Phrase-Based Affordance Detection via Cyclic Bilateral Interaction

Liangsheng Lu, Wei Zhai, Hongchen Luo, Yu Kang, Senior Member, IEEE, and Yang Cao, Member, IEEE

Abstract—Affordance detection, which refers to perceiving objects with potential action possibilities in images, is a challenging task since the possible affordance depends on the person's purpose in real-world application scenarios. The existing works mainly extract the inherent human-object dependencies from image/video to accommodate affordance properties that change dynamically. In this article, we explore to perceive affordances from a vision-language perspective, and consider the challenging phrase-based affordance detection task, i.e., given a set of phrases describing the potential actions, all the object regions in a scene with the same affordance should be detected. To this end, we propose a cyclic bilateral consistency enhancement network (CBCE-Net) to align language and vision features in a progressive manner. Specifically, the presented CBCE-Net consists of a mutual guided vision-language module that updates the common features of vision and language in a progressive manner, and a cyclic interaction module that facilitates the perception of possible interaction with objects in a cyclic manner. In addition, we extend the public purpose-driven affordance dataset (PAD) by annotating affordance categories with short phrases. The extensive contrastive experimental results demonstrate the superior performance of our method over nine typical methods from four relevant fields in terms of both objective metrics and visual quality.

Impact Statement—Affordance learning is concerned with the possible set of actions that an environment can offer to an actor. At present, almost all work focus on utilizing cues of human-object interactions in visual media such as images and videos to learn affordances. Few work explores affordance using language information, which is crucial to building more intelligent agents because we live in a multimodal world. In this paper, we propose a phrase-based affordance detection task combining vision and language modalities. We first build a vision-language dataset based on the public Purpose-driven Affordance Dataset (PAD). Then we propose a method to fuse and align multi-modal features. Our method outperforms relevant methods from several other fields in terms of both objective metrics and visual quality. The proposed dataset and method can facilitate building more intelligent agents that can better comprehend the surrounding environment. For instance, the phase-based affordance detection can be used in building more intelligent home service systems. The agents can understand instructions of instructions without specific entities.

Index Terms—Affordance, deep learning, segmentation, vision-language, visual affordance detection.

I. INTRODUCTION

The term “Affordance” is used to describe the interactions between humans, animals, and their environment. In other words, affordance implies the complementary between the animal and the environment [1]. The comprehension of affordance is essential for intelligent agents to move from passive perceptual systems, such as those trained for image recognition to active ones, which embody agents capable of sensing and interacting with their environments, from cooking in the kitchen to grasping objects [2]. Thus, investigating the affordance of objects leads agents to interact with environments better. In the field of computer vision, techniques not only need abilities to understand the content in a scene but also require the capability to infer possible interactions between humans, animals, and the corresponding environments [3]. Recently, affordance has drawn remarkable attention and has been widely explored in various application fields. For instance, the theory of affordance is applied to design more intelligent and more robust robotic systems in complex and dynamic environments [4]. As a result, perceiving the affordance of objects has a broad range of applications in several fields, such as action recognition [5], [6], [7], scene parsing [8], [9], [10], and robot grasping [11], [12], etc.

Previous work mainly addressed applications of affordance in the vision-only fields. Some works [13], [14], [15] construct mapping relationships between objects representations and affordance categories. However, affordances are closely associated with the environment and actors, limiting the model's capacity to generalize in new unseen scenarios, which results in incorrect perception and localization. To solve the above problem, some other work perceives affordance objects by mining human–object interaction cues from videos or images [15], [16], [17], [18] and transferring them to target images which enables models to better cope with the effects of dynamic changes of affordance and keep well generalization ability in new unseen environments. Different from the above work, this

2691-4581 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
article attempts to explore the potential of language pairs in affordance detection tasks from a vision-language perspective. We consider utilizing a series of phrases combinations to describe affordances of an object and then generate segmentation masks from the corresponding images. This process is also consistent with application scenarios where real intelligent agents receive information in multiple modalities from different sources to jointly perceive the affordances of objects.

Therefore, we propose a phrase-based affordance detection task in this article. That is, input a set of textual phrases describing affordances and an image, then the objects that can afford the corresponding affordance are expected to be segmented precisely (as shown in Fig. 1). Choosing natural language phrases to describe affordances without considering the specific object categories is suitable in practical application scenarios. In practice, humans often communicate with each other with incomplete sentences and pass out some common-sense information that does not need to be explicitly pointed out. Similarly, when a human interacts with an agent, the given instructions are likely incomplete [19]. For example, a human asks the agent to “pour me some water” while whether to use a cup or a bowl to pour the water is not indicated explicitly. This shows that learning common-sense knowledge, such as affordance will lead to more intelligent agents.

Nevertheless, affordance is a special property different from the semantic category. One object may have multiple affordances, while one affordance may associate with multiple objects. For example, the affordance of “Bed” includes two different actions: “Sit” and “Lie,” while “Chair” can also afford action “Sit.” This may lead to great differences in the visual representation of different objects referred to by the same textual descriptions. Fig. 2(a) shows the differences between traditional vision-language (V-L) tasks and affordance-related V-L tasks. Such variations in the color, texture, and shape of objects would render affordance-related V-L task more difficult in the alignment of textual and visual features compared to conventional multimodal tasks [20], [21], [22], [23], [24]. The differences can lead to significant divergence in the distribution of textual and visual representations in feature space. It may be difficult for the deep-learning-based network to align features from these two modalities through a single learning step. Because the distribution of associated visual features seems to be irregular for the same textual representation, the cross-modal affordance consistency is difficult to capture if the network only updates features once a time. To tackle this problem, as shown in Fig. 2(b), we design a cyclic bilateral update mechanism. Our model updates visual and linguistic features with the leverage of another modality to enhance the intermodal affordance consistency progressively in a bilateral and cyclic manner. The intermodal affordance consistency is gradually enhanced after several cyclic alignments.

To this end, we propose a cyclic bilateral consistency enhancement network (CBCE-Net), which consists of three main modules: 1) vision guide language module (VLM), 2) language guide vision module (LVM), and 3) cyclic interaction module (CIM). Utilizing the attention mechanism, VLM learns the importance of linguistic features in each visual region and derives new linguistic features. Then LVM uses the output of the textual feature from VLM with aggregating multilevel information to guide the generation of new visual features. VLM and LVM operations are repeated several times in CIM module to enhance the intermodal semantic consistency in a cyclic and bilateral manner.

The current affordance datasets lack explicit descriptions of affordance using natural language. To address this issue, based on the previously proposed Purpose-driven Affordance Dataset (PAD) dataset [15], we annotate associated short phrases according to affordance categories, as this is more suitable for practical application scenarios. With the leverage of WordNet [25], which is a hierarchical lexical database, we annotate the text of affordance from four different but closely related perspectives: 1) potential actions, 2) functions, 3) appearance features, and 4) the environment. We name the resulting PAD-based dataset with annotated natural language as PAD-Language dataset (PAD-L).

In summary, our contributions are fourfold as follows.

1) We propose a new challenging phrase-based affordance detection task. Inputting a set of text phrases and an image containing related objects; the corresponding segmentation masks are expected to be generated. This makes more intelligent agents better comprehend humans’ intentions during interactions with humans and locate specific objects in the scene even if the instructions do not indicate the specific category of the object.

2) We design a novel CBCE-Net to effectively extract the affordance information from the given set of text phrases and then segment the corresponding object regions in the given image. Our model can effectively solve the vision-language alignment issue caused by the multiplicity property of affordance. The text and vision information could interact and align well with each other in a cyclic and bilateral manner even though the visual appearance, texture, and color are highly diverse.

3) We annotate affordance categories using natural language phrases based on the existing dataset PAD. A new affordance dataset with natural language descriptions named PAD-L is constructed, which extends objects’ affordances from limited categories to unconstrained natural language. The new dataset can be used in various downstream tasks.
Fig. 2. Task and method differences between affordance-related vision-language task and traditional ones. (a) shows the problems caused by the multiplicity property of affordance. In traditional vision-language (V-L) tasks, the appearances of objects with the same language descriptions are generally similar, while the differences are significant in affordance-related V-L task. For images on the left, objects referred to by the same entity phrases are similar in color, shape, and texture; nevertheless, the opposite is true for images on the right. In (b), We compare our method with conventional methods. In traditional methods, vision features are close enough in distance in feature space, leading to easier alignments. However, for affordance-related V-L task, vision and language features are cluttered in feature space. We design a cyclic and bilateral mechanism to cope with these problems by enhancing iter-modal semantic consistency in a progressive manner (See Section IV for details).

4) Compared with nine different approaches chosen from four relevant fields (salient detection, affordance detection, semantic segmentation, and referring segmentation), our model achieves the best result in both subjective and objective terms, which is able to serve as a strong baseline for future work.

The rest of the article is organized as follows. Section II illustrates the previous work related to phrase based affordance detection task. Section III describes details to annotate text phrases. The novel proposed CBCE-Net is introduced in Section IV. Section V shows the experiment results and analysis on PAD-L. Finally, Section VI concludes this article.

II. RELATED WORK

A. Affordance Learning

Visual affordance has been extensively studied in computer vision and robotics communities because of its close association with action recognition, scene parsing, and human–robot interaction. Many approaches have been proposed to perceive the visual affordance of objects in scenes. Hassan et al. [26] proposed a Bayesian network-based affordance detection method that exploits the attribute of the object, actor, and environment. Grabner et al. [27] utilized a human skeleton 3-D model to learn the action of sitting on a chair to infer whether an object can afford “sitting” action or not.

With the development of deep neural networks, many deep learning-based methods have been proposed. Inspired by semantic segmentation [28] approaches, affordance learning is extensively studied at the pixel level. Sawatzky et al. [13] proposed a weakly supervised affordance segmentation method to predict the fine segmentation masks by effectively leveraging the weakly labeled data, which is annotated in image-level and keypoints level. Nguyen et al. [29] considered affordance segmentation as an object detection task. They employ the existing object detection methods to obtain a set of candidate bounding box proposals. Afterward, atrous convolutions is used to generate the final fine masks. Zhao et al. [14] proposed an end-to-end model to exploit symbiotic relationship between multiple affordances with the combinational relationship between affordance and objectness to produce affordance maps.

In addition to using the features of objects themselves, some recent work has leveraged auxiliary information to learn visual affordance [30], [31]. Fang et al. [9] proposed a method to learn the affordance of unseen objects from expert demonstration videos. Their model extracts feature embedding vectors from demonstration videos to predict the same objects’ interaction regions and action labels in a given image. Luo et al. [15] proposed an one-shot detection method to detect affordance in unseen scenarios. Their model first extracts intention information from support images. Then, the intention is transferred to query images to detect objects capable of affording the intention. To this end, they construct a new PAD, which compensates for the lack of rich scenes in the previous datasets.

Unlike all the work mentioned above, where affordance is explored only in visual mediums, we attempt to investigate affordance detection involving natural language. Inputting a set of phrases that describe affordances and an image, the corresponding objects are expected to be segmented, which meets the
realistic scenarios where robots receive information in multiple modalities from multiple sources.

**B. Referring Expression Grounding**

Giving a piece of text, referring expression grounding task is aimed at comprehending the natural language content with locating the corresponding regions in the input image. Many efforts achieve localization at bounding box level [32], [33], [34]. Yu et al. [35] proposed a modular network which decomposes the input natural language description into subject, location, and relationship attributes to improve the localization performance. Liu et al. [36] adopted graph models with an attention mechanism to capture the relationship between the object regions in the given image. In association with visual affordance, Mi et al. [37], [38] investigated the use of natural language to guide visual affordance detection. Their model first extracts the intention in the natural language then locates referred objects in the given image at the bounding box level.

Many approaches have been proposed at pixel level. In [20], [22], [39], the multimodal features from CNN and LSTM [40] are directly concatenated together and input into the fully convolutional network to generate the final pixelwise masks. These methods do not exploit the intramodal and intermodal relationships explicitly. More recent work uses self-attention and cross-attention mechanisms for linguistic and visual information. Ye et al. [41] proposed a cross-modal self-attention module that can adaptively focus on essential words in referring expressions and important regions in images to capture the long-range dependencies between linguistic and visual features. They Hu et al. [23] designed a bidirectional relationship inferring network to model the relationship between linguistic and visual features. Liu et al. [36] proposed a model that first perceives all the entities in the image according to the entity and attribution words in expressions, then infers the location of the target object with the words that represent the relationship. Jing et al. [24] first gets the position prior of referred objects using the language and image, then generates segmentation masks on the basis of the obtained results.

Compared to the mentioned referring segmentation task, our proposed one has some significant differences. At first, the inherent multiplicity feature of affordance leads to a much more significant variation in the visual representation of objects referred to by the same text than in traditional ones, which masks the alignment of linguistic and visual features more difficult. Second, our phrases only describe affordance without presenting entity words. Therefore, it is unable to utilize relationships between entities as in [24], [36], and [41] to localize objects. We employed short phrases to express affordances rather than long sentences to meet practical scenarios. This disables us from leveraging the text context information to capture the relationship between linguistic and visual features like the operations in [21], [23], [41], and [36]. The work mentioned above also utilized natural language to learn visual affordances [37], [38]. Our work differs from them in the following ways. First, we consider the inherent multiplicity problem of affordance and involve richer indoor and outdoor scenes, which meet affordance’s definition and are suitable for practical applications. Second, our model can generate a more precise pixel-level segmentation mask instead of a bounding box, limiting the ability to capture the inherent shape of objects. The accurate shape offers downstream tasks, such as “robot grasping” richer geometric features to facilitate potential actions. In addition, technically, unlike their two-stage strategy, which relies heavily on the accuracy of intention extraction from natural language at first, we propose an end-to-end framework to enable multimodal information to interact with each other adequately in a cyclic and bilateral manner.

### III. Language Annotations

This section shows the process to get our language annotations based on the PAD dataset. The whole complete PAD dataset could be found at https://github.com/lhc1224/OSAD_Net. We describe details of the annotation process where we consider affordances from four perspectives with the assistance of WordNet, which is a hierarchical lexical database and could be explored at https://wordnet.princeton.edu/. After that, we show statistics, including overall perspective, image to phrase number according to affordance categories, and the word cloud of annotations of the proposed PAD-language dataset.

Instead of describing affordances with grammatically coherent sentences, we find it more effective to use several phrases closer to the actual application scenarios to describe the affordance. People tend to use short instructions to computers or robots in daily life rather than long sentences. Moreover, Short phrases are more representative than complicated sentences to describe affordances because the annotations are relatively subjective, i.e., the descriptions just involve actions and do not contain any objective actors and objects. So, compared to long sentences with a rigid grammatical structure, such as “subject + verb + object,” short phrases are more flexible and extensible. Typically, the number of words in short phrases is from 1 to 5, and average of PAD-L is 2.36 (See Fig. 4).

As the inherent property of an object, the term “affordance” is related to a set of possible actions that are able to manipulate objects. Therefore, the linguistic descriptions of affordance must be tightly relevant to these potential actions or the functionalities for human use. In addition, the environment in which objects are located and their appearance features may also affect the affordances of objects. Most of the affordances associated with our daily life are related to these aspects. Therefore, for better descriptions, we consider short phrases from several different perspectives: 1) the actions that can be potentially performed on the object. 2) The function of the object. 3) The appearance features related to the actions or functionalities. 4) The environment that has capabilities to afford possible interactions between actors and objects.

1) **Potential Actions:** Different from other properties of objects, affordance holds one noteworthy distinction. That is, an object may have several different affordances, while different objects may have identical affordances. It is difficult to focus on this issue while describing affordance using natural language. To this end, we make the phrase descriptions based on the
affordance categories rather than object categories in the PAD dataset. To explore more expressions for actions, we utilize a widely used large lexical tool WordNet [25] to assist the annotation process. WordNet is a hierarchical lexical database that groups verbs, nouns, adjectives, and adverbs into sets of cognitive synonyms and *synsets* associated by semantic relations. Other words or phrases expressing a similar sense could be found easily in WordNet for a specific word. To describe affordances, actions can be generally indicated by a *synset* rather than individual verbs. The semantic relationship between words in WordNet can be represented by a tree diagram. Two typical keyword centered tree diagram examples are shown in Fig. 3.

For a specific action, the phrases on nodes in the tree diagram can greatly enrich the descriptions for affordances. As shown in Fig. 3, it is reasonable to consider that “soccer ball” and “punching bag” have the affordance “deliver a blow by foot,” “drive or propel with foot,” “strike out,” etc. and affordance *Throw* could be extended to phrases, such as *propel through the air*, *deliver*, or *pass*, etc., rather than a single word. It is worth noting that we adopt multiple verb tenses instead of only in original form to exhibit more diverse application scenarios and get more natural language phrases.

2) Function: Object function is an intrinsic property of an object independent of the users. In addition, functionality understanding plays a crucial role in human–machine interactions. Sensing function of an object is essential to building a more intelligent computer vision system. Therefore, functions of an object are also included in our phrase annotations. For instance, the object “Umbrella” has the function of “sheltering from the wind and rain,” object “Knife” has the function of “cut,” object “Drum” has the function of “make sound.” We annotate functions of objects in PAD dataset using simple phrases instead of involving in more details.

3) Appearance Feature: Visual appearance and geometric characteristics can be regarded as the physical basis of affordances. For instance, the middle of a *cup* is depressed resulting in its ability to “hold water.” *Soccer ball* is “spherical” in appearance causing it to have the ability to “roll.” Besides, in practice, one may not know the specific category of an object but one can infer its affordances by its appearance features.
4) Environment: In the most accepted affordance definitions, the environment plays an important role. The term “affordance” is thought to reveal the complementary nature of the animal and the environment [1]. In our textual annotations, we incorporate descriptions of the environment. Sometimes, specific affordances are available only when the object is located in a particular environment. For instance, a “soccer ball” generally only exhibits the affordance “play” in an outdoor environment and “chopsticks” are usually found at the dining table or the kitchen. With considering environments, the description of affordance becomes more complete. More examples can be found in the Appendix.

Fig. 4 shows some overall statistics of the proposed PAD-L dataset, which is constructed based on the previous PAD dataset containing 4002 images from 31 affordance categories and 72 object categories. We split these images into 75% training and 25% test. For a single image in the PAD dataset, we randomly select a set of phonetic numbers (the number is 4 in this article) from the candidate annotations to build a new extended version with text information considered. The statistic shows that PAD-L contains rich phrases in a variety of scenarios.

IV. METHOD

A. Problem Description

Taking a set of affordance descriptions \( P = \{P_1, P_2, \ldots, P_n\} \) that describes affordance \( A_m \) and an affordance-related image \( I \), which contains multiple objects \( \{O_1, O_2, \ldots, O_n\} \), the phrase-based affordance detection task aims to get the segmentation masks \( M \) of the object \( O_m \), which can afford the affordance \( A_m \) in the image. We define the input as \( n \) query phrases and an image \( I \) in each batch.

We need encoders to encode visual and linguistic features, respectively. Afterward, information from the two distinct modalities is aligned and fused. Then, the result features are input to a module to learn the consistency between the two modalities. After adequate alignment and fusion, we use a segmentation module to generate the final masks.

B. Visual and Linguistic Features Extraction

As shown in Fig. 5(a), our model takes a set of phrases and an affordance-related image as inputs. In the image branch, a CNN backbone (e.g., ResNet101 [42]) is used to extract multilevel image features. The output of the third, fourth, and fifth stages of CNN backbone are denoted as \( \{I_1, I_4, I_5\} \) with channel dimension of 512, 1024, and 2048, respectively. Afterward, \( 1 \times 1 \) convolution layers are employed to transform the multilevel visual features to the same size of \( \mathbb{R}^{H \times W \times C_t} \).

In the textual branch, the language features \( L = \{L_1, L_2, \ldots, L_n\} \) is extracted using a language encoder (e.g., LSTM [40]), where \( n \) is the number of phrases. The parameters of embedding layer is initialized using GloVe word embeddings [43]. After encoding by the language encoder, the resulting linguistic features are applied in a max pooling operation and get a global language representation \( L_0 \in \mathbb{R}^{C_t} \), where \( C_t \) is the language feature dimension interacts with the multilevel visual features and feed into the proposed CIM.

Afterward, a bilinear fusion [44] is adopted to fuse different level visual features with linguistic feature \( L_0 \), in order to incorporate more spatial information, we concatenate a 8-D spatial coordinate feature which is denoted as \( P \in \mathbb{R}^{H \times W \times 8} \) with the resulting fused multimodel feature before to get final fused features \( \{F_0^i, F_0^j, F_0^5\} \in \mathbb{R}^{H \times W \times (C_t + C_i + 8)} \), which could be defined as follows:

\[
\{F_0^i = \text{concat}(f(I_i, L_0), P)\}_{i=3,4,5}
\]

where \( \text{concat}(\cdot, \cdot) \) represents the concatenation operation along the channel dimension and \( f \) denotes bilinear fusion operation. In this article, \( H = 40, W = 40 \).

C. Cyclic Interaction Module

Compared to traditional V-L tasks, the apparent difference of objects referred to by the same language descriptions in different images could be significant because of the multiplicity property of affordance.

The vast differences in visual appearance lead to substantial divergence in the distribution of textual and visual features in feature space. To generate accurate and consistent representations of the target object and the given affordance description phrases, the feature representation for one modality is enhanced several times adaptively guided by the other modality in CIM, which is indicated in Fig. 5(a). CIM consists of bilateral interaction operations between the two modalities to learn the consistency step by step, which leads to adequate fusion and alignment.

Specifically, we propose a VLM to enhance the linguistic feature representation with the guidance of visual features and an LVM to get improved visual feature representations. The cyclic interaction process is illustrated as follows (\( i = 3, 4, 5 \) and \( j, k \in \{3, 4, 5\} \setminus \{i\} \) in the equations):

\[
L_1^i = \text{VLM}(L_0, F_0^i)
\]

\[
F_j^i = \text{LVM}(L_1^i, F_0^j, F_0^k)
\]

\[
L_2^i = \text{VLM}(L_1^i, F_1^i)
\]

\[
F_2^i = \text{LVM}(L_2^i, F_1^i, F_1^k).
\]

D. Vision Guide Language Module

The architecture of the proposed VLM is illustrated in Fig. 5(b). To update the language feature representation \( L_{m+1}^i \), we leverage the previous fused feature \( F_0^i \) to guide the transformation of the previous language feature \( L_m^i \) in VLM.

For a language feature \( L_m^i \in \mathbb{R}^{C_t} \) and fused visual feature \( F_m^i \in \mathbb{R}^{H \times W \times C_v} \), we can compute the elementwise correlations using inner product

\[
S_m^i = \phi(F_m^i)\theta(L_m^i)^T
\]

where \( \theta \) and \( \phi \) are \( 1 \times 1 \) convolution layers to transform the feature to have the same dimension where \( \theta(L_m^i) \in \mathbb{R}^{1 \times C} \), \( \phi(F_m^i) \in \mathbb{R}^{H \times W \times C} \). The affinity map \( S_m^i \in \mathbb{R}^{H \times W \times 1} \) captures the correlation information of the given features. Then we employ scale and softmax operations following the scaled dot-product attention practice in [46] to normalize and reshape the
Fig. 5. Architecture of our proposed CBCE-Net. CBCE-Net first uses DeepLab Resnet101 [45] and LSTM [40] to extract multilevel visual and linguistic features, respectively. Subsequently, combining spatial coordinate, multilevel multimodal features are generated through bilinear fusion operations (see Section IV-B for details). Afterwards, fused features are fed into CIM to enhance the semantic consistency in a cyclic and bilateral manner (see Section IV-C for details). In CIM, we design a VLM (see Section IV-D) and an LVM (see Section IV-E) to update visual and linguistic features bilaterally with the guidance of each other. VLM is shown in part (b) in the top right corner and LVM is illustrated in part (c). Note that in LVM, the original feature is shown at the top left corner, which is denoted as $F_i^m$, and the updated feature is at the output, denoting as $F_i^{m+1}$. At last, an ASPP module [shown in part (d)] receives the final concatenated fused features and generates predicted masks (see Section IV-F).

E. Language Guide Vision Module

Previous work [22], [24], [47] on V-L tasks demonstrates that the information exchange among multilevel features benefits the vision and language interaction process a lot. Therefore, leveraging multilevel information, we propose a novel LVM to update the visual features under the guidance of linguistic features. The operation is shown in Fig. 5(c).

The updated linguistic feature $L_i^{m+1} \in \mathbb{R}^{1 \times C}$ contains rich multimodal context information of $F_i^m$. We utilize $L_i^{m+1}$ to select the relevant information from other two level features $F_j^m$ and $F_k^m$ after necessary transformations. The final aggregated global context feature $F_i^{m+1}$ is obtained by adding $F_i^m$ and relevant information from other two levels

$$F_i^{m+1} = F_i^m + \sum_{k \in \{3,4,5\} \setminus \hat{i}} \sigma(\text{conv}(L_i^{m+1})) \odot F_m^k$$

where $\sigma(\cdot)$ denotes sigmoid function.

affinity map to produce global affinity attention map $A_i^m \in \mathbb{R}^{1 \times HW}$. This process is shown as follows:

$$A_i^m = \text{Softmax} \left( \frac{S_i^m}{\sqrt{C}} \right).$$

where $S_i^m$ is the affinity score of the $i$-th feature map.

Affinity map is reshaped to send to the next level for further processing. Subsequently, the reshaped original fused visual feature $F_i^m$ is multiplied by the attention map $A_i^m$ along the channel dimension to generate the attention feature map $A_i^m \in \mathbb{R}^{1 \times C}$, followed by a convolution layer and a $L_2$ normalization operation as follows:

$$L_i^{m+1} = ||\text{conv}(\text{concat}(L_i^m, A_i^m))||_2$$

where $\text{conv}(\cdot), || \cdot ||_2$ denote $1 \times 1$ convolution, concatenate, and $L_2$ normalization operations, respectively.
F. Segmentation Module

The segmentation module aims to produce the final fine segmentation mask. At first, we obtain a concatenated feature $F^C_2$ which contain multilevel information

$$F^C_2 = \text{concat}(F^3_2, F^4_2, F^5_2).$$  

(10)

Next, as shown in Fig. 5(d), we utilize an ASPP module [45] to capture multiscale information. The structure of ASPP consists of five parallel subnetworks. The first one learns global information by employing global average pooling operation, while the remaining four branches apply atrous convolutions with multiple dilation rates of $\{1, 3, 7, 11\}$, respectively. In the parallel branches, the depthwise separable convolution is applied to reduce the model complexity. After that, the multiscale features are concatenated together. Finally, a $1 \times 1$ convolution and an upsampling operation are adopted to generate the final fine mask $P_m$ with the exact resolution and channel dimension as the input image.

During training, we adopt the sigmoid binary cross entropy (BCE) loss as a minimized objective, which is defined on the predicted output $P_{\text{mask}}$ and the ground truth segmentation mask $G$ as follows:

$$L = \sum_{i=1}^{H \times W} G(i) \log(P_{\text{mask}}(i)) + (1 - P_{\text{mask}}(i)) \log(1 - G(i))$$

(11)

where $i$ is the elements of the ground-truth mask and $H \times W$ denotes the size of the ground-truth mask.

V. EXPERIMENTS

This section elaborates on the experiments’ details, including experiment settings, results, and analysis. Section V-A presents the evaluation metrics and comparison methods we choose. In Section V-B, we describe the implementation details of our experiments. Section V-C analyzes the results of our model. Section V-D demonstrates the ablation study.

A. Settings

We choose five broadly used metrics to comprehensively evaluate the performance of different methods, i.e., intersection over union (IoU) [28], F-measure ($F_\beta$) [52], E-measure ($E_\phi$) [53], Pearson’s correlation coefficient (PCC) [54], and mean absolute error (MAE) [55]. More details could be found in the Appendix.

To illustrate the superiority of our model, we compare several different kinds of methods, which involve two salient detection models (BASNet [48], CPD [49]), two affordance detection models (OSAD-Net [15], OAFFD [14]), two semantic segmentation models (PSPNet [50], DeepLabV3+ [51]), and three referring segmentation models (CMSA [41], BRINet [23], and CMPC [36]). Referring segmentation is the most relevant task as it has almost the same format of input and output as our task. Two typical segmentation tasks: semantic segmentation and salient object detection, are considered as comparisons because our task aims to segment the objects with specific affordance in the scenes. Affordance detection task is considered as it has the same goals as ours. More details can be found in the Appendix.

B. Implementation Details

Our method is implemented using TensorFlow. For visual feature extraction, we choose DeepLab-ResNet101 network [45] which is pretrained on PASCAL-VOC dataset [56] as the backbone. We use the outputs of the DeepLab blocks Res3, Res4, and Res5 as the input multilevel visual features $\{I_3, I_4, I_5\}$. The parameters of the backbone are fixed in the training phase. During training, the input images are randomly clipped from $360 \times 360$ to $320 \times 320$ with a random horizontal flipping. The multilevel visual feature dimension $C_1$ is set to be 1000 in this article.

Meanwhile, for linguistic feature extraction, we first adopt the GloVe word embeddings [43] pretrained on Common Crawl (840B tokens) to initialize the parameters of embedding layers then an LSTM is employed as the language feature extractor. The LSTM shares parameters to embed each phrase. The corresponding phrases to each image are selected from phrase annotations according to affordance categories. The number of phrases for each image is set to be 4, and each phrase is embedded to a vector of $C_1 = 1000$ dimensions.

We train the model using the Adam optimizer [57]. The learning rate is initialized as $2.5 \times 10^{-4}$ with a weight decay of $5 \times 10^{-4}$ with gradually decreasing by a polynomial policy with a power of 0.9. The model is trained on an NVIDIA RTX3080 GPU for 100 epochs with a batch size of 1 and the training duration is about 8 h.

C. Results Analysis

We conduct a comprehensive and thorough analysis of the proposed model in this section.

Comparison With Other Methods: We use the same backbone, input data augmentation strategy, and size mentioned above for all comparison methods. The input of BASNet, CPD, OAFFD, PSPNet, and DLABV3+ is a single query image. The input of OSAD is a support image with a bounding box of human/object and five query images as the original article. The input of CMSA, BRINet, and CMPC is a single query image with a piece of text. The input of ours is a single query image and several phrases. Methods from referring segmentation use the total PAD-L dataset, and methods from other vision-only methods only use the image part of PAD-L.

The results of objective metrics are shown in Table I. It reveals that our method surpasses all other methods in all metrics. Especially in terms of IoU, $E_\phi$, and $F_\beta$, our model achieves 2.4%, 2.0%, and 1.8% performance improvements, respectively. Notably, the table shows that methods involving multimodalities generally achieve better performance than methods only using one modal features because of additional language information guidance. The use of a cyclic interaction mechanism provides our method with better alignments. We also present the subjective visualization results in Fig. 6. Our method generates more precise segmentation masks closer to

\footnote{The pretrained DeepLab-ResNet101 model can be downloaded at https://drive.google.com/drive/folders/0B_rootXHuwsZ0E4Mjh1ZU5xZVU?resourcekey=0-9U12e1br1d6jgmd6UdgUQILink.}
TABLE I
EXPERIMENTAL RESULTS OF 10 MODELS (BASNNet, CPD, OSAD, OAFFD, PSPNet, DePPalabV3+ (DLabV3+) CMSA, BRINet, AND CMPC) IN TERMS OF FIVE METRICS (IOU (↑), $E_\alpha$ (↑), $P_{CC}$ (↑), AND MAE (↓))

| Methods          | BASNet [48] | CPD [49] | OSAD [15] | OAFFD [14] | PSPNet [50] | DePPalabV3+ [51] | CMSA [41] | BRINet [23] | CMPC [36] | Ours |
|------------------|-------------|----------|-----------|------------|-------------|-------------------|-----------|-------------|-----------|------|
| Year             | 2019        | 2019     | 2021      | 2020       | 2017        | 2018              | 2021      | 2020        | 2021      |      |
| IoU (↑)          | 0.491       | 0.496    | 0.554     | 0.459      | 0.464       | 0.509             | 0.571     | 0.579       | 0.579     | 0.582|
| $E_\alpha$ (↑)   | 0.752       | 0.744    | 0.777     | 0.714      | 0.692       | 0.761             | 0.799     | 0.795       | 0.806     | 0.822|
| PCC (↑)          | 0.557       | 0.626    | 0.662     | 0.565      | 0.573       | 0.638             | 0.711     | 0.710       | 0.706     | 0.713|
| MAE (↓)          | 0.086       | 0.083    | 0.083     | 0.098      | 0.138       | 0.064             | 0.063     | 0.061       | 0.062     | 0.061|
| $P_{CC}$ (↑)     | 0.571       | 0.573    | 0.639     | 0.521      | 0.503       | 0.631             | 0.644     | 0.653       | 0.650     | 0.665|

Bold and underline indicate the best and the second-best scores, respectively.

the ground truth than other methods. It indicates that our model can effectively capture the relationship between vision and language. Compared with the multimodal method CMPC, one of the best-performing methods for referring segmentation, our approach introduces fewer noises in the background because of more accurate alignment between cross-modal information. For other methods, unexpected objects may be segmented incorrectly because of the absence of necessary language guidance, and some failure cases are shown in the figure.

We also show the IoU scores of all methods in every affordance category in Table II. It further demonstrates the superior performance of our proposed model. Our model achieves the best performance in almost all categories except in “Beat,” “Bounce,” and “Brush” classes. The highest IoU score (0.766) is in the affordance class “Beat,” which only contains the object “drum” with similar simple and regular shapes. The lowest IoU score (0.443) appears in class “Brush” including object “toothbrush,” which is small and has complicated geometry. It shows that for objects with large, simple, and regular shapes, our model will get higher IoU scores, while it is slightly underperforming for small objects with more complicated shapes.

Single Image With Different Descriptions: To better comprehend the surrounding scenes, when there are multiple objects with different affordances in the same image, our model is expected to be able to segment the corresponding regions based on the natural language descriptions. Some examples are shown in Fig. 7. These examples show that our proposed model can align vision and language information correctly with language changes.

Multiple Images With Same Descriptions: In practice, multiple objects may have the same affordance, although with significant appearance variations in terms of color, shape, and texture. Therefore, our model is expected to identify corresponding objects in different images regardless of these variations.
TABLE II
RESULTS OF DIFFERENT METHODS ON THE PAD-L FOR EACH AFFORDANCE CATEGORY

| Classes       | BASNet [48] | CPD [49] | OSAD [15] | OAFFD [14] | PSPNet [50] | DLABV3+ [51] | CMSA [41] | BRINet [21] | CMPC [136] | Ours         |
|---------------|-------------|-----------|-----------|------------|-------------|--------------|-----------|-------------|------------|--------------|
| Beat          | 0.548       | 0.625     | 0.808     | 0.562      | 0.577       | 0.671        | 0.813     | 0.835       | 0.779      | 0.766        |
| Bounce        | 0.362       | 0.524     | 0.601     | 0.376      | 0.427       | 0.564        | 0.606     | 0.642       | 0.652      | 0.616        |
| Brush         | 0.275       | 0.369     | 0.427     | 0.267      | 0.292       | 0.395        | 0.449     | 0.450       | 0.440      | 0.443        |
| Contain-1     | 0.290       | 0.393     | 0.463     | 0.313      | 0.355       | 0.404        | 0.493     | 0.489       | 0.481      | 0.508        |
| Contain-2     | 0.447       | 0.518     | 0.573     | 0.419      | 0.483       | 0.539        | 0.608     | 0.609       | 0.618      | 0.634        |
| Contain-3     | 0.485       | 0.543     | 0.593     | 0.482      | 0.511       | 0.555        | 0.631     | 0.629       | 0.635      | 0.656        |
| Cut           | 0.448       | 0.511     | 0.557     | 0.446      | 0.472       | 0.524        | 0.594     | 0.587       | 0.595      | 0.621        |
| Fork          | 0.433       | 0.490     | 0.538     | 0.431      | 0.455       | 0.507        | 0.575     | 0.567       | 0.574      | 0.603        |
| Hit           | 0.420       | 0.475     | 0.531     | 0.421      | 0.446       | 0.500        | 0.561     | 0.552       | 0.562      | 0.590        |
| Jump          | 0.395       | 0.438     | 0.502     | 0.388      | 0.404       | 0.458        | 0.523     | 0.520       | 0.526      | 0.556        |
| Kick          | 0.409       | 0.450     | 0.516     | 0.400      | 0.410       | 0.471        | 0.531     | 0.533       | 0.536      | 0.567        |
| Lie           | 0.442       | 0.476     | 0.541     | 0.425      | 0.439       | 0.491        | 0.547     | 0.553       | 0.554      | 0.579        |
| Lift          | 0.445       | 0.480     | 0.546     | 0.429      | 0.443       | 0.494        | 0.549     | 0.554       | 0.557      | 0.581        |
| Look Out      | 0.448       | 0.484     | 0.549     | 0.433      | 0.447       | 0.499        | 0.552     | 0.557       | 0.558      | 0.583        |
| Mix           | 0.465       | 0.488     | 0.542     | 0.428      | 0.449       | 0.488        | 0.541     | 0.541       | 0.547      | 0.568        |
| Pick Up       | 0.469       | 0.488     | 0.541     | 0.427      | 0.448       | 0.486        | 0.538     | 0.538       | 0.546      | 0.567        |
| Play-1        | 0.483       | 0.498     | 0.553     | 0.435      | 0.461       | 0.503        | 0.551     | 0.551       | 0.559      | 0.578        |
| Play-2        | 0.497       | 0.513     | 0.562     | 0.447      | 0.476       | 0.517        | 0.561     | 0.564       | 0.570      | 0.589        |
| Play-3        | 0.493       | 0.519     | 0.572     | 0.452      | 0.483       | 0.525        | 0.570     | 0.572       | 0.578      | 0.596        |
| Play-4        | 0.499       | 0.519     | 0.574     | 0.451      | 0.484       | 0.526        | 0.571     | 0.579       | 0.578      | 0.596        |
| Ride          | 0.502       | 0.518     | 0.575     | 0.455      | 0.486       | 0.528        | 0.574     | 0.575       | 0.580      | 0.597        |
| Roll Dough    | 0.500       | 0.518     | 0.576     | 0.454      | 0.486       | 0.530        | 0.574     | 0.576       | 0.580      | 0.598        |
| Rolling       | 0.500       | 0.515     | 0.570     | 0.456      | 0.479       | 0.525        | 0.579     | 0.580       | 0.588      | 0.603        |
| Scoop         | 0.501       | 0.511     | 0.564     | 0.451      | 0.473       | 0.517        | 0.572     | 0.573       | 0.580      | 0.596        |
| Shelter       | 0.495       | 0.504     | 0.556     | 0.445      | 0.465       | 0.514        | 0.574     | 0.575       | 0.581      | 0.595        |
| Sit           | 0.499       | 0.505     | 0.559     | 0.446      | 0.469       | 0.516        | 0.572     | 0.574       | 0.581      | 0.595        |
| Swing         | 0.494       | 0.499     | 0.555     | 0.440      | 0.461       | 0.507        | 0.572     | 0.574       | 0.581      | 0.596        |
| Take Photo    | 0.494       | 0.499     | 0.555     | 0.441      | 0.461       | 0.508        | 0.573     | 0.574       | 0.581      | 0.596        |
| Throw         | 0.491       | 0.498     | 0.550     | 0.438      | 0.458       | 0.510        | 0.571     | 0.576       | 0.580      | 0.594        |
| Wear-1        | 0.492       | 0.499     | 0.551     | 0.441      | 0.462       | 0.513        | 0.574     | 0.581       | 0.584      | 0.597        |
| Wear-2        | 0.491       | 0.496     | 0.553     | 0.439      | 0.459       | 0.510        | 0.571     | 0.579       | 0.577      | 0.593        |

IoU is used as the evaluation metric. Bold and underline indicate the best and the second-best scores, respectively.

**Fig. 7.** Single image with Different descriptions. When multiple objects with various affordances appear in the same image, our model is expected to highlight the correct object according to the description phrases. The phrases in red indicate the results in the second column, while the blue ones refer to the blue objects in the third column.
Some examples are illustrated in Fig. 8. From the examples, we can find that for the same set of phrases, corresponding referred objects described by the phrases could be highlighted correctly, which proves that our model can cope with the multiplicity property of affordance.

### D. Ablation Study

In this section, we conduct ablation study to investigate the effect of different modules and hyper-parameter settings. We consider the following factors: the number of input phrases, the language encoder method, and cyclic times of CIM module.

**Number of Text Phrases:** To explore the influence of the number of input phrases, we set the phrase number to be $N = 1, 2, 3, 4, 5, 6$, respectively. The results are shown in Table III. The results illustrate that the phrase number influences all five metrics. It suggests that taking four phrases as input makes the model fuse information the most effectively. Contrary to intuition, performance does not continue to improve with the number of phrases increasing after the number getting four. Our model may reach to a bottleneck after that point.

**Language Encoder Method:** We consider to explore the effect of different language encoder methods. We replace LSTM with two different popular pre-trained language models bidirectional encoder representation from transformers (BERT) [58] and embeddings from language models (ELMo) [59]. The results are illustrated in Table IV. It is shown that LSTM outperforms the other two language encoder methods in this task. The possible reason is that the latter two pretrained language models are more suitable for long sentences because of rich text context. However, the text descriptions of affordances in the proposed PAD-L are all short phrases that may limit their capabilities.

**Number of Cycles:** The semantic consistency is enhanced in a cyclic and bilateral manner. To investigate the effect of the number of cycles, we repeat CIM several times. The results are shown in Table V. It is demonstrated that more cyclic times do not necessarily lead to better performance. We set cycling once as the baseline. When CIM repeats twice, the performance outperforms the baseline in several metrics. However, the performance is not as good as the baseline when cycle three times. Our model may get stuck in overfitting as the number of cycles increases.
VI. CONCLUSION

In this article, we propose a novel phrase-based affordance detection task. At first, based on the previously proposed PAD dataset, we annotate the affordance categories using short phrases to construct a new multimodal dataset, PAD-L dataset. Then to align text features and vision features better, we adopt a novel cyclic and bilateral mechanism to cope with the problem caused by the inherent multiplicity property of affordance. Specifically, we design a CBCE-Net, which consists of three main modules: VLM, LVM, and CIM to improve feature representations cyclically and bilaterally. Compared with nine relevant methods, our model achieves the best results in terms of all five evaluation metrics. Compared with nine relevant methods, our model achieves the best results in all five-evaluation metrics and can serve as a strong baseline for future research. Besides, the generated precise pixel-level segmentation masks could be used to support many downstream tasks. The results could be used to extract rich information, such as location and geometric features, which could support tasks like robot’s grasping. The critical points for grasping of objects can be learned from the information for further mechanical movements.

Our approach also has some limitations. At first, our method may not achieve satisfactory results in more complicated scenes than what we explored in this article, such as scenes containing many small objects. To improve our method, we could adopt a locate-then-segment framework to locate objects [60] then generate the mask. Secondly, Our approach aims at detecting all possible objects in the image and cannot detect the one that best fits the intention. We can introduce a sorting mechanism to segment the best match object.

In the future, based on PAD-L, more works could be done to explore the combination of multimodal applications and affordance. For instance, exploring affordance detection in videos with natural language instructions would be a promising topic.

REFERENCES

[1] J. J. Gibson, “The theory of affordances,” in The Ecological Approach to Visual Perception. Hilddale, NJ, USA: Lawrence Erlbaum Associates, 1977, pp. 67–82.
[2] T. Nagarajan, C. Feichtenhofer, and K. Grauman, “Grounded human-object interaction hotspots from video,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 8688–8697.
[3] G. J. J., The Ecological Approach to Visual Perception: Classic Edition. London, England, U.K.: Psychology Press, 2014.
[4] T. E. Horton, A. Chakraborty, and R. S. Amant, “Affordances for robots: A brief survey,” AVANT: Pismo Awangardi Filozoficzno-Naukowej, vol. 2, pp. 70–84, 2012.
[5] S. Qi, S. Huang, P. Wei, and S.-C. Zhu, “Predicting human activities using stochastic grammar,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 1164–1172.
[6] J. Earley, “An efficient context-free parsing algorithm,” Commun. ACM, vol. 13, no. 2, pp. 94–102, 1970.
[7] X. Li, W. Zhai, and Y. Cao, “A tri-attention enhanced graph convolutional network for skeleton-based action recognition,” IET Comput. Vis., vol. 15, no. 2, pp. 110–121, 2021.
[8] T. Bagautdinov, A. Ahlai, F. Fleuret, P. Fua, and S. Savarese, “Social scene understanding: End-to-end multi-person action localization and collective activity recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4315–4324.
[9] K. Fang, T.-L. Wu, D. Yang, S. Savarese, and J. J. Lim, “Demo2vec: Reasoning object affordances from online videos,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 2139–2147.
[10] K. Zhu, W. Zhai, and Y. Cao, “Self-supervised tuning for few-shot segmentation,” in Proc. 29th Int. Conf. Int. Joint Conf. Artif. Intell., 2021.
[11] N. Yamamoto et al., “A brief review of affordance in robotic manipulation research,” Adv. Robot., vol. 31, no. 19–20, pp. 1086–1101, 2017.
[12] Y. Shi et al., “Modeling of everyday objects for semantic grasp,” in Proc. IEEE 23rd Int. Symp. Robot Hum. Interactive Comm., 2014, pp. 750–755.
[13] J. Sawatzky and J. Gall, “Adaptive binarization for weakly supervised affordance segmentation,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops, 2017, pp. 1383–1391.
[14] Y. Shao, Y. Cao, and Y. Kang, “Object affordance detection with relationship-aware network,” Neural Comput. Appl., vol. 32, no. 18, pp. 14321–14333, 2020.
[15] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, “One-shot affordance detection,” in Proc. Int. Joint Conf. Artif. Intell., 2021, pp. 895–901.
[16] H. Kjellström, J. Romero, and D. Kragic, “Visual object-action recognition: Inferring object affordances from human demonstration,” Comput. Vis. Image Understanding, vol. 115, no. 1, pp. 81–90, 2011.
[17] W. Zhai, H. Luo, J. Zhang, Y. Cao, and D. Tao, “One-shot object affordance detection in the wild,” Int. J. Comput. Vis., to be published, doi: 10.1007/s11263-022-01642-4.
[18] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, “Learning visual affordance grounding from demonstration videos,” 2021, arXiv:2108.05675.
[19] H. Chen, H. Tan, A. Kunzt, M. Bansal, and R. Alterovitz, “Enabling robots to understand incomplete natural language instructions using commonsense reasoning,” in Proc. IEEE Int. Conf. Robot. Automat., 2020, pp. 1963–1969.
[20] R. Hu, M. Rohrbach, and T. Darrell, “Segmentation from natural language expressions,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 108–124.
[21] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille, “Recurrent multimodal interaction for referring image segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 1271–1280.
[22] R. Li et al., “Referring image segmentation via recurrent refinement networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5745–5753.
[23] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu, “Bi-directional relationship inferring network for referring image segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 4424–4433.
[24] Y. Jing, T. Kong, W. Wang, L. Wang, L. Li, and T. Tan, “Locate then segment: A strong pipeline for referring image segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 9858–9867.
[25] G. A. Miller, “Wordnet: A lexical database for English,” Commun. ACM, vol. 38, no. 11, pp. 39–41, 1995.
[26] M. Hassan and A. Dharmaratne, “Attribute based affordance detection from human-object interaction images,” in Image and Video Technology. Berlin, Germany: Springer, 2015, pp. 220–232.
[27] H. Grabner, J. Gall, and L. Van Gool, “What makes a chair a chair?” in Proc. Conf. Comput. Vis. Pattern Recognit., 2011, pp. 1529–1536.
[28] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3431–3440.
[29] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsigaridas, “Object-based affordances detection with convolutional neural networks and dense conditional random fields,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst., 2017, pp. 5908–5915.
[30] S. Thermos, G. T. Papadopoulos, P. Daras, and G. Potamianos, “Deep affordance-grounded sensorimotor object recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6167–6175.
[31] H. Wang, W. Liang, and L.-F. Yu, “Transferring objects: Joint inference of container and human pose,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2923–2934.
[32] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, “Generation and comprehension of unambiguous object descriptions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 11–20.
[33] L. Yu, H. Tan, M. Bansal, and T. L. Berg, “A joint speaker-listener-reinforcer model for referring expressions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 7282–7290.
[34] X. Liu, Z. Wang, J. Shao, X. Wang, and H. Li, “Improving referring expression grounding with cross-modal attention-guided erasing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 1950–1959.
[35] L. Yu et al., “Mattnet: Modular attention network for referring expression comprehension,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1307–1315.
[36] S. Liu, T. Hui, S. Huang, Y. Wei, B. Li, and G. Li, “Cross-modal progressive comprehension for referring segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 9, pp. 4761–4775, Sep. 2021.
[37] J. Mi, S. Tang, Z. Deng, M. Goerner, and J. Zhang, “Object affordance based multimodal fusion for natural human-robot interaction,” Cogn. Syst. Res., vol. 54, pp. 128–137, 2019.

[38] J. Mi et al., “Intention-related natural language grounding via object affordance detection and intention semantic extraction,” Front. Neurorobot., vol. 14, 2020, Art. no. 26.

[39] E. Margfroy-Tuay, J. C. Pérez, E. Botero, and P. Arbeláez, “Dynamic multimodal instance segmentation guided by natural language queries,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 630–645.

[40] A. Graves, “Long short-term memory,” in Supervised Sequence Labelling With Recurrent Neural Networks. Berlin, Germany: Springer, 2012, pp. 37–45.

[41] L. Ye, M. Rochan, Z. Liu, X. Zhang, and Y. Wang, “Referring segmentation in images and videos with cross-modal self-attention network,” IEEE Tran. Pattern Anal. Mach. Intell., vol. 44, no. 7, pp. 3719–3732, 2021.

[42] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[43] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in Proc. Conf. Empirical Methods Natural Lang. Process., 2014, pp. 1532–1543.

[44] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, “Mutant: Multimodal Tucker fusion for visual question answering,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2612–2620.

[45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2018.

[46] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 1–11.

[47] D. Wu, J. Fu, T. Mei, and Y. Rui, “Multi-level attention networks for visual question answering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4709–4717.

[48] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, “Basnet: Boundary-aware salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 7479–7489.

[49] Z. Wu, L. Su, and Q. Huang, “Casced partial decoder for fast and accurate salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 3907–3916.

[50] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2881–2890.

[51] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, arXiv:1706.05587.

[52] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, “Contour detection and hierarchical image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 5, pp. 898–916, May 2011.

[53] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, “Enhanced-alignment measure for binary foreground map evaluation,” in Proc. Int. Joint Conf. Artif. Intell., 2018.

[54] O. Le Meur, P. Le Callet, and D. Barba, “Predicting visual fixations on video based on low-level visual features,” Vis. Res., vol. 47, no. 19, pp. 2483–2498, 2007.

[55] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, “Saliency filters: Contrast based filtering for salient region detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2012, pp. 733–740.

[56] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, 2010.

[57] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Representations, 2015.

[58] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019.

[59] J. Sarzynska-Wawer et al., “Detecting formal thought disorder by deep contextualized word representations,” Psychiatry Res., vol. 304, 2021, Art. no. 114135.

[60] P. Wu, W. Zhai, and Y. Cao, “Background activation suppression for weakly supervised object localization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022.