Learning Visual Affordance Grounding From Demonstration Videos

Hongchen Luo, Graduate Student Member, IEEE, Wei Zhai, Jing Zhang, Senior Member, IEEE, Yang Cao, Member, IEEE, and Dacheng Tao, Fellow, IEEE

Abstract—Visual affordance grounding aims to segment all possible interaction regions between people and objects from an image/video, which benefits many applications, such as robot grasping and action recognition. Prevailing methods predominantly depend on the appearance feature of the objects to segment each region of the image, which encounters the following two problems: 1) there are multiple possible regions in an object that people interact with and 2) there are multiple possible human interactions in the same object region. To address these problems, we propose a hand-aided affordance grounding network (HAG-Net) that leverages the aided clues provided by the position and action of the hand in demonstration videos to eliminate the multiple possibilities and better locate the interaction regions in the object. Specifically, HAG-Net adopts a dual-branch structure to process the demonstration video and object image data. For the video branch, we introduce hand-aided attention to enhance the region around the hand in each video frame and then use the long short-term memory (LSTM) network to aggregate the action features. For the object branch, we introduce a semantic enhancement module (SEM) to make the network focus on different parts of the object according to the action classes and utilize a distillation loss to align the output features of the object branch with that of the video branch and transfer the knowledge in the video branch to the object branch. Quantitative and qualitative evaluations on two challenging datasets show that our method has achieved state-of-the-art results for affordance grounding. The source code is available at: https://github.com/lhc1224/HAG-Net.

Index Terms—Deep learning, learning from demonstrations, visual affordance grounding, weakly supervised learning.

I. INTRODUCTION

For an object in the scene, it is essential to identify what it is and the interactions it may have with people. The potential interactions between an object and an agent are known as affordances, which was introduced by the ecological psychologist Gibson [1] to describe the possible “action possibilities” provided by the environment. The

Fig. 1. Challenges of affordance grounding. We show some examples of affordance with multiple possibilities. (a) Different interactions occur at different locations of the same object. (b) Different interactions occur at a similar position of the same object.

“action possibilities” reflect all possible interactions between the object and the agent. In particular, perceiving object affordance is a valuable capability and has a wide range of applications in action recognition [2], [3], robot grasping [4], [5], human–robot interaction [6], autonomous driving [7], scene understanding [8], [9], and so on.

Since affordance is a highly relevant attribute for human–object interaction, mining human–object interaction information from videos is a promising approach to learn the object’s affordance. This article focuses on affordance grounding from demonstration videos, i.e., learning regions of human–object interaction from demonstration videos and transferring them to static objects for application.

However, since affordance is a complementary property of the animal and the environment [1], the multiple potential complementarities of the environment and the animal lead to multiple possibilities for affordance: 1) there are multiple possible regions in an object where people interact. As shown in Fig 1(a), the same object may imply two different interactions, “pull” and “rotate,” and the interaction occurs at different locations and 2) there are multiple possible human interactions in the same region. As shown in Fig. 1(b), two different interactions, “rotate” and “touch,” can happen at the same position on the same object. Existing methods [10], [11], [12], [13] typically establish the mapping relationship between apparent features and affordance labels, which struggle to capture the affordance-related context and perceive the affordance in multiple possibilities.

Manuscript received 8 June 2021; revised 11 December 2021, 4 February 2023, and 15 May 2023; accepted 17 July 2023. Date of publication 11 September 2023; date of current version 30 October 2024. This work was supported by the National Key Research and Development Program of China under Grant 2020AAA0105701. (Corresponding author: Yang Cao.)

Hongchen Luo, Wei Zhai, and Yang Cao are with the School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui 230026, China (e-mail: lhc12@mail.ustc.edu.cn; wzhai056@ustc.edu.cn; forest@ustc.edu.cn).

Jing Zhang and Dacheng Tao are with the Sydney AI Centre and the School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia (e-mail: jing.zhang1@sydney.edu.au; dacheng.tao@sydney.edu.au).

Digital Object Identifier 10.1109/TNNLS.2023.3298638
At the same time, the study [14] found that, in a particular scene, the object’s affordance is uniquely determined by the intention of the human action. For example, in a demonstration video, human hands and objects interact frequently, and human action intention can be inferred from hand position and action, providing an essential clue to solve the above two problems. For the first question, we can infer the region where the person interacts with the object by the hand’s position, such as “touch” and “pull” in Fig. 1(a). For the second question, we eliminate the ambiguity of affordance by inferring human intention through human hand actions, such as in 1(b); according to the hand action in the video, we can infer whether the object’s affordance is “touch” or “rotate.” To achieve this objective, we leverage both the position and action of the hand to extract affordance clues and solve multiple possibilities of affordance grounding by establishing a mapping relationship between hand action intentions and different object parts.

This article presents a hand-aided affordance grounding network (HAG-Net) that uses both hand position and action cues to learn the affordance of objects from demonstration videos (as shown in Fig. 2). To better capture the affordance cues from the demonstration videos, we propose a hand-related selection network to select the keyframes, which primarily reflect the interaction between the hand and the object. During learning the object’s affordance from the demonstration videos (i.e., training phase), the hand-related features in the video frames are enhanced to enable the network to focus more on interaction-related object regions and thus extract affordance-related key features from the videos. Meanwhile, in transferring the affordance knowledge from the video branch to the object branch, we consider both hand position and action information to mitigate the two challenges in affordance grounding and enhance the perception and localization of affordance regions in the object branch of the network. Since the region of human–object interaction is related to the category of action, we design a semantic enhancement module (SEM) that considers both the action category and appearance features to jointly adjust the feature of the object image so that the network can better focus on the feature regions related to the action category. In the affordance grounding phase (i.e., testing phase), only the static object image and affordance label are given, and the network infers the affordance region of the object. The contributions of this article are given as follows.

1) We propose an HAG-Net to use the clues provided by the hands in the demonstration videos. By leveraging both the hand’s position and action, HAG-Net can eliminate the multiple possibilities typically encountered in the affordance grounding task.

2) We propose the SEM, which incorporates both action category and appearance features to modulate the object image’s feature representation jointly. This enhancement enables more precise and effective localization of the relevant interaction regions.

3) Experimental results on the two most challenging affordance grounding datasets have demonstrated the superiority of the proposed model against the state-of-the-art.

II. RELATED WORK

A. Affordance Grounding

Affordance grounding is to detect the regions of the object where interactions may occur. Early works [10], [11], [12] are mainly on image-based supervised segmentation tasks, outputting the affordance label of each pixel for a given image. However, these tasks rely on pixel-level labels, and these methods cannot learn how humans interact with objects.

Another part of the early works represents affordance as a pose in which humans interact with objects. Yao et al. [15] used a clustering method to find all possible object functionalities and represented it in the form of a pose in which humans interact with musical instruments. Wang et al. [16] used variational autoencoders (VAEs) to construct their model to predict affordance poses. Li et al. [17] proposed a 3-D generative model to predict physically plausible and physically feasible human poses in a given 3-D scene. In contrast to the previous methods, our approach emphasizes the potential interactions among different object regions instead of solely focusing on the human pose.

Many works solve the affordance grounding by learning the interactions between humans and objects in a video. Koppula and Saxena [10] proposed a generative model to ground the affordance into the form of the object’s spatial position and time trajectory. Fang et al. [18] used the demonstration video’s feature embedding to predict the target object’s interaction region and proposed the Online Product Review Dataset for Affordance (OPRA) dataset. Nagarajan et al. [19] learned the interaction between humans and objects by observing videos. Unlike the above methods, our approach involves the explicit inference of human action intentions from the cues presented by the position and action of the human hand. This enables the specification of the unique affordance for a given scene and eliminates any ambiguity that may arise from multiple possible affordances.

B. Hand and Affordance

There are a large number of works in hand-based egocentric action recognition. Hand detection, segmentation, and tracking technologies continue to develop, and their research results are
applied to model actions and activities. Kjellström et al. [8] learned the affordance of objects from the human demonstration and extracted hand position, orientation, and articulated pose for embedding action features. Stark et al. [23] obtained the affordance clues from the interaction between the human hand and the objects in the trainset and then determined the functions of the objects according to the affordance cue features. Sun et al. [24] significantly improved the detection accuracy of the interactive objects by learning hand-movement trajectories and statistical knowledge in training data. Song et al. [25] predicted the human intention by observing the human/object interaction. Pieropan et al. [26] directly represented objects as their interactions with human hands for modeling human activity. In this article, we use the clues provided by the hand to learn the affordance of the objects. The above works utilize the preacquired hand-related information, e.g., hand position, motion direction, and movement speed as auxiliary clues for tasks such as affordance classification/detection/reasoning. However, it is difficult to collect the video data with hand information in practice, which obscures the large-scale training for the existing methods.

Different from these works, this article addresses the problem of weakly supervised affordance grounding from video demonstrations and only uses the action class labels as supervision. In particular, our proposed method only uses the bounding box corresponding to hand position as auxiliary information obtained from input images by a pretrained detection network but not preacquired. Therefore, our proposed method is more universal in practical application. Moreover, inferring the affordance region from images containing only static scenes is also a more challenging task under the weak supervision of action labels. To address this issue, this article proposes an HAG-Net, which includes a hand-aided attention module to perceive hand regions and aggregate hand-movement features from demonstration video and an SEM to combine the appearance features with action cues and ground the action-related parts of an object from the input images.

C. Learning From Demonstrations

By learning how humans interact with objects through demonstrations, the robot can perceive the affordance of the objects. Furthermore, when facing with objects in different environments, the robot can mimic human actions to interact with the objects through the knowledge learned from human demonstrations. Recently, there have been a great number of works on learning from demonstrations [18], [27], [28]. Schulman et al. [29] proposed a method based on nonrigid trajectory transfer for adapting the demonstrated trajectory from the training geometry to the test geometry, enabling the robot to tie different types of knots in the rope automatically. Chu et al. [27] learned different haptic affordances by demonstrating learning to provide a robot with an example of successful interaction with a given object. Aleotti and Caselli [30] performed the task of robot grasp planning well by combining the limited and automatic 3-D shape segmentation of human demonstrators for object recognition and semantic modeling. Fang et al. [18] proposed to learn from online product review videos and then transfer the knowledge from the videos to the target image to learn the relevant regions of human–object interaction. This article also considers learning object affordance from demonstration videos. However, we restrict using the action labels as supervision to learn the mapping relationships from static objects to the object features that interacted in the videos.

III. METHOD

A. Problem Description

In this article, we aim to learn objects’ affordance through demonstration videos. As shown in Fig. 2, the human–object interaction features are extracted from the demonstrative videos and transferred to the static object images during training. The process utilizes only affordance labels as supervision. During testing, only the static object images and affordance labels are given, and the network infers the affordance region of the object. Before learning affordance, we first detect the hand in each video frame (some detection results are shown in Fig. 3), and a specific selection network is used to select the keyframes where the hand interacts with the objects. After preprocessing the video datasets, we introduce an HAG-Net, as shown in Fig. 4. In the affordance learning phase (training phase), we leverage the affordance cues of the object from the demonstration videos and transfer them to a static object. Section III-C2 presents the details about hand-aided attention for enhancing hand position information and a special distillation loss for transferring hand position and action information to static object branches. With the combination of hand position and action cues, the network can better tackle the two challenges of affordance grounding and

---

Fig. 3. Hand-aided mask. (a) Some hand detection results, we use the trained YOLOv3 pruning model [21], [22] to detect the position of the hand on each frame of the two datasets of OPRA [18] and EPIC [20]. (b) Shadow regions are the surrounding regions of the hand enhanced by hand-aided attention.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
achieve superior localization results. Furthermore, we introduce an SEM to enhance affordance-related feature regions in object images. Finally, in the affordance grounding phase (test phase), we feed an object image and an action label into the HAG-Net and output a heatmap of the region on the object where the action may appear using visualization techniques.

B. Preprocessing

Since human–object interactions frequently occur in the demonstration videos, the hand position and actions can offer valuable clues for locating the affordance region of interest. Both OPRA [18] and EPIC [20] have no hand annotations. However, many existing works on object detection [21], [36] allow us to detect the human hand accurately. We choose the YOLOv3 pruning model [21], [22] trained on the Oxford hand dataset [32] to detect the hands in each frame of the two datasets. If the number of detected hands is greater than two, we select the two hands with the highest confidence. Some detection results are shown in Fig. 3(a).

Due to the potential presence of unhelpful video frames in conveying human–object interaction information, which could negatively affect affordance grounding, a hand-related selection network is employed to pinpoint keyframes that capture hand-object interactions while filtering out frames that fail to provide meaningful affordance cues. The framework is similar to the video branch in Fig. 4(a-1). Since hand position and action cues play an essential role in affordance grounding, we also introduce hand-aided attention in the keyframe selection process to make the network pay attention to hand-aided information. We send the feature maps after CNN into hand-aided attention to enhance the relevant features of human hands, thus making the selected frames more discriminative. We will introduce how hand-aided attention can better use hand information in Section III-C1. To capture the action information of the hands in the demonstration videos, we use the long short-term memory (LSTM) [34], [35] to aggregate the temporal features. After training, we randomly choose a starting position, select eight consecutive frames in demonstration videos, and feed into the network, and the frame with the highest confidence is selected. Only frames with correct action predictions or confidence above a certain threshold (typically 0.3) are kept. Repeating this process for each video yields several discriminative keyframes that provide affordance-related information.

C. Affordance Learning

After preprocessing the video, we introduce the HAG-Net. It is divided into two parts, as shown in Fig. 4. Fig. 4(a) shows the process of affordance learning of objects from demonstration videos, and Fig. 4(b) shows the process of affordance grounding. This section mainly introduces the process of affordance learning, and we will describe the affordance grounding in Section III-D. During the training phase, we divide the network into two branches to learn how humans interact with the objects in the videos and transfer the knowledge to the object branch.

For the video branch, let a video contain T frames $V = \{f_1, f_2, \ldots, f_T\}$ with an afforded action class $c$. First, we use a ResNet50 [33] backbone to extract the features of each frame, and the features of the video are represented as $X = \{X_1, X_2, \ldots, X_T\}$. We then input $X$ into hand-aided attention (in Section III-C1) to obtain the enhanced feature $M = \{M_1, M_2, \ldots, M_T\}$. After enhancing the relevant features of the human hand, we input them into the LSTM [34] for aggregating the hand action features

$$H(V) = \text{LSTM}(M_1, M_2, \ldots, M_T)$$  \hspace{1cm} (1)

where $H$ is the feature representation after aggregation. Finally, $H$ is sent to a classifier to predict the action class. For the object branch, we first extract the features of the object image using the same backbone. Then, the SEM $\text{SEM}(\cdot)$ is introduced to enhance the features. In addition, we introduce a projection layer $\text{Proj}(\cdot)$ to learn the object feature representation $\tilde{X}$ when the interaction occurs

$$\tilde{X} = \text{Proj}(\text{SEM}(X, c))$$  \hspace{1cm} (2)

$X$ is the feature map of the object image after ResNet50 [33]. Finally, we introduce a distillation loss ($L_{\text{distill}}$) to align

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
the feature space of static objects to the feature space of human–object interaction, from the affordance of the objects in the demonstration video to the static objects. To ensure that the feature representation of the object after the projection layer has the same classification performance as the video branch, we send the output of the projection layer to the same LSTM and action classifier and calculate the classification loss of the object branch. The total loss ($L_{\text{total}}$) is mainly composed of the following three parts:

$$L_{\text{total}} = \lambda_{\text{vcls}}L_{\text{vcls}} + \lambda_{\text{distill}}L_{\text{distill}} + \lambda_{\text{vcls}}L_{\text{vcls}}$$

where $\lambda_{\text{vcls}}$, $\lambda_{\text{distill}}$, and $\lambda_{\text{vcls}}$ are hyperparameters that balance the loss items. $L_{\text{vcls}}$ and $L_{\text{distill}}$ represent the cross-entropy loss of video branch and object branch for action classification, respectively. In the following, we present the details of hand-aided attention, SEM, projection layer, and distillation loss.

1) Hand-Aided Attention: The outcomes of Section III-B are utilized to enhance the characteristics of hand regions. In order to capture contextual details of the hands and objects, the identified hand bounding box is enlarged while eliminating the possible occluded hand regions. The expansion and exclusion regions are controlled by parameters $\alpha_1$ ($0 < \alpha_1 < 1$) and $\alpha_2$ ($0 < \alpha_2 < 1$). As shown in Fig. 3(b), $h$ and $w$ are the height and width of the hand bounding box, respectively. The expansion region is the bounding box enclosed by $(1+\alpha_1) \times w$ and $(1+\alpha_1) \times h$, while the exclusion region is the bounding box enclosed by $(1-\alpha_2) \times w$ and $(1-\alpha_2) \times h$. The hand mask (Mask$_h$) used in our network is shown in the right part of Fig. 3(b). We set $\alpha_2$ to 0.4 when calculating the hand-aided attention in the selection network and do not exclude the hand region, while in the affordance learning phase, since it will affect the extraction of affordance clues, we set $\alpha_2$ to 0.4 accordingly. After obtaining Mask$_h$, we calculate the masked feature $X^M_i$ by multiplying it with the feature map $X_i$ of the video frame via a positionwise dot product operator $\odot$

$$X^M_i = X_i \odot \text{Mask}_h.$$ (4)

Then, the feature representation of each frame and the mask of the hand are, respectively, passed through the L2 pooling layer (L2P(·)) and added to obtain the enhanced feature $M_i$

$$g_i = \text{L2P}(X_i), \quad g^M_i = \text{L2P}(X^M_i)$$ (5)

$$M_i = g_i + g^M_i$$ (6)

where $g_i$ and $g^M_i$ are the feature representations of the video frame and the masked feature after pooling, respectively.

2) Semantic Enhancement Module: The challenges of transferring objects from videos to static object images arise from appearance and viewpoint variance. Inspired by [36], an SEM is developed to improve the network’s ability to transfer affordance cues from video branches by enhancing object affordance class-related regions.

As shown in Fig. 4(a-2), the SEM enhances the object branch’s feature representations using a groupwise feature enhancement mechanism [37], [38]. This mechanism modulates the features of object images jointly according to global features and action categories, promoting the network’s focus on affordance-related features. The feature map of size $C \times H \times W$ is divided into $G$ groups along the channel dimension.

$$X^k = \text{Softmax}(\text{fc}(\text{one-hot}(c)))$$ (7)

where one-hot($\cdot$) is the one-hot representation of action class. $X^k$ is divided into $G$ groups. Spatial contextual cues are constructed using pixels and relationships between categories to reduce noisy information in the context and increase interpretability [39]. To achieve this objective, we use a global average pooling to aggregate the global statistical feature from $X^k$ and add it with $X^k_{\text{class}}$ to obtain the semantic vector $S^k$ of the $k$th group

$$S^k = \frac{1}{H \times W} \sum_{i=1}^{H \times W} X^k_i + X^k_{\text{class}}.$$ (8)

Then, we calculate the correlation coefficient $\epsilon_k$ between $X^k$ and $S^k$ and normalize it at each position as follows:

$$\epsilon^k = \frac{X^k_i \odot S^k_i}{\sigma_k + \epsilon}$$ (9)

$$\mu^k = \frac{1}{H \times W} \sum_{j=1}^{H \times W} \epsilon^k_j, \quad \sigma^k = \sqrt{\frac{1}{H \times W} \sum_{j=1}^{H \times W} (\epsilon^k_j - \mu^k)^2}$$ (10)

where $\odot$ means the elementwise multiplication operation, $\mu^k$ and $\sigma^k$ represent the mean and standard deviation of the group of coefficients, respectively, and $\epsilon$ is a regularization constant. After that, $\gamma$ and $\beta$ are introduced to scale and shift the normalized value $\tilde{c}_{ik}$:

$$a^k_i = \gamma^k \tilde{c}^k_i + \beta^k,$$ (11)

where $k$ denotes the $k$th group of parameters. Finally, $a^k_i$ is input to a sigmoid function $\sigma(\cdot)$ and multiply it with the original feature representation $X^k_i$ to get the enhanced feature representation $\tilde{X}^k_i$:

$$\tilde{X}^k_i = X^k_i \odot \sigma(a^k_i).$$

3) Projection Layer and Distillation Loss: The proposed approach includes a projection layer and a distillation loss, which aim to align the feature representations of the object and the video branches. This alignment decreases the feature space distance between the object in the static image and the object involved in the human–object interaction visible in the video. During aligning the feature representation of the interaction between humans and objects to that of static objects, we mainly focus on two factors: 1) the highest confidence frame $X_t$ in the input video contains the affordance clues of the hand’s position and 2) the average feature of the input contains the hand context. Specifically, we determine $t^*$ using the following equation:

$$t^* = \min_{i \in \{1, \ldots, T\}} L_{\text{fcls}}(\text{LSTM}(M_1, \ldots, M_T), c)$$ (11)

where $L_{\text{fcls}}$ is the cross-entropy loss for classification using the hidden state of LSTM at time $t$ (each frame of input). The average feature of the video is represented as $\bar{X}_V$: $X_V = (1/T) \sum_{t=1}^{T} X_t$. $L_{\text{distill}}$ is utilized to enable the projection
layer to align the feature representation of static and interactive objects
\[
L_{\text{distil}}(\tilde{X}, X_V, X_r) = \lambda_1 \times \|L2P(\tilde{X}) - L2P(X_r)\|_2 + \lambda_2 \times \|L2P(\tilde{X}) - L2P(X_V)\|_2
\]
where \(L2P(\cdot)\) denotes the L2 pooling operation and \(\lambda_1\) and \(\lambda_2\) are the loss weights. \(L_{\text{distil}}\) is the key to ensuring that the object image is mapped to the feature space of character interactions.

D. Affordance Grounding

The affordance grounding process aims to infer interaction regions for all possible actions on an object image with only a static image input. For each action class \(c\), the model generates a heatmap of the position where the human interacts with the object, as shown in Fig. 4(b). In this article, we use the Grad-CAM [40] feature visualization technique to establish the mapping between action labels to regions related to human–object interaction. In addition, in the ablation study in Section IV-D, alternative visualization techniques are evaluated, indicating that Grad-CAM obtains superior results. The detailed approach is given as follows: 1) for a given object image encoding \(X\) and action class \(c\), the sensitivity of the action with respect to each channel of the feature map’s last layer is computed to obtain an attention mask; 2) unlike Grad-CAM, we apply a weighted linear combination of the last layer of the feature map directly with the attention mask; and 3) the feature map is further processed through the ReLU activation function to focus on the positive pixels associated with a particular category:
\[
H_r(X) = \sum_k \text{ReLU} \left( \frac{\partial Y^c}{\partial X^k} \odot X^k \right)
\]
where \(\odot\) is the elementwise multiplication operator, \(X^k\) is the \(k\)th channel of the input embedding, \((\partial Y^c/\partial X^k)\) is a 2-D attention mask, and \(H_r(X)\) represents the final interaction heatmap for the given action class.

IV. EXPERIMENTS

In our experiments, we explore the following questions:

1) \(Q1\): Does our method outperform other weakly or fully supervised methods on the affordance grounding task? (See Section IV-C.)

2) \(Q2\): Does our method have excellent generalization performance on unseen objects? (See Section IV-C.)

3) \(Q3\): What is the influence of each module of our model and different visualization strategies on affordance grounding? (See Section IV-D.)

For evaluating our model’s effectiveness based on objective metrics and subjective visualization, we selected the OPRA [18] dataset for third-person perspective and the EPIC [20] dataset for first-person perspective. In addition, we conduct an ablation study where we assess varied visualization approaches and investigate the influence of each module of our model on affordance grounding.

A. Experimental Setup

1) Datasets: Our main goal is to use human demonstration videos to learn how people interact with objects. We need datasets that contain a large number of demonstration videos in which people interact with various objects. To this end, we conducted our experiments on the following two datasets.

1) \(\text{OPRA} [18]\): Fang et al. [18] proposed the OPRA dataset, which aims to use demonstration videos for object affordance inference. Each sample contains a video, an object image, an affordance class, and annotations of the interacting regions on the image. It includes 16763 training samples and 3798 test samples, covering seven different action categories.

2) \(\text{EPIC-KITCHENS (EPIC)} [20]\): The dataset contains a large number of egocentric videos of activities in kitchens. Each clip contains an action label and an object. Moreover, each frame has a bounding box of the object interacting with the person. This article uses the data annotated in [19], which contains 9236 training samples and 3520 test samples, covering 20 different action categories.

2) Evaluation Metrics: To evaluate the results of different models comprehensively, we choose four metrics from [41], including Kullback–Leibler divergence (KLD) [41], similarity metric (SIM) [42], AUC-J (AUC-J) [43], and NSS [44].

1) \(\text{KLD} [41]\): It is used to measure the distribution difference between the prediction map and the target map. Given a prediction map \(P\) and a ground-truth map \(Q^D\), \(\text{KL}(\cdot)\) is computed as follows:
\[
\text{KL}(P, Q^D) = \sum_i Q^D_i \log \left( \frac{Q^D_i}{\epsilon + P_i} \right)
\]
where \(\epsilon\) is a regularization constant.

2) \(\text{SIM} [42]\): The SIM measures the similarity between the prediction and the ground truth. Given a prediction map \(P\) and a continuous ground-truth map \(Q^D\), \(\text{SIM}(\cdot)\) is computed as the sum of the minimum values at each pixel, after normalizing the input maps
\[
\text{SIM}(P, Q^D) = \sum_i \min(P_i, Q^D_i)
\]
where \(\sum_i P_i = \sum_i Q^D_i = 1\).

3) \(\text{AUC-J} [43]\): AUC-J is a variant of AUC proposed by Judd et al. [43]. It measures the relative prediction map values at ground-truth locations.

4) \(\text{NSS} [44]\): The normalized scanpath saliency measures the correspondence between the prediction map and the ground truth, and it treats false positives and false negatives symmetrically. Given a prediction map \(P\) and a binary ground-truth map \(Q^D\), \(\text{NSS}(\cdot)\) computes the average normalized prediction at ground-truth locations
\[
\text{NSS}(P, Q^D) = \frac{1}{N} \sum_i \tilde{P} \times Q^D_i
\]
where \(N = \sum_i Q^D_i\) and \(\tilde{P} = ((P - \mu(P))/\sigma(P)))\).
3) Implementation Details: Each video in the OPRA [18] dataset has an object image associated with it, while the EPIC [20] dataset does not. Thus, for the EPIC dataset, we use the provided bounding box to crop the object from the video frame according to [19] and randomly select an image whose class matches the object class in the video. For the OPRA dataset [18], we set $\alpha_1$ to 0.2, $\alpha_2$ to 0.5, and the number of groups $G$ of the SEM to 2. For the EPIC dataset [20], we set $\alpha_1$ to 0.2, $\alpha_2$ to 0.3, and the number of groups $G$ of the SEM to 32. We also discuss the impact of different hyperparameters on the model’s performance in Section IV-D.

Due to the complexity of EPIC’s background and the distinct features between objects in the video and the object image, we substitute triplet loss for $L_\text{distinct}$ features between objects in the video and the object of interest on the model’s performance in Section IV-D.

Due to the complexity of EPIC’s background and the same category and greater distances between objects of different categories. By reducing the gap between features belonging to the same category, the network can better focus on interacting objects and disregard irrelevant backgrounds. $L_\text{distill}$ is calculated as follows:

$$L_\text{distill} = \lambda_1 \times \max[0, d(L2P(X'), L2P(\tilde{X}))]$$

$$- d(L2P(X'), L2P(\tilde{X}')) + M]$$

$$+ \lambda_2 \times \max[0, d(L2P(X_0), L2P(\tilde{X}))]$$

$$- d(L2P(X_0), L2P(\tilde{X}')) + M]. \quad (17)$$

We set $\lambda_\text{ech} = \lambda_\text{ech} = 1$ on the two datasets. For OPRA dataset, we set $\lambda_1 = 1$, $\lambda_2 = 0.2$, and $\lambda_\text{distill} = 0.1$. For the EPIC dataset, we set $\lambda_1 = 1$, $\lambda_2 = 0.5$, and $\lambda_\text{distill} = 1$. We carry out all experiments on 1080Ti and set the batch size and learning rate to 128 and $1e^{-4}$, respectively. During training in the hand-related selection network, the video input is eight frames, while during affordance learning from demonstration videos, the video input is three frames. Furthermore, we set the stride of the last two residual stages of ResNet50 to 1 and used dilated convolutions in the convolutional layers. In this way, the output spatial resolution of ResNet50 is 1/8 of the input.

B. Contenders

We compare the performance of our method with several state-of-the-art methods on OPRA [18] and EPIC [20]. It is worth noticing that we select a series of saliency detection models as the comparison methods because the human visual system has the ability to quickly orient attention to the most informative parts of visual scenes and the research on salient object detection is derived from this ability of the human visual system to extract the most attention-grabbing objects from the image [49], [50]. These saliency detection models include Egogaze [45], Mnet [46], DeepgazeII [47], and Salgan [48]. The generated heatmaps by them represent the parts of the object first noticed by the human visual system. We directly use the trained saliency detection models for test. In addition, we also choose the Hotspot [19], Demo2vec [18], and Img2heatmap [19] models, where our model, Hotspot, and the saliency detection models are all weakly supervised, while Demo2vec and Img2heatmap are fully supervised. All these methods are briefly described as follows.

1) Center Bias: It generates a Gaussian heatmap at the center of an image. Thus, it is a simple baseline for a dataset with features of central bias.

2) Egogaze [45]: It is a hybrid gaze prediction model that exploits both the visual saliency of bottom-up and task-dependent attention transition. It is the first work to explore the attention transition model in the egocentric gaze prediction task and achieves state-of-the-art results.

3) Mnet [46]: Unlike previous works that predict saliency maps directly from the last layer of convolution neural network, the model fuses features extracted from different layers of the CNN. Their method contains three main blocks: feature extraction CNN, feature encoding network (weighting of low and high features), and a prior learning network. The model achieves promising results in all datasets for saliency detection.

4) DeepgazeII [47]: Unlike other saliency models, DeepGazeII does not perform additional fine-tuning of the VGG features and only trains some output layers to predict saliency on top of VGG.

5) Salgan [48]: It introduces the adversarial training mechanism of GAN for salient object prediction, which consists of two main parts: one predicts saliency maps based on the input image and the other discriminates whether the input is the prediction result or the ground-truth. It explores the application of GAN to salient object detection and achieves excellent results on relevant datasets.

6) Hotspot [19]: It is a weakly supervised way to learn the affordance of an object through video, and affordance grounding is achieved only through action labels.

7) Demo2vec [18]: The approach discussed is a supervised learning method that leverages coded representations from demonstration videos to infer objects’ affordances accurately. Specifically, the network processes an object image alongside a demonstration video and generates predictions regarding the video’s action category and corresponding interacting regions within the object image.

8) Img2heatmap [19]: The encoder is a VGG16 [51] pre-trained on Imagenet [52]. The decoder is a mirrored VGG16, where the max pooling is replaced by the upsampling operation. The network’s final output is a heatmap of the same size as the input, and the number of channels is the same as the action class.

C. Results Analysis

In this section, we compare the ability of different models for affordance grounding on two datasets and their generalization abilities on unseen objects. We also discuss the compatibility of our model for datasets with first- and third-person view videos. Finally, we discuss the effectiveness of our model for handling affordance with multiple possibilities.

1) Affordance Grounding Results: Table 1 (left) shows the comparison of the HAG-Net with the state-of-the-art methods on the OPRA [18] and EPIC [20] datasets. Our
results of the HAG-Net and comparative models, including saliency detection methods (Egogaze [45], Mlnet [46], DeepgazeII [47], and Salgan [48]) and Hotspot [19], and two affordance detection methods using mask labels as strong supervision during training, including Demo2vec [18] and Img2Heatmap [19], on the OPRA [18] and EPIC [20] datasets. K, S, A, and N represent the evaluation metrics KLD [41], SIM [42], AUC-J [43], and NSS [44], respectively. ↑/↓ indicate that higher/lower results are better.

| Method     | Dataset | Affordance Grounding | Generalization to Novel Objects |
|------------|---------|----------------------|---------------------------------|
|            | OPRA [18] | EPIC [20]             | OPRA [18] | EPIC [20] |
|            | Time (s) | K ↓ S ↑ A ↑ N ↑       | K ↓ S ↑ A ↑ N ↑       |
| Center bias| -       | 11.132 0.205 0.325 0.323 | 10.660 0.222 0.634 0.333 |
| Egogaze    | 0.026   | 2.428 0.245 0.640 0.247 | 2.241 0.273 0.614 0.281 |
| Mlnet      | 0.006   | 4.022 0.284 0.763 0.607 | 6.116 0.318 0.746 0.809 |
| DeepgazeII | 3.760   | 1.897 0.296 0.720 0.496 | 1.352 0.394 0.751 0.888 |
| Salgan     | 0.002   | 2.116 0.309 0.769 0.659 | 1.510 0.395 0.774 0.978 |
| Hotspot    | 0.087   | 1.427 0.362 0.806 0.907 | 1.258 0.404 0.785 0.923 |
| Ours       | 0.085   | 1.409 0.365 0.812 0.948 | 1.209 0.414 0.801 1.045 |
| Img2Heatmap| 0.014   | 1.473 0.455 0.821 0.894 | 1.400 0.359 0.794 0.925 |
| Demo2vec   | -       | 1.197 0.482 0.847 1.170 | - | - | - | - |

![Fig. 5](image-url)  
Visualizations of affordance heatmaps generated by different methods, including Hotspot [19], Saliency (Egogaze [45], Mlnet [46], DeepgazeII [47], and Salgan [48]), Img2Heatmap [19], and Demo2vec [18].

Our method surpasses all other weakly supervised methods in all metrics and is close to the supervised Demo2vec [18] and Img2Heatmap [19]. It proves that our method utilizes the affordance cues provided by hand position and action can achieve promising results. We also visualize the heatmaps generated by different methods in Fig. 5. Our method generates heatmaps closer to ground truth than those of Hotspot and saliency detection models. Moreover, there is no large response on object parts unrelated to actions. The results on some objects are even better than those of the supervised Img2Heatmap [19] and Demo2vec [18] methods. It shows that our method transfers the affordance cues from the hand to the static object, making the network pay more attention to the regions related to the affordance while suppressing the regions unrelated to the action.

2) Generalization to Novel Objects: To verify the generalization ability of our method on new objects, we redivde the datasets according to [19]. The results are shown in Table I (right). On the EPIC [20] dataset, our model outperforms all other methods in all metrics. On the OPRA [18] dataset, our method is superior to other methods in most metrics. It demonstrates that our method can predict the region of interaction with unseen objects. It may be due to the cues our method provides from hand motion and position information enabling the network to pay more attention to the local details of the particular affordance class of objects, thus having better generalization performance.

3) Compare the Results of Merging Two Datasets: To test the efficacy of our method on a complex multiview dataset, we combined the OPRA and EPIC datasets [18], [20].
TABLE II
RESULTS OF DIFFERENT METHODS ON THE MIXTURE OF OPRA [18] AND EPIC [20] DATASETS, I.E., A DATASET WITH IMAGES AT DIFFERENT VIEWS

| Dataset      | OPRA [18] + EPIC [20] |
|--------------|------------------------|
| Method       | K ↑ | S ↑ | A ↑ | N ↑ |
| Center bias  | 10.968 | 0.211 | 0.629 | 0.327 |
| Egogaze [45] | 2.361 | 0.255 | 0.635 | 0.327 |
| Minet [46]   | 4.735 | 0.297 | 0.757 | 0.676 |
| Deepgaze II  | 1.705 | 0.341 | 0.730 | 0.629 |
| Salgan [47]  | 1.902 | 0.389 | 0.771 | 0.768 |
| Hotspot [19] | 1.435 | 0.366 | 0.769 | 0.825 |
| Ours         | 1.378 | 0.373 | 0.786 | 0.901 |
| Img2heatmap [19] | 1.444 | 0.358 | 0.769 | 0.811 |

Fig. 6. Rank list. We rank the 20 results of different methods on the tasks of “affordance grounding,” “generalization to novel object,” and “merging two datasets” in Tables I and II, where (i, j) indicates how many metrics that the model i are ranked the jth. The red letter denotes the average rank.

experimental results are presented in Table II, which demonstrates the superiority of our method in all evaluation metrics. This indicates that our approach is capable of transferring hand position and action information to stationary objects within a complex dataset. It is because our method focuses more on hand-object interactions, is not too sensitive to changes in viewpoint, and can be more robust to complex backgrounds.

We compared the results of “affordance grounding,” “generalization to novel objects,” and “merging two datasets” for three task settings with 20 metrics, ranked the results of various methods on each metric in each task setting, and summarized the result as a matrix, where each element (i, j) indicates how many metrics the model i are ranked the jth. Fig. 6 shows the promising outcomes of our method across various task settings, demonstrating significant robustness and generalization ability.

4) Comparison on 12 Most Frequent Classes of EPIC: Due to the highly unbalanced data distribution in the EPIC dataset, some categories have less than 1% of the total number of samples. Such imbalance may heavily influence the experimental outcomes by prioritizing large-sample classes. In this study, we evaluate our method against others in the 12 most frequent categories to demonstrate its efficacy when confronted with a large number of samples. The experimental findings are concisely shown in Table III, indicating that our approach outperforms alternative methods in identifying regions where people interact with objects in the majority of the samples.

5) Comparison of the Action Classification Accuracy: Since our method and hotspot [19] only use action class as supervision during affordance learning, the accuracy of action recognition affects the results of affordance grounding. We compare the performance of action classification between ours and hotspot [19] on the EPIC [20] dataset. We only choose the eight most frequent classes and calculate the confusion matrix, as shown in Fig. 7. Our method performs better on “open,” “close,” “take,” “put,” and “wash,” demonstrating that it can better distinguish different actions by combining hand position and action-related affordance cues to improve the result of action classification.

6) Affordance With Multiple Possibilities: We also show the visual heatmap results to demonstrate that our model can address the challenge of multiple possibilities of affordance well. As shown in Fig. 8(a), different actions interact with different regions of the same object. Our model can localize to different object regions based on the action category. Fig. 8(b) shows that different interactions occur in the same region of the object, and our model can also accurately localize to the region associated with the affordance of the interaction. It demonstrates that our model can reason about human
intents through human actions and positions and address the challenge of ambiguity in affordance.

7) Running Time: Table I shows the runtimes of various models tested on a 1080ti GPU. Salgan exhibits the shortest execution time, while HAG-Net is on par with the recent Hotspot. However, Demo2vec is not included due to its unavailability as an open-source model. In the future, we will consider optimization of the model to achieve faster inference.

D. Ablation Studies

1) Impact of Visualization Strategy: In this section, we compare two more advanced visualization methods, i.e., XGrad-CAM++ [54], to verify the effectiveness of Grad-CAM [40] used in our method for affordance grounding both subjectively and objectively.

1) XGrad-CAM++ [53]: This article presents two axioms, sensitivity and consistency, and improves XGrad-CAM to adhere to these principles. It can effectively identify relevant regions associated with a specific class

$$w_k = \sum_i \sum_j \frac{X_{ij}^k}{\partial x_{ij}} \odot \frac{\partial Y^c}{\partial x_{ij}} \quad (18)$$

$$H_{c}^{xgradcam}(X) = \sum_k \text{ReLU}(w_k \odot X^k). \quad (19)$$

2) Grad-CAM++ [54]: Grad-CAM++ is an enhanced version derived from Grad-CAM, offering improved visual interpretation and localization of CNN. It provides a way to measure the importance of each pixel in the feature map to the overall decision of the CNN by introducing a pixel weighting of the gradients of the output to get better visualization results, which is calculated as follows:

$$\alpha_{ij}^k = \frac{\frac{\partial Y^c}{\partial x_{ij}}}{2 \frac{\partial Y^c}{\partial x_{ij}} + \sum_a \sum_b X_{ab}^k \odot \frac{\partial Y^c}{\partial x_{ij}}} \quad (20)$$

$$w_k = \sum_i \sum_j \alpha_{ij}^k \odot \text{ReLU} \left( \frac{\partial Y^c}{\partial x_{ij}} \right) \quad (21)$$

$$H_{c}^{xgradcam++}(X) = \sum_k \text{ReLU}(w_k \odot X^k). \quad (22)$$

The experimental results are shown in Table IV. As can be seen, using Grad-CAM in our method outperforms the two more advanced visualization methods. In the affordance grounding task setting, it outperforms XGrad-CAM [53] and Grad-CAM++ [54] by 13.3% and 18.9% in terms of NSS [44] metrics on the OPRA [18] dataset, respectively. On the EPIC [20] dataset, it exceeds XGrad-CAM by 14.6% and surpasses Grad-CAM++ by 30.8% in terms of NSS metrics. These results demonstrate that our method can obtain more accurate region localization. Some visualization results are shown in Fig. 9. The heatmaps generated by XGrad-CAM and Grad-CAM++ are large and contain many irrelevant interaction regions. At the same time, our results can better focus on affordance-related regions, probably because their calculations focus more on the object as a whole. In contrast, our method can retain local details to better focus on affordance-related local regions of the object.

2) Effectiveness of Different Modules: In this section, we investigate the impact of each module in our method on affordance grounding.

1) Random (a): In the preprocessing phase, a single frame is randomly chosen from the input and the process is iterated three times. We also remove the hand-aided attention during the affordance learning phase.

2) Random (b): In preprocessing, a single frame is randomly chosen from the input and the process is iterated three times. We keep the network structure for affordance learning as in Fig. 4(a).

3) w/o Hand-Aided Attention (a): We remove the hand-aided attention in the selection network during the preprocessing process and also remove the hand-aided attention in the HAG-Net.

4) w/o Hand-Aided Attention (b): We remove the hand-aided attention in the selection network during the preprocessing but keep it for affordance learning, as shown in Fig. 4.

5) w/o Hand-Aided Attention (c): There is no change in the preprocessing stage. However, during the affordance learning phase, the hand-aided attention of HAG-Net is removed.

6) w/o Select Frame: We input eight video frames into our HAG-Net without selecting keyframes.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE IV
COMPARISON OF DIFFERENT VISUALIZATION APPROACHES, INCLUDING GRAD-CAM++ [54], XGRAD-CAM [53], AND GRAD-CAM [40] FOR OBTAINING THE OBJECT AFFORDANCE REGION

| Dataset                  | Affordance Grounding | Generalization to Novel Objects |
|--------------------------|----------------------|---------------------------------|
|                          | OPRA [18]            | EPIC [20]                       | OPRA [18]            | EPIC [20]                       |
| Method                   | K ↓ S ↑ A ↑ N ↑      | K ↓ S ↑ A ↑ N ↑                | K ↓ S ↑ A ↑ N ↑      | K ↓ S ↑ A ↑ N ↑                |
| HAG-Net & Grad-CAM++ [54]| 1.482 0.356 0.770 0.797 | 1.319 0.400 0.750 0.799 | 1.436 0.366 0.773 0.797 | 1.322 0.398 0.758 0.786 |
| HAG-Net & XGrad-CAM [53] | 1.464 0.359 0.782 0.837 | 1.274 0.407 0.775 0.912 | 1.417 0.368 0.785 0.832 | 1.274 0.404 0.785 0.907 |
| HAG-Net & Grad-CAM (Ours)| 1.409 0.365 0.812 0.948 | 1.209 0.414 0.801 1.048 | 1.366 0.373 0.817 0.927 | 1.197 0.412 0.820 1.084 |

Table V
ABLATION RESULTS ON BOTH OPRA [18] AND EPIC [20] DATASETS. FOR DETAILS, PLEASE REFER TO SECTION IV-D

| Dataset                  | OPRA [18] | EPIC [20] |
|--------------------------|-----------|-----------|
| Method                   | K ↓ S ↑ A ↑ N ↑ | K ↓ S ↑ A ↑ N ↑ |
| Random (a)               | 1.446 0.358 0.798 0.877 | 1.281 0.403 0.772 0.922 |
| Random (b)               | 1.436 0.360 0.802 0.900 | 1.273 0.403 0.779 0.944 |
| w/o hand-aided attention (a) | 1.424 0.363 0.806 0.916 | 1.253 0.404 0.789 0.960 |
| w/o hand-aided attention (b) | 1.420 0.363 0.808 0.914 | 1.248 0.405 0.791 0.974 |
| w/o hand-aided attention (c) | 1.416 0.364 0.809 0.928 | 1.235 0.408 0.795 0.996 |
| w/o select frame         | 1.418 0.364 0.806 0.934 | 1.239 0.408 0.793 0.990 |
| Max score                | 1.437 0.359 0.801 0.890 | 1.261 0.403 0.787 0.966 |
| Average                  | 1.443 0.361 0.799 0.885 | 1.264 0.404 0.781 0.968 |
| w/o SEM                  | 1.419 0.363 0.809 0.917 | 1.241 0.407 0.797 1.001 |
| Ours                     | 1.409 0.365 0.812 0.948 | 1.209 0.414 0.801 1.045 |

Fig. 9. Heatmaps by using different visualization approaches including Grad-CAM++ [54], XGrad-CAM [53], and Grad-CAM [40] in our method.

7) Max Frame: We use the frame with the highest confidence when calculating the distillation loss.
8) Average Frames: We take the average of the input three frames when calculating the distillation loss.
9) w/o SEM: We remove the SEM from the object branch of the affordance learning network.

Table V presents the ablation study outcomes. The findings indicate that hand-aided attention is essential for affordance grounding. Specifically, the proposed model shows a significant enhancement of up to 8.1% and 13.3% (NSS) when compared to random (a) on the OPRA [18] and EPIC [20] datasets, respectively. It proves that the hand’s position and action provide essential clues for the affordance of learning objects from the demonstration videos. In the setting without selecting keyframes, where we input the original video frames into the network for training, it can still achieve good results but falls behind ours. For example, our method achieves a gain of 1.5% NSS on OPRA and 5.6% NSS on EPIC over the model “w/o select frame,” respectively. These results demonstrate that the preprocessing step of selecting keyframes enables the network to learn more critical cues, leading to better results.

Compared to using only the frame with the highest confidence, our method achieves relative improvements of 6.5% and 8.2% (NSS) on OPRA and EPIC datasets, respectively.
Furthermore, compared to using only the input average features, our method achieves relative improvements of 7.1% and 8.0% (NSS) on the two datasets. These results validate that calculating the distillation loss of the object and video branches from the highest confidence frame and the average feature of the input can better transfer the affordance cues of the video to the static objects. Compared with the model without SEM, our method achieves 3.4% and 4.4% (NSS) relative improvements on OPRA and EPIC, respectively. SEM can allow the object branch to focus on different parts based on appearance and action class information, thereby improving the results of affordance grounding.

3) Different Semantic Sources for the SEM: To investigate the impact of semantic input on the SEM module used to enhance affordance-related features, we replaced the affordance labels utilized in the SEM with the prediction results of the video branch. The experimental results are shown in Table VI. It shows that using more accurate affordance labels can improve the model performance. The SEM can effectively adapt the object branch’s network focus class-related feature regions according to the affordance category.

4) Effectiveness of Hand-Aided Attention in Other Models: We apply a preprocessing strategy and hand-aided attention to the hotspot model [19], investigating the role of hand cues in affordance grounding. As shown in Table VII, the results demonstrate that utilizing information from human hands can enhance the performance of existing approaches. Using preprocessed frames as input to the Hotspot results in a 1.7% increase in performance, while incorporating hand-aided attention yields a 1.9% improvement. Nonetheless, our method achieves superior performance due to the utilization of affordance labels in modulating the target image’s features and considering affordance clues derived from hand actions and positions. As a result, our approach excels in identifying human–object interactions and efficiently learning affordance-related contextual information.

5) Different Hyperparameters: We explore the impact of the different number of groupings on the model’s performance in the SEM, and the experimental results are shown in Table VIII. It shows that different numbers of groupings do not have a significant impact on model performance, i.e., the best results on the OPRA dataset are achieved at $G = 2$, while the best results on the EPIC dataset are achieved at $G = 32$. It may be because each group in the SEM captures specific semantics, and the EPIC action interaction categories are more fine-grained and require more groups to capture different class-related features. Hence, a larger number of groups delivers better results. We also investigate the impact of $\alpha_1$ and $\alpha_2$ in the hand-aided attention on model performance, and the experimental results are shown in Tables IX and X. $\alpha_1$ mainly controls the region size where the hand occlusion is removed. It suggests that removing the hand occlusion can effectively suppress irrelevant information and make the network focus on the features of the object region around the hand. For $\alpha_1$ taken from 0.3 to 0.5, the model performance is negatively affected by introducing too much irrelevant information, such as occlusion. We set it to 0.2 on both datasets. $\alpha_2$ controls the area of the object where the hand interacts. $\alpha_2$ in the OPRA dataset is set to 0.5, and in the EPIC dataset, it is set to 0.3, leading to the best performance. In the third-person OPRA dataset, the scale of the hand in the image is generally small, and a larger $\alpha_2$ needs to be used to effectively cover
interaction-related regions, while EPIC generally has a larger scale for the hand and a smaller $\alpha_2$ of 0.3 can effectively enhance human–object interaction-related regions.

V. CONCLUSION

To achieve the goal of localizing all potential regions of interaction involving objects, this article proposes a novel HAG-Net for learning the affordance of objects from demonstration videos. The HAG-Net leverages the information provided by the position and action of the hand, which addresses the problem of multiple possibilities in affordance grounding. Furthermore, the proposed SEM incorporates both action category and appearance features to adjust the object image’s feature representation jointly, enabling more effective localization of interaction-related regions. Experiments on two public datasets demonstrate that our method achieves state-of-the-art results for affordance grounding.

VI. DISCUSSION

In this section, we discuss the work’s weaknesses, the HAG-Net’s potential applications, and future research directions.

A. Weakness

Although our HAG-Net has achieved good affordance grounding performance, some limitations should be addressed in future work. The current solution lacks end-to-end capability and requires a distinct preprocessing stage. Aiming for an integrated affordance grounding approach, we plan to optimize frame selection and contextual feature selection within a singular framework. In addition, the efficiency of LSTM training is suboptimal. As such, we plan to explore transformer structures as they have demonstrated proficiency in handling sequential data, according to literature [55].

B. Potential Applications

There may be some potential applications of our method as listed in the following.

1) Our method can provide candidate operation areas for robot grasping. This vision-based robot grasping problem is a highly researched topic in robotics [56]. By mimicking human actions, the robot can effectively select relevant areas that can be manipulated for grasping.

2) Our method can assist visual perception agents to understand the scene [16] comprehensively. For example, the agent needs to know the semantic category of each part in a scene and how it interacts with people and feedback on the environment. Moreover, it can deliver a goal-oriented understanding of the object and the environment.

3) Visual affordance grounding can also be used in virtual reality [57], [58] applications, where the object affordance heatmap can highlight the regions where humans interact with an object, together with some virtually displayed information, e.g., affordance indicators or warning signs.

4) Object affordance information can be used to improve object design and manufacture in the physical domain. By incorporating human–object interaction habits and characteristics, objects can be optimized to enhance potential human–object interactions [59].

C. Future Research Directions

There are several promising directions for future research on this task.

1) Dataset: In the future, we will consider constructing a more extensive and richer dataset containing videos from first- and third-person perspectives [60], [61], complex backgrounds, diverse interactions, various object classes, and human pose annotations [62], [63].

2) Generalization: It is also worthwhile to further investigate how to improve the generalization ability [64] of affordance grounding methods, e.g., locating the affordance-related regions precisely among unseen objects is of practical importance [65].

3) Compatibility: First-person perspective videos offer a distinct vantage point of object interaction, attention, and intentionality, whereas third-person videos offer an objective perspective of human actions. This dissimilarity presents a substantial challenge when addressing compatibility between first- and third-person videos. To improve affordance grounding ability, it is worth exploring datasets comprising first- and third-person videos [66], [67].

4) Transferability: Third-person videos are not always suitable for robot manipulation tasks that require the agent to observe the environment and objects from a first-person perspective. Therefore, transferring affordance grounding knowledge from third-person to first-person scenarios is crucial in such cases. However, transferring affordance grounding knowledge between different perspectives [68] is still underexplored and of practical meaning.

5) Multi-Information Assistance: The affordance grounding task is related to many visual factors, and to obtain better results in practice, we need to investigate affordance grounding methods with the aid of other information, such as the introduction of texture cues [69], [70], depth cues, and human-related information [62].

6) Combined With Other Related Techniques: The optimization of HAG-Net can benefit from integrating strategies from various domains. Several researchers have proposed video object segmentation/saliency detection techniques [71], [72], [73] to identify objects in videos that engage with humans, alongside hand position and action cues. This approach can eliminate irrelevant background regions and locate human–object interaction-related areas more precisely, ultimately leading to enhanced prediction accuracy.

REFERENCES

[1] J. J. Gibson, “The theory of affordances,” Hilldale, PA, USA, 1977, vol. 1, no. 2, pp. 67–82.
[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 248–255.

[53] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, “Axiom-based grad-CAM: Towards accurate visualization and explanation of CNNs,” 2020, arXiv:2008.02312.

[54] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2018, pp. 839–847.

[55] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in vision: A survey,” 2021, arXiv:2101.01169.

[56] K. Fang et al., “Learning task-oriented grasping for tool manipulation from simulated self-supervision,” Int. J. Robot. Res., vol. 39, nos. 2–3, pp. 202–216, Mar. 2020.

[57] E. Fujinawa, S. Yoshida, Y. Koyama, T. Narumi, T. Tanikawa, and M. Hirose, “Computational design of hand-held VR controllers using haptic shape illusion,” in Proc. 23rd ACM Symp. Virtual Reality Softw. Technol., Nov. 2017, pp. 1–10.

[58] J. Kim, H. Oh, W. Kim, S. Choi, W. Son, and S. Lee, “A deep motion sickness predictor induced by visual stimuli in virtual reality,” IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 2, pp. 554–566, Feb. 2022.

[59] N. Zhao, Y. Cao, and R. W. H. Lau, “What characterizes personalities of graphic designs?” ACM Trans. Graph., vol. 37, no. 4, pp. 1–15, Aug. 2018.

[60] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, “Charades-ego: A large-scale dataset of paired third and first person videos,” 2018, arXiv:1804.09626.

[61] Y. Yang, W. Zhai, H. Luo, Y. Cao, L. Jiebo, and Z.-J. Zha, “Grounding 3D object affordance from 2D interactions in images,” in Proc. Int. Conf. Comput. Vis., 2023, pp. 1–25.

[62] J. Zhang, Z. Chen, and D. Tao, “Towards high performance human keypoint detection,” Int. J. Comput. Vis., vol. 129, no. 9, pp. 2639–2662, 2021.

[63] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, “ViTPose: Simple vision transformer baselines for human pose estimation;” in Proc. Adv. Neural Inf. Process. Syst., 2022, pp. 38571–38584.

[64] K. Zhu, Y. Cao, W. Zhai, J. Cheng, and Z.-J. Zha, “Self-promoted prototype refinement for few-shot class-incremental learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 6797–6806.

[65] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, “One-shot affordance detection,” in Proc. IJCAI, 2021, pp. 895–901.

[66] H. Yu, M. Cai, Y. Liu, and F. Lu, “First- and third-person video co-analysis by learning spatial-temporal joint attention,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 6, pp. 6631–6646, Jun. 2023.

[67] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, “Actor and observer: Joint modeling of first and third-person videos,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7396–7404.

[68] Y. Li, T. Nagarajan, B. Xiong, and K. Grauman, “Ego-exo: Transferring visual representations from third-person to first-person videos,” 2021, arXiv:2104.07905.

[69] Y. Zhai et al., “LEARNING VISUAL AFFORDANCE GROUNDED FROM DEMONSTRATION VIDEOS,” 2021.

[70] X. Dong, Y. Guo, Y. Gao, and B. Li, “Axiom-based grad-CAM: Towards accurate visualization and explanation of CNNs,” 2020, arXiv:2008.02312.

[71] J. Han, L. Yang, D. Zhang, X. Chang, and X. Liang, “Reinforcement cutting-agent learning for video object segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 9080–9089.

[72] D. Zhang, J. Han, L. Yang, and D. Xu, “SPFTN: A joint learning framework for localizing and segmenting objects in weakly labeled videos,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 475–489, Feb. 2020.

[73] Y. Ji, H. Zhang, Z. Jie, L. Ma, and Q. M. Jonathan Wu, “CASNet: A cross-attention Siamese network for video salient object detection,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 6, pp. 2676–2690, Jun. 2021.

**Luo et al.**: Learning Visual Affordance Grounding from Demonstration Videos

**Chattopadhay et al.**: Axiom-based grad-CAM: Towards accurate visualization and explanation of CNNs

**Fu et al.**: Axiom-based grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks

**S. Khan et al.**: Transformers in Vision: A Survey

**Yang et al.**: Grounding 3D object affordance from 2D interactions in images

**Zhai et al.**: One-shot affordance detection

**Y. Cao et al.**: First- and third-person video co-analysis by learning spatial-temporal joint attention

**Zhang et al.**: Towards-high performance human keypoint detection

**X. Dong et al.**: ViTPose: Simple vision transformer baselines for human pose estimation

**Zhai et al.**: Self-promoted prototype refinement for few-shot class-incremental learning

**Luo et al.**: Learning Visual Affordance Grounding from Demonstration Videos

**Chattopadhay et al.**: Axiom-based grad-CAM: Towards accurate visualization and explanation of CNNs

**Fu et al.**: Axiom-based grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks

**S. Khan et al.**: Transformers in Vision: A Survey

**Yang et al.**: Grounding 3D object affordance from 2D interactions in images

**Zhai et al.**: One-shot affordance detection

**Y. Cao et al.**: First- and third-person video co-analysis by learning spatial-temporal joint attention

**Zhang et al.**: Towards-high performance human keypoint detection

**X. Dong et al.**: ViTPose: Simple vision transformer baselines for human pose estimation

**Zhai et al.**: Self-promoted prototype refinement for few-shot class-incremental learning

**Luo et al.**: Learning Visual Affordance Grounding from Demonstration Videos

**Chattopadhay et al.**: Axiom-based grad-CAM: Towards accurate visualization and explanation of CNNs

**Fu et al.**: Axiom-based grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks

**S. Khan et al.**: Transformers in Vision: A Survey

**Yang et al.**: Grounding 3D object affordance from 2D interactions in images

**Zhai et al.**: One-shot affordance detection

**Y. Cao et al.**: First- and third-person video co-analysis by learning spatial-temporal joint attention

**Zhang et al.**: Towards-high performance human keypoint detection

**X. Dong et al.**: ViTPose: Simple vision transformer baselines for human pose estimation

**Zhai et al.**: Self-promoted prototype refinement for few-shot class-incremental learning

**Luo et al.**: Learning Visual Affordance Grounding from Demonstration Videos

**Chattopadhay et al.**: Axiom-based grad-CAM: Towards accurate visualization and explanation of CNNs

**Fu et al.**: Axiom-based grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks

**S. Khan et al.**: Transformers in Vision: A Survey

**Yang et al.**: Grounding 3D object affordance from 2D interactions in images

**Zhai et al.**: One-shot affordance detection

**Y. Cao et al.**: First- and third-person video co-analysis by learning spatial-temporal joint attention