)(g).

[21] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16611–16621, 2021.

[22] Y. Di, F. Manhardt, G. Wang, X. Ji, N. Navab, and F. Tombari, "So-pose: Exploiting self-occlusion for direct 6d pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12396–12405, 2021.

[23] L. Saadi, B. Besbes, S. Kramm, and A. Bensrhair, "Optimizing rgb-d fusion for accurate 6dof pose estimation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2413–2420, 2021.

[24] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "Ffb6d: A full flow bidirectional fusion network for 6d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3003–3013, 2021.

[25] 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 *Asian conference on computer vision*, pp. 548–562, Springer, 2012.

[26] 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 *2011 international conference on computer vision*, pp. 858–865, IEEE, 2011.

[27] M. Gonzalez, A. Kacete, A. Murienne, and E. Marchand, "L6dnet: Light 6 dof network for robust and precise object pose estimation with small datasets," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2914–2921, 2021.

[28] A. Collet and S. S. Srinivasa, "Efficient multi-view object recognition and full pose estimation," in *2010 IEEE International Conference on Robotics and Automation*, pp. 2050–2055, IEEE, 2010.

[29] A. Collet, M. Martinez, and S. S. Srinivasa, "The moped framework: Object recognition and pose estimation for manipulation," *The international journal of robotics research*, vol. 30, no. 10, pp. 1284–1306, 2011.

[30] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "Cosypose: Consistent multi-view multi-object 6d pose estimation," in *European Conference on Computer Vision*, pp. 574–591, Springer, 2020.

[31] J. Fu, Q. Huang, K. Doherty, Y. Wang, and J. J. Leonard, "A multi-hypothesis approach to pose ambiguity in object-based slam," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7639–7646, IEEE, 2021.

[32] X. Liu, R. Jonschkowski, A. Angelova, and K. Konolige, "Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11602–11610, 2020.

[33] N. D. F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo," in *European Conference on Computer Vision*, 2008.

[34] Engin, Tola, Christoph, Strecha, Pascal, and Fua, "Efficient large-scale multi-view stereo for ultra high-resolution image sets," *Machine Vision and Applications*, vol. 23, no. 5, pp. 903–920, 2011.

[35] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent mvsnet for high-resolution multi-view stereo depth inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5525–5534, 2019.

[36] K. Luo, T. Guan, L. Ju, H. Huang, and Y. Luo, "P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10452–10461, 2019.

[37] J. Yang, W. Mao, J. M. Alvarez, and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4877–4886, 2020.

[38] X. Gu, Z. Fan, S. Zhu, Z. Dai, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[39] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6d object pose estimation using 3d object coordinates," in *European conference on computer vision*, pp. 536–551, Springer, 2014.

[40] Z. Li, G. Wang, and X. Ji, "Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7678–7687, 2019.