MAAL: Multimodality-Aware Autoencoder-based Affordance Learning for 3D Articulated Objects

Yuanzhi Liang[1], Xiaohan Wang[2], Linchao Zhu[2], and Yi Yang[2]

[1]ReLER Lab, AAII, University of Technology Sydney
[2]CCAI, Zhejiang University
yuanzhi.Liang@student.uts.edu.au {xiaohan.wang, zhulinchao, yangyics}@zju.edu.cn

Abstract

Inferring affordance for 3D articulated objects is a challenging and practical problem. It is a primary problem for applying robots to real-world scenarios. The exploration can be summarized as figuring out where to act and how to act. Correspondingly, the task mainly requires producing actionability scores, action proposals, and success likelihood scores according to the given 3D object information and robotic information. Current works usually directly process multi-modal inputs with early fusion and apply critic networks to produce scores, which leads to insufficient multi-modal learning ability and inefficiently iterative training in multiple stages. This paper proposes a novel Multimodality-Aware Autoencoder-based affordance Learning (MAAL) for the 3D object affordance problem. It is an efficient pipeline, trained in one go, and only requires a few positive samples in training data. More importantly, MAAL contains a MultiModal Energized Encoder (MME) for better multi-modal learning. It comprehensively models all multi-modal inputs from 3D objects and robotic actions. Jointly considering information from multiple modalities, the encoder further learns interactions between robots and objects. MME empowers the better multi-modal learning ability for understanding object affordance. Experimental results and visualizations, based on a large-scale dataset PartNet-Mobility, show the effectiveness of MAAL in learning multi-modal data and solving the 3D articulated object affordance problem.

1. Introduction

Recently, robots have been widely used in various applications in manufacturing, transportation, and other industries. Toward diverse tasks, a fundamental requirement is to interact with objects by robots. To this end, the robots need to understand real-world objects, use grippers or other manipulators in the robotic system, and interact with given objects in a given scenario. As a primary problem, the object affordance problem [21, 14] is conceptualized and summarized as the first step for the interaction of robots and objects. It aims to figure out where and how to interact with an object by the robot in a given environment. Many works [4, 29] propose various solutions to solve the affordance problem. However, due to the diversity of instances and complexity of practical robotic scenarios, the problem is still far from being resolved.

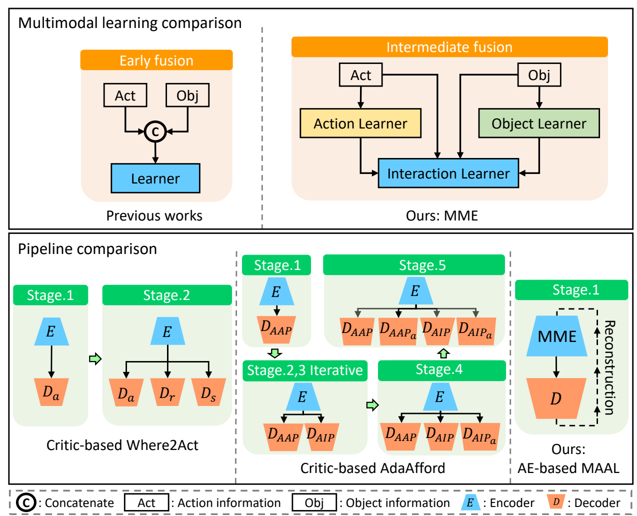

Figure 1. Comparison of methods. MAAL contains a MME module, which provides better multi-modal learning ability. Besides, previous methods with critics or decoders require multiple training stages. MAAL pipeline only contains one step and is trained in one go, which is more efficient.

Specifically, recent works focus on the affordance problem of interacting with 3D articulated objects [30, 9]. Mo et al. [28] introduce a solid benchmark for learning to manipulate articulated objects. They construct a large-scale

3D articulated object dataset and formulates a standard benchmark for the 3D articulated object affordance problem. Wang et al. [45] consider the kinematic and dynamic uncertainties of objects. They design multiple critics to improve the understanding of hidden kinematic information in articulated objects. More works [29, 53] continuously emerge, pushing the frontier of solving the 3D object affordance problem.

Moreover, previous works can be concluded as early fusion [22] for learning multi-modal data and critic-based learning [28, 45] for 3D object affordance. Specifically, they usually concatenate all data (e.g., the point cloud of a 3D object, the robot gripper direction, etc.) as inputs. Then, multiple critics or decoders, trained by classification loss according to labels (negative or positive) initially, are introduced to leverage supervision for other networks.

The straightforward idea leads to significant advancements but still has two defeats. **First**, learning of inputs neglects the correlation between multi-modal data. In the 3D object affordance problem, the input data are from various modalities (i.e., object modality and robot modality). The relationships and interactions between objects and robots are valuable clues for understanding affordance [14, 21]. However, as shown in Fig 1, direct concatenation, as in [28, 45], considering as an early fusion operation [22], would miss the correlation between inputs [27, 49]. This leads that the multi-modal inputs and their interaction may not be sufficiently learned by the previous works. **Second**, the critic-based pipeline is not efficient enough. It requires adequately labeled samples to teach the critics to distinguish the difference between negative samples and positive samples [51, 52]. However, as in [28], training data of articulated object affordance are sampled from $SE(3)$ space, and most actions fail during manipulation. This means most of the samples are negative. For example, sometimes, only 1% [28] of the data are positive samples for pulling action. Training of critic-based methods needs all the samples for training and consumes larger training time. Moreover, critics or decoders need to be trained independently. Then, they will be fixed or iteratively updated with the training of other networks, as shown in Fig 1. The training procedure with multiple stages further increases the overall training time.

To overcome above defeats, we present a novel solution named Multimodality-Aware Autoencoder-based affordance Learning (MAAL). In MAAL, a MultiModal Energized Encoder (MME) is introduced to handle multi-modal inputs in the affordance problem. MME energizes the multi-modal learning ability to understand 3D object affordance. Then, rather than a critic-based designation, MAAL leverages the deep autoencoder (AE) [11, 16] to solve the affordance problem and achieve better training efficiency.

Toward better multi-modal learning, MME is proposed to comprehensively understand data from various modalities and fused features at different levels. Specifically, it involves three branches, carefully designed for learning information in object modality, robot modality, and their interactions. This empowers MAAL to pursue a better understanding of affordance from different perspectives in modalities. Moreover, rather than directly concatenating all data and applying early fusion for various modalities, our encoder considers the correlation between inputs and fuses multi-level features according to the modalities. This can formulate better multi-modal learning than simply early fusion, as proved in [49, 31, 5].

Furthermore, MAAL introduces AE [11] pipeline to solve the 3D affordance problem more efficiently. AE can learn the valuable pattern [51, 42, 52] in high-dimensional data points without labeled examples [15, 13, 6]. This property leads AE can only use positive samples to learn specific valuable patterns from datasets. This also induces the better computational efficiency of the AE pipeline in solving the affordance problem. Besides, rather than learning representations with multiple critics, it only uses reconstruction loss [52, 51] as supervision. The overall pipeline can be trained in one go without multiple training steps for different parts. All these advantages lead that MAAL can achieve better training efficiency than previous critic-based works.

In addition to the above encoder, our MAAL has an action memory and an action decoder, which are used to formulate the AE pipeline. More than applying AE, MAAL specifically considers the properties of 3D object affordance, which takes object information as known conditions and aims to produce action proposals. Correspondingly, MAAL takes multi-modal data as inputs and only reconstructs action proposals as outputs. This leads the network to concentrate on learning action information and the interaction between robots and objects rather than remembering object information and overfitting to some points in objects. Overall, MAAL fully considers the multi-modal inputs, leverages the AE pipeline, and formulates a novel framework for learning 3D articulated object affordance.

Our main contribution can be summarized as follows:

1. We propose a novel pipeline named Multimodality-Aware Autoencoder-based affordance Learning (MAAL). It is an efficient framework for solving the 3D object affordance problem. MAAL does not need multiple training steps and only requires a few data samples compared to previous methods.

2. We propose MultiModal Energized Encoder (MME) to handle the multi-modal information and their interaction in the 3D object affordance problem. The proposed encoder comprehensively learns data in all modalities and provides better multi-modal learning ability.

3. Without bells and whistles, our method outperforms all current methods in both F-score and sample success rate. Visualizations also show the effectiveness of our MAAL.

2. Related Work

3D object affordance: In the field of robotics, 3D object affordance is an important area of many practical applications. Before manipulating objects in reality, the robots need to understand what and where can be acted at first, which can be contributed to the exploration of affordance [10]. Recently, many works have emerged to explore this problem. [21] and [37] leverage the CNN network to produce the affordance area of the affordance map, which is used for indicating the grasping operations of robots. Jiang et al. [17] propose to constrain the consistency between hand contact points and object contact regions. The contact points of the robot hand are required to be close to the shape of the object's surface. Then, Mo et al. [28] provide a large-scale dataset and benchmark. The authors also predict affordance maps to indicate the actionability of robots at every point of objects. 3DAffordanceNet [7] explore another interesting problem and introduces a dataset for the functional understanding for 3D objects. Moreover, AdaAfford [45] goes further with the affordance predictions, considers the information hidden in the 3D shapes, and mines important kinematic and dynamic factors in 3D interactions. Through better modeling of the kinematic uncertainties, AdaAfford improves the performance of manipulating objects within fewer action steps. The significant advancements in [28] and [45] should be admired, but these works also contain defeats. All previous works utilize multiple decoders or critics to predict the probability of actionability (separately training three networks in [28] and four networks in [45]). The method design is complex and requires many data samples for training. In this work, we propose an AE-based pipeline to solve the problem efficiently.

Deep autoencoder: Deep autoencoder (AE) [11, 16, 42, 43] is a widely used structure for representation learning. It aims to represent and reconstruct the same inputs and is generally supervised by a reconstruction loss. It shows outstanding ability in representing and understanding high-dimension data. In this paper, we apply the idea of AE in learning 3D interaction, which can solve the 3D interaction problem in one go without training multiple decoders or critics in different steps.

Multi-modal Learning: Many tasks (e.g., VQA [3, 24], gesture generation [50, 25], video representation [23]) involve multi-modal inputs and require the network to handle the multi-modal problems [22, 44]. These problems usually entail the understanding of various knowledge [48] and require the proper handling of diverse inputs. Generally, the network needs to handle data samples with various modalities, which may possess different distributions and semantics. Methods usually need to fuse data or features for further learning. Formally, there are three kinds of strategies [22, 20] to fuse multi-modal data: early fusion, late fusion, and inter-media fusion. Early fusion means fusing data samples before specific learning. Methods [1, 22, 12] with early fusion usually combine raw data without considering the connection between data samples or fuse embedded features in low dimensional space. This strategy may be useful if the multi-modal data are conditionally independent [39, 32, 34]. However, the performances for highly correlated data samples or features would be lower [27]. Moreover, late fusion [20, 40, 46, 18] indicates the independent learning data sample before the last module, which is used for decision-making (e.g., classifier, retrieval projector). This leads the network can understand each modality better and avoid accumulating uncorrelated errors [36]. However, the advantages of late fusion in multi-modal tasks are insignificant [36, 12, 41] compared with early fusion. Finally, intermediate fusion [22, 5] is the most commonly used strategy in recent multi-modal learning. It flexibly fuses different data samples at different levels and designs explicit modules to model different modalities adaptively. Many works [49, 8, 19] with intermediate fusion achieve better performances in various multi-modal tasks. We propose a MultiModal Energized Encoder (MME) to provide better multi-modal learning for 3D object affordance considering intermediate fusion for modalities. The better design of the encoder module supports MAAL to achieve higher performances in affordance learning.

3. Preliminary

Following the problem settings in [28], the 3D affordance problem can be generally formulated as where and how to act for a given 3D object. During training, 3D object information and interactive points are given as inputs. The methods are required to produce actionability scores for corresponding points, action proposals, and success likelihoods for proposals, respectively.

In detail, each input sample involves four kinds of data: x_o, x_p, x_a, and x_h. Specifically, x_o indicates the 3D object information represented by the 3D point cloud. $x_o \in \mathbb{R}^{\mathcal{O}\times 3}$, where $\mathcal{O}$ is the dimension of point clouds. x_p is the interactive point, and $x_p \in x_o$. x_a means an interaction proposal and can be described by gripper orientation $x_a \in SO(3)$. Finally, given gripper orientation x_a, articulated object x_o, and point x_p to the simulator, x_h is the part motion. It can indicate whether the action is successfully manipulated or not after simulation.

In this task, methods are required to:

- Given an object (x_o) and interactive point (x_p), produce an actionability score ϕ.
- Given an object (x_o) and interactive point (x_p), produce an action proposal ρ.
- Given an object (x_o), interactive point (x_p), and action proposal (x_a), produce a success likelihood score σ.

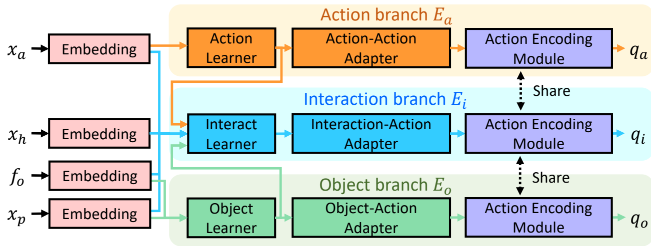

Figure 2. Structure of our MME. It contains three branches for learning different modalities. Features of different modalities with different levels are carefully fused in the interaction branch. MME provides better multi-modal learning for 3D object affordance. f_o is extracted by PointNet++ from x_o.

4. Method

We propose a Multimodality-Aware Autoencoder-based affordance Learning (MAAL) to solve the 3D object affordance problem. Specifically, MAAL contains three parts: a MultiModal Energized Encoder (MME), an action memory, and an action decoder. MME is proposed to learn multi-modal information, model the interaction and provide a comprehensive understanding of the inputs of the 3D object affordance problem. Then, action memory is used to record action information. Outputs from the encoder are taken as retrieval queries and are used to select items in the memory. Finally, given the aggregations of selected items from memory, the action decoder is proposed to reconstruct the corresponding actions.

4.1. MultiModal Energized Encoder

We propose MultiModal Energized Encoder (MME). MME empowers better multi-modal learning ability and solves the 3D affordance problem more effectively. Specifically, two kinds of modalities (object modality and action modality) and their interaction should be understood. Object modality mainly includes the point cloud of 3D objects and the points of the object for interaction. The action modality contains the gripper directions of the robot. Then, to model the interactions, object data, action data, and motion data from the simulator should be jointly considered. Although all the data are collected from the 3D space, there are still domain gaps among modalities: 1) Dimensional variations. The point cloud data in object modality has a dimension of $\mathbb{R}^{10000\times3}$. The gripper direction in robotic modality is a vector in $\mathbb{R}^{3\times3}$. 2) Physical property differences. Point cloud data are scalar values that indicate spatial information of objects. Robotic modality data are vectors and indicate the direction of the action. 3) Distinct networks in representation. Different encoders or embedding layers are utilized to process various inputs, resulting in features with varying distributions, further enlarging the gaps between modalities. In our work, as shown in Fig 2, rather than directly processing all modalities by early fusion, MME contains multiple branches of networks to handle different modalities and carefully fuses features to learn the interaction.

First, following [45, 28], we use PointNet++ [35] network to encode the 3D point cloud of the object into feature f_o, where $f_o \in \mathbb{R}^C$ and C is the dimension of the feature. Then four embedding layers are introduced to embed action x_a, motion x_h, object feature f_o, and point x_p, respectively. All embedding layers learn individually and are built by two fully-connected layers.

Then, as shown in Fig. 2, we have three branches to learn multi-modal features and their interaction separately: the action branch E_a, object branch E_o, and interaction branch E_i. Each branch contains a learner module and an adapter module. Learner modules aim to learn information, particularly for each modality and interaction. Then, the adapters convert features from learners to adapt the action encoding module. Different branches in MME help the network to learn affordance with different perspectives. The network is encouraged to mine valuable clues for object affordance from every modality separately. This leads to comprehensive multi-modal modeling and would not neglect any modalities.

Specifically, in the action branch, the action learner module is proposed to learn features after embedding and is constructed by three fully-connected layers. Similarly, in the object branch, the embedded features from f_o and x_p are given to an object learner module. The object learner contains a batch normalization layer and three fully-connected layers. Moreover, the interaction branch takes all information from modalities and aims to learn the interaction between objects and robots further. It contains a bilinear network to model the interaction between features from the action learner and object learner. A residual connection block is also involved in merging features from all modalities. This designation introduces the better ability for multimodal fusion [42, 49]. Features from different levels are considered and fused in the module. This provides a better understanding of information in multiple modalities.

Then, the adapters are introduced in the pipeline, which consists of two fully-connected layers. Finally, a shared encoding module generates query features from the different branches, denoted as q_a, q_o, and q_i, respectively. The procedure of MME can be formulated as follows:

$$q_a = E_a(x_a), \tag{1}$$
$$q_o = E_o(x_o, x_p), \tag{2}$$
$$q_i = E_i(x_a, x_o, x_p, x_h, \theta_a, \theta_o). \tag{3}$$

where θ_a and θ_o are the features extracted from the action learner and interact learner, respectively. The feature dimension of all queries is C. More details are presented in the supplementary.

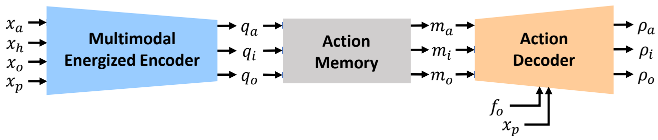

Figure 3. An overview of our Multimodality-Aware Autoencoder-based affordance Learning (MAAL). MAAL contains three parts: MultiModal Energized Encoder (MME), action memory, and action decoder. The encoder produces query feature q. The memory module receives queries, selects items, and aggregates them as m. Action decoder takes action information (f_o and x_p) and features m as inputs and reconstructs corresponding action x_a as ρ.

Moreover, other works directly use concatenated data (e.g., $[f_o, x_p, x_a, x_h]$ in [45], where $[*]$ is the concatenate operation.) as inputs. Taking all data as a whole, different modalities are learned equivalently. Comparatively, our encoder considers the learning of different modalities and their interaction. The encoder fuses multi-modal data at different levels and forms a comprehensive understanding. This leads our encoder to possess better multi-modal learning ability than the early fusion methods [28, 45].

4.2. Multimodality-aware Autoencoder-based Affordance Learning:

We propose Multimodality-Aware Autoencoder-based affordance Learning (MAAL). MAAL provides a more efficient pipeline to solve the affordance problem. As shown in Fig. 3, more than MME, we leverage a memory module M and a decoder module D to construct an AE pipeline. The memory module aims to prevent the "over-generalized" problem [11] in the original AE framework (only with an encoder and a decoder). Though only trained with positive samples, the original AE may also reconstruct negative samples with low reconstruction error during evaluation. By introducing a content-addressable memory, we do not directly provide encoder outputs to the decoder for reconstruction. The representation from the encoder is used as a query to retrieve the most relevant item in action memory. Then, the selected memory features are aggregated and provided to the MAAL decoder. The memory module is a widely used strategy in AE, which has been applied and discussed in many works [38, 33, 2].

As shown in Fig 3, given q_a, q_i, and q_o, the memory module addresses memory items and aggregates them as m_a, m_i, and m_o, respectively. $m_a = M(q_a)$, $m_i = M(q_i)$, and $m_o = M(q_o)$. Finally, the decoder network is introduced to reconstruct action information. Given object information (f_o and x_p), it reconstructs the actions ρ_o, ρ_a, and ρ_i according to features m_a, m_i, and m_o, respectively. $\rho_o = D(m_a, f_o, x_p)$, $\rho_a = D(m_a, f_o, x_p)$, and $\rho_i = D(m_i, f_o, x_p)$. To be noticed, the decoder network also takes object information as inputs. This is because the 3D affordance problem treats object information as known conditions. Under the real scenario, the robots have to know the object information and then produce actions to interact. Moreover, the decoder is constructed by two batch normalization layers and five fully-connected layers. More details will be offered in the supplementary.

Generally, MAAL is not expected to memorize and reconstruct the objects precisely. The memory module only needs to record and represent action information. Given features selected by queries, the decoder is responsible for reconstructing action information according to known object information.

4.3. Training and Evaluation

The overall loss function $\mathcal{L}$ can be formulated as follows:

$$\mathcal{L} = \|x_a - \rho_o\| + \|x_a - \rho_a\| + \|x_a - \rho_i\| \tag{4}$$

where $\|*\|$ indicates the ℓ_2 distances of input actions x_a and action proposals ρ from every branch. The overall training loss consists of reconstruction losses for three queries, respectively. Only a single and end-to-end training step is required in our work, as in Fig. 1.

During the evaluation, the final goal of the affordance problem requires predicting action proposal ρ by given object information, actionability score ϕ by given object information, and success likelihood score σ by given action proposal and object information. The action proposal can be directly produced by reconstruction result ρ_o in MAAL. However, ϕ and σ are hard to be obtained directly through MAAL. They can be estimated according to reconstruction errors. Meanwhile, the reconstruction error in MAAL is an absolute error [26], which indicates that it may be variant by different data splits. To overcome this problem, we additionally utilize the k-nearest-neighbor (KNN) algorithm to produce ϕ and σ.

In detail, we train the KNN algorithm using the average reconstruction error in the validation set. For every sample in the validation set, we have data x_a^{v}, x_o^{v}, x_p^{v}, and x_h^{v}, which indicate action, object, point, and motion data, respectively. Then, by MAAL, we achieve corresponding action proposals in the validation set, which are denoted as ρ_o^{v}, ρ_a^{v}, ρ_i^{v}. Thus, the reconstruction error e^{v} for a given sample in the validation set can be written as: $e^{\mathrm{v}} = (\|x_a^{\mathrm{v}} - \rho_o^{\mathrm{v}}\| + \|x_a^{\mathrm{v}} - \rho_a^{\mathrm{v}}\| + \|x_a^{\mathrm{v}} - \rho_i^{\mathrm{v}}\|)/3$. Then, we denote the KNN model as $\mathcal{K}$. $\mathcal{K}$ is trained by reconstruction error e^{v} from all the samples (including both positive and negative samples) and corresponding labels (binary labels indicate whether the actions can be successfully manipulated or not).

During the evaluation, we first achieve ρ_o^{t} by testing object data x_o^{t} and x_p^{t}. Then, the reconstructed action results of ρ_o^{t} can be calculated by:

$$m_a^{\mathrm{t}} = M(E_a(\rho_o^{\mathrm{t}})), \tag{5}$$

$$m_i^{\mathrm{t}} = M(\rho_o^{\mathrm{t}}, x_o^{\mathrm{t}}, x_p^{\mathrm{t}}, x_h^{\mathrm{t}}, E_a(\rho_o^{\mathrm{t}}), E_o(x_o^{\mathrm{t}}, x_p^{\mathrm{t}})), \tag{6}$$

$$\rho_a^{\mathrm{t}} = D(m_a^{\mathrm{t}}, x_o^{\mathrm{t}}, x_p^{\mathrm{t}}), \tag{7}$$

$$\rho_i^{\mathrm{t}} = D(m_i^{\mathrm{t}}, x_o^{\mathrm{t}}, x_p^{\mathrm{t}}). \tag{8}$$

where x_h^{t} is padded by zero. ρ_a^{t} and ρ_i^{t} are reconstruction results for ρ_o^{t} with action and interaction branches for testing. Then, for the current test sample, the actionability score $\phi = \mathcal{K}(\|\rho_o^{\mathrm{t}} - \rho_o^{\mathrm{t}}\| + \|x_o^{\mathrm{t}} - \rho_i^{\mathrm{t}}\|)/2)$. Similarly, for evaluating actions x_a^{t} in the test set, we can achieve reconstruction results ϱ_a^{t}, ϱ_i^{t}, and ϱ_a^{t} for x_a^{t}, respectively. Then, the success likelihood score can be computed as $\sigma = \mathcal{K}((\|x_a^{\mathrm{t}} - \varrho_a^{\mathrm{t}}\| + \|x_a^{\mathrm{t}} - \varrho_o^{\mathrm{t}}\| + \|x_a^{\mathrm{t}} - \varrho_i^{\mathrm{t}}\|)/3)$.

5. Experiment

In this section, we discuss all the details of our method design and task settings, evaluate our method with various metrics, and show the superiority and effectiveness of our work.

Implementation Details: Instead of training multiple critics and iterative training, all training procedures of our MAAL can be operated in one go. Specifically, the encoder, memory, and decoder modules are trained and updated at the same stage. Adam optimizer is used to optimize the networks within the learning rate 0.001 and weight decay 0.00001. More details about the network design will be presented in the supplementary. The memory module is implemented following [11], which has been widely used in many works [38, 33, 2]. We set memory size N as 200, and the dimension C is 128. Ablations will be offered in Sec. 5.1. Other settings (e.g., training data generation, gripper data processing, simulator settings, etc.) follow [45]. Additionally, during evaluation, the number of nearest neighbors of the KNN classifier is 500. Due to space limitations, more details of network designs and ablations will be offered in supplementary. We will also provide more details and update the results of real-world experiments on Github [1].

Datasets: We experiment with all methods and operate comparisons based on PartNet-Mobility dataset [30]. It is a large-scale and standard dataset for 3D articulated object affordance problems and has been widely used in previous works [28, 45, 53, 29]. The action simulation is operated through SAPIEN simulator [47]. In this dataset, 972 articulated 3D objects within 15 object categories are used for conducting 3D object affordance tasks. There are ten classes for training and five classes for testing. Besides, the validation set is also split and contains ten categories same as the training set. For better comparison, we separately report the results for shapes with training categories and shapes with unseen novel categories, which are marked as "train cat." and " test cat." in tables, respectively. The data split is constructed following [28, 45]. Moreover, the 3D articulated object affordance task has six pre-defined actions ("pushing", "pushing up", "pushing left", "pulling", "pulling up" and "pulling left"). For a fair comparison, categories are split into "pushing all" and "pulling all" actions following [28, 45]. All actions are parameterized in the $SE(3)$ space according to the robot gripper poses. Corresponding to the actions, the training and test data samples are generated by the simulator.

Dataset	Method	F-score (%)	Sample-Succ (%)
Pushing All (train cat.)	Where2Act [28]	66.29	27.33
	AdaAfford [45]	73.21	32.50
	MAAL	**76.63**	**34.25**
Pushing All (test cat.)	Where2Act [28]	52.38	21.04
	AdaAfford [45]	65.50	26.20
	MAAL	**69.88**	**28.34**
Pulling All (train cat.)	Where2Act [28]	48.76	6.40
	AdaAfford [45]	53.80	8.18
	MAAL	**59.26**	**10.47**
Pulling All (test cat.)	Where2Act [28]	40.88	5.71
	AdaAfford [45]	42.35	6.02
	MAAL	**43.57**	**6.67**

Table 1. The performance of the different methods for the 3D affordance problem in PartNet-Mobility dataset. Our method outperforms other methods in both data splits and metrics and also produces better action proposals than AdaAfford.

Moreover, we also apply settings in [45] to evaluate some special categories and further show the effectiveness. We sample data from the doors category from pulling actions and faucet categories from pushing actions following [45]. This data split further shows the ability of methods to handle kinematic ambiguity. Besides, we also visualize the actionability scores to plot affordance heatmaps following [28, 45], which further prove the effectiveness of MAAL.

Evaluation Metrics: To evaluate and compare methods, we apply the two standard metrics in the affordance task as in [28, 45], which are F-score for success likelihood score and sample-success-rate (Sample-Succ) for action proposals. Since the generated actions are randomly sampled, the positive and negative samples may not be balanced. Thus, in [28], the authors introduced an F-score to balance precision and recall for unbalanced samples. Then, Sample-Succ reflects the quality of proposals. It calculates the proportion of successfully manipulated actions among action proposals. Following [28, 45], we generate 100 candidates to compute the metric. We First select 100 points according to the actionability score ϕ in the given testing object. Then, we produce query q_o according to the object and point information and generate an action proposal. We experiment 10 times per testing object and report the average values of both metrics.

[1] https://github.com/akira-l/MAAL

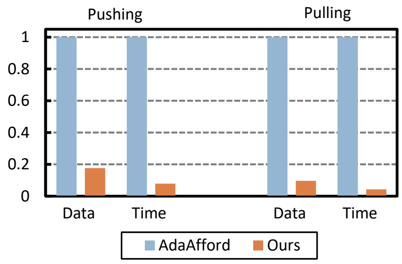

Figure 4. Comparison of data usage and training time. To better show the differences, we assume the data usage and training time of AdaAfford as 100% and calculate the relative percentages of MAAL compared with AdaAfford. Our method only consumes a small part of data samples and training times.

5.1. Results and Analysis

Comparisons with State-of-the-art Methods: As shown in Table 1, we first compare MAAL with previous works with four data splits following [45, 28]. Our method outperforms other methods in all data splits and metrics. The higher results reveal the effectiveness of our method. The comparison shows the advantages of our method in two aspects. The higher values of F-score indicate that our method assesses the actions better. This proves that the reconstruction error from MAAL works well for evaluating actions. Without any critics and multiple training stages, MAAL can perform and even better complete this task. Besides, MAAL also achieves better performances in Sample-Succ. This reveals that the quality of our proposals is also better than the previous works. Moreover, in another data split from [45], our method also achieves better results, as shown in Tab 2. The performance gain reveals the effectiveness of our MAAL in solving the kinematic ambiguity.

Statistic for Data Usage: Due to the properties of AE, our MAAL only takes the positive samples (successfully manipulated actions in simulation) as inputs. To show the efficiency of our data usage, we statistic the percentage of positive samples in all training data. We produce data samples following [28, 45] three times and calculate the average proportion. Comparatively, our method only uses positive samples and is more efficient. As shown in Fig. 4, Our method only takes 17.69% data of AdaAfford for training pushing action. Meanwhile, in pulling action, the positive samples are mere 9.63%, and our method only requires such limited data samples. Moreover, our method also possesses lower training time. We compute the average time of 100 training epochs of different methods, as in Fig. 4. Due to the training procedure with multiple stages and more data samples, the training time of AdaAfford is 23.34 and 12.72 times than ours. All these results show the efficiency of our method.

Comparisons with Different Action Proposals: To compare the quality of action proposals, we take action proposals and actionability scores from different methods separately and combine them for comparison. Specifically, as shown in Tab. 3, the action proposals are provided by different methods. Where2Act-P and AdaAfford-P indicate using the action proposal parts in these methods. Where2Act-C and Adaafford-C mean using critics in these works, which are responsible for predicting confidence for action proposals. The action proposal from MAAL can be directly achieved by ρ_o, and we score the action proposals by reconstruction errors as in 4.3. Then, we select the top-100 action proposals by corresponding scoring modules and compute the Sample-Succ of selected actions.

Dataset	Method	F-score (%)	Sample-Succ (%)
Pulling Door	Where2Act [28]	58.26	12.84
	AdaAfford [45]	69.34	17.62
	MAAL	**70.39**	**18.27**
Pushing Faucet	Where2Act [28]	78.14	36.35
	AdaAfford [45]	81.62	39.89
	MAAL	**81.82**	**40.06**

Table 2. Comparison of categories selected by [45]. MAAL still achieves better results in these relatively harder categories.

Given proposals from different methods, action selections by MAAL achieve a higher or comparable success rate compared with others. This indicates that MAAL possesses a high ability to assess and score actions compared with other methods. Besides taking proposals from MAAL, other methods also achieve better Sample-Succ values. The results further reflect that the proposal quality of our method is higher than others.

Ablation Study for the Multi-modal Learning: We compare different multi-modal learning as shown in Tab. 4. Experiments for using individual branches (only action branch, only object branch, and only interaction branch) and using the combinations of branches (action branch + object branch, action branch + interaction branch, and object branch + interaction branch) are provided.

Due to the comprehensive learning of multi-modal data, our method performs best among all the combinations. Learning with more modalities can improve the ability of the encoder. As in Tab. 4, the designation with only interaction outperform the designation with single modalities. Meanwhile, due to the intermediate fusion with other modalities, the interaction branch combines with another branch and outperforms the encoder only with the interaction branch. All the results prove the effectiveness of our method design. These may also reveal the necessity of multi-modal learning in 3D affordance. With better multi-modal learning, the network can better model and understand the affordance of a given object.

Furthermore, we modify our encoder with early fusion. We remain all three branches in the encoder but do not provide features from the action and object learner to the interaction branch. This leads the encoder to degrade to an early fusion-based method but still considers multi-modal learn-

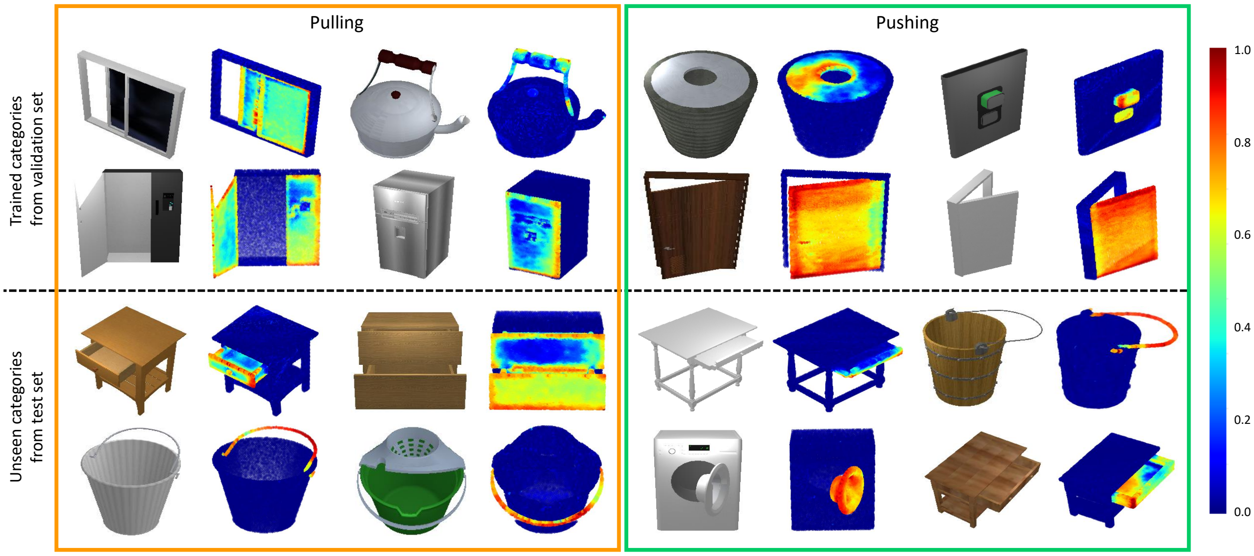

Figure 5. Visualization of affordance heatmap. All objects are from the test set. The heatmap is plotted by per-pixel action scores and produced by reconstruction error of action proposals from MAAL. Our method can effectively solve the 3D affordance problem and outperform the previous work.

Method		Sample-Succ (%)
Action Proposal	Actionability Score	
Where2Act-P [28]	Where2Act-C [28]	27.33
	AdaAfford-C [45]	28.58
	MAAL	28.67
AdaAfford-P [45]	Where2Act-C [28]	30.90
	AdaAfford-C [45]	32.50
	MAAL	32.36
MAAL	Where2Act-C [28]	31.50
	AdaAfford-C [45]	33.44
	MAAL	34.25

Table 3. Comparison of different combinations of methods. The higher performances prove that MAAL possesses a higher ability to evaluate actionability scores and generate high-quality proposals.

Multi-modal Learning Method	F-score (%)	Sample-Succ (%)
only action branch	32.47	13.54
only object branch	53.42	21.75
only interaction branch	58.74	24.01
action branch + object branch	59.87	23.88
action branch + interaction branch	73.26	32.55
object branch + interaction branch	75.54	33.89
All branches	**76.63**	**34.25**

Table 4. Combinations of learning different modalities. MAAL jointly considers object modality and action modality and further learn the interaction from both modalities. The comprehensive multi-modal learning by MAAL achieves better performance in the comparison.

ing. Then, the performance decreases by 8.31% in F-score compared with ours. All results reveal that our encoder is effective in multi-modal learning. The idea of intermediate fusion also improves learning ability.

5.2. Visualization for Affordance Predictions

We showcase the affordance predictions by heatmap as Fig. 5. The value of each pixel is calculated by the actionability score of MAAL following [45]. The visualized results show the effectiveness of MAAL in learning 3D object affordance. The actionable point in 3D objects can be correctly predicted by MAAL. Besides, we visualize different shapes with different categories from the validation set and test set in Fig. 5. For the unseen categories in the test set, our method can also understand the 3D object affordance and produce high confidence for actionable points. This further reveals the generalization of our MAAL.

6. Conclusion

This paper proposes a simple and data-efficient pipeline for the 3D affordance problem, named Multimodality-Aware Autoencoder-based affordance Learning (MAAL). MAAL contains three parts: MultiModal Energized Encoder(MME), action memory, and action decoder. We specifically design the encoder for multi-modal learning in 3D object affordance. The previous work usually directly applies early fusion to process multi-modal data. Comparatively, in our work, MME provides a comprehensive understanding of multi-modal learning and boosts the multi-modal learning ability for 3D affordance. In the experiment, the comparisons reveal the effectiveness of our method. MAAL outperforms former methods in different data splits, conditions, and metrics.

7. Acknowledgement

This work was supported in part by the Australian Research Council (ARC) under Grant DP200100938.

References

[1] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*, 2019.

[2] Peng An, Zhiyuan Wang, and Chunjiong Zhang. Ensemble unsupervised autoencoders and gaussian mixture model for cyberattack detection. *Information Processing & Management*, 59(2):102844, 2022.

[3] Yalong Bai, Jianlong Fu, Tiejun Zhao, and Tao Mei. Deep attention neural tensor network for visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–35, 2018.

[4] Jessica Borja-Diaz, Oier Mees, Gabriel Kalweit, Lukas Hermann, Joschka Boedecker, and Wolfram Burgard. Affordance learning from play for sample-efficient policy learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6372–6378. IEEE, 2022.

[5] Said Yacine Boulahia, Abdenour Amamra, Mohamed Ridha Madi, and Said Daikh. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6):121, 2021.

[6] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022.

[7] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[8] Changxing Ding and Dacheng Tao. Robust face recognition via multimodal deep face representation. *IEEE transactions on Multimedia*, 17(11):2049–2058, 2015.

[9] Ben Eisner, Harry Zhang, and David Held. Flowbot3d: Learning 3d articulation flow to manipulate articulated objects. *arXiv preprint arXiv:2205.04382*, 2022.

[10] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977.

[11] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019.

[12] Hatice Gunes and Massimo Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE international conference on systems, man and cybernetics*, volume 4, pages 3437–3443. IEEE, 2005.

[13] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 8(3):331–368, 2022.

[14] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *ACM Computing Surveys (CSUR)*, 54(3):1–35, 2021.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[16] Chaoqun Hong, Jun Yu, Jian Wan, Dacheng Tao, and Meng Wang. Multimodal deep autoencoder for human pose recovery. *IEEE transactions on image processing*, 24(12):5659–5670, 2015.

[17] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11107–11116, 2021.

[18] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550, 2013.

[19] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[20] Bahador Khaleghi, Alaa Khamis, Fakhreddine O Karray, and Saiedeh N Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, 14(1):28–44, 2013.

[21] Mia Kokic, Johannes A Stork, Joshua A Haustein, and Danica Kragic. Affordance detection for task-specific grasping using deep learning. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 91–98. IEEE, 2017.

[22] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[23] Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. Efficient multimodal fusion via interactive prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2023.

[24] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. Vrr-vg: Refocusing visually-relevant relationships. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10403–10412, 2019.

[25] Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. Seeg: Semantic energized co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10473–10482, 2022.

[26] OL Mangasarian and RR Meyer. Absolute value equations. *Linear Algebra and Its Applications*, 419(2-3):359–367, 2006.

[27] Héctor P Martínez and Georgios N Yannakakis. Deep multimodal fusion: Combining discrete events and continuous signals. In *Proceedings of the 16th International conference on multimodal interaction*, pages 34–41, 2014.

[28] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.

[29] Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2o-afford: Annotation-free large-scale object-object affordance learning. In *Conference on Robot Learning*, pages 1666–1677. PMLR, 2022.

[30] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019.

[31] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015.

[32] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 801–816. Springer, 2016.

[33] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381, 2020.

[34] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544, 2015.

[35] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[36] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

[37] Joseph Redmon and Anelia Angelova. Real-time grasp detection using convolutional neural networks. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1316–1322. IEEE, 2015.

[38] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

[39] Nicu Sebe, Ira Cohen, Ashutosh Garg, and Thomas S Huang. *Machine learning in computer vision*, volume 29. Springer Science & Business Media, 2005.

[40] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[41] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, 2005.

[42] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[43] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[44] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. In *NeurIPS*, 2022.

[45] Yian Wang, Ruihai Wu, Kaichun Mo, Jiaqi Ke, Qingnan Fan, Leonidas J Guibas, and Hao Dong. Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 90–107. Springer Nature Switzerland Cham, 2022.

[46] Di Wu, Lionel Pigou, Pieter-Jan Kindermans, Nam Do-Hoang Le, Ling Shao, Joni Dambre, and Jean-Marc Odobez. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1583–1597, 2016.

[47] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11097–11107, 2020.

[48] Yi Yang, Yueting Zhuang, and Yunhe Pan. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering*, 22(12):1551–1558, 2021.

[49] Dong Yi, Zhen Lei, and Stan Z Li. Shared representation learning for heterogenous face recognition. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–7. IEEE, 2015.

[50] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, 39(6), 2020.

[51] Jianbo Yu, Xiaoyun Zheng, and Shijin Wang. A deep autoencoder feature learning method for process pattern recognition. *Journal of Process Control*, 79:1–15, 2019.

[52] Chaoning Zhang, Chenshuang Zhang, Junha Song, John Seon Keun Yi, Kang Zhang, and In So Kweon. A survey on masked autoencoder for self-supervised learning in vision and beyond. *arXiv preprint arXiv:2208.00173*, 2022.

[53] Yan Zhao, Ruihai Wu, Zhehuan Chen, Yourong Zhang, Qingnan Fan, Kaichun Mo, and Hao Dong. Dualafford: Learning collaborative visual affordance for dual-gripper object manipulation. *arXiv preprint arXiv:2207.01971*, 2022.