This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Affordance Grounding from Demonstration Video to Target Image

Joya Chen Difei Gao Kevin Qinghong Lin Mike Zheng Shou†
Show Lab, National University of Singapore
{joyachen,qinghonglin}@u.nus.edu {daniel.difei.gao,mike.zheng.shou}@gmail.com

Abstract

*Humans excel at learning from expert demonstrations and solving their own problems. To equip intelligent robots and assistants, such as AR glasses, with this ability, it is essential to ground human hand interactions (*i.e.*, affordances) from demonstration videos and apply them to a target image like a user's AR glass view. This video-to-image affordance grounding task is challenging due to (1) the need to predict fine-grained affordances, and (2) the limited training data, which inadequately covers video-image discrepancies and negatively impacts grounding. To tackle them, we propose Affordance Transformer (Afformer), which has a fine-grained transformer-based decoder that gradually refines affordance grounding. Moreover, we introduce Mask Affordance Hand (MaskAHand), a self-supervised pre-training technique for synthesizing video-image data and simulating context changes, enhancing affordance grounding across video-image discrepancies. Afformer with MaskAHand pre-training achieves state-of-the-art performance on multiple benchmarks, including a substantial 37% improvement on the OPRA dataset. Code is made available at https://github.com/showlab/afformer.*

1. Introduction

Humans frequently learn from observing others interact with objects to enhance their own experiences, such as following an online tutorial to operate a novel appliance. To equip AI systems with this ability, a key challenge lies in comprehending human interaction across videos and images. Specifically, a robot must ascertain the points of interaction (*i.e.*, affordances [19]) in a demonstration video and apply them to a new target image, such as the user's view through AR glasses.

This process is formulated as video-to-image affordance grounding, recently proposed by [13], which presents a more challenging setting than previous affordance-related tasks, including affordance detection [11, 59], action-to-image grounding [18,40,41,44], and forecasting [16,35,36].

†Corresponding Author.

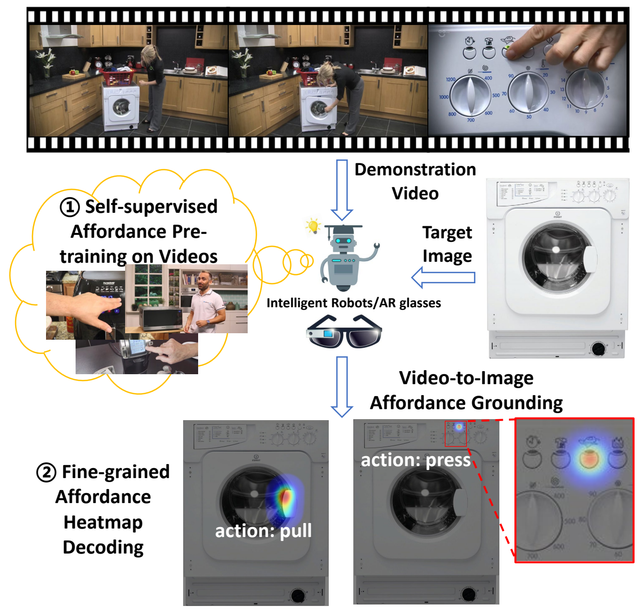

Figure 1. This figure demonstrates the video-to-image affordance grounding task, which aims to identify the area of human hand interaction (*i.e.*, affordance) in a demonstration video and map it to a target image (*e.g.*, AR glass view). Our contributions include (1) proposing a self-supervised pre-training approach for affordance grounding, and (2) establishing a new model that excels remarkably in fine-grained heatmap decoding.

The complexity of this setting stems from two factors: (1) Fine-grained grounding: Unlike conventional grounding tasks that usually localize coarse affordance positions (*e.g.*, identifying all buttons related to "press"), video-to-image affordance grounding predicts fine-grained positions specific to the query video (*e.g.*, only buttons pressed in the video). (2) Grounding accross various video-image discrepancies: Demonstration videos and images are often captured in distinct environments, such as a store camera's perspective versus a user's view in a kitchen, which complicates the grounding of affordances from videos to images. Moreover, annotating for this task is labor-intensive, as it necessitates thoroughly reviewing the entire video, correlating it with the image, and pinpointing affordances. As a result, affordance grounding performance may be limited by insufficient data on diverse video-image discrepancies.

To enable fine-grained affordance grounding, we propose a simple yet effective Affordance transformer (Afformer), which progressively refines coarse-grained predictions into fine-grained affordance grounding outcomes. Previous methods [13,40,44] either simply employ large stride upsampling or deconvolution for coarse-grained affordance heatmap prediction (*e.g.*, $8 \times 8 \rightarrow 256 \times 256$ [13]), or just evaluate at low resolution (*e.g.*, 28×28 [40, 44]). As a result, these methods struggle with fine-grained affordance grounding, particularly when potential affordance regions are closely situated (*e.g.*, densely packed buttons on a microwave). Our approach employs cross-attention [4, 5] between multi-scale feature pyramids, to facilitate gradual decoding of fine-grained affordance heatmaps.

To address the limited data issue that inadequately covers video-image differences and hampers affordance grounding performance, we present a self-supervised pre-training method, Masked Affordance Hand (MaskAHand), which can leverage vast online videos to improve video-to-image affordance grounding. MaskAHand can automatically generate target images from demonstration videos by masking hand interactions and simulating contextual differences between videos and images. In the generated target image, the task involves estimating the interacting hand regions by watching the original video, thereby enhancing the similarity capabilities crucial for video-to-image affordance grounding. Our approach uniquely simulates context changes, an aspect overlooked by previous affordance pretraining techniques [18, 36]. Furthermore, we also rely less on external off-the-shelf tools [1,39,48,50] used in [18,36].

We conducted comprehensive experiments to evaluate our Afformer and MaskAHand methods on three video-to-image affordance benchmarks, namely OPRA [13], EPIC-Hotspot [44], and AssistQ [57]. Compared to prior architectures, our most lightweight Afformer variant achieves a relative 33% improvement in fine-grained affordance grounding (256×256) on OPRA and relative gains of 10% to 20% for coarse-grained affordance prediction (28×28) on OPRA and EPIC-Hotspot. Utilizing MaskAHand pre-training for Afformer results in zero-shot prediction performance that is comparable to previously reported fully-supervised methods on OPRA [13], with fine-tuning further enhancing the improvement to 37% relative gains. Moreover, we demonstrate the advantage of MaskAHand when the data scale for downstream tasks is limited: on AssistQ, with only approximately 600 data samples, MaskAHand boosts Afformer's performance by a relative 28% points.

2. Related Work

Visual Affordance. The term "affordance" was coined in 1979 [17] to refer to potential action possibilities offered by the environment. With the progress of computer vision, learning visual affordances has gained significant attention in several domains, such as 2D/3D human-object interaction [9, 11, 46, 50], robotic manipulation/grasping [22, 42, 43], and instructional video understanding [13, 44]. Affordance detection [11, 59], forecasting [16, 35, 36], and grounding [13, 18, 40, 41, 44] are some of the well-known visual affordance tasks.

Affordance Grounding. Affordance grounding tasks are primarily divided into two categories: (1) Action-to-Image Grounding [18,40,41,44], which aims to identify image regions corresponding to a specific action query (*e.g.*, "press" $\rightarrow$ all microwave buttons). (2) Video-to-Image Affordance Grounding [13, 40, 44], which involves predicting the interaction region and associated action label in a target image, based on a video (*e.g.*, pressing the power button on a microwave in a video $\rightarrow$ the same region in a target image). Our paper concentrates on the latter problem, which presents greater challenges than action-to-image grounding.

Predicting Affordance Heatmap. Affordance regions have typically been represented by pixel-wise heatmaps or masks in prior research [11,13,18,40,41,44,59]. To predict these pixel-wise affordance maps, upsampling and deconvolution with a large stride (*e.g.*, $32\times$) are frequently employed [13, 18, 59]. Additionally, some weakly-supervised approaches [40, 41, 44] utilize Grad-CAM [49] to estimate the affordance heatmap through gradient activations. However, these techniques struggle to generate fine-grained affordance heatmaps, as they only learn affordance in low-resolution feature maps (*e.g.*, 7×7).

We get inspiration from multi-scale segmentation transformers [5, 6, 55, 58] to design a pyramid transformer decoder, which performs cross attention across pyramid feature levels to gradually refine the affordance heatmap estimations. This architecture can better support fine-grained affordance heatmap prediction.

Self-supervised Affordance Pre-training. Affordance regions within or across objects can be highly diverse and irregular in shape, complicating the prediction process. Several affordance pre-training methods [18, 36] have been introduced to improve affordance heatmap prediction. HoI-forecast [36] generates self-supervised annotations for the current frame using future frames by: (1) employing hand-object interaction detection [50], skin segmentation [48], and hand landmark estimation [1] to produce affordance points, and (2) utilizing SIFT [39] matching to map affordance points back to the current frame. HandProbes [18] also leverages [50] to select frames for "masked hand-object intersection prediction".

In contrast to HoI-forecast [36] and HandProbes [18], which primarily focus on minimally altered egocentric images/videos [31] and rely on detecting common objects, our proposed MaskAHand enhances affordance grounding from video to image across substantial contextual variations, eliminating the need for common object detection.

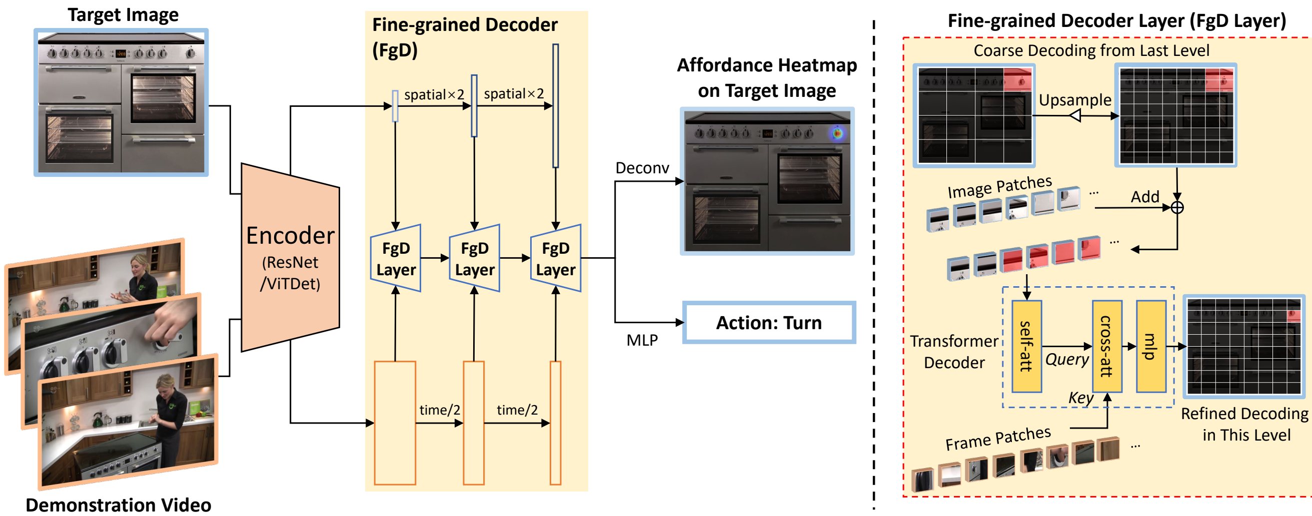

Figure 2. Our proposed Afformer is a simple yet effective model for video-to-image affordance grounding. The Afformer takes a demonstration video and a target image as inputs and produces an affordance heatmap on the target image. It employs an encoder to extract video and image features, followed by a multi-scale, transformer-based decoder to progressively refine fine-grained affordance heatmaps.

3. Affordance Transformer

We introduce Affordance Transformer (Afformer), a novel model for video-to-image affordance grounding. See Figure 2, our model employs a multi-scale, transformer-based decoder to handle fine-grained affordance grounding. We now proceed to discuss the Afformer in detail.

3.1. Formulation

We present the problem formulation for video-to-image affordance grounding [13]. Given a demonstration video $V \in \mathbb{R}^{t\times 3\times h^V\times w^V}$ and a target image $I \in \mathbb{R}^{3\times h^I\times w^I}$ as inputs, the goal of video-to-image affordance grounding is to predict an affordance heatmap $H \in \mathbb{R}^{h^I\times w^I}$ over the target image, accompanied by an action label $A \in \{1, \dots, c\}$, where c denotes the number of action classes. Here, t represents the number of sampled video frames, h and w indicate the height and width of the image or video, respectively. The Afformer $\boldsymbol{F}$ can be described as a function that maps the demonstration video and target image to the affordance heatmap and action label:

$$\boldsymbol{F}(V, I) \rightarrow (H, A). \tag{1}$$

3.2. Afformer Architecture

As shown in Figure 2, the Afformer $\boldsymbol{F}$ consists of a video/image shared encoder $\boldsymbol{F_e}$, a fine-grained decoder $\boldsymbol{F_d}$, and a prediction head $\boldsymbol{F_h}$, *i.e.* $\boldsymbol{F} = \boldsymbol{F_h} \circ (\boldsymbol{F_d} \circ \boldsymbol{F_e})$.

Shared Encoder. Our Afformer employs a shared encoder for both video and image feature representation, mitigating overfitting issues in small video-to-image affordance grounding datasets [13, 44, 57]. During the encoding phase, we only utilize a spatial network [12, 21] to process images and sampled video frames, reducing computational costs caused by temporal modeling. To enable multi-scale decoding, we preserve multi-scale features during the encoding process. This can be expressed as:

$$\boldsymbol{F_e}(V) \rightarrow \{E_l^V\}, \boldsymbol{F_e}(I) \rightarrow \{E_l^I\}, \tag{2}$$

where $\{E_l^V\}$ ($\{E_l^I\}$) represents the set of features extracted from the video (image) input at different scales l. Their shapes are $E_V^l \in \mathbb{R}^{t\times C\times h_l^V\times w_l^V}$ and $E_I^l \in \mathbb{R}^{C\times h_l^I\times w_l^I}$. C is the feature dimension. Following [32], l is the stride with respect to the input scale, *e.g.* $l \in \{2, 3, 4, 5\}$. We adopt multi-scale feature pyramid networks such as ResNet with FPN [21, 32] and ViTDet [12, 29] as the encoder.

Fine-grained Affordance Decoder. With multi-scale encodings $\{E_l^V\}$ and $\{E_l^I\}$, our fine-grained decoder produces heatmap decoding $D_{l_{min}}$ by

$$\boldsymbol{F_d}(\{E_l^V\}, \{E_l^I\}) \rightarrow D_{l_{min}}. \tag{3}$$

where l_{min} represents the minimal stride, which corresponds to the pyramid level of the highest feature resolution. For example, if the encoder uses $l_{min} = 2$, then $D_{l_{min}}$ will be in $\mathbb{R}^{C\times(h^I/2^2)\times(w^I/2^2)}$. Previous methods [13, 44] typically use single-step, small-scale decoding (*e.g.*, $l_{min} = 5$) and a large deconvolution to predict large-scale heatmaps. However, this approach loses spatial information and produces coarse predictions. In comparison, our decoder utilizes a multi-scale strategy to gradually decode the heatmap, which leads to more detailed affordance heatmaps.

We first outline single-scale decoding before extending it to multi-scale. Video-to-image affordance grounding can be considered a process where, at each spatial location in the target image, we search the video to determine if the location corresponds to an affordance region. Consequently, this process can be intuitively modeled as a cross-attention operation, with image encodings serving as the query and video encodings as the key and value. Unlike previous studies [13,44], our method explicitly incorporates spatial modeling, making it more suitable for heatmap grounding. The core operation of multi-head cross-attention (MCA) is

$$\boldsymbol{MCA}(\hat{Q}_l, \hat{K}_l, \hat{V}_l) = \sigma(\frac{\hat{Q}_l(\hat{K}_l)^{\mathrm{T}}}{\sqrt{C}} + R_l)\hat{V}_l W_l^{MCA}, \quad (4)$$

where $\hat{Q}_l, \hat{K}_l, \hat{V}_l$ denote flattened, layer-normalized, and linearly mapped query, key, and value features from E_l^I, E_l^V, E_l^V, respectively. R_l represents the decomposed relative positional encoding [29,30]. W^{MCA} is a learnable linear mapping, and σ is the softmax operation along the sequence of the $\hat{K}_l$ axis. For simplicity, we only display a single attention head here, while the actual attention head used is a common $C/64$. According to [53], the complete transformer decoder layer also consists of a multi-head self-attention layer (MSA) and an MLP layer (MLP). We perform MSA on image features, rather than video features, to reduce memory usage, as video feature sequences are longer due to the temporal dimension. The single-scale decoding D_l is obtained by

$$D_l = flatten(E_l^I) + \boldsymbol{MSA}(\hat{Q}_l), \quad (5)$$

$$D_l = D_l + \boldsymbol{MCA}(D_l, \hat{K}_l, \hat{V}_l), \quad (6)$$

$$D_l = D_l + \boldsymbol{MLP}(D_l). \quad (7)$$

We proceed with multi-scale decoding. The query in Equation 5 should incorporate heatmap decoding at the pyramid level $(l+1)$ for coarse-to-fine refinement. Furthermore, we find that maintaining a fixed resolution video pyramids stabilize training. We select $l = 3$ to balance resolution and semantics for video pyramids. Meanwhile, as the encoder only processes videos spatially, we apply temporal sampling to video features to aggregate temporal information before decoding, as depicted in Figure 2. Consequently, the multi-scale decoding can be expressed as

$$D_l = flatten(E_l^I) + \boldsymbol{MSA}(\hat{Q}_l + UP(D_{l+1})), \quad (8)$$

$$D_l = D_l + \boldsymbol{MCA}(D_l, \hat{K}_l^{C3D}, \hat{V}_l^{C3D}), \quad (9)$$

$$D_l = D_l + \boldsymbol{MLP}(D_l). \quad (10)$$

where UP denotes to the nearest spatial upsampling used in FPN [32]. $C3D$ denotes spatial-temporal convolution [52] with a kernel size of 3 and a stride of 2 applied only in the temporal dimension. By multi-scale decoding we can get the final fine-grained heatmap decoding $D_{l_{min}}$.

Heatmap and Action Prediction. The predictor computes $\boldsymbol{F_h}(D_{l_{\min}}) \rightarrow H, A$. We utilize appropriate deconvolutional layers to reconstruct the heatmap H by decoding $D_{l_{\min}}$ (*e.g.*, $C\times64\times64 \mapsto 1\times256\times256$). A MLP network and adaptive 2D pooling are employed for action classification A. Thus, we have concisely presented the Afformer as $\boldsymbol{F}(V, I) \rightarrow H, A$.

Training Loss Function. Our Afformer predicts action logits for c classes and the corresponding $h^I \times w^I$ heatmap. In a supervised setting, Afformer is trained with action classification loss $\mathcal{L}_a$ and heatmap regression loss $\mathcal{L}_h$. The former uses multi-class cross-entropy loss, while the latter employs KL divergence as the heatmap loss [2,13]:

$$\mathcal{L}_h = \sum_i^{h^I}\sum_j^{w^I} 1_{g_{i,j}>0}\sigma_g(g)_{i,j}\log\frac{\sigma_g(g)_{i,j}}{\sigma_h(h)_{i,j}}, \quad (11)$$

where g represents the ground-truth heatmap, $\sigma_g(g)_{i,j} = g_{i,j}/\sum g$ normalizes by sum, and $\sigma_h(h)_{i,j} = \exp(h_{i,j})/\sum\exp(h)$ normalizes by softmax. The total loss for optimization is $\mathcal{L} = \mathcal{L}_a + \mathcal{L}_h$.

Our proposed Afformer attains fine-grained heatmap decoding via multi-scale cross-attention, making it an effective model for video-to-image affordance grounding. However, the limited available training data [13, 44, 57] ($< 20k$) restricts context variation and may impair the Afformer's ability to ground affordances across video-image discrepancies. While some modified cross-attention techniques [23,34] aim to address the low-data issue [23], they focus on feature aspects and are limited to few-shot settings. We propose to tackle the problem from data, yielding a more versatile solution applicable to various challenges. In the following section, we introduce a self-supervised pre-training method for video-to-image affordance grounding.

4. Masked Affordance Hand Grounding

We introduce Masked Affordance Hand (MaskAHand), a self-supervised pre-training approach for video-to-image affordance grounding. As depicted in Figure 3, our method leverages the consistency between hand interaction and affordance regions, approximating affordance grounding to hand interaction grounding. MaskAHand can be viewed as an extension of the "masked image modeling" paradigm [20,54,56] to "masked hand grounding". We now elaborate on the MaskAHand method in detail.

4.1. Formulation

As described in Section 3.1, the video-to-image affordance grounding can be expressed as Equation 1: $\mathbf{F}(V, I) \rightarrow H, A$. Since action prediction is a well-established problem in action recognition [14,15,25,52,60], we focus on the challenging task of affordance heatmap

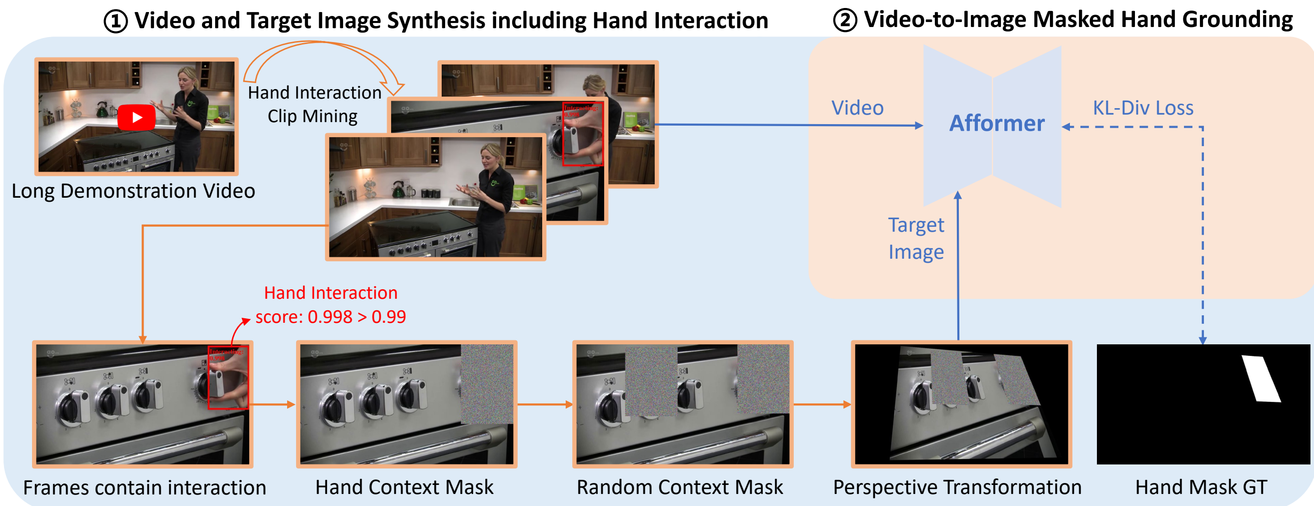

Figure 3. MaskAHand pre-training focuses on "video-to-image masked interaction hand grounding", acting as a proxy task for video-to-image affordance grounding. There are two steps: (1) video and target image synthesis including hand interaction, and (2) training Afformer with the generated data to learn "video-to-image masked interaction hand grounding". GT stands for "ground-truth".

grounding, denoted by $\boldsymbol{F}(V, I) \rightarrow H$. However, annotating for this heatmap is labor-intensive, as it necessitates thoroughly reviewing the entire video, correlating it with the image, and pinpointing affordances. Consequently, current video-to-image affordance grounding datasets contain limited training samples, with fewer than $20k$ in total.

We consider solving the limited data problem in the task $\boldsymbol{F}(V, I) \rightarrow H$. We observe that all affordance regions are interacted with by hands. The human hand exhibits a distinct visual pattern, making it easier to detect than irregular affordance regions. The interaction state can also be readily distinguished using the hand interaction detector [50]. Consequently, we focus on a related task, $\boldsymbol{F'}(V, I) \rightarrow H'$, which is simpler to gather data for, where H' represents the "imagined" hand in the target image that interacts with the affordance region shown in the video. As demonstrated, the capabilities of $\boldsymbol{F'}$ and $\boldsymbol{F}$ are closely related, allowing us to obtain $\boldsymbol{F}$ by fine-tuning the $\boldsymbol{F'}$ network or even considering $\boldsymbol{F'}$ as $\boldsymbol{F}$ for zero-shot grounding.

However, training $\boldsymbol{F'}(V, I) \rightarrow H'$ still needs the demonstration video V, the target image I, and interaction hand annotation heatmap H' (*e.g.*, hand box mask). But thanks to our approximation for the original task, the data preparation becomes much simpler and allows us to do self-supervised pre-training. We refer to our approach as Masked Affordance Hand (MaskAHand), illustrated in Figure 3 and described in the following section.

4.2. Affordance-related Data Synthesis

Hand Interaction Detection. Our MaskAHand relies solely on a hand interaction detector, eliminating the need for an object detector as required by [18, 36]. We employ a Faster R-CNN [32, 47] trained on the 100DOH dataset [8,28,50,51] to detect hand bounding boxes and output binary hand states (*i.e.*, interacting or not). Our trained detector achieves an 84.9 AP in hand interaction detection on the 100DOH test set, demonstrating its accuracy and reliability for synthesizing hand interaction data.

Hand Interaction Clip Mining. We extract multiple hand interaction clips from a long demonstration video, each containing 32 consecutive frames and regarding as V, guaranteeing the presence of an interacting hand in at least one frame. To avoid redundancy, we apply a stride of 16 frames between successive clips. We only set a high interaction score threshold 0.99 to reduce false positives.

Target Image Synthesis and Transformation. Inspired by SuperPoint [10], we synthesize the target image I from V by simulating video-image context changes, as illustrated in Figure 3. The process consists of four steps: (1) Select the corresponding frame from V involving the interaction hand; (2) Apply a mask to conceal the hand, as the target image should not include it, making the hand context mask M_h larger than the detected hand box by a factor of > 1 (*e.g.*, 1.5 times); (3) Introduce a random context mask M_r with the same scale as M_h to enhance MaskAHand pre-training difficulty, preventing the model from simply predicting the masked region; (4) Apply a random perspective transformation to simulate perspective change between the demonstration video and target image (*e.g.*, egocentric vs. exocentric).

4.3. Video-to-Image Masked Hand Grounding

Given a masked and perspective-transformed interaction frame as the target image I and a mined interaction clip as the video V, the Afformer without action predictor $\boldsymbol{F}'(V, I)$

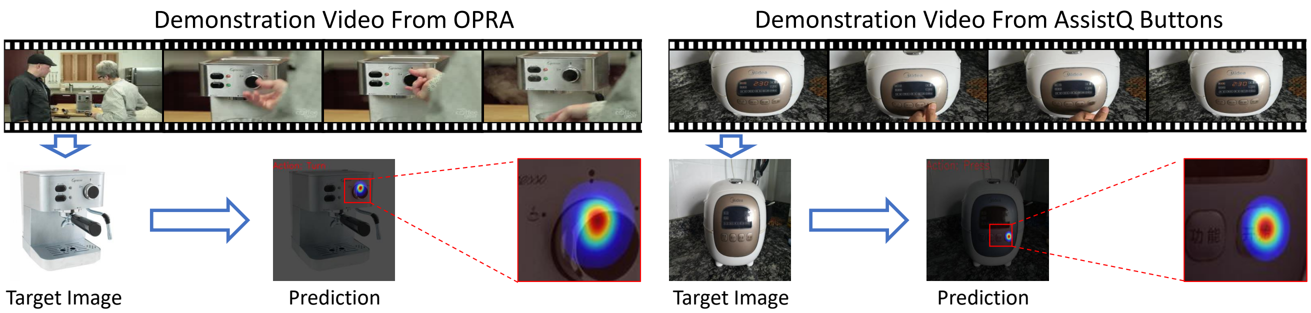

Figure 4. Afformer's video-to-image affordance grounding visualization. MaskAHand's visualization is in the supplementary material.

takes them as input and produces a heatmap H'. Unlike the supervised setting, the ground truth for H' is derived from the detected hand box mask, subjected to the same perspective transformation as I (see Figure 3). During MaskAHand pre-training, the network is tasked with video-to-image masked hand grounding, which requires observing the video to "match" the unmasked context in the target image and predict the precise hand box position. Thus, this pre-training shares abilities essential for video-to-image affordance grounding, including context matching between the video and image to ground affordances in the image.

For MaskAHand pre-training, we still utilize the KLD loss, as shown in Equation 11. The ground-truth hand box masks are gaussian blurred, following video-to-image affordance grounding [13]. After training Afformer with MaskAHand, we can directly perform zero-shot evaluation on video-to-image affordance grounding, or fine-tune using supervised data and subsequently conduct evaluation.

5. Experiments

5.1. Experimental settings

Datasets. We conduct experiments on three datasets:

• **OPRA** [13] consists of YouTube product review videos for appliances (*e.g.*, washing machine, stove). Each data sample consists of a video clip V (*e.g.*, holding a frying pan), paired with a target image I (*e.g.*, ads picture of the frying pan), an action label A (*e.g.*, holding) belonging to a total of 7 actions, and ten annotated points on the image representing the affordance region (*e.g.*, ten points around the frying pan handle). The ten points are always produced as ground-truth heatmap H by applying a gaussian blur with a kernel size of 3 [13, 40, 44]. The dataset comprises roughly $16k$ training samples and $4k$ testing samples, each represented in the form of (V, I, A, H).

• **EPIC-Hotspot** [44] is made up of EPIC-Kitchens [8], which contains egocentric videos of kitchen activities. EPIC-Kitchens provides an action label A and annotations for interacted objects in video V, but no target image. EPIC-Hotspot chooses one frame in V to be the target image I that corresponds to the object class and appearance. Follow [44] for more information. They crowdsource annotations for ground-truth heatmaps H after I is chosen, yielding $1.8k$ annotated instances across 20 actions and 31 objects. The data sample format in EPIC-Hotspot is (V, I, A, H), which is the same as OPRA.

• **AssistQ [57]** is a benchmark to solve user egocentric query [27] according to instructional videos. It includes fine-grained button regions for a wide range of everyday devices (*e.g.*, microwave), which require precise affordance prediction to distinguish very close buttons. In AssistQ [57], we consider the instructional video to be the active video and the user view to be the inactive image. Using transcript timestamps, we divide the instructional video into multiple clips V and manually annotate the interacted button on the inactive image I. We finally get 650 training samples from 80 videos and 91 testing samples from 20 videos, and each of sample contains active video clip, inactive image, and active-to-inactive button bounding-boxes. The ground-truth heatmap H is generated using gaussian blur for the button center point map. Because the action class in AssistQ is mostly limited to "press" or "turn", the data sample format is (V, I, H) without action.

We report results on the test sets of these datasets and perform ablation studies on the largest OPRA dataset.

Implementation Details. We train the Afformer model with a ResNet encoder using a batch size of 16 and $5k$ iterations, employing the AdamW optimizer [38] and a cosine annealing learning rate scheduler [37] with an initial learning rate of 3×10^{-4}. As per [13], we set the spatial size of both images and videos to 256. The ground-truth heatmap is generated using annotation points (box center for AssistQ) mapped to a Gaussian blur kernel size of $\sqrt{256 \times 256}/3$, following [13,44]. For the Afformer with a ViTDet encoder, we adjust the learning rate to 2×10^{-4} and the spatial size to 1024 to accommodate the pre-trained positional encodings from [29]. All encoders are initialized with COCO [33] detection weights [29, 47], following [13]. These hyperparameters remain consistent across all datasets, including MaskAHand pre-training.

Evaluation. We report saliency metrics [3] as KLD, SIM, and AUC-J. Please refer to [13, 44] for more details.

Method	Variants		OPRA (256 × 256)	
			Heatmap KLD ↓	Action Top-1 Acc ↑
Demo2Vec [13]	LSTM		3.45	20.41
	ConvLSTM		3.31	30.20
	TSA + ConvLSTM		3.34	38.47
	Motion + TSA + ConvLSTM		2.34	40.79
Naive Baseline (Ours)	ResNet-50-Deconv		2.20	45.66
Afformer (Ours)	ViTDet-B-Afformer		**1.51**	**52.27**
	ResNet-50-Afformer		**1.55**	**52.14**
MaskAHand (Ours)	Zero-shot	ResNet-50-Deconv	2.89	n/a
		ResNet-50-Afformer	**2.36**	n/a
	Fine-tune	ResNet-50-Deconv	1.74	48.93
		ResNet-50-Afformer	**1.48**	**52.50**

Table 1. Video-to-image affordance grounding performance of our Afformer and MaskAHand models on the OPRA dataset (fine-grained, 256×256): Afformer reduces heatmap KLD errors by over 30%; MaskAHand's zero-shot pre-training results are comparable to [13] (2.36 *vs.* 2.34); further fine-tuning yields the best performance on OPRA.

Method	Method	OPRA (28 × 28)			EPIC (28 × 28)		
		KLD ↓	SIM ↑	AUC-J ↑	KLD ↓	SIM ↑	AUC-J ↑
Weakly Supervised	EGOGAZE [24, 44]	2.43	0.25	0.65	2.24	0.27	0.61
	MLNET [7, 44]	4.02	0.28	0.76	6.12	0.32	0.75
	DEEPGAZEII [26, 44]	1.90	0.30	0.72	1.35	0.39	0.75
	SALGAN [44, 45]	2.12	0.31	0.77	1.51	0.40	0.77
	Hotspot [44]	1.42	0.36	0.81	1.26	0.40	0.79
	HAG-Net (+Hand Box) [40]	1.41	0.37	0.81	1.21	0.41	0.80
Self-supervised Zero-shot	Center Bias (action agnostic)	11.13	0.21	0.63	10.66	0.22	0.63
	MaskAHand (action agnostic)	1.86	0.28	0.76	1.32	0.37	0.76
Supervised	Img2heatmap [44]	1.47	0.36	0.82	1.40	0.36	0.80
	Demo2Vec [13]	1.20	0.48	0.85	n/a	n/a	n/a
	Afformer (Ours)	**1.05**	**0.53**	**0.89**	**0.97**	**0.56**	**0.88**

Table 2. Performance of Afformer and MaskAHand models on OPRA and EPIC-Hotspot datasets (coarse-grained, 28×28). MaskAHand can surpass many weakly-supervised methods in KLD. Afformer achieve the best performance among supervised methods.

Method	KLD ↓	SIM ↑	Top-1 Acc ↑
Afformer	1.13	0.44	0.41
+MaskAHand	**1.01(-12%)**	**0.54(+23%)**	**0.57 (+28%)**

Table 3. Results on AssistQ Buttons [57]. "Acc" refers to the accuracy of button classification.

5.2. Main Results

Fine-grained Video-to-image Affordance Grounding. In Table 1, the prior method Demo2Vec [13] achieves a KLD error of 2.34 on OPRA, which is comparable to our naive ResNet-50 [21] + deconvolution baseline with 2.20 KLD. However, our proposed Afformer significantly reduces KLD error. Utilizing ResNet-50 and ViTDet-B [29] backbones, Afformer attains 1.55 and 1.51 KLD, respectively, surpassing previous results by over 30%. We attribute Afformer's success to its better design for fine-grained affordance heatmaps, which also boosts action classification accuracy.

We also assess MaskAHand pre-training as Table 1 reveals, surprisingly, its zero-shot results are already comparable to Demo2Vec (2.36 *vs.* 2.34), demonstrating its effectiveness as a proxy task for supervised video-to-image affordance tasks. Furthermore, fine-tuning the MaskAHand pre-trained Afformer on OPRA leads to the lowest KLD error of 1.48, a 37% improvement over Demo2Vec.

Coarse-grained Affordance Grounding on OPRA, EPIC-Hotspots. We adopt the evaluation protocol from [44] to assess our method's performance on coarse-grained affordance grounding at low resolution (28×28). We downsample the standard resolution 256×256 prediction heatmap to 28×28 using bilinear interpolation during both training and inference phases. Table 2 demonstrates that Afformer, despite not being explicitly designed for lower resolution, outperforms other methods. Moreover, the self-supervised MaskAHand zero-shot results surpass some weakly-supervised approaches.

Video-to-image Grounding on small-scale Data. We train our Afformer and fine-tune MaskAHand models on AssistQ Buttons, which contains only 600+ training samples, posing challenges for data-hungry deep neural networks. As demonstrated in Table 3, MaskAHand pre-training substantially reduces the heatmap KLD error and increases SIM metric, increasing the relative button classification accuracy by 28%. Thus, MaskAHand self-supervised pre-training is a viable option when dealing with limited video-to-image affordance grounding data.

Spatial I	Spatial V	KLD ↓
8^2	8^2	1.88
16^2		1.73
32^2		**1.65**
64^2		**1.65**
32^2	16^2	1.63
	32^2	**1.62**
	64^2	1.63

(a) Single pyramid. I: image, V: video.

Spatial I	Spatial V	KLD ↓
32^2	32^2	**1.62**
	$16^2 \rightarrow 32^2$	1.63
32^2	32^2	1.62
$16^2 \rightarrow 32^2$		**1.60**
$8^2 \rightarrow 16^2 \rightarrow 32^2$	32^2	1.59
$16^2 \rightarrow 32^2 \rightarrow 64^2$		**1.57**
$8^2 \rightarrow ... \rightarrow 64^2$		**1.57**

(b) Multiple pyramids. I: image, V: video.

Spatial	Temporal	KLD ↓	Mem ↓
16^2 ↓ 32^2 ↓ 64^2	T	1.57	-0.0 G
	$T \rightarrow \frac{T}{2} \rightarrow \frac{T}{2}$	1.56	-2.7 G
	$T \rightarrow \frac{T}{2} \rightarrow \frac{T}{4}$	**1.55**	-3.8 G

(c) Temporal pyramids and reduced memory.

Table 4. Ablation study results for Afformer on the OPRA dataset at a 256×256 scale, comparing default settings (image spatial pyramids: $16^2 \rightarrow 32^2 \rightarrow 64^2$, video spatial pyramid: single 32^2, no temporal downsampling, and one cross/self-attention module in decoding). Enhanced by multi-scale, high-resolution decoding, performance significantly improves (1.57 vs. 1.88), while temporal pyramids further reduce KLD error and decrease GPU memory consumption.

# Hand Mask ($\#M_h <= 1$)	# Random Mask ($\#M_r$)	Mask Scale	Zero-shot KLD
0	0	n/a	4.35
1	0	1.0×	4.29
1	0	1.5×	2.68
1	0	2.0×	2.54
1	0	3.0×	3.42
1	1	1.5×	**2.48**
1	1	2.0×	2.75
1	2	2.0×	2.98

(a) Ablations on masking ratio and number of masks.

Masking	Distortion	Zero-shot KLD	Fine-tune KLD
$\#M_h = 1$, $\#M_r = 1$, 1.5×	0	2.48	1.53
	0.25	2.40	1.50
	0.5	**2.36**	**1.48**
	1.0	n/a	n/a

(b) Ablations on perspective transformation. "n/a": network divergence.

Table 5. Ablation studies of MaskAHand pre-training on OPRA.

5.3. Ablation Studies

Afformer Fine-grained Decoder. We evaluate our fine-grained decoder on the OPRA dataset (256×256 resolution). Table 4(a) reveals that the highest image pyramid resolution (64^2) reduces the KLD error; however, for videos, a 32^2 resolution is more effective, potentially due to weaker semantics in high-resolution pyramids. Table 4(b) indicates that preserving a fixed video pyramids when building low-to-high resolution image pyramids also decreases KLD error. As per Table 4(c), constructing video temporal pyramids results in considerable memory savings and slight performance enhancement. These findings suggest the significance of our fine-grained decoder in heatmap decoding.

Context Masking in MaskAHand. We examine the context masking effects in MaskAHand (Table 5(a)). Without masking ($\#M_h = 0$ and $\#M_r = 0$), pre-training degenerates into hand saliency detection (Figure 3), providing no benefit to video-to-image affordance grounding and resulting in a large KLD error 4.35. A similar outcome occurs with hand masking only: when $\#M_h = 1$ and $\#M_r = 0$, the network can directly predict the masked region for a low training loss, yielding a meaningless result (KLD 4.29).

However, increasing the hand mask region to $1.5\times$ significantly reduces the zero-shot KLD error to 2.68, indicating that the network learns a useful representation for video-to-image affordance grounding. Extending the mask scale to $3.0\times$ causes performance degradation, likely due to the overly large masked region creating challenging context matching between video and image. An additional random mask offers a similar effect as enlarging the hand mask but with more diversity, preventing the network from merely predicting a region within the hand mask. Therefore, we use $\#M_h = 1$, $\#M_r = 1$, and $1.5\times$ hand box masking as the default MaskAHand pre-training setting.

Perspective Transformation in MaskAHand. Our context masking strategy is to enhance the grounding ability across the video-image context differences. However, it is also crucial to consider perspective transformation, such as enabling simulation of egocentric and exocentric views. As shown in Table 5(b), perspective transformation can lead to performance improvements when the distortion ratio is in a reasonable range.

6. Conclusion

In this paper, we introduce the Affordance Transformer (Afformer), a simple and effective model for video-to-image affordance grounding, utilizing multi-scale decoding to generate fine-grained affordance heatmaps. We also propose a pre-training technique, Masked Affordance Hand (MaskAHand), that employs a proxy task of masked hand interaction grounding, facilitating data collection while benefiting video-to-image affordance grounding. Our extensive experiments show Afformer significantly outperforms previous methods, and MaskAHand pre-training impressively improves performance on small datasets.

Acknowledgment This project is supported by the National Research Foundation, Singapore under its NRFF Award NRF-NRFF13-2021-0008, and Mike Zheng Shou's Start-Up Grant from NUS. The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore.

References

[1] Finger counter. https://www.computervision.zone/courses/finger-counter/, 2022.

[2] Alexandre Bruckert, Hamed R. Tavakoli, Zhi Liu, Marc Christie, and Olivier Le Meur. Deep saliency models : The quest for the loss function. *Neurocomputing*, 453:693–704, 2021.

[3] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 740–757, 2019.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.

[5] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022.

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, pages 17864–17875, 2021.

[7] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. In *ICPR*, pages 3488–3493, 2016.

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The dataset. In *ECCV*, pages 753–771, 2018.

[9] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *CVPR*, pages 1778–1787, 2021.

[10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Workshop*, pages 224–236, 2018.

[11] Thanh-Toan Do, Anh Nguyen, and Ian D. Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *ICRA*, pages 1–5, 2018.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[13] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2vec: Reasoning object affordances from online videos. In *CVPR*, pages 2139–2147, 2018.

[14] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *NeurIPS*, 2022.

[15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6201–6210, 2019.

[16] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *J. Vis. Commun. Image Represent.*, pages 401–411, 2017.

[17] J. J. Gibson. The theory of affordances. *The people, place, and space reader*, 1979.

[18] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *CVPR*, pages 3283–3293, 2022.

[19] Mohammed Hassanin, Salman H. Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *ACM Comput. Surv.*, pages 47:1–47:35, 2021.

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[22] Thomas E Horton, Arpan Chakraborty, and Robert St Amant. Affordances for robots: a brief survey. *AVANT*, pages 70–84, 2012.

[23] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *NeurIPS*, pages 4005–4016, 2019.

[24] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *ECCV*, pages 754–769, 2018.

[25] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv*, 1705.06950, 2017.

[26] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv:1610.01563*, 2016.

[27] Weixian Lei, Difei Gao, Yuxuan Wang, Dongxing Mao, Zihan Liang, Lingmin Ran, and Mike Zheng Shou. Assistsr: Task-oriented video segment retrieval for personal AI assistant. In *EMNLP*, pages 319–338, 2022.

[28] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *ECCV*, pages 639–655, 2018.

[29] Yanghao Li, Hanzi Mao, Ross B. Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, pages 280–296, 2022.

[30] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, pages 4804–4814, 2022.

[31] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *NeurIPS*, 2022.

[32] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944, 2017.

[33] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014.

[34] Weidong Lin, Yuyan Deng, Yang Gao, Ning Wang, Jinghao Zhou, Lingqiao Liu, Lei Zhang, and Peng Wang. CAT: cross-attention transformer for one-shot object detection. *arXiv:2104.14984*, 2021.

[35] Miao Liu, Siyu Tang, Yin Li, and James M. Rehg. Forecasting human-object interaction: Joint prediction of motor attention and actions in first person video. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, pages 704–721, 2020.

[36] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *CVPR*, pages 3272–3282, 2022.

[37] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017.

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[39] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, pages 91–110, 2004.

[40] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning visual affordance grounding from demonstration videos. *arXiv:2108.05675*, 2021.

[41] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *CVPR*, pages 2252–2261, 2022.

[42] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *CoRL*.

[43] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *ICRA*, pages 6169–6176, 2021.

[44] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *ICCV*, pages 8687–8696, 2019.

[45] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv:1701.01081*, 2017.

[46] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV)*, pages 401–417, 2018.

[47] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017.

[48] Frerk Saxen and Ayoub Al-Hamadi. Color-based skin segmentation: An evaluation of the state of the art. In *ICIP*, pages 4467–4471, 2014.

[49] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[50] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, pages 9866–9875, 2020.

[51] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*, pages 7396–7404, 2018.

[52] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 6000–6010, 2017.

[54] Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval. In *CVPR*, pages 3313–3322, 2022.

[55] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, pages 568–578, 2021.

[56] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan L. Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, pages 14648–14658, 2022.

[57] Benita Wong, Joya Chen, You Wu, Stan Weixian Lei, Dongxing Mao, Difei Gao, and Mike Zheng Shou. Assistq: Affordance-centric question-driven task completion for egocentric assistant. In *ECCV*, pages 485–501, 2022.

[58] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, pages 12077–12090, 2021.

[59] Wei Zhai, Hongchen Luo, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot object affordance detection in the wild. *Int. J. Comput. Vis.*, pages 2472–2500, 2022.

[60] David Junhao Zhang, Kunchang Li, Yali Wang, Yunpeng Chen, Shashwat Chandra, Yu Qiao, Luoqi Liu, and Mike Zheng Shou. Morphmlp: An efficient mlp-like backbone for spatial-temporal representation learning. In *ECCV*, pages 230–248, 2022.