HOIST-Former: Hand-held Objects Identification, Segmentation, and Tracking in the Wild

Supreeth Narasimhaswamy[1], Huy Anh Nguyen[1], Lihan Huang[1], and Minh Hoai[1,2]

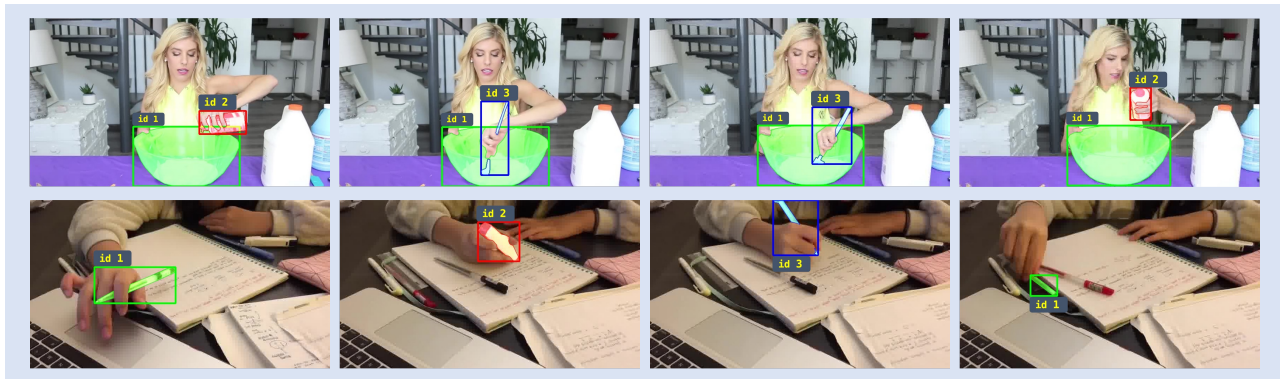[1]Stony Brook University, USA, [2] VinAI Research, Vietnam

Figure 1. Identification, segmentation, and tracking of hand-held objects.

Abstract

We address the challenging task of identifying, segmenting, and tracking hand-held objects, which is crucial for applications such as human action segmentation and performance evaluation. This task is particularly challenging due to heavy occlusion, rapid motion, and the transitory nature of objects being hand-held, where an object may be held, released, and subsequently picked up again. To tackle these challenges, we have developed a novel transformer-based architecture called HOIST-Former. HOIST-Former is adept at spatially and temporally segmenting hands and objects by iteratively pooling features from each other, ensuring that the processes of identification, segmentation, and tracking of hand-held objects depend on the hands' positions and their contextual appearance. We further refine HOIST-Former with a contact loss that focuses on areas where hands are in contact with objects. Moreover, we also contribute an in-the-wild video dataset called HOIST, which comprises 4,125 videos complete with bounding boxes, segmentation masks, and tracking IDs for hand-held objects. Through experiments on the HOIST dataset and two additional public datasets, we demonstrate the efficacy of HOIST-Former in segmenting and tracking hand-held objects. Project page: https://supreethn.github.io/research/hoistformer/index.html

1. Introduction

Humans primarily use their hands to interact with their surroundings, making the ability to segment and track hand-held objects crucial for understanding and interpreting human interactions with the environment. From monitoring a factory worker navigating through assembly tasks to evaluating the skill set of a resident doctor performing intricate medical operations, the dynamic interplay between hands and objects forms the core of many activities. Segmenting hand-held objects allows computer vision systems to identify the focal points of action, while tracking these objects over time provides a coherent understanding of sequential and complex actions. This combined capability is particularly crucial in scenarios involving multiple similar objects, as it requires the system to differentiate and monitor the path of each item to deliver contextually rich, actionable insights.

In this paper, we study the problem of jointly segmenting and tracking objects that are held and moved by hands in unconstrained videos, as illustrated in Fig. 1. Specifically, given an input video composed of a sequence of frames, we consider all portable objects that are held by hands at any point within these frames. Suppose there are a number of such object instances. For each object instance, our objective is to produce a series of binary segmentation masks corresponding to each frame, such that the mask for a particular frame is empty if the object instance is not being

held by a hand in that frame. Note that our study is limited to portable objects that can be held and moved by hand, excluding non-portable objects like furniture.

Segmenting and tracking hand-held objects involves three complex sub-tasks: first, identifying the object in the grasp of a hand from among several; second, accurately segmenting that object; and third, maintaining its track throughout the video. Identifying hand-held objects is challenging because the mere overlapping of hand and object segments does not confirm a hold, as they may overlap in a 2D view without actual 3D contact. The segmentation task is complicated by heavy occlusion of objects by the hands, resulting in non-contiguous segments and the need to account for various object shapes and appearances in an open-world setting, regardless of category. Tracking is made difficult by the rapid movement of hands, which can drastically alter the position of the object from one frame to the next, potentially causing incorrect associations of object identity over time. Furthermore, while an object can be visible throughout the video, being hand-held is not a persistent characteristic; an object might be held at one moment, released the next, and picked up again subsequently. At times when the object is not in hand, segmentation and tracking should cease, notwithstanding its visibility. Despite these breaks in continuity, the object should maintain a consistent identifier throughout the video. Some of the challenges described here are illustrated in Fig. 1.

To address the aforementioned challenges of **H**and-held **O**bjects **I**dentification **S**egmentation and **T**racking, we propose HOIST-Former. This model builds on the transformer-based image and video segmentation method Mask2Former [7, 8], enhancing it with an innovative decoder architecture designed to overcome its limitations. Although Mask2Former is a leading method for object segmentation and tracking, its reliance on a predefined set of object categories makes it unsuitable for the segmentation and tracking of arbitrary hand-held objects in an open-world setting. Furthermore, Mask2Former's methodology, grounded solely on categorical membership and object visibility, is inadequate in scenarios where segmentation and tracking need to be initiated, paused, and resumed based on additional criteria, such as the hand-held status of an object, which is the central concern of this paper.

HOIST-Former addresses the limitations of Mask2Former with a novel Hand-Object Transformer decoder, which iteratively localizes hands and hand-held objects by mutually pooling features, effectively conditioning the identification and segmentation of the hand-held objects based on the appearance of hands and their surrounding context. Specifically, from a given set of video frames, a backbone network extracts low-resolution spatio-temporal features. These features are then gradually upscaled by a pixel decoder to produce high-resolution, per-pixel spatio-temporal embeddings. Finally, the Hand-Object Transformer decoder utilizes these high-resolution embeddings to generate spatio-temporal segmentation masks for both hands and the objects they are holding.

To train and evaluate HOIST-Former, we have collected and annotated a large-scale in-the-wild video dataset, named HOIST, a contribution of this work. Specifically, for each hand-held object in the video, we annotate its segmentation mask and assign a tracking instance ID that persists throughout the video. Our dataset comprises 4,228 videos with approximately 85,000 frames in total. The HOIST dataset includes numerous videos featuring hand-held objects within challenging and unconstrained environments, which can be used to train robust methods for hand-held object segmentation and tracking.

Experiments conducted on the HOIST dataset, along with two other datasets, reveal that HOIST-Former achieves superior results in the segmentation and tracking of hand-held objects.

2. Related Work

Hand Analysis. Hand analysis is crucial for many computer vision applications, so various problems have been studied. For example, there are works on detecting hands in images [3, 21, 22, 26, 29, 34, 35, 46, 55]. There are also prior works that estimate hand contact [2, 28, 30] and localize the contact objects in images using bounding boxes [39]. Some works analyze hands by estimating their poses [4, 5, 16, 19, 24, 25, 37, 38, 50, 56, 57], tracking them in videos [18, 27, 40, 42–44, 51], and even generating them [32, 33]. However, none of these works address segmenting and tracking hand-held objects.

Some works address the task of jointly estimating and tracking hand and object poses in videos [6, 14, 15, 23]. However, they do not focus on segmenting or tracking hand-held objects as their main goal is pose estimation. Moreover, these methods typically deal with egocentric videos in constrained indoor environments with simple backgrounds and limited object categories. There is prior work on segmenting hands and hand-held objects [52], but tracking is not addressed. Our work is more related to video hand-held object segmentation [12], but this task requires segmentation masks of the objects in the first frame as an additional input and propagates these masks in subsequent frames. In contrast, our method only requires video frames as input to jointly segment and track hand-held objects. Another related problem is Video Instance Segmentation (VIS) [49]. However, VIS methods such as Mask2Former [7] segment and track objects from a predefined categories and are unsuitable for localizing arbitrary hand-held objects in open-world settings. Conversely, the proposed method HOIST-Former can segment and track arbitrary hand-held objects.

Datasets. One of our key contributions is the creation of a novel annotated dataset specifically designed for segmenting and tracking hand-held objects, which offers unique features not found in existing datasets. While many hand datasets are available, they are unsuitable for our purpose, either lacking video data [29–31, 41] or missing annotations for hand-held objects [18]. The 100DOH [39] is a video dataset that includes bounding box annotations for hands and hand-contact objects, but it lacks tracking annotations and randomly samples frames for annotation. EgoHOS [52] provides segmentation masks for hands and hand-held objects in frames from egocentric videos. However, it does not offer consistent tracking annotations, assigning the same instance ID to different objects held by a hand at different times, thereby limiting its utility for our purpose.

Our dataset shares similarities with the VISOR benchmark [12], which annotates Epic-Kitchen videos [11] with segmentation masks and tracking IDs for objects interacting with hands. However, VISOR includes all types of interacting objects, not just those held by hands. Another limitation of VISOR is its focus solely on egocentric kitchen videos, whereas our HOIST dataset includes a wide range of diverse, unconstrained videos from various settings. While there are general-purpose video object segmentation datasets like [45, 48, 49], they do not specifically annotate hand-held objects; we manually selected 80 videos featuring hand-held objects from these datasets for additional evaluation. However, this quantity is insufficient for training, highlighting the necessity of our dataset.

3. HOIST-Former

This section describes HOIST-Former, a novel network designed to jointly segment and track hand-held objects. The network takes as input a video $\mathcal{V} \in \mathbb{R}^{T \times H \times W \times 3}$, consisting of T frames with spatial dimensions $H \times W$ and three color channels. For each object O in the video that is held by a hand in at least one of the frames, the network outputs a binary 3D tensor $\mathbf{M} \in \{0,1\}^{T \times H \times W}$, representing the spatio-temporal locations of the object. If object O is not held by a hand in a frame, then its corresponding 2D binary mask will be devoid of any segment, as our task is solely on segmenting and tracking the object when it is being held. Note that object O retains a unique instance ID throughout the video, even if certain 2D segmentation masks become empty at times due to complete occlusion or temporary disruption of the hand-held status.

Inspired by the success of Mask2Former in video instance segmentation, we designed HOIST-Former with a similar overall architectural framework, consisting of three main components: a backbone network, a pixel decoder, and a transformer decoder, as depicted in Fig. 2. First, the input video is fed into the backbone network to extract low-resolution features. These features are then upscaled by the
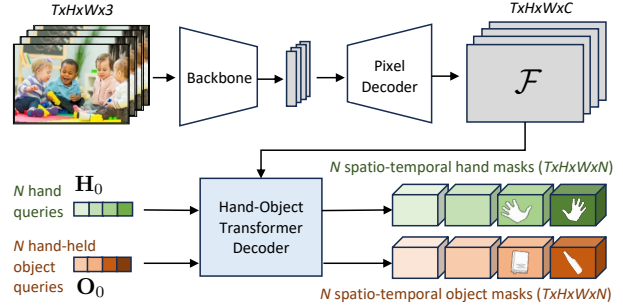

Figure 2. HOIST-Former consists of a backbone network, a pixel decoder, and a transformer decoder. The input video is initially processed through the backbone network and the pixel decoder to generate high-resolution spatio-temporal features $\mathcal{F}$. The transformer decoder operates on $\mathcal{F}$, decoding a set of N hand queries and their corresponding object queries, resulting in N spatio-temporal hand masks and corresponding object masks.

pixel decoder to generate high-resolution spatio-temporal features $\mathcal{F}$. The transformer decoder operates on $\mathcal{F}$, processing hand and object queries iteratively. These queries, starting as an initial set of learnable C-dimensional feature vectors representing potential hands or hand-held objects in the video, are iteratively updated by the transformer decoder. The spatio-temporal binary mask predictions for both object and hand tracks are decoded from these hand and object queries in conjunction with the high-resolution spatio-temporal features $\mathcal{F}$.

In the remainder of this section, we will describe the innovative transformer decoder of HOIST-Former, called Hand-Object Transformer Decoder. Following this, we will describe how HOIST-Former can be trained.

3.1. Hand-Object Transformer Decoder

The Hand-Object Transformer Decoder is an innovative component, designed to systematically determine the positions of hands and hand-held objects through an iterative and collaborative feature pooling process. This effectively conditions the identification and segmentation of hand-held objects based on the appearance of hands and their immediate environment. This innovative transformer decoder allows us to segment and track arbitrary hand-held objects in an open-world setting, satisfying selection criteria that extend beyond categorical membership and object visibility.

Given the spatio-temporal features $\mathcal{F}$, we start with N learnable hand queries $\mathbf{H}_0 \in \mathbb{R}^{N \times C}$ and object queries $\mathbf{O}_0 \in \mathbb{R}^{N \times C}$. These queries function similarly to region proposals [36] and can generate spatio-temporal segmentation masks for hands and hand-held objects by attending to the features $\mathcal{F}$. Similar to Mask2Former [7], these queries are processed by L transformer decoder layers to produce segmentation masks. However, unlike Mask2Former, our focus is on segmenting objects based on their interaction
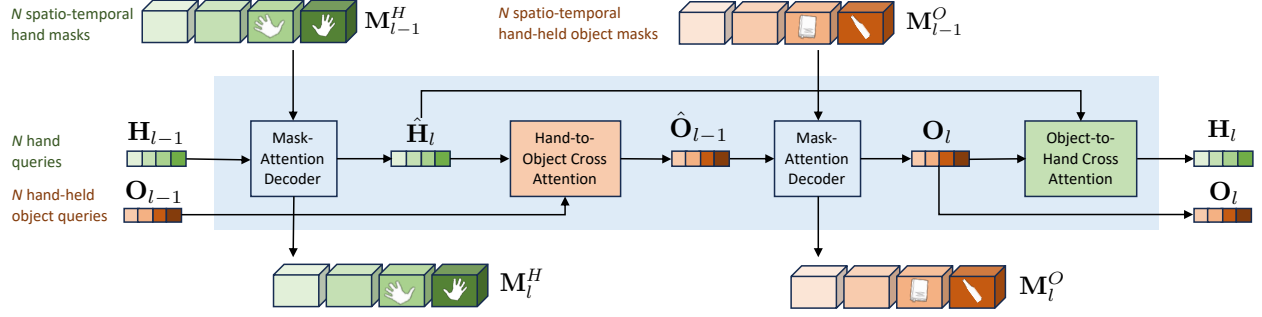
Figure 3. The Hand-Object Transformer Decoder features a network architecture with L layers. This figure demonstrates the operational flow of a single layer, which includes two mask attention modules and two cross-attention modules. The inputs for this layer consist of N sets of four elements each: a hand query, an object query, a spatio-temporal hand mask, and a spatio-temporal object mask. The outputs of this layer are the correspondingly updated versions of these entities.

with hands, regardless of their category or visibility. We therefore condition object segmentation on the appearance of hands and their surrounding context. Conversely, identifying hand-held objects aids in localizing hands. Therefore, we condition the hand segmentation upon hand-held objects. These dual tasks are achieved by mutually pooling information between hand and object queries.

The operational flow of the Hand-Object Transformer Decoder is illustrated in Fig. 3. It encompasses four principal operations: an initial mask attention operation, succeeded by a cross-attention operation, another mask attention operation, and finally concluding with a second cross-attention operation. The formal representation of these four steps is provided in the following equations:

$$\hat{\mathbf{H}}_l, \mathbf{M}_l^H := MaskAtt(\mathbf{H}_{l-1}, \mathbf{M}_{l-1}^H \mid f^{H \to H}), \quad (1)$$

$$\hat{\mathbf{O}}_{l-1} := CrossAtt(\mathbf{O}_{l-1}, \hat{\mathbf{H}}_l \mid f^{H \to O}), \quad (2)$$

$$\mathbf{O}_l, \mathbf{M}_l^O := MaskAtt(\hat{\mathbf{O}}_{l-1}, \mathbf{M}_{l-1}^O \mid f^{O \to O}), \quad (3)$$

$$\mathbf{H}_l := CrossAtt(\hat{\mathbf{H}}_l, \mathbf{O}_l \mid f^{O \to H}) \quad (4)$$

In the above, $CrossAtt$ and $MaskAtt$ refer to the cross-attention and mask-attention modules, respectively. To elaborate, the function $CrossAtt(\mathbf{X}, \mathbf{Y} | f)$ takes two inputs, X and Y, where f symbolizes a trio of linear functions $f_Q(\cdot), f_K(\cdot), f_V(\cdot)$ (for query, key, value) with learnable parameters. The function $CrossAtt$ draws information from Y to X, resulting in an updated version $\mathbf{X}'$ of $\mathbf{X}$. This process is defined as follows:

$$\mathbf{X}' = softmax\left(f_Q(\mathbf{X}) f_K(\mathbf{Y})^T\right) f_V(\mathbf{Y}). \quad (5)$$

The function $MaskAtt(\mathbf{X}, \mathbf{M} | f)$ operates with two inputs: the query set X and the collection of spatio-temporal binary masks $\mathbf{M}$. It outputs the revised queries X' and the updated masks $\mathbf{M}'$. Also in this context, f represents a set of three linear functions: $f_Q(\cdot), f_K(\cdot)$, and $f_V(\cdot)$, each characterized by learnable parameters. The update equation

for the queries is:

$$\mathbf{X}' = softmax(\mathcal{M} + f_Q(\mathbf{X}) f_K(\mathcal{F})^T) f_V(\mathcal{F}) + \mathbf{X}. \quad (6)$$

In the above, $\mathcal{F}$ is the high-resolution spatio-temporal feature maps from the pixel decoder. The 4D attention mask $\mathcal{M}$ is determined by the set of 3D binary masks $\mathbf{M}$. The value of $\mathcal{M}$ at location (t, y, x, n) is:

$$\mathcal{M}(t, y, x, n) = \begin{cases} 0 & \text{if } \mathbf{M}(t, y, x, n) = 1, \\ -\infty & \text{otherwise.} \end{cases} \quad (7)$$

Another output of the $MaskAtt(\mathbf{X}, \mathbf{M}|f)$ function is the set of updated 3D masks $\mathbf{M}'$, which are obtained by using dot products between query features $\mathbf{X}$ and spatio-temporal features $\mathcal{F}$. We refer the reader to Cheng et al. [7] for more details.

Note that the Hand-Object Transformer Decoder is composed of four Attention modules: two cross-attention and two mask-attention modules. Each of these modules is equipped with three linear functions corresponding to query, key, and value, resulting in a total of 12 linear functions, each featuring learnable weights.

3.2. Training Losses

We train HOIST-Former using the multi-task loss:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{mask} + \mathcal{L}_{dice}. \quad (8)$$

The term $\mathcal{L}_{cls}$, $\mathcal{L}_{mask}$, and $\mathcal{L}_{dice}$ denote the class loss, mask loss, and dice losses, respectively. Both the mask loss and dice loss comprise a linear combination of individual losses calculated for hands, objects, and contact masks

$$\mathcal{L}_{mask} = \lambda_1 \mathcal{L}_{mask}^H + \lambda_2 \mathcal{L}_{mask}^O + \lambda_3 \mathcal{L}_{mask}^C, \quad (9)$$

$$\mathcal{L}_{dice} = \lambda_4 \mathcal{L}_{dice}^H + \lambda_5 \mathcal{L}_{dice}^O + \lambda_6 \mathcal{L}_{dice}^C. \quad (10)$$

The mask loss and dice loss each include a loss term specific to the contact mask, which represents the interaction area between a hand-held object and the hand holding
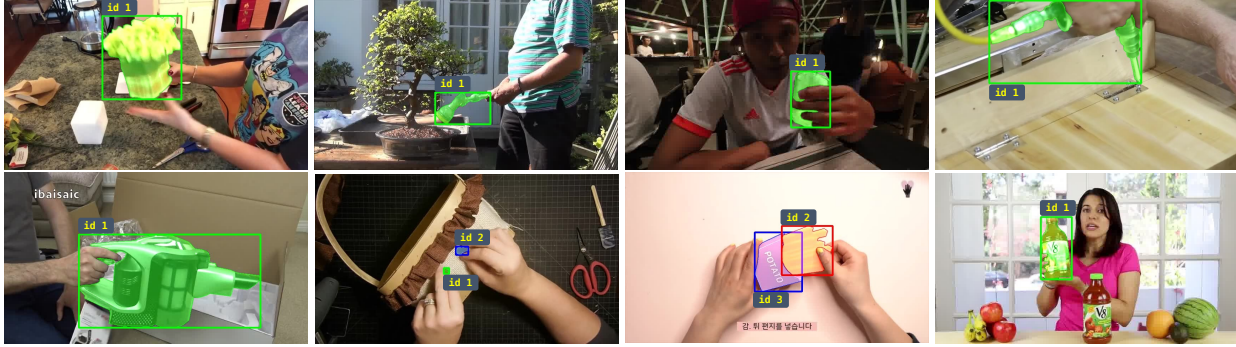
Figure 4. **Sample frames from HOIST.** HOIST dataset contains videos with diverse scenes, camera views, object sizes, and occlusions.

it. This contact mask is derived after acquiring the spatio-temporal hand and object masks at each decoder layer l. Encoding the key interaction zone between hands and objects, the contact mask plays a vital role in accurately localizing hand-held objects. During training, we incorporate contact losses to guide HOIST-Former's focus towards areas where hands and objects make contact. The effectiveness of including contact losses is demonstrated through empirical results in our experimental section.

The λ_i's in Eq. (9) and Eq. (10) are tunable hyperparameters that control the relative strength of individual losses.

4. HOIST Dataset

This section describes a novel and challenging dataset we have collected to develop and evaluate hand-held object segmentation and tracking methods.

Dataset Source. Our goal is to compile a sufficiently large and diverse video dataset to develop methods for detecting, segmenting, and tracking hand-held objects. Specifically, we aim for the dataset to satisfy several criteria. First, it should include videos of everyday activities in diverse indoor and outdoor environments. Second, the videos should feature people interacting with a wide range of objects. Third, we aim to develop methods suitable for unconstrained videos showcasing hands interacting with multiple objects, involving scenarios where objects are held, released, and picked up in various sequences; hence, the dataset should include such types of videos. To fulfill these requirements, we selected YouTube videos from the 100DOH dataset [39], which consists of 100K frames from 27.3K videos. While 100DOH provides bounding box annotations for hands and hand-contact objects, it lacks segmentation or tracking annotations.

To construct HOIST, we initially focused on 100DOH frames featuring portable object annotations. We identified the publicly available YouTube videos corresponding to these frames as potential candidates. For each selected frame, we used shot boundary information to locate contiguous segments within the corresponding video, eliminat-

	Train	Valid	Test	Total
# Videos with no object	345	0	0	345
# Videos with 1 object	1896	99	178	2173
# Videos with 2 objects	1117	61	100	1278
# Videos with ≥ 3 objects	367	22	43	432
# Videos - total	3725	182	321	4228
# Frames	74527	3470	5973	83970
# Object Instances	5393	310	522	6225

Table 1. **HOIST dataset statistics.**

ing any duplicate segments in the process. The videos were post-processed to an approximate length of three seconds each. We extracted the videos at a rate of six frames per second, resulting in about 18 frames per video, and resized the frames to maintain a shorter side of 480 pixels. In total, our dataset comprises 4,228 videos with 83,970 frames.

Annotation and Statistics. In the HOIST dataset, we annotate every instance of hand-held objects in the videos. Specifically, we provide annotations for each object's bounding box, segmentation mask, and a unique instance ID to facilitate tracking. We omit the object's bounding box and mask annotations in frames where the object is not held by a hand. The annotation process begins with the manual annotation of bounding boxes and tracking IDs for hand-held objects. Subsequently, we divide the videos into train, validation, and test sets, adhering to the splits defined in the 100DOH dataset. For the test and validation videos, we manually annotate the segmentation masks of hand-held objects. In contrast, for the training videos, we generate segmentation masks by applying the Segment Anything [20] model to the manually annotated bounding boxes. Table 1 provides some key statistics of the HOIST dataset and Fig. 4 illustrates some sample annotated frames.

5. Experiments

This section outlines the extensive experiments we conducted on several datasets. We begin by describing the datasets used, the evaluation metrics employed, and the

Method	First identification		Continued tracking at Frame t		Evaluation dataset		
	Box	Mask	Box	Mask	HOIST	VISOR	UVO
BL-A	N/A	GT	N/A	STCN	48.9	33.1	40.4
BL-B	GT	SAM	N/A	STCN	41.6	23.7	33.7
BL-C	100DOH	SAM	N/A	STCN	23.3	8.0	15.9
BL-D	GT	SAM	GT + IoUTracker	SAM	4.2	10.0	1.8
BL-E	100DOH	SAM	100DOH + IoUTracker	SAM	1.2	5.2	1.3
BL-F	GT	SAM	GT + StrongSORT	SAM	20.2	1.3	14.5
BL-G	100DOH	SAM	100DOH + StrongSORT	SAM	4.7	0.5	2.0
BL-H	GT	SAM	GT + StrongSORT++	SAM	19.3	1.3	14.4
BL-I	100DOH	SAM	100DOH + StrongSORT++	SAM	5.4	0.5	3.5
BL-J	GT	SAM	GT + ByteTrack	SAM	26.0	8.1	18.4
BL-K	100DOH	SAM	100DOH + ByteTrack	SAM	7.8	3.2	2.8
BL-L	GT	SAM	GT + MixFormer	SAM	39.7	24.0	29.7
BL-M	100DOH	SAM	100DOH + MixFormer	SAM	22.0	11.0	18.1
BL-N	GT	SAM	GT + GTR	SAM	25.8	21.3	21.8
Mask2Former	N/A	N/A	N/A	N/A	51.8	42.8	63.1
HOIST-Former	N/A	N/A	N/A	N/A	**56.4**	**46.6**	**66.7**

Table 2. Average Precision (AP) for diverse approaches to identifying, segmenting, and tracking hand-held objects. The first 14 methods (from BL-A to BL-N) vary in their strategies for initially identifying and then continuously tracking the bounding box and segmentation mask at Frame t. We consider several alternatives, from automatic methods that leverage state-of-the-art techniques in hand-held object detection (100DOH [39]), object segmentation (SAM [20]), segmentation mask propagation (STCN [9]), and bounding box tracking (IoUTracker [1, 17], StrongSORT [13], ByteTrack [53], MixFormer [10], GTR [54]), to hypothetical scenarios where an oracle supplies ground truth bounding boxes or segmentation masks. This comprehensive analysis results in an extensive list of methods, as detailed in the table; however, all are outperformed by HOIST-Former by a wide margin.

Method	AP
HOIST-Former	56.4
HOIST-Former w/o Hand-to-Object Attn.	52.7
HOIST-Former w/o Object-to-Hand Attn.	53.5
HOIST-Former w/o Contact Loss	54.8

Table 3. **Ablation study outcomes** for HOIST-Former evaluated on the HOIST dataset's test data. Removing either the Hand-to-Object Attention module, the Object-to-Hand Attention module, or the Contact Loss would significantly degrade the performance of HOIST-Former.

training details.

5.1. Datasets

In addition to evaluating our method's effectiveness on the proposed HOIST dataset, we also perform experiments on selected videos from the VISOR [12] and UVO [45] datasets that are amendable for hands and hand-held objects evaluation. This section details the filtering steps we employed to select and prepare these videos.

VISOR is an egocentric video semantic segmentation dataset derived from EPIC-KITCHEN [11], centered around active objects involved in the user's actions. The dataset offers two types of annotations: manually curated, high-quality sparse annotations, and dense annotations gen-

erated through interpolation. Due to the unreliability of dense annotations, we only use the sparse annotations.

In its sparse annotations, each VISOR video clip typically encompasses three actions with a total of six annotated frames. From the annotations, we identified objects in direct contact with hands. Initially, we manually reviewed the object categories and names, excluding immobile categories like stovetops, dishwashers, and freezers. For certain remaining categories, we applied a further filter to eliminate excessively large objects, setting the mask area threshold for an "overly large" object at 0.3. Upon a detailed review of the dataset, we corrected several annotation errors, particularly those involving gloves and hands. The final step is to exclude object instances not in contact with hands by checking the overlap between diluted hands and instance masks. The VISOR dataset is divided into two subsets: originally, the Train subset had 5322 clips, and the Valid subset had 1251 clips. After our filtering steps, we are left with 5022 clips in the Train subset, featuring 31.4K frames and 34.5K hand-held object instances with mask annotations, and 1162 clips in the Valid subset, encompassing 7.3K frames with 8K hand-held object instances.

UVO is a general-purpose video segmentation dataset featuring annotations for various object categories, along with their segmentation masks and tracking IDs. To extract

videos containing hand-held objects for segmentation and tracking, we employ a three-step process. First, we select a subset of UVO object categories likely to be hand-held, excluding categories like vehicles, people, and furniture. Second, we manually review videos under these potential hand-held object categories, retaining only those that actually feature hand-held objects. Third, we adjust the segmentation masks of these hand-held objects, setting them to empty masks in frames where they are not held by hands. This process results in 90 videos from the training set and 80 from the validation set. Given the insufficiency of 90 videos for training purposes, we opt to use only the 80 videos from the validation set for evaluation.

5.2. Evaluation metric and training details

We measure the joint performance of hand-held object segmentation and tracking using the Average Precision (AP). We consider a detected spatio-temporal mask $\mathbf{M}$ as a true positive if its Intersection over Union (IoU) with a ground-truth spatio-temporal mask is greater than 0.5. Specifically, given a detected spatio-temporal mask $\mathbf{M} = (\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_T)$ and a ground-truth spatio-temporal mask $\mathbf{M}^{gt} = (\mathbf{M}_1^{gt}, \mathbf{M}_2^{gt}, \cdots, \mathbf{M}_T^{gt})$, we compute the IoU as follows:

$$\text{IoU}\left(\mathbf{M}, \mathbf{M}^{gt}\right) = \frac{\sum_{t=1}^{T} \left| \mathbf{M}_t \cap \mathbf{M}_t^{gt} \right|}{\sum_{t=1}^{T} \left| \mathbf{M}_t \cup \mathbf{M}_t^{gt} \right|}. \qquad (11)$$

We implement HOIST-Former using Detectron2 [47]. We set the hyperparmeters λ_2, λ_5 in Eq. (9) and Eq. (10) to be 5. We set the rest of λ_i's to be 0.001. We train HOIST-Former on eight 80GB GPUs using AdamW optimizer with an initial learning rate of 0.0001.

5.3. Comparison Methods

Our objective is to effectively identify, segment, and track hand-held objects. Any viable approach must address these functions in some manner. We evaluate HOIST-Former against a comprehensive range of methods, each representing different approaches to these tasks.

The first task involves locating a hand-held object's bounding box. The second requires generating a segmentation mask, and the third entails maintaining identity consistency across frames, either by matching detected object instances or through propagation. These tasks hinge on the initial identification and subsequent tracking of the bounding box and segmentation mask at Frame t. We consider two options for obtaining the initial bounding box: automatic detection using the 100DOH detector [39], or using a human-annotated ground truth bounding box. For initial segmentation, one option is using SAM with the bounding box as a prompt; alternatively, methods that use ground truth masks are also considered. For the continued tracking of the segmented object at Frame t, the state-of-the-art

video object segmentation method STCN [9] can be employed. Alternatively, this process can be split into tracking the bounding box first and then applying SAM segmentation, with the tracked box serving as the prompt. This division leads to further decisions regarding how the bounding box at Frame t is detected and linked to its previous appearance, with options including the automatic 100DOH detector or the oracle-based ground truth bounding box. To connect detections, we experiment with IoUTracker [1, 17], StrongSORT [13], ByteTrack [53], GTR [54] and MixFormer [10], combining both automatic and oracle-based methods in detection, segmentation, and tracking.

In contrast, HOIST-Former operates by producing a spatio-temporal binary segmentation mask for each hand-held object, treating the video as a singular entity and thus avoiding the need to distinctly separate these tasks. This sets HOIST-Former apart from the other baseline and oracle methods discussed. A notable comparison in terms of operation is the Mask2Former method, which we also benchmark against here.

5.4. Experimental Results

Tab. 2 reports the performance of various methods for identifying, segmenting, and tracking hand-held objects. This analysis, detailed in Sec. 5.3, includes comparisons with Mask2Former and HOIST-Former. Among the methods excluding Mask2Former and HOIST-Former, BL-A emerges as the most effective. However, BL-A is only viable when a user manually identifies and segments the hand-held object in the first frame, a task requiring significant time and effort. Choosing to only draw the bounding box leads to a notable performance drop of 7.3% on the HOIST dataset (Method BL-B). If the initial bounding box is determined using the 100DOH detector instead of manual annotation (Method BL-C), performance drastically decreases, halving the AP. This highlights the critical need for accurate hand-held object bounding box detection, a task where the 100DOH detector falls short. The other methods in Tab. 2, following a similar approach of bounding box detection, linking, and segmentation using SAM with bounding box prompts, are also evaluated. BL-J, assuming ground truth bounding boxes, performs best among these. However, when ground truth bounding boxes are replaced by those detected by 100DOH, as in method BL-K, there is a significant decline in performance. MixFormer proved to be the most effective among various linking methods tested.

HOIST-Former emerges as the leading method, significantly surpassing others, especially those ranging from BL-A to BL-M. This achievement is particularly noteworthy given that some methods have the advantage of accessing ground truth bounding boxes or segmentation masks, a privilege not afforded to HOIST-Former. Mask2Former ranks as the second-best method. Notably, this is not the

Figure 5. **Illustrative qualitative results of HOIST-Former**. Each row displays selected frames from a single video. Within each row, a distinct hand-held object is assigned a unique tracking ID and is consistently represented in the same color.

standard Mask2Former network designed for segmenting and tracking predefined object categories. Rather, it is a re-trained version of the vanilla Mask2Former architecture, specifically adapted for a single consolidated class of hand-held objects. While Mask2Former outperforms many contenders in Tab. 2, it is still surpassed by HOIST-Former. This success of HOIST-Former is attributed to its advanced Hand-Object Transformer Decoder, which proficiently pools information and bases decisions on both hands and objects–key factors for determining hand-held status. Both Mask2Former and HOIST-Former models are trained using the HOIST training data. For the VISOR dataset, characterized by a much sparser set of video frames compared to HOIST's framerates, the models are trained on VISOR training data. The performances reported in the table for VISOR reflect this specialized training.

5.5. Ablation studies and qualitative results

A key innovation in HOIST-Former lies in its utilization of hand context to ascertain hand-held status. This novel concept is embodied in the Hand-Object Transformer Decoder, a unique Transformer decoder featuring two cross-attention modules: Hand-to-Object and Object-to-Hand, crafted for bidirectional context integration and decision-making. Additionally, the significance of this mutual context is highlighted by considering the contact boundary between hand and object segmentation, reinforced through the implementation of Contact Loss. These elements are purposefully integrated to ensure the network has sufficient information for accurate decision-making. Our ablation study, detailed in Tab. 3, evaluates the criticality of these components. The results clearly demonstrate that removing either the Hand-to-Object Attention module, the Object-to-Hand Attention module, or the Contact Loss significantly impacts HOIST-Former's performance.

Fig. 5 shows results of HOIST-Former on two videos. The first row shows a case where the object is segmented only when it is hand-held. The second row shows a case where the object is asssigned the same tracking ID by

Figure 6. **Some failure cases of HOIST-Former**. The first row corresponds to two frames from different videos. The last row corresponds to two frames from the same video.

HOIST-Former even after the object disappears for a while and contacts the hand later. Fig. 6 shows some failure cases. The top row highlights cases where only a part of the object is segmented; the unsegmented part of the object contains extremely thin region and therefore hard to segment. The second row shows a case where a different object is assigned a previously assigned ID.

6. Conclusions

In this paper, we tackled the task of identifying, segmenting, and tracking hand-held objects. We introduced HOIST-Former, an innovative transformer-based architecture, adept at segmenting hands and objects by pooling features based on their positions and context. This approach is further refined with a contact loss that emphasizes areas where hands contact objects. We also presented the HOIST dataset, comprising 4,125 in-the-wild videos with comprehensive annotations. Our experiments on HOIST and two other public datasets showcased HOIST-Former's effectiveness in hand-held object segmentation and tracking.

References

[1] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2017. 6, 7

[2] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *The European Conference on Computer Vision (ECCV)*, 2020. 2

[3] Patrick Buehler, Mark Everingham, Daniel P Huttenlocher, and Andrew Zisserman. Long term arm and hand tracking for continuous sign language tv broadcasts. In *Proceedings of the British Machine Vision Conference*, 2008. 2

[4] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[5] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[6] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[7] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2former for video instance segmentation. 2021. 2, 3, 4

[8] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 2

[9] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *Advances in Neural Information Processing Systems(NeurIPS)*, 2021. 6, 7

[10] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13608–13618, 2022. 6, 7

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. 2021. 3, 6

[12] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 2, 3, 6

[13] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deep-sort great again. *IEEE Transactions on Multimedia*, 2023. 6, 7

[14] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[15] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[16] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[17] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 6, 7

[18] Mingzhen Huang, Supreeth Narasimhaswamy, Saif Vazir, Haibin Ling, and Minh Hoai. Forward propagation, backward regression and pose association for hand tracking in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[19] Dong Uk Kim, Kwang In Kim, and Seungryul Baek. End-to-end detection and pose estimation of two interacting hands. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 5, 6

[21] M. Kölsch and M. Turk. Robust hand detection. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, 2004. 2

[22] M Pawan Kumar, Andrew Zisserman, and Philip HS Torr. Efficient discriminative learning of parts-based models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2009. 2

[23] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[24] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[25] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 2

[26] Arpit Mittal, Andrew Zisserman, and P. H. S. Torr. Hand detection using multiple proposals. In *Proceedings of the British Machine Vision Conference*, 2011. 2

[27] Franziska Mueller, Dushyant Mehta, Oleksandr Sotny-chenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE/ CVF International Conference on Computer Vision Workshops (ICCVW)*, 2017. 2

[28] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[29] Supreeth Narasimhaswamy, Zhengwei Wei, Yang Wang, Justin Zhang, and Minh Hoai. Contextual attention for hand detection in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[30] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Hoai. Detecting hands and recognizing physical contact in the wild. In *Advances in Neural Information Processing Systems*, 2020. 2

[31] Supreeth Narasimhaswamy, Thanh Nguyen, Mingzhen Huang, and Minh Hoai. Whose hands are these? hand detection and hand-body association in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[32] Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, and Saayan Mitra. Text-to-hand-image generation using pose- and mesh-guided diffusion. In *IEEE/CVF International Conference on Computer Vision (ICCV), International Workshop on Observing and Understanding Hands in Action*, 2023. 2

[33] Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, Saayan Mitra, and Minh Hoai. Hand-iffuser: Text-to-image generation with realistic hand appearances. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[34] Eng-Jon Ong and Richard Bowden. A boosted classifier tree for hand shape detection. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2004. 2

[35] Pramod Kumar Pisharady, Prahlad Vadakkepat, and Ai Poh Loh. Attention based detection and recognition of hand postures against complex backgrounds. *Internation Journal on Computer Vision*, 2013. 2

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. 3

[37] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2017. 2

[38] Viktor Rudnev, Vladislav Golyanik, Jiayi Wang, Hans-Peter Seidel, Franziska Mueller, Mohamed Elgharib, and Christian Theobalt. Eventhands: Real-time neural 3d hand pose estimation from an event stream. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[39] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 5, 6, 7

[40] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, Daniel Freedman, Eyal Krupka, Andrew Fitzgibbon, Shahram Izadi, and Pushmeet Kohli. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015. 2

[41] Roy Shilkrot, Supreeth Narasimhaswamy, Saif Vazir, and Minh Hoai. WorkingHands: A hand-tool assembly dataset for image segmentation and activity mining. In *Proceedings of British Machine Vision Conference*, 2019. 3

[42] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[43] Srinath Sridhar, Franziska Mueller, Michael Zollhoefer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[44] Robert Y. Wang and Jovan Popović. Real-time hand-tracking with a color glove. *ACM Transactions on Graphics*, 2009. 2

[45] W. Wang, M. Feiszli, H. Wang, and D. Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3, 6

[46] Ying Wu, Qiong Liu, and Thomas S Huang. An adaptive self-organizing color segmentation algorithm with application to robust real-time human hand localization. In *Proceedings of the Asian Conference on Computer Vision*, 2000. 2

[47] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 7

[48] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. 2018. 3

[49] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[50] Linlin Yang, Shicheng Chen, and Angela Yao. Semihand: Semi-supervised hand pose estimation with consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[51] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias

Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020. 2

[52] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, 2022. 2, 3

[53] Yifu Zhang, Pei Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, 2021. 6, 7

[54] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 6, 7

[55] Xiaojin Zhu, Jie Yang, and Alex Waibel. Segmenting hands of arbitrary color. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2000. 2

[56] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. 2

[57] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2