IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Multi-Keypoint Affordance Representation for Functional Dexterous Grasping

Fan Yang, Dongsheng Luo, Wenrui Chen*, Jiacheng Lin, Junjie Cai,
Kailun Yang, Zhiyong Li, and Yaonan Wang

***Abstract*— Functional dexterous grasping requires precise hand-object interaction, going beyond simple gripping. Existing affordance-based methods primarily predict coarse interaction regions and cannot directly constrain the grasping posture, leading to a disconnection between visual perception and manipulation. To address this issue, we propose a multi-keypoint affordance representation for functional dexterous grasping, which directly encodes task-driven grasp configurations by localizing functional contact points. Our method introduces Contact-guided Multi-Keypoint Affordance (CMKA), leveraging human grasping experience images for weak supervision combined with Large Vision Models for fine affordance feature extraction, achieving generalization while avoiding manual keypoint annotations. Additionally, we present a Keypoint-based Grasp matrix Transformation (KGT) method, ensuring spatial consistency between hand keypoints and object contact points, thus providing a direct link between visual perception and dexterous grasping actions. Experiments on public real-world FAH datasets, IsaacGym simulation, and challenging robotic tasks demonstrate that our method significantly improves affordance localization accuracy, grasp consistency, and generalization to unseen tools and tasks, bridging the gap between visual affordance learning and dexterous robotic manipulation. The source code and demo videos are publicly available at https://github.com/PopeyePxx/MKA.**

I. Introduction

Functional dexterous grasping enables robots to execute complex object manipulations from human instructions. Unlike simple grasping, it requires a dexterous hand to adapt grasp postures and contact different object regions according to the task, involving fine physical interactions between the fingers and the object. For example, in *"Hold Drill"*, all five fingers firmly grasp the drill head, whereas in *"Press Drill"*, the index finger presses the switch while the others stabilize the handle. The core challenge lies in inferring task-relevant contact regions and grasp postures from visual perception.

This work was partially supported by the National Key R&D Program of China under Grant 2022YFB4701400/2022YFB4701404, the National Natural Science Foundation of China under Grant 62273137, 62473139, No. U21A20518, and No. U23A20341, the Hunan Provincial Research and Development Project under Grant 2025QK3019, the Hunan Science Fund for Distinguished Young Scholars under Grant 2024JJ2027, and the Open Research Project of the State Key Laboratory of Industrial Control Technology, China (Grant No. ICT2025B20). (*Corresponding author: Wenrui Chen. E-mail: chenwenrui@hnu.edu.cn.*)

F. Yang, D. Luo, J. Cai, W. Chen, K. Yang, and Z. Li are with the School of Artificial Intelligence and Robotics, Hunan University, Changsha 410012, China. (E-mail: ysyf293@hnu.edu.cn.)

J. Lin is with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China.

W. Chen, K. Yang, Z. Li, and Y. Wang are also with the National Engineering Research Center of Robot Visual Perception and Control Technology, Hunan University, Changsha 410082, China.

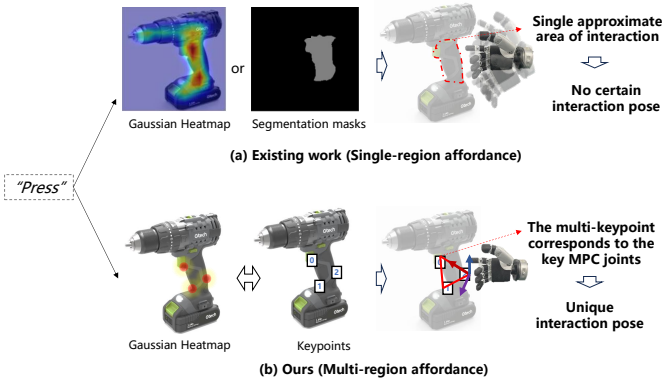

Fig. 1: Comparison between existing affordance-based grasping methods and our proposed Multi-Keypoint Affordance representation. (a) Existing methods identify only a rough interaction region, leading to uncertain interaction poses. (b) Our method localizes multiple keypoints corresponding to dexterous hand joints, enabling a precise and constrained grasping posture.

In the field of vision, affordance-based methods [1], [2], [3], [4], [5] have been widely explored to predict potential human interaction regions. Among them, deep-learning-based approaches estimate heatmaps [4], [5] or segmentation masks [2], [3] to indicate feasible interaction areas. However, existing methods [6], [7] can only provide coarse region predictions given an image and a task. A rough affordance map cannot specify the exact interaction posture, leading to uncertainty in the grasping motion and insufficient constraints for functional dexterous grasping, as shown in Fig. 1(a). Therefore, how to find a novel visual representation that not only identifies task-relevant contact areas but also directly constrains the dexterous grasping posture, ensuring a well-defined interaction between the hand and the object, remains a challenging problem.

Keypoint-based representations offer a potential solution by structuring high-dimensional visual data into a compact and interpretable form. Prior work [8], [9], [10], [11] often decomposes grasping into object and environment keypoints. For example, KETO [10] defines grasp, functional, and operation points, while SKP [11] specifies five surface keypoints for parallel grasping. However, these approaches suffer from limited generalization—keypoints are either manually designed for specific tasks, rely heavily on simulation, or require extensive manual annotations, raising data collection costs.

To address these issues, VRB [12] learns contact points and motion trajectories from human operation videos, im-

proving generalization and applicability, though it still depends on post-processing and yields indirect visual representations. Recent advances in Large Vision Models (LVMs) enhance object feature extraction. For instance, ReKep [13] automatically detects candidate keypoints via LVMs, filters them with vision-language models, and directly guides robotic operations, improving task generalization and linking vision more directly to action.

Despite successes, the above methods primarily focus on simple two-finger pinch grasps and do not extend to dexterous grasping tasks. In dexterous grasping, keypoints must not only determine the grasping location but also constrain the entire hand configuration, ensuring functional stability, as shown in Fig. 1(b). Achieving this goal introduces three key challenges: (1) Fine-grained feature extraction: Dexterous grasping involves small, detailed interaction regions between fingers and the object. How can part-level keypoint features be extracted from the object? (2) Data annotation cost: Dexterous grasping requires precise keypoint annotations, which are costly to acquire. How can reliance on manual annotation be reduced? (3) Keypoint correspondence: Establishing a consistent mapping between object keypoints and hand keypoints is essential for stable grasping. How can robust keypoint correspondence be ensured?

To address the challenges, we propose the Multi-KeyPoint Affordance representation for Functional Dexterous Grasping. By localizing multiple keypoints on the object and the hand, a unique dexterous grasping posture with clear constraints is determined. First, we introduce the Contact-guided Multi-Keypoint Affordance (CMKA) learning, which leverages LVMs for fine-grained affordance feature extraction. The CMKA supervises Egocentric images using hand-object interaction regions in Exocentric images as contact priors via CAM [14], guiding keypoint learning towards meaningful functional contact areas and eliminating the need for manual keypoint annotations. Unlike existing methods that only predict contact regions or keypoint locations without providing actionable grasp execution, we propose the Keypoint-based Grasp matrix Transformation (KGT) to bridge perception and control by explicitly computing the relative pose between the hand and the object. Specifically, KGT leverages the geometric relationship among three semantically meaningful keypoints-the wrist, functional finger (index or thumb), and little finger MCP joints-which form a unique triangular structure capturing the relative contact posture. Experiments across 6 tasks and 18 tool shapes on the public FAH dataset [15], achieving an improvement of 45.35% over the state-of-the-art method in the KLD metric. In both IsaacGym [16] and real robot experiments, we successfully establish the geometric constraint relationship between tool and hand keypoints.

The main contributions of this work are as follows:

- A multi-keypoint affordance representation is proposed, which constrains dexterous grasping postures through the geometric relationships of keypoints in the hand-object interaction region, directly establishing a link between vision and dexterous grasping actions.
- CMKA, a multi-keypoint affordance localization method based on a weakly-supervised framework, and KGT, a keypoint-based hand-object relative pose transformation method, are introduced, leveraging existing human interaction image data and LVMs for learning, effectively reducing data costs, and enabling functional dexterous grasping.
- The proposed algorithm is validated in both simulation and real robot experiments, demonstrating its ability to directly map tasks to grasping actions while exhibiting good generalization across tasks and objects, especially excelling in complex functional grasping scenarios.

II. Related Work

A. Object Representation for Dexterous Grasping

Grasping and manipulation are fundamental topics in robotics. Traditional methods [17], [18], [19] often rely on six Degrees of Freedom (6DoF) poses to represent objects for parallel gripper tasks. However, these methods are insufficient for dexterous grasping, which requires handling precise contact points on functional regions and complex interactions, going beyond simple object poses that only describe overall position and orientation. Early methods such as rigid body modeling [20], [21] and template matching [22], [23] are task-specific and lack generalization, limiting their applicability to diverse tasks. Recent studies have focused on object structure-based grasp affordance representations, such as ContactDB [24], which annotates object-finger contact relationships; the method in [25], which maps contact points to finger regions and intent codes; and F2F [26], which uses knowledge graphs to associate functional object parts with functional fingers. While these methods improve task performance, they depend on ideal perception systems that assume precise segmentation or localization of functional regions—an assumption rarely achievable in real-world settings. In contrast, we propose an object representation specifically designed for dexterous grasping. The object is abstracted as a set of semantically meaningful keypoints on its surface. This structured representation captures functional regions and supports executable grasp synthesis. It provides a bridging link between affordance perception and dexterous grasping control, eliminating the need for idealized perception inputs required by the existing methods [25], [26].

B. Keypoint Representation and Robotic Manipulation

Keypoint-based methods have been widely applied in computer vision tasks such as face recognition [27], [28], human pose estimation [29], and tracking [30], where keypoints typically serve as low-level features or part-level object descriptors. In robotics, keypoints provide compact representations of the environment and objects. For example, KETO [10] and SKP [11] define different types or fixed numbers of keypoints to describe specific tasks, but these methods often lack generalizability across tasks. Recently, ReKep [13] introduced a more generalizable manipulation framework by leveraging Large Visual Models (LVMs) [31], [32] to extract candidate keypoints and vision-language models to filter task-relevant

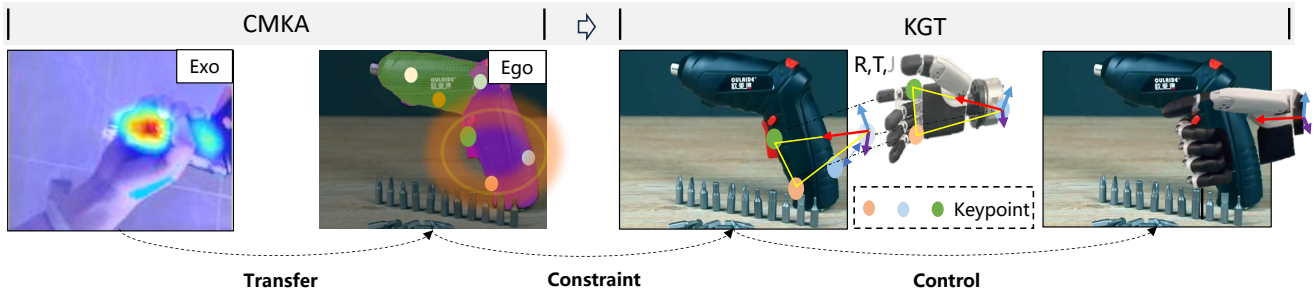

Fig. 2: The key process of learning and connecting visual perception to functional dexterous grasping actions. The Contact-guided Multi-Keypoint Affordance (CMKA) module learns from human demonstrations in Exo images and transfers this knowledge to Ego views, predicting three keypoints constrained by functional grasp priors. These keypoints are then used by the Keypoint-based Grasp matrix Transformation (KGT) module to compute the hand-object relative pose ((R, T)) to control the grasping task.

keypoints for operational guidance. However, ReKep [13] focuses on simple parallel gripper tasks and requires additional reasoning steps, making it unsuitable for dexterous manipulation. Inspired by human hand interactions [33], we propose a multi-keypoint representation based on the wrist, functional fingers, and the little finger. This design directly constrains dexterous grasping postures, providing effective and robust solutions for complex manipulation tasks.

C. Visual Affordance and Interaction

Visual affordance learning explores potential object regions for specific actions and is a key topic in robotic grasping. Early fully supervised methods [34], [35] relied on large-scale annotated datasets, which were both expensive and time-consuming to create. To reduce annotation costs, recent research has shifted toward weakly supervised methods, leveraging keypoints [36], [37] or image-level labels [4], [38], [39]. In this work, we utilize human interaction images to supervise Ego-view images through contact features, significantly reducing training data costs by leveraging existing resources. Existing affordance methods for robotic manipulation, such as VRB [12], learn contact points and trajectories from human operation videos, whereas Robo-ABC [40] generates hand-object contact datasets to enable zero-shot generalization. Similarly, GAT [7] proposes pixel-level affordance learning to capture precise regions. However, those methods often rely on post-processing and additional modules, with coarse affordance regions that lack the fine-grained constraints needed for dexterous grasping.

III. Methodology

In this work, we propose a complete framework that establishes a direct visual representation for functional dexterous grasping with cross-task and cross-object generalization. As illustrated in Fig. 2, the proposed Contact-guided Multi-Keypoint Affordance (CMKA) module learns from human operation experience in exocentric (Exo) images and transfers this knowledge to egocentric (Ego) images, localizing three keypoints constrained by functional dexterous grasping (see Sec. III-A). Using these keypoints, the Keypoint-based Grasp matrix Transformation (KGT) method computes the hand-object relative pose, obtaining the rotation and translation parameters (R, T) required for grasp execution (see Sec. III-B). During inference, the system requires only an Ego image and the affordance class to predict the three functional keypoints on the object.

A. Contact-guided Multi-KeyPoint Affordances Learning

To identify the keypoint regions on the object surface where the fingers should make contact, robust fine-grained feature extraction is required. To achieve this, we first extract multi-level visual features from the Ego and Exo view images using the large vision model DINOv2 [31], which serves as the base visual representation for the following stages. As illustrated in the blue-shaded region of Fig. 3, we then apply LVM-based Multi-Scale Clustering (LMSC) to extract candidate keypoints from different parts of the object surface based on these features (see Sec. III-A.1). In the green-shaded region, we design an affordance-aware keypoint weighting and feature extraction mechanism, where weights based on affordance class (*e.g.*, *"Press"*) are assigned to candidate keypoints, guiding both the selection of the most relevant keypoints and their feature extraction from the Ego view (see Sec. III-A.2). In the yellow-shaded region, we leverage the hand-object interaction feature extraction capability of the weakly-supervised LOCATE framework [38] for contact geometry knowledge transfer from Exo to Ego view. A cosine similarity loss is applied to supervise the keypoints selection, ensuring that the selected keypoints are concentrated around meaningful hand-object contact regions (see Sec. III-A.3).

1) LVM-based Multi-Scale Clustering Module (LMSC): Inspired by ReKep [13], we propose the LMSC module to improve keypoints selection in ego-view images, as illustrated in Fig. 4. Given an input image, we first apply the Segment Anything Model (SAM) [32] to generate a set of region masks $\{M_i\}_{i=1}^{S}$.

To improve the perception of fine-grained structural details and contact point features in the object, we then apply a Hierarchical Feature Fusion (HFF) mechanism to integrate multi-level and multi-scale features. Specifically, we extract features F_{lm} (m=1,2,3) from the last three layers of DINOv2. We then perform the fusion:

$$F = \sum_{m=1}^{3} \alpha_m \cdot \varnothing(F_{lm}), \tag{1}$$

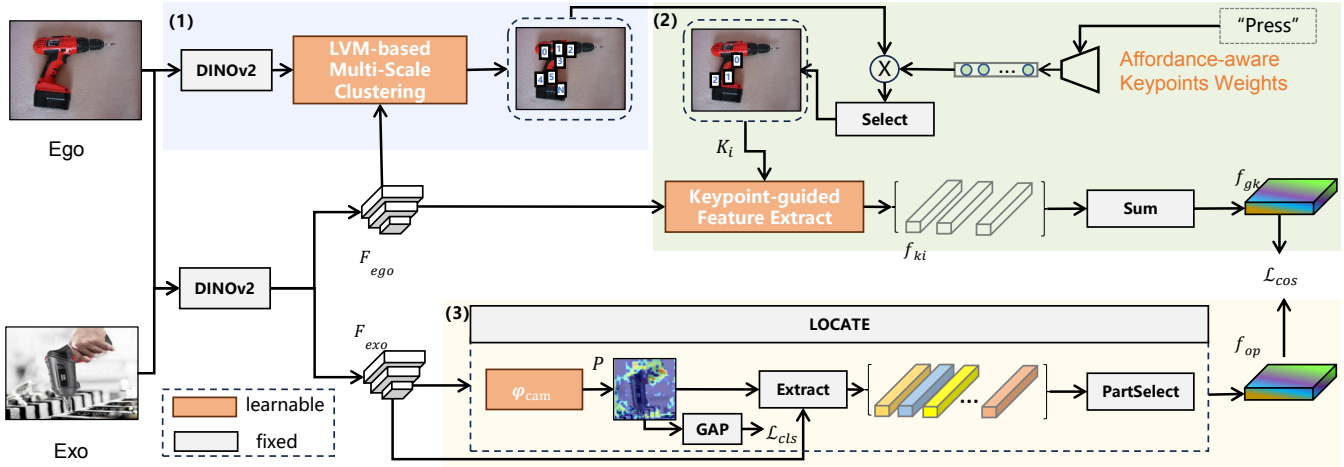

Fig. 3: Framework of the proposed CMKA: (1) LVM-based Multi-Scale Clustering for candidate keypoint extraction; (2) Affordance-aware keypoint weighting and feature extraction from Ego view; (3) Contact geometry knowledge transfer from Exo to Ego view.

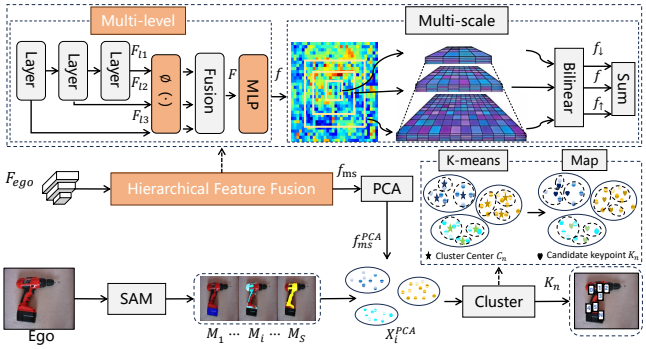

Fig. 4: Candidate keypoint generation using the LMSC module.

where $\varnothing(\cdot)$ denotes a linear layer followed by normalization, and α_m are learnable weights. The fused feature F is further processed by an MLP to obtain the dense feature f.

To capture multi-scale context, we generate bilinearly aligned upsampled features $f_\uparrow$ and downsampled features $f_\downarrow$ versions of f, and sum them with the original to obtain f_{ms}:

$$f_{\text{ms}} = f + f_\uparrow + f_\downarrow. \tag{2}$$

Then, we apply Principal Component Analysis (PCA) to f_{ms} to obtain a reduced feature map $f_{\text{ms}}^{\text{PCA}}$, from which we extract region-wise features X_i^{PCA}.

Lastly, we introduce a two-step Cluster module, as illustrated in Fig. 4, to extract candidate keypoints from X_i^{PCA}. Specifically, we first apply K-means clustering on each X_i^{PCA} to obtain J cluster centers per region, resulting in a total of $N{=}S{\times}J$ cluster centers across all S segmented regions:

$$\{C_n\}_{n=1}^{N} = \{C_{ij}\}_{i=1,j=1}^{S,J} = \text{K-means}(\{X_i^{\text{PCA}}\}_{i=1}^{S}). \tag{3}$$

These cluster centers are then mapped back to the pixel space by selecting the nearest feature vector from X_i^{PCA}, yielding the candidate keypoints $\{K_n\}_{n=1}^{N}$:

$$K_n = \arg\min_{x \in X_i^{\text{PCA}}} \|x - C_n\|^2. \tag{4}$$

If clustering is not applicable (*e.g.*, due to insufficient pixels), the geometric centroid of M_i is used as a fallback.

2) Affordance-aware Keypoint Feature Extraction from Ego View: To extract keypoint features from the Ego view image, we define a set of learnable weights $W{\in}\mathbb{R}^{T\times N}$, where T represents the number of affordance classes and N is the number of candidate keypoints. These weights are multiplied with the candidate keypoints K_n to select the final three keypoints K_i (where $i{=}1,2,3$) for feature extraction from the corresponding regions in the Ego view image.

For the selected keypoints K_i, we define a circular region centered at each keypoint with a radius r and extract features from these regions, denoted as F_{ki}.

To align the features from the Ego and Exo views in a unified feature space, we project the features of the selected three keypoints and sum them to obtain the final keypoint feature f_{gk}, which encapsulates the contact geometry information from the Ego view:

$$f_{gk} = \sum_{i=1}^{3} \text{proj}(F_{ki}), \tag{5}$$

where $\text{proj}(\cdot)$ denotes a linear projection layer.

3) Contact Geometry Knowledge Transfer: To supervise egocentric keypoints selection, we define an affordance-specific activation function φ_{cam} [14], which consists of a feed-forward layer, two convolutional layers, and a $1{\times}1$ class-aware convolution. Given the exocentric feature map F_{exo}, φ_{cam} produces the affordance localization map $P{\in}\mathbb{R}^{C\times H\times W}$, where C is the number of affordance classes. We apply global average pooling on P to obtain class-wise affordance scores, which are optimized using a cross-entropy loss L_{cls}.

In addition, we adopt the Extract and PartSelect modules from LOCATE [38] to obtain the object-part prototype feature f_{op}, by filtering part embeddings through clustering and metric, resulting in representative contact features.

Next, we calculate the cosine similarity loss L_{cos} between the Exo prototype features f_{op} and the global Ego keypoint features f_{gk}:

$$L_{\text{cos}} = 1 - \frac{f_{\text{op}} \cdot f_{\text{gk}}}{\|f_{\text{op}}\| \|f_{\text{gk}}\|}. \tag{6}$$

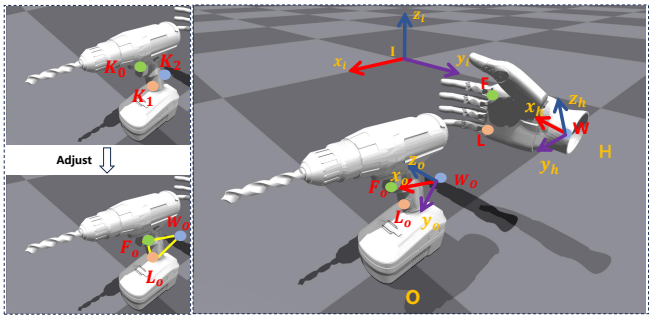

Fig. 5: Illustration of KGT method in IsaacGym [16], showing the keypoints on the object and the hand (functional finger, little finger, and wrist) and their role in coordinate frame construction.

The final loss is the combination of the classification loss and the cosine similarity loss, ensuring that the contact geometry knowledge is accurately transferred between the two views:

$$L = L_{\text{cls}} + L_{\text{cos}}. \tag{7}$$

B. Keypoint-based Grasp Matrix Transformation

After obtaining the three keypoints K_i on the object, we apply the KGT to obtain the relative pose transformation matrix (R,T) between the dexterous hand and the tool. Specifically, as shown in the Fig. 5 (a), we take K_0 as the reference point, determine the direction from K_0 to K_1, and form a plane using K_0, K_1, and K_2. Based on the hand model (yellow triangle), we adjust the keypoints, resulting in the corrected contact points positions in the world coordinate system F_o, L_o, and W_o.

Then, as shown in the Fig. 5(b), we define the object coordinate system O with W_o as the origin, the x-axis as $\mathbf{x}_o = \frac{\overrightarrow{W_oF_o}}{\|\overrightarrow{W_oF_o}\|}$, the z-axis as $\mathbf{z}_o = \frac{\overrightarrow{W_oF_o} \times \overrightarrow{W_oL_o}}{\|\overrightarrow{W_oF_o} \times \overrightarrow{W_oL_o}\|}$, and the y-axis as $\mathbf{y}_o = \mathbf{z}_o \times \mathbf{x}_o$, leading to the object rotation matrix in the world coordinate system:

$$R_O^I = [\mathbf{x}_o, \mathbf{y}_o, \mathbf{z}_o]. \tag{8}$$

At the same time, we obtain the transformation matrix between the world coordinate system I and the object coordinate system O:

$$T_O^I = \begin{bmatrix} R_O^I & W_o \\ 0 & 1 \end{bmatrix}. \tag{9}$$

Similarly, the hand coordinate system H is defined with W as the origin, the x-axis as $\mathbf{x}_h = \frac{\overrightarrow{WF}}{\|\overrightarrow{WF}\|}$, the z-axis as $\mathbf{z}_h = \frac{\overrightarrow{WF} \times \overrightarrow{WL}}{\|\overrightarrow{WF} \times \overrightarrow{WL}\|}$, and the y-axis as $\mathbf{y}_h = \mathbf{z}_h \times \mathbf{x}_h$, where F, L, and W represent the keypoints positions on the hand in the world coordinate system, yielding the hand rotation matrix:

$$R_H^I = [\mathbf{x}_h, \mathbf{y}_h, \mathbf{z}_h]. \tag{10}$$

The relative rotation matrix between the hand and the object is then computed as:

$$R = (R_O^I)^{-1} R_H^I, \tag{11}$$

while the translation vector is given by:

$$T = (T_O^I)^{-1}(W - W_o). \tag{12}$$

IV. Experiments

A. Experiment Setup

Datasets: The public challenging FAH benchmark [15] is a large-scale affordance-annotated dataset specifically designed for hand-object interactions. It contains 6 functional grasp affordances (such as *"Press"*, *"Click"*, *"Hold"*, *etc.*). and 18 tools, with 5,858 images spanning both Exo and Ego views. The dataset provides image-level affordance labels for weakly supervised learning and annotations for coarse dexterous grasp gestures targeting specific "*Task Tool*" pairs. However, its test set only includes heatmap annotations for functional finger contact regions. To address this limitation, we annotate two additional contact points (little finger and wrist projection areas). Specifically, polygons with up to five points are constructed around finger keypoints within a radius of 5*mm*, and Gaussian kernels are applied at each point to generate dense annotations. During training, point annotations are added to the object regions in each Ego-view image to distinguish foreground and background during segmentation with SAM [32].

Implementation Details: Experiments are conducted on an NVIDIA RTX A6000 GPU. The model is trained using the SGD optimizer with a learning rate of 0.01 over 15 iterations. Images are resized to a resolution of 448×448.

Metrics: For affordance grounding, we adopt Kullback-Leibler Divergence (KLD), Similarity (SIM), and Normalized Scanpath Saliency (NSS) following prior work [4], [38].

For functional grasping, we measure the functional grasp success rate (FGS) as defined in [25]: a grasp is successful if the hand holds the object stably for at least ten seconds and correctly performs the intended action on the tool's functional area.

In addition, we introduce the 2D-to-Physical Contact Consistency (TPC) metric to evaluate whether the predicted 2D keypoints, when projected into 3D, fall within the functional contact region. TPC is defined as the ratio of the number of hits n to the three selected keypoints.

B. Results of Functional Affordance Grounding

In this section, we present qualitative and quantitative results to demonstrate the effectiveness of our method on the FAH test set [15]. As weakly supervised methods for multi-region affordance localization are scarce in the state of the art, we use ReKep* [13], a keypoint prediction method, as our baseline.

Quantitative Results. As shown in Tab. I, our method significantly outperforms ReKep* [13] across multiple metrics. Specifically, it improves KLD by 45.35%, increases SIM by 54.19%, and improves NSS by 101.63%. These improvements stem from ReKep*'s lack of adaptation to dexterous grasping. While ReKep* [13] originally relies on manually selected keypoints, its modified version ReKep* [13] randomly generates three keypoints without explicit modeling of functional contact regions. In contrast, our method employs a learnable weighting mechanism to generate keypoints specifically for dexterous grasping, ensuring their alignment with functional contact regions.

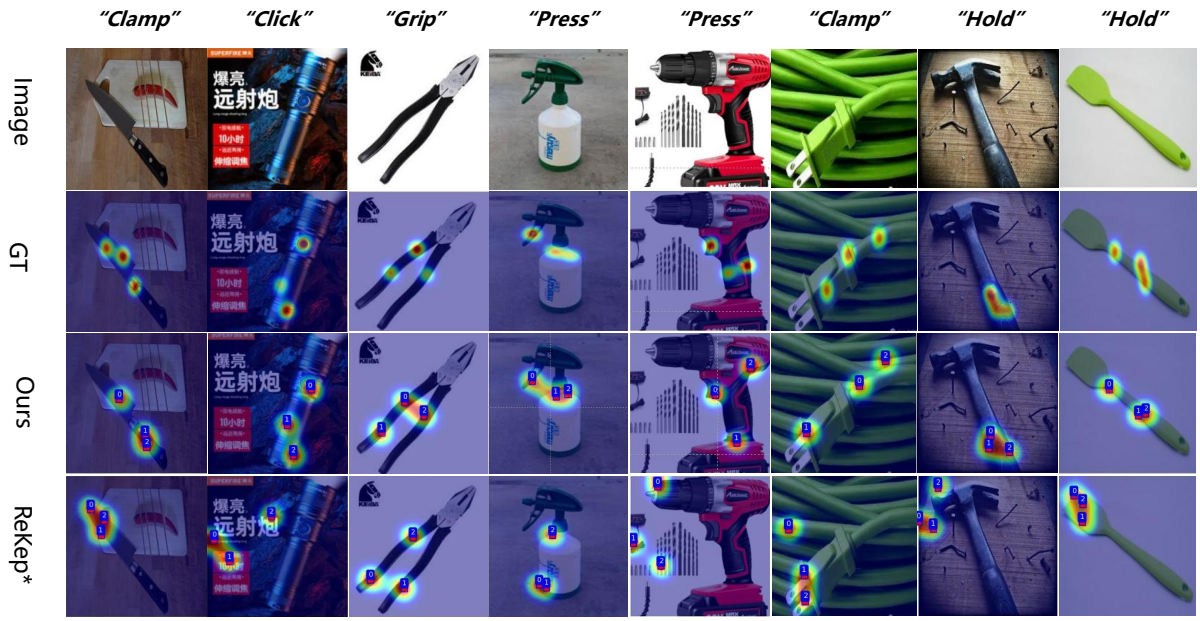

Fig. 6: Qualitative comparison between our approach and the state-of-the-art multi-keypoint affordance grounding method (ReKep* [13]) on the FAH test set [15]. Our proposed method predicts keypoints that are more concentrated in the contact areas and captures the geometric information of the grasping posture.

TABLE I: Comparison to state-of-the-art method on the FAH test set [15]. The **best** results are highlighted in bold. (↑/↓ means higher/lower is better).

Model	KLD (↓)	SIM (↑)	NSS (↑)
ReKep* [13]	9.213	0.203	0.429
Ours	**5.035**	**0.313**	**0.865**

Hyperparameter Sensitivity Analysis. To evaluate the impact of the total number of candidate keypoints $N{=}S{\times}J$ on model performance, we systematically test different combinations of region numbers $S{\in}\{2,3,4\}$ and cluster centers per region $J{\in}\{2,3,4,5\}$, using KLD, SIM, and NSS as evaluation metrics (see Fig. 7). Experimental results show that the configuration $S{=}3$, $J{=}4$ (*i.e.*, $N{=}12$) achieves the best performance across all three metrics.

In general, moderately increasing the number of candidate keypoints helps improve spatial coverage and semantic diversity, enabling the model to better perceive task-relevant regions. However, an excessive number of keypoints may introduce redundancy and noise, thereby hindering the learning of discriminative weights. Notably, under the condition $S{=}3$, the performance of $J{=}3$ is not only lower than that of $J{=}4$, but also slightly worse than $J{=}2$. This indicates that under weak supervision, the model benefits more from structurally coherent and well-distributed candidate arrangements rather than simply increasing the number of keypoints.

Ablation Study. The object priors provided by SAM [32] are crucial for constraining keypoint proposals to objects in the scene rather than the background. Thus, we focus on analyzing the critical visual feature extraction network in CMKA. As shown in Tab. II, DINOv2 [31], as the backbone network combined with our designed Hierarchical Feature Fusion (HFF) module, achieves the best performance. In

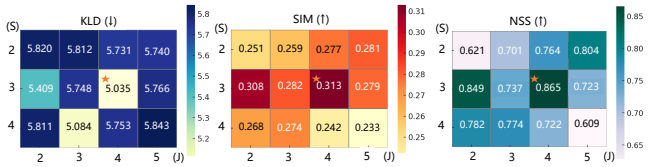

Fig. 7: Performance heatmaps for combinations of region number S and cluster number J. The best results are consistently achieved at $S{=}3$, $J{=}4$ (marked by ★).

TABLE II: Ablation study on different feature extractors. FFL: feed-forward layer. HFF: hierarchical feature fusion.

DINOv2	DINO-ViT	HFF	FFL	KLD (↓)	SIM (↑)	NSS (↑)
✓		✓		**5.035**	**0.313**	**0.865**
✓			✓	5.517	0.302	0.67
	✓	✓		5.807	0.267	0.77
	✓		✓	6.075	0.253	0.65

the backbone network, DINOv2 generates clearer features compared to DINO-ViT [41], better distinguishing fine-grained object regions and leading to improved performance. Furthermore, compared to replacing HFF with a simple fully connected network, HFF, with its multi-layer and multi-scale feature mapping, demonstrates superior potential.

Qualitative Analysis. As shown in Fig. 6, we compare visibility grounding results from Ground Truth (GT), our method, and the baseline ReKep* [13]. Our method accurately localizes keypoints within the hand-object contact region while preserving the spatial relationships among the functional finger, little finger, and wrist, ensuring a meaningful distribution for dexterous grasping. For example, in *"Click Flashlight"*, keypoints correctly align with the thumb and little finger, while in *"Press Drill"*, they align with the index finger, little finger, and wrist. In contrast,

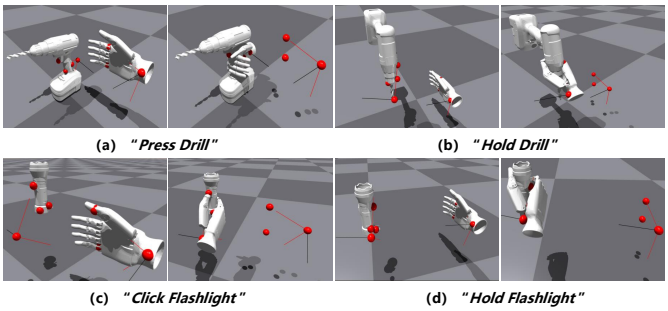

Fig. 8: Visualization of initial and final hand-object states in IsaacGym [16] for different "*Task Tool*" combinations. The red spheres represent the three keypoints used for grasp transformation.

ReKep* [13], which relies on manual post-processing, fails to constrain keypoints within the contact region and lacks spatial consistency, resulting in scattered and less reliable affordance localization.

C. Evaluation of Keypoint-Based Grasp Transformation

To validate the effectiveness of the keypoint-based dexterous grasp transformation method KGT, we conduct experiments on four "*Task Tool*" combinations: *"Press Drill"*, *"Hold Drill"*, *"Click Flashlight"*, and *"Hold Flashlight"*. As shown in Fig. 8, we visualized the initial and final hand-object states in the simulation environment IsaacGym [16]. The results demonstrate that our method accurately computes the grasp transformation matrix, enabling precise hand-object interaction across different task-tool combinations with varying initial hand-object relative postures. For functional interaction tasks, such as *"Press Drill"* and *"Click Flashlight"*, the method ensures correct contact between the functional fingers and the target components. For general grasping tasks, such as *"Hold"*, our method achieves a natural grasp, ensuring a reasonable hand posture.

D. Performance in Real-World Scenarios

As shown in Fig. 9 (a), the real-world platform consists of an Inspire hand, a UR5 industrial robotic arm, an Intel RealSense camera, a tool rack, and a control computer. To address real-world uncertainties, we introduce keypoint relative position calibration annotations based on the Inspire model during the grasping process, as shown in Fig. 9(b).

In the experiments, we evaluated common daily tools unseen during training, including three *"Task-Tool"* pairs with strict functional grasping requirements: *"Click Flashlight"*, *"Press Drill"*, and *"Press Spraybottle"* (Fig. 9(c)). The process includes four steps: (1) CMKA localizes three affordance keypoints and estimates their planar relationship; (2) depth values at the 2D locations are retrieved from the RGB-D depth map to obtain 3D coordinates (see the bottom of the second row); (3) keypoints are adjusted on the hand palm based on calibration (see the yellow triangles in the third row); (4) KGT computes the wrist grasp pose $W(R,T)$, combined with a coarse grasp angle J from FAH [15], to execute the grasp. Results show that our method effectively

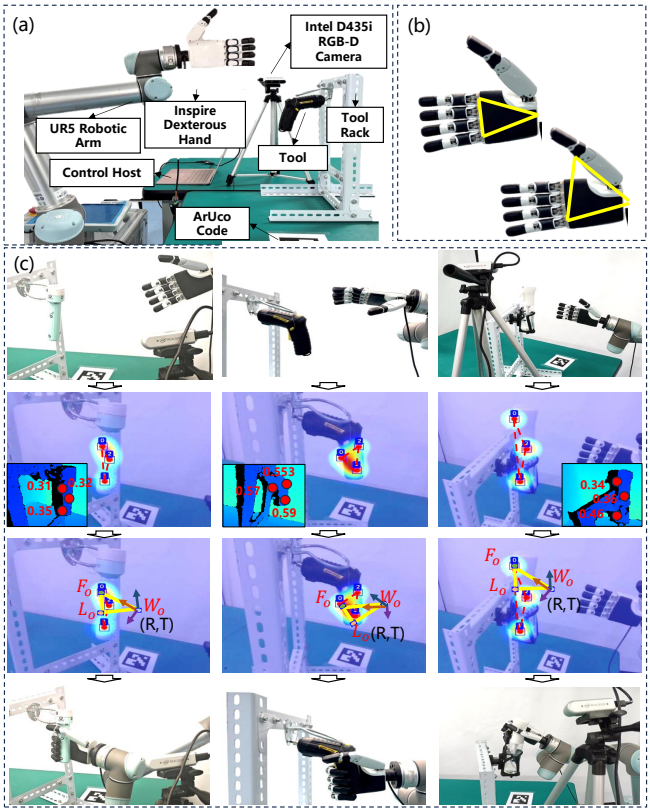

Fig. 9: Experiments with three typical "*Task Tool*" in real-world scenarios. (a) Hardware setup. (b) Calibration standards illustrating the geometric relationship between keypoints when the functional finger is the index finger (up) and the thumb (down). (c) Functional grasping steps across the 1st to 4th rows: Initial, Localization, Adjustment, and Control.

TABLE III: Results of FGS / TPC (%) on three representative *"Task-Tool"* in real-world scenarios.

Method	*Click Flashlight*	*Press Drill*	*Press Spraybottle*
GAAF-Dex [15]	60 / –	0 / –	0 / –
Ours	**80 / 66.7**	**60 / 100**	**40 / 66.7**

bridges perception and execution, enabling precise finger-object alignment in complex tasks like *"Press"* and *"Click"*.

Furthermore, due to the lack of direct methods combining perception and dexterous grasping, we compared our method with the state-of-the-art GAAF-Dex [15] by the functional grasp success rate (FGS). As shown in Tab. III, cross the three complex tasks, our method achieves an average FGS of 60%, compared to 20% for GAAF-Dex, corresponding to an absolute improvement of 40 percentage points. GAAF-Dex relies on the similarity between initial and target grasp orientations, succeeding only in cases like *"Click Flashlight"*, but failing in others due to the lack of rotation adaptation. In contrast, our method handles arbitrary initial poses.

Tab. III reports the 2D-to-Physical contact consistency (TPC) results. Our method achieves 100% on *"Press Drill"* and 66.7% on both *"Click Flashlight"* and *"Press Spraybottle"*, showing reliable contact alignment across tasks, though performance varies due to challenges like depth ambiguity

and transparency. While high TPC does not ensure high FGS, the results demonstrate good generalization of our method and provide a solid basis for further improvement with feedback and online adjustment.

V. Conclusion and Future Work

This work proposes a keypoint-based affordance representation for functional dexterous grasping. By leveraging human experience data for weak supervision and integrating the CMKA module with large visual models, the proposed method achieves precise multi-point contact localization with pose constraints for functional grasping, effectively reducing annotation costs and improving generalization. The KGT method enables the mapping of dexterous hand postures to object keypoints, ensuring a direct connection between perception and action. Experimental results demonstrate that our method outperforms existing approaches in both localization accuracy and functional grasp success rate. Real-world experiments further show that, while 2D vision alone provides reliable perception, it may be insufficient to ensure robust grasp execution under real-world uncertainties.

In the future, we aim to utilize multimodal information to enhance the accuracy and stability of multi-point 3D localization in real-world scenarios.

References

[1] J. J. Gibson, "The theory of affordances," *Hilldale, USA*, vol. 1, no. 2, pp. 67–82, 1977.

[2] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos, "Affordance detection of tool parts from geometric features," in *Proc. ICRA*, 2015, pp. 1374–1381.

[3] R. Xu, F.-J. Chu, C. Tang, W. Liu, and P. A. Vela, "An affordance keypoint detection network for robot manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2870–2877, 2021.

[4] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "Learning affordance grounding from exocentric images," in *Proc. CVPR*, 2022, pp. 2242–2251.

[5] L. Xu, Y. Gao, W. Song, and A. Hao, "Weakly supervised multimodal affordance grounding for egocentric images," in *Proc. AAAI*, vol. 38, no. 6, 2024, pp. 6324–6332.

[6] T. Nguyen *et al.*, "Language-conditioned affordance-pose detection in 3D point clouds," in *Proc. ICRA*, 2024, pp. 3071–3078.

[7] G. Li *et al.*, "Learning precise affordances from egocentric videos for robotic manipulation," *arXiv preprint arXiv:2408.10123*, 2024.

[8] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013.

[9] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, "KPAM: KeyPoint affordances for category-level robotic manipulation," in *Proc. ISRR*, 2019, pp. 132–157.

[10] Z. Qin, K. Fang, Y. Zhu, L. Fei-Fei, and S. Savarese, "KETO: Learning keypoint representations for tool manipulation," in *Proc. ICRA*, 2020, pp. 7278–7285.

[11] Z. Luo, W. Xue, J. Chae, and G. Fu, "SKP: Semantic 3D keypoint detection for category-level robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5437–5444, 2022.

[12] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, "Affordances from human videos as a versatile representation for robotics," in *Proc. CVPR*, 2023, pp. 1–13.

[13] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, "ReKep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation," *arXiv preprint arXiv:2409.01652*, 2024.

[14] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proc. CVPR*, 2018, pp. 1325–1334.

[15] F. Yang *et al.*, "Learning granularity-aware affordances from human-object interaction for tool-based functional grasping in dexterous robotics," *arXiv preprint arXiv:2407.00614*, 2024.

[16] V. Makoviychuk *et al.*, "Isaac gym: High performance GPU based physics simulation for robot learning," in *Proc. NeurIPS*, 2021.

[17] S. Srivastava, E. Fang, L. Riano, R. Chitnis, S. Russell, and P. Abbeel, "Combined task and motion planning through an extensible planner-independent interface layer," in *Proc. ICRA*, 2014, pp. 639–646.

[18] S. Tyree *et al.*, "6-DoF pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark," in *Proc. IROS*, 2022, pp. 13 081–13 088.

[19] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6D pose estimation and tracking of novel objects," in *Proc. CVPR*, 2024, pp. 17 868–17 879.

[20] C. Rosales, R. Suárez, M. Gabiccini, and A. Bicchi, "On the synthesis of feasible and prehensile robotic grasps," in *Proc. ICRA*, 2012, pp. 550–556.

[21] S. El-Khoury, R. De Souza, and A. Billard, "On computing task-oriented grasps," *Robotics and Autonomous Systems*, vol. 66, pp. 145–158, 2015.

[22] C. Gabellieri *et al.*, "Grasp it like a pro: Grasp of unknown objects with robotic hands based on skilled human expertise," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2808–2815, 2020.

[23] M. Kokic, D. Kragic, and J. Bohg, "Learning task-oriented grasping from human activity datasets," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3352–3359, 2020.

[24] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays, "ContactDB: Analyzing and predicting grasp contact via thermal imaging," in *Proc. CVPR*, 2019, pp. 8709–8719.

[25] T. Zhu, R. Wu, J. Hang, X. Lin, and Y. Sun, "Toward human-like grasp: Functional grasp by dexterous robotic hand via object-hand semantic representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 521–12 534, 2023.

[26] F. Yang *et al.*, "Task-oriented tool manipulation with robotic dexterous hands: A knowledge graph approach from fingers to functionality," *IEEE Transactions on Cybernetics*, vol. 55, no. 1, pp. 395–408, 2025.

[27] M. Mayo and E. Zhang, "3D face recognition using multiview keypoint matching," in *Proc. AVSS*, 2009, pp. 290–295.

[28] S. Berretti, B. Ben Amor, M. Daoudi, and A. Del Bimbo, "3D facial expression recognition using sift descriptors of automatically detected keypoints," *The Visual Computer*, vol. 27, pp. 1021–1036, 2011.

[29] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *Proc. FG*, 2017, pp. 468–475.

[30] S. Chan, X. Zhou, and S. Chen, "Robust adaptive fusion tracking based on complex cells and keypoints," *IEEE Access*, vol. 5, pp. 20 985–21 001, 2017.

[31] M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, vol. 2024, 2024.

[32] A. Kirillov *et al.*, "Segment anything," in *Proc. ICCV*, 2023, pp. 3992–4003.

[33] J. Hang *et al.*, "DexFuncGrasp: A robotic dexterous functional grasp dataset constructed from a cost-effective real-simulation annotation system," in *Proc. AAAI*, vol. 38, no. 9, 2024, pp. 10 306–10 313.

[34] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Object-based affordances detection with convolutional neural networks and dense conditional random fields," in *Proc. IROS*, 2017, pp. 5908–5915.

[35] Y. Yang, W. Zhai, H. Luo, Y. Cao, J. Luo, and Z.-J. Zha, "Grounding 3D object affordance from 2D interactions in images," in *Proc. ICCV*, 2023, pp. 10 871–10 881.

[36] J. Sawatzky, A. Srikantha, and J. Gall, "Weakly supervised affordance detection," in *Proc. CVPR*, 2017, pp. 5197–5206.

[37] J. Sawatzky and J. Gall, "Adaptive binarization for weakly supervised affordance segmentation," in *Proc. ICCVW*, 2017, pp. 1383–1391.

[38] G. Li, V. Jampani, D. Sun, and L. Sevilla-Lara, "LOCATE: Localize and transfer object parts for weakly supervised affordance grounding," in *Proc. CVPR*, 2023, pp. 10 922–10 931.

[39] T. Nagarajan, C. Feichtenhofer, and K. Grauman, "Grounded human-object interaction hotspots from video," in *Proc. ICCV*, 2019, pp. 8687–8696.

[40] Y. Ju, K. Hu, G. Zhang, G. Zhang, M. Jiang, and H. Xu, "Robo-ABC: Affordance generalization beyond categories via semantic correspondence for robot manipulation," in *Proc. ECCV*, vol. 15099, 2024, pp. 222–239.

[41] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel, "Deep ViT features as dense visual descriptors," *arXiv preprint arXiv:2112.05814*, 2021.