Seeing Through the Occluders: Robust Monocular 6-DOF Object Pose Tracking via Model-guided Video Object Segmentation

Leisheng Zhong[1], Yu Zhang[1], Hao Zhao[1], An Chang[1], Wenhao Xiang[2], Shunli Zhang[3], and Li Zhang[1]

***Abstract*—To deal with occlusion is one of the most challenging problems for monocular 6-DOF object pose tracking. In this paper, we propose a novel 6-DOF object pose tracking method which is robust to heavy occlusions. When the tracked object is occluded by another object, instead of trying to detect the occluder, we seek to see through it, as if the occluder doesn't exist. To this end, we propose to combine a learning-based video object segmentation module with an optimization-based pose estimation module in a closed loop. Firstly, a model-guided video object segmentation network is utilized to predict the accurate and full mask of the object (including the occluded part). Secondly, a non-linear 6-DOF pose optimization method is performed with the guidance of the predicted full mask. After solving the current object pose, we render the 3D object model to obtain a refined, model-constrained mask of the current frame, which is then fed back to the segmentation network for processing the next frame, closing the whole loop. Experiments show that the proposed method outperforms the state-of-arts by a large margin for dealing with heavy occlusions, and could handle extreme cases which previous methods would fail.**

***Index Terms*—Computer Vision for Other Robotic Applications; Visual Tracking; Virtual Reality and Interfaces**

I. Introduction

TRACKING the 6-DOF pose of a known rigid object in monocular video sequences is a fundamental problem in 3D computer vision [1]. Many popular applications depend on robust and accurate pose tracking algorithms, including robotic perception and manipulation, augmented reality (AR), and human-computer interaction [2]–[9]. With only monocular input, it is very challenging to robustly and accurately determine the translation and rotation of the object in unconstrained 3D environment. An even more difficult occasion is to successfully track the object with some unknown occluders moving in front of it. While the occlusion situation is very common in real applications (e.g., the object is often occluded by human hands in AR applications), this problem is still largely unsolved. Previous approaches try to alleviate the influence of the occluders by detecting them and removing the corresponding pixels from the subsequent processes [10], [11]. However, these approaches will not work for heavy occlusions, in which case a large proportion of the object is occluded, leaving very few useful information of the non-occluded part. In this paper, we consider this problem in a different way. Instead of trying to detect the occluders, we seek to see through it. As shown in Fig. 1, we try to predict the accurate full mask of the object even when it is heavily occluded. The reason why this could be done is that we have strong priors for the object shape and its previous location. Given the geometrical and temporal priors of the object, it is possible to interpret its full mask in the current image, even if it is occluded. In order to achieve this goal, we design a novel 6-DOF object tracking pipeline (as demonstrated in Fig. 2), which is based on two main observations: First is to build a mutual guidance loop of video object segmentation and 6-DOF object pose estimation, and second is to combine learning and optimization in a complementary manner.

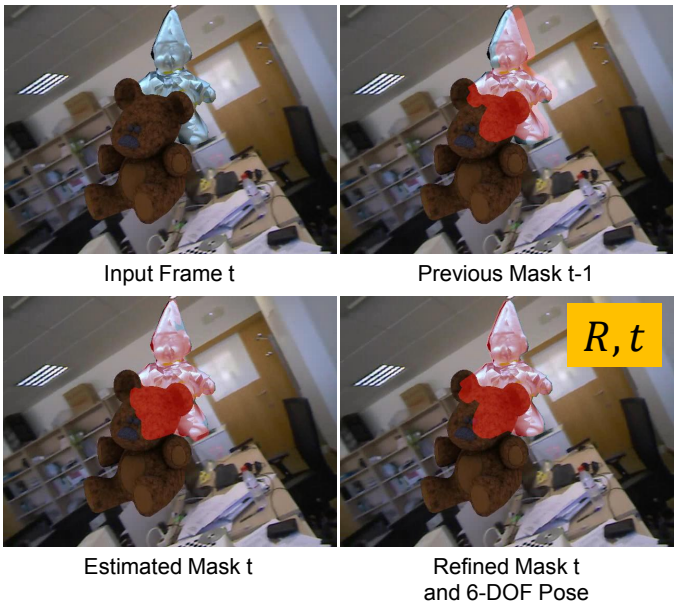

Fig. 1: Our method could see through the occluder and accurately predict the full mask of the target object together with its 6-DOF pose.

Manuscript received: February 23, 2020; Revised: May 12, 2020; Accepted: June 8, 2020.

This paper was recommended for publication by Editor Eric Marchand upon evaluation of the Associate Editor and Reviewers' comments. This work is supported by the National Natural Science Foundation of China under Grant No.61871248 and No.61976017, and the Beijing Natural Science Foundation under Grant No.4202056. *(Corresponding author: Li Zhang.)*

[1]Leisheng Zhong, Yu Zhang, Hao Zhao, An Chang and Li Zhang are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. {zls13, zhang-yu15, zhao-h13, cal3}@mails.tsinghua.edu.cn, chinazhangli@mail.tsinghua.edu.cn

[2]Wenhao Xiang is with Systems Engineering Research Institute, CSSC, Beijing, 100094, China. xiangwh2018@163.com

[3]Shunli Zhang is with the School of Software Engineering, Beijing Jiaotong University, Beijing 100044, China. slzhang@bjtu.edu.cn

Digital Object Identifier (DOI): see top of this page.

A. The Mutual Guidance of Video Object Segmentation and 6-DOF Object Pose Estimation

The first observation of this paper is that video object segmentation and 6-DOF object pose estimation could be guided by one another. The 2D segmentation of the target object provides foreground/background probabilities in the image, which could lead the 3D object model to rigidly transform towards the best fitting of the projected contour and the probabilistic contour. Consequently, the 6-DOF pose of the object could be estimated through iterative optimization. On the other hand, with the current estimated pose and the 3D object model, a refined and model-constrained segmentation mask could be obtained, and serves as a guidance for segmenting the object in the next frame. Since the target object is a rigid body, the 2D object shape in the image is constrained by the shape of the 3D object model. As a result, after the pose estimation step, the refined and model-constrained segmentation mask will be more accurate than the network prediction, which would benefit the segmentation process for the next frame.

Full segmentation (also referred to as *amodal segmentation* [12], which aims to segment the whole object including the occluded part) under occlusion is difficult, but would be easier with the guidance from the 6-DOF pose estimation result of the previous frame. Also, accurate 6-DOF pose estimation under occlusion is difficult, but would be easier with the guidance from the full segmentation mask. The mutual guidance of these two modules perfectly utilizes both the temporal and geometrical priors of the target object, which contributes to a novel occlusion-resistant 6-DOF object tracking pipeline.

B. Improving Robustness and Accuracy by Combining Learning and Optimization

Our second key observation is that learners and optimizers could be combined with each other in a complementary manner. As also mentioned in the previous work [13], learning-based methods tend to be more robust when encountering challenging situations, while optimization-based methods could converge to a better accuracy. Both robustness and accuracy are crucial in 6-DOF object tracking applications, so we seek to combine these two kinds of methods together. Specifically, we decouple the whole problem into two successive stages: the segmentation stage, and the pose estimation stage. Since deep neural networks have achieved great success in image and video segmentation, we perform the segmentation stage using a learning-based method. In order to obtain higher accuracy for 6-DOF pose estimation, we utilize an optimization-based method in the second stage to estimate the pose parameters and refine the segmentation mask. Combining these two stages in a closed loop, the proposed method benefits from both the robustness of learning-based methods and the high precision of optimization-based methods.

C. Contribution

The contributions of our method are:

(1) We propose a novel 6-DOF object pose tracking pipeline based on the mutual guidance of segmentation and pose estimation. The proposed tracker is able to handle heavy occlusions by seeing through the occluders. To the best of our knowledge, we are the first to handle occlusions in 6-DOF tracking by interpreting the full mask information behind the occluders.

(2) We improve the robustness and accuracy of our tracker by combining a learning-based segmentation module with a optimization-based pose estimation module in a complementary manner.

II. Related Work

A. Monocular 6-DOF Object Pose Tracking

In early years, feature-based methods and edge-based methods have been popular for the task of monocular 6-DOF object pose tracking [1]. However the so-called region-based methods have proved to achieve state-of-art performance in recent years. Region-based methods estimate the pose of the object by maximizing the discrimination between the statistical foreground and background appearance models [14]. Starting from the famous work PWP3D [14], a lot of region-based methods have been proposed. The original gradient descent based optimization is replaced by a more efficient Gauss-Newton-like optimization strategy in [15]. The global region model used in [14] is also replaced by a localized region model in [16] to improve the performance in cluttered scenes. A more recent work [17] further extends the localized model by introducing the so-called tclc-histograms. The author summarizes their previous works [15], [17] in [18] and expands them by providing a mathematical explanation of the iteratively reweighted Gauss-Newton optimization. Later, the authors in [19] propose a hybrid tracker by combining the region-based energy function with a photometric term, which could be beneficial for handling symmetrical objects. Recently, deep learning based methods have also been proposed for pose refinement [20], [21]. Although these methods are designed for refining the pose estimation of single-shot estimators, they could also be utilized as trackers.

B. Image and Video Object Segmentation

Deep learning based methods have achieved great success in the fields of image and video object segmentation. For image segmentation, a lot of network architectures have been proposed (including instance and semantic segmentation), such as SegNet [22], Mask-RCNN [23], DeepLab [24], [25], etc. A more closely related topic to this work is the so-called *Amodal Segmentation*, which aims to segment both the visible and occluded parts of the object. Li *et al.* [12] proposes the first amodal instance segmentation method by manually adding occlusions to modal image segmentation datasets. Qi *et al.* [26] proposes an amodal instance segmentation method with a augmented version of the KITTI dataset. Zhu *et al.* [27] introduces a semantic amodal segmentation method that includes annotations of the full extent of the objects, semantic labels, visible edges and depth order. As for video object segmentation, some recent methods segment the target relying only on each single frame without temporal information [28],

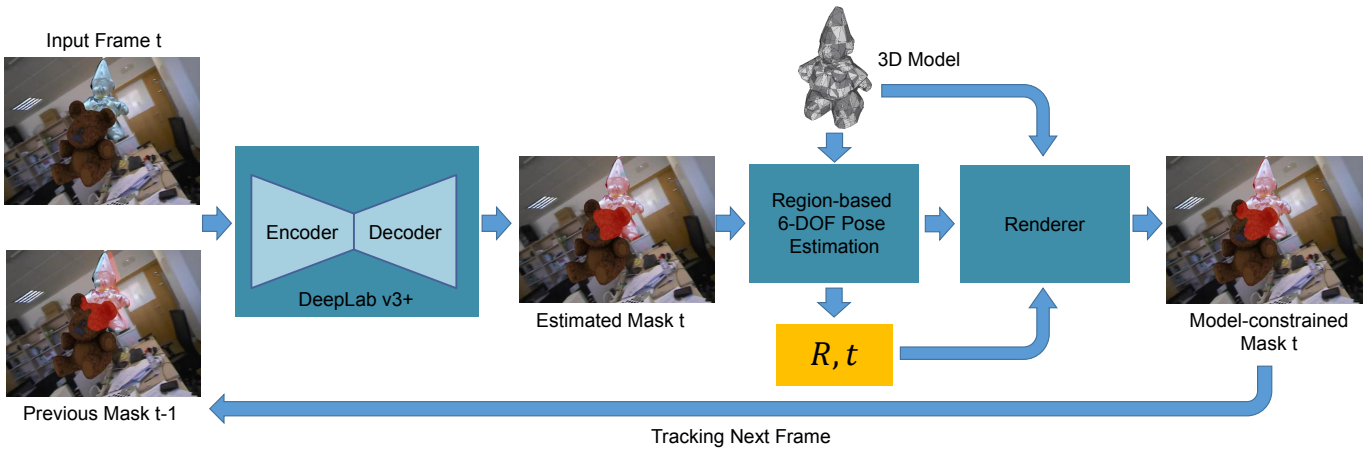

Fig. 2: The proposed 6-DOF object pose tracking pipeline. This novel pipeline contains a model-guided video object segmentation stage and a segmentation-guided 6-DOF pose estimation stage, forming a closed loop.

[29]. However, it is beneficial to utilize the temporal consistency in continuous video frames, as proposed in MaskTrack [30] and some following works such as [31], which perform mask propagation from timestamp $t-1$ to timestamp t. We also follow this idea in the proposed video object segmentation stage.

C. Occlusion Handling

To handle occlusion is one of the most difficult problems in monocular 6-DOF object tracking. To tackle this issue, a modified on-line update rule for the statistical models is introduced in [32], but the new rule only considers the model update step, and is not guaranteed to work for complex occluders and heavy occlusions. In [10], the occluded area is detected by comparing the current image to a realistically rendered template image, which requires high-quality textured object model. The authors in [15] try to handle mutual occlusions by tracking both the target and occluder, which is only possible when the 3D model of the occluder is given. An edge-based occlusion handling strategy is proposed in [11], which assigns less weights to the contour points that are far away from image edges. These occlusion handling strategies could handle certain degrees of occlusion, but would still fail when encountering heavy occlusions.

III. Method

In this section, we describe the proposed novel 6-DOF object tracking framework in detail. The overview of the proposed mutual guidance pipeline is first introduced in Sec. III-A. Then we present the model-guided video object segmentation stage and the segmentation-guided 6-DOF pose estimation stage in Sec. III-B and Sec. III-C respectively.

A. The Proposed Mutual Guidance Pipeline for 6-DOF Object Tracking

The overview of the proposed mutual guidance pipeline for 6-DOF object tracking is summarized in Fig. 2. The whole pipeline consists of two stages: the model-guided video object segmentation stage, and the segmentation-guided 6-DOF pose estimation stage. Firstly, given the current frame and the previous predicted (and refined) full mask of the object, an encoder-decoder network is applied to estimate the full object mask in the current frame. Secondly, the estimated full object mask is sent into a region-based 6-DOF pose estimation module as the probabilistic foreground/background maps. Together with the known 3D object model, an efficient second-order optimization step is performed to estimate the current 6-DOF pose of the object. After that, a refined mask could be obtained by rendering the 3D object model with the estimated pose parameters. The refined, model-constrained mask is then fed back to the segmentation network for tracking the next frame, closing the whole loop. We present these two stages in detail in the following two subsections.

B. Model-guided Video Object Segmentation

In the first stage, we utilize the idea of mask propagation and design a video object segmentation module which is able to see through occluders. The key is to exploit the temporal prior from the previous frame and the geometrical prior from the 3D object model. Although the target object is occluded, the previous full mask gives very useful clues about its current position, which enables the network to predict the current full mask. On the other hand, the network could also learn the geometrical structure of the target object through numerous synthetic training data (see below), which also helps to predict the current mask with better precision.

A recent state-of-art semantic image segmentation network, DeepLab v3+ [25] is employed in order to achieve the best segmentation performance. We have made several modifications to the original network. Firstly, the input to the network is modified to 4 channels (current frame RGB + previous mask) instead of only the current frame. Secondly, instead of predicting hard semantic labels, we extract the soft probabilities after the final softmax layer as the predicted soft object mask, which is then sent into the 6-DOF pose estimation stage. Using soft probabilities instead of hard labels is crucial in obtaining good tracking performance in our task, which will be demonstrated in the experiment part. Inspired by [30], we also render synthetic data for off-line training. Finally, to make the trained model adapt better to the specific scene, an on-line

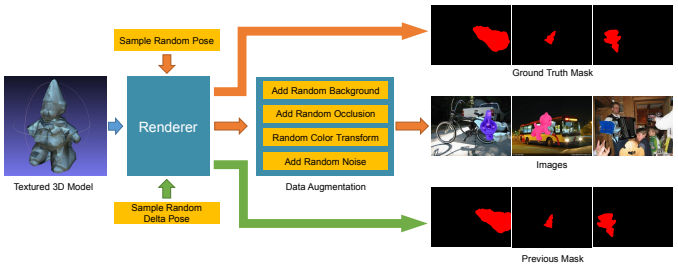

Fig. 3: Training data generation in the off-line training step.

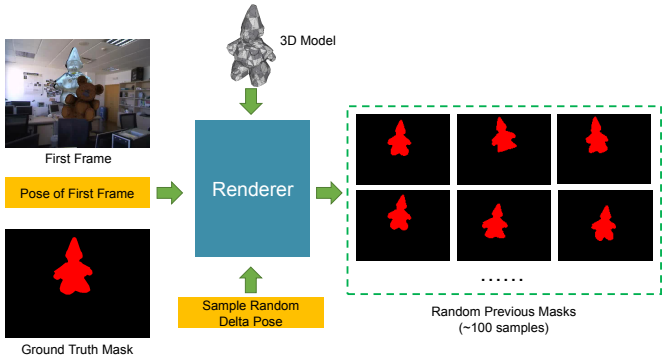

Fig. 4: Training data generation in the on-line fine-tuning step.

fine-tuning step is applied using only the first frame of the test video.

1) Off-line Training: The off-line training step is illustrated in Fig. 3. We only use synthetic data for network training. The training samples are generated by rendering the textured 3D model with randomly sampled 6-DOF poses $\mathbf{p} = (\omega_1, \omega_2, \omega_3, t_1, t_2, t_3)^T \in \mathbb{R}^6$. The rendered RGB image is further processed with the following augmentations: (1) Add random background using pictures from the Pascal VOC dataset [33]; (2) Add random occlusion using the 3D models from the LineMod dataset [34]; (3) Add random color transform to the occluder in order to make the network generalize better to unknown occluders; (4) Add random Gaussian noise to the image. The rendered foreground mask is used as the ground-truth object mask. Next, for each rendered image and ground-truth mask, we sample a random delta pose $\Delta\mathbf{p} = (\Delta\omega_1, \Delta\omega_2, \Delta\omega_3, \Delta t_1, \Delta t_2, \Delta t_3)^T \in \mathbb{R}^6$ to simulate the movement from the previous frame to the current frame. Then we render the 3D object model with the simulated previous pose $\mathbf{p}_0 = \mathbf{p} + \Delta\mathbf{p}$. The rendered foreground mask is then saved as the previous mask.

Each of our training samples contains a current image $I = \{R, G, B\}$, a current ground-truth mask M, and a simulated previous mask M_0. The input to the network is a 4-channel tensor $\{R, G, B, M_0\}$, and the desired output is M. By default, standard Cross Entropy loss is used as the loss function. We render around $\sim 10^4$ training samples for each object in the off-line training step.

2) On-line Fine-tuning: To further improve the segmentation quality, we employ an on-line fine-tuning step using the first frame of the test video. In 6-DOF tracking tasks, the pose of the first frame is usually given for initialization. So we could utilize the first frame to fine-tune the network trained on synthetic data and make it generalize better to the specific video. The on-line fine-tuning step is demonstrated in Fig. 4.

Although only one single real image is available for fine-tuning, we could randomly sample more than one possible previous masks. To this end, given the ground-truth pose of the first frame $\mathbf{p}_{gt} = (\omega_1, \omega_2, \omega_3, t_1, t_2, t_3)^T \in \mathbb{R}^6$, we randomly sample $\sim$ 100 delta poses $\Delta\mathbf{p} = (\Delta\omega_1, \Delta\omega_2, \Delta\omega_3, \Delta t_1, \Delta t_2, \Delta t_3)^T \in \mathbb{R}^6$. Then we render the 3D model with the simulated previous poses $\mathbf{p}_0 = \mathbf{p}_{gt} + \Delta\mathbf{p}$ and obtain multiple previous masks. After that, the network is further fine-tuned using the first video frame and the multiple simulated previous masks.

C. Segmentation-guided 6-DOF Pose Estimation

After obtaining the current full mask of the target object from the video object segmentation network, we perform 6-DOF pose estimation based on the 3D model, the predicted segmentation mask, and the previous pose. The 6-DOF pose tracking stage follows the region-based statistical formulation [14]:

$$E(\mathbf{p}) = -\sum_{\mathbf{x}\in\Omega} \log[H_e(\Phi(\mathbf{x}(\mathbf{p})))P_f(\mathbf{x}) + (1 - H_e(\Phi(\mathbf{x}(\mathbf{p}))))P_b(\mathbf{x})] \quad (1)$$

where $\mathbf{p} = (\omega_1, \omega_2, \omega_3, t_1, t_2, t_3)^T \in \mathbb{R}^6$ is the 6-DOF object pose. Ω is the image region, $\mathbf{x} = (x, y)^T \in \Omega$ are the pixel locations in the image. $\Phi(\mathbf{x})$ is the signed distance function, and the object contour $\mathbf{C}$ is defined as the zero level-set of $\Phi(\mathbf{x})$: $\mathbf{C} = \{\mathbf{x} | \Phi(\mathbf{x}) = 0\}$. H_e is the smoothed Heaviside step function. $P_f(\mathbf{x})$ and $P_b(\mathbf{x})$ are the posterior maps for the foreground and the background respectively. In the foreground region, each pixel $\mathbf{x}$ corresponds to the projection of the a 3D model point:

$$\mathbf{x} = \pi(K\mathbf{X}) = \pi[K(R\mathbf{X}_0 + \mathbf{t})] \quad (2)$$

where $\mathbf{X} = (X, Y, Z)^T$ is the coordinate of the 3D model point in the camera coordinate frame, and $\mathbf{X}_0 = (X_0, Y_0, Z_0)^T$ is the coordinate in the object coordinate frame. $R \in \mathbb{SO}(3)$ is the rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ is the translation vector, which could be calculated from $\mathbf{p}$. $K \in \mathbb{R}^{3\times3}$ is the intrinsic matrix of the camera, and $\pi(\mathbf{X}) = (X/Z, Y/Z)^T$. The energy function (1) is calculated once for each pixel $\mathbf{x}$. The correspondence between pixel position $\mathbf{x}$ and 3D model point $\mathbf{X}$ is obtained by back-projecting $\mathbf{x}$ to the 3D model.

The aim of region-based pose tracking is to minimize the energy function Eq. (1). The basic idea of Eq. (1) is to maximize the discrimination between statistical foreground and background appearance models via direct optimization of the 6-DOF pose parameters [14]. The posteriors $P_f(\mathbf{x})$ and $P_b(\mathbf{x})$ represent 2D image statistics of foreground/background regions, and the $H_e(\Phi(\mathbf{x}(\mathbf{p})))$ and $1 - H_e(\Phi(\mathbf{x}(\mathbf{p})))$ represent the soft spatial partition of foreground/background regions (defined by the projected contour of the 3D model). When these two components best match (i.e., the pixels inside the contour best fit the foreground statistics and the pixels outside the contour best fit the background statistics), the energy function is minimized and the optimal pose parameters are obtained.

In all of the previous region-based methods [14]–[19], the posterior maps $P_f(\mathbf{x})$ and $P_b(\mathbf{x})$ are calculated using color histograms:

$$\begin{aligned} P_f(\mathbf{x}) &= P(M_f|\mathbf{y}) \\ &= \frac{P(M_f)\,P(\mathbf{y}|M_f)}{P(M_f)\,P(\mathbf{y}|M_f) + P(M_b)\,P(\mathbf{y}|M_b)} \end{aligned} \tag{3}$$

where $\mathbf{y} = I(\mathbf{x})$ are the RGB values of pixel $\mathbf{x}$, $P(\mathbf{y}|M_f)$ and $P(\mathbf{y}|M_b)$ are the color histograms of the foreground region and the background region, $P(M_f)$ and $P(M_b)$ are the region priors. $P_b(\mathbf{x})$ is in the similar form with $P_f(\mathbf{x})$. Since previous methods use simple color histograms to calculate the foreground and background probabilities, they would fail when encountering heavy occlusions, because the color histograms would be corrupted by the occluder.

In this paper, we propose to generate robust foreground and background probability maps from the video object segmentation network. By replacing the simple color histograms with a robust video object segmentation network, our method obtains much better performance compared to previous works in heavy occlusions. More specifically, in our method, the foreground and background probability maps are calculated as:

$$P_f(\mathbf{x}) = Soft(M_{pred}) \tag{4}$$

$$P_b(\mathbf{x}) = 1 - Soft(M_{pred}) \tag{5}$$

where M_{pred} is the predicted object mask from the segmentation network, and $Soft(M_{pred})$ indicates the soft probabilities extracted from the output of the softmax layer.

To minimize Eq. (1), we use a similar Gauss-Newton-based pose optimization strategy as in [18] by rewriting the energy function Eq. (1) as a non-linear iteratively re-weighted least squares problem:

$$E(\mathbf{p}) = \frac{1}{2}\sum_{\mathbf{x}\in\Omega}\psi(\mathbf{x})F^2(\mathbf{x},\mathbf{p}) \tag{6}$$

where

$$\begin{aligned} F(\mathbf{x},\mathbf{p}) = -\log[&H_e(\Phi(\mathbf{x}(\mathbf{p})))P_f(\mathbf{x}) \\ &+ (1 - H_e(\Phi(\mathbf{x}(\mathbf{p}))))P_b(\mathbf{x})] \end{aligned} \tag{7}$$

and $\psi(\mathbf{x}) = \frac{1}{F(\mathbf{x},\mathbf{p})}$.

Then the non-linear optimization problem could be iteratively solved by fixing and alternatingly updating the weights $\psi(\mathbf{x})$. The Jacobian is calculated as:

$$\begin{aligned} \mathbf{J}(\mathbf{x}) &= \frac{\partial F(\mathbf{x},\mathbf{p})}{\partial \mathbf{p}} \\ &= -\frac{P_f - P_b}{H_e(\Phi(\mathbf{x}))\,P_f + (1 - H_e(\Phi(\mathbf{x})))\,P_b} \\ &\quad \times \frac{\partial H_e(\Phi(\mathbf{x}))}{\partial \mathbf{p}} \end{aligned} \tag{8}$$

and

$$\frac{\partial H_e(\Phi(\mathbf{x}))}{\partial \mathbf{p}} = \frac{\partial H_e}{\partial \Phi}\frac{\partial \Phi}{\partial \mathbf{x}}\frac{\partial \mathbf{x}}{\partial \mathbf{p}} \tag{9}$$

where $\frac{\partial H_e}{\partial \Phi} = \delta_e(\Phi)$ is the smoothed Dirac delta function, $\frac{\partial \Phi}{\partial \mathbf{x}} = \left[\frac{\partial \Phi}{\partial x}, \frac{\partial \Phi}{\partial y}\right]$ is calculated using centered finite differences. $\frac{\partial \mathbf{x}}{\partial \mathbf{p}}$ can be derived from Eq. (2), and the details can be found in [16], [18].

The Hessian is then approximated using first-order derivatives [18]:

$$\mathbf{H}(\mathbf{x}) = \psi(\mathbf{x})\,\mathbf{J}(\mathbf{x})^T\mathbf{J}(\mathbf{x}) \tag{10}$$

which leads to the optimal Gauss-Newton update:

$$\mathbf{\Delta p} = -\left(\sum_{\mathbf{x}\in\Omega}\mathbf{H}(\mathbf{x})\right)^{-1}\sum_{\mathbf{x}\in\Omega}\mathbf{J}(\mathbf{x})^T \tag{11}$$

After updating the pose $\mathbf{p}$, we re-render the 3D object model to obtain a refined, model-constrained full mask of the target object, and feed this refined mask back to the segmentation network for tracking the next frame. Now we have closed the whole 6-DOF tracking loop.

IV. Experiments

In this section, we provide detailed evaluation of the proposed method. We first clarify some implementation details. Then we present a careful ablation study to compare the performance of some different settings in the video object segmentation network. Finally, we compare the overall 6-DOF tracking performance of our method with some state-of-art methods on a challenging dataset.

A. Implementation Details

For training the video object segmentation network, the DeepLab v3+ [25] network with ResNet as backbone is utilized. We use SGD with batch size of 8. A polynomial learning policy with initial learning rate of 0.005 is applied. The momentum and weight decay are set to 0.9 and 0.0005, respectively. The network is trained for 30k iterations for the off-line training and 300 iterations for the on-line fine-tuning. The fine-tuning step takes about 10 min in practice. The input and output resolution is 513×513.

For sampling the delta pose $\mathbf{\Delta p}$, random Gaussian perturbations with standard deviation of $10mm$ and 5° are added to the translation and rotation parameters respectively.

The network is trained on a single Nvidia Titan Xp GPU, and the inference time is about 30 ms per frame. The non-linear pose optimization is performed on CPU and costs about 20 ms per frame. As a result, our tracker could run at 20 Hz.

For evaluation, we choose the Rigid Pose Dataset [35], which provides semi-synthetic video sequences of 6 different objects (including both well-textured and texture-less objects, as well as symmetrical objects) under a variety of realistic conditions. The dataset composes of original sequences, noisy sequences and occluded sequences (18 video sequences in total). These sequences are featured with wide-range rotation and translation, object variability and background cluttering. We first perform an ablation study on this dataset to test the segmentation results with different settings. Next, we conduct a full evaluation of our method for 6-DOF tracking and compare with some state-of-arts.

B. Ablation Study on Video Object Segmentation

In this subsection, we test the proposed video object segmentation network with some different settings on Rigid Pose Dataset [35]. We use two kinds of loss functions: the standard Cross Entropy loss and the L1 loss. By using the Cross Entropy loss, the segmentation is formulated as a pixel labeling (classification) task. And by using L1 loss, we treat the segmentation problem as a regression task. We also test a baseline method in which the input tensor is without the previous mask, and a different version of our method without the online fine-tuning step. We evaluate using the standard mIoU metric: intersection-over-union of the estimated segmentation and the ground-truth binary mask. When using L1 loss, a hard threshold of 0.5 is used in order to calculate the mIoU metric. The results are summarized in Table I.

TABLE I: The mIoU scores obtained by different variants of our method on Rigid Pose Dataset.

	Cross Entropy	w/o finetune	L1	w/o prev. mask
Val	96.47	96.54	93.35	88.16
Test	97.68	95.89	92.84	87.68

Obviously, when the previous mask is not available, the segmentation results are poor owing to the lack of temporal prior. Compared to the complete version of our method, the results without fine-tuning step is similar on validation set, but poorer on test set, because of the lack of specific background information in these test videos. On the other hand, as discussed in Sec. III-B and Sec. III-C, when trained by Cross Entropy loss, using the soft probabilities extracted from the softmax layer would lead to better tracking performance than using the final hard segmentation results. The reason is that soft probabilities preserve more information than hard thresholding. Then it is natural to expect that even better precision could be obtained using regression instead of classification. However, after repeated experiments, we find that the results of using L1 loss are inferior to that of using Cross Entropy loss. Part of the reason might be that the regression network is much more difficult to train compared to the classification network.

Some of the segmentation results are illustrated in Fig. 5. Note that the 'Soft Seg' and 'Hard Seg' results are both obtained using the Cross Entropy loss, but the former is extracted from the output of the softmax layer without further thresholding. Comparing to the ground-truth segmentation ('GT Seg'), the soft segmentation obtains the best segmentation quality. The hard segmentation results lose some details due to the thresholding operation. Since we only require foreground/background probabilities for 6-DOF pose optimization, it is not necessary to use the determined hard labels as in the segmentation tasks. As a result, we choose to use the soft segmentation in the following pose optimization step. It is also very interesting that our segmentation network is able to predict the full mask not only under partial occlusions (1st and 2nd rows), but also under very strong occlusions (in which almost all of the target object is invisible, 3rd and 4th rows). As discussed in previous sections, the strong temporal and geometrical priors have made it possible to see through the occluders even under extreme cases. This advantage would greatly extend the application of 6-DOF object tracking, since to the best of our knowledge, previous methods could not handle this kind of strong occlusions. It is also clear that the segmentation results without the previous mask are very poor, and could not see through the occluders. The results using L1 loss are also worse than that of using Cross Entropy loss. Therefore, we believe using the soft probabilities and the Cross Entropy loss is the best choice for our task.

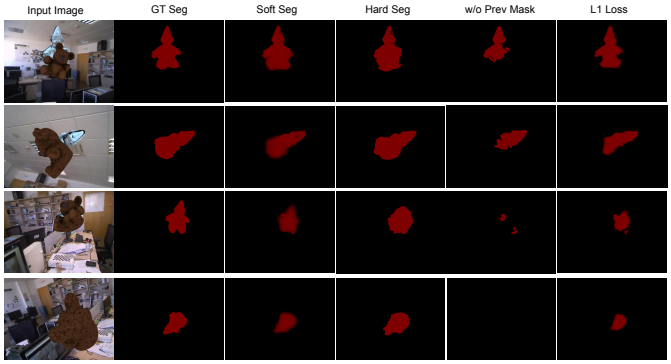

Fig. 5: Comparison of the segmentation results using different inputs and loss functions.

C. Evaluation on Rigid Pose Dataset

Next, we evaluate the overall tracking performance on the Rigid Pose Dataset [35]. We use the same evaluation metric as in [35]. When tracking is lost, the tracker is reset to the ground truth. We measure the tracking success rate (SR) throughout the entire sequence, which is defined as the proportion of frames that are successfully tracked (in %).

We compare our method with the other 4 state-of-art methods: (1) PWP3D [14], a region-based 6-DOF object tracking method; (2) Region Tracker [32], an improved version of PWP3D; (3) Direct Tracker [10], a 6-DOF tracker based on direct model alignment; (4) Hybrid Tracker [19], a hybrid 6-DOF object tracker which combines statistical constraints and photometric constraints together. Since this paper mainly focuses on handling heavy occlusions in monocular 6-DOF object tracking, we consider and discuss the performance on the original+noisy sequences and the performance on occluded sequences separately.

1) The original and noisy sequences: The evaluation results on original and noisy sequences of Rigid Pose Dataset are summarized in Table II. The proposed method obtains comparable results with the state-of-arts. Specifically, our method performs very well for the two most poor-textured objects (edge and clown), because the lack of texture makes it easier to precisely segment the target from the background. Here our method (as well as PWP3D) performs not very well for 'soda', because it is a symmetrical object (cylinder). The reason is that region-based methods are inherently not suitable for tracking the pose of symmetrical objects, due to the fact that they only rely on the object contour, which could be ambiguous in the case of symmetrical objects. This problem could be partly solved by incorporating a photometric term in the energy function, as proposed in the Hybrid Tracker [19]. It is also suggested to narrow the range of rotation around the axis

TABLE II: Evaluation results on original and noisy sequences of Rigid Pose Dataset. (Tracking Success Rate in %, best scores are in bold.)

Method	soda		soup		clown		candy		cube		edge		average
	orig	noisy	orig	noisy	orig	noisy	orig	noisy	orig	noisy	orig	noisy	
[14]	84.4	84.4	96.2	95.5	96.4	89.4	83.6	83.9	83.9	74.3	84.6	83.7	86.7
[32]	96.0	**97.0**	95.0	**98.0**	95.0	96.0	94.0	**96.0**	92.0	93.0	93.0	93.0	94.8
[10]	94.0	89.8	96.1	95.9	98.0	94.9	94.2	90.8	**98.0**	96.2	94.7	91.3	94.5
[19]	**96.4**	95.2	**98.3**	**98.0**	98.3	96.9	**94.7**	94.5	97.1	96.1	96.1	95.7	**96.5**
Ours	87.0	87.7	97.3	97.3	**98.6**	**98.6**	92.7	92.7	97.3	**97.1**	**97.8**	**98.0**	95.2

TABLE III: Evaluation results on occluded sequences of Rigid Pose Dataset. (Tracking Success Rate in %, best scores are in bold.)

Method	soda	soup	clown	candy	cube	edge	average
[14]	44.0	44.0	43.8	39.2	37.8	38.9	41.3
[32]	73.0	82.0	81.0	84.0	76.0	64.0	76.7
[10]	**76.3**	86.7	84.1	80.9	80.6	68.1	79.5
[19]	75.1	79.0	84.1	83.6	73.0	66.9	77.0
Ours	75.8	**88.9**	**89.6**	**87.9**	**92.3**	**95.6**	**88.4**

of symmetry when generating training samples in order to deal with symmetrical objects [36], which might be an useful technique when constructing an end-to-end trainable pipeline.

2) The occluded sequences: We aim to see through the occluders and handle heavy occlusions in this paper, so more importantly, we want to evaluate the performance of our method on the occluded sequences. The results on the occluded sequences of Rigid Pose Dataset are demonstrated in Table III. The proposed method performs significantly better than the previous state-of-art methods, which proves the effectiveness of the proposed mutual guidance pipeline. Specifically, Our method obtains the best SR scores on 5 out of the 6 occluded videos (the only exception is the symmetrical 'soda' object) and improves the average SR score by 8.9%. In this paper, we have provided a possible solution for these extreme cases. When the occlusion is too strong, it is impossible to successfully track the object only based on the very small non-occluded region. Therefore, we propose to see through the occluder to interpret the extra underlying information behind it, with the help of the model-guided video object segmentation network. In our opinion, the improvement in these extreme cases could extend the application of 6-DOF tracking in more complex and difficult situations.

We also note that if hard segmentation results are used in the pose estimation step, the performance drops by a large margin. As shown in Table IV, the average SR scores when using hard segmentation decrease by more than 5% compared to using soft probabilities.

TABLE IV: Comparison of using soft segmentation and hard segmentation on Rigid Pose Dataset. (Tracking Success Rate in %, best scores are in bold.)

Method	original + noisy	occluded
Hard Segmentation	90.1	82.5
Soft Segmentation	**95.2**	**88.4**

D. Evaluation on ACCV14 Dataset

In order to show the performance of our method on real data, we conduct another set of experiments on the ACCV14 Dataset [38]. The ACCV14 Dataset is a real-world RGB-D dataset featuring heavy occlusions. Here we only use RGB images for evaluation. We measure accuracy as in [38] as the fraction of frames where the object pose was correctly tracked. We compare the performance of our method with the state-of-art RGB method [19] and two RGB-D methods [37], [38] in Table V. While using only monocular information, the proposed method achieves much better accuracy than [19] and [37]. However, our method is still not as accurate as [38], which shows the gap between monocular methods and RGB-D methods in dealing with real data. The results also demonstrate the effectiveness of our fine-tuning step, which contributes to the domain adaptation from synthetic training data to real video sequence.

TABLE V: Evaluation results on ACCV14 Dataset. (Accuracy in %, best scores are in bold, second best scores are underlined. Ours-f: Our method without fine-tuning.)

Method	Input	Cat1	Cat2	Sam1	Sam2	Tool1	Tool2	average
[19]	RGB	86.0	82.7	79.4	80.6	78.7	74.9	80.4
[37]	RGB-D	66.8	44.2	72.0	33.7	54.7	59.4	58.9
[38]	RGB-D	**100.0**	**99.4**	**96.3**	**92.3**	**88.8**	**100.0**	**96.2**
Ours-f	RGB	89.5	87.2	91.0	88.5	81.8	89.1	87.9
Ours	RGB	<u>93.5</u>	<u>91.3</u>	<u>93.3</u>	<u>90.5</u>	<u>85.1</u>	<u>92.9</u>	<u>91.1</u>

E. Discussions

Although the proposed method achieves competitive or even superior performance compared to the state-of-art methods, it still has some limitations. Firstly, although our method obtains superior performance in handling occlusions, the results in the original and noisy sequences of Rigid Pose Dataset show that the performance of our method is only similar to those of the compared methods. The results on ACCV14 Dataset also show that our method is not as accurate as the method utilizing depth information. Part of the reason is that the encoder-decoder architecture in the segmentation network produces relatively low-resolution segmentation results (which are upsampled by 4 to meet the input resolution [25]). This is a common problem in learning-based methods, but we will try to further improve the precision in our future works. Secondly, for now, a specific segmentation model is trained for each object, which lacks generalization ability to work for new objects. Considering that the pose estimation stage is universal to all objects, it would be better if we could train a general segmentation model that is able to work for more than one object (e.g. for a category of objects, or even for universal objects).

V. Conclusion

In this paper, we have presented a robust 6-DOF object pose tracker by seeing through the occluders. The proposed tracker is able to handle very heavy occlusions in which previous methods would fail. The key is to form a mutual guidance loop of the video object segmentation stage and the

6-DOF pose estimation stage, and to combine the learning-based and optimization-based methods in a complementary manner. Experiments have shown that our method could achieve competitive performance on non-occluded sequences and significantly better robustness on occluded sequences. We believe the improved performance of our method in heavy occlusion cases could help to extend the application of 6-DOF object tracking in more complex situations.

REFERENCES

[1] V. Lepetit and P. Fua, *Monocular model-based 3D tracking of rigid objects*. Now Publishers Inc, 2005.

[2] K. Eckenhoff, Y. Yang, P. Geneva, and G. Huang, "Tightly-coupled visual-inertial localization and 3-d rigid-body target tracking," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1541–1548, 2019.

[3] K. M. Judd and J. D. Gammell, "The oxford multimotion dataset: Multiple se (3) motions with ground truth," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 800–807, 2019.

[4] C. Choi and H. I. Christensen, "Real-time 3d model-based tracking using edge and keypoint features for robotic manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2010, pp. 4048–4055.

[5] L. Chen, F. Zhou, Y. Shen, X. Tian, H. Ling, and Y. Chen, "Illumination insensitive efficient second-order minimization for planar object tracking," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017.

[6] A. Petit, E. Marchand, and K. Kanani, "A robust model-based tracker combining geometrical and color edge information," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2013, pp. 3719–3724.

[7] A. Loesch, S. Bourgeois, V. Gay-Bellile, and M. Dhome, "Generic edgelet-based tracking of 3d objects in real-time," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 6059–6066.

[8] J. P. Lima, F. Simões, L. Figueiredo, and J. Kelner, "Model based markerless 3d tracking applied to augmented reality," *Journal on 3D Interactive Systems*, vol. 1, 2010.

[9] Y. Park, V. Lepetit, and W. Woo, "Multiple 3d object tracking for augmented reality," in *IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, 2008, pp. 117–120.

[10] L. Zhong, M. Lu, and L. Zhang, "A direct 3d object tracking method based on dynamic textured model rendering and extended dense feature fields," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2302–2315, 2018.

[11] B. Wang, F. Zhong, and X. Qin, "Pose optimization in edge distance field for textureless 3d object tracking," in *Computer Graphics International Conference*. ACM, 2017, p. 32.

[12] K. Li and J. Malik, "Amodal instance segmentation," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 677–693.

[13] D. J. Tan, N. Navab, and F. Tombari, "Looking beyond the simple scenarios: Combining learners and optimizers in 3d temporal tracking," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 11, pp. 2399–2409, 2017.

[14] V. A. Prisacariu and I. D. Reid, "Pwp3d: Real-time segmentation and tracking of 3d objects," *International Journal of Computer Vision*, vol. 98, no. 3, pp. 335–354, 2012.

[15] H. Tjaden, U. Schwanecke, and E. Schömer, "Real-time monocular segmentation and pose tracking of multiple objects," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 423–438.

[16] J. Hexner and R. R. Hagege, "2d-3d pose estimation of heterogeneous objects using a region based approach," *International Journal of Computer Vision*, vol. 118, no. 1, pp. 95–112, 2016.

[17] H. Tjaden, U. Schwanecke, and E. Schömer, "Real-time monocular pose estimation of 3d objects using temporally consistent local color histograms," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 124–132.

[18] H. Tjaden, U. Schwanecke, E. Schömer, and D. Cremers, "A region-based gauss-newton approach to real-time monocular multiple object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1797–1812, 2019.

[19] L. Zhong and L. Zhang, "A robust monocular 3d object tracking method combining statistical and photometric constraints," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 973–992, 2019.

[20] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "Deepim: Deep iterative matching for 6d pose estimation," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 683–698.

[21] F. Manhardt, W. Kehl, N. Navab, and F. Tombari, "Deep model-based 6d pose refinement in rgb," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 800–815.

[22] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[25] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[26] L. Qi, L. Jiang, S. Liu, X. Shen, and J. Jia, "Amodal instance segmentation with kins dataset," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3014–3023.

[27] Y. Zhu, Y. Tian, D. Metaxas, and P. Dollár, "Semantic amodal segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1464–1472.

[28] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 221–230.

[29] K.-K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "Video object segmentation without temporal information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 6, pp. 1515–1530, 2018.

[30] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2663–2672.

[31] X. Li, Y. Qi, Z. Wang, K. Chen, Z. Liu, J. Shi, P. Luo, X. Tang, and C. C. Loy, "Video object segmentation with re-identification," *arXiv preprint arXiv:1708.00197*, 2017.

[32] S. Zhao, L. Wang, W. Sui, H.-y. Wu, and C. Pan, "3d object tracking via boundary constrained region-based model," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 486–490.

[33] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[34] 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 (ACCV)*. Springer, 2012, pp. 548–562.

[35] K. Pauwels, L. Rubio, J. Diaz, and E. Ros, "Real-time model-based rigid object pose estimation and tracking combining dense and sparse visual cues," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2347–2354.

[36] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3828–3836.

[37] 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 (ECCV)*. Springer, 2014, pp. 536–551.

[38] A. Krull, F. Michel, E. Brachmann, S. Gumhold, S. Ihrke, and C. Rother, "6-dof model based tracking via object coordinate regression," in *Asian Conference on Computer Vision (ACCV)*. Springer, 2014, pp. 384–399.