Multi-view Self-supervised Object Segmentation

Wenxuan Ma[1], Liming Zheng[1], Yinghao Cai[2,*], Tao Lu[2], Shuo Wang[2]

***Abstract*— Robots often operate in open-world environments, where the capability to generalize to new scenarios is crucial for robotic applications such as navigation and manipulation. In this paper, we propose a novel multi-view self-supervised framework (MVSS) to adapt off-the-shelf segmentation methods in a self-supervised manner by leveraging multi-view consistency. Pixel-level and object-level correspondences are established through unsupervised camera pose estimation and cross-frame object association to learn feature embeddings that the same object are close to each other and embeddings from different objects are separated. Experimental results show that it only needs to observe the RGB-D sequence once without any annotation, our proposed method is able to adapt existing methods in new scenarios to achieve performance close to that of supervised segmentation methods.**

I. INTRODUCTION

Robots often operate in open-world environments, where the capability to discover and segment novel objects is crucial for robotic tasks such as grasping, manipulation, and navigation. As a fundamental task in robotics and computer vision, object segmentation has been explored for many years. Although deep learning-based approaches [1]–[4] have achieved great progress in object segmentation, they often require large amounts of labeled data, which is both time-consuming and expensive to obtain. Moreover, real-world scenarios often involve multiple objects arranged and placed in various ways, the performance of supervised object segmentation may suffer in real-world scenarios.

In recent years, self-supervised learning has been applied to various computer vision tasks such as image classification, object detection, semantic segmentation, and depth estimation [5]–[7]. With reasonable prior knowledge as self-supervised cues, self-supervised learning methods can significantly reduce the need for labeled data and improve the model generalization. As we know, there are a lot of natural laws in the world, such as the motion coherence [8], [9], spatio-temporal consistency [10]–[12] and photometric consistency [7]. By leveraging these self-supervised cues, the learning process could be much more data efficient and cost effective.

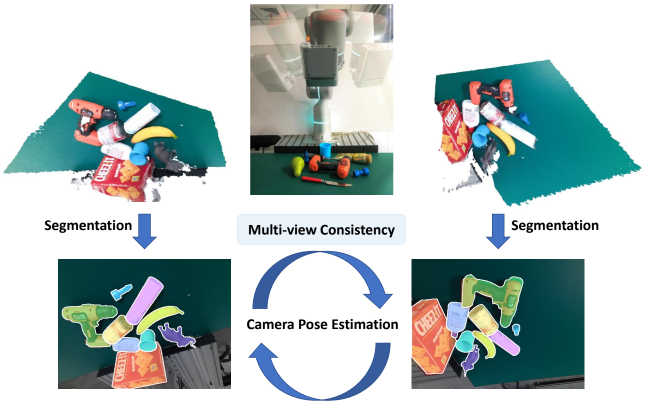

Fig. 1. Illustration of multi-view consistency. We expect the segmentation results of the model to be consistent under multi-view observations. Pixel-level and object-level correspondences are established through unsupervised camera pose estimation and cross-frame object association to learn feature embeddings that the same object are close to each other and embeddings from different objects are separated.

In this paper, we propose a method to self-supervisedly improve object segmentation performance through multi-view observations. Due to the irregularity and diversity of object stacking in unstructured scenes, the observation of target objects may be affected by severe occlusion and observation angle differences. Observation differences from different perspectives will lead to large changes in the extracted feature representations. However, when humans observe a scene from multiple perspectives, they usually naturally associate the observations from multiple perspectives, and thus continuously learn to establish a consistent semantic relationship. As shown in Fig. 1, our approach incorporates the concept of multi-view consistency as the constraint in self-supervised learning. We assume that multiple views of the same object should produce similar segmentation results and feature embeddings. We use this assumption as a regularization term to ensure consistency in the generated segmentation results. In order to leverage the information contained in multi-view observation, it is necessary to establish the pixel-level and object-level correspondences from different perspectives. Many methods such as optical flow estimation [13], [14] can be used here. However, optical flow based methods may suffer under large motions. In order to obtain more accurate correspondence, we introduce an unsupervised pose estimation network to obtain relative poses between different views, which does not require ground truth during training

*This work is supported in part by the National Key Research and Development Program of China under Grant 2021ZD0114505 and the National Natural Science Foundation of China under Grant U1913201 and 62273342.

[1] Wenxuan Ma and Liming Zheng are with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 101408, China. mawenxuan2021@ia.ac.cn

[2] Yinghao Cai, Tao Lu and Shuo Wang are with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. Yinghao Cai is also with the Centre for Artificial Intelligence and Robotics (CAIR), the Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences (HKISI-CAS). Corresponding author: Yinghao Cai. yinghao.cai@ia.ac.cn

and can be continuously optimized during observation.

After obtaining the relative poses, we then re-project the segmentation masks across views and perform ROI matching for object association to establish object-level correspondence, in which matching pairs are considered robust objects, and unmatched pairs are mis-segments. The object matching pairs obtained in this way are spatially close and semantically similar. Through this process, we obtain robust object-level correspondence to learn feature embeddings for segmentation. This provides robust positive and negative samples for self-supervised contrastive learning. Therefore, we introduce a multi-view contrastive loss that encourages the feature embeddings of matching object pairs to be as close as possible between different frames, while keeping the embeddings of unmatched pairs separated. Despite the potential noisy pseudo-labels, our network can still learn discriminative features through multi-view observations, resulting in improved segmentation performance in a self-supervised manner.

Experimental results demonstrate that the proposed multi-view self-supervised (MVSS) framework significantly improves object segmentation performance without any human involvement. We believe that our approach has the great potential to adapt the performance of object segmentation methods in new scenarios while reducing the need for labeled data in various applications.

II. RELATED WORK

A. Object Segmentation

Object segmentation refers to the process of extracting object masks from input images. In recent years, there has been remarkable progress in object segmentation. Top-down methods detect the objects first and then segment them to obtain masks [1], [15]–[17], while bottom-up methods cluster pixels into segments without relying on object proposals [18]–[20].

Object segmentation provides high-level visual perception capacity for robotic manipulation tasks. The ability to segment previously unseen objects is crucial in unstructured robotic manipulation scenarios where robots may encounter various environments and objects. Xie et al. [21] first generate initial segmentation masks from depth information, and then use RGB information for refinement. Xie et al. [22] further utilize both RGB and depth images to generate pixel-level feature embeddings. Clustering is then performed to achieve object segmentation. The network in [22] is trained with metric learning in an end-to-end manner. Back et al. [23] presents unseen object amodal instance segmentation (UOAIS) for robotic manipulation in cluttered scenes. A Hierarchical Occlusion Modeling (HOM) scheme is proposed in [23] to reason about the occlusions. While these methods are demonstrated to be effective in unseen object segmentation, their deployment onto robots is hindered by the inability to optimize further without additional labels. Recently, Segment Anything Model (SAM) [24] is proposed for zero-shot image segmentation. SAM is trained on an extensive dataset of 1 billion masks and 11 million images. The segmentation model is designed and trained to be promptable. However, SAM may still produce over-segmentation of the objects which highlights the necessity for subsequent refinement to ensure its effectiveness in real-world applications.

B. Self-supervised Learning

Self-supervised learning methodologies leverage the inherent characteristics of unlabeled data to automatically generate data labels for subsequent learning processes. A wide range of pretext tasks have been designed to provide self-supervised signals [5]–[7], [25]–[27]. Self-supervised learning also shown great promise in the field of image segmentation. LOST [28] and FreeSOLO [29] use pre-trained models to generate pseudo labels for segmentation. Therefore, the performance of [28], [29] is inherently limited by the pre-trained model. OGC [30] and LSMOL [31] rely on motion cues instead of pre-trained models to generate pseudo labels. The segmentation network is then trained through iterative optimization to form positive feedbacks. While they perform satisfactory in scenes with moving object, distinguish foreground and background in static scenes remains a challenge.

C. Contrastive Learning

Contrastive learning is widely used in self-supervised learning. Examples of contrastive learning methods include SimCLR [5], Moco [6], SwAv [32] and SimSiam [33]. Typically, these methods use instance discrimination as the proxy task, expecting different views of the same image to have similar feature representations. Beyond operating at the image level, contrastive learning can be also applied at the point level [26] or region level [27] to obtain fine-grained supervision signals, which is helpful for segmentation tasks. For example, DenseCL [26] expects the same local region in different augmented views of the same image to have similar feature representations. Reco [27] expects the pixel-level features in the same object region to be as close as possible to the mean features of the category. By utilizing fine-grained supervised signals, these methods successfully improve the performance of the segmentation task.

III. Multi-view Self-supervised Segmentation Network

In this paper, we aim to adapt an off-the-shelf object segmentation network in a self-supervised manner by leveraging multi-view observations. When humans observe a scene from multiple viewpoints, knowledge of the ego-motion allows humans to effectively associate the observations. Robot is desired to have the same capability to utilize the multi-view consistency especially in indoor scenes where objects and scenes are mostly stationary. For example, it is expected that a pixel belonging to the same object should have the same predicted segmentation labels and consistent feature representations across views.

The overall architecture of our proposed MVSS framework is shown in Fig. 2. Taking a sequence of RGB-D

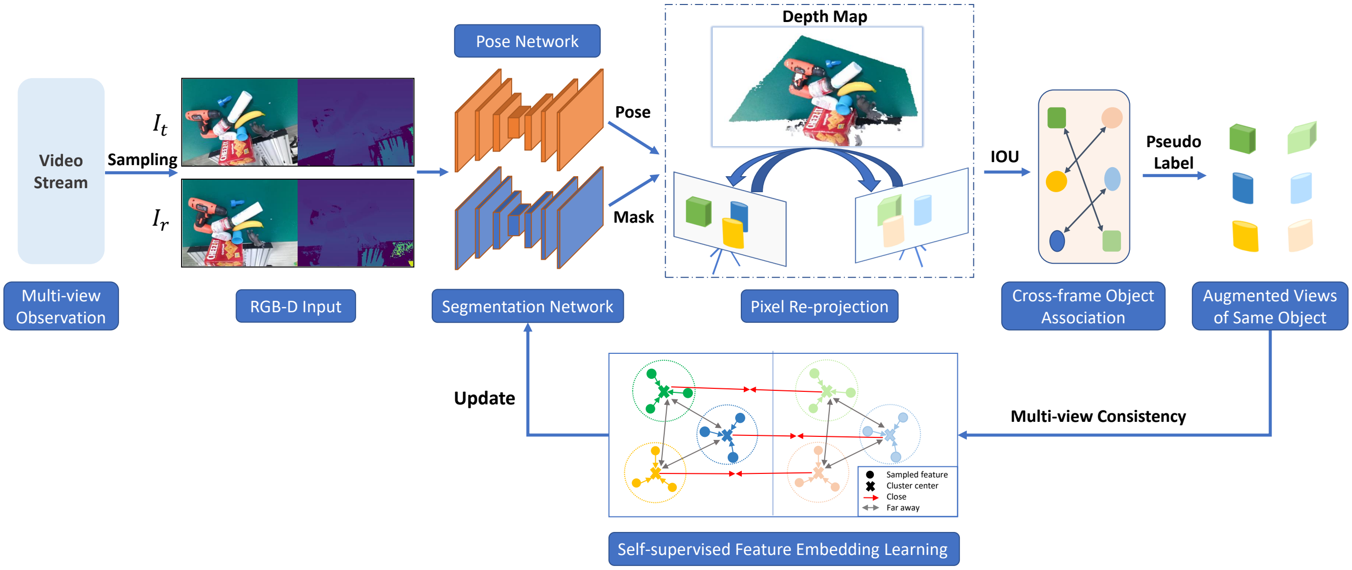

Fig. 2. The architecture of our proposed MVSS framework. The model iteratively samples a reference frame I_r and a target frame I_t as input, and utilizes a camera pose estimation network and a segmentation network to obtain relative poses and segmentation masks. The segmentation masks of the two views are re-projected according to the depth map and the relative pose, then cross-frame object association is performed in which matching mask pairs are considered as object pairs, and unmatched masks are considered as mis-segments. The multi-view observations of matching objects are regarded as augmented sample pairs. Our proposed multi-view consistency loss is utilized here to perform self-supervised feature embedding learning to improve the performance of the segmentation network. Note that the pose network accepts RGB images as input and can be continuously trained in an unsupervised manner when the input includes depth images.

images, we iteratively sample the reference frame I_r and the target frame I_t as input. There are mainly two branches of the MVSS framework: the segmentation network and the camera pose estimation network. Pixel-level and object-level correspondences are then established through camera pose estimation and cross-frame object association to learn feature embeddings that the same object are close to each other and embeddings from different objects are separated. It can be observed in experimental results that by utilizing both pixel-level and object-level correspondences through multi-view consistency, our proposed method is able to adapt existing segmentation methods to learn discriminative features for segmentation in new scenarios to achieve performance close to that of supervised methods.

A. Camera Pose Estimation

We utilize the camera pose estimation network from SfM-Learner [7] to estimate the relative camera poses between pairs of RGB frames. Specifically, we take two RGB images, the reference frame and the target frame as input, from which the network estimates the camera pose of the target frame relative to the reference frame. The network is trained in an unsupervised manner by minimizing the photometric reconstruction loss. We use the depth image obtained by the RGB-D camera in computing the photometric loss, which differs from the original SfMLearner where the pose estimation network and the single-view depth CNN are jointly trained. During testing, only RGB images are needed to estimate the camera relative poses.

We denote $I_r(p)$ as the reference frame, $I_t(p)$ as the target frame, and denote $\hat{I}_r(p)$ as the image obtained by differentially projecting the reference frame $I_r(p)$ onto the camera view of $I_t(p)$ through the depth map rendering. $\hat{E}_r(p)$ is the explainability prediction network [7] indicating the network's belief in where pixel correspondence is successfully modeled. The view synthesis objective is weighted correspondingly by:

$$L_{vs} = \sum_p \hat{E}_r(p) |I_t(p) - \hat{I}_r(p)| \tag{1}$$

Since there is no direct supervision for $\hat{E}_r$, a regularization term $L_{reg}(\hat{E}_r)$ is added to encourage non-zero predictions by minimizing the cross-entropy loss with constant label 1 at each pixel. The explainability network is jointly and simultaneously trained with the pose network. With l indexes over different image scales, the total loss function of the pose estimation network is formulated as follows:

$$L_{pose} = \sum_l L^l_{vs} + \lambda_e L_{reg}(\hat{E}^l_r), \tag{2}$$

B. Pixel Re-projection

To effectively leverage the information provided by the multiple views, it is necessary to first establish the pixel correspondence between different views. This involves projecting the depth map onto a point cloud and obtaining its coordinates in the camera coordinates using the intrinsic matrix, i.e. depth map rendering. Once we have the relative pose of the camera, the point cloud can be projected onto camera coordinate of the other view:

$$\begin{bmatrix}u_t\\v_t\\1\end{bmatrix} = K\begin{bmatrix}X_t\\Y_t\\Z_t\\1\end{bmatrix} = KT_{r\to t}\begin{bmatrix}X_r\\Y_r\\Z_r\\1\end{bmatrix} = KT_{r\to t}d_rK^{-1}\begin{bmatrix}u_r\\v_r\\1\end{bmatrix} \tag{3}$$

where $[u_r, v_r]$ and $[u_t, v_t]$ are the pixel coordinates in the reference frame and target frame, respectively. $[X_r, Y_r, Z_r]$ and $[X_t, Y_t, Z_t]$ are the corresponding camera coordinates. K is the intrinsic matrix of the camera. d_r is the depth value of $[u_r, v_r]$. $T_{r\to t}$ is the projection matrix from the reference frame to the target frame obtained from the aforementioned camera pose estimation.

C. Cross-frame Object Association

In cross-frame object association, we project the segmentation results between two frames through pixel re-projection introduced in the previous section. Assuming there are M objects in the reference frame and N objects in the target frame, the intersection over union (IoU) for each pair of objects is calculated as shown in Fig. 3. We employ the Hungarian algorithm [34] for object association. Hungarian algorithm addresses the unique constraint where the optimal assignment can be obtained. Additionally, object pairs with IoU scores below a threshold λ are considered false matches.

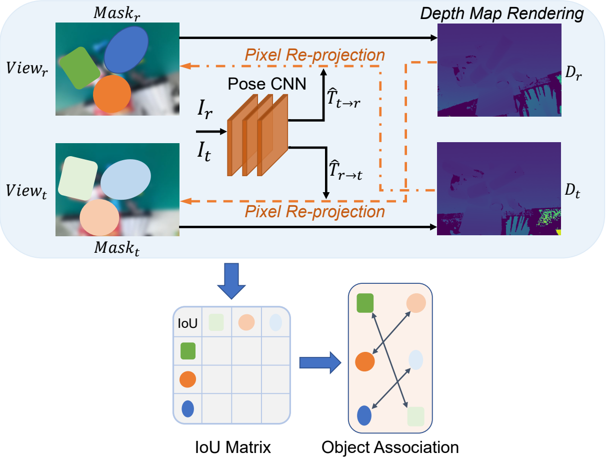

Fig. 3. The process of cross-frame object association. The segmentation masks $Mask_r$ and $Mask_t$ and the relative pose $\hat{T}_{r\to t}$ and $\hat{T}_{t\to r}$ are obtained by the segmentation network and pose network, respectively. We then re-project the mask to another view through depth map rendering. The Intersection over Union (IOU) matrix of all object pairs between the two frames are calculated for cross-frame object association.

After cross-frame object association, matched object pairs are treated as the same object for self-supervised feature embedding learning. Besides, unmatched segments may also include objects that should not be considered as backgrounds during subsequent training, which may lead to performance degradation. Here, we aim to produce consistent segmentation results for well-segmented objects across different views. The pixels from the unmatched segments will not be sampled in the subsequent loss calculation. This process continues until these challenging samples can establish bidirectional matches in specific frames. Through this way, the self-supervised feature learning is able to learn discriminative features of objects for improved segmentation.

D. Self-supervised Feature Embedding Learning

We aim to achieve consistent and continuous segmentation results across multiple views. Ideally, the feature embeddings of the same object should be similar across frames, while the feature embeddings of different objects should be far apart, which we consider to be an important metric for optimization. Suppose N object pairs are obtained through cross-frame object association, we denote the mean feature embeddings of all $2N$ object regions as $z_1, ...z_{2N}$, where z_{2n} and z_{2n-1} represent a matching pair. The cosine similarity between $\ell 2$ normalized $\boldsymbol{u}$ and $\boldsymbol{v}$ is denoted as $sim(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}/\|\boldsymbol{u}\|\|\boldsymbol{v}\|$. The multi-view consistency loss, or the multi-view contrastive loss, is formulated as follows:

$$L_{i,j} = -log\frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N}\mathbb{1}_{[k\neq i]}exp(sim(z_i, z_k)/\tau)} \tag{4}$$

$$L_{mvss} = \frac{1}{N}\sum_{n=1}^{N}[L_{2n-1,2n} + L_{2n,2n-1}] \tag{5}$$

where τ is the temperature parameter. Furthermore, we utilize dense contrastive learning to obtain more fine-grained supervision signals. The region contrastive loss proposed in [27] is extended to fit our experimental setting. Assuming there are N matched object pairs in the two frames, we perform pixel-to-region contrast on the sampled pixels and objects. We denote R_q^n as the set of sampled pixel representations r_q from object n. $r_k^{n,+}$ is the representation of the positive key, which is the mean representation of object n. R_k^n is the negative key set including pixel representations sampled from other objects besides object n. Then, the dense contrastive loss can be formulated as:

$$r_k^{n,+} = \frac{1}{|R_q^n|}\sum_{r_q \in R_q^n} r_q \tag{6}$$

$$L_{dense} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{|R_q^n|}\sum_{r_q\in R_q^n}$$
$$-log\frac{exp(r_q\cdot r_k^{n,+}/\tau)}{exp(r_q\cdot r_k^{n,+}/\tau) + \sum_{r_k^-\sim R_k^n} exp(r_q\cdot r_k^-/\tau)} \tag{7}$$

We also incorporate an intra-cluster loss similar to [20], which improves the stability of clustering by pushing the feature embeddings of all pixels belonging to the same object to the corresponding cluster center. Suppose P pixels are sampled on each object, and x_i^n is the feature embedding of the i-th pixel that belongs to object n. μ^n is the average of the pixel embeddings of the n-th object, d represents the cosine distance, α is the margin. Then the intra-cluster loss function is formulated as:

$$\mu^n = \frac{\sum_{i=1}^{P} x_i^n}{||\sum_{i=1}^{P} x_i^n||} \tag{8}$$

$$L_{intra} = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{P} \frac{\mathbb{1}\{d(\mu^n, x_i^n) - \alpha \geq 0\} d^2(\mu^n, x_i^n)}{\sum_{i=1}^{P} \mathbb{1}\{d(\mu^n, x_i^n) - \alpha \geq 0\}} \tag{9}$$

The total loss function is formulated as follows:

$$L_{total} = \lambda_{intra} L_{intra} + \lambda_{mvss} L_{mvss} + \lambda_{dense} L_{dense} + \lambda_{pose} L_{pose} \tag{10}$$

IV. Experiments

In this section, we evaluate the performance of the proposed MVSS framework. We employ UCN [20], a state-of-the-art class-agnostic RGB-D object segmentation network, as our segmentation backbone. UCN achieves class-agnostic segmentation through two steps of feature extraction and clustering. The network is pre-trained on a simulated indoor object dataset TOD [22]. Furthermore, we utilize the unsupervised pose estimation network from SfMLearner [7] to estimate the relative camera pose for pixel re-projection and object association. Finally, we apply our MVSS (multi-view self-supervised) framework on the GraspNet-1Billion Dataset [35] to evaluate the effectiveness of our method.

A. Dataset and Metrics

GraspNet-1Billion Dataset. The GraspNet-1Billion [35] dataset includes 88 daily objects with high-quality 3D mesh models. The images are collected from 190 cluttered scenes, each contributing 256 real-world RGB-D images, for a total of $48,640$ images. Annotations of the GraspNet-1Billion dataset includes grasp poses, instance masks, camera poses and object 6D poses.

To simulate real-world online learning scenarios, we sample two images including a reference frame I_r and a target frame I_t during each iteration. Denoting the frame sampling gap as T, then $256 - T$ frame pairs can be sampled from each video. All images are resized to 640×480 pixels during training and evaluation.

Evaluation Metrics. We first optimize our network without using any human annotations, and then use the following metrics to evaluate the performance of segmentation network and pose estimation network.

- *Overlap P/R/F* Following previous works [20]–[22], we use the object precision, recall and F-measure as the metrics to evaluate the performance of the object segmentation network. Additionally, the Boundary P/R/F is used to evaluate the accuracy of the segmentation boundaries. Notably, the widely used IoU in semantic segmentation, which is closely correlated with F-measure, is not presented in evaluation.
- *Absolute Trajectory Error* (ATE) [36] directly measures the difference between the camera pose of the estimated trajectory and the ground truth. Following [7], we employ ATE as the metric to evaluate the performance of our pose estimation network.

B. Implementation Details

In our experiments, we set the hyperparameters $\lambda_{mvss} = \lambda_{dense} = 0.3$ and $\lambda_{intra} = \lambda_{pose} = 10$. We employ the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a learning rate of 10^{-5} for both segmentation network and pose estimation network. The batch size is set to 2 (I_r and I_t). The frame sampling gap n is set to 10. The threshold λ is set to a small value of 0.2. For the multi-view consistency loss and the dense contrastive loss, we set the temperature parameter τ to 0.5. With regard to the dense contrastive loss, we sample 256 anchor points with an equal number of negative samples each time. For the intra-cluster loss, the number of sampled points P and margin α are set to 1000 and 0.02, respectively.

C. Pose Estimation

The pose estimation network [7] is trained in an unsupervised manner. During each optimization step, the network takes two RGB images - the reference frame and the target frame - as input, with the target depth map being required to calculate the photometric reconstruction loss. Notably, the network learns to estimate the relative camera pose between the two input RGB images without relying on any ground truth supervision. At the beginning of training, we optimize the pose estimation network alone on the first 10 videos observed by MVSS, and then jointly optimize with the segmentation network through multi-view observations. As shown in Fig 4, with the pose estimation network, we are able to achieve fairly stable pose estimation performance without any ground-truth camera pose, which is crucial in real-world applications.

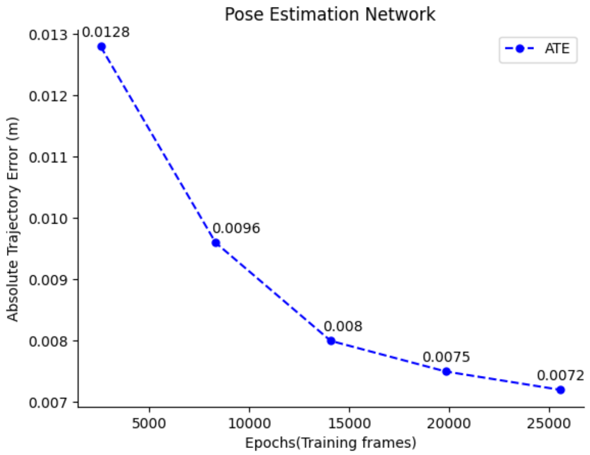

Fig. 4. Performance evaluation of pose estimation as the increase of training frames.

Furthermore, to mitigate the impact of estimation errors on potential false matches, we simultaneously estimate the relative pose between the reference frame and the target frame from two directions. The inverse matrix for these estimations is calculated to evaluate the pose error. If the error exceeds a predefined threshold, the estimation is considered inaccurate and no subsequent matching is performed.

TABLE I

COMPARISON WITH STATE-OF-THE-ART METHODS

Method	Overlap			Boundary		
	P	R	F	P	R	F
UCN (Pretrained on simulated TOD)	77.3	52.2	52.9	54.1	37.3	36.0
SD-Mask R-CNN (Fully-supervised)	88.6	76.7	82.1	56.9	50.7	53.4
UCN (Fully-supervised)	87.4	94.8	90.9	77.6	76.5	77.0
MVSS (Fully-supervised)	**84.3**	**87.0**	**85.6**	**70.2**	**73.4**	**71.5**
MVSS (w/o dense loss)	74.6	83.1	78.5	47.2	49.5	48.1
MVSS (Self-supervised)	**79.3**	**84.2**	**81.6**	**53.4**	**48.6**	**50.7**

TABLE II

EVALUATION OF DIFFERENT FRAME SAMPLING GAPS

Sampling	Overlap			Boundary		
Gap	P	R	F	P	R	F
1	68.7	78.1	73.0	38.5	45.1	41.4
4	71.3	78.6	74.7	43.3	46.8	44.8
7	73.2	81.4	77.0	45.7	48.7	47.0
10	**79.3**	**84.2**	**81.6**	**53.4**	**48.6**	**50.7**
13	75.4	81.9	78.4	48.7	50.8	49.5
16	74.3	80.6	77.3	47.0	49.6	48.1

D. *Object Segmentation*

In this section, we compare our methods with the state-of-the-art methods on GraspNet-1Billion Dataset. UCN [20] and SD-Mask R-CNN [17] are both representative methods for class-agnostic object segmentation, which we train supervised on the dataset. As shown in Table I, UCN pretrained on the simulated dataset TOD significantly degrades on the real-world dataset GraspNet-1Billion. However, after observing the RGB-D sequence once without any manual annotation, our proposed MVSS significantly improves the performance of the F-measure of UCN (52.9% to 81.6%), even close to the result of the supervised SD-Mask RCNN (82.1%).

It is worth noting that all supervised methods are trained on the training set for multiple epochs until convergence, while the self-supervised MVSS only trains for one epoch. The best result comes from the fully-supervised UCN while the fully-supervised MVSS takes the second place. Although the self-supervised MVSS achieves satisfactory segmentation results, due to the noises in pseudo-labels, pixel embeddings produced by MVSS may aggregate around some specific object parts for some objects. This occasionally leads to instances of over-segmentation during the subsequent clustering process. For example, objects like bottles might be split into segments representing the body and cap. However, MVSS produces consistent segmentation results across frames. The overall occurrence of the over-segmentation is largely reduced.

E. *Ablation Study*

In this section, we perform experiments to analyze the performance of different components of our MVSS framework.

Frame Sampling Gap. Selecting an appropriate sampling gap is important for multi-view segmentation. A small sampling gap leads to similarities in observations, constraining the advantages of the multi-view information. On the contrary, a large sampling gap may result in occlusions of the scene. The previous object may be occluded from the current view due to the camera movement. In Table II, We evaluate the influence of different sampling gaps and find that a sampling gap of 10 frames produces the best results, which is used by default for all experiments.

Multi-View Consistency Loss. In our experiments, we find that the introduced multi-view consistency loss L_{mvss}, which associates the similarity of cross-frame object pairs is crucial in segmentation. Experimental results indicate that without incorporating the multi-view consistency loss, the experiments all fail due to the noises in the pseudo-labels. Here, we provide a more detailed discussion on our proposed multi-view consistency loss.

The key to the effectiveness of self-supervised contrastive learning is to learn mutual information from the positive samples. Anchors are paired with corresponding positive and negative samples. Positive samples are usually augmented samples (e.g. SimCLR [5], Moco [6]), cluster means [32], or embeddings of the corresponding region of the anchor [26]. Negative samples, on the other hand, are other unrelated pairs randomly sampled from the dataset.

Some approaches such as Swav [32], BYOL [37] and SimSiam [33] rely solely on positive samples for contrastive learning and still achieve promising results. The key of these methods is to ensure a stable correlation between the anchor and positive samples. The sample positions are adjusted in the embedding space during iterations, where positive samples are getting closer to each other. Models trained by self-supervised contrastive learning can be used as backbone for downstream tasks and generally have good generalization performance.

In our method, the object matching pairs obtained by pixel re-projection and cross-frame object association yield object pairs with spatial proximity and semantic similarity. Other object pairs are naturally spatially and semantically far apart. Despite the initial pseudo-label noises, our multi-view contrastive learning effectively converges, generating meaningful feature representations. Combining the pixel-level and object-level correspondences via pixel re-projection and cross-frame object association helps improve the effectiveness of positive and negative sample selection, thereby enabling the segmentation network to converge more effi-

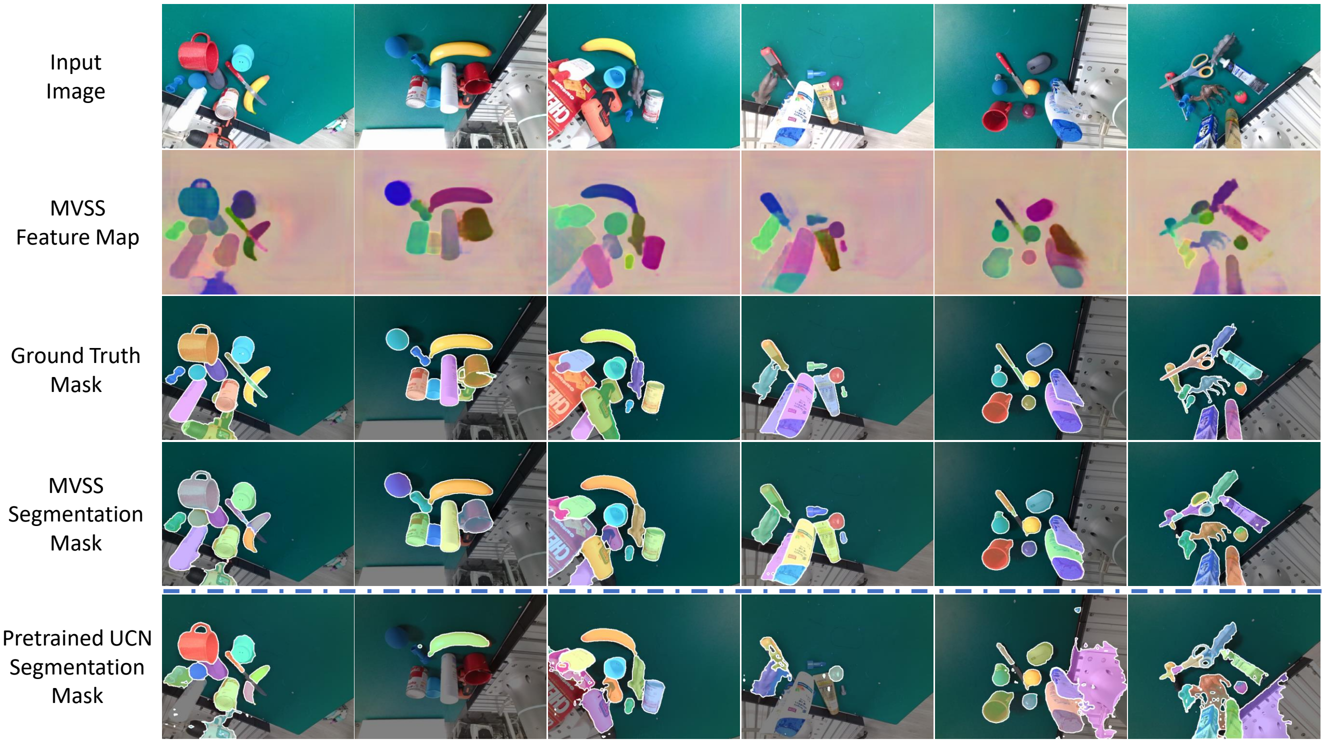

Fig. 5. Examples of the feature maps and segmentation masks obtained by MVSS and UCN pretrained on TOD [20]. It is observed that the pre-trained UCN yields mis-segmented samples in real-world scenes, even only a few objects can be correctly segmented in some scenes. Incorporating our proposed MVSS framework, the network is able to utilize multi-view observations for self-supervised feature embedding learning, achieving improved accuracy and consistency without the need for annotations.

ciently in the embedding space.

Dense Contrastive Loss. We study the effectiveness of the proposed dense contrastive loss, as shown in Table I. It is observed in Table I that the integration of the dense contrastive loss brings improvements on all metrics. Note that the optimization objective of the dense contrastive loss is similar to the multi-view consistency loss. Both objectives aim to push the embeddings of the same objects to be as close as possible, while keeping the embeddings of different objects far apart. The improvement of the performance indicates that fine-grained supervision information is useful for segmentation task.

Coefficients of Losses. Here we discuss the coefficients of the four loss functions L_{mvss}, L_{dense}, L_{intra} and L_{pose}. The pose estimation network loss L_{pose} is independent of the other three terms, the coefficient λ_{pose} is set to a fixed value of 10. Note that L_{intra} is highly correlated with the performance of the clustering, we also set the coefficient λ_{intra} to 10. In experiments, we find that similar results are obtained when λ_{mvss} ranges from 0.2 to 0.8. Considering that L_{dense} and L_{mvss} have similar optimization objectives, we finally set λ_{dense} and λ_{mvss} to the same value of 0.3.

V. CONCLUSIONS

In this paper, we proposed a novel multi-view self-supervised framework (MVSS) for self-adaptive object segmentation. Through pose estimation and cross-frame object association, object-level and pixel-level correspondences were established to optimize the multi-view consistency. The spatial consistency of object matching pairs naturally provides stable positive and negative samples, which provides robust supervision for contrastive learning, enabling the network to learn discriminative feature representations and multi-view consistent segmentations. Experimental results show that our proposed MVSS achieves performance close to supervised methods in novel scenes by observing RGB-D sequences without any human annotations. If we apply methods such as optical flow, which do not require depth information to obtain pixel-level correspondences, MVSS can be utilized to segmentation methods that use only RGB data. We consider that this self-supervised consistency is also useful for top-down methods and class-aware segmentation methods, enabling off-the-shelf segmentation networks to further improve the performance from multi-view observations without any human annotation.

REFERENCES

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[4] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[7] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised learning of depth and ego-motion from video,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[8] X. Wang, A. Jabri, and A. A. Efros, “Learning correspondence from the cycle-consistency of time,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2566–2576.

[9] A. Eitel, N. Hauff, and W. Burgard, “Self-supervised transfer learning for instance segmentation through physical interaction,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4020–4026.

[10] R. Qian, T. Meng, B. Gong, M.-H. Yang, H. Wang, S. Belongie, and Y. Cui, “Spatiotemporal contrastive video representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6964–6974.

[11] Y. He, W.-C. Chiu, M. Keuper, and M. Fritz, “Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4837–4846.

[12] X. Li, S. Liu, S. De Mello, X. Wang, J. Kautz, and M.-H. Yang, “Joint-task self-supervised learning for temporal correspondence,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[13] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “Flownet 2.0: Evolution of optical flow estimation with deep networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[14] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, “High accuracy optical flow estimation based on a theory for warping,” in *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV 8*. Springer, 2004, pp. 25–36.

[15] L. Yi, W. Zhao, H. Wang, M. Sung, and L. J. Guibas, “Gspn: Generative shape proposal network for 3d instance segmentation in point cloud,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3947–3956.

[16] W. Wang, R. Yu, Q. Huang, and U. Neumann, “Sgpn: Similarity group proposal network for 3d point cloud instance segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2569–2578.

[17] M. Danielczuk, M. Matl, S. Gupta, A. Li, A. Lee, J. Mahler, and K. Goldberg, “Segmenting unknown 3d objects from real depth images using mask r-cnn trained on synthetic data,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7283–7290.

[18] X. Wang, S. Liu, X. Shen, C. Shen, and J. Jia, “Associatively segmenting instances and semantics in point clouds,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4096–4105.

[19] J. Lahoud, B. Ghanem, M. Pollefeys, and M. R. Oswald, “3d instance segmentation via multi-task metric learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9256–9266.

[20] Y. Xiang, C. Xie, A. Mousavian, and D. Fox, “Learning rgb-d feature embeddings for unseen object instance segmentation,” in *Conference on Robot Learning*. PMLR, 2021, pp. 461–470.

[21] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, “The best of both modes: Separately leveraging rgb and depth for unseen object instance segmentation,” in *Conference on robot learning*. PMLR, 2020, pp. 1369–1378.

[22] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, “Unseen object instance segmentation for robotic environments,” *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1343–1359, 2021.

[23] S. Back, J. Lee, T. Kim, S. Noh, R. Kang, S. Bak, and K. Lee, “Unseen object amodal instance segmentation via hierarchical occlusion modeling,” in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 5085–5092.

[24] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, “Segment anything,” *arXiv preprint arXiv:2304.02643*, 2023.

[25] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin, and C. C. Loy, “Self-supervised scene de-occlusion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3784–3792.

[26] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, “Dense contrastive learning for self-supervised visual pre-training,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3024–3033.

[27] S. Liu, S. Zhi, E. Johns, and A. J. Davison, “Bootstrapping semantic segmentation with regional contrast,” in *International Conference on Learning Representations*, 2022.

[28] O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. Pérez, R. Marlet, and J. Ponce, “Localizing objects with self-supervised transformers and no labels,” November 2021.

[29] X. Wang, Z. Yu, S. De Mello, J. Kautz, A. Anandkumar, C. Shen, and J. M. Alvarez, “Freesolo: Learning to segment objects without annotations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 176–14 186.

[30] Z. Song and B. Yang, “OGC: Unsupervised 3D Object Segmentation from Rigid Dynamics of Point Clouds,” in *NeurIPS*, 2022.

[31] Y. Wang, Y. Chen, and Z. Zhang, “4d unsupervised object discovery,” in *NeurIPS*, 2022.

[32] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[33] X. Chen and K. He, “Exploring simple siamese representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.

[34] J. Munkres, “Algorithms for the assignment and transportation problems,” *Journal of the society for industrial and applied mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[35] H.-S. Fang, C. Wang, M. Gou, and C. Lu, “Graspnet-1billion: A large-scale benchmark for general object grasping,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

[36] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of rgb-d slam systems,” in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 573–580.

[37] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, “Bootstrap your own latent-a new approach to self-supervised learning,” *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.