Rethinking Amodal Video Segmentation from Learning Supervised Signals with Object-centric Representation

Ke Fan¹†, Jingshi Lei¹†, Xuelin Qian¹∗, Miaopeng Yu¹, Tianjun Xiao²∗, Tong He²
Zheng Zhang², Yanwei Fu¹
¹Fudan University ²Amazon Web Service
{kfan21, jslei21}@m.fudan.edu.cn, {xlqian, mpyu19, yanweifu}@fudan.edu.cn
{tianjux, htong, zhaz}@amazon.com

Abstract

Video amodal segmentation is a particularly challenging task in computer vision, which requires to deduce the full shape of an object from the visible parts of it. Recently, some studies have achieved promising performance by using motion flow to integrate information across frames under a self-supervised setting. However, motion flow has a clear limitation by the two factors of moving cameras and object deformation. This paper presents a rethinking to previous works. We particularly leverage the supervised signals with object-centric representation in real-world scenarios. The underlying idea is the supervision signal of the specific object and the features from different views can mutually benefit the deduction of the full mask in any specific frame. We thus propose an Efficient object-centric Representation amodal Segmentation (EoRaS). Specifically, beyond solely relying on supervision signals, we design a translation module to project image features into the Bird’s-Eye View (BEV), which introduces 3D information to improve current feature quality. Furthermore, we propose a multi-view fusion layer based temporal module which is equipped with a set of object slots and interacts with features from different views by attention mechanism to fulfill sufficient object representation completion. As a result, the full mask of the object can be decoded from image features updated by object slots. Extensive experiments on both real-world and synthetic benchmarks demonstrate the superiority of our proposed method, achieving state-of-the-art performance. Our code will be released at https://github.com/kfan21/EoRaS.

1. Introduction

Deep learning has demonstrated remarkable success in various computer vision tasks. Nevertheless, neural net-works are limited to learning visible patterns in the data, and are typically challenged in reasoning about the broader and unseen components. Currently, most researches in object detection and segmentation tasks concentrate on enhancing the visible part’s performance, leaving few studies on inferring occluded information. Conversely, humans possess an innate ability to imagine and extrapolate, enabling us to easily complete an occluded part of an image based on prior knowledge. This critical capacity is instrumental in advanced deep learning models for real-world scenarios, such as medical diagnosis and autonomous driving. Thereby, the central issue addressed in this paper is the video amodal segmentation task, which aims to deduce an object’s complete mask, whether it is partially obscured or not.

Prior studies on image amodal segmentation [22, 30, 32]...
are over-reliance on prior knowledge, which actually hampers the model’s generalization abilities, resulting in limited improvements under complex circumstances. For video amodal, Yao et al. [35] proposed that the occluded part of the current frame may appear in other frames, and therefore, information from all frames should be collected to fill in the occluded regions of any specific frame. While this method achieves promising results under the self-supervised setting, it fails when camera motion exists, as 2D warping is used to make connections within different frames, leading to distorted signals.

This paper aims to propose a better approach for video amodal segmentation by rethinking the importance of using supervised signals with object-centric representation. Such object-centric representations reflect the compositional nature of our world, and potentially facilitate supporting more complex tasks like reasoning about relations between objects. While signals such as motion flow and shape priors have shown promising results, they are limited by moving cameras and complicated object types respectively. In contrast, recent advances [6, 13, 21] in video object segmentation produce highly accurate object masks that are less sensitive to moving cameras, making them better suited as supervision signals. Surprisingly, such video object masks have not been fully exploited before.

To this end, we propose a novel approach that learns video amodal segmentation not only from observed object supervision signals in the current frame (shape prior) but also from integrated information of object features under different views (view prior). Our motivation is clearly shown in Fig. 1. By using visual patterns of other views to explain away occluded object parts in the current frame [31], our approach gets rid of optical flow and eliminates the shortcomings of mere reliance on shape priors. Our model is highly effective, even in complex scenarios.

In this paper, we propose a novel supervised method for the video amodal segmentation task that leverages a multi-view fusion layer based temporal module and a Bird’s-Eye View (BEV) feature translation network. Rather than relying on warping the amodal prediction into the next frame using optical flow or using shape priors alone, we enhance the current frame features by incorporating feature information from different viewpoints and leveraging the supervision signals simultaneously. Specifically, we first extract front-view features from the videos using FPN50 [18]. Then, we employ a translation network to transform these front-view features into bird’s-eye view features, which bring in 3D information through the usage of the intrinsic matrix. In contrast to some related work [28] extracting object-centric 3D representation by object reconstruction, the acquisition of BEV feature is simpler, faster, and easier to train. As each frame is equivalent to a unique view, features from both different frames and the BEV space, which carry shape information about the occluded part, are further utilized. We repurpose the vanilla object-centric representations [19] – object slots to integrate those information, which is accomplished by our novel multi-view fusion layer. Finally, we refine the front-view features using the updated object slots containing object information from multiple views and decode the full mask. Compared to previous methods [35], our model can handle scenarios with 3D viewing angle changes or complex object shapes better by leveraging shape knowledge and integrating information across multiple views simultaneously. To evaluate our method, we conduct extensive experiments on real-world and synthetic amodal benchmarks. The results demonstrate that our model achieves outstanding performance compared to comparable models and effectively demonstrates the efficacy of our architecture.

In summary, our main contributions are listed below. (1) Our contribution lies in formulating the video amodal segmentation task using supervised signals for the first time. Our model efficiently learns the shape and view priors, enabling it to handle complex scenarios with ease. (2) We propose a novel approach to learning object-centric representations through a multi-view fusion layer based temporal module equipped with a set of object slots, which achieves significant improvement in the correlation of information from different views. (3) We introduce the novel concept of bird’s-eye view features in our amodal task, which provides front-view features with 3D information, resulting in consistent benefits. (4) By utilizing the bird’s-eye view generator and multi-view fusion layer based temporal module, our algorithm achieves remarkable improvement on both real-world and synthetic amodal benchmarks, highlighting the novelty of our approach.

2. Related Work

Amodal segmentation is a more challenging task than instance segmentation because it requires predicting the full shape of occluded objects through the visible parts. While previous literature has focused on using shape priors effectively through multi-level coding [22], variational autoencoder [14], shape prior memory codebook construction [32], mixing feature decoupling [16] or Bayesian model [29], relying solely on shape priors can lead to poor empirical performance due to distribution shifts between training data and real scenarios. To address this issue, [35] leverages spatiotemporal consistency and dense object motion to explain away occlusion. Although their work has made progress in video amodal segmentation, optical flow can cause object deformation in the presence of camera motion. In contrast, our proposed architecture introduces a novel approach that does not require optical flow and utilizes bird’s-eye view features to bring in 3D information that enhances the learning of front-view features.
Object-centric learning aims at identifying all the objects from raw input for better understanding the complex scenes. Existing object-centric learning methods can be categorized into unsupervised and supervised methods. While unsupervised methods use image/scene reconstruction to extract object representations from images/scenes [2, 19, 25], supervised methods represent each object as a query embedding and pay much attention to obtaining a great initialization [3, 4, 7, 9, 13, 34]. Our EoRaS is more related to the supervised method in terms of constructing a set of learnable queries as an information container.

BEV maps generation requires to generate semantic maps in bird’s-eye view space. Due to a lack of high-quality annotated data, most of the early work adopts weak supervision by utilizing stereo information [20, 21] or obtaining pseudo label [27]. Others directly translate semantic segmentation maps from image space into bird’s-eye view space [6, 26]. With the advent of large-scale annotated datasets, research on supervised methods has also made some progress. [23] and [24] respectively take advantage of dense transformer layer and 1D sequence-to-sequence translation to learn a map representation. [1] and [17] instead blend features from multi-camera images to construct BEV map. In our EoRaS, the bird’s-eye view feature is utilized to integrate 3D information into the front-view feature. To the best of our knowledge, it’s the first attempt to incorporate the BEV translation module in the amodal segmentation task.

3. Methodology

This paper focuses on the video amodal segmentation task. Specifically, given a video sequence \( \{I_t\}_{t=1}^T \) with \( K \) objects, EoRaS aims to predict the full mask \( \{M^k_t\} \) of each object in all frames, where \( k \) is the index of objects. In our EoRaS, visible masks \( \{V^k_t\} \) also serves as supervision but will not be utilized at the test phase.

3.1. Architecture

The overall architecture of EoRaS is shown in Figure 2. Our EoRaS is mainly comprised of four modules: (i) the feature encoding module which extracts the front-view feature \( f^k_t \) from the input frames; (ii) the BEV translation network which converts the front-view features into bird’s-eye view angle \( b^k_t \) using the camera intrinsic matrix \( K \) and neural network; (iii) the multi-view fusion layer based temporal module which utilizes the object slots updated through the forward and backward streams to integrate the feature information from different views and fulfill the completion of each front-view feature; and (iv) the deconvolution network that estimates the full masks and visible masks of the current frame simultaneously.

**Feature Encoding Module** In this module, FPN50 [18] pretrained on ImageNet [5] is used to extract features from the input frames. These features are obtained from a frontal perspective and capture a lot of information but will fail to make inferences about the missing parts of the objects.

\[
f^k_t = FPN(I^k_t)
\] (1)

**BEV Translation Network** The features from the bird’s-eye view (BEV) are widely used and work well in autonomous driving research. Recall that features from different perspectives are likely to contain the missing part information and contribute to the full mask deduction of the
current frame. As obstruct doesn’t exist in BEV space unless objects are stacked on top of each other, it is reasonable to introduce BEV as a special perspective to promote the completion of the front-view feature. Consider a horizontally placed camera, for each frame, a 3D volume feature \( V_{3D} \) is constructed in the camera coordinate system. As the BEV feature generation just involves the current frame, we omit the subscript of time \( t \) and the object index \( k \) for simplicity.

Denote the camera intrinsic matrices as \( K \), we first focus on a single point \((x, y, z)\) in the camera coordinate space. By utilizing intrinsic matrices, this point can be easily projected into the image/feature plane and we denote its coordinate as \((u, v)\):

\[
\begin{pmatrix}
\lambda u \\
\lambda v
\end{pmatrix} = K \begin{pmatrix}
x \\
y \\
z
\end{pmatrix}
\tag{2}
\]

We use bilinear interpolate to obtain the feature at \((u, v)\) from the corresponding front-view feature \( f \). The obtained value at \((u, v)\) will act as the volume feature at \((x, y, z)\).

As shown in Figure 3a, the 3D volume in the camera coordinate will be rasterized into a group of points \( p_{ijk} = (x_i, y_j, z_k) \), where \( 1 \leq i \leq m, 1 \leq j \leq n, 1 \leq k \leq h \) and \( x, y, z \) represent the direction of width, depth, and height, respectively. The value of \( V_{3D} \) will be gotten by simply repeating the above process for each point. Further, by stacking the feature of the volume obtained from different channels together, we will get \( V_{3D} \in \mathbb{R}^{c \times m \times n \times h} \). Since our goal is to acquire BEV features, \( V_{3D} \) is rearranged to \( \mathbb{R}^{h \times m \times c} \) and sent to a lightweight CNN for compression along height dimension:

\[
b_k^t = \text{CNN}(V_{3D}, \text{reshape}(c, m, n))
\tag{3}
\]

**Multi-view Fusion Layer based Temporal Encoder** As the occluded part of a specific view may potentially appear in other frames, we can make full use of the information in each frame (equivalent to different perspectives) to refine the completion of the object shape. Specially, inspired by DETR [3] and Slot Attention [19], we would like to generate an object-centric feature utilizing both front-view and BEV representations.

A direct method is to follow [19], which uses ConvGRU to aggregate temporal information. However, the cost of nested recurrent slot computation to gather the object information from each frame is expensive when processing videos. Here, we propose a more efficient attention-based encoder architecture named Multi-view Fusion Layer. Generally, in such a layer, three \( N \)-layer object attentions which is a non-recurrent variant of slot attention are carefully designed and closely connected. And features from different views and object slots serve as the inputs.

In particular, as shown in Figure 3b, each object attention layer (ObjAttention\((SP, SR)\)) is stacked by self-attention, cross-attention, and feedforward network and serves as information fusion network. The variable absorbing the missing shape information during the fusion process is named shape receiver (SR), and another one is dubbed shape provider (SP) as it offers extra shape patches. The total forward process in ObjAttention is formulated as,

\[
\hat{SR} = SR + \text{Attention}(SR, SR, SR) \tag{4}
\]

\[
\hat{SR} = \hat{SR} + \text{Attention}(SP, \hat{SR}, SP) \tag{5}
\]

\[
\text{output} = \text{MLP}(\hat{SR}) \tag{6}
\]

where \( \text{Attention}(K, Q, V), \text{MLP}(\cdot) \) denotes multi-head attention module and two-layer feedforward network, respectively. And we omit all normalization layers. ObjAttention first enhances the SR representation by renewing information contained in itself, then extracts fresh properties from the SP.
On the other hand, similar to [3], a set of object slots $S_0 \in \mathbb{R}^{n_s \times d}$ is initialized before the videos enter. $n_s$ denotes the number of slots and $d$ is the feature dimension. In our model, $S_0$ is set to be learnable and serves as a container that gathers shape information from various views.

With the above preparations, we now go to the detail of our multi-view fusion layer. For each frame, we take advantage of the $S_{t-1}$ from the last frame which includes object shape information from previous frames, and provide it with the fresh characters from the front-view $f_t^k$ and BEV feature $b_t^k$ under current perspective at first:

$$S_t' = \text{ObjAttention}(SR = S_{t-1}, SP = b_t^k) \quad (7)$$

$$S_t = \text{ObjAttention}(SR = S_t', SP = f_t^k) \quad (8)$$

Then, the updated slots will provide clues about the occluded part and help complete the front-view features of the current frame by setting the front-view features as SR in the object attention layer. Thus, we inversely enhance the front-view feature using the object slots by:

$$\hat{f}_t^k = \text{ObjAttention}(SR = \hat{f}_t^k, SP = S_t) \quad (9)$$

Deconvolution Network The deconvolution network (DeConv) is served as the mask predictor and takes the updated front-view features as input since it shares the same perspective with the full mask to be predicted. In our experiments, we just construct several de-convolutional layers for this module.

$$\hat{M}_t^k, \hat{V}_t^k = \text{DeConv}(\hat{f}_t^k) \quad (10)$$

where $\hat{M}_t^k$ and $\hat{V}_t^k$ are the full and visible mask predictions of the current frame, respectively.

Bi-directional Prediction Cold start problem exists under the above framework since the first few frames may not be informative enough. Thus, backward prediction is added to solve this problem. We simply concatenate the forward and backward features, and send them to the final deconvolution network.

3.2. Loss Function for EoRaS

Our EoRaS is designed as an end-to-end framework and trained with the focal loss (Focal()) using both full mask and visible mask as supervision signals. Note that the discard of the visible mask loss will not heavily damage the model performance, as shown in Tab. 4. The overall loss function is

$$\mathcal{L}_{\text{full}} = \sum_{t=1}^{T} \sum_{k=1}^{K} \text{Focal}(\hat{M}_t^k, M_t^k) \quad (11)$$

$$\mathcal{L}_{\text{vis}} = \sum_{t=1}^{T} \sum_{k=1}^{K} \text{Focal}(\hat{V}_t^k, V_t^k) \quad (12)$$

$$\mathcal{L} = \mathcal{L}_{\text{full}} + \lambda \cdot \mathcal{L}_{\text{vis}}. \quad (13)$$
Figure 5: Qualitative comparison between our EoRaS and competitors. From top to down, the three rows are from Movi-B, Movi-D, and KITTI, respectively.

<table>
<thead>
<tr>
<th>DATASET</th>
<th>METHODS</th>
<th>mIoU_{full}</th>
<th>mIoU_{occ}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movi-B</td>
<td>VM</td>
<td>59.19</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Convex</td>
<td>64.21</td>
<td>18.42</td>
</tr>
<tr>
<td></td>
<td>PCNET [36]</td>
<td>65.79</td>
<td>24.02</td>
</tr>
<tr>
<td></td>
<td>AISFormer [30]</td>
<td>77.34</td>
<td>43.53</td>
</tr>
<tr>
<td></td>
<td>SaVos-sup. [35]</td>
<td>70.72</td>
<td>33.61</td>
</tr>
<tr>
<td></td>
<td>BiLSTM [12]</td>
<td>77.93</td>
<td>46.21</td>
</tr>
<tr>
<td></td>
<td>EoRaS (Ours)</td>
<td>79.22</td>
<td>47.89</td>
</tr>
<tr>
<td>Movi-D</td>
<td>VM</td>
<td>56.92</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Convex</td>
<td>60.18</td>
<td>16.48</td>
</tr>
<tr>
<td></td>
<td>PCNET [36]</td>
<td>64.35</td>
<td>27.31</td>
</tr>
<tr>
<td></td>
<td>AISFormer [30]</td>
<td>67.72</td>
<td>33.65</td>
</tr>
<tr>
<td></td>
<td>SaVos-sup. [35]</td>
<td>60.61</td>
<td>22.64</td>
</tr>
<tr>
<td></td>
<td>BiLSTM [12]</td>
<td>68.43</td>
<td>36.00</td>
</tr>
<tr>
<td></td>
<td>EoRaS (Ours)</td>
<td>69.44</td>
<td>36.96</td>
</tr>
<tr>
<td>KITTI</td>
<td>VM</td>
<td>74.75</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Convex</td>
<td>78.62</td>
<td>8.29</td>
</tr>
<tr>
<td></td>
<td>PCNET [36]</td>
<td>81.58</td>
<td>17.90</td>
</tr>
<tr>
<td></td>
<td>AISFormer [30]</td>
<td>86.42</td>
<td>51.04</td>
</tr>
<tr>
<td></td>
<td>SaVos-sup. [35]</td>
<td>83.09</td>
<td>37.33</td>
</tr>
<tr>
<td></td>
<td>BiLSTM [12]</td>
<td>86.68</td>
<td>49.95</td>
</tr>
<tr>
<td></td>
<td>EoRaS (Ours)</td>
<td>87.07</td>
<td>52.00</td>
</tr>
</tbody>
</table>

Table 1: The performance of EoRaS on real-world and synthetic video amodal benchmarks.

the ground truth visible mask as amodal prediction; (2) **Convex**, take the convex hull of the visible mask as the amodal mask; (3) **PCNET** [36], a self-supervised image-level amodal completion method by in turn recovering occlusion ordering and completing amodal masks and content; (4) **AISFormer** [30], an image-level amodal segmentation model equipped with a transformer-based mask head and achieves the new state-of-the-art recently; (5) **SaVos** [35], a recent state-of-the-art method in the field of self-supervised video amodal segmentation and is modified to supervised version by removing the 2D warping and bringing the supervised signal for fair comparison (We also did additional experiments involving warping operation, but the experiment results are quite inferior); (6) **BiLSTM** [12], a variant of our proposed method for which we keep the same FPN50 backbone but utilize BiLSTM to aggregate temporal information across frames.

**Implementations** Results on all datasets are reported in terms of mIoU metrics for both full mask and occluded regions. Since most amodal segmentation algorithms use the visible mask or the bounding boxes of the visible part as model input, the estimation results of the visible area may be more confident, and the mIoU of the occluded area can better reflect the model performance. On all datasets, the mIoU metric of the occluded part is only computed on those partially occluded objects. We use AdamW as optimizer with batch size 4 for 50 epochs. The learning rate is set to $1e^{-5}$ on Movi datasets and $1e^{-4}$ on the KITTI dataset. Exponential learning rate decay is used where the decay rate is 0.95. The weight decay is $5e^{-4}$. And the $\gamma$ in focal loss is set to 2. We set $\lambda = 1$, $n_s = 8$, $N = 2$ and train our model on four Tesla T4 GPUs using PyTorch.

### 4.2. Results on Movi Datasets

As shown in Table 1, compared with supervised SaVos, our EoRaS achieves extremely significant performance improvements on both Movi datasets. In particular, by applying our algorithm, the prediction of the full mask of the objects in the two datasets is improved by 8.50% and 8.83%, respectively. The improvements are more remarkable in the prediction of occluded parts. For the performance on
Table 2: Ablation study of our temporal and bev modules on Movi-B (left) and Movi-D (right) dataset.

<table>
<thead>
<tr>
<th>No.</th>
<th>Designs</th>
<th>Metrics</th>
<th>No.</th>
<th>Designs</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Temporal</td>
<td>Bi-direction</td>
<td>BEV</td>
<td>mIoU&lt;sub&gt;full&lt;/sub&gt;</td>
<td>mIoU&lt;sub&gt;occ&lt;/sub&gt;</td>
</tr>
<tr>
<td>1</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>76.93</td>
<td>44.55</td>
</tr>
<tr>
<td>2</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>78.66</td>
<td>46.83</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>78.42</td>
<td>46.51</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>79.22</td>
<td>47.89</td>
</tr>
</tbody>
</table>

4.3. Results on KITTI Dataset

The experiment results on the KITTI dataset are shown in Table 1. For objects in real scenes, our EoRaS can still exceed all the current state-of-the-art methods. Compared with the image-level baseline, we achieve 0.65% and 0.96% improvement for the full and occluded mask prediction, respectively. For the supervised SaVos, EoRaS achieves enormous promotion, ~4% on the full shape and ~15% on the missing part. Furthermore, other video-level baselines consistently underperform our EoRaS by ~2% on the deduction of the occluded part. Qualitative comparison in Figure 5 clearly exhibits the great precision of EoRaS. The above evidence is sufficient enough to prove the effectiveness of EoRaS under weakly supervised settings.

5. Further Analysis

Effectiveness of Temporal and BEV Modules As shown in Table 2, on Movi-B dataset, our temporal module brings about ~2.3% performance improvement in occluded part prediction. After plugging in the BEV module, the occluded mIoU is further improved by 1.06%. Additionally, Bi-direction prediction also plays an important role in our model as it brings 1.38% performance improvement for the missing part deduction. On the Movi-D dataset, the improvements brought in by those modules are also significant, as presented in the right table. Some visualizations derived from different architectures are presented in Figure 6. It’s clear that both temporal and BEV modules are capable of improving the smoothness and shape similarity of full masks. These experiments fully prove the effectiveness of the modules proposed in this paper, and also verify the correctness of our hypothesis that feature information from different perspectives can benefit the completion of object shape in any specific frame/view.

Sensitivity Analysis of Slot Number To analyze the sensi-
Different choices of λ We conduct experiments to analyze the effect of λ on the performance of our model, and the results are presented in Table 4. First of all, the utilization of visible masks in supervision signals will benefit the model training as also shown in previous amodal segmentation algorithms. But the way that EoRaS differs lies in the insensitivity to the choice of λ once the visible mask is added, which demonstrates the superiority of EoRaS.

Open Set Segmentation To evaluate the capacity of out-of-distribution generalization, we conduct open set segmentation experiments on Movi-D and KITTI datasets. Models are pretrained on the relatively simple Movi-B dataset. As presented in Table 5, EoRaS achieves the best accuracy among all competitors. Concretely, compared with supervised SaVos, EoRaS outperforms by at least 6%, showing strong dominance. Again, the image-level SOTA algorithm underperforms EoRaS by ~2% on the occluded part deduction, indicating that the integration of information from different views indeed benefits the generalization ability.

Test-time Assistance by Ground Truth Visible Mask (GTVM) The same as SaVos and PCNET, we explore the utilization of GTVM at the test phase. On the one hand, the post-processing (PP), including taking the intersection of the predicted full mask and GTVM, is feasible. On the other hand, containing partial shape information, GTVM may be capable of serving as a shape guidance (SG) for mask completion. To this end, we simply train our model with the concatenation of images and visible masks. The experimental results are presented in Table 6. Overall, the introduction of GTVM brings in huge benefits, which is inline with [35, 36]. Despite the usage of GTVM in those algorithms, our EoRaS still outperforms them by a large margin (see Table 1), suggesting the powerful function.

6. Conclusion

In this paper, we proposed a brand-new pipeline named EoRaS to cope with the video amodal segmentation task. Based on the assumption that both the supervision signals (shape prior) and the features from different perspectives (view prior) will benefit the deduction of the full mask under any specific view, the multi-view fusion layer based temporal encoder and BEV translation network are designed to integrate 3D information and front-view shape patches from different frames respectively in an object-centric pattern. Utilizing those modules, our EoRaS eliminates the optical flow usage and the over-reliance on shape priors, achieving high efficiency even in complex scenarios. We conduct experiments on both real-world and synthetic video amodal benchmarks, including Movi-B, Movi-D, and KITTI datasets. The empirical results demonstrate that our EoRaS achieves the new state-of-the-art performance.


