Learning Affordance Grounding from Exocentric Images

Hongchen Luo\(^{1\dagger}\) Wei Zhai\(^{1\dagger} \) Jing Zhang\(^{2\dagger}\) Yang Cao\(^{1,4\dagger}\) Dacheng Tao\(^{3,2}\)

\(^{1}\) University of Science and Technology of China
\(^{2}\) The University of Sydney
\(^{3}\) JD Explore Academy
\(^{4}\) Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

\{lhc12, wzhai056\}@mail.ustc.edu.cn, jing.zhang@sydney.edu.au, forrest@ustc.edu.cn, dacheng.tao@gmail.com

Abstract

Affordance grounding, a task to ground (i.e., localize) action possibility region in objects, which faces the challenge of establishing an explicit link with object parts due to the diversity of interactive affordance. Human has the ability that transform the various exocentric interactions to invariant egocentric affordance so as to counter the impact of interactive diversity. To empower an agent with such ability, this paper proposes a task of affordance grounding from 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. To this end, we devise a cross-view knowledge transfer framework that extracts affordance-specific features from exocentric interactions and enhances the perception of affordance regions by preserving affordance correlation. Specifically, an Affordance Invariance Mining module is devised to extract specific clues by minimizing the intra-class differences originated from interaction habits in exocentric images. Besides, an Affordance Co-relation Preserving strategy is presented to perceive and localize affordance by aligning the co-relation matrix of predicted results between the two views. Particularly, 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 in terms of objective metrics and visual quality. Code: github.com/lhc1224/Cross-View-AG.

1. Introduction

The goal of affordance grounding is to locate the region of “action possibilities” of an object. For an intelligent agent, it is necessary to know not only what the object is but also to understand how it can be used [11]. 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 [1, 33, 34, 38]. It has a wide range of applications for robot grasping, scene understanding, action prediction [12, 13, 19, 23, 28, 31, 47, 49].

As affordance is a dynamic property closely related to the interaction between humans and environment [13], it is difficult to understand how to interact with objects and establish an explicit link between the objects’ intrinsic properties and affordances [29]. However, humans can easily perceive the object’s affordance region by observing exocentric human-object interactions, and give an egocentric definition. As shown in Fig. 1, although different persons hold the racket in different positions due to their individual habits, the human observer can perceive swingable regions determined by the intrinsic properties (e.g., the long handle...
Exocentric interactions can be decomposed into affordance-specific features $M$ and differences in individual habits $E$. (b) There are co-relations between affordances, e.g., “Cut with” inevitably accompanies “Hold” and is independent of the object category (knife and scissors). Such co-relation is common between objects. In this paper, we mainly consider extracting affordance-specific cues $M$ from diverse interactions while preserving the affordance co-relations to enhance the perceptual capability of the network.

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

To address this problem, we propose a cross-view knowledge transfer framework to extract affordance-specific features from exocentric interactions and transfer them to egocentric view. Specifically, we first devise an Affordance Invariance Mining (AIM) module to decompose the exocentric human-object interactions into the affordance representations determined by objects’ intrinsic properties and the differences originated from individual habits (as shown in Fig. 2 (a)). We use low rank matrix decomposition\cite{10,18,22,24} to minimize the intra-class differences caused by diverse interactions to obtain affordance-specific cues. Furthermore, there is a correlation between the object affordances (as shown in Fig. 2 (b)), which can be adopted to establish the link between different affordances to reduce the uncertainty caused by multiple affordances regions on the object. Therefore, we present a novel Affordance Correlation Preserving (ACP) strategy to perceive and localize the affordance region by aligning the co-relation matrix of prediction results from two views.

Despite the advances in affordance learning, the existing datasets\cite{8, 29, 32, 35, 43} still bear limitations in terms of affordance/object category, image quality, and scene complexity. To carry out a comprehensive study, this paper proposes an affordance grounding dataset named AGD20K, consisting of 20,061 exocentric images and 3,755 egocentric images from 36 affordance categories. The contrastive experiments against several representative methods are performed on the AGD20K dataset. The results demonstrate the superiority of our proposed method in capturing the intrinsic property of objects and suppressing the interactive diversity of affordance.

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

2. Related Works

2.1. Visual Affordance Grounding

The goal of affordance grounding is to locate the region of “action possibilities” of an object. Numerous works\cite{4, 7, 8, 20, 27, 35, 48, 50} mainly build upon super-
vised approaches to establish mapping relations between local regions of objects and affordance. Sawatzky et al. [42, 43] adopt an Expectation-Maximization algorithm [6] to achieve weakly supervised affordance detection using only a few key points. Nagarajan et al. [33] exploit only affordance labels to ground the interactions from the videos. In contrast to [33], 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 knowledge transfer 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

The emergence of the relevant datasets drives the development of affordance grounding, as shown in Table 1. For example, Sawatzky et al. [43] select video frames from CAD120 [19] to construct a weakly supervised affordance detection dataset, using only cropped out object regions but in inferior image quality. Other affordance-related datasets [4, 8, 32, 35, 41] face the problems of small scale and low affordance/object category diversity and do not consider human actions to reason about the affordance regions. PAD dataset [29] considers the inference of human purpose from support images of human-object interactions and transfers to a group of query images but does not provide part-level affordance labels. In contrast to the above works, we explicitly consider exocentric-to-egocentric viewpoint transformations and collect a much larger scale of images, with richer affordance/object categories and part-level annotations, which are more useful and applicable to real-world application domains.

2.3. Learning View Transformations

The existing learning-view transformation works start from the theory of mirror neurons [40], which adopts embedding learning to generate perspective invariant representations from paired data, and leverage it for tasks such as action recognition and video summarization under egocentric view [16, 39, 44, 45]. For example, Li et al. [25] extract key egocentric signals from the egocentric 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, which is challenging due to the uncertainty caused by various interactions and the multiple affordance regions that objects contain.

3. Method

Our goal is to ground the object affordance regions in egocentric images. During training, given a group of exocentric images. During training, we first use Resnet50 [14] to extract the features for affordance grounding is shown in Fig. 3. During training, we first use Resnet50 [14] to extract the features of exocentric and egocentric images to obtain $Z_{exo} = \{Z_1, ..., Z_N\}$ and $Z_{ego}$, respectively. We then present the Affordance Invariance Mining (AIM) module (see in Sec. 3.1) to extract affordance-specific clues ($F_{exo}$) from the exocentric features. Meanwhile, we use two convolutional layers to map the egocentric feature to the embedding space consistent with the exocentric view: $F_{ego} = Conv(Z_{ego})$. Subsequently, the features of the two branches ($F_{exo}$ and $F_{ego}$) are fed into the same convolution layer to obtain features $D_{exo}$ and $D_{ego}$ respectively. To ensure the affordance knowledge can be transferred to the egocentric view, we average the $D_{exo}$ through the global average pooling (GAP) layer to obtain the $f_{exo}$ and pass the $D_{ego}$ through the GAP layer to get the $f_{ego}$, and align $f_{exo}$ and $f_{ego}$ using L2 loss $L_{KT}$. Then, $f_{exo}$ and $f_{ego}$ are fed into the same fully connected layer to obtain the affordance prediction. Finally, we propose an Affordance Co-relation Preserving (ACP) strategy (see in Sec. 3.2) to enhance the network’s perception of affordance by aligning the co-relation matrix of the outputs of the two views. During testing, we feed the egocentric object images into the network only through the egocentric branch, and then use the CAM [51] technique to obtain the affordance regions of the object (see in Sec. 3.3).

3.1. Affordance Invariance Mining Module

As shown in Fig. 3, we decompose the interactions in exocentric images into affordance-specific features $M$ and individual differences $E$. Inspired by low-rank matrix decomposition [10, 18, 22], we represent the $M$ as the multiplication of a dictionary matrix $W$ and a corresponding matrix $H$, where the dictionary bases represent the sub-features of human-object interaction, and minimize $E$ by iterative op-
Figure 3. Overview of the proposed cross-view knowledge transfer affordance grounding framework. It mainly consists of an Affordance Invariance Mining (AIM) module and an Affordance Co-relation Preservation (ACP) strategy. The AIM module (see in Sec. 3.1) aims to obtain invariant affordance representations from diverse exocentric interactions. The ACP strategy (see in Sec. 3.2) enhances the network’s affordance perception by aligning the co-relation of the outputs of the two views.

After several iterations, we get the output \( M = WH \), and reshape it to \( M_{exo} = \{ M_1, ..., M_N \}, M_i \in R^{c \times h \times w} \). Finally, we use a convolution layer to map it to the residual space and sum it with the \( Z \) to get the final output \( F_{exo} \):

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

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

\[
W^{(0)} \leftarrow \alpha W^{(0)} + (1 - \alpha) \bar{W},
\]

where \( \bar{W} \) is the average over each mini-batch.

### 3.2. Affordance Co-relation Preserving Strategy

As shown in Fig. 3, we feed the feature representations of the two branches \( f_{exo} \) and \( f_{ego} \) into the same fully connected layer respectively to obtain the affordance category prediction scores \( s \) and \( q \):

\[
s = FC(f_{exo}), \quad q = FC(f_{ego}).
\]
Train | Carry | Sit on | Drink with | Skis | Snowboard | Surfboard | Bench | Couch | Skateboard | Bottle | Wine glass | Cup

Figure 4. The properties of the AGD20K dataset. (a) Some examples from the dataset. (b) The distribution of categories in AGD20K. (c) The word cloud distribution of affordances in AGD20K. (d) 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.

Then, we align the affordance co-relation between the exocentric and egocentric views by calculating the cross-entropy loss

\[ L_{ACP} \] of the co-relation matrix of the prediction scores of the two branches:

\[
p_j = \frac{\exp(s_j/T)}{\sum_k^N \exp(s_k/T)} , \quad q_j = \frac{\exp(g_j/T)}{\sum_k^N \exp(g_k/T)} , \quad (6)
\]

\[
P = pp^T , \quad Q = qq^T , \quad (7)
\]

\[
L_{ACP} = - \sum_j^N \sum_k^N p_{jk} \log(Q_{jk}) . \quad (8)
\]

where \( T \) is used to control the degree of attention paid to the correlations between negative labels. \( P_{jk} \) and \( Q_{jk} \) denote the correlation between classes \( j \) and \( k \) in the prediction results. Finally, the total loss can be calculated as:

\[
L = \lambda_1 L_{cls} + \lambda_2 L_{ACP} + \lambda_3 L_{KT} , \quad (9)
\]

where \( \lambda_1 , \lambda_2 \) and \( \lambda_3 \) are hyper-parameters to balance the classification loss, ACP loss and \( L_{KT} \) loss. \( L_{cls} \) is the sum of the cross-entropy losses of the classification results of the two branches, and \( L_{KT} \) is loss of cross-view affordance knowledge transfer: \( L_{KT} = \| f_{exo} - f_{ego} \| . \)

3.3. Inference

Our test procedure only requires an egocentric object image and an affordance label as input to predict the affordance region. We utilize the class activation mapping [51] by computing a weighted sum of the feature maps \( D^i \) of the last convolutional layer to obtain the affordance region heatmap:

\[
Y^{Ca} = \sum_i w_i^{Ca} D^i , \quad (9)
\]

where \( C_a \) is the affordance class, \( D^i \) is the \( i \)-th layer feature map, and \( w_i^{Ca} \) is the weight corresponding to the \( i \)-th neuron under the \( C_a \) category.

4. Dataset

Dataset Collection. The exocentric images are mainly obtained from HICO [3] and COCO [26]. We select images from the HICO dataset according to the verb category and the COCO dataset according to the object category. Then, we manually remove images with ambiguous interactions. To enrich the diversity of the dataset, we further collect 2, 112 exocentric images from free-license websites. Meanwhile, We collect 3, 755 egocentric images from the Internet with free use license according to object categories. Some examples are shown in Fig. 4 (a).
The results of different methods on AGD20k. 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 \( \diamond \) defines the relative improvement of our method over other methods. “Dark red”, “Orange” and “Purple” represent saliency detection, weakly supervised object localization and affordance grounding models, respectively.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Seen</th>
<th>Unseen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KLD ↓</td>
<td>SIM ↑</td>
</tr>
<tr>
<td>Minet [5]</td>
<td>5.197 ± 7.4%</td>
<td>0.280 ± 19.3%</td>
</tr>
<tr>
<td>DeepGazeII [21]</td>
<td>1.858 ± 71.2%</td>
<td>0.280 ± 19.3%</td>
</tr>
<tr>
<td>EgoGaze [17]</td>
<td>4.185 ± 66.2%</td>
<td>0.227 ± 47.1%</td>
</tr>
<tr>
<td>EIL [30]</td>
<td>1.931 ± 20.4%</td>
<td>0.285 ± 17.2%</td>
</tr>
<tr>
<td>SPA [36]</td>
<td>5.528 ± 72.2%</td>
<td>0.221 ± 51.1%</td>
</tr>
<tr>
<td>TS-CAM [9]</td>
<td>1.842 ± 16.5%</td>
<td>0.260 ± 28.5%</td>
</tr>
<tr>
<td>Hotspots [33]</td>
<td>1.773 ± 14.3%</td>
<td>0.278 ± 20.4%</td>
</tr>
<tr>
<td>Ours</td>
<td>1.538 ± 0.017</td>
<td>0.334 ± 0.001</td>
</tr>
</tbody>
</table>

Figure 5. Visual affordance heatmaps on the AGD20K dataset. We select the prediction results of representative methods of affordance grounding (Hotspots [33]), weakly supervised object localization (EIL [30]), and saliency detection (DeepGazeII [21]) for presentation.

Dataset Annotation. We select 36 affordance classes commonly used in real-world application scenarios and assign labels to each image based on the interaction between human and object in each exocentric image. Given the object class contained in each affordance class, we assign affordance labels based on the object class in the egocentric images. The testing process requires pixel-level labels to calculate objective metrics. As the annotation approach in [8], we take the form of points for regions of interaction, in which the dense points are for regions of frequent interaction and vice versa. Then, heatmaps of affordance regions can be obtained from the points as [8]. Some annotation examples are shown in Fig. 4 (a).

Statistic Analysis. To obtain deeper insights into our AGD20K dataset, we show its important features from the following aspects. The distribution of categories in the dataset is shown in Fig. 4 (b), which shows that the dataset contains a wide range of affordance/object categories in diverse scenarios. The affordance word cloud is shown in Fig. 4 (c). The confusion matrix of affordance and object categories is shown in Fig. 4 (d). It shows a multi-to-multi relationship between affordance and object categories, posing a significant challenge for the affordance grounding task. See supplementary materials for more details.

5. Experiments
5.1. Benchmark Setting

To provide a comprehensive evaluation, we choose three commonly used metrics Kullback-Leibler Divergence (KLD) [2]. SIMilarity (SIM) [46] and Normalized Scanpath Saliency (NSS) [37]. Our model is implemented
The experimental results are shown in Table 2. Our method achieves the best results in both “Seen” and “Unseen” settings. Taking KLD as the metric, our method improves 17.2% compared to the best saliency model, 16.5% over the best weakly supervised object localization (WSOL) model, and 13.3% over the affordance grounding model in the “Seen” setting. Our method with the “Unseen” setting improves 10.2% compared to the best saliency model, surpasses the best WSOL model by 15.1%, and exceeds the affordance grounding model by 10.4%. It indicates that our method can effectively transfer the affordance knowledge from the exocentric view to the object in the egocentric view and has a good generalization ability for unseen objects.

In addition, we visualize the affordance maps in “Seen” and “Unseen” settings, as shown in Fig. 5. It shows that our method can obtain more accurate prediction results for affordance grounding. While “Sit on” contains objects with different appearances (“bed” and “chair”), our method can capture the common features of the affordance region and obtain better prediction results. Since “bottle” has two different affordances, “Open” and “Hold”, the network predicts different affordance regions. In the “Unseen” setting, the “knife” has two different affordances, “Hold” and “Cut
with”. Our method can locate different affordance regions based on the learned affordance knowledge, demonstrating its superior generalization capability.

5.3. Ablation Study

The ablation study results are shown in Table 3. It confirms the ability of the AIM module to learn affordance-specific features from diverse exocentric interactions, which play a significant role in improving network performance. The ACP strategy improves more obviously than $L_{KT}$, indicating that the preservation of affordance co-relation can more effectively improve the network’s ability to perceive and locate affordance regions. In addition, we investigate the influence of different hyper-parameter settings of $T$ in the ACP strategy (shown in Fig. 6 (left)), the rank $r$ of the dictionary matrix $W$ in the AIM module (shown in Fig. 6 (middle)), and different exocentric images $N$ (see Fig. 6 (right)). It can be seen that the performance of the model is more influenced by $T$, while the rank $r$ does not have a significant impact on the results. The number of exocentric images taken from 2 to 3 has a larger impact on the model. For $N = 1$, our model still outperforms most contenders.

5.4. Performance Analysis

Different Classes. The KLD metrics on some representative categories are shown in Table 4. “Hold” and “Swing” both contain diverse object categories with different appearances. “Drink with” and “Hold” contain overlapped object categories but have completely different affordance regions. Objects of “Lie on” are generally labeled with a larger region, while those of “Brush with” are generally smaller. Our model exceeds others regarding different aspects of the challenge, which confirms its robustness.

Different Scales. We divide the test set into “Big”, “Middle” and “Small” according to the ratio of mask to the whole image, and show the results of the NSS metrics.

Figure 7. Different scales. We split the test set into three subsets of “Big”, “Middle” and “Small” according to the ratio of mask to the whole image, and show the results of the NSS metrics.

Table 5. Different sources. “Exo” means that training only uses exocentric images, “Exo&Ego” means that training uses both exocentric and egocentric images.

<table>
<thead>
<tr>
<th>Method</th>
<th>Source</th>
<th>KLD ↓</th>
<th>SIM ↑</th>
<th>NSS ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIL [30]</td>
<td>Exo</td>
<td>1.931</td>
<td>0.285</td>
<td>0.522</td>
</tr>
<tr>
<td></td>
<td>Exo&amp;Ego</td>
<td>2.156</td>
<td>0.321</td>
<td>0.747</td>
</tr>
<tr>
<td>SPA [36]</td>
<td>Exo</td>
<td>5.528</td>
<td>0.221</td>
<td>0.357</td>
</tr>
<tr>
<td></td>
<td>Exo&amp;Ego</td>
<td>4.312</td>
<td>0.252</td>
<td>0.494</td>
</tr>
<tr>
<td>TS-CAM [9]</td>
<td>Exo</td>
<td>1.842</td>
<td>0.260</td>
<td>0.336</td>
</tr>
<tr>
<td></td>
<td>Exo&amp;Ego</td>
<td>1.707</td>
<td>0.290</td>
<td>0.622</td>
</tr>
<tr>
<td>Ours</td>
<td>Exo&amp;Ego</td>
<td>1.538</td>
<td>0.334</td>
<td>0.927</td>
</tr>
</tbody>
</table>

Different Sources. The results for different sources are shown in Table 5. It shows that using both exocentric and egocentric images improves most methods, but the improvement is limited. Our method still surpasses all models, showing that the knowledge transfer from explicitly cross-views is effective in learning from exocentric diverse interactions to egocentric invariant affordance representation.

Limitations. Our method still has limitations, e.g., the predicted affordance maps may contain intermediate background regions when multiple objects appear and irrelevant background regions may be activated for slender object. In the future, we will refer to [36] to refine the generated results to obtain more accurate results.

6. Conclusion

In this paper, we make an attempt to address a new challenging task named affordance grounding from exocentric view. Specifically, we propose a novel cross-view knowledge transfer framework that can extract invariant affordance from diverse exocentric interactions and transfer it to egocentric view. We establish a large affordance grounding dataset named AGD20K, which contains 20K well-annotated images, serving as a pioneer testbed for the task. Our model outperforms representative models from related areas and can serve as a strong baseline for future research.

Broader Impacts. The research on affordance grounding from exocentric view will advance the realization of embodied intelligence. However, harmful human demonstrations (risky behaviors) may lead to negative guidance for the agent, which should be prohibited by strict legislation.

Acknowledgments. This work was supported by National Key R&D Program of China under Grant 2020AAA0105701, National Natural Science Foundation of China (NSFC) under Grants 61872327 and ARC FL-170100117.
References


[32] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning visual affordance grounding from


