ISAR: A Benchmark for Single- and Few-Shot Object Instance Segmentation and Re-Identification

Nicolas Gorlo
gorlon@ethz.ch

Kenneth Blomqvist
kblomqvist@mavt.ethz.ch

Francesco Milano
francesco.milano@mavt.ethz.ch

Roland Siegwart
rsiegwart@ethz.ch

Autonomous Systems Lab, ETH Zürich

Abstract

*Most object-level mapping systems in use today make use of an upstream learned object instance segmentation model. If we want to teach them about a new object or segmentation class, we need to build a large dataset and retrain the system. To build spatial AI systems that can quickly be taught about new objects, we need to effectively solve the problem of single-shot object detection, instance segmentation and re-identification. So far there is neither a method fulfilling all of these requirements in unison nor a benchmark that could be used to test such a method. Addressing this, we propose **ISAR**, a benchmark and baseline method for single- and few-shot object **I**nstance **S**egmentation **A**nd **R**e-identification, in an effort to accelerate the development of algorithms that can robustly detect, segment, and re-identify objects from a single or a few sparse training examples. We provide a semi-synthetic dataset of video sequences with ground-truth semantic annotations, a standardized evaluation pipeline, and a baseline method. Our benchmark aligns with the emerging research trend of unifying Multi-Object Tracking, Video Object Segmentation, and Re-identification.*

1. Introduction

Object-level scene understanding is fundamental to the development of robust and effective spatial AI systems. The ability for an AI system to accurately perceive objects within a scene is crucial to many applications in robotics, augmented reality, navigation, and autonomous vehicles.

Working towards the goal of general-purpose, spatial AI systems, the research community has become increasingly interested in developing object-level mapping pipelines [2, 15, 25, 31, 35, 42, 51, 64–66, 72, 83, 85, 90] for use in robotics, augmented reality and other spatial AI applications. All of these pipelines assume known object models [66] or an upstream object segmentation pipeline, either to assign a semantic class to static parts of the scene [25, 64], or to deal with dynamic objects separately from the static background mapping and reconstruction [2, 15, 31, 35, 42, 65, 72, 83, 85, 90]. Additionally, systems have been proposed which incrementally build databases of objects and their geometry to help with object manipulation [23, 45]. Such systems also rely on object instance segmentation. Crucially, this object-level segmentation is currently produced using learned object segmentation models such as Mask R-CNN [27], Sharp-Mask [57], PSPNet [91] or Detic [96]. Such learned object segmentation models require classes to be known at training and dataset building time. Fine-tuning them to handle a new class requires annotating thousands of diverse examples and many expensive optimization iterations. To build general-purpose robots and spatial AI systems that can deal with arbitrary objects, one needs to be able to teach them about new objects at run-time. Teaching the spatial AI system and adding new objects to an object tracking and segmentation pipeline should be as easy as selecting the new objects through a user interface. Effectively, this means solving the problem of single- or few-shot object instance segmentation and re-identification.

So far, the computer vision community has studied the problems of few-shot semantic segmentation, Video Object Segmentation (VOS) and Re-identification (re-ID) individually. While tremendous progress has been made on all of these tasks independently, we can still not teach our spatial AI systems effectively about new objects. Few-shot semantic segmentation systems use dense object segmentation masks, which are expensive to obtain, and they do not make use of the temporal structure in the video data that spatial AI systems inherently operate on. VOS methods perform extremely well, and make full use of video data, but require an initial dense segmentation mask. Further, they do not deal with re-identifying the objects across different *scene contexts*, which we define as the surrounding environment in which an object is recorded, including the pose of the objects in the environment. As a result, the in-the-wild usability of VOS methods for the previously mentioned tasks is limited.

In summary, existing methods for object instance segmentation and re-identification either do not use the temporal structure of video data, rely on initial dense segmentation masks and/or struggle to re-identify objects across different scenes. Consequently, there is an apparent gap which could be filled by combining few-shot semantic segmentation, VOS and re-ID methods into one approach.

To address this gap, we propose **ISAR**, a benchmark and baseline method for single- and few-shot object **I**nstance **S**egmentation **A**nd **R**e-identification. Our benchmark is designed to accelerate research progress towards robust algorithms that will enable teaching spatial AI systems through a single or a few sparse training examples. We aim to unify few-shot semantic segmentation, Video Object Segmentation (VOS) and Re-identification (re-ID) – traditionally mostly separately studied research topics. This reflects the emerging trend in recent years towards combining these topics [48]. As our goal is to build spatial AI systems that can deal with objects in any configuration, including moving objects, we aim to make it hard to rely purely on spatial information as the primary means of object description. Instead, we advocate the development of vision-based object-centric methods, which do not rely on the context in which the object is first perceived. Through this approach, we strive to advance the representations of object instances, making them more detailed and distinguishable, which in turn may enable more robust re-identification and object-level mapping of dynamic environments. Our semi-synthetic benchmark dataset is recorded using the Habitat AI Simulator [50, 73] with scenes of Replica [71] and the Habitat-Matterport 3D research dataset [62], using objects from the YCB dataset [8]. The dataset and evaluation code are available at https://nicogorlo.github.io/isar_wacv24/.

To summarize, our contributions are the following: 1. A semi-synthetic dataset of video sequences with high-quality ground-truth semantic annotations; 2. A standardized evaluation pipeline to measure the performance of different methods on this task against each other; 3. A baseline method to segment and re-identify objects.

2. Background

In this section, we cover some of the work done on related problems, and show that none of these fully address the needs of modern object-centric spatial AI systems.

2.1. Video Object Segmentation

The primary goal of Video Object Segmentation (VOS) is to segment object instances in a video sequence. VOS has been categorized into four major subcategories: Semi-Supervised, Unsupervised, Referring and Interactive VOS.

Semi-Supervised VOS requires propagating a mask, provided in the first frame, to subsequent frames in a video. Facilitated by various benchmark datasets [6, 7, 33, 56, 58, 86], large steps in performance have been achieved [5, 13, 38, 68]. In particular, recent methods [11, 43, 77] have shown very strong performance on these benchmarks. However, their reliance on a costly annotation in the initial frame and their dependence on the object being salient and staying in the same context reduces their in-the-wild applicability for spatial AI systems, as the initial annotation is hard to provide in an online setting and objects are not guaranteed to be *salient* (*i.e*., prominent in the video frame).

In **Unsupervised VOS** the goal is to segment the salient objects in a video clip without any labeling cues or training. Using the same benchmark datasets as Semi-Supervised VOS and some specialized ones [19, 37, 53, 78], equally impressive steps in performance of methods have been achieved [14, 46, 49]. However, unsupervised VOS methods by definition rely on the fact that the objects are salient and the overall scene context does not change. Consequently, it is not applicable when these assumptions are not met. However, in in-the-wild scenarios, there no guarantees that the assumptions are met.

In **Video Instance Segmentation** (VIS), the goal is to segment all instances in a video from a set of object categories. YouTube-VIS [87], the most popular benchmark for VIS is built on the YouTube-VOS [86] benchmark dataset and therefore suffers from the same problem of few reappearing and non-salient objects as YouTube-VOS. Further, it is constrained to the predefined object categories.

The novel task of **Referring VOS** replaces the segmentation in the initial frame of Semi-Supervised VOS with a language prompt. The benchmarks Youtube-VOS [67] and DAVIS-2017 [29] have been extended with natural language prompts describing the objects. A number of methods have already achieved promising results on this task [4, 29, 36, 67, 80, 81].

In **Interactive VOS** [6, 7] the initial full mask of the semi-supervised scenario is replaced with interactive user inputs, given as scribbles, to refine the video object segmentation throughout the video. The interactive user input is provided in up to 8 rounds for the frame with the worst prediction among candidate frames. The strong performance of state-of-the-art [10, 12] methods on this task shows that one does not have to rely on expensive mask annotations like Semi-Supervised VOS.

While many benchmarks have been created for different types of VOS [6, 7, 16, 19, 37, 53, 56, 58, 59, 78, 86, 87], these mostly focus on segmenting salient objects, which rarely move out of frame or reappear in a different context. In addition, recently it was shown on MOSE [16] and OVIS [59], datasets for VOS and VIS in complex scenes that state-of-the-art VOS and VIS methods, achieving near perfect scores on DAVIS 2016 [56], struggle with heavy occlusion and more complex scenes. This strengthens the case that current VOS and VIS benchmark datasets do not properly mimic an in-the-wild scenario. However, even in the MOSE and OVIS benchmark datasets, the scene context mostly stays the same. Therefore, methods developed for these tasks are not fit to deal with reappearing, dynamic objects or for identifying previously seen objects in a new context.

Our benchmark, separating annotated and annotation-free scenes into distinct scene contexts, goes beyond the current scope of tasks in the field. This benchmark is designed to fuel the advancement of methods that combine VOS with re-identification across varying scene contexts. Instead of providing full segmentation masks, we provide point- and bounding box annotations as hints. This maintains some of the advantages of Interactive VOS, preventing methods from relying on costly segmentation.

2.2. Vehicle and Person Re-identification

Instance re-identification has been mainly explored within vehicle and person re-identification, as well as for face recognition. Facilitated by datasets for person [24, 28, 63, 84, 92–94] and vehicle [32, 40, 44, 74, 88] re-identification, impressive accuracy has been achieved in distinguishing and re-identifying vehicles and people, despite few differences among the instances that need to be distinguished from one another. State-of-the-art methods [1,22,34,52,60,76,79,95,98] now effectively solve the task with little error on the most popular datasets.

However, direct application of these methods to other object classes is infeasible, as the methods rely on training on large datasets of the same class. These challenging circumstances demand the development of techniques that can perform efficiently even with scarce data or no prior class information. Further, these methods only tackle re-identification and not object instance segmentation from videos.

As we deal with re-identifying objects of any class rather than specific classes, our definition of re-ID differs from these standard re-ID frameworks in that we do provide single- or few-shot labels. In standard re-ID frameworks, no labels are provided, but rather all objects of a specific class need to be re-identified.

2.3. Few-shot Semantic Segmentation

Few-shot semantic segmentation (FSS) [17, 20, 21, 55, 69, 75] methods predict pixelwise masks for novel classes given a few annotations. Most methods are based on metric learning [17]. Later works have introduced support query features and attention mechanisms [41, 70, 89], others have introduced fine-tuning [97], memory modules [82], and learned classifiers [47].

FSS methods are typically evaluated on the PASCAL-5$^{\mathrm{i}}$ [18, 69] and the COCO-20 [39] datasets, which hold out a certain number of classes that are used for few-shot evaluation. In these datasets, the training set actually contains some instances of the test classes already, but those are not labeled. FFS methods focus on simply segmenting the object class and do not detect or re-identify the same instances of an object. In the test phase, a full segmentation mask is provided, instead of a sparse prompt like a bounding box or pixel coordinate. They also operate on individual images instead of video, which rules out the use of methods which leverage temporal structure in the data. For an object-level SLAM type application, figuring out how to rely on sparser expert labels and how to make the best use of video data is necessary.

3. Problem Formalization

In the following, we denote with $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{eval}}$ respectively an annotated training set and an evaluation set, both of which we assume to consist of one or more ordered image sequences. Let the train sequences be indexed with $i \in \{1, \dots, \Omega^{\text{train}}\}$ and the evaluation sequences be indexed with $j \in \{1, \dots, \Omega^{\text{eval}}\}$. In particular, we define a training sequence $\mathcal{I}_i^{\text{train}} = \{\mathbf{I}_1^{\text{train}}, \mathbf{I}_2^{\text{train}}, ..., \mathbf{I}_{N_i}^{\text{train}}\} \in \mathcal{D}^{\text{train}}$ as a sequence of N_i posed RGB or RGB-D images $\mathbf{I}_r^{\text{train}} = (f_r, p_r)$ (*i.e.*, tuples of image f_r and 6-DoF pose p_r in the world coordinate frame), with $r \in \{1, \dots, N_i\}$, and an evaluation sequence $\mathcal{I}_j^{\text{eval}} = \{\mathbf{I}_1^{\text{eval}}, \mathbf{I}_2^{\text{eval}}, ..., \mathbf{I}_{M_j}^{\text{eval}}\} \in \mathcal{D}^{\text{eval}}$ as a sequence of M_j posed RGB or RGB-D images. Each training sequence $\mathcal{I}_i^{\text{train}}$ contains K_i annotated objects. Each annotation consists of the image index in which the annotation occurs, a bounding box given by the 2D image coordinates of the top left and bottom right corners, and a point that is part of the object. Let $\mathcal{A}_i^{\text{train}}$ denote the set of all annotations for sequence $\mathcal{I}_i^{\text{train}}$. For the train data annotation, there are two scenarios: a single-shot scenario and a multi-shot scenario. In the single-shot scenario, there is only one train sequence ($|\mathcal{D}^{\text{train}}| = 1$) and there is only one annotation per object. In the multi-shot scenario, there can be both more than one train sequence and more than one annotation per sequence. The evaluation sequences only contain ground-truth mask annotations for evaluation.

For each object O_k we seek to learn a mapping

$$\begin{aligned} F_k :\ & \{\mathcal{I}_i^{\text{train}}\}_{i \leq \Omega^{\text{train}}} \times \{\mathcal{A}_i^{\text{train}}\}_{i \leq \Omega^{\text{train}}} \times \\ & \{\mathbf{I}_j^{\text{eval}}(0), ..., \mathbf{I}_j^{\text{eval}}(t^\star)\} \in \mathcal{I}_j^{\text{eval}} \\ & \mapsto \mathbf{M}_{k,j}^{\text{eval}}(t^\star), \end{aligned} \tag{1}$$

such that the discrepancy between $\mathbf{M}_{k,j}^{\text{eval}}(t^\star)$ and $\mathbf{M}_{k,j,gt}^{\text{eval}}(t^\star)$ is minimized for all frames in each evaluation sequence $\mathcal{I}_j^{\text{eval}}$ and all objects O_k, $k \leq K$. Here $\mathbf{M}_{k,j}^{\text{eval}}(t^\star)$ denotes the predicted binary instance segmentation mask of object O_k in the sequence $\mathcal{I}_j^{\text{eval}}$ at time $t^\star$ and $\mathbf{M}^{\text{eval}}_{k,j,gt}(t^\star)$ its corresponding ground-truth annotation. Therefore, the goal is to sequentially produce pixelwise masks on the evaluation sequences, given the training sequence and the annotations.

The desired properties of the mappings F_k include: 1. *Consistency with annotations*: For every object instance O_k, the mapping F_k, if applied to a train sequence $\mathcal{I}_i^{\text{train}}$ should produce segmentation masks that are consistent with the provided annotations. 2. *Temporal consistency*: Given the temporal nature of the input video sequences, the mapping F_k should produce segmentation masks that are temporally consistent. In other words, for each object instance O_k, the predicted instance segmentation masks across different frames in a sequence should correspond to the same object. 3. *Spatial Consistency*: Both the predicted segmentation masks and its boundary should match the ground-truth segmentation mask. 4. *Generalization to different scenes*: The mappings F_k should be able to generalize to entirely different scene contexts in the evaluation data and should therefore be able to produce coherent segmentation masks in all the sequences $\{\mathcal{I}_j^{\text{eval}}\}_{j\in\{1,\dots,\Omega^{\text{eval}}\}}$ regardless of the scene context that the sequences may be recorded in.

We measure adherence to these properties with evaluation metrics introduced in Section 5. The overall objective of this task is to derive such mappings F_k that fulfill the desired properties. Importantly, the entire train sequence and the associated sparse annotation data can be used to form object representations that enable re-identification of the objects in a different scene context. An example of such object representations is outlined in Section 6.1.

For in-the-wild applicability, methods tackling this task should not pre-train on any data contained in the dataset. However, they may and are expected to make use of general pretraining on other data. In the case of our benchmark, for instance, pre-training on the Replica [71] or Matterport 3D datasets [62] is not allowed nor is using data containing the YCB [8] objects, as the objects are contained in our dataset.

4. Dataset

The benchmark contains 24 *test cases*, each consisting of a training set $\mathcal{D}^{\text{train}}$ of one sequence and an evaluation set $\mathcal{D}^{\text{eval}}$ of 1 to 5 evaluation sequences. For each test case, annotations for a single-shot and a multi-shot scenario are provided. In the single-shot scenario, a single annotation per object is provided, while in the multi-shot scenario there are between 6 and 14 annotations per object, varying across test cases. This is to test how different algorithms perform with a varying amount of supervision. Each test case contains up to 8 objects to track. In total, the dataset consists of 84 combined train and evaluation sequences. The length of each sequence is 400 frames at a frame-rate of 30 Hz. In Table 1 we compare its size to commonly used VOS benchmark datasets. Note that, in contrast to these, our dataset is only meant for *evaluating*, not training models and that, unlike for instance YouTube-VOS [86], its focus is not on achieving large scale. In particular, the amount of data provided for each object is chosen so as to prevent using the training sequences to overfit to the characteristics of the benchmark.

Benchmark	# instances per sequence (avg.)	# Annotations
Ours	3.52	124,681
DAVIS 2016 [56]	1	3,455
DAVIS 2017 [58]	2.56	10,474
YouTube-VOS [86]	1.74	197,272

Table 1. Comparison of dataset size to the three most common VOS benchmarks. # Annotations is the total number of unique object annotations. # instances per sequence (avg.) is the average number of annotated object instances per sequence

As our benchmark is designed to test single- and few-shot methods, we limit the amount of data available in training sequences and rather propose to test methods across a wide range of test scenarios. Methods tackling this task can make use of external data, as long as they do not contain data from the Matterport, Replica or YCB datasets.

The scenes in which the video sequences are recorded are indoor scenes of the Replica [71] and the Habitat-Matterport 3D research dataset [62]. We use data from the YCB dataset [8] as objects to be segmented and re-identified. A few sample frames of the dataset can be seen in Figure 1.

4.1. Data Recording

The scenes are recorded using the Habitat AI Simulator [50, 73]. To generate coherent sequences, we manually set the 6-DoF poses of keyframes and subsequently interpolate the positions using B-splines and the orientation using spherical quadrangular interpolation of the quaternion orientations with cubic splines. This results in smooth trajectories with continuous linear and angular accelerations. Using semantic annotations of the Replica [71] and Habitat Matterport 3D research dataset [62] allows us to circumvent expensive manual labeling and yields pixel-perfect annotations while retaining near photo-realistic data.

4.2. Attributes

Taking inspiration from the DAVIS dataset [56], we assign attributes to every training/evaluation sequence. These attributes are chosen to point out the strengths and weaknesses of methods tackling the benchmark. The attributes and their number of occurrences are shown in Table 2.

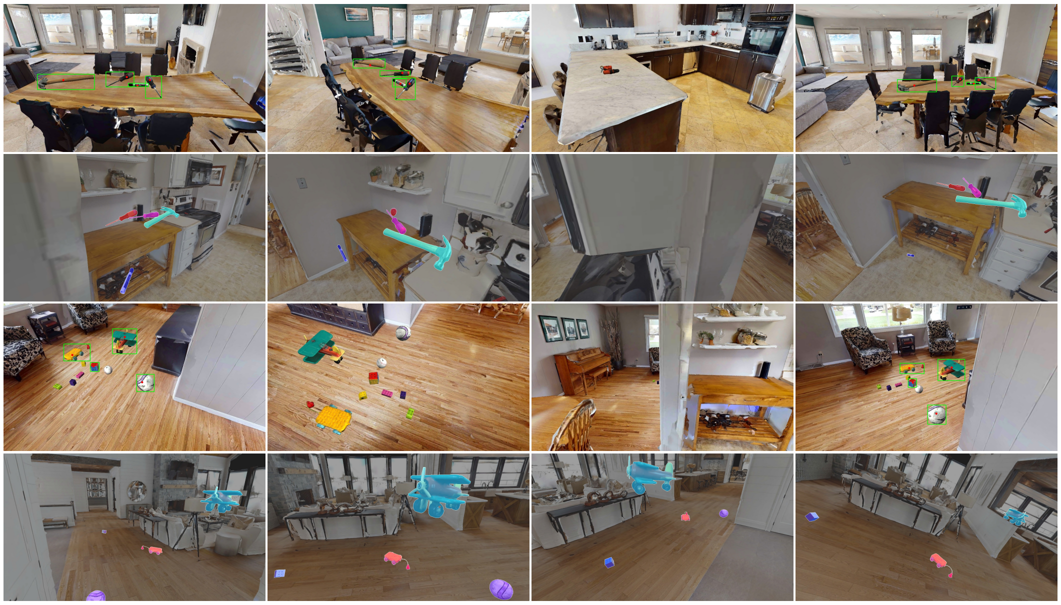

Figure 1. Samples from the dataset. The first and third rows show examples of sparsely annotated train sequences. The second and fourth row show dense ground truth segmentations from the evaluation set. In both cases, the train sequence is static, while the shown evaluation sequence is dynamic (pen falling off the countertop; toy airplane flying through the room). Best viewed in color.

ID	Description	#
DYN	The scene contains dynamic objects	26
CLT	The scene is cluttered with other objects	78
CLA	The scene contains multiple distinct objects of the same class (*e.g.*, 2 different cans)	43
SML	The mean ratio between object bounding box and image area is smaller than 0.005	58
SMF	There are frames with a ratio between the object bounding box and the image area smaller than 0.001	73
FST	The recording contains fast camera movements, defined as linear velocity $> 1.5\,\mathrm{m/s}$ or angular velocity $> 1.0\,\mathrm{rad/s}$	64

Table 2. Attributes of the dataset scenes with associated descriptions and number of occurrences.

5. Evaluation

To facilitate the comparison of different methods, we first draw a distinction between two possible scenarios: one where the object is visible within a frame, and the other where the object is not present in the frame. This categorization is designed so that we can separately compute evaluation metrics for each scenario. In the scenario where the object instance k is visible (the number of pixels in the mask $\mathbf{M}_{k,j,gt}(\hat{t})$ is larger than zero), we compute four different metrics:

1. The *Jaccard index* $\mathcal{J}$ (Intersection over Union) defined by

$$\mathcal{J} = \frac{\mathbf{M}_{k,j}(\hat{t}) \cap \mathbf{M}_{k,j,gt}(\hat{t})}{\mathbf{M}_{k,j}(\hat{t}) \cup \mathbf{M}_{k,j,gt}(\hat{t})} \in [0,1] \tag{2}$$

   measures the accuracy of the mask prediction, when the object is visible. Here, $\mathbf{M}_{k,j}(\hat{t})$ is the predicted binary segmentation mask of the object in a single frame and $\mathbf{M}_{k,j,gt}(\hat{t})$ is the corresponding ground truth mask.

2. The *boundary F_1-score* $\mathcal{F}$. For this we define the boundary of a segmentation mask $\mathbf{B}$ as $\boldsymbol{\partial}\mathbf{B} = \overline{\mathbf{B}} \setminus \mathbf{B}^{\circ}$, where $\overline{\mathbf{B}}$ denotes the closure of $\mathbf{B}$ and $\mathbf{B}^{\circ}$ its interior. These are the boundary pixels of the segmentation masks. The F_1-score of the boundary of the two segmentation masks is calculated with

$$\mathcal{F} = \frac{2\ \mathrm{TP}}{2\ \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{3}$$

   where $\mathrm{TP} = |\boldsymbol{\partial}\mathbf{M}_{k,j}(\hat{t}) \cap \boldsymbol{\partial}\mathbf{M}_{k,j,gt}(\hat{t})|$, $\mathrm{FP} = |\{\boldsymbol{\partial}\mathbf{M}_{k,j}(\hat{t}) \notin \boldsymbol{\partial}\mathbf{M}_{k,j,gt}(\hat{t})\}|$ and $\mathrm{FN} = |\{\boldsymbol{\partial}\mathbf{M}_{k,j,gt}(\hat{t}) \notin \boldsymbol{\partial}\mathbf{M}_{k,j}(\hat{t})\}|$. In the scenario of spatial perception, this measure is of particular interest, as it quantifies to what degree a mask either leaks into the background or undershoots and fails to cover the entire object. Preventing such leakage is crucial for building up a good 3D representation of an object.

Both for $\mathcal{J}$ as well as $\mathcal{F}$ we calculate the mean over an entire sequence and report this as the final measure, with higher values indicating better performance. These two measures follow DAVIS [56]. Additionally, we define:

3. The *visible misclassification rate* $J_{\text{mis, v}}$ as the ratio of frames, where $\mathcal{J} < t_{\text{mis}} = 0.4$. This measures the ratio of frames where at most a small part of the object was correctly classified.

4. The *visible false detection rate* $J_{\text{fd, v}}$: the ratio of frames, where $\mathcal{J} < t_{\text{fd}} = 0.1$ and $|\mathbf{M}_{k,j}| > |\mathbf{M}_{k,j,gt}| \times \mathcal{J}$. This measures the ratio of frames in which an object that is not the same instance as the correct object is segmented, even though the correct object is visible.

When the object is not visible ($|\mathbf{M}_{k,j,gt}| = 0$), we measure the *non-visible false detection rate* $J_{\text{fd, n}}$ as the ratio of frames, where $|\mathbf{M}_{k,j}| > 0$.

6. Baseline Method

We provide a baseline method for the benchmark that methods can be compared against. It relies on building feature descriptors of an object instance and on instance segments from the Segment Anything Model (SAM) [30]. The feature descriptors are used to re-identify the objects in other scene contexts by constructing a simple classifier for each object.

In this section we describe our method, provide an evaluation on our benchmark and perform an ablation study of the different components of our method.

6.1. Pipeline

Our method builds on pixel-wise features, which we use to describe our objects. An overview of the pipeline can be seen in Figure 2.

In the initial step, SAM is prompted with the bounding box and point prompt of the annotations in the training sequence. Inspired by [3], we leverage image features learned on a massive, diverse dataset to describe object instances. Therefore, each of the images is passed through a pretrained DINOv2 [54] backbone to extract a feature tensor $T_{\text{DINO}} \in \mathbb{R}^{40\times 40\times 1024}$. This tensor is upsampled to the image size, using bicubic upsampling. A linear maximal margin classifier (SVM) is trained for each object, using the dense DINOv2 features that lie within the predicted mask on the train images as positive samples; as negative samples, we use DINOv2 feature vectors sampled from outside of the predicted mask. To prevent overfitting to the current scene context, we include a set of DINOv2 feature vectors previously extracted from a diverse set of random images as negative examples. For frames in the evaluation dataset, we similarly compute upsampled DINOv2 feature vectors and use the classifier to determine instance membership pixel-wise. Optionally, the resulting masks are refined using SAM. To do this, we compute the dot product between all (pixel-wise) features and the SVM weights for an object. This effectively scores each pixel in terms of similarity to the object instance. Then, we select the 5 largest local maxima of this dot product to prompt SAM. This results in multiple mask proposals by SAM. These mask proposals are allocated to the object instances using linear assignment, maximizing the dot product of the SVM weights for an object and the feature vectors that lie beneath a mask proposal. Finally, if the average of the features that are covered by a mask lie on the negative side of the decision boundary of an assigned SVM, the mask is discarded.

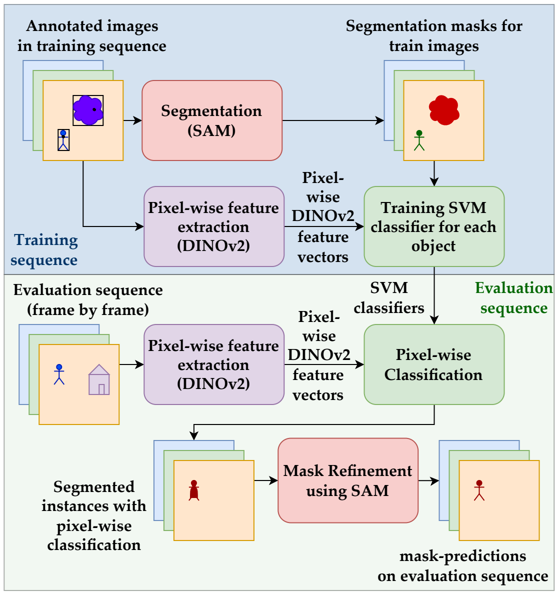

Figure 2. Pipeline of the baseline method. First, on annotated images of a training sequence, the point and bounding box prompts are transformed into masks using SAM [30]. Second, DINOv2 [54] features are extracted on these images. Using the masks, a dataset of positive and negative DINOv2 features is created for each annotated object. This dataset is used to train a linear SVM for each object. These linear SVMs are sequentially applied to the dense upsampled DINOv2 features of images in evaluation sequences to determine instance membership. Finally, the resulting masks are refined with SAM.

The benefits of the method include the fact that it does not rely on external training with segmentation or video data and purely uses out of the box models. Further, it is agnostic to the class of the objects.

6.2. Quantitative Results

We evaluate the baseline method on the proposed benchmark. The $\mathcal{J}$ with respect to scene attributes can be seen in Figure 3. The method performs best in scenarios with little clutter and no objects of the same semantic class and achieves lowest performance in scenes with clutter, objects of the same class, and small objects. The evaluation metrics averaged over all evaluation sequences can be seen in Table 3.

	$\mathcal{J}$	$\mathcal{F}$	$J_{fd,v}$	$J_{fd,n}$	J_{mis}
single shot	0.2686	0.1249	0.1319	0.0998	0.7092
multi shot	0.3224	0.1474	0.0629	0.0401	0.6545

Table 3. Performance of the baseline method expressed in the metrics of Section 5. Averaged over all evaluation sequences.

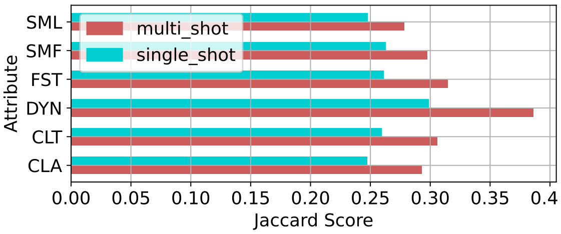

Figure 3. Mean $\mathcal{J}$ score of the baseline method with respect to scene attributes.

6.3. Qualitative Results

Using SAM to transform the point and bounding-box prompts into masks in general works well, however even small failures lead to inconsistencies over the entire evaluation sequence, as no additional data from the training sequence is used. A good and a bad example can be seen in Figure 4. This is especially apparent when the object is composed of multiple parts. Performance at this stage is critical, as negative features to train the SVM are sampled from the area outside the mask. Therefore, a mask that does not capture the entire object might lead to failure later on. However, this was observed in few sequences.

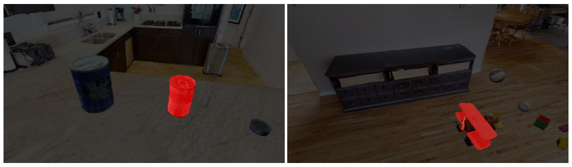

Figure 4. Qualitative results showing the effectiveness of SAM as a way to transform the given prompts into initial segmentations. Left: good example, right: bad example.

While the segmentation quality on the evaluation data is high (in frames where the correct instance is segmented, the $\mathcal{J}$ score is generally above 0.9), the mask quality varies from frame to frame, as no temporal information is taken advantage of.

6.4. Failure Cases

The high rate of misclassifications (frames where $\mathcal{J} < 0.4$) shows that the robustness of the method can be improved. Especially in scenarios where the object is small and few pixels are available, the DINOv2-model [54] fails to extract enough meaningful information to re-identify an object. Further, as the segmentation method (SAM [30]) is not optimized with respect to the the boundary F_1-measure $\mathcal{F}$, the latter is very low.

Moreover, as addressed in the previous paragraph, predictions in two subsequent frames may be entirely different due to no temporal knowledge being used.

Lastly, in the single-shot setting the baseline sometimes completely fails with objects of the same semantic class (for instance it fails in a test case that requires distinguishing between a flat and a crosshead screwdriver that are both small in the image). This is caused by the SVM classifier not being aware of the other instance in the scene. Therefore, the features of the very similar looking screwdriver might lie on the same side of the decision boundary as the one to track, as no negative examples of the same class exist. A way to improve this would be to combine the method with state-of-the-art VOS methods (*e.g.*, [11, 43, 77]) to track an object while it is visible in subsequent frames. This then may be used in the training sequence to generate additional data for the object's SVMs to be trained on. It could also be beneficial in the evaluation sequences to improve robustness.

6.5. Ablation on Method Components

We perform ablations on the different components in our baseline method. First, we discuss the choice of object descriptors. We test CLIP [61], SAM [30], and DINOv2 [54] features. For the SAM features, we used a similar pipeline as in Section 6.1, simply replacing the DINOv2 feature tensor T_{DINO} with the result of the SAM backbone $T_{\text{SAM}} \in \mathbb{R}^{64\times 64\times 256}$.

To be able to use CLIP feature vectors as descriptors, we take a different approach, as they are global and not local like the features resulting from the SAM or DINOv2 backbones. This method relies on CLIP features [61] extracted from object bounding box proposals. The bounding box proposals are generated by OW-DETR [26] which is a method for open world object detection based on DETR [9]. First, on annotated frames of the training sequence, CLIP features are extracted from the image cropped at the bounding box annotation. For each frame in the evaluation sequence OW-DETR bounding box proposals are generated. CLIP features are extracted from each bounding box and compared under the cosine similarity metric with the train features. The bounding box corresponding to the features that achieve the highest similarity is selected. If the maximum similarity $S_{c,max} \in \mathbb{R}$ is smaller than a threshold

$t_{sim} \in \mathbb{R}$, no object is selected. From the selected bounding box, a segmentation mask is inferred by prompting SAM with the bounding box.

We noticed that this performs badly when object instances of the same or a similar semantic class are present, as CLIP features mostly encode class-level knowledge. Further, it is sensitive to viewpoint changes, struggles with small objects and the best threshold t_{sim} varies by the type of object.

In contrast to the SAM and CLIP features, DINOv2 features used in combination with a SVM as described above are a better fit not only for distinguishing objects belonging to two different semantic classes, but also for differentiating between two objects of the same class. This approach using DINOv2 features consistently outperforms SAM and CLIP features on almost every sequence in the benchmark. The discrepancy in how these features encode semantic knowledge can be shown by visualizing the first three principal components of the pixel-wise upsampled feature vectors in an RGB image. The dimensionality reduction is fit to the masked features of an object instance. An example can be seen in Figure 5. The DINOv2 features used in the way that we outlined seem to be significantly better object descriptors than the SAM features.

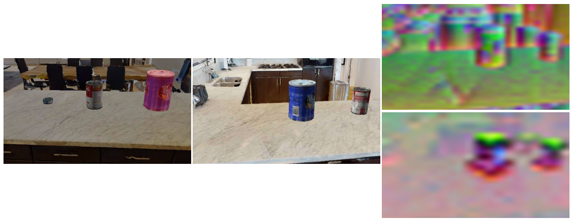

Figure 5. An example of how different models encode instance-level knowledge. Left: Image with segmentation mask used to fit a PCA transformation. Middle: Inference image (different perspective). top right: Largest three principal components of the upsampled SAM features, visualized on the entire image. bottom right: Largest three principal components of the upsampled DINOv2 features, visualized on entire image. Note that the DINOv2 features better encode the class- and instance membership of the pixels of the larger can.

Additionally, we discuss the use of the mask-refinement step using SAM outlined in Section 6.1. Performing this step greatly improves the boundary F_1-score $\mathcal{F}$ and increases the method's ability to deal with objects of the same class. However, it struggles with objects that are assembled of multiple parts, as SAM proposes masks for each part individually. An example illustrating the strengths and weaknesses of the mask-refinement step can be seen in Figure 6.

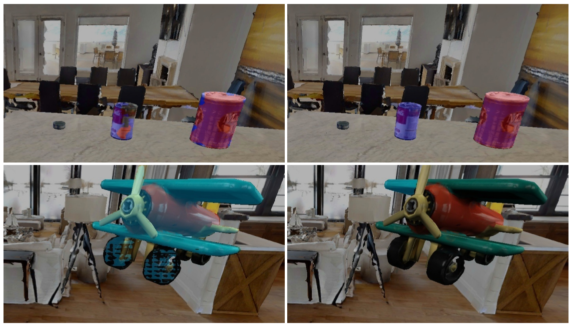

Figure 6. Selected examples of the benefits and drawbacks of using the final mask-refinement step with SAM. Left: Mask predictions using no SAM refinement. Right: Mask predictions with SAM refinement. Notice that SAM refinement helps with multiple object instances of the same class, but decreases the performance when the object consists of multiple parts.

7. Conclusion

We presented a benchmark for few-shot Video Object Instance Segmentation and Re-Identification. The goal of the benchmark is to improve object segmentation and re-identification in different scene contexts. It offers various different scenarios: single-object/multi-object segmentation, single-shot/few-shot annotation data in the training sequences, RGB/RGB-D data and 6-DoF camera poses/no camera poses available to the user. The dataset and evaluation code will be released upon publication.

We also developed a method as a baseline solution for the task, tackling the problem by forming DINOv2 [54]-based instance descriptors and training a SVM classifier in a single- or few-shot manner (depending on the scenario). We ablate on the descriptor used in the baseline method and show that among three popular state-of-the-art vision models (DINOv2, SAM and CLIP), DINOv2 embeddings are best at encoding instance knowledge.

Future work might focus on better leveraging motion and correlation in the video sequence. Another avenue would be building up richer, perhaps 3D, representations on-the-fly of the objects and using those for detection, tracking, segmentation and outlier filtering. Another possibility would be the use of a representation not directly based on point features, for example by modeling local neighborhoods or patches of features on the objects and using those to build a better, less ambiguous representation. Our goal is to build more malleable and teachable spatial AI systems and we hope to apply these techniques in such downstream systems.

8. Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017008 (Harmony).

References

[1] Yan Bai, Yihang Lou, Feng Gao, Shiqi Wang, Yuwei Wu, and Ling-Yu Duan. Group-Sensitive Triplet Embedding for Vehicle Reidentification. *IEEE Transactions on Multimedia*, 20(9):2385–2399, Sept. 2018. Conference Name: IEEE Transactions on Multimedia. 3

[2] Berta Bescos, Carlos Campos, Juan D Tardós, and José Neira. Dynaslam ii: Tightly-coupled multi-object tracking and slam. *IEEE robotics and automation letters*, 6(3):5191–5198, 2021. 1

[3] Kenneth Blomqvist, Lionel Ott, Jen Jen Chung, and Roland Siegwart. Baking in the feature: Accelerating volumetric segmentation by rendering feature maps, 2022. 6

[4] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-End Referring Video Object Segmentation with Multimodal Transformers, Apr. 2022. arXiv:2111.14821 [cs]. 2

[5] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation, 2017. 2

[6] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv:1803.00557*, 2018. 2

[7] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv:1905.00737*, 2019. 2

[8] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE Robotics & Automation Magazine*, 22(3):36–52, 2015. 2, 4

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020. 7

[10] Bowen Chen, Huan Ling, Xiaohui Zeng, Gao Jun, Ziyue Xu, and Sanja Fidler. ScribbleBox: Interactive Annotation Framework for Video Object Segmentation, Aug. 2020. arXiv:2008.09721 [cs]. 2

[11] Ho Kei Cheng and Alexander G. Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model, 2022. 2, 7

[12] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular Interactive Video Object Segmentation: Interaction-to-Mask, Propagation and Difference-Aware Fusion, Mar. 2021. arXiv:2103.07941 [cs]. 2

[13] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking Space-Time Networks with Improved Memory Coverage for Efficient Video Object Segmentation, Oct. 2021. arXiv:2106.05210 [cs]. 2

[14] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. SegFlow: Joint Learning for Video Object Segmentation and Optical Flow, Sept. 2017. arXiv:1709.06750 [cs]. 2

[15] Linyan Cui and Chaowei Ma. Sof-slam: A semantic visual slam for dynamic environments. *IEEE access*, 7:166528–166539, 2019. 1

[16] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. *arXiv preprint arXiv:2302.01872*, 2023. 2

[17] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, 2018. 3

[18] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 3

[19] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting More Attention to Video Salient Object Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8546–8556, June 2019. ISSN: 2575-7075. 2

[20] Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 701–719. Springer, 2022. 3

[21] Zhibo Fan, Jin-Gang Yu, Zhihao Liang, Jiarong Ou, Changxin Gao, Gui-Song Xia, and Yuanqing Li. Fgn: Fully guided network for few-shot instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9172–9181, 2020. 3

[22] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised Pre-training for Person Re-identification, Apr. 2021. arXiv:2012.03753 [cs] version: 2. 3

[23] Fadri Furrer, Tonci Novkovic, Marius Fehr, Abel Gawel, Margarita Grinvald, Torsten Sattler, Roland Siegwart, and Juan Nieto. Incremental object database: Building 3d models from multiple partial observations. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6835–6842. IEEE, 2018. 1

[24] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, pages 262–275, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. 3

[25] Margarita Grinvald, Fadri Furrer, Tonci Novkovic, Jen Jen Chung, Cesar Cadena, Roland Siegwart, and Juan I. Nieto. Volumetric instance-aware semantic mapping and 3d object discovery. *CoRR*, abs/1903.00268, 2019. 1

[26] Akshita Gupta, Sanath Narayan, K. J. Joseph, Salman H. Khan, Fahad Shahbaz Khan, and Mubarak Shah. OW-DETR: open-world detection transformer. *CoRR*, abs/2112.01513, 2021. 7

[27] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[28] Srikrishna karanam, Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, and Richard J. Radke. A systematic

evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):523–536, 2019. 3

[29] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions, 2019. 2

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 6, 7

[31] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J. Davison. vmap: Vectorised object mapping for neural field slam, 2023. 1

[32] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 3

[33] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, Ondrej Drbohlav, Linbo He, Yushan Zhang, Song Yan, Jinyu Yang, Gustavo Fernández, Alexander Hauptmann, Alireza Memarmoghadam, Álvaro García-Martín, Andreas Robinson, Anton Varfolomieiev, Awet Haileslassie Gebrehiwot, Bedirhan Uzun, Bin Yan, Bing Li, Chen Qian, Chi-Yi Tsai, Christian Micheloni, Dong Wang, Fei Wang, Fei Xie, Felix Jaremo Lawin, Fredrik Gustafsson, Gian Luca Foresti, Goutam Bhat, Guangqi Chen, Haibin Ling, Haitao Zhang, Hakan Cevikalp, Haojie Zhao, Haoran Bai, Hari Chandana Kuchibhotla, Hasan Saribas, Heng Fan, Hossein Ghanei-Yakhdan, Houqiang Li, Houwen Peng, Huchuan Lu, Hui Li, Javad Khaghani, Jesus Bescos, Jianhua Li, Jianlong Fu, Jiaqian Yu, Jingtao Xu, Josef Kittler, Jun Yin, Junhyun Lee, Kaicheng Yu, Kaiwen Liu, Kang Yang, Kenan Dai, Li Cheng, Li Zhang, Lijun Wang, Linyuan Wang, Luc Van Gool, Luca Bertinetto, Matteo Dunnhofer, Miao Cheng, Mohana Murali Dasari, Ning Wang, Ning Wang, Pengyu Zhang, Philip H. S. Torr, Qiang Wang, Radu Timofte, Rama Krishna Sai Gorthi, Seokeon Choi, Seyed Mojtaba Marvasti-Zadeh, Shaochuan Zhao, Shohreh Kasaei, Shoumeng Qiu, Shuhao Chen, Thomas B. Schön, Tianyang Xu, Wei Lu, Weiming Hu, Wengang Zhou, Xi Qiu, Xiao Ke, Xiao-Jun Wu, Xiaolin Zhang, Xiaoyun Yang, Xuefeng Zhu, Yingjie Jiang, Yingming Wang, Yiwei Chen, Yu Ye, Yuezhou Li, Yuncon Yao, Yunsung Lee, Yuzhang Gu, Zezhou Wang, Zhangyong Tang, Zhen-Hua Feng, Zhijun Mai, Zhipeng Zhang, Zhirong Wu, and Ziang Ma. The eighth visual object tracking vot2020 challenge results. In Adrien Bartoli and Andrea Fusiello, editors, *Computer Vision – ECCV 2020 Workshops*, pages 547–601, Cham, 2020. Springer International Publishing. 2

[34] Ratnesh Kumar, Edwin Weill, Farzin Aghdasi, and Parthsarathy Sriram. Vehicle Re-Identification: an Efficient Baseline Using Triplet Embedding, Aug. 2019. arXiv:1901.01015 [cs]. 3

[35] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022. 1

[36] Dezhuang Li, Ruoqi Li, Lijun Wang, Yifan Wang, Jinqing Qi, Lu Zhang, Ting Liu, Qingquan Xu, and Huchuan Lu. You Only Infer Once: Cross-Modal Meta-Transfer for Referring Video Object Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):1297–1305, June 2022. Number: 2. 2

[37] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M. Rehg. Video segmentation by tracking many figure-ground segments. In *2013 IEEE International Conference on Computer Vision*, pages 2192–2199, 2013. 2

[38] Xiaoxiao Li, Yuankai Qi, Zhe Wang, Kai Chen, Ziwei Liu, Jianping Shi, Ping Luo, Xiaoou Tang, and Chen Change Loy. Video object segmentation with re-identification, 2017. 2

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[40] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2167–2175, 2016. 3

[41] Lizhao Liu, Junyi Cao, Minqian Liu, Yong Guo, Qi Chen, and Mingkui Tan. Dynamic extension nets for few-shot semantic segmentation. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1441–1449, 2020. 3

[42] Yu Liu, Yvan Petillot, David Lane, and Sen Wang. Global localization with object-level semantics and topology. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4909–4915, 2019. 1

[43] Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. Learning Quality-aware Dynamic Memory for Video Object Segmentation, July 2022. arXiv:2207.07922 [cs]. 2, 7

[44] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3230–3238, 2019. 3

[45] Shiyang Lu, Rui Wang, Yinglong Miao, Chaitanya Mitash, and Kostas Bekris. Online object model reconstruction and reuse for lifelong improvement of robot manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 1540–1546. IEEE, 2022. 1

[46] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See More, Know More: Unsupervised Video Object Segmentation with Co-Attention Siamese Networks, Jan. 2020. arXiv:2001.06810 [cs]. 2

[47] Zhihe Lu, Sen He, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 8741–8750, 2021. 3

[48] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial Intelligence*, 293:103448, Apr. 2021. 2

[49] Sabarinath Mahadevan, Ali Athar, Aljoša Ošep, Sebastian Hennen, Laura Leal-Taixé, and Bastian Leibe. Making a Case for 3D Convolutions for Object Segmentation in Videos, Aug. 2020. arXiv:2008.11516 [cs]. 2

[50] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 4

[51] John McCormac, Ronald Clark, Michael Bloesch, Andrew J. Davison, and Stefan Leutenegger. Fusion++: Volumetric object-level SLAM. *CoRR*, abs/1808.08378, 2018. 1

[52] Xingyang Ni and Esa Rahtu. FlipReID: Closing the Gap between Training and Inference in Person Re-Identification, May 2021. arXiv:2105.05639 [cs] version: 1. 3

[53] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1187–1200, June 2014. 2

[54] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 6, 7, 8

[55] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23641–23651, 2023. 3

[56] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 2, 3, 4, 6

[57] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 75–91. Springer, 2016. 1

[58] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 2, 4

[59] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8):2022–2039, 2022. 2

[60] Rodolfo Quispe, Cuiling Lan, Wenjun Zeng, and Helio Pedrini. AttributeNet: Attribute Enhanced Vehicle Re-Identification, Aug. 2021. arXiv:2102.03898 [cs] version: 2. 3

[61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 7

[62] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 2, 4

[63] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 17–35, Cham, 2016. Springer International Publishing. 3

[64] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2020. 1

[65] Martin Rünz and Lourdes Agapito. Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4471–4478. IEEE, 2017. 1

[66] Renato F. Salas-Moreno, Richard A. Newcombe, Hauke Strasdat, Paul H.J. Kelly, and Andrew J. Davison. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1352–1359, June 2013. ISSN: 1063-6919. 1

[67] Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified Referring Video Object Segmentation Network with a Large-Scale Benchmark. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 208–223, Cham, 2020. Springer International Publishing. 2

[68] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized Memory Network for Video Object Segmentation, July 2020. arXiv:2007.08270 [cs]. 2

[69] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017. 3

[70] Mennatullah Siam, Naren Doraiswamy, Boris N Oreshkin, Hengshuai Yao, and Martin Jagersand. Weakly supervised few-shot object segmentation using co-attention with visual and semantic embeddings. *arXiv preprint arXiv:2001.09540*, 2020. 3

[71] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal,

Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 4

[72] Michael Strecke and Jorg Stuckler. Em-fusion: Dynamic object-level slam with probabilistic data association. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2019. 1

[73] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 4

[74] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8797–8806, 2019. 3

[75] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):1050–1065, 2020. 3

[76] Guangcong Wang, Jianhuang Lai, Peigen Huang, and Xiaohua Xie. Spatial-Temporal Person Re-identification, Dec. 2018. arXiv:1812.03282 [cs] version: 1. 3

[77] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Chuanxin Tang, Xiyang Dai, Yucheng Zhao, Yujia Xie, Lu Yuan, and Yu-Gang Jiang. Look Before You Match: Instance Understanding Matters in Video Object Segmentation, Dec. 2022. arXiv:2212.06826 [cs]. 2, 7

[78] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, Nov 2015. 2

[79] Mikolaj Wieczorek, Barbara Rychalska, and Jacek Dabrowski. On the Unreasonable Effectiveness of Centroids in Image Retrieval, Apr. 2021. arXiv:2104.13643 [cs] version: 1. 3

[80] Dongming Wu, Xingping Dong, Ling Shao, and Jianbing Shen. Multi-Level Representation Learning with Semantic Alignment for Referring Video Object Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4986–4995, June 2022. ISSN: 2575-7075. 2

[81] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as Queries for Referring Video Object Segmentation, Mar. 2022. arXiv:2201.00487 [cs]. 2

[82] Zhonghua Wu, Xiangxi Shi, Guosheng Lin, and Jianfei Cai. Learning meta-class memory for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 517–526, 2021. 3

[83] Linhui Xiao, Jinge Wang, Xiaosong Qiu, Zheng Rong, and Xudong Zou. Dynamic-slam: Semantic monocular visual localization and mapping based on deep learning in dynamic environment. *Robotics and Autonomous Systems*, 117:1–16, 2019. 1

[84] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3415–3424, 2017. 3

[85] Binbin Xu, Wenbin Li, Dimos Tzoumanikas, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Mid-fusion: Octree-based object-level multi-instance dynamic slam. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5231–5237. IEEE, 2019. 1

[86] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *CoRR*, abs/1809.03327, 2018. 2, 4

[87] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 2

[88] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3973–3981, 2015. 3

[89] Xianghui Yang, Bairun Wang, Kaige Chen, Xinchi Zhou, Shuai Yi, Wanli Ouyang, and Luping Zhou. Brinet: Towards bridging the intra-class and inter-class gaps in one-shot segmentation. *arXiv preprint arXiv:2008.06226*, 2020. 3

[90] Jun Zhang, Mina Henein, Robert Mahony, and Viorela Ila. Vdo-slam: a visual dynamic object-aware slam system. *arXiv preprint arXiv:2005.11052*, 2020. 1

[91] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1

[92] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 868–884, Cham, 2016. Springer International Publishing. 3

[93] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015. 3

[94] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild, 2017. 3

[95] Zhedong Zheng, Tao Ruan, Yunchao Wei, Yi Yang, and Tao Mei. VehicleNet: Learning Robust Visual Representation

for Vehicle Re-identification. *IEEE Transactions on Multimedia*, 23:2683–2693, 2021. arXiv:2004.06305 [cs]. 3

[96] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 350–368. Springer, 2022. 1

[97] Kai Zhu, Wei Zhai, and Yang Cao. Self-supervised tuning for few-shot segmentation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1019–1025, 2021. 3

[98] Zhihui Zhu, Xinyang Jiang, Feng Zheng, Xiaowei Guo, Feiyue Huang, Weishi Zheng, and Xing Sun. Viewpoint-Aware Loss with Angular Regularization for Person Re-Identification, Dec. 2019. arXiv:1912.01300 [cs, eess] version: 1. 3