

Grounded Affordance from Exocentric View

Hongchen Luo^{1*} · Wei Zhai^{1*} · Jing Zhang² · Yang Cao^{1,3} ✉ ·
 Dacheng Tao²

Received: date / Accepted: date

Abstract Affordance grounding aims to locate objects’ “action possibilities” regions, an essential step toward embodied intelligence. Due to the diversity of interactive affordance, *i.e.*, the uniqueness of different individual habits leads to diverse interactions, which makes it difficult to establish an explicit link between object parts and affordance labels. Human has the ability that transforms various exocentric interactions into invariant egocentric affordance to counter the impact of interactive diversity. To empower an agent with such ability, this paper proposes a task of affordance grounding from the exocentric view, *i.e.*, given exocentric human-object interaction and egocentric object images, learning the affordance knowledge of the object and transferring it to the egocentric image using only the affordance label as supervision. However, there is some “interaction bias” between personas, mainly regarding different regions and views. To this end, we devise a cross-view affordance knowledge transfer framework that extracts affordance-specific features from exocentric interactions and transfers them to the egocentric view to solve the above problems. Furthermore, the perception of affordance regions is enhanced by preserving affordance correlations. In addition, an affordance grounding dataset named AGD20K is constructed by collecting and labeling over 20K images from 36 affordance categories. Experimental results demonstrate that our method outperforms the representative models regarding objective

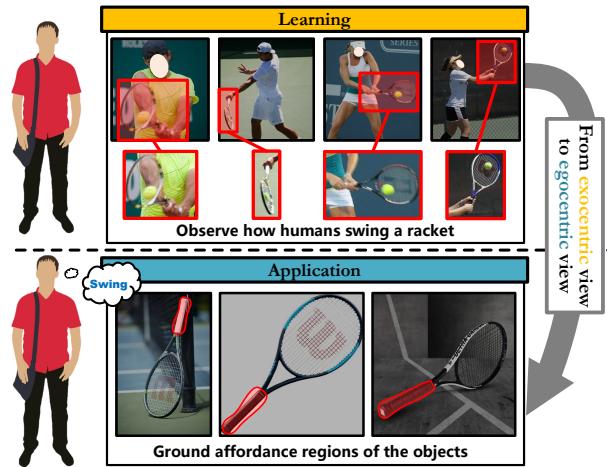


Fig. 1 Observation. By observing the exocentric diverse interactions, the human learns affordance knowledge determined by the object’s intrinsic properties and transfers it to the egocentric view.

metrics and visual quality. The code is available via: github.com/lhc1224/Cross-View-AG.

Keywords Affordance Grounding · Knowledge Transfer · Benchmark · Exocentric View · Egocentric View

1 Introduction

Affordance grounding aims to locate an object’s region of “action possibilities”. For an intelligent agent, it is necessary to know not only what the object is but also to understand how it can be used (Gibson 1977). Perceiving and reasoning about possible interactions in local regions of objects is the key to the shift from passive perception systems to embodied intelligence systems that actively interact with and perceive their environment (Bohg et al. 2017; Nagarajan and Grauman 2020). It has a wide range of applications for robot grasping, scene understanding, and action prediction (Grabner

Hongchen Luo (lhc12@mail.ustc.edu.cn)

Wei zhai (wzhai056@ustc.edu.cn)

Jing Zhang (jing.zhang1@sydney.edu.au)

✉ Yang Cao (forrest@ustc.edu.cn)

Dacheng Tao (dacheng.tao@gmail.com)

¹University of Science and Technology of China, Hefei, China

²The University of Sydney, Sydney, Australia

³Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

*Hongchen Luo and Wei Zhai contributed equally.

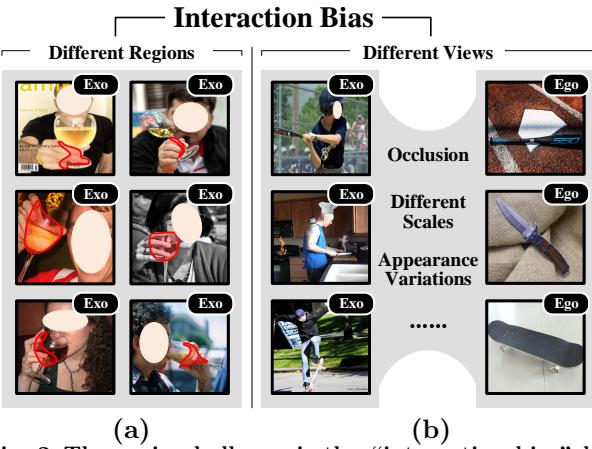


Fig. 2 The main challenge is the “interaction bias” between the personas. (a) “interaction bias” between regions, *i.e.* different habits leading to regional variations in interaction. (b) “interaction bias” between views, due to occlusion, different scales, and apparent variations, the affordance regions of the two views cannot align directly.

et al. 2011; Hassanin et al. 2021; Koppula et al. 2013; Luo et al. 2021a; Mandikal and Grauman 2021; Yang et al. 2021; Zhang and Tao 2020).

Affordance is a dynamic property closely related to the interaction between humans and environments (Hassanin et al. 2021). As shown in Fig. 1, the uniqueness of different individual habits leads to diverse interactions, making it difficult to understand how to interact with objects and establish links between object parts and affordance labels (Luo et al. 2021b). In contrast, humans can easily perceive the object’s affordance region by observing diverse exocentric human-object interactions and giving a unique egocentric definition. Although different persons hold the racket in different positions due to their habits, the observer can distinguish swingable regions determined by the intrinsic properties (such as the long handle structure) of the racket from a collection of interacting images and transfer the knowledge to the egocentric view, thereby creating a bridge between the object part and the affordance category.

To empower an agent with this ability to perceive the invariant egocentric affordance from various exocentric interactions, this paper first proposes a task of affordance grounding from the exocentric view, *i.e.*, given exocentric human-object interactions and egocentric images, learning affordance knowledge and transferring it to object images by only using affordance labels as supervision. During testing, the model predicts the affordance region for a specific object with the input of an egocentric image and a particular affordance label.

Bringing this power to real-world scenes would be a major leap forward, but doing so includes an issue with interpersonal “**interaction bias**” that has both subjective and objective aspects (as shown in Fig. 2). The subjective aspect is due to differences in individual habits leading to diverse interaction regions, making it

challenging to locate the affordance region accurately, as shown in Fig 2 (a). Despite such diversity, examining multiple human-object interaction images enables an exploration of objects’ generic affordance regions. Thus, it is viable to disintegrate different interactions into the affordance regions dictated by objects’ intrinsic features and human habit disparities, as shown in Fig. 3 (a). This paper employs non-negative matrix factorization (NMF) methodology (Lee and Seung 2000), to reduce the variability of human habits and obtain affordance features. The objective aspect refers to the occlusion, appearance, and scale differences, making it difficult to directly align the features in the affordance region between views (as shown in Fig. 2 (b)). This paper considers obtaining the cross-view matching matrix by densely measuring the similarity between the affordance-related exocentric feature and the egocentric feature. Then, adaptively adjust it to acquire the affordance feature representation in the egocentric view (as shown in Fig. 3 (b)). Furthermore, there is a co-relation between affordance categories, which is independent of the semantic classes of objects (as shown in Fig. 3 (c)). This paper enhances the network’s ability to perceive affordance regions by preserving the co-relation between affordance categories.

In this paper, we propose a **cross-view affordance knowledge transfer framework**. First, an **Affordance Invariance Mining (AIM)** module is introduced to extract affordance regions from diverse exocentric interaction regions by employing non-negative matrix factorization. Then, a **Cross-view Feature Transfer (CFT)** module is introduced to transfer them to the egocentric view by densely matching. Finally, an **Affordance Co-relation Preserving (ACP)** strategy enhances the network’s ability to perceive affordance regions by aligning the co-relation of affordance categories from both views. Specifically, the AIM module decomposes the human-object interaction into affordance-related features M and personal habit differences E . The non-negative matrix factorization technique minimizes E to obtain exocentric affordance features. In alternating iterations, the dictionary bases of the AIM module store the affordance features. The CFT module computes the cross-view matching matrix by evaluating the similarity between the dictionary bases and the features associated with each egocentric pixel. The bases and matrix are updated to accommodate differences between exocentric and egocentric views by adapting them based on egocentric features, which enables the transfer of exocentric affordance knowledge to the egocentric branch. Finally, the ACP strategy uses cross-entropy loss to align the co-relation matrices of the exocentric and egocentric branches to enhance the network’s ability to perceive and locate affordance regions.

Despite the advances in affordance learning, the existing datasets (Fang et al. 2018; Myers et al. 2015; Nguyen et al. 2017; Sawatzky et al. 2017) still bear lim-

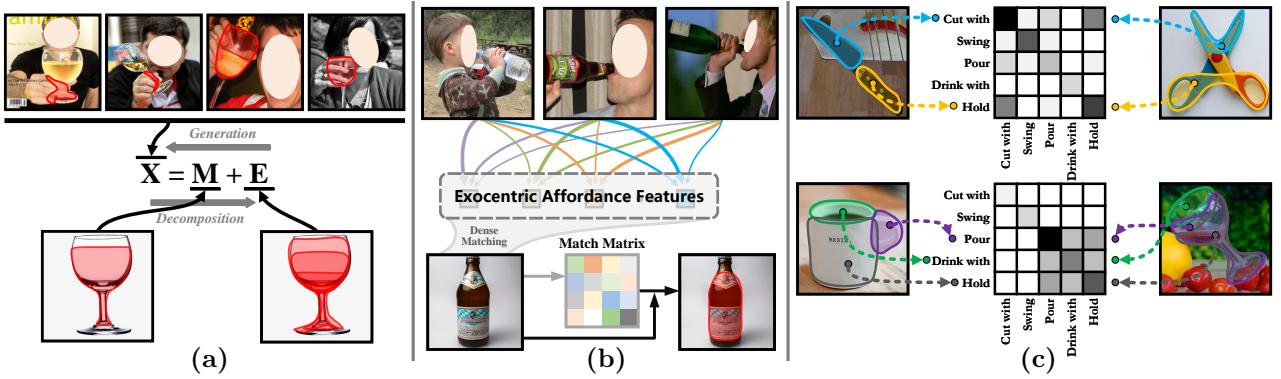


Fig. 3 Motivation. (a) Exocentric interactions can be decomposed into affordance-specific features M and differences in individual habits E (Sect. 3.1). (b) The model obtains the cross-view match matrix by densely matching the exocentric affordance features with each egocentric location feature and thus obtains the egocentric affordance feature based on the match matrix in an attentive manner (Sect. 3.2). (c) There are co-relations between affordances which are very common due to the multiple affordance properties of objects. This paper enhances the network’s ability to perceive affordance regions by aligning co-relations between views (Sect. 3.3).

itations in terms of affordance/object category, image quality, and scene complexity. To carry out a comprehensive study, this paper presents an affordance grounding dataset named **AGD20K**, consisting of 20,061 exocentric images and 6,060 egocentric images from 36 affordance categories. A comparative analysis of the AGD20K dataset uses eight prominent models across four related domains. The outcomes illustrate the superior efficacy of our approach in apprehending the intrinsic characteristics of objects and reducing the variability of affordance interaction. In summary, our primary contributions are:

- 1) We present an affordance grounding task from the exocentric view and establish a large-scale AGD20K benchmark to facilitate the research for empowering the agent to capture affordance features from exocentric interactions.
- 2) We propose a novel cross-view affordance knowledge transfer framework in which the affordance knowledge is acquired from exocentric human-object interactions and transferred to egocentric views while preserving the correlation between affordances, thus achieving better perception and localization of interactive affordance.
- 3) Experiments on the AGD20K dataset demonstrate that our method outperforms representative methods in several related fields and can serve as a strong baseline for future research.

This paper builds upon our conference version (Luo et al. 2022), which has been extended in three distinct aspects. **Firstly**, we provide a deeper insight into the problem of interpersonal “interaction bias” in the task caused by different regions and views. **Secondly**, we introduce a cross-view feature transfer module to effectively align affordance knowledge under the egocentric view by dense comparison. **Thirdly**, we extend the dataset from multiple aspects and conduct more experiments regarding multiple attributes to comprehensively analyze the model’s performance.

The remainder of this paper is organized as follows: Sect. 2 provides a brief review of existing related studies. Sect. 3 describes the pipeline of the proposed model and its details. In Sect. 4, we introduce the collection, annotation process, and statistical analysis of the AGD20K dataset. Sect. 5 describes the experimental setting and provides comprehensive results and analysis. In Sect. 6, we present the conclusions, limitations, and potential applications of this work.

2 Related Work

2.1 Visual Affordance Learning

Visual affordance research regards affordance perception as a computer vision issue that relies on images or videos. It employs machine learning or deep learning techniques to detect, segment, or ground “action possibilities” regions on objects (Hassanin et al. 2018). Table 1 lists recent works in affordance classification, detection, segmentation, and reasoning. Numerous previous studies (Chuang et al. 2018; Do et al. 2018; Fang et al. 2018; Nguyen et al. 2017; Zhao et al. 2020) primarily rely on supervised methods to create connections between local regions of objects and their corresponding affordances. Sawatzky and Gall (2017); Sawatzky et al. (2017) achieve weakly supervised affordance detection using only a few key points. Deng et al. (2021) expand affordance detection to 3D scenes. Luo et al. (2021b); Zhai et al. (2022) explore human purpose-driven object affordance detection in unseen scenarios. Mi et al. (2019, 2020) and Lu et al. (2022b) investigate affordance detection/segmentation in multimodal scenes. Nagarajan et al. (2019) exploit only affordance labels to ground the interactions from the videos. In contrast to (Nagarajan et al. 2019), our goal is to empower the agent to learn affordance knowledge from exocentric human-object interactions. To this end, we propose an explicit cross-view affordance knowledge trans-

Table 1 Comparison of affordance-related works. **Interaction:** the manner in which the drive model discovers affordance. **View:** the viewpoint of the input data. **CV:** cross-view. **SL:** supervised learning. **US:** whether valid on new unseen objects. **Rep:** representation of affordance.

| Paper | Interaction | View | CV | SL | US | Rep. | Dataset | Format | Task |
|-------------------------------------|-------------|---------|----|--------|----|--------------------|---------------------------|--------|----------------|
| Stark et al. (2008) | None | Exo | ✗ | Fully | ✗ | Obj | ETHZ Shape | 2D | Detection |
| Kjellström et al. (2011) | Vision | Exo | ✗ | Fully | ✗ | - | NORB | 2D | Classification |
| Koppula and Saxena (2014) | Vision | Exo | ✗ | Fully | ✗ | Obj & Tra | CAD-120 | RGBD | Grounding |
| Fouhey et al. (2015) | Vision | Exo | ✗ | Fully | ✗ | Scene | NYUv2, UIUC | RGBD | Segmentation |
| Myers et al. (2015) | None | Exo | ✗ | Fully | ✗ | Part | UMD, IIT-AFF | RGBD | Segmentation |
| Nguyen et al. (2016) | None | Exo | ✗ | Fully | ✗ | Part | UMD, IIT-AFF | RGBD | Segmentation |
| Do et al. (2018) | None | Exo | ✗ | Fully | ✗ | Part | UMD, IIT-AFF | RGBD | Segmentation |
| Zhao et al. (2020) | None | Exo | ✗ | Fully | ✗ | Part | UMD, IIT-AFF | RGBD | Segmentation |
| Lakani et al. (2017) | None | Exo | ✗ | Fully | ✗ | Part | UMD, IIT-AFF | RGBD | Segmentation |
| Srikantha and Gall (2016) | None | Exo | ✗ | Weakly | ✗ | Part | Srikantha and Gall (2016) | RGBD | Segmentation |
| Sawatzky et al. (2017) | None | Exo | ✗ | Weakly | ✗ | Part | | RGBD | Segmentation |
| Chuang et al. (2018) | Multimodal | Ego | ✗ | Fully | ✗ | Scene | ADE-AFF | RGB | Segmentation |
| Fang et al. (2018) | Vision | Exo | ✗ | Fully | ✗ | Part | OPRA | RGB | Grounding |
| Nagarajan et al. (2019) | Vision | Exo/Ego | ✗ | Weakly | ✓ | Part | OPRA, EPIC | RGB | Grounding |
| Luo et al. (2021a) | Vision | Exo/Ego | ✗ | Weakly | ✓ | Part | OPRA, EPIC | RGB | Grounding |
| Deng et al. (2021) | None | Exo | ✗ | Fully | ✗ | Point cloud & Part | 3D AffordanceNet | 3D | Segmentation |
| Luo et al. (2021b) | Vision | Exo | ✗ | Fully | ✓ | Obj | PAD | RGB | Segmentation |
| Mi et al. (2020) | Language | Exo | ✗ | Fully | ✗ | Obj | Mi et al. (2019) | RGB | Detection |
| Mi et al. (2019) | Language | Exo | ✗ | Fully | ✗ | Obj | Mi et al. (2019) | RGB | Detection |
| Lu et al. (2022b) | Language | Exo | ✗ | Fully | ✗ | Obj | PAD-L | RGB | Segmentation |
| Luo et al. (2022) & Ours | Vision | Exo/Ego | ✓ | Weakly | ✓ | Part | AGD20K | RGB | Grounding |

Table 2 Statistics of related datasets and the proposed AGD20K dataset. **Part:** part-level annotation. **HQ:** high-quality annotation. **BG:** the background is fixed or from general scenarios. **Exo&Ego:** whether to transfer from exocentric to egocentric view. **#Obj:** number of object classes. **#Aff:** number of affordance classes. **#Img:** number of images.

| | Dataset | Pub. | Year | link | HQ | Part | BG | Exo&Ego | #Obj. | #Aff. | #Img. |
|---|------------------------------|-----------|------|------|----|------|---------|---------|-------|-------|--------|
| 1 | UMD (Myers et al. 2015) | ICRA | 2015 | Link | ✗ | ✓ | Fixed | ✗ | 17 | 7 | 30,000 |
| 2 | (Sawatzky et al. 2017) | CVPR | 2017 | Link | ✗ | ✓ | Fixed | ✗ | 17 | 7 | 3,090 |
| 3 | IIT-AFF (Nguyen et al. 2017) | IROS | 2017 | Link | ✗ | ✓ | General | ✗ | 10 | 9 | 8,835 |
| 4 | ADE-Aff (Chuang et al. 2018) | CVPR | 2018 | Link | ✓ | ✓ | General | ✗ | 150 | 7 | 10,000 |
| 5 | PAD (Luo et al. 2021b) | IJCAI | 2021 | Link | ✓ | ✗ | General | ✗ | 72 | 31 | 4,002 |
| 6 | PADv2 (Zhai et al. 2022) | IJCV | 2022 | Link | ✓ | ✗ | General | ✗ | 103 | 39 | 30,000 |
| 7 | AGD20k (Ours) | This Work | 2022 | Link | ✓ | ✓ | General | ✓ | 50 | 36 | 26,117 |

fer framework that extracts affordance knowledge determined by the intrinsic properties of objects from multiple exocentric interactions and transfers it into egocentric images.

2.2 Visual Affordance Dataset

Numerous datasets supporting affordance-related tasks are available as summarized in Table 2. Myers et al. (2015) introduce a large-scale RGB-D dataset containing pixel-level affordance labels and corresponding ranks. Nguyen et al. (2017) construct the IIT-AFF dataset considering a more complex background of practical applications. Sawatzky et al. (2017) select video frames to construct a weakly supervised affordance detection dataset, using only cropped-out object regions but in inferior image quality. Luo et al. (2021b) consider inference human’s purpose from support images and transfers to a group of query images. Then, Lu et al. (2022b) consider multimodal scenarios and extended the PAD dataset to a phase-based affordance detection dataset, but both failed to provide part-level affordance labels. Other affordance datasets (Chuang et al. 2018; Fang et al. 2018; Myers et al. 2015; Nguyen et al. 2017) suffer from the problems of small scale and low affordance/object category diversity and do not consider human actions to reason about the affordance regions.

Chuang et al. (2018) take the physical world and social norms into account and construct the ADE-Affordance dataset. However, they do not consider the requirement for an intelligent agent to observe and learn from the exocentric view and transfer it to the egocentric view. In contrast to the above works, we explicitly consider exocentric-to-egocentric view transformations and collect a much larger scale of images, with richer affordance/object classes and part-level annotations, which are more valuable in developing affordance perception approaches towards practical real-world applications.

2.3 Learning View Transformations

The existing learning-view transformation works almost start from the mirror neurons theory (Rizzolatti and Craighero 2004), which adopts embedding learning to generate perspective invariant feature representations from paired data and then leverage it for tasks such as action recognition and video summarization under egocentric view (Fan et al. 2017; Ho et al. 2018; Lu et al. 2022a; Regmi and Shah 2019; Sigurdsson et al. 2018; Soran et al. 2014). Sigurdsson et al. (2018) construct a large-scale video dataset containing first- and third-person pairs, while they use the data to learn joint embeddings in a weakly supervised setting to align the two domains, thus effectively transferring knowledge from

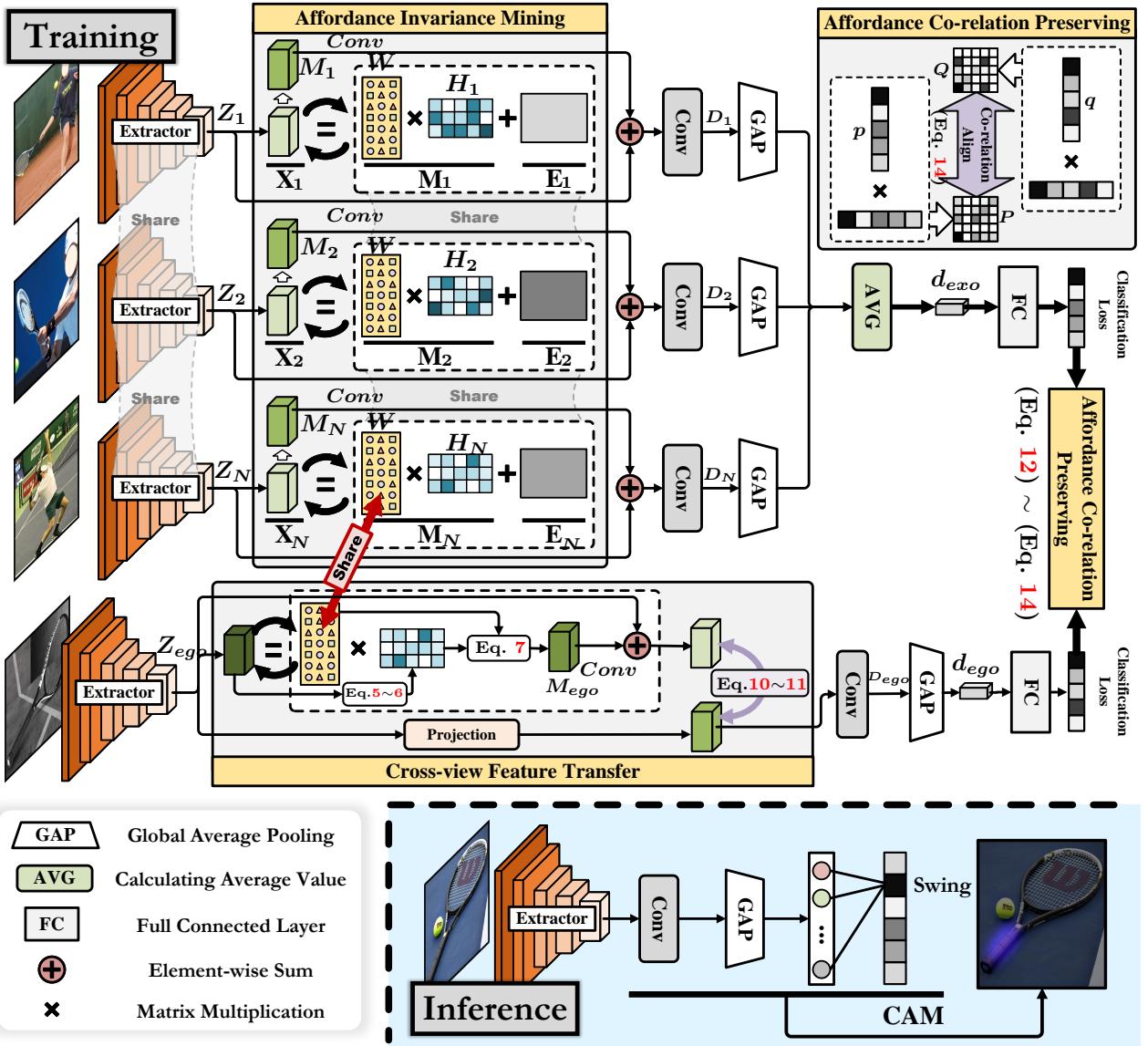


Fig. 4 Overview of the cross-view affordance knowledge transfer framework. It mainly consists of an Affordance Invariance Mining (AIM) module (Sect. 3.1), a Cross-view Feature Transfer (CFT) module (Sect. 3.2) and an Affordance Co-relation Preserving (ACP) strategy (Sect. 3.3).

third to first person. Fan et al. (2017) proposes to build person-level correspondence across perspectives while introducing a novel semi-Siamese CNN architecture to address this challenge. Li et al. (2021) extract key egocentric signals from the exocentric view dataset during pre-training and distill them to the backbone to guide feature learning in the egocentric video task. In contrast to the above works, we aim to extract affordance knowledge from the diverse exocentric human-object interactions and transfer it to the egocentric view. It is challenging due to the uncertainty caused by various interactions and objects' multiple affordance regions.

3 Method

The goal of the cross-view affordance grounding task is to locate the object affordance regions in egocentric images. During training, given a group of exocentric im-

ages $\mathcal{I}_{exo} = \{I_1, \dots, I_N\}$ (N is the number of exocentric images) and an egocentric image I_{ego} , the network uses only affordance labels as supervision, to learn affordance knowledge from exocentric images and transfer it to egocentric images. During testing, only given an egocentric image I_{ego} and the affordance label C_a , the model outputs the affordance region on the object. Our cross-view affordance transfer framework for affordance grounding is shown in Fig. 4. During training, we first use Resnet50 (He et al. 2016) to extract the features of exocentric and egocentric images to obtain $\mathcal{Z}_{exo} = \{Z_1, \dots, Z_N\}$ and Z_{ego} , respectively. Then, the Affordance Invariance Mining (AIM) module is introduced to extract affordance-specific clues ($\mathcal{F}_{exo} = \{F_1, \dots, F_N\}$) from the exocentric features (see in Sect. 3.1). Subsequently, the Cross-view Feature Transfer (CFT) module is proposed to transfer the affordance features extracted from the exocentric to the

egocentric view (see in Sect. 3.2). Afterward, the features of the two branches (\mathcal{F}_{exo} and F_{ego}) are fed into the same convolution layer to obtain features D_{exo} and D_{ego} respectively. We feed the D_{exo} through the global average pooling (GAP) layer to obtain the d_{exo} and pass the D_{ego} through the GAP layer to get the d_{ego} . Later, d_{exo} and d_{ego} are fed into the same fully connected layer to obtain the affordance prediction. Finally, the Affordance Co-relation Preserving (ACP) strategy is devised to enhance the network’s perception of affordance by aligning the co-relation matrix of the outputs of the two views (see in Sect. 3.3). During testing, we feed the egocentric object images into the network only through the egocentric branch and then use the CAM (Zhou et al. 2016) technique to obtain the affordance regions of the object (see in Sect. 3.4). Table 3 exhibits operation symbols, while Table 4 displays symbol dimensions, domains of definition, and meanings in the methods.

3.1 Affordance Invariance Mining Module

Human-object interaction varies across regions due to differences in human behavior, presenting challenges in obtaining complete feature representation for affordances from a single image. Nevertheless, it is possible to explore the generic features of the affordance regions from multiple human-object interaction images. These multiple interactions can be deconstructed using affordance invariant features and individual habits. The affordance invariance mining (AIM) module seeks to extract affordance invariant features from exocentric images depicting human-object interaction.

As shown in Fig. 4, the exocentric interactions are decomposed into affordance-specific features M and individual differences E . The aim is to improve the affordance feature representation M by reducing individual variation in habits E . Inspired by low-rank matrix factorization (Geng et al. 2021; Kolda and Bader 2009; Lee and Seung 2000), we represent the M as the multiplication of a dictionary matrix W and a coefficient matrix H , where the dictionary bases represent the subfeatures of human-object interaction and minimize E by iterative optimization to obtain a reconstructed affordance representation M . Specifically, for the input Z_i , we first reduce its dimensionality with a convolution layer and a ReLU layer to ensure the non-negativity of the input and then reshape them into $X_i \in \mathbb{R}^{c \times hw}$ (c , h and w are the channels, length, and width of the feature maps respectively). We use non-negative matrix factorization (NMF) (Lee and Seung 2000) to update the dictionary and the coefficient matrices. Consequently, X_i is decomposed into two non-negative matrices W and H_i . Here $W \in \mathbb{R}^{c \times r}$ is the dictionary matrix shared by all exocentric features, while $H_i \in \mathbb{R}^{r \times hw}$ is the coefficient matrix of each exocentric feature, and r is the rank of the low-rank matrix W . To update H_i and

W in parallel, we concatenate $\mathcal{X}_{exo} = \{X_1, \dots, X_N\}$ and $\mathcal{H} = \{H_1, \dots, H_N\}$ to obtain $X \in \mathbb{R}^{c \times Nhw}$ and $H \in \mathbb{R}^{r \times Nhw}$. The optimization process is as follows:

$$\min_{W,H} \|X - WH\|, \quad s.t. \quad W_{ab} \geq 0, H_{bk} \geq 0. \quad (1)$$

W and H are updated according to the following rules:

$$H_{ab} \leftarrow H_{ab} \frac{(W^T X)_{ab}}{(W^T W H)_{ab}}, \quad W_{ab} \leftarrow W_{ab} \frac{(X H^T)_{ab}}{(W H H^T)_{ab}}. \quad (2)$$

After several iterations, we get the output $M = WH$, and reshape it to $\mathcal{M}_{exo} = \{M_1, \dots, M_N\}$. Finally, we use a convolution layer (f) to map it to the residual space and sum it with the \mathcal{Z} to get the final output $\mathcal{F}_{exo} = \{F_1, \dots, F_N\}$:

$$F_i = \text{ReLU}(Z_i + f(M_i)), \quad i \in [1, N]. \quad (3)$$

In each batch of training, we update the initial dictionary matrix $W^{(0)}$ such that it can accumulate the statistical prior of the common subfeature of human-object interaction, i.e.,

$$W^{(0)} \leftarrow \alpha W^{(0)} + (1 - \alpha) \bar{W}, \quad (4)$$

where \bar{W} is the average over each mini-batch, and $\alpha \in [0, 1]$ is the momentum, which is set to 0.9 by default.

3.2 Cross-view Feature Transfer Module

The different views, scale, and occlusion cause difficulties in transferring affordance features from the exocentric to the egocentric view, and direct distillation may lead to lost spatial information and hard-to-perceive local detail features. Since the AIM module contains compact and comprehensive affordance region-specific detail cues after several iterations, we propose utilizing it as the initial value to densely compare the similarity of the egocentric feature with the dictionary base, ultimately achieving the cross-view matching matrix. Further, the dictionary bases are updated to adapt to the variation in viewpoint and scale in the egocentric branches by alternate iterative optimization with the egocentric features. It leads to improved optimization of the affordance region features activated in the egocentric. Finally, alignment is conducted to maintain the unique local details of the egocentric features during affordance knowledge transfer.

The CFT modoule is shown in Fig. 4. Specifically, Z_{ego} is passed through the convolution layer and ReLU to ensure the non-negativity of the egocentric features X_{ego} . The cross-view matching matrix H_{ego} is generated by performing dense comparisons between the egocentric feature pixels and the dictionary base W , followed by a Softmax activation for normalization:

$$X_{ego} = \text{ReLU}(f(Z_{ego})), \quad (5)$$

Table 3 Meaning of the operation

| Operation | Meanings |
|-------------|------------------------------------------|
| Conv f | Convolution operation |
| Max | Take the maximum value along the channel |
| Softmax | Softmax operation |
| GAP | Global average pooling |
| ReLU | ReLU activation function |
| Project | Project layer |
| $\ \cdot\ $ | L_2 loss |

$$H_{ego} = \text{Softmax}(X_{ego}^T W). \quad (6)$$

Subsequently, the values of W and H_{ego} are alternately adjusted utilizing an iterative optimization of the NMF of Eq. 2, enabling the adaptive tuning of the values of W to the egocentric branch features. Then, the activation features M_{ego} are obtained by reconstructing W and H_{ego} . Finally, M_{ego} is reshaped to $\mathbb{R}^{c \times h \times w}$ and mapped to the same dimension as Z_{ego} and augmented with the Z_{ego} to obtain the activation feature \tilde{F}_{ego} :

$$M_{ego} = W H_{ego}, \quad (7)$$

$$\tilde{F}_{ego} = \text{ReLU}(f(M_{ego}) + Z_{ego}). \quad (8)$$

To ensure that the egocentric features maintain their unique local details during the functional knowledge transfer, we perform an alignment with the egocentric branch features. Specifically, Z_{ego} is inputted to the project layer Project(\cdot) to produce the feature representation \bar{F}_{ego} :

$$\bar{F}_{ego} = \text{Project}(Z_{ego}). \quad (9)$$

Then, the maximum response value of each pixel region is selected for alignment:

$$\tilde{V}_{ego} = \text{Max}(\tilde{F}_{ego}), \quad \bar{V}_{ego} = \text{Max}(\bar{F}_{ego}), \quad (10)$$

where \tilde{V}_{ego} and \bar{V}_{ego} represent the channel maximum response values for each pixel of \tilde{F}_{ego} and \bar{F}_{ego} , respectively. Finally, the alignment is calculated as the L_2 loss between \tilde{V}_{ego} ($\tilde{V}_{ego} \in \mathbb{R}^{h \times w}$) and \bar{V}_{ego} ($\bar{V}_{ego} \in \mathbb{R}^{h \times w}$):

$$L_{KT} = \|\tilde{V}_{ego} - \bar{V}_{ego}\|. \quad (11)$$

3.3 Affordance Co-relation Preserving Strategy

There is a co-relation between the categories of affordances, and capturing such co-relation can enhance the network's capability to accurately perceive object affordances. Therefore, the affordance co-relation preserving (ACP) strategy is designed to exploit the co-relation between affordances, as shown in Fig. 4. First, we feed the feature representations of the two branches (d_{exo}

Table 4 The dimensions, domains of definition, and meanings of the symbols used in the proposed approach. Dim.: Dimensions.

| | Dim. | Domains | Meanings |
|-------------------|---------------------------|----------------------|------------------------------------------|
| I_i/I_{ego} | $3 \times 224 \times 224$ | $[-1, 1]$ | Exocentric/Egocentric image |
| Z_i/Z_{ego} | $2048 \times w \times h$ | $[-\infty, +\infty]$ | Exocentric/Egocentric feature |
| X_i | $c \times w \times h$ | $[0, +\infty]$ | Z_i after dimensionality reduction |
| X | $c \times Nhw$ | $[0, +\infty]$ | X_i reshape after concatenating |
| W | $c \times r$ | $[0, +\infty]$ | Dictionary matrix |
| H | $r \times Nhw$ | $[0, +\infty]$ | Coefficient matrix |
| M | $c \times Nwh$ | $[0, +\infty]$ | Reconstructed from W and H |
| F_i | $c \times Nhw$ | $[0, +\infty]$ | Output of the AIM module |
| X_{ego} | $c \times wh$ | $[0, +\infty]$ | Z_{ego} after dimensionality reduction |
| H_{ego} | $r \times hw$ | $[0, +\infty]$ | Coefficient matrix for CFT module |
| M_{ego} | $c \times hw$ | $[0, +\infty]$ | Reconstructed from W and H_{ego} |
| \tilde{F}_{ego} | $c \times h \times w$ | $[0, +\infty]$ | Activate feature for egocentric |
| \hat{F}_{ego} | $c \times h \times w$ | $[0, +\infty]$ | Project feature |
| D_{exo} | $1024 \times h \times w$ | $[-\infty, +\infty]$ | F_i after convolution |
| D_{ego} | $1024 \times h \times w$ | $[-\infty, +\infty]$ | F_{ego} after convolution |
| s/g | N_c | $[-\infty, +\infty]$ | Prediction scores |
| P/Q | $N_c \times N_c$ | $[0, 1]$ | Co-relation matrix |

and d_{ego}) into a shared fully connected layer to obtain the prediction scores s and g . Then, we align the affordance co-relation between the exocentric and egocentric views by calculating the cross-entropy loss (Hinton et al. 2015) L_{ACP} of the co-relation matrix of the prediction scores of the two branches:

$$p_j = \frac{\exp(s_j/T)}{\sum_k^{N_c} \exp(s_k/T)}, \quad q_j = \frac{\exp(g_j/T)}{\sum_k^{N_c} \exp(g_k/T)}, \quad (12)$$

$$P = p^T p, Q = q^T q, \quad (13)$$

$$L_{ACP} = - \sum_j^{N_c} \sum_k^{N_c} P_{jk} \log(Q_{jk}). \quad (14)$$

where T is a hyper-parameter used to control the degree of attention paid to the correlations between negative labels (Hinton et al. 2015). P_{jk} and Q_{jk} denote the correlation between categories j and k in the prediction results. Finally, the total loss can be calculated as:

$$L = \lambda_1 L_{cls} + \lambda_2 L_{ACP} + \lambda_3 L_{KT}. \quad (15)$$

The hyper-parameters λ_1 , λ_2 and λ_3 are utilized to balance the losses L_{cls} , L_{ACP} and L_{KT} . L_{cls} represents the total cross-entropy losses associated with the classification outcomes generated by both branches.

3.4 Inference

During the testing process, given only an egocentric image and an affordance label, the network localizes the corresponding affordance region (as shown in Fig. 4). Specifically, we utilize the class activation mapping (CAM) (Zhou et al. 2016) by computing a weighted sum of the feature maps D_{ego}^i of the last convolutional: $Y^{C_a} = \sum_i w_i^{C_a} D_{ego}^i$, where C_a is the affordance class, D_{ego}^i is the i -th layer feature map of D_{ego} , and $w_i^{C_a}$ is the weight of the fully connected layer corresponding to the i -th neuron and the C_a category.

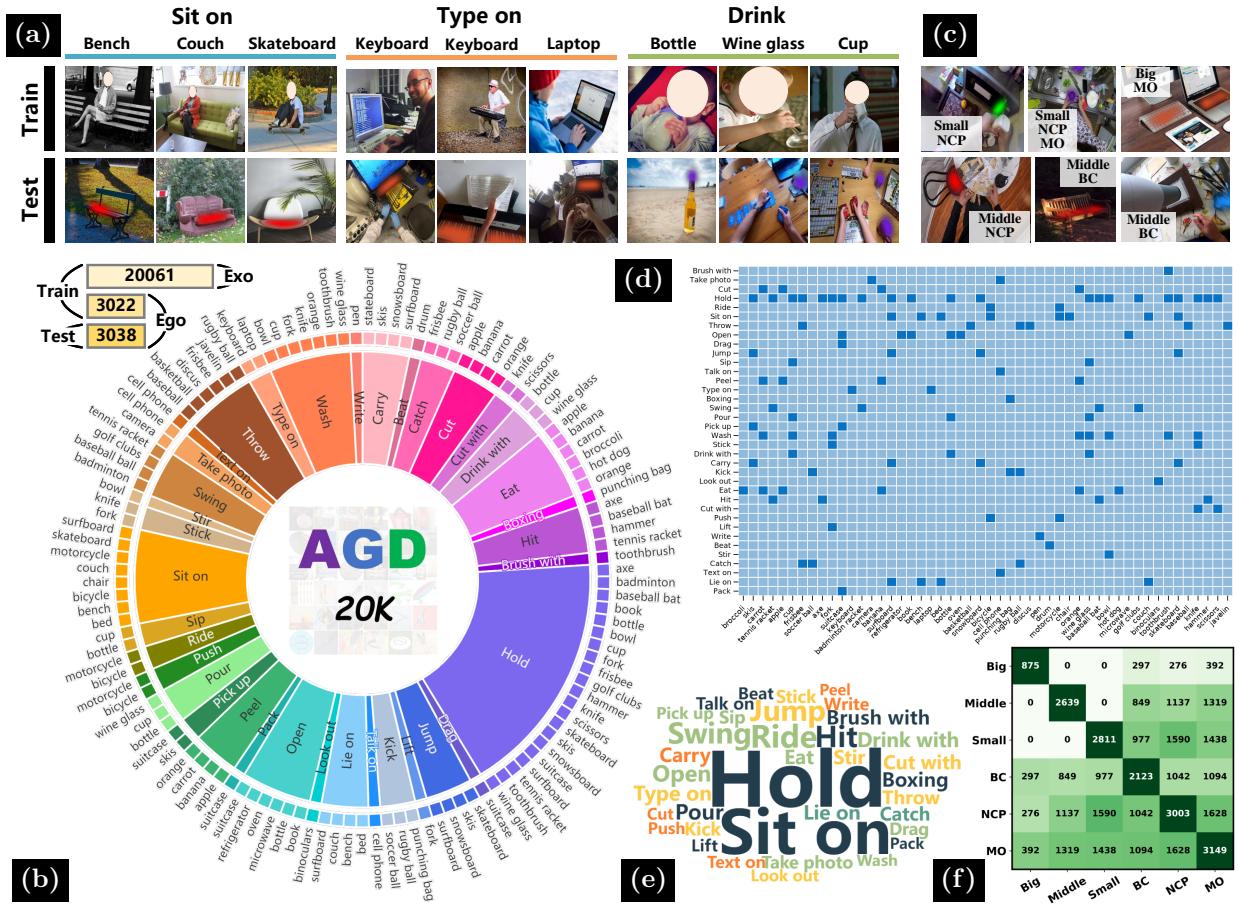


Fig. 5 Some examples and properties of AGD20K. (a) Some examples of training and test sets in AGD20K. (b) The distribution of categories in AGD20K. (c) Examples of the different attributes in the AGD20K test set. (d) The confusion matrix between the affordance category and the object category in AGD20K, where the horizontal axis denotes the object category, and the vertical axis denotes the affordance category. (e) The word cloud distribution of affordances in AGD20K. (f) The distribution of co-occurring attributes of the AGD20K test set.

4 Dataset

4.1 Dataset Collection

Based on the interactions that often occur in human daily life and the commonly used objects, we select 36 affordance categories, including indoor and outdoor scenarios in different weather conditions. The exocentric imagery is sourced primarily from COCO (Lin et al. 2014) and HICO (Chao et al. 2018), supplemented by data from PAD (Luo et al. 2021b), OPRA (Fang et al. 2018), and UCF101 (Soomro et al. 2012). Images from the HICO, COCO, and PAD are sorted based on affordance categories, then manually filtered to remove images exhibiting ambiguous interactions. The datasets mentioned above lack adequate examples of images for some interactions, and some affordance categories have no corresponding examples. We expand the dataset by selecting appropriate images from the OPRA (Fang et al. 2018) and UCF101 (Soomro et al. 2012) datasets to address this issue. We select video frames depicting human-object interactions with extended durations from various videos, including numerous diverse and intricate examples within the same interaction process.

To enrich the diversity of the dataset, we download and filter 2,112 exocentric images from free-license websites according to interaction and object categories.

There are two main components for the egocentric image: (1) Perception of how objects interact when the intelligent agent enters a new scene. In this instance, the egocentric image does not contain human-object interactions. (2) The perception of the object being interacted with and the interactable region of the surrounding objects when the intelligent body interacts with the object, in which case the egocentric image contains the human-object interaction. We download and select 4,744 images from the free-license websites for the first case. For the second case, we collect 1,316 images from EPIC-KITCHENS (Damen et al. 2018), Ego4dD (Grauman et al. 2022), THU-READ (Tang et al. 2017) etc, covering a wide range of complex human-object interactions in different environments. Compared to the original dataset, we add 989 and 1,316 images for the first and second cases, respectively. Thus, the number of images for both cases in the test set is approximately equivalent. These two components represent the com-

plete egocentric views within the AGD20K, encompassing the agent’s perception and interaction with its environment. Examples are shown in Fig. 5 (a).

4.2 Dataset Annotation

For each exocentric image, we assign affordance and object category labels based on the human-object interaction observed in the image. Given the object class contained in each affordance class, we assign affordance labels based on the object class in the egocentric images. Fig. 5 (c) shows the confusion matrix between the affordance and the object categories. We adopt heatmaps as part-level labels in the test set to better describe the “action possibilities” (*i.e.*, affordance). Specifically, we refer to the OPRA dataset (Fang et al. 2018) for annotating interaction regions, and the annotation routine from previous visual saliency works (Bylinskii et al. 2015, 2018; Judd et al. 2012). By observing the interactions between humans and objects in the exocentric images, we label the egocentric images with points of different densities according to the probability of interaction between the human and object regions. In generating the mask, we apply a Gaussian blur to each labeled point and normalize it to obtain the affordance heatmaps. Some examples are shown in Fig. 5 (a).

4.3 Statistic Analysis

To obtain deeper insights into our AGD20K dataset, we show its essential features from the following aspects. The distribution of categories in the dataset is shown in Fig. 5 (b), which shows that the dataset contains a wide range of affordance/object categories in diverse scenarios. The affordance and object categories confusion matrix is shown in Fig. 5 (d). It shows a multi-to-multi relationship between affordance and object categories, posing a significant challenge for the affordance grounding task. Fig. 5 (e) shows the word cloud statistics of AGD20K, implying an unbalanced data distribution, which also satisfies the fact that different interactions occur at different frequencies in the real world scenario. We divide the test set into three subsets, “Big”, “Middle”, and “Small”, according to the scale of the affordance region, *i.e.*, “Big”: if the proportion of the mask to the whole image is greater than 0.1, “Middle”: if the ratio is between 0.03 and 0.1, and “Small” for the remaining data. Furthermore, we split the test set into three subsets, “BC” (Background Clutter) (Swain and Ballard 1991), “NCP” (Negative Central Position) (Lv et al. 2022) and “MO” (Multiple Objects), according to the image background complexity, distance from the center, and whether it contains multiple objects. Fig. 5 (C) shows some examples of the attributes, while Fig. 5 (f) illustrates the confusion matrix for correlating different attributes in the test

set. Note that one image may have multiple attributes, increasing the difficulty of locating affordance regions.

5 Experiments

5.1 Metrics

Previous works mainly segment precise affordance regions (Chuang et al. 2018; Luo et al. 2021b; Myers et al. 2015; Nguyen et al. 2017), while the cross-view affordance grounding task considers a weakly supervised setting that predicts the affordance heatmap using only the affordance category label. Referring to the hotspots grounding-related works (Bylinskii et al. 2018; Fang et al. 2018; Liu et al. 2022; Nagarajan et al. 2019), we adopt heatmaps to give a better description of the “action possibilities” (*i.e.*, affordance) and use **KLD** (Bylinskii et al. 2018), **SIM** (Swain and Ballard 1991), and **NSS** (Peters et al. 2005) to evaluate the probability distribution correlation between the predicted affordance heatmap and Ground Truth (GT).

- **Kullback-Leibler Divergence (KLD)** (Bylinskii et al. 2015) measures the distribution difference between the prediction (P) and the ground truth (Q). It is computed as follows:

$$KLD(P, Q^D) = \sum_i Q_i^D \log \left(\epsilon + \frac{Q_i^D}{\epsilon + P_i} \right), \quad (16)$$

where ϵ is a regularization constant.

- **Similarity (SIM)** (Swain and Ballard 1991) measures the similarity between the prediction map (P) and the continuous ground truth map (Q^D). It is computed as follows:

$$SIM(P, Q^D) = \sum_i \min(P_i, Q_i^D), \quad (17)$$

where $\sum_i P_i = \sum_i Q_i^D = 1$.

- **Normalized Scanpath Saliency (NSS)** (Peters et al. 2005) measures the correspondence between the prediction map (P) and the ground truth (Q^D). It is computed as follows:

$$NSS(P, Q^D) = \frac{1}{N} \sum_i \hat{P} \times Q_i^D, \quad (18)$$

where $N = \sum_i Q_i^D$, $\hat{P} = \frac{P - \mu(P)}{\sigma(P)}$. $\mu(P)$ and $\sigma(P)$ are the mean and standard deviation, respectively.

5.2 Comparison Methods

To comprehensively evaluate the effectiveness of our approach in cross-view affordance grounding tasks, we choose eight advanced methods in four relevant fields, such as saliency detection (**DeepGazeII** (Kümmerer

Table 5 The results of different models on the original/additional test set. We compare the results of eight models, DeepGazeII (Kümmerer et al. 2016), EgoGaze , EIL (Mai et al. 2020), SPA (Pan et al. 2021), TS-CAM (Gao et al. 2021), BAS (Wu et al. 2021), Hotspots (Nagarajan et al. 2019) and Cross-view-AG (Luo et al. 2022), on the original test set and the newly added test set. The score before the slash represents the results of the original test set while the score after the slash represents the results of the additional test set. The best results are in **bold**.

| Method | Seen | | | Unseen | | |
|---------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|----------------------|-----------------------------|
| | KLD ↓ | SIM ↑ | NSS ↑ | KLD ↓ | SIM ↑ | NSS ↑ |
| DeepGazeII | 1.858 / 1.910 | 0.280 / 0.259 | 0.623 / 0.678 | 1.990 / 2.032 | 0.256 / 0.243 | 0.597 / 0.707 |
| EgoGaze | 4.185 / 4.194 | 0.227 / 0.222 | 0.333 / 0.438 | 4.285 / 4.537 | 0.211 / 0.193 | 0.350 / 0.401 |
| EIL | 1.931 / 1.903 | 0.285 / 0.274 | 0.522 / 0.778 | 2.167 / 2.141 | 0.227 / 0.226 | 0.330 / 0.577 |
| SPA | 5.528 / 5.779 | 0.221 / 0.232 | 0.357 / 0.506 | 7.425 / 7.376 | 0.169 / 0.193 | 0.262 / 0.390 |
| TS-CAM | 1.842 / 1.930 | 0.260 / 0.238 | 0.336 / 0.496 | 2.104 / 2.176 | 0.201 / 0.196 | 0.151 / 0.267 |
| BAS | 1.925 / 1.951 | 0.279 / 0.241 | 0.702 / 0.763 | 2.216 / 2.226 | 0.226 / 0.208 | 0.531 / 0.570 |
| Hotspots | 1.773 / 2.136 | 0.278 / 0.208 | 0.615 / 0.368 | 1.994 / 2.377 | 0.237 / 0.179 | 0.577 / 0.243 |
| Cross-view-AG | 1.538 / 1.684 | 0.334 / 0.296 | 0.927 / 1.071 | 1.787 / 1.936 | 0.285 / 0.249 | 0.829 / 0.905 |
| Ours | 1.478 / 1.576 | 0.342 / 0.314 | 1.012 / 1.228 | 1.749 / 1.848 | 0.281 / 0.258 | 0.897 / 1.050 |

Table 6 The results of different methods on the mixed testset. The best results are in **bold**. “Seen” means that the training set and the test set contain the same object categories, while “Unseen” means that the object categories in the training set and the test set do not overlap. The ◇ defines the relative improvement of our method over other methods.

| Method | Seen | | | Unseen | | |
|---------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| | KLD ↓ | SIM ↑ | NSS ↑ | KLD ↓ | SIM ↑ | NSS ↑ |
| DeepGazeII | 1.899 ◇18.1% | 0.264 ◇22.0% | 0.663 ◇76.5% | 2.020 ◇9.3% | 0.246 ◇6.1% | 0.677 ◇50.2% |
| EgoGaze | 4.195 ◇62.9% | 0.223 ◇44.4% | 0.409 ◇186.1% | 4.467 ◇59.0% | 0.198 ◇31.8% | 0.387 ◇162.8% |
| EIL | 1.914 ◇18.8% | 0.276 ◇16.7% | 0.708 ◇65.3% | 2.148 ◇14.7% | 0.226 ◇15.5% | 0.509 ◇99.8% |
| SPA | 5.719 ◇72.8% | 0.229 ◇40.6% | 0.466 ◇151.1% | 7.399 ◇75.2% | 0.186 ◇40.3% | 0.351 ◇189.7% |
| TS-CAM | 1.909 ◇18.5% | 0.243 ◇32.5% | 0.453 ◇158.3% | 2.129 ◇14.0% | 0.205 ◇27.3% | 0.277 ◇267.1% |
| BAS | 1.945 ◇20.1% | 0.251 ◇28.3% | 0.748 ◇56.4% | 2.223 ◇17.6% | 0.213 ◇22.5% | 0.560 ◇81.6% |
| Hotspots | 2.104 ◇26.1% | 0.215 ◇49.8% | 0.356 ◇228.7% | 2.332 ◇21.4% | 0.184 ◇41.8% | 0.245 ◇315.1% |
| Cross-view-AG | 1.647 ◇5.6% | 0.306 ◇5.3% | 1.032 ◇13.4% | 1.895 ◇3.3% | 0.259 ◇0.8% | 0.884 ◇15.0% |
| Ours | 1.555±0.007 | 0.322±0.001 | 1.170±0.020 | 1.832±0.005 | 0.261±0.003 | 1.017±0.012 |

Table 7 Parameters (M) and inference time (s) for all models.

| Method | DeepGazeII | EgoGaze | EIL | SPA | TS-CAM | BAS | Hotspots | Cross-view-AG | Ours |
|------------|------------|---------|-------|-------|--------|-------|----------|---------------|-------|
| Param. (M) | 20.44 | 46.53 | 42.41 | 69.28 | 85.86 | 53.87 | 132.64 | 120.03 | 82.27 |
| Time (s) | 3.760 | 0.026 | 0.019 | 0.081 | 0.023 | 0.057 | 0.087 | 0.023 | 0.022 |

et al. 2016), EgoGaze (Huang et al. 2018)), weakly supervised object localization (WSOL) (EIL (Mai et al. 2020), SPA (Pan et al. 2021), TS-CAM (Gao et al. 2021), BAS (Wu et al. 2021)), affordance grounding (Hotspots (Nagarajan et al. 2019)) and Cross-view affordnace knowledge transfer (Cross-view-AG (Luo et al. 2022)) for comparison. For the saliency detection models, we use models trained on the saliency datasets and test in the same way as (Nagarajan et al. 2019). For the weakly supervised object localization models, we only utilize exocentric images for training.

- **DeepGazeII** (Kümmerer et al. 2016): Unlike other saliency models, it does not perform additional fine-tuning of the VGG features and only trains some output layers to predict saliency on top of VGG (Simeonyan and Zisserman 2014).
- **EgoGaze** (Huang et al. 2018): The proposed model is a hybrid approach that merges task-dependent attention transitions and bottom-up saliency prediction to generate gaze predictions.
- **EIL** (Mai et al. 2020): It introduces a novel adversarial erasing technique jointly exploring highly

response class-specific areas and less discriminative regions to obtain a complete object region.

- **SPA** (Pan et al. 2021): It explores how to extract object structure information during training and proposes a structure-preserving activation method that leverages the structure information incorporated in the convolutional features for WSOL task.
- **TS-CAM** (Gao et al. 2021): It proposes a token semantic coupled attention map to take full advantage of the self-attention mechanism in visual transformer for long-range dependency extraction.
- **BAS** (Wu et al. 2021): The proposed model enhances the accuracy of foreground map generation through an activation mapping constraint module. This module helps in learning predicted maps by restraining background activation, resulting in more precise predictions.
- **Hotspots** (Nagarajan et al. 2019): It is a weakly supervised way to learn the affordance of an object through video, and affordance grounding is achieved only through action labels.
- **Cross-view-AG** (Luo et al. 2022): The model extracts affordance invariance cues from diverse ex-

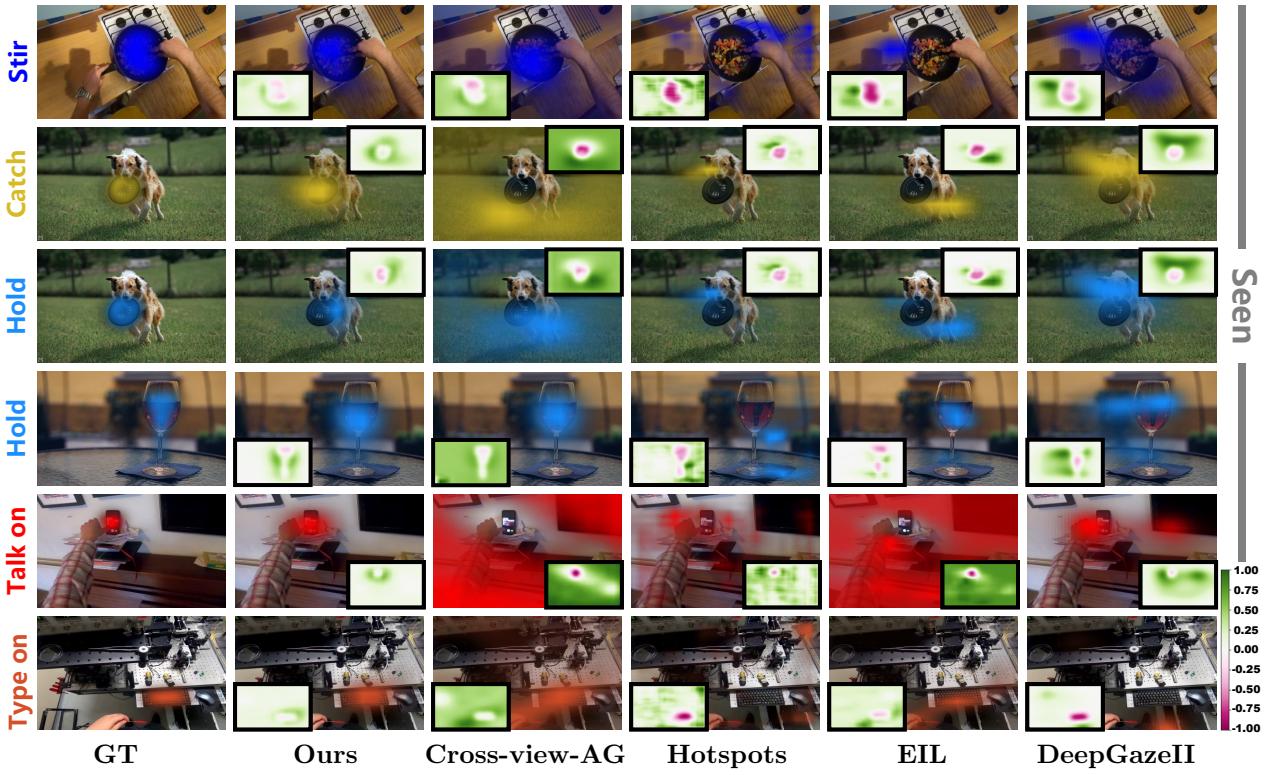


Fig. 6 Visualization of the difference maps at the “Seen” setting. We select results from representative models, such as Cross-view-AG (Luo et al. 2022), Hotspots (Nagarajan et al. 2019), EIL (Mai et al. 2020) and DeepGazeII (Kümmerer et al. 2016) for presentation. The difference maps represent the difference between the prediction and the GT.

ocentric interactions and transfer it to egocentric view. Furthermore, It improves the perception of affordance regions through the preservation of affordance co-relation

5.3 Implementation Details

Our model is implemented in PyTorch and trained with the SGD optimizer. With random horizontal flipping, the input images are randomly cropped from 256×256 to 224×224 . We train the model for 35 epochs on a single NVIDIA 3090ti GPU with an initial learning rate of $1e-3$. The hyper-parameters λ_1 , λ_2 , and λ_3 are set to 1, 0.5, and 0.5, respectively. We set the batch size to 32, and the number of exocentric images N is set to 3. The hyper-parameter T in the ACP strategy is set to 1. The dictionary matrix W ’s rank r and the number of iterations in the AIM module are set to 64 and 6, respectively. The number of channels of input features in the AIM module is 64. We divide the dataset into “Seen” and “Unseen”, in which “Seen” means that the training and test sets contain the same class of objects, while “Unseen” indicates that the training and test sets contain different classes of objects. The “Unseen” split can be used to evaluate the generalization ability of the models. We use Resnet50 (He et al. 2016) as the backbone while other advanced backbones (Liu et al. 2021; Zhang et al. 2023) can be explored in future

work. We use the same data augmentation method and batch size for the other methods. For weakly supervised object localization models (EIL (Mai et al. 2020), SPA (Pan et al. 2021), TS-CAM (Gao et al. 2021) and BAS (Wu et al. 2021)), the input is exocentric images during training, while the test input is egocentric images. For Hotspots (Nagarajan et al. 2019), three images are randomly sampled from the exocentric images during training and used as input to the video branch.

5.4 Quantitative and Qualitative Comparisons

Table 5 shows the performance of the different models on the original test set and the newly supplemented test set. The scores before and after the slash denote the outcomes on the original and newly established test sets. For models such as DeepGazeII (Kümmerer et al. 2016), BAS (Wu et al. 2021), and Cross-view-AG (Luo et al. 2022), which perform better in the original test set, they all show a performance drop on the new test set in most metrics. EIL (Mai et al. 2020) explores the whole object through adversarial erasure techniques. Thus there is no particular impact on performance for attributes such as complex backgrounds. However, it is not easy to obtain accurate part-level localization. Compared to the previous state-of-the-art Cross-view-AG (Luo et al. 2022) framework, the performance of our model degrades less on the new test set. Moreover, the evidence is more ob-

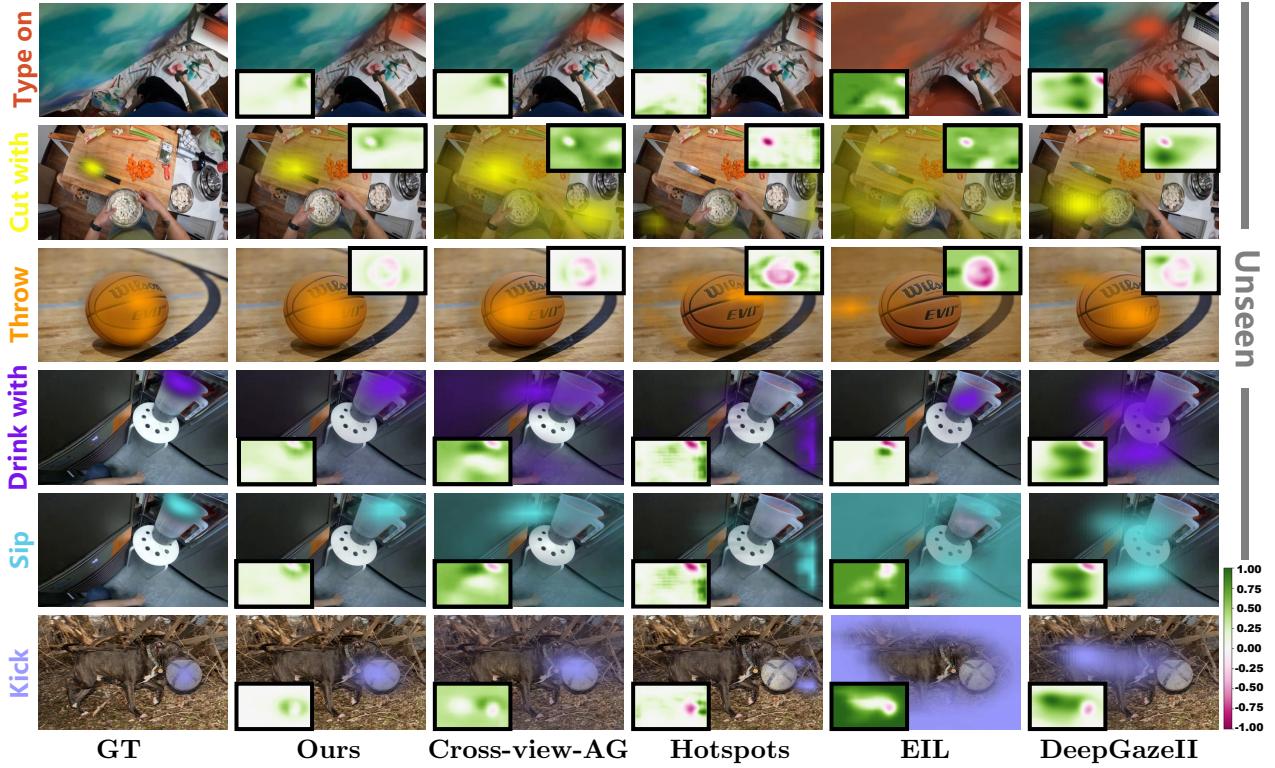


Fig. 7 Visual affordance heatmaps and difference maps at the “Unseen” setting. We select results from representative models, such as Cross-view-AG (Luo et al. 2022), Hotspots (Nagarajan et al. 2019), EIL (Mai et al. 2020) and DeepGazeII (Kümmerer et al. 2016) for presentation. The difference maps represent the difference between the prediction and the GT.

vious in the “Unseen” setting (see in Table 5 right), demonstrating the better generalization ability of our model in handling complex scenarios.

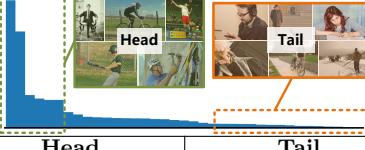
The subsequent experiments and the corresponding analyses are conducted on a combined dataset consisting of a mixture of the original and the new test sets. Table 6 shows the results for different related models, and it is evident that our approach achieves the best results for all metrics on both the Seen and Unseen settings. Taking KLD as the metric, our method improves **18.1%** compared to the best saliency model, **18.5%** over the best weakly supervised object localization (WSOL) model, **26.1%** over the affordance grounding model, and **5.6%** over the best cross-view affordance knowledge transfer model in the “Seen” setting. Our method on the “Unseen” setting improves **9.3%** compared to the best saliency model, surpasses the best WSOL model by **14.0%**, exceeds the affordance grounding model by **21.4%**, and outperforms the cross-view affordance knowledge transfer model by **3.3%**. Table 7 presents the various models’ parameter numbers and inference times. The number of parameters in our model is noticeably lower than Hotspots (Nagarajan et al. 2019) and Cross-view-AG (Luo et al. 2022), and comparable to TS-CAM (Gao et al. 2021). However, our model still has a larger number of parameters than the other methods, and thus, future work should aim to further reduce its size. Furthermore, the inference time of our model is comparable to that of various other methods.

Fig. 6 and Fig. 7 present the affordance maps for the “Seen” and “Unseen” settings, along with the discrepancy maps between the predicted outcomes and the actual ground truth. Compared to other models, our method can more accurately locate the affordance region of the object. Specifically, the red areas in the difference map are generally small and light, which indicates that our method can generate a complete affordance region of the object. For images with complex backgrounds (*e.g.*, Fig. 6, row 5 and Fig. 7, row 6), the model also can locate the affordance regions more accurately, indicating that our method generalizes well in complex scenes. For the cases where an object belongs to more than one affordance class or the same affordance contains multiple object classes with vastly different appearances, the model can accurately find the corresponding region, demonstrating that our method can effectively address the challenges posed by multiple possibilities of affordances.

5.5 Performance Analysis

Long Tail Distribution. The AGD20K dataset exhibits a long-tailed distribution characterized by significant data imbalance, as presented in Table 8. To validate whether our model can perform better on a small number of samples, we select the “Head” and “Tail” classes and test them separately. The results are shown

Table 8 Long Tail Distribution. We divide AGD20K into two subsets (“Head” and “Tail”, according to the number of images in the affordance class), and test the performance of the models in the two subsets separately.

| Method |  | | | | | |
|---------------|-----------------------------------------------------------------------------------|--------------|--------------|--------------|--------------|--------------|
| | Head | | | Tail | | |
| | KLD ↓ | SIM ↑ | NSS ↑ | KLD ↓ | SIM ↑ | NSS ↑ |
| DeepGazeII | 1.921 | 0.250 | 0.636 | 1.925 | 0.269 | 0.662 |
| EgoGaze | 4.132 | 0.222 | 0.414 | 4.370 | 0.206 | 0.366 |
| EIL | 1.980 | 0.269 | 0.658 | 2.036 | 0.258 | 0.622 |
| SPA | 4.176 | 0.226 | 0.424 | 7.997 | 0.205 | 0.368 |
| TS-CAM | 1.789 | 0.257 | 0.698 | 2.130 | 0.213 | 0.126 |
| BAS | 2.210 | 0.225 | 0.550 | 1.988 | 0.233 | 0.661 |
| Hotspots | 2.110 | 0.209 | 0.324 | 2.204 | 0.205 | 0.309 |
| Cross-view-AG | 1.691 | 0.288 | 0.931 | 1.812 | 0.284 | 0.928 |
| Ours | 1.607 | 0.303 | 1.045 | 1.713 | 0.300 | 1.080 |

in Table 8. Our model outperforms the other methods in the “Head” and “Tail” subsets. This improvement may stem from implementing the ACP strategy, which maintains the relationship between affordance classes and amplifies the network’s ability to recognize classes with minimal data.

Different Sources. To validate that the superiority of our method is not due to the additional egocentric images, we conducted retraining of the weakly supervised object localization models with both exocentric and egocentric images as input. Table 9 shows the results, indicating that using exocentric and egocentric images simultaneously enhances most approaches. However, the benefits of additional samples are limited and do not lead to significant gains. Our approach outperforms all these models, demonstrating that explicit knowledge transfer can efficiently transfer affordance knowledge to egocentric perspectives and attain more precise localization results.

Different Classes. Fig. 8 shows the results of the KLD metrics for each category in both “Seen” and “Unseen” settings, with deeper colors indicating better performance. Our model achieves the best results under most affordance categories, demonstrating our method’s superiority in locating affordance regions. In the “Seen” setting, our model obtains more accurate results for both the affordance categories “Hold” and “Cut with”, where there are some co-relations, demonstrating that our approach can enhance the network’s perception of affordance regions by aligning the co-relation of the two views. For affordance categories such as “Open” and “Carry”, where the interaction habits of different humans are quite diverse, our method still exceeds all other models, validating the effectiveness of the AIM module in extracting affordance-specific features for localization. In the “Unseen” setting, our model achieves promising results for the categories “Pick up”, “Sit on”, etc, with large variations between the object appear-

Table 9 Different sources. “Exo” means simply using exocentric images, while “Both” means using both exocentric and egocentric images in training.

| Method | Source | Seen | | | Unseen | | |
|-------------|--------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | KLD ↓ | SIM ↑ | NSS ↑ | KLD ↓ | SIM ↑ | NSS ↑ |
| EIL | Exo | 1.914 | 0.276 | 0.708 | 2.148 | 0.226 | 0.509 |
| | Both | 1.983 | 0.294 | 0.849 | 2.078 | 0.239 | 0.662 |
| SPA | Exo | 5.719 | 0.229 | 0.466 | 7.399 | 0.186 | 0.351 |
| | Both | 5.057 | 0.248 | 0.540 | 6.389 | 0.216 | 0.478 |
| TS-CAM | Exo | 1.909 | 0.243 | 0.453 | 2.129 | 0.205 | 0.277 |
| | Both | 1.882 | 0.254 | 0.511 | 2.124 | 0.206 | 0.272 |
| BAS | Exo | 1.945 | 0.251 | 0.748 | 2.223 | 0.213 | 0.560 |
| | Both | 2.188 | 0.263 | 0.734 | 2.002 | 0.225 | 0.826 |
| Ours | Both | 1.555 | 0.322 | 1.170 | 1.832 | 0.261 | 1.017 |

Table 10 Ablation study. We examine the effect of the AIM module, CFT module and ACP strategy on results.

| AIM CFT ACP | Seen | | | Unseen | | |
|-------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | KLD ↓ | SIM ↑ | NSS ↑ | KLD ↓ | SIM ↑ | NSS ↑ |
| ✓ | 1.765 | 0.289 | 0.878 | 2.003 | 0.247 | 0.725 |
| | 1.680 | 0.304 | 0.974 | 1.973 | 0.249 | 0.765 |
| ✓ | 1.659 | 0.313 | 1.009 | 1.936 | 0.248 | 0.828 |
| ✓ ✓ | 1.602 | 0.320 | 1.089 | 1.896 | 0.256 | 0.912 |
| ✓ ✓ | 1.641 | 0.317 | 1.048 | 1.912 | 0.251 | 0.874 |
| ✓ ✓ ✓ | 1.555 | 0.322 | 1.170 | 1.832 | 0.261 | 1.017 |

ances in the training and test sets, demonstrating that our method has a strong generalization ability to new object categories.

Different Attributes. Fig. 9 shows the results for all models on different attributes (“Big”, “Middle”, “Small”, “BC”, “NCP”, and “MO”). Our approach outperforms other models in almost all settings. In the challenging “Big” subset, our method still produces relatively strong results in terms of the NSS metric. Although the results are decreasing in the KLD and SIM metrics, they outperform the other methods. In the three subsets of “BC”, “MO”, and “NCP”, our model achieves superior results to the other models, indicating that our method is more robust and can accurately locate the affordance regions of the object.

5.6 Ablation Study

To investigate the impact of the AIM module, the CFT module, and the ACP strategy, we evaluate all combinations, as shown in Table 10. Note that since the matching of affordance regions in the CFT module requires an optimized dictionary base matrix in the AIM module, the CFT module’s presence must depend on the AIM module. It indicates that the AIM module extracts invariant affordance features from the diverse interactions of exocentric views, which enables the model to extract affordance-related features quickly. Meanwhile, the ACP strategy can enhance the perception of the affordance region by aligning the co-relation of the two branches. We make the T-SNE feature visualization of our model and baseline, as shown in Fig. 11. It

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|-----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|---------------|---------------|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| 1. 1.004 | 1.273 | 2.438 | 1.871 | 1.790 | 1.410 | 2.482 | 3.292 | 1.808 | 1.327 | 2.310 | 1.967 | 1.649 | 1.449 | 1.622 | 2.240 | 1.535 | 2.045 | 0.849 | 1.247 | 2.290 | 2.135 | 2.579 | 1.708 | 1.808 | 1.910 | 2.786 | 1.360 | 2.478 | 2.103 | 2.389 | 2.626 | 1.750 | 2.181 | 1.200 | 2.328 | |
| 2. 4.397 | 3.264 | 4.927 | 3.395 | 3.573 | 3.792 | 5.193 | 5.203 | 4.649 | 3.936 | 4.734 | 4.263 | 3.593 | 3.135 | 3.331 | 4.758 | 4.595 | 4.594 | 2.783 | 4.107 | 3.499 | 4.701 | 4.782 | 3.761 | 4.641 | 3.991 | 6.555 | 3.949 | 6.723 | 5.026 | 5.246 | 5.520 | 3.583 | 5.012 | 3.939 | 4.141 | |
| 3. 1.113 | 1.048 | 2.973 | 2.120 | 1.539 | 1.341 | 2.492 | 3.678 | 1.957 | 1.359 | 2.247 | 1.870 | 2.001 | 1.072 | 1.451 | 2.242 | 1.956 | 1.784 | 1.355 | 1.226 | 2.427 | 2.155 | 3.296 | 2.385 | 1.871 | 1.958 | 3.000 | 1.281 | 2.486 | 1.982 | 2.775 | 2.357 | 1.734 | 1.482 | 1.383 | 1.926 | |
| 4. 9.669 | 7.185 | 8.601 | 4.535 | 5.829 | 5.264 | 5.817 | 7.576 | 8.289 | 5.497 | 6.099 | 3.459 | 5.802 | 9.461 | 9.832 | 3.543 | 10.011 | 8.425 | 2.531 | 5.431 | 4.005 | 9.426 | 4.318 | 6.587 | 10.907 | 11.740 | 11.485 | 5.732 | 6.748 | 6.263 | 7.499 | | | | | | |
| 5. 1.330 | 1.398 | 2.672 | 1.830 | 1.695 | 1.665 | 1.665 | 2.518 | 3.049 | 2.022 | 1.729 | 2.295 | 1.641 | 1.884 | 1.403 | 1.520 | 2.415 | 1.231 | 1.924 | 1.223 | 1.692 | 2.539 | 2.180 | 2.741 | 1.889 | 2.015 | 1.848 | 2.786 | 1.560 | 2.420 | 2.282 | 2.532 | 2.821 | 1.610 | 2.215 | 1.629 | 2.265 |
| 6. 1.033 | 1.206 | 2.332 | 1.852 | 1.655 | 1.236 | 2.425 | 2.863 | 1.538 | 0.897 | 2.221 | 1.876 | 1.462 | 1.487 | 1.593 | 1.716 | 1.233 | 1.779 | 1.330 | 1.179 | 2.585 | 1.640 | 3.337 | 4.954 | 1.691 | 2.050 | 3.029 | 1.331 | 2.536 | 1.766 | 2.469 | 2.363 | 1.151 | 1.696 | 1.608 | 2.030 | |
| 7. 1.680 | 1.415 | 2.693 | 2.091 | 1.616 | 1.674 | 2.756 | 3.727 | 2.129 | 1.685 | 2.349 | 2.026 | 1.854 | 1.600 | 1.878 | 2.077 | 1.619 | 2.309 | 1.452 | 1.645 | 2.594 | 2.127 | 2.883 | 2.138 | 2.123 | 2.208 | 3.094 | 1.734 | 2.564 | 2.282 | 2.532 | 2.821 | 1.610 | 2.215 | 1.629 | 2.265 | |
| 8. 1.134 | 0.879 | 2.105 | 1.842 | 1.431 | 1.162 | 2.239 | 3.937 | 1.524 | 1.112 | 2.074 | 1.691 | 1.653 | 0.991 | 1.294 | 2.339 | 1.540 | 1.711 | 1.100 | 1.148 | 2.382 | 1.761 | 2.646 | 1.754 | 1.606 | 1.688 | 2.312 | 1.329 | 2.557 | 2.105 | 2.465 | 2.783 | 1.458 | 1.482 | 1.121 | 2.420 | |
| 9. 1.177 | 0.842 | 2.442 | 1.724 | 1.100 | 1.145 | 2.001 | 2.791 | 1.433 | 1.043 | 2.042 | 1.576 | 1.489 | 0.816 | 1.197 | 2.222 | 1.196 | 1.536 | 1.085 | 1.041 | 2.064 | 1.760 | 2.756 | 1.735 | 1.451 | 1.597 | 2.263 | 1.324 | 2.279 | 1.495 | 1.694 | 2.039 | 1.125 | 1.105 | 1.267 | 1.659 | |

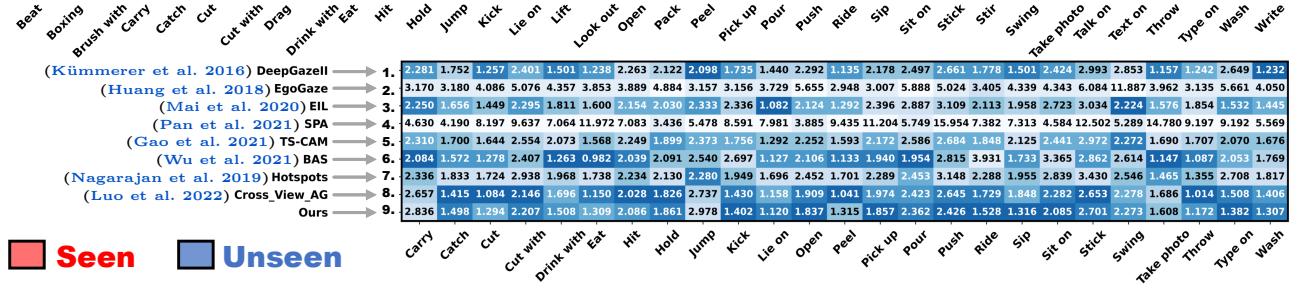
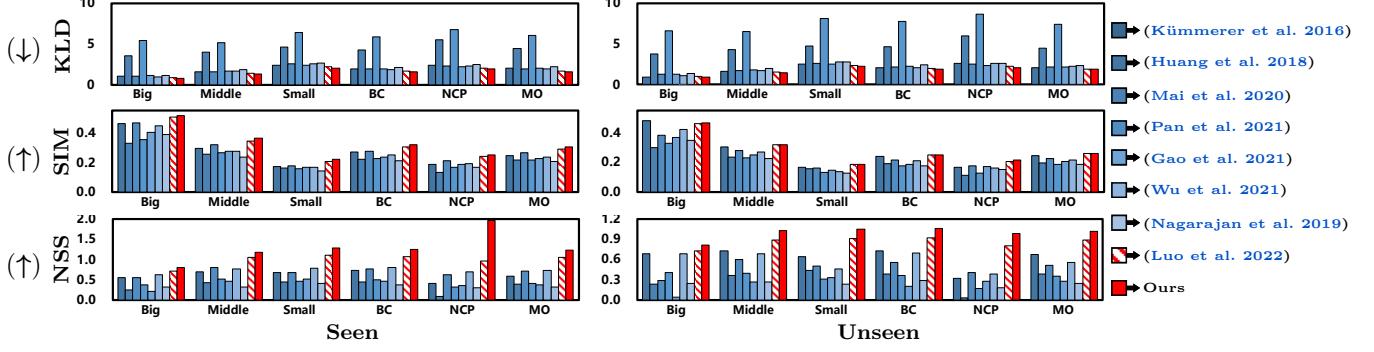


Fig. 8 The results of the different methods on the AGD20K for each affordance category. We calculate the KLD metrics for each affordance category in both “Seen” and “Unseen” settings, with darker colors representing better model performance.



shows that our method has more explicit discriminative boundaries and can accurately distinguish different affordance features. To analyze the features mined by the AIM module from multiple exocentric images, we visualized the correlation coefficient matrix H of the AIM module (as shown in Fig. 12). The maps represent the attention maps generated by different dictionary bases in W , indicating that different dictionary bases focus on different regions during human-object interaction. When the human drinks, the AIM module can focus on the region where the mouth is in contact with the bottle. The activated face can provide valuable contextual clues for reasoning about the action of “Drink”. At the same time, there is a co-relation between “Drink” and “Hold”, so the AIM module also activates the region of hand interaction. It indicates that different AIM module bases focus on interaction-related features and jointly form a better representation. To verify the ability of the CFT module to transfer affordance-related features, we visualize the output features of the CFT module. We take the mean value of the channels for egocentric features and obtain the results shown in Fig. 13. The CFT module can locate the interaction regions corresponding to affordance, thus providing more reliable features for cross-view knowledge transfer. To evaluate the ACP strategy

for object availability co-relation mining, we visualize the co-relation matrix of the output of the egocentric branch in the testing phase separately (shown in Fig. 10). The left figure shows that our model could better capture the co-relation between the affordance classes “Drink with” and “Hold” of the cup, and suppress irrelevant category predictions. The right figure shows that our approach can explore the possible existence of multiple affordance classes (“Hold” and “Hit”) in the same interaction region.

Fig. 14 (a) shows the influence of T in the ACP strategy on the model. T has a smoothing effect on the category correlation distribution and plays a preservation role for affordance co-relation. The performance is more sensitive to changes in T and has a greater impact at larger values. Fig. 14 (b) and (c) show the effect of the channel dimension c of the features and the rank r of the dictionary matrix W in the AIM module, respectively. The value of c has no significant impact on the results, and the model achieves a slightly better performance at $c = 512$. A larger number of channels may increase the complexity of the optimization and lead to a decrease in model performance. Different ranks represent the number of bases of the interaction subfeatures. The best results are obtained when $r = 64$. A smaller r (for example $r = 8$) may lead to poor results due

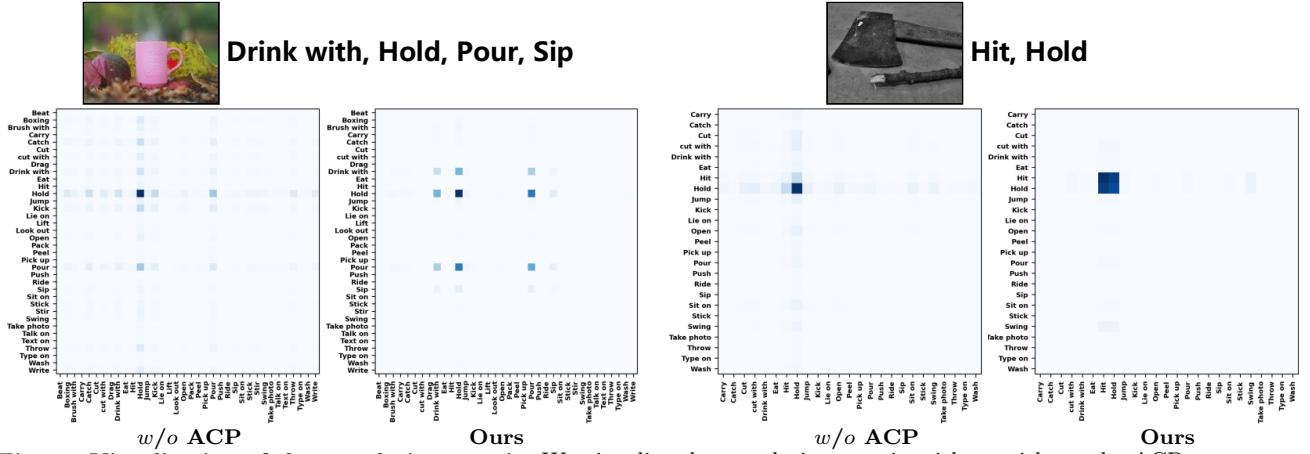


Fig. 10 Visualization of the co-relation matrix. We visualize the co-relation matrix with or without the ACP strategy.

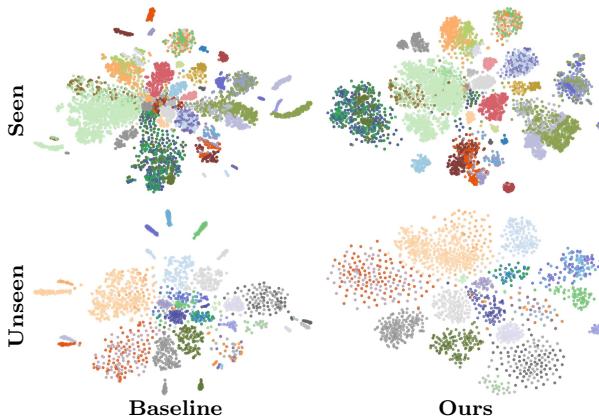


Fig. 11 T-SNE visualization results. The T-SNE results for the baseline without any modules and our model. The first row is the “Seen” setting and the second row is the “Unseen” setting. Due to the small number of images in the “Unseen” setting, the results in the second row are sparse.

to the number of bases being too small to represent the interactions’ sub-features fully. Moreover, a larger r leads to worse results possibly due to the redundancy of information caused by redundant bases. Fig. 14 (d) shows the impact of the number of exocentric images on model performance, which has a relatively positive effect on model performance as N increases from 1 to 3. It indicates that the AIM module can capture affordance-specific cues from multiple images, playing a critical role in affordance region prediction.

6 Conclusion and Discussion

In this paper, we make the first attempt to address a challenging task named affordance grounding from the exocentric view. Specifically, we propose a novel cross-view affordance knowledge transfer framework that can extract invariant affordance from diverse exocentric interactions and transfer it to an egocentric view. We also establish a large affordance grounding dataset named AGD20K, which contains 20K well-annotated images, serving as a pioneer testbed for the task. Besides, we

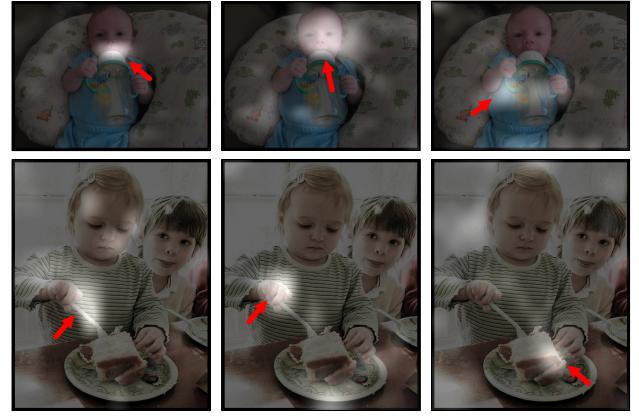


Fig. 12 Visualization of the coefficient matrix H for the last iteration of the AIM module. We choose maps activated by different bases.



Fig. 13 Visualization of the CFT module outputs.

expand the scale of the test set from a wide range of different attributes, making the test set more challenging and more suitable for real-world application scenarios. Our model outperforms eight representative models from four related areas and can serve as a strong baseline for future research.

Future Directions. Future research could focus on more precisely localizing human body parts (such as the hands, feet, mouth, *etc*) to interact with objects and recognizing each local region that the human body interacts with based on an exocentric view. Moreover, object affordance area localization could be studied in multimodal scenarios which contain language or audio data. Additionally, exploring the development of various prompts employing large models (Kirillov et al. 2023; Ramesh et al. 2022; Shen et al. 2023; Zhou et al.

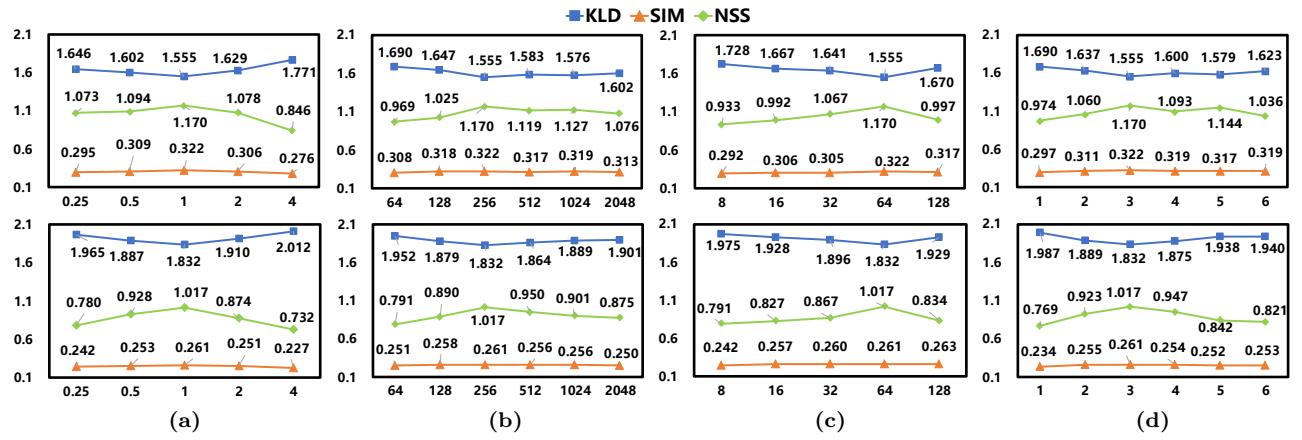


Fig. 14 Different Hyper-parameters. We explore (a) the effects of different T values in the ACP strategy, (b) different number of channels c in the input features in the AIM module, (c) different rank r of the dictionary matrix in the AIM module, and (d) different number of exocentric images N on the model performance. The first and second rows show the results of the experiments with the “Seen” and “Unseen” settings respectively.



Fig. 15 Failure cases. Some examples of the model’s failure in slender structures, complex structures and indistinct front and back views.

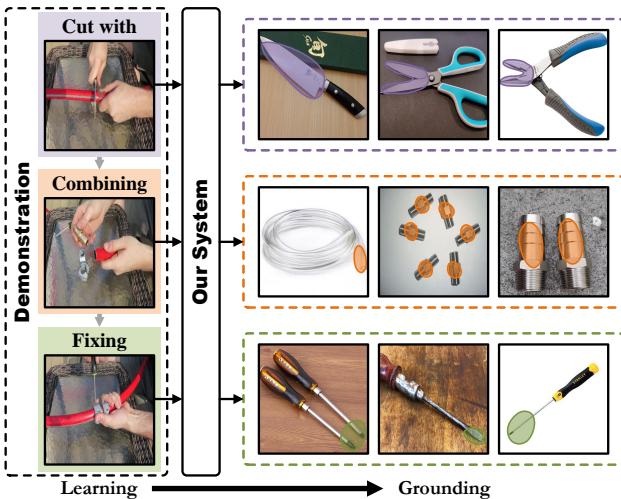


Fig. 16 Potential Applications: Learning from demonstrations. Our system can quickly extract the interaction region from the human demonstration image and locate it in the egocentric image.

2022) to enhance the identification of an object’s affordance with greater efficiency is worth examining.

Weakness. Fig. 15 shows some failure cases. Our model may activate irrelevant background when the structure is lengthy, thin, and complex or when the background and foreground cannot clearly distinguish. Future work could focus on enhancing the associated affordance class regions during training, ignoring the irrelevant background regions (Wu et al. 2021), and adjusting the obtained results according to affordance properties to produce more accurate predictions (Pan et al. 2021).

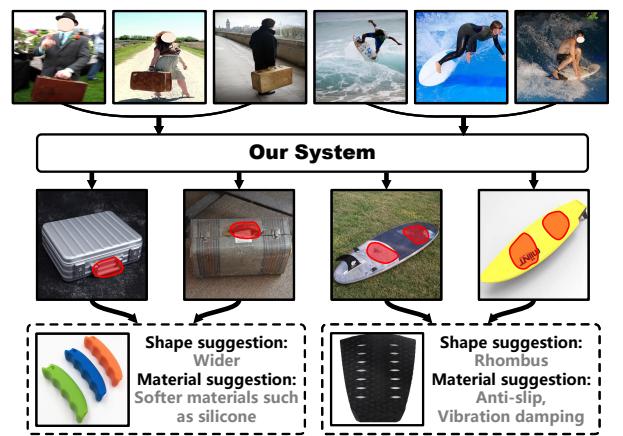


Fig. 17 Potential Applications: Industrial Manufacturing. Our system provides the ability to help design the product’s shape and material according to the interaction habits during the manufacturing process.

Potential Applications. **1) Learning from demonstrations:** As shown in Fig. 16, our system empowers the agent to efficiently acquire affordance-related knowledge from the human-object interaction in the exocentric view and to locate it in the egocentric image, which enables the agent to operate in the first-person view (Fang et al. 2018; Nagarajan et al. 2019). **2) Industrial Manufacturing:** As shown in Fig. 17, our system can explore the object’s interaction regions from multiple exocentric views of the human-object interaction and design more suitable shapes and materials to improve product quality during the manufacturing process (Lau et al. 2016).

References

- Bohg J, Hausman K, Sankaran B, Brock O, Kräig D, Schaal S, Sukhatme GS (2017) Interactive perception: Leveraging action in perception and perception in action. IEEE Transactions on Robotics 33(6):1273–1291

- Bylinskii Z, Judd T, Borji A, Itti L, Durand F, Oliva A, Torralba A (2015) Mit saliency benchmark
- Bylinskii Z, Judd T, Oliva A, Torralba A, Durand F (2018) What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 41(3):740–757
- Chao YW, Liu Y, Liu X, Zeng H, Deng J (2018) Learning to detect human-object interactions. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, pp 381–389
- Chuang CY, Li J, Torralba A, Fidler S (2018) Learning to act properly: Predicting and explaining affordances from images. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 975–983
- Damen D, Doughty H, Farinella GM, Fidler S, Furnari A, Kazakos E, Moltisanti D, Munro J, Perrett T, Price W, et al. (2018) Scaling egocentric vision: The epic-kitchens dataset. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp 720–736
- Deng S, Xu X, Wu C, Chen K, Jia K (2021) 3d affordancenet: A benchmark for visual object affordance understanding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 1778–1787
- Do TT, Nguyen A, Reid I (2018) Affordancenet: An end-to-end deep learning approach for object affordance detection. In: *2018 IEEE international conference on robotics and automation (ICRA)*, IEEE, pp 5882–5889
- Fan C, Lee J, Xu M, Singh KK, Yong JL (2017) Identifying first-person camera wearers in third-person videos. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*
- Fang K, Wu TL, Yang D, Savarese S, Lim JJ (2018) Demo2vec: Reasoning object affordances from online videos. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 2139–2147
- Fouhey DF, Wang X, Gupta A (2015) In defense of the direct perception of affordances. arXiv preprint arXiv:150501085
- Gao W, Wan F, Pan X, Peng Z, Tian Q, Han Z, Zhou B, Ye Q (2021) Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp 2886–2895
- Geng Z, Guo MH, Chen H, Li X, Wei K, Lin Z (2021) Is attention better than matrix decomposition? arXiv preprint arXiv:210904553
- Gibson JJ (1977) The theory of affordances. Hilldale
- Grabner H, Gall J, Van Gool L (2011) What makes a chair a chair? In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp 1529–1536
- Grauman K, Westbury A, Byrne E, Chavis Z, Furnari A, Girdhar R, Hamburger J, Jiang H, Liu M, Liu X, et al. (2022) Ego4d: Around the world in 3,000 hours of egocentric video. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 18995–19012
- Hassanin M, Khan S, Tahtali M (2018) Visual affordance and function understanding: A survey. arXiv
- Hassanin M, Khan S, Tahtali M (2021) Visual affordance and function understanding: A survey. *ACM Computing Surveys (CSUR)* 54(3):1–35
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 770–778
- Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. arXiv preprint arXiv:150302531
- Ho HI, Chiu WC, Wang YCF (2018) Summarizing first-person videos from third persons' points of view. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp 70–85
- Huang Y, Cai M, Li Z, Sato Y (2018) Predicting gaze in egocentric video by learning task-dependent attention transition. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp 754–769
- Judd T, Durand F, Torralba A (2012) A benchmark of computational models of saliency to predict human fixations
- Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, et al. (2023) Segment anything. arXiv preprint arXiv:230402643
- Kjellström H, Romero J, Krägíć D (2011) Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding* 115(1):81–90
- Kolda TG, Bader BW (2009) Tensor decompositions and applications. *SIAM review* 51(3):455–500
- Koppula HS, Saxena A (2014) Physically grounded spatio-temporal object affordances. In: *European Conference on Computer Vision (ECCV)*, Springer, pp 831–847
- Koppula HS, Gupta R, Saxena A (2013) Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research* 32(8):951–970
- Kümmerer M, Wallis TS, Bethge M (2016) Deepgaze ii: Reading fixations from deep features trained on object recognition. arXiv preprint arXiv:161001563
- Lakani SR, Rodríguez-Sánchez AJ, Piater J (2017) Can affordances guide object decomposition into semantically meaningful parts? In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, pp 82–90
- Lau M, Dev K, Shi W, Dorsey J, Rushmeier H (2016) Tactile mesh saliency. *ACM Transactions on Graph-*

- ics (TOG) 35(4):1–11
- Lee DD, Seung HS (2000) Algorithms for non-negative matrix factorization. In: NIPS
- Li Y, Nagarajan T, Xiong B, Grauman K (2021) Ego-exo: Transferring visual representations from third-person to first-person videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 6943–6953
- Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV), Springer, pp 740–755
- Liu S, Tripathi S, Majumdar S, Wang X (2022) Joint hand motion and interaction hotspots prediction from egocentric videos. arXiv preprint arXiv:220401696
- Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022
- Lu J, Zhou Z, Zhu X, Xu H, Zhang L (2022a) Learning ego 3d representation as ray tracing. arXiv preprint arXiv:220604042
- Lu L, Zhai W, Luo H, Kang Y, Cao Y (2022b) Phrase-based affordance detection via cyclic bilateral interaction. arXiv preprint arXiv:220212076
- Luo H, Zhai W, Zhang J, Cao Y, Tao D (2021a) Learning visual affordance grounding from demonstration videos. arXiv preprint arXiv:210805675
- Luo H, Zhai W, Zhang J, Cao Y, Tao D (2021b) One-shot affordance detection. arXiv preprint arXiv:210614747
- Luo H, Zhai W, Zhang J, Cao Y, Tao D (2022) Learning affordance grounding from exocentric images. arXiv preprint arXiv:220309905
- Lv Y, Zhang J, Dai Y, Li A, Barnes N, Fan DP (2022) Towards deeper understanding of camouflaged object detection. arXiv preprint arXiv:220511333
- Mai J, Yang M, Luo W (2020) Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 8766–8775
- Mandikal P, Grauman K (2021) Learning dexterous grasping with object-centric visual affordances. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 6169–6176
- Mi J, Tang S, Deng Z, Goerner M, Zhang J (2019) Object affordance based multimodal fusion for natural human-robot interaction. Cognitive Systems Research 54:128–137
- Mi J, Liang H, Katsakis N, Tang S, Li Q, Zhang C, Zhang J (2020) Intention-related natural language grounding via object affordance detection and intention semantic extraction. Frontiers in Neurorobotics p 26
- Myers A, Teo CL, Fermüller C, Aloimonos Y (2015) Affordance detection of tool parts from geometric features. In: 2015 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 1374–1381
- Nagarajan T, Grauman K (2020) Learning affordance landscapes for interaction exploration in 3d environments. Advances in Neural Information Processing Systems 33:2005–2015
- Nagarajan T, Feichtenhofer C, Grauman K (2019) Grounded human-object interaction hotspots from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 8688–8697
- Nguyen A, Kanoulas D, Caldwell DG, Tsagarakis NG (2016) Detecting object affordances with convolutional neural networks. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 2765–2770
- Nguyen A, Kanoulas D, Caldwell DG, Tsagarakis NG (2017) Object-based affordances detection with convolutional neural networks and dense conditional random fields. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 5908–5915
- Pan X, Gao Y, Lin Z, Tang F, Dong W, Yuan H, Huang F, Xu C (2021) Unveiling the potential of structure preserving for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 11642–11651
- Peters RJ, Iyer A, Itti L, Koch C (2005) Components of bottom-up gaze allocation in natural images. Vision research 45(18):2397–2416
- Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M (2022) Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:220406125
- Regmi K, Shah M (2019) Bridging the domain gap for ground-to-aerial image matching. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp 470–479
- Rizzolatti G, Craighero L (2004) The mirror-neuron system. Annu Rev Neurosci 27:169–192
- Sawatzky J, Gall J (2017) Adaptive binarization for weakly supervised affordance segmentation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 1383–1391
- Sawatzky J, Srikantha A, Gall J (2017) Weakly supervised affordance detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- Shen Y, Song K, Tan X, Li D, Lu W, Zhuang Y (2023) Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. arXiv preprint arXiv:230317580

- Sigurdsson GA, Gupta A, Schmid C, Farhadi A, Alahari K (2018) Actor and observer: Joint modeling of first and third-person videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 7396–7404
- Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556
- Soomro K, Zamir AR, Shah M (2012) Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:12120402
- Soran B, Farhadi A, Shapiro L (2014) Action recognition in the presence of one egocentric and multiple static cameras. In: Asian Conference on Computer Vision, Springer, pp 178–193
- Srikantha A, Gall J (2016) Weakly supervised learning of affordances. arXiv preprint arXiv:160502964
- Stark M, Lies P, Zillich M, Wyatt J, Schiele B (2008) Functional object class detection based on learned affordance cues. In: International conference on computer vision systems, Springer, pp 435–444
- Swain MJ, Ballard DH (1991) Color indexing. International Journal of Computer Vision (IJCV) 7(1):11–32
- Tang Y, Tian Y, Lu J, Feng J, Zhou J (2017) Action recognition in rgb-d egocentric videos. In: 2017 IEEE International Conference on Image Processing (ICIP), IEEE, pp 3410–3414
- Wu P, Zhai W, Cao Y (2021) Background activation suppression for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- Yang Y, Ni Z, Gao M, Zhang J, Tao D (2021) Collaborative pushing and grasping of tightly stacked objects via deep reinforcement learning. IEEE/CAA Journal of Automatica Sinica 9(1):135–145
- Zhai W, Luo H, Zhang J, Cao Y, Tao D (2022) One-shot object affordance detection in the wild. International Journal of Computer Vision (IJCV)
- Zhang J, Tao D (2020) Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. IEEE Internet of Things Journal 8(10):7789–7817
- Zhang Q, Xu Y, Zhang J, Tao D (2023) VitaeV2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. International Journal of Computer Vision (IJCV) pp 1–22
- Zhao X, Cao Y, Kang Y (2020) Object affordance detection with relationship-aware network. Neural Computing and Applications 32(18):14321–14333
- Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2921–2929
- Zhou K, Yang J, Loy CC, Liu Z (2022) Learning to prompt for vision-language models. International Journal of Computer Vision (IJCV) 130(9):2337–2348