Predicting Task-Driven Attention via Integrating Bottom-Up Stimulus and Top-Down Guidance

Zhixiong Nan, Jingjing Jiang, Xiaofeng Gao, *Graduate Student Member, IEEE*, Sanping Zhou, *Member, IEEE*, Weiliang Zuo, *Member, IEEE*, Ping Wei, *Member, IEEE*, and Nanning Zheng, *Fellow, IEEE*

***Abstract*— Task-free attention has gained intensive interest in the computer vision community while relatively few works focus on task-driven attention (TDAttention). Thus this paper handles the problem of TDAttention prediction in daily scenarios where a human is doing a task. Motivated by the cognition mechanism that human attention allocation is jointly controlled by the top-down guidance and bottom-up stimulus, this paper proposes a cognitively-explanatory deep neural network model to predict TDAttention. Given an image sequence, bottom-up features, such as human pose and motion, are firstly extracted. At the same time, the coarse-grained task information and fine-grained task information are embedded as a top-down feature. The bottom-up features are then fused with the top-down feature to guide the model to predict TDAttention. Two public datasets are re-annotated to make them qualified for TDAttention prediction, and our model is widely compared with other models on the two datasets. In addition, some ablation studies are conducted to evaluate the individual modules in our model. Experiment results demonstrate the effectiveness of our model.**

***Index Terms*— Human attention, task-driven.**

I. Introduction

NEARLY 10^8-10^9 bits data arrive human retina [1], [2] per second, conveying an overwhelming amount of information that is impossible for the human visual system to process all data equally. The attention mechanism filters relevant data from irrelevant noises, which essentially reduces the amount of data processing and enables the visual system to allocate limited neural computing resources to the most important parts, making it effortless to perceive and understand the visual world [3]. As the importance of attention, visual attention has been intensively studied in the computer vision community since the 1980s. However, the study mainly focuses on the task-free attention that is attracted by the salient stimuli in a bottom-up manner. Motivated by this situation, this paper handles the problem of predicting task-driven attention that additionally involves the top-down guidance.

The task is at the core of AI. Artificial General Intelligence (AGI) system factually targets to tackle a wide range of tasks in the human-like manner [4]. Some researchers argue that objects in the world are designed for humans to finish various tasks and satisfy various requirements. Actually, many studies [5]–[7] have demonstrated that a human's behavior and attention are, to a large extent, guided by the task. For example, under the guidance of the task 'taking the water', a human pays attention to the cup and water dispenser; under the guidance of the task 'sweep the floor', a human pays attention to the broom and dustpan. Therefore, it is significant to study task-driven attention.

One potential application is that a robot infers a human's attention so that to assist the human. For patients suffering from the spinal cord injury, it is difficult for them to perform some extremely-easy tasks in daily living. In this kind of case, if a robot could infer human attention using the camera configured on it, the robot is able to assist the patient to grasp, release, and move the attention objects. As shown in Fig. 1, in a human-robot coexistence scenario, the images captured from the robot's view are taken as the input of the model to predict human attention, based on which the robot could take action to assist the human. Targeting for this kind of application, this paper studies the task-driven attention prediction in daily living scenarios from the third perspective (i.e., robot's perspective). We believe this study will benefit various intelligent applications.

In the past decades, eye fixation and saliency are two synonyms of human attention, which mainly refers to the regions or objects that 'pop out' from the surrounding environments due to their salient features of color, size, motion, texture, edge, etc. Since this kind of attention is purely driven by the data in a bottom-up manner, it is called task-free attention (i.e., free viewing attention). Realizing the importance of tasks in guiding human attention, some researches have started to study Task-Driven Attention (TDAttention) in recent years [5], [8], [9], and the tasks are defined as searching for two specific object categories [8], multiple target objects [9] or playing the game [5]. Compared with current studies on TDAttention, our work presents two-fold differences. On the one hand, current studies focus on inferring TDAttention from the first perspective (i.e., the human's perspective) where the attention

Manuscript received December 29, 2020; revised July 22, 2021 and September 6, 2021; accepted September 7, 2021. Date of publication September 24, 2021; date of current version September 30, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 62006180 and Grant 62088102 and in part by the Fundamental Research Funds for the Central Universities under Grant sxxj022020039. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hairong Qi. *(Corresponding author: Nanning Zheng.)*

Zhixiong Nan, Jingjing Jiang, Sanping Zhou, Weiliang Zuo, Ping Wei, and Nanning Zheng are with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: nnzheng@mail.xjtu.edu.cn).

Xiaofeng Gao is with the Center for Vision, Cognition, Learning, and Autonomy, University of California at Los Angeles, Los Angeles, CA 90095 USA.

Digital Object Identifier 10.1109/TIP.2021.3113799

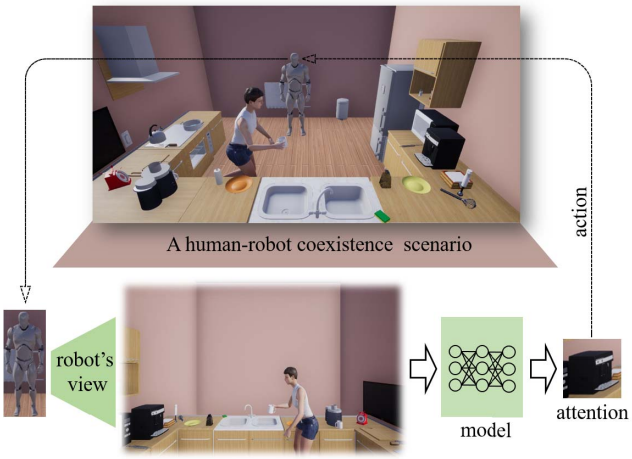

Fig. 1. The problem handled in this paper is predicting human attention from a robot's view, in a human robot-coexistence scenario where the human is doing a task. The potential application is that the robot predicts human attention and then takes action to assist the human.

is defined as the inside-image locations or objects that attract human attention, while our work infers TDAttention from the third perspective (i.e., the robot's perspective) as shown in Fig. 1. On the other hand, current studies mainly focus on a single task (e.g., visual searching), while our work predicts human attention in various tasks in daily life (e.g., make coffee, mop floor, take medicine, etc.).

From the perspective of methodology, though it is a common view that attention is simultaneously controlled by both bottom-up and top-down components [5], [10], [11], methods for both task-free attention and task-driven attention prediction usually ignore modeling top-down guidance. The main idea of task-free attention prediction methods is firstly constructing a neural network to predict a saliency map, and the network is then optimized to predict attention by minimizing the 'distance' of the predicted saliency map with the provided ground truth [12]–[14]. Methods for TDAttention prediction have not converged to a unified framework. For example, Zelinsky *et al.* [8] discretized the image into 160 grids and formulated attention prediction as a 160-class classification problem; Yang *et al.* [9] predicted attention by selecting maximally-rewarding eye fixation location via inverse reinforcement learning. In current studies, though some so-called top-down cues like object context and object semantic relationship have been considered, few works directly model the task to facilitate attention inference.

Based on the above observations, this paper proposes a deep task-driven attention prediction model that integrates the bottom-up stimulus with top-down guidance. Human pose and motion cues, which are tightly linked with attention, are fused with the raw RGB image cue to serve as the bottom-up stimulus. For the top-down task guidance, the coarse-grained and fine-grained task information are embedded as the top-down feature. Our proposed neural network model is composed of four modules, bottom-up feature fusion module (BU), top-down task guidance module (TD), spatial-temporal inference module (ST), and decoder module (DC). Given an image sequence as the input, the images are processed by the BU module to extract the bottom-up feature maps. In the TD module, the top-down feature is integrated with the bottom-up feature, and the task constraint is imposed to guide the network to update the integration feature under the guidance of the task. The TD module outputs a sequence of strengthened feature maps, which are further processed by the ST module to explore the relation of features in the temporal dimension. Finally, in the DC module, the feature maps are decoded as attention probability maps.

To test our model, two public datasets collected in real daily activities, including the CAD120 [15] and TIA [16] datasets, are selected for testing. In the experiments, six classic and state-of-the-art methods are compared with our method on the two datasets, and the experiment results show that our method outperforms other methods. Several ablation experiments are also conducted to test the individual modules of our model, and the experiment results validate the effectiveness of the individual modules.

The contributions of this paper are as follows: **1)** This paper predicts human attention from the third perspective, while most current studies predict human attention from the first perspective. **2)** This study infers human attention in various tasks in daily activities, while most current studies focus on a single task. **3)** This paper constructs a top-down and bottom-up collaborative mechanism to predict human attention. **4)** We manually annotated the attention objects in two public datasets to make them qualified for TDAttention prediction. The annotations of attention objects, the information of datasets and the code are available at *https://github.com/xjtu-nzx/Predicting-Task-Drivern-Attention*.

II. Related Work

A. A Brief History of Human Attention Study

Human attention study has a long history in diverse domains such as philosophy, psychology, cognitive science, and computer science. Dating back to the fourth century, attention is originally studied by philosophers, and they held the idea that attention is linked with awareness and consciousness [11]. Coming to the nineteenth century, attention begins to be studied in the psychology domain. James, known as the father of psychology, linked attention to the data compression and noted that attention should answer the question of 'what are the objects of interest' [17]. In the middle of the twentieth century, Broadbent, a well-known cognitive scientist, described attention as a 'early selection' mechanism that acts likes a filter to select the relevant information and discard the irrelevant information [18]. In contrast to the 'early selection' theory, Deutsch *et al.* proposed the 'late selection' model [19], with the idea that all information is acquired at the early stage and the relevant information is selected at the late stage. In the 1980s, the famous 'feature integration' theory was proposed by Treisman and Gelade [20], arguing that an object attracts human attention by the integration of features like color, size, and orientation.

Different from the above-mentioned domains, researchers in the computer science domain started to study attention at the end of twentieth century, and the studies focus on computational models rather than theory analysis. After widely reviewing the literature regarding attention in the computer

science domain, we summarize the related works from three different perspectives: 1) bottom-up and top-down, 2) pixel-level and object-level, and 3) first perspective and third perspective. The related works are detailed as follows.

B. Bottom-Up Attention and Top-Down Attention

Overall speaking, a considerable amount of works study bottom-up attention, while few works attempt to predict top-down attention. Bottom-up attention is based on the assumption that conspicuous visual features 'pop-out' from the background and involuntarily capture human attention. Top-down attention emphasizes the effect of high-level information (e.g., tasks and goals) on attention allocation.

For bottom-up attention, many excellent models have been proposed. In the early time, inspired by the 'feature integration' theory [20], many methods predict attention by integrating low-level features like color, edge, texture, intensity, orientation, and size. For example, in 1998, Itti *et al.* proposed a model that firstly extracts three low-level features (color, intensity, and orientation) from an image to generate three saliency feature maps, which are then fused to predict attention [21]. Before the deep learning era, these kind of methods are dominant in the computer vision community. Since 2012, with the rapid development of Deep Neural Network (DNN), DNN based methods [12], [13], [22]–[24] have replaced the traditional methods. The pipeline of DNN based methods is firstly predicting a saliency map using a DNN model, and then minimizing the loss functions that measure the 'distance' of the predicted saliency map with the provided ground truth map. Currently, the focus of research is exploring more effective network architectures and feature fusion mechanisms. For example, Liu *et al.* [23] adopted the multi-branch network architecture, and Wang and Shen [24] fused the feature maps in different layers to learn robust features. Recent approaches begin to take advantage of other modalities, such as auditory information [25]–[27], to model human attention.

For top-down attention, an early work [28] predicted the eye fixation locations of observers who performs the task of searching for people in images, and the attention map was estimated by integrating the output of three models, including saliency model, target appearance model, and scene context model. In 2012, Borji *et al.* [5] predicted human attention in the scenario that the human is engaged in video games. A Bayesian model fusing multi-modal information, including global context, previous saccades, and previous motor actions, was proposed to handle the problem. In recent years, studies concerning top-down attention mainly focus on the visual search task. For example, the task defined in [8] is searching images for two kinds of target objects, including microwave and clock; The task defined in [9] is searching for 18 categories of target objects. One common point of these works is that the top-down attention is defined as the inside-image objects or locations that are predicted from a human's view (i.e., first perspective) in the situation that a human is observing images. In this paper, top-down attention is defined as the objects that are predicted from a robot's view (i.e., third perspective) in human-robot coexistence scenarios. In addition, this paper considers diverse tasks involving in daily life, rather than a single task.

C. Pixel-Level Attention and Object-Level Attention

Eye fixation or saliency is the synonym of pixel-level attention, which is modeled as a heat map, and each pixel of the heat map has a probability of representing human attention. Saliency object is the synonym of object-level attention, which is modeled as a binary map, where pixels belonging to the attention object are assigned with the value of '1' while background pixels are assigned with the value of '0'.

Because the related works concerning pixel-level attention have been included in the above 'bottom-up attention' part, here we mainly review the works regarding object-level attention. One of the earliest works studying saliency objects was introduced by Liu *et al.* [29] in 2007, and the authors proposed a CRF model that integrates three kinds of features, including multi-scale contrast feature, center-surround histogram feature, and color spatial distribution feature. Inspired by the success of DNN models, many DNN based methods [30]–[32] are proposed for saliency object detection. The core idea is similar to that of pixel-level attention estimation, and the research point lies in exploring the effective network architecture to learn the informative feature representations. The disadvantage of these methods is that the network weakens the detailed information in the deep layers of the network. After 2016, the majority of saliency object detection methods are based on the Fully Convolutional Networks (FCN) [33]. Compared with DNN, the advantage of FCN is that it enables the network to recover spatial details of images, allowing the model to make use of both high-level semantic features and low-level fine details to detect saliency objects. Therefore, existing studies intensively explore the FCN based network architectures, such as the single-stream architecture [34], [35], and the multi-stream architecture [36]. In addition, the 'skip-connection' architecture is also widely used to fuse multi-scale feature maps to strengthen the feature representation [14], [37]–[39]. The recent approach proposed by [40] learns the residual of each side-output so that to refine the side-output for a fine-grained prediction. Some other efforts are also made in recent works. For example, the work [41] attempts to detect saliency objects while reducing the size of the training dataset; The work [42] targets to increase the speed of saliency object detection by modeling spatial-temporal interactions.

D. First Perspective Attention and Third Perspective Attention

When a human observer is observing an image, the locations or objects that attract the attention of the observer are defined as the first perspective attention. Comparatively, the third perspective attention is predicted using the images/videos captured from a robot's perspective (as illustrated in Fig. 1) or a fixed monitoring perspective.

Related works concerning the first perspective attention have been involved in the above parts, thus here we mainly review the studies of the third perspective attention. In the early time, the third perspective attention is studied in the relatively simple

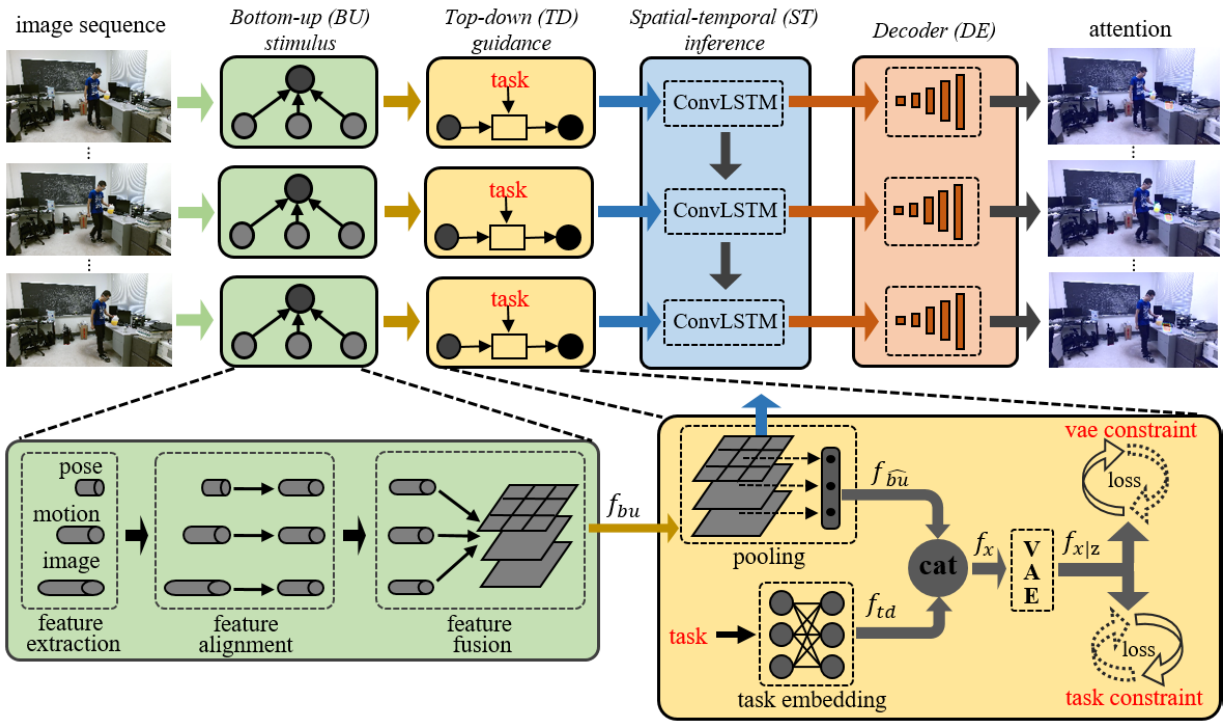

Fig. 2. The overview of our proposed model. The model takes an image sequence as the input, and outputs the attention probability maps of the input image sequence. Our model mainly consists of four modules. The *BU* module extracts the bottom-up features such as human pose and motion. The *TD* module integrates the task feature with bottom-up features, and imposes the task constraint on the integration feature to update the bottom-up feature map f_{bu} in the back-propagation manner. The *ST* module updates f_{bu} using the Convolution Long Short Term Memory network. The *DE* module up-samples f_{bu} to the attention probability maps.

scenarios where humans are restricted with limited head and body movements, and the detailed facial information (e.g., pupil, eye, and face) of a human is fully observed [43], [44]. The idea of mainstream methods is using the appearance features from the human pupil, eye, and/or face to regress a gaze direction [45], or using the appearance features to fit a known model to estimate human gaze [46]. To make the model to be applicable to complex scenes where humans are moving freely and the detailed facial information is not always available, Recasens *et al.* [47] proposed a more complex benchmark dataset named GazeFollow in 2015. In the following years, a series of excellent models are proposed [48]–[50]. For these studies, attention is mainly predicted in a bottom-up manner, and the top-down task is not directly involved. In recent years, another active research direction in third perspective attention is joint attention (i.e., shared attention) [51], [52], which is defined as the attention that multiple humans are paying to. The joint attention usually happens in social interaction scenarios involving more than one human subject.

III. Method

In this section, we introduce our method by starting with the overview of our proposed deep neural network model, followed by detailing the individual modules in the model.

A. Overview

As shown in Fig. 2, given an image sequence $\mathcal{I} = \{i_t | t = 1, 2, \ldots, T\}$ as input, the output is the human attention $\mathcal{A} = \{a_t | t = 1, 2, \ldots, T\}$, where a_t represents the human attention that corresponds to the image i_t. Our proposed model is composed of four modules, including *bottom-up stimulus (BU)* module, *top-down guidance (TD)* module, *spatial-temporal inference (ST)* module, and *Decoder (DE)* module. In the *BU* module, human pose and motion features are fused with the image feature to provide bottom-up cues for attention prediction. In the *TD* module, the task information is embedded as a feature to integrate with the bottom-up cues, and the task is also formulated as a constraint to guide the model to predict the task-driven attention. In the *ST* module, the bottom-up feature maps are inputted to ConvLSTM networks to learn the more robust feature representation by exploring the spatial-temporal relation. In the *DE* module, the feature maps are upsampled to probability maps that signal the possible locations of human attention.

B. Bottom-Up Stimulus

Previous studies have demonstrated that human attention is, to a large extent, stimulated by salient features in a bottom-up manner. Therefore, the skeleton and optical-flow features that signal human pose and motion are extracted from given images to serve as the bottom-up features. The motivation is that human pose and motion significantly indicate human attention. In general, when a human's hand is reaching to, or a human's body is approaching an object, the object is most likely to be the attention object. For example, a human intending to microwave food needs to walk (motion) to the microwave and open (pose) it, where the microwave is the attention object.

Let i_t be an input image at time t with the size of $3 \times H \times W$ (3 channels, H pixels in height, and W pixels in

width). To extract human pose feature, the human skeleton detector introduced in [53] is first used to detect human skeleton, which is then encoded as a $H \times W$ binary skeleton mask where human skeleton pixels are set as "1" and other pixels are set as "0". The human skeleton mask in i_t is denoted as s_t. To extract human motion feature, two consecutive images at time t and $t-1$ are used to compute optical flow feature map, which is with the size of $2 \times H \times W$ and denoted as m_t. By separately applying the resnet [54] on i_t, s_t and m_t, we obtain the image feature (f_i), human pose feature (f_p) and human motion feature (f_m) to represent the bottom-up cues. f_i, f_p, and f_m are with the same size of $512 \times \frac{H}{32} \times \frac{W}{32}$.

Since f_i, f_p, and f_m represent multi-modal features, to better utilize and fuse these features, we refer to the self-attention mechanism introduced in [55] to align the three features. The aligned features are fused by adding them up. The fusion feature serves as the bottom-up feature f_{bu}:

$$f_{bu} = f_v + f_s + f_m \tag{1}$$

C. Top-Down Guidance

The importance of high-level task information for guiding human attention and behavior in a top-down manner has been intensively emphasized in previous literature [4], [6], [10], [28]. Therefore, in our work, the task information is encoded to integrate with the bottom-up cues. A task is usually composed of several sub-tasks. As a result, all frames in a video has the same task label, but they may have different sub-task label. The task label is a coarse-grained cue, while the sub-task label is a fine-grained cue.

Given the task label and sub-task label of an image, the BERT model [56] is used to encode the two labels as two features, which are then fused as the top-down task feature f_{td} with the size of $512 \times 1 \times 1$. The BERT allows to explore the semantic relation of different words, contributing to construct a powerful feature representation to encode the task information. To fuse f_{td} with the bottom-up feature f_{bu} defined in Eq.1, an average pooling operator is firstly applied on f_{bu} to obtain a new bottom-up feature $f_{\hat{bu}}$ with the size of $512 \times 1 \times 1$. Then, the fusion feature f_x is computed as follows.

$$f_x = \mathcal{F}_c(f_{td}, f_{\hat{bu}}) \tag{2}$$

where $\mathcal{F}_c$ represents the concatenation operator.

To obtain a better representation of fusion feature, the VAE (Variational Autoencoder) model is adopted to firstly learn a latent feature f_z based on f_x. Then, the reconstruction feature $f_{x|z}$ is computed based on f_z:

$$f_{x|z}, \mu, \sigma = \mathcal{N}_{vae}(f_x) \tag{3}$$

where $\mathcal{N}_{vae}$ represents the VAE network, and μ and σ are the parameters to represent the Gaussian distribution of the latent feature f_z.

Two constraints are involved in our model to strengthen the feature representation and task guidance. The first is the reconstruction constraint that encourages the network to minimize the difference between f_x and $f_{x|z}$. The reconstruction constraint assists in learning a robust representation for the fusion feature to weaken the gap between top-down feature f_{td} (linguistic modality) and bottom-up feature $f_{\hat{bu}}$ (visual modality).

The second is the task constraint, which is based on the reconstruction feature $f_{x|z}$. By applying a classification network on $f_{x|z}$, a task is predicted:

$$y = \mathcal{N}_p(f_{x|z}) \tag{4}$$

where $\mathcal{N}_p$ represents the task prediction network, and y represents the predicted task. The task constraint encourages y to be identical with the ground truth, guiding the network to predict the task-driven attention.

We note that the above two constraints are not directly imposed on f_{bu} that is actually used for attention prediction. However, the backward propagation of the two constraints can update the parameters of the whole network so that f_{bu} is updated to convey both top-down and bottom-up information.

D. Spatial-Temporal Inference

Both spatial and temporal cues are essential for inferring human attention. Actually, human pose and motion in the temporal axis convey richer information than that in a single image. Therefore, a spatial-temporal inference module is used in our model, aiming at exploring the semantic context in both spatial and temporal dimensions.

Given the input image sequence $\mathcal{I} = \{i_t | t = 1, 2, \ldots, T\}$, the bottom-up features for individual images are firstly computed by Eq. 1, obtaining the bottom-up feature sequence $\mathcal{S} = \{f_{bu}^t | t = 1, 2, \ldots, T\}$. $\mathcal{S}$ is then taken as the input of the ConvLSTM (Convolution Long Short Term Memory) [57] network, which outputs a new feature sequence $\mathcal{S}'$:

$$\mathcal{S}' = \mathcal{N}_{st}(\mathcal{S}) \tag{5}$$

where $\mathcal{N}_{st}$ represents the spatial-temporal inference network.

E. Decoder

The decoder module takes $\mathcal{S}'$ defined in Eq. 5 as the input, outputting a sequence of pixel-level human attention probability map $\mathcal{A} = \{a_t | t = 1, 2, \ldots, T\}$:

$$\mathcal{A} = \mathcal{N}_{de}(\mathcal{S}') \tag{6}$$

$\mathcal{N}_{de}$ is a deconvolution network. The size of feature map in $\mathcal{S}'$ is $512 \times \frac{H}{32} \times \frac{W}{32}$, and the size of probability map in $\mathcal{A}$ is $1 \times H \times W$.

Compared with the pixel-level attention, object-level attention presents many advantages in human-robot interaction scenarios. Therefore, an object detector is applied on the input image sequence $\mathcal{I}$, obtaining the object detection result $\mathcal{O} = \{o_t | t = 1, 2, \ldots, T\}$, which is then integrated with $\mathcal{A}$ defined in the Eq. 6 to infer object-level human attention $\mathcal{A}_o$:

$$\mathcal{A}_o = \mathcal{F}_i(\mathcal{A}, \mathcal{O}) \tag{7}$$

where $\mathcal{F}_i$ represents inference function, which will be detailed in the following section. We note that our model outputs the human attention probability map $\mathcal{A}$ defined in Eq. 6 if the object detection is not applied.

IV. Learning and Inference

The process of learning is actually optimizing the loss function, which is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{overlap} + \lambda_2 \mathcal{L}_{pixel} + \lambda_3 \mathcal{L}_{vae} + \lambda_4 \mathcal{L}_{task} \tag{8}$$

where λ_1, λ_2, λ_3 and λ_4 are weights for the individual loss. For simplicity, the losses here are defined for a single image, and are detailed as follows.

Let $a \in [0, 1]^{1\times H\times W}$ be pixel-wise human attention probability map that is predicted by the model, $\tilde{a} \in \{0, 1\}^{1\times H\times W}$ be the ground truth of pixel-wise human attention map. $\tilde{a}$ is a binary map with attention region assigned with '1' and non-attention region assigned with '0'.

$\mathcal{L}_{overlap}$ is defined as:

$$\mathcal{L}_{overlap} = 1 - \frac{2||a \odot \tilde{a}||_1}{||\mathrm{a}||_1 + ||\tilde{a}||_1} \tag{9}$$

$\mathcal{L}_{overlap}$ represents the overlap of a and $\tilde{a}$, measuring the difference between a and $\tilde{a}$ from an overall perspective.

$\mathcal{L}_{pixel}$ is defined as:

$$\mathcal{L}_{pixel} = -\frac{\sum_{i,j}((1-\omega)\tilde{a}_{ij}\log a_{ij} + \omega(1-a_{ij})\log(1-a_{ij}))}{W \times H} \tag{10}$$

where ω represents the area ratio of attention region in $\tilde{a}$, a_{ij} is the $(i, j)^{th}$ element of a, and $\tilde{a}_{ij}$ is the $(i, j)^{th}$ element of $\tilde{a}$. $\mathcal{L}_{pixel}$ is a weighted cross entropy that measures the difference between a and $\tilde{a}$ from the pixel-wise perspective. Compared with the standard cross entropy, the weighted cross entropy presents the advantage in weakening the imbalance effect of attention region and non-attention region.

$\mathcal{L}_{vae}$ is defined as:

$$\mathcal{L}_{vae} = -\mathrm{f}\,(x|z)\cdot\log f\,(x) - (1-\mathrm{f}\,(x|z))\cdot\log(1-f(x)) - \frac{1}{2}\left(1+\log\sigma^2-\mu^2-e^{\sigma}\right) \tag{11}$$

where f_x is the fusion feature defined in Eq. 2, and $f(x|z)$, μ and σ are defined in Eq. 3.

$\mathcal{L}_{task}$ is defined as:

$$\mathcal{L}_{task} = \mathcal{F}_{ce}(y, \tilde{y}) \tag{12}$$

where y is the predicted task defined in Eq. 4, $\tilde{y}$ is the ground truth of task, and $\mathcal{F}_{ce}$ is the standard function to compute cross-entropy loss. The purpose of $\mathcal{L}_{task}$ is to encourage the model to predict task-driven attention by involving the top-down task guidance into the back propagation of the network.

During the inference, a new image sequence is processed by the four modules in our model, outputting a sequence of pixel-wise human attention probability maps as defined in Eq. 6. For simplicity, we take a single image i_t at time t in the sequence as the example, and the corresponding pixel-wise human attention probability map is denoted as a_t. To explain the inference function defined in Eq. 7, let $o_t = \{o_t^n | n = 1, 2, \ldots, N\}$ be N objects detected in i_t. The goal of the inference function is estimating an object that is most likely to be the attention object. The idea is computing the scores of objects in o_t and choose the object with the highest score. The score of an object o_t^n is computed by:

$$S(o_t^n) = \frac{\sum_{(i,j)\in o_t^n} a_t^{ij}}{\mathcal{F}_a(o_t^n)} \tag{13}$$

where a_t^{ij} is the $(i, j)^{th}$ element of a_t, $\mathcal{F}_a$ is the function to compute the area of o_t^n. Actually, $S(o_t^n)$ reflects the quantitative overlapping proportion of o_t^n and a_t.

We know that an object detector outputs the objects and their probabilities. Let T_o be the threshold of object detection. We set $T_o = 0.6$, which means that the objects whose probabilities exceed 0.6 will be taken $o_t = \{o_t^n | n = 1, 2, \ldots, N\}$. If T_o is set as a small value, then a large number of objects are detected, which increases the difficulty for our model to select the correct one as the attention object. In contrast, if T_o is set as a large value, a small number of objects are detected, leading to that the detected objects might not contain the attention object. Therefore, $T_o = 0.6$ is a tradeoff.

V. Experiments

A. Dataset

CAD-120 [15] and TIA [16] datasets are used for evaluating our model. For both datasets, third-person-view videos are collected from various daily-life scenes, including kitchen, office, corridor, classroom, lobby, elevator entrance, etc. The scenes are diverse in the illumination, furniture configuration, and object placement. The human subjects perform tasks freely, without motion or pose constraint. Therefore, the pose and motion present the diversity. In the CAD-120 [15] dataset, the same task share the same background while human subjects are different. In the TIA [16] dataset, some tasks have more than one background, and human subjects are different. The main challenge is that scenes are complex and open, making it difficult to locate attention objects from various objects in scenes. In addition, attention objects might be far from the human subject, or be partially overlapped with other objects, which further increases the challenge. The CAD-120 [15] dataset and the TIA [16] dataset present the differences in video numbers, tasks, and human subjects. Another difference is that the videos in the CAD-120 [15] dataset are captured by the cameras that are relatively near to human subjects, while the videos in the TIA [16] dataset are captured by the cameras with diverse distances to human subjects. In addition, the video resolution in the CAD-120 [15] dataset is 640×480, while it is 1920×1280 in the TIA [16] dataset.

The CAD-120 dataset [15] contains 124 videos of 4 human subjects performing 10 tasks, including *making cereal, taking medicine, stacking objects, unstacking objects, microwaving food, picking objects, cleaning objects, taking food, arranging objects, having a meal*. In each video, a human is performing a task that can be further decomposed into several sub-tasks. In each video frame, task-related objects, task label and sub-task labels are annotated. To make the dataset available for task-driven attention prediction, we additionally annotated attention objects in all video frames. After filtering some unqualified data and annotations, 116 videos covering 9 tasks

TABLE I
INFORMATION OF BASELINES

Num	Citation	Title	Year	Jour/Conf
1	TDOD [58]	What object should I use?-task driven object detection	2019	CVPR
2	HAI [59]	Learning to infer human attention in daily activities	2020	PR
3	WATL [47]	Where are they looking?	2015	NIPS
4	FHPE [60]	Fine-grained head pose estimation without keypoints	2018	CVPR
5	JFR [61]	Joint 3d face reconstruction and dense alignment with position map regression network	2018	ECCV
6	MSOD [62]	Multi-scale interactive network for salient object detection	2020	CVPR

are used in the experiments. To separate the human subjects in training set and testing set, the videos of 3 human subjects are used for training, and the videos of 1 human subject are used for testing.

The TIA dataset [16] contains 809 videos of 14 human subjects performing 14 tasks, including *sweep floor, mop floor, write on blackboard, clean blackboard, use elevator, pour liquid from jug, make coffee, read book, throw trash, microwave food, use computer, search drawer, move bottle to dispenser*, and *open door*. In each video frame, task-related objects and task label are annotated. We additionally annotated attention objects and sub-task labels in all video frames. After filtering some unqualified data and annotations, 780 videos are used in our experiments. We guarantee that the human subjects in the training set and testing set are not overlapping in all tasks, and the ratio of the training set and the testing set is 2.9:1.

B. Implementation Details

H and W involved in the model are set as 224. Therefore, the size of input image is $3 \times 224 \times 224$. If the size of an original image was not $3 \times 224 \times 224$, the image will be resized to $3 \times 224 \times 224$ at first. The cues of human skeleton s_t and optical flow m_t are extracted in the off-line manner. Image feature f_i, human pose feature f_p and human motion feature f_m are extracted using ResNet18 [54] in the on-line manner. The "BatchNorm2d" and "ReLU" operators are applied on f_i, f_p and f_m, guaranteeing that the minimum value of feature maps is 0. These features are with the same size of $512 \times 7 \times 7$. In the top-down guidance module, the dimension of the original task feature f_{td} from BERT model is 768. To fuse with the 512-dimension bottom-up feature $f_{\hat{bu}}$, f_{td} is converted to 512-dimension using fully connection layers. In the spatial-temporal inference module, we set $T = 8$. In the decoder module, five deconvolution layers are used to upsample the feature maps. The model is implemented with PyTorch. During the learning, $\lambda_1 = \lambda_2 = 1$. Because the magnitude of $\mathcal{L}_{vae}$ and $\mathcal{L}_{task}$ is larger than other losses, we set $\lambda_3 = \lambda_4 = 0.1$ to balance the effect of different losses. The total loss is used to train the network in the end-to-end manner. The videos in the datasets are preprocessed by a sliding window mechanism to trim each video into sequences, and each sequence contains T images.

C. Baselines

Six baselines from related fields are selected. We classify the six baselines into three categories, including task-driven attention /object detection baseline, third perspective attention estimation baseline, and saliency estimation baseline. Some information of six baselines are summarized in Tab. I.

1) Task-Driven Attention/Object Detection Baselines: The following two baselines are mostly similar to our work.

a) Task-driven object detection (TDOD) [58]: In this work, a set of objects in the image are firstly detected using a standard object detector. Then, a Gated Graph Neural Network (GGNN) model is proposed to estimate the probability of each object being preferred for each task. We note that the object ground truth is used to substitute the object detection. Thus the performance is supposed to be higher than that of the original method.

b) Human attention inference (HAI) [59]: In this work, the input is the image sequence and human skeleton cue. By involving the one-hot task encoding information in the model, attention objects are predicted.

2) Third Perspective Attention Estimation Baselines: Our model predicts human attention from the third perspective. The following three baselines reveal the third perspective of human attention by human gaze direction. Thus we select these models as the baselines.

a) Where are they looking (WATL) [47]: In this work, the authors propose a model that utilizes the input image and human head location in the image to predict human gaze direction.

b) Fine-grained head pose estimation (FHPE) [60]: In this work, the model takes the image as input, outputting the Euler angles (yaw, pitch, and roll) of the head to indicate human gaze direction.

c) Joint 3d face reconstruction (JFR) [61]: In this work, the model takes the raw image and human face bounding box as input, and the output is the dense (more than 40K) aligned face key points. These dense points are compared with a pre-trained model to compute the camera matrix, which is further combined with 68 key points on the face to estimate the gaze direction.

3) Saliency Estimation Baseline: Human attention is closely related to saliency. Therefore, we select a saliency estimation baseline.

a) Multi-scale salient object detection (MSOD) [62]: In this work, the model extracts multi-level features from an RGB image and integrates multi-scale information from a specific level to generate the final saliency prediction.

D. Metric

For human attention prediction methods, their outputs can be classified into four categories: 1) attention direction

TABLE II

COMPARISON WITH BASELINES ON THE CAD120 DATASET. T1 TO T9 CORRESPOND TO THE TASKS OF T1: ARRANGING OBJECTS, T2: CLEANING OBJECTS, T3: MAKING CEREAL, T4: MICROWAVING FOOD, T5: PICKING OBJECTS, T6: STACKING OBJECTS, T7: TAKING FOOD, T8: TAKING MEDICINE, AND T9: UNSTACKING OBJECTS

Methods	T1	T2	T3	T4	T5	T6	T7	T8	T9	Overall
WATL [47]	0.45	0.40	0.63	0.70	0.23	0.76	0.31	0.45	0.63	0.54
JFR [61]	0.30	0.74	0.82	0.59	0.41	0.71	0.36	0.80	0.80	0.66
FHPE [60]	0.19	0.80	0.67	0.73	0.31	0.51	0.48	0.55	0.57	0.59
HAI [59]	0.35	0.97	0.90	**0.95**	0.69	0.88	**0.89**	0.78	0.85	0.85
MSOD [62]	0.31	0.97	0.54	0.66	0.67	0.51	**0.89**	0.45	0.61	0.64
TDOD [58]	**0.81**	0.90	0.60	0.82	**0.88**	0.46	0.81	**0.85**	0.87	0.82
Our	0.37	**0.98**	**0.93**	0.92	0.67	**0.89**	**0.89**	0.84	**0.91**	**0.87**

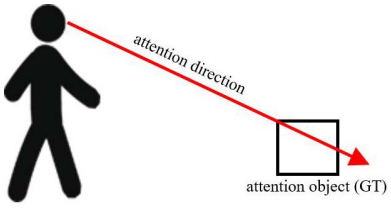

Fig. 3. Metric visualization.

(e.g., FHPE [60] and JFR [61]), 2) attention point (e.g., WATL [47]), 3) attention object (e.g., TDOD [58] and HAI [59]) and 4) attention probability map (e.g., MSOD [62]). An attention direction can be computed using an attention point or attention object. Note that an attention probability map can be combined with object detection to obtain an attention object, thus an attention direction can also be computed using an attention probability map. We can find that attention direction is the basic information that all methods share. Therefore, it is fair and comparable to use attention direction to evaluate the baseline methods. For all baseline methods, an attention direction is firstly computed based on their own output. As shown in Fig. 3, given a testing image, if the attention direction line intersects with the ground truth of the attention object, human attention in this image is counted to be correctly estimated. The ratio of correctly estimated images to all testing images is taken as the metric.

The output of FHPE [60] and JFR [61] is attention direction indicated by human head/face orientation. The output of these two baselines is directly used as the attention direction line shown in Fig. 3. The output of TDOD [58], HAI [59] and our model is human attention object. For these three methods, the line, which starts from the human head and passes through the center point of the predicted attention object, is used as the attention direction line. The output of WATL [47] is the attention point. The line, which starts from the human head and passes through the predicted attention point, is used as the attention direction line. The output of MSOD [62] is a saliency probability map, which is same as the attention map defined in Eq. 6. Therefore, the procedure to compute the attention direction line is same as that of our method.

E. Experiment Results and Analysis

Four kinds of experiments are conducted. Firstly, six baselines are compared with our model on the CAD-120 and TIA datasets. Secondly, top-down task guidance experiments are performed to evaluate the effects of different task encoding strategies and the performance improvement after applying the task-guidance module. Thirdly, bottom-up stimulus experiments are conducted to evaluate the effects of different bottom-up cues. Fourthly, different temporal encoding categories and temporal durations are evaluated to explore a robust spatial-temporal inference model. The results and analysis are detailed as follows.

1) Compare With Baselines: Table II shows the performance of the baseline models and our model on the CAD-120 dataset, and some qualitative examples are shown in Fig. 4. We compute the overall performance of each model on all tasks as well as the performance on the individual task. Our model outperforms all baselines on the overall performance, and achieves the highest performance on the majority of individual tasks. The advantage benefits from our proposed bottom-up and top-down integration model. On the one hand, the pose and motion cues are utilized in the bottom-up stimulus module. The pose and motion directly indicate the objects a human is reaching and moving to, and these objects tend to be the attention objects in common cases. For example, as shown in Fig. 4, many attention objects are revealed by the locations and orientations of the human's head, arms, and hands. On the other hand, the top-down task guidance module provides the task embedding information to assist the model to accurately locate the task-driven attention objects.

In Table II, we note the proposed model exhibits poor performance on the tasks of T1: arranging objects and T5: picking objects. The reason lies in the high failure of object detection. In these two tasks, attention objects are not detected in many video frames, so that object-level attention predictions in these frames are failed, even though pixel-level attention probability maps are correctly estimated. The failure cases will be further discussed in the following *Dicussion* section.

Table III shows comparison results on the TIA dataset, and some qualitative examples are shown in Fig. 5. Similar to the result on the CAD dataset, we can observe that our model also presents a distinct advantage on the overall tasks and the majority of individual tasks, demonstrating the general applicability of our proposed bottom-up and top-down integration model. As shown in Fig. 5, our model could successfully predict the attention objects that are far from the human and the attention objects with small sizes, even in the scenes with complex backgrounds and noisy objects. In task T4: clean board, our model behaves badly. The reason is the object 'eraser' in this task is not well detected.

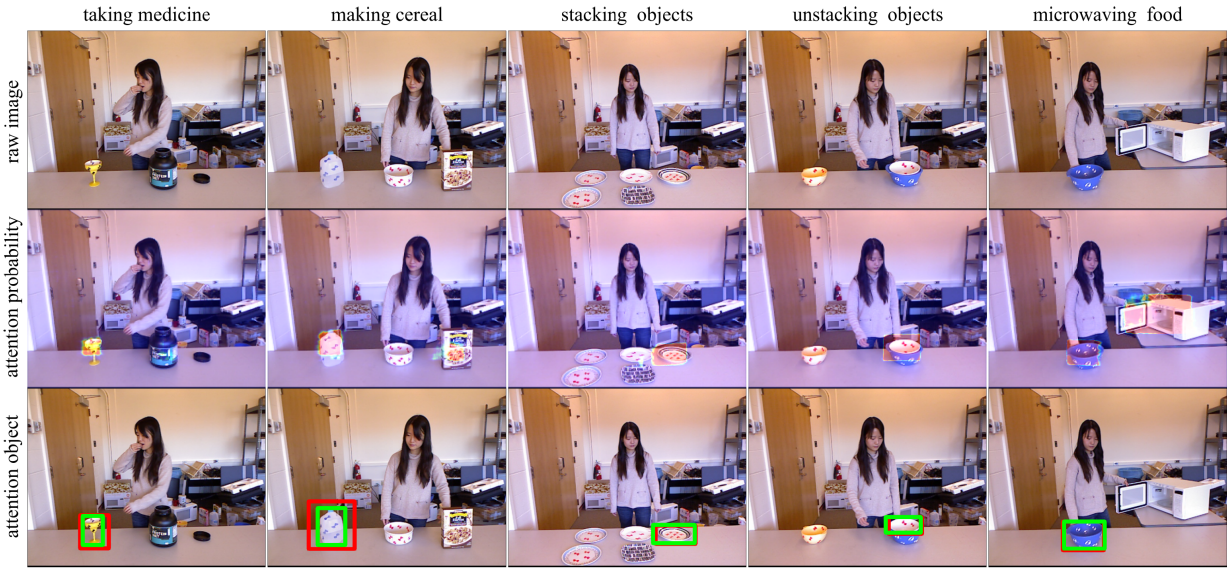

Fig. 4. The qualitative results of our model on the CAD dataset. Different columns correspond to the examples in different tasks. In each column, the raw image (top row), predicted attention probability map (middle row), predicted attention object (green box in bottom row) and ground truth attention object (red box in bottom row) are shown.

TABLE III

COMPARISON WITH BASELINES ON THE TIA DATASET. T1 TO T14 CORRESPOND TO THE TASKS OF TASKS. T1: SWEEP FLOOR, T2: MOP FLOOR, T3: WRITE ON BOARD, T4: CLEAN BOARD, T5: USE ELEVATOR, T6: POUR LIQUID, T7: MAKE COFFEE, T8: READ BOOK, T9: THROW TRASH, T10: HEAT FOOD, T11: USE COMPUTER, T12: SEARCH DRAWER, T13: MOVE BOTTLE, AND T14: OPEN DOOR

Methods	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	Overall
WATL [43]	0.46	0.27	0.45	0.68	0.80	0.32	0.37	0.64	0.57	0.79	0.11	0.52	0.63	0.24	0.51
JFR [61]	0.24	0.36	0.33	0.38	0.83	0.17	0.41	0.77	0.35	0.59	0.53	0.79	0.56	0.58	0.52
FHPE [60]	0.16	0.22	0.10	0.24	0.73	0.01	0.01	0.77	0.06	0.02	0.53	0.45	0.06	0.57	0.30
HAI [59]	0.68	0.70	0.51	0.23	0.68	0.65	0.91	**0.99**	**0.90**	0.94	**1.00**	**1.00**	**0.99**	**0.78**	0.80
MSOD [62]	**0.69**	0.60	0.41	0.20	0.85	0.78	0.63	**0.99**	0.57	0.90	0.41	**1.00**	0.79	0.64	0.68
TDOD [58]	0.28	0.84	**0.82**	**0.90**	0.90	**0.89**	0.88	0.90	0.85	0.90	0.91	0.89	0.89	0.03	0.84
Our	0.67	**0.90**	0.60	0.22	**0.96**	0.86	**0.96**	**0.99**	0.89	**0.97**	**1.00**	**1.00**	**0.99**	0.77	**0.86**

TABLE IV

EFFECT OF TASK ENCODING CATEGORIES

Encoding category	TIA	CAD120
sub-task	0.858	0.838
task	0.856	0.850
sub-task & task	**0.863**	**0.869**

TABLE V

EFFECT OF TOP-DOWN TASK GUIDANCE

Bottom-up cues	Task	TIA	CAD120
✓	×	0.832	0.823
✓	✓	**0.863**	**0.869**

2) Effect of Top-Down Task Guidance: Firstly, we are interested in studying how different task encoding categories affect the performance, and the experiment results are shown in Table IV. Three categories, including sub-task encoding, task encoding and sub-task&task encoding, are tested. The task encoding is a *coarse-grained* category that encodes top-down guidance using the task label, and all frames in a video share the same task label. The sub-task encoding is a *fine-grained* category that encodes top-down guidance using the sub-task label, and different frames in a video may have different sub-task labels. The sub-task & task encoding is a category that involves both *coarse-grained* and *fine-grained* task information. We can observe in Table IV that the sub-task & task encoding category outperforms the other two categories.

Secondly, we are interested in studying how much the task guidance module would improve the performance of human attention prediction. Therefore, we design an ablation experiment to test the performance in the following two cases: (1) only using the bottom-up cues and (2) integrating the bottom-up cues with the top-down task guidance. Table V shows the results on the TIA and CAD120 datasets. We can observe that, after adding the task guidance module, the performance of our model is significantly improved on both datasets. The result matches the cognition mechanism that high-level task information plays an important role in predicting human attention. The result also verifies the effectiveness of our proposed top-down task guidance module.

3) Effect of Bottom-Up Stimulus: In the bottom-up stimulus module, two cues are used, including human motion and human pose. We are interested in verifying how much these cues contribute to the performance improvement. Therefore, we design an ablation experiment. The experiment results are shown in Table VI. We can observe that our model achieves the highest performance when all bottom-up cues are used, demonstrating the effectiveness of each bottom-up cue. Another subtle phenomenon is that human pose cue seems to

Fig. 5. The qualitative results of our model on the TIA dataset. Different columns correspond to the examples in different tasks. In each column, the raw image (top row), predicted attention probability map (middle row), predicted attention object (green box in bottom row) and ground truth attention object (red box in bottom row) are shown.

TABLE VI
EFFECT OF BOTTOM-UP CUES

Motion	Pose	TIA	CAD120
×	×	0.845	0.815
✓	×	0.842	0.830
×	✓	0.848	0.836
✓	✓	**0.863**	**0.869**

TABLE VII
EFFECT OF DIFFERENT SPATIAL-TEMPORAL INFERENCE MODULES

Modules	TIA	CAD120
Conv3D	0.860	0.853
ConvGRU	0.855	0.842
ConvLSTM	**0.863**	**0.869**

TABLE VIII
EFFECT OF TEMPORAL DURATION

Duration	TIA	CAD120
1	0.845	0.828
3	0.857	0.847
5	0.843	0.834
8	**0.863**	**0.869**
10	0.859	0.850
15	0.855	0.860

be more helpful than human motion cue. Human pose conveys the position information of human body parts, including arms and hands. When an attention object is near to a human, the human usually uses hands to move to and operate on the attention objects, so that the attention objects are, to a large extent, revealed by human pose. In addition, the human pose potentially contains the information of gaze (i.e., where a human is gazing at). As we know, when an attention object is far from a human, the attention object is generally indicated by the human gaze. Different from the human pose, human motion emphasizes the information of human body moving direction, which is not that helpful than human pose.

4) Effect of Spatial-Temporal Inference: Besides the bottom-up stimulus module and top-down guidance module, our model also consists of another important module named the spatial-temporal inference module. The function of this module is to fuse the spatial information of different frames in an image sequence. Spatial-temporal inference can be implemented using different networks. To find an appropriate network, three different deep neural networks, including ConvLSTM, convGRU, and conv3D, are tested. As shown in Table VII, the convLSTM network exhibits best performance on both datasets.

An important factor influencing the performance of the spatial-temporal inference module is the temporal duration of an image sequence. To test the effect of the temporal duration, the performance of our model with different durations are computed on both datasets. The results are shown in Tab. VIII. We note that setting duration as 1 equals to disabling the spatial-temporal inference. We can observe that our model achieves the best performance when the duration is 8, and too long or too short duration leads to the degradation of performance. The potential reason is that the temporal relation can hardly be learned from an image sequence with the short duration. A long duration will inevitably involve more complex relations that are difficult to learn. In addition, too long duration involves more parameters of the module, making it difficult to train the model.

F. Discussion

1) Attention Object Is Not Detected: For attention objects that are heavily occluded, easily confused with background, extremely small, or with sparse training samples, they might not be detected. In this kind of case, our model outputs none attention object (as shown in Fig. 6(a)) or false attention object (as shown in Fig. 6(b)). We can observe the predicted pixel-level probability maps are correctly estimated, but the attention object predictions are failed due to the failure of

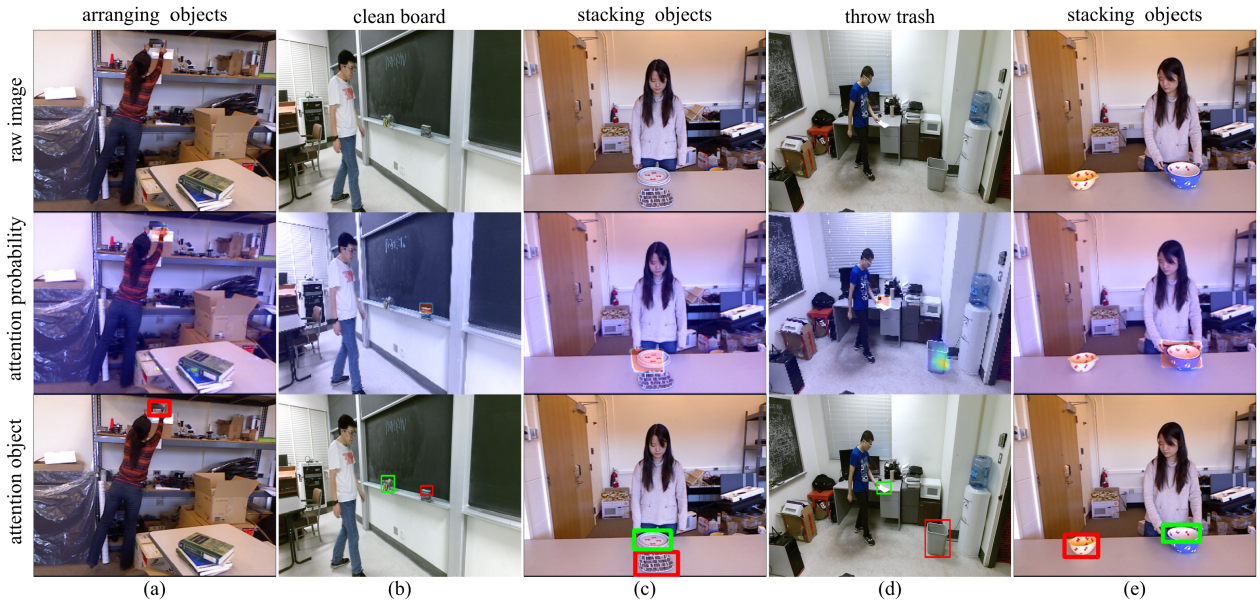

Fig. 6. Typical failure cases. (a) and (b) are the examples that the failure object detection leads to the failure of attention object prediction. (c) is the example that the information in the image is not enough to confirm the attention objects when multiple objects present high probabilities of being attention object. (d) and (e) are the examples that human pose falsely indicates the attention objects, when the ground truth attention objects are far from the human. Green boxes are predicted attention objects and red boxes are ground truth attention objects.

object detection. We note that the focus of this paper is not object detection, thus only a commonly-used object detector is adopted. If the object detector is optimized according to the scenarios, the performance will be further improved.

2) Attention Object Is Near to or Overlapped With Other Objects: As shown in Fig. 6(c), both plates exhibit the high probabilities of being the attention object. This is a difficult situation because the bottom-up cues (i.e., human pose and motion) cannot obviously indict which plate is the attention object, thus our model outputs an inaccurate attention probability map. Actually, the top-down task information is helpful in this kind of situation. Our proposed top-down guidance module benefits to improve the attention prediction performance as shown in Tab. V, but it does not work in all scenarios, and we hope this case could inspire researchers to explore a more effective way to make use of the top-down task information.

3) Attention Object Is Misguided By Human Hands: In most cases, human hands, as a part of human pose cue, directly reveal the attention objects. However, in some complex scenarios (e.g., attention object is far from human), human hand cue is sometimes hysteretic when predicting the attention object. In contrast, the attention object is indicated in advance by the human body moving direction or human gaze. As shown in Fig. 6(d) and (e), the ground truth attention objects (red boxes) are not the objects that are near to human hands, but the objects the human is gazing at or moving to. Our model outputs the wrong prediction. One main reason is that the model learns to fit the majority of data in a dataset, while the cases shown in Fig. 6(d) and (e) account for a few proportions in the dataset.

VI. Conclusion and Future Work

The problem handled by this paper is motivated by the application of human attention prediction in a human-robot coexistence scenario where the robot predicts human attention objects to assist disabled or injured people to do some simple tasks like grasping and moving objects. Inspired by the cognition mechanism that human attention is controlled by both bottom-up and top-down information, the top-down task guidance is computationally modeled to integrate with the bottom-up cues to predict task-driven attention. Many experiments are conducted to evaluate our model, and the experiment results demonstrate that the proposed model exhibits a significant advantage in performance, robustness, and general applicability on two public datasets.

Several important conclusions are drawn through this study. We list the conclusions here and hope they could benefit the related studies in the community. First of all, the task is the core of daily human activities. As long as a human is not insensible, the human always allocates the attention under the guidance of the intended task in mind. Therefore, it is of tremendous research significance to study task-driven attention. Secondly, for task-driven attention, an object with salient color, texture, size, or motion might not be an attention object. Conversely, it is to a large extent indicated by task-related information and human own information such as human pose and motion. Thirdly, the task is abstract and difficult to computationally modeled. Though our paper proposes a task encoding method and fuses the task encoding feature with the bottom-up cues, there is still a long way to explore more effective and cognitively-explanatory models. Fourthly, when encoding task information, the category fusing both the task and sub-task encoder behaves better than that encodes the individual task or sub-task because it utilizes both the coarse-grained and fine-grained task information. Fifthly, for the bottom-up cues, the human pose presents to be effective because it contains the position information of human body parts and indicates human gaze, and these information are important for predicting task-driven attention objects.

The goal of this study is to build a human-like and cognitively-explanatory model to predict TDAttention. Though the idea of the proposed model is identical with the human attention allocation mechanism that is controlled by both top-down guidance and bottom-up stimulus, there is still a long way to develop more effective and explanatory models. In a scenario where a human is doing a task, the semantic relations between the human, objects, and the scene covey informative cues for attention prediction. Therefore, our future work would model the scenario semantic relations to enrich the bottom-up module. In addition, future work will also focus on how to involve more top-down task information and seamlessly integrate it with the bottom-up cues.

REFERENCES

[1] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[2] K. Koch *et al.*, "How much the eye tells the brain," *Current Biol.*, vol. 16, no. 14, pp. 1428–1434, Jul. 2006.

[3] M. Carrasco, "Visual attention: The past 25 years," *Vis. Res.*, vol. 51, no. 13, pp. 1484–1525, 2011.

[4] K. R. Thórisson, J. Bieger, T. Thorarensen, J. S. Sigurðardóttir, and B. R. Steunebrink, "Why artificial intelligence needs a task theory," in *Proc. Int. Conf. Artif. General Intell.* Cham, Switzerland: Springer, 2016, pp. 118–128.

[5] A. Borji, D. N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 470–477.

[6] A. L. Yarbus, "Eye movements and vision," *Quart. Rev. Biol.*, 1967.

[7] G. T. Buswell, *How People Look at Pictures: A Study of the Psychology and Perception in Art*. Chicago, IL, USA: Univ. Chicago Press, 1935.

[8] G. Zelinsky *et al.*, "Benchmarking gaze prediction for categorical visual search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–9.

[9] Z. Yang *et al.*, "Predicting goal-directed human attention using inverse reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 193–202.

[10] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Rev. Neurosci.*, vol. 3, no. 3, pp. 201–215, 2002.

[11] M. Mancas, V. P. Ferrera, N. Riche, and J. G. Taylor, *From Human Attention to Computational Attention*, vol. 2. New York, NY, USA: Springer, 2016.

[12] J. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, and N. E. Oconnor, "Shallow and deep convolutional networks for saliency prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 598–606.

[13] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu, "DeepFix: A fully convolutional neural network for predicting human eye fixations," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4446–4456, Sep. 2017.

[14] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7264–7273.

[15] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *Int. J. Robot. Res.*, vol. 32, no. 8, pp. 951–970, Jul. 2013.

[16] P. Wei, Y. Liu, T. Shu, N. Zheng, and S.-C. Zhu, "Where and why are they looking? Jointly inferring human attention and intentions in complex tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6801–6809.

[17] H. Spencer, *The Principles of Psychology*, vol. 1. New York, NY, USA: Appleton, 1895.

[18] D. E. Broadbent, "A mechanical model for human attention and immediate memory," *Psychol. Rev.*, vol. 64, no. 3, p. 205, 1957.

[19] J. A. Deutsch and D. Deutsch, "Attention: Some theoretical considerations," *Psychol. Rev.*, vol. 70, no. 1, p. 80, 1963.

[20] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[21] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[22] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 262–270.

[23] N. Liu, J. Han, T. Liu, and X. Li, "Learning to predict eye fixations via multiresolution convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 2, pp. 392–404, Feb. 2018.

[24] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[25] A. Tsiami, P. Koutras, and P. Maragos, "STAViS: Spatio-temporal AudioVisual saliency network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4766–4776.

[26] X. Min, G. Zhai, J. Zhou, X.-P. Zhang, X. Yang, and X. Guan, "A multimodal saliency model for videos with high audio-visual correspondence," *IEEE Trans. Image Process.*, vol. 29, pp. 3805–3819, 2020.

[27] G. Wang, C. Chen, D.-P. Fan, A. Hao, and H. Qin, "From semantic categories to fixations: A novel weakly-supervised visual-auditory saliency detection approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15119–15128.

[28] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva, "Modelling search for people in 900 scenes: A combined source model of eye guidance," *Vis. Cognit.*, vol. 17, nos. 6–7, pp. 945–978, 2009.

[29] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[30] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3183–3192.

[31] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1265–1274.

[32] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 660–668.

[33] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[34] P. Hu, B. Shuai, J. Liu, and G. Wang, "Deep level sets for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2300–2309.

[35] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 212–221.

[36] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4019–4028.

[37] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 202–211.

[38] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3203–3212.

[39] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8150–8159.

[40] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, and Y. Fu, "Reverse attention-based residual network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3763–3776, 2020.

[41] Z. Wu, S. Li, C. Chen, A. Hao, and H. Qin, "A deeper look at image salient object detection: Bi-stream network with a small training dataset," *IEEE Trans. Multimedia*, early access, Dec. 23, 2020, doi: 10.1109/TMM.2020.3046871.

[42] C. Chen, G. Wang, C. Peng, Y. Fang, D. Zhang, and H. Qin, "Exploring rich and efficient spatial temporal interactions for real-time video salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3995–4007, 2021.

[43] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: Passive eye contact detection for human-object interaction," in *Proc. 26th Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2013, pp. 271–280.

[44] K. A. F. Mora, F. Monay, and J.-M. Odobez, "EYEDIAP: A database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras," in *Proc. Symp. Eye Tracking Res. Appl.*, Mar. 2014, pp. 255–258.

[45] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4511–4520.
[46] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling, "A 3D morphable eye region model for gaze estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 297–313.
[47] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 199–207.
[48] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 162–175, Jan. 2019.
[49] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 51–60.
[50] S. Park, A. Spurr, and O. Hilliges, "Deep pictorial gaze estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 721–738.
[51] L. Fan, Y. Chen, P. Wei, W. Wang, and S.-C. Zhu, "Inferring shared attention in social scene videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6460–6468.
[52] O. Sumer, P. Gerjets, U. Trautwein, and E. Kasneci, "Attention flow: End-to-end joint attention estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3327–3336.
[53] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.
[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[55] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[57] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.
[58] J. Sawatzky, Y. Souri, C. Grund, and J. Gall, "What object should I use?—Task driven object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7605–7614.
[59] Z. Nan *et al.*, "Learning to infer human attention in daily activities," *Pattern Recognit.*, vol. 103, Jul. 2020, Art. no. 107314.
[60] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2074–2083.
[61] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3D face reconstruction and dense alignment with position map regression network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 534–551.
[62] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9413–9422.

Zhixiong Nan received the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2019. From 2017 to 2019, he was a Joint Ph.D. Student with the Center for Vision, Cognition, Learning, and Autonomy (VCLA), University of California, Los Angeles (UCLA). He is currently an Assistant Professor with the College of Artificial Intelligence, Xi'an Jiaotong University. His research interests include human attention estimation, human intention prediction, and autonomous cars.

Jingjing Jiang received the B.S. degree in mechanical engineering from Xi'an Jiaotong University in 2017, where she is currently pursuing the Ph.D. degree with the Institute of Artificial Intelligence and Robotics. Her research interests include computer vision and deep learning.

Xiaofeng Gao (Graduate Student Member, IEEE) received the B.E. degree in electronic engineering from Fudan University in 2017. He is currently pursuing the Ph.D. degree with the Department of Statistics, University of California, Los Angeles. His research interests include human–machine interaction and computer vision.

Sanping Zhou (Member, IEEE) received the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China, in 2020. From 2018 to 2019, he was a Visiting Ph.D. Student with the Robotics Institute, Carnegie Mellon University. He is currently an Assistant Professor with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include machine learning, deep learning, and computer vision, with a focus on medical image segmentation, person re-identification, salient object detection, image classification, and visual tracking.

Weiliang Zuo (Member, IEEE) received the B.E. degree in electrical engineering and the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2010 and 2018, respectively.

From 2016 to 2017, he was a Visiting Ph.D. Student with the School of Electrical and Computer Engineering, Georgia Institute of Technology (Georgia Tech). He is currently an Assistant Professor with the College of Artificial Intelligence, Xi'an Jiaotong University. His current research interests include array and statistical signal processing, pattern recognition, and medical imaging processing.

Ping Wei (Member, IEEE) received the B.E. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China. From 2016 to 2017, he was a Postdoctoral Researcher with the Center for Vision, Cognition, Learning, and Autonomy (VCLA), University of California, Los Angeles (UCLA). He is currently an Associate Professor with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include computer vision, machine learning, and computational cognition.

Nanning Zheng (Fellow, IEEE) received the Ph.D. degree from Keio University, Japan, in 1985. He is currently a Professor and the Director of the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests include computer vision, pattern recognition, computational intelligence, image processing, and hardware implementation of intelligent systems. He became a member of the Chinese Academy of Engineering in 1999. Since 2000, he has been the Chinese Representative on the Governing Board of the International Association for Pattern Recognition.