Pattern Recognition 103 (2020) 107314

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Learning to infer human attention in daily activities

Zhixiong Nan[a], Tianmin Shu[b], Ran Gong[b], Shu Wang[b], Ping Wei[a,*], Song-Chun Zhu[b], Nanning Zheng[a]

[a] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, PR China
[b] University of California, Los Angeles, Los Angeles, CA 90024, USA

ARTICLE INFO

Article history:
Received 10 August 2019
Revised 18 February 2020
Accepted 24 February 2020
Available online 26 February 2020

Keywords:
Human attention
Deep neural network
Attentional objects

ABSTRACT

The first attention model in the computer science community is proposed in 1998. In the following years, human attention has been intensively studied. However, these studies mainly refer human attention as the image regions that draw the attention of a human (outside the image) who is looking at the image. In this paper, we infer the attention of a human inside a third-person view video where the human is doing a task, and define human attention as attentional objects that coincide with the task the human is doing. To infer human attention, we propose a deep neural network model that fuses both low-level human pose cue and high-level task encoding cue. Due to the lack of appropriate public datasets for studying this problem, we newly collect a video dataset in complex Virtual-Reality (VR) scenes. In the experiments, we widely compare our method with three other methods on this VR dataset. In addition, we re-annotate a public real dataset and conduct the extensional experiments on this real dataset. The experiment results validate the effectiveness of our method.

1. Introduction

Attention is an important topic in the computer vision field. In the past 20 years, saliency map estimation [52] and saliency object estimation [51] are two intensively-studied problems concerning visual attention, with the goal of estimating saliency regions or saliency objects in an image that draw the attention of the human (*outside the image*) who is looking at the image. In this paper, we study the attention of the human (*inside a video*), we call it Inside-video human attention.

Saliency-based visual attention has wide applications in video tracking [17], image retrieval [30], and scene rendering [14], while the main advantage of inside-video human attention estimation lies in its significance for human-robot interaction, which has promising applications in various facets of society like the elderly care [47], education [25], and military [13]. In a typical human-robot interaction scenario in daily life, a robot is installed with a camera capturing a video, inside which a human is performing daily activities. In this kind of scenarios, inferring human attention from the robot's view equals to inferring the attention of a human inside a video (Inside-video human attention). Elderly care is a potential and valuable application of human-robot interaction. As we know, it is laborious for the elderly people to perform some simple activities such as open the refrigerator, lift a cup, and move a bottle. To enable the robot to assist the human, it is necessary for the robot to infer human attentional objects. For example, a human is going to take an apple from a refrigerator, when the human is approaching the refrigerator, the robot could infer that human attentional object is the refrigerator, so that the robot could assist the human to open the refrigerator door in advance.

To infer human attention, the foremost thing is to make clear what the human attention is. Originally, attention is a concept in philosophy. Nowadays, it is well known as a concept in psychology. One dominant definition in psychology is that attention is the process of attending to objects [42]. This definition indicates that the attention is based on objects. Actually, some studies [7,8,37] in psychophysics and biology fields as well as some inter-discipline studies in neuro image filed [63] and brain image field [34] also claim the object-based attention. Especially, Chou [8] provides the evidence of object-based attention. These studies provide the strong theory support for defining human attention as objects. Another widely accepted definition in psychology is that attention is something that happens in the mind - a mental "inside" which is linked with the perceivable "outside" [43]. This definition indicates that attention is related with the high-level mental information in human mind. When a human is doing a task, the task is a kind of high-level information in the mind, guiding human attention. For example, the juicer tends to draw human attention in the task of

* Corresponding author.
E-mail addresses: nanzhixiong@stu.xjtu.edu.cn, 729887877@qq.com (Z. Nan), pingwei@mail.xjtu.edu.cn (P. Wei).

https://doi.org/10.1016/j.patcog.2020.107314
0031-3203/

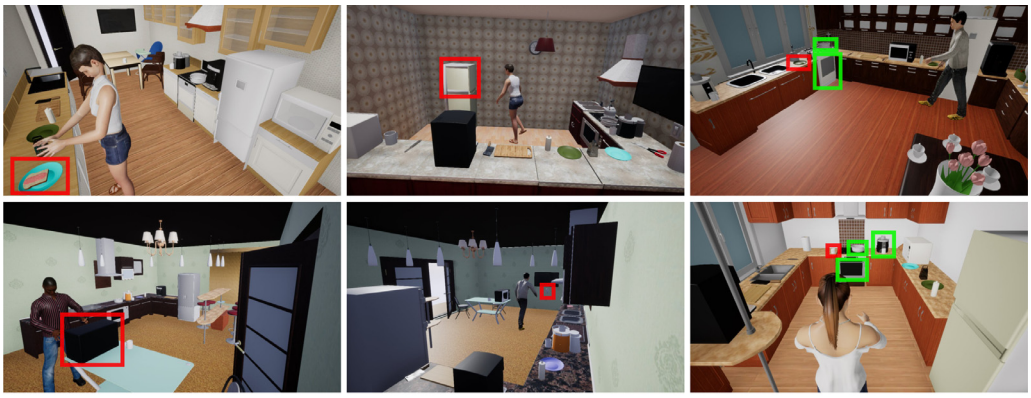

Fig. 1. The attentional objects (denoted as red bounding boxes) in three typical situations. (a) Easy situation where human gaze or human pose significantly indicates the attentional objects. (b) Moderate situation where human gaze is not available but human pose conveys the sufficient information for inferring the attentional objects. (c) Hard situation where the attentional objects can not be estimated only depending on human pose and human gaze because both cues indicate multiple possible attentional objects (denoted as green bounding boxes). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

"make juice", while the coffee machine tends to draw human attention in the task of "make coffee".

Based on these studies, we define human attention as the attentional objects that coincide with the task a human is doing. With a task in the mind, a human finishes the task by doing several sub-tasks in certain temporal order. For example, when a human is doing the task of "take the water from the drinking fountain", the human firstly finds the cup, then goes to the drinking fountain, and finally takes the water. To finish each sub-task, a human behaves purposely to operate on or approach to the attentional objects. For example, when the human is doing the sub-task of "finding the cup", the human uses the hand to catch the attentional object *cup*. When the human is doing the sub-task of "going to the drinking fountain", the human walks to approach to the attentional object *drinking fountain*. For inside-video human attention estimation, we have a basic assumption that attentional objects locate inside videos/images.

The intuitive method to infer human attention is to estimate where a human is gazing at. It is true in some easy situations that human gaze significantly signals the attentional objects. As shown in Fig. 1(a), human gaze conveys sufficient cues to infer the attentional objects. However, human gaze does not absolutely indicate attentional objects, since in many cases a human is not necessarily gazing at attentional objects all the time. For example, when a human is walking to a drinking fountain to take the water, though the human's attentional object is the drinking fountain, the human could gaze at other objects in this procedure. In addition, human gaze estimation usually heavily depends on human facial information [35,45], but the facial information is often unavailable when a human moves naturally in uncontrolled scenes. As shown in Fig. 1(b), human faces are not observable, so it is difficult to estimate the human gaze. On the contrary, in these situations, human pose is available and significantly indicates the attentional objects. In most cases, the object, a human's hand is reaching to or a human's body is approaching to, is most likely to be the attentional object. However, human pose can not accurately signals attentional objects in all situations. For example, during the transition of two sub-tasks, the attentional objects are hardly signaled by human pose. One main reason is that "what a human thinks" goes ahead of "what a human does", so at the shifting time point from one sub-task to another sub-task, the attentional object might have changed while human pose still signals the attentional object in the previous sub-task. In some more complex and challenging situations, attentional objects can not be revealed even if both human gaze and pose are available. As shown in Fig. 1(c), even assuming that the human pose and human gaze are known, we still can not correctly infer the attentional objects because the human is facing a large number of objects and every object is possible to be the attentional object. In these cases, to correctly infer human attention, we need to resort to invisible high-level task information. For example, when a human is facing a juicer, a pot, and a stove at the same time, if the task was making juice, the attentional object is most likely to be the juicer, if the task was cooking soup, the attentional object is most likely to be the pot, and if the task was making pizza, the attentional object is most likely to be the stove.

Based on these observations, we propose a deep neural network model that fuses both visible low-level human pose cue and invisible high-level task encoding cue. The low-level human pose conveys the rich information of human body key joints, and the high-level task encoding cue is organized as a graph which encodes a task as several sub-tasks. By integrating the low-level human pose cue with the high-level task encoding cue, our model exhibits impressive robustness and effectiveness.

To validate our model, we conduct the intensive experiments for the comparison with other methods and for the ablation study of our method. We collect a new VR dataset and re-annotate a public real dataset. The experiments on both datasets validate the effectiveness of our method.

Our contributions are three-fold: (1) We propose and define a problem of inferring inside-video human attention that is different from the traditional human (outside images) attention. (2) We propose a model that integrates the low-level visible human pose cue with the high-level invisible task encoding information. (3) We collect and annotate a large-scale dataset in Virtual-Reality scenes and re-annotate a public real dataset. To our best knowledge, our newly collected dataset is the first VR dataset for inferring the task-driven inside-video human attention, and the dataset will be publicly released.

2. Related work

In this section, we review three related works. For each work, we first explain how it differs from our work, then briefly introduce the classical methods for solving the problem, finally analyze the datasets that are widely used for studying the problem.

2.1. Visual attention.

Visual attention mainly refers to the eye fixation saliency object [31,53,55] or saliency map [2,23,54], which signals the regions of an image where the human observer would pay attention at the first glance. The ground truth of saliency map or saliency object is usually obtained by the eye-tracking equipment that records the eye fixations of the observer looking at the image. Therefore, the visual attention is the attention of a human outside images. In this paper, we infer the attention of a human inside videos.

The classical pipeline is firstly predicting a saliency map and then minimizing the loss that signals the difference between the estimated saliency map and the ground truth. To predict a saliency map, early works use the single stream network to extract the feature map. However, the single stream network can not extract multiple-scale cues. Therefore, the multiple stream network is also proposed [22]. Feature map extraction is important for many visual problems, some researchers have proposed excellent models for feature extraction in unsupervised framework [12,60]. Recently, motivated by the study showing that early layers in a network capture low-level detail information while later layers capture high-level semantic information [6], one novel architecture, which is termed as skip-layer, is proposed to extract feature by combining the features from different layers [27,50]. A detailed survey for saliency object detection can be found in the work [49].

MIT1003 [24], TORONTO [3], PASCAL-S [28], and DUT-OMRON [62] are four widely used datasets. These datasets are proposed for studying the attention of a human outside images. Therefore, they are not suitable for inferring the inside-video human attention.

2.2. Human gaze

Human gaze is roughly categorized as first-person view gaze [15,32] and third-person view gaze [36,39,56]. For the first-person view gaze, the typical scenario is that a human is equipped with the wearable sensors, the data (images or videos) collected by the wearable sensors are used for the gaze estimation. For the third-person view gaze, a camera is installed in a fixed place, capturing videos or images that are used for the gaze estimation. In this paper, we estimate the attention of a human inside third-person view videos, which is related with the third-person view human gaze. The difference is that human gaze is usually defined as a direction [35,64,65] indicating where a human is physically gazing at, while human attention is the task-driven attentional objects.

Human gaze estimation methods usually operate in the bottom-up manner. Low-level visual features extracted from human pupil, eye, and/or face are fed to a model to regress a gaze direction [35,45] or used to fit to a known model [48,59] to estimate the most possible gaze. For example, the work [45] extracts the visual feature from the full-face image using convolution neural network, then the feature is fed into the fully connected layers to regress the 2D gaze location or 3D gaze direction. The work [48] consists of the offline and online stages. During the offline stage, a generic 3D eye-face model, which describes the relationship between the eyeball and facial landmarks, is learned. During the online stage, the facial landmarks are fit to the offline-learned model to estimate eyeball, the eyeball is then combined with the 3D geometric eye model to infer the human gaze.

EYE-DIAP [18], MPIIGaze [65], and Columbia Gaze [44] are three benchmark datasets for human gaze estimation. These datasets are collected in simple scenarios and the humans are restricted with limited head and body movements. The humans involved in the EYE-DIAP dataset [18] are gazing at a point on a screen or a floating target in the nearby space while keeping their heads static or with slight movements. MPIIGaze dataset [65] is collected in the single scenario that humans are gazing at the front camera of a laptop. When collecting Columbia Gaze dataset [44], the humans are controlled with the maximum 30° horizontal head rotation. As a result, the detailed facial information (such as pupil, eye, and face) of a human is fully observed in these datasets. To stride to large and complex scenes where humans are moving freely and the detailed facial information is not always available, some challenging and natural datasets like GazeFollow [39], Flickr gaze [36] and VideoGaze [40] are proposed. However, these datasets either lack object-level annotations or do not involve high-level task information. Therefore, they are not suitable for inferring the object-based and task-driven human attention.

2.3. Human object interaction

Human object interaction (HOI) involves two tasks, HOI classification and HOI detection. Given an image, the goal of HOI classification is to estimate a binary label for each HOI category, while the goal of HOI detection is to estimate a triplet of the human, object and HOI label [5]. Objects involved in the HOI detection usually refer to the objects that a human is directly interacting with at the current time. However, attentional objects might be far away from a human that the human is not currently interacting with or gazing at.

The typical methods for HOI detection firstly propose some HOI candidates, and then score each candidate based on the visual features that encode the relationship of human, object, and action. Constructing HOI feature is significant for HOI detection. Some early works extract features by encoding the spatial relation between human skeleton keypoints and the object [26,57]. These features are handcrafted. Recently, benefiting from the success of deep learning and the availability of large-scale HOI datasets, deep learning methods are used for feature extraction. For example, to represent the human-object relationship, the work [38] extracts the CNN feature of the bounding box enclosing both the human and the object. In [61], a novel feature is extracted by combining human gaze feature with human pose and object feature.

HICO-DET [5] and V-COCO [20] are two benchmark datasets for HOI detection and classification. They are not suitable for studying the task-driven human attention in videos for two reasons: (1) the datasets are composed of still images, which do not have temporal information; (2) the annotations are limited to the objects that a human is directly interacting with, while attentional objects might be the objects that a human is not directly interacting with at the current time.

3. Approach

In this section, we introduce our method by starting with the overview of our proposed deep neural network model. We then detail the architecture and data flow of our model. Finally, the loss functions are introduced.

3.1. Overview

Fig. 2 shows the overview of our model. The input is an image sequence, for simplicity, three images are shown in the figure to represent the input image sequence. The output is the attentional objects (denoted by red bounding boxes) in each image of the image sequence. Our model mainly consists of four modules: encoder, 3D convolution, task encoding module, and decoder. For clarification, we summarize some important denotations in Tab. 1. Each image I in the image sequence, together with the human pose h extracted from I, are served as the input of the encoder. The output m_{en} of each encoder for all images are concatenated together to serve as the input of the 3D convolution module. The output m_{3d} of the 3D convolution module is further processed by the task

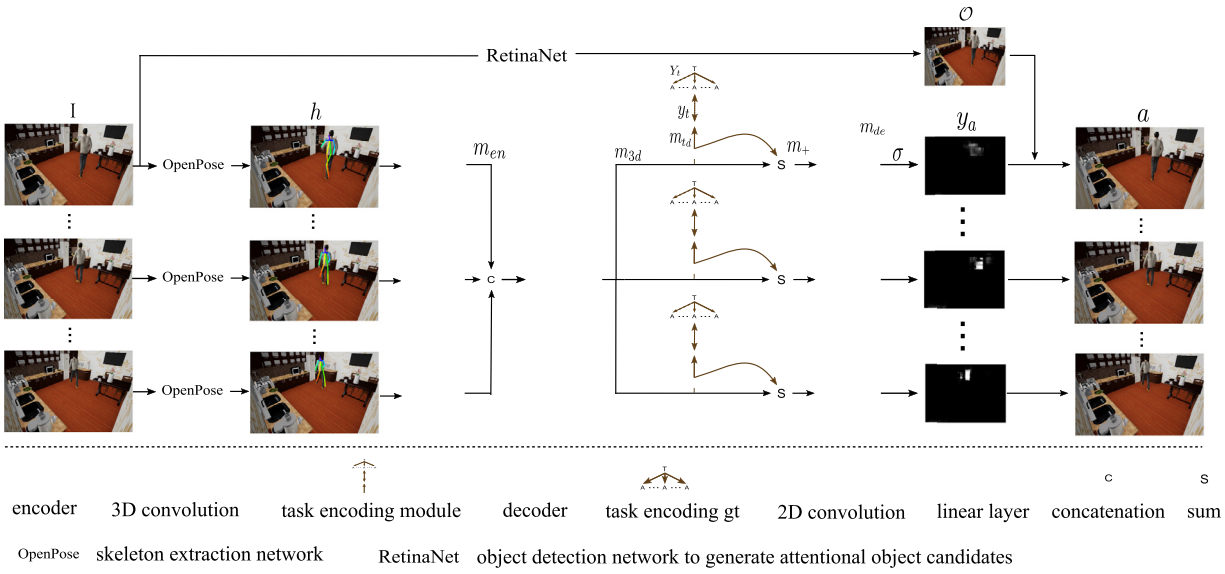

Fig. 2. Overview of our method. Given an image sequence as input, the purpose of our model is to output human attention a (denoted as red bounding boxes) in each image of the input sequence. To this end, the image I and the human skeleton h serve as the input of the model. The model is mainly composed of four modules: encoder module, 3D convolution module, decoder module, and task encoding module. m_{en} is the output of the encoder module for each image, the m_{en} for all images in the input image sequence are concatenated and processed by the 3D convolution module, generating the feature map m_{3d} which is further processed by a 2D convolution network to output the feature map m_{td}. m_{3d} and m_{td} are summed as m_+, which is processed by the decoder and the activation function σ to generate the attention map y_a. The human attention a is jointly inferred by attention heat map y_a and the attentional object candidates $\mathcal{O}$ (obtained by an object detection network named RetinaNet). In this figure, for the simplicity, the attentional object candidate generation network (RetinaNet) is only illustrated for one image of the input sequence. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
The summary of denotations.

Denotations	Representations
I	Input image
h	Human pose
m_{en}	Encoder feature map
m_{3d}	3D convolution feature map
m_{td}	Task-driven feature map
y_t	Task encoding prediction
Y_t	Task encoding ground truth
m_+	Fusion feature map
m_{de}	Decoder feature map
y_a	Attention heat map
$\mathcal{O}$	Attentional object candidates
a	Human attention/ attentional objects

encoding module, obtaining the feature map m_{td}. m_{3d} and m_{td} are summed as the fusion feature map m_+, which is taken as the input of the decoder. The output m_{de} of the decoder is activated by the sigmoid function σ, generating the attention heat map y_a. The human attention a is jointly inferred by the attentional object candidates $\mathcal{O}$ and attention heat map y_a.

3.2. Network architecture design motivation

The base architecture of our neural network model is "Encoder+Decoder", which is inspired by some classic works of semantic segmentation [1,33] and saliency estimation [53]. The "Encoder+Decoder" architecture is effective to extract the intrinsic feature, and can regress the attention map with the same size as the input image. To explore the temporal information in the image sequence, we configure a 3D convolution module between the encoder and decoder. Therefore, the backbone of our architecture is "Encoder+3D Convolution+Decoder".

However, the backbone architecture does not consider the high-level task information. Therefore, as shown in Fig. 2, we add a 2D convolution network on the top of 3D convolution feature map m_{3d} to extract the task-driven feature map m_{td}, which goes through a linear layer to generate the task encoding prediction y_t. The motivation of this architecture is two-fold. On one hand, we can compute the loss between the task encoding prediction y_t and its ground truth Y_t, the backward propagation of the loss can update the parameters of 3D convolution module and encoder module, guiding the network to predict the task-driven feature map. On the other hand, the feature map m_{td} conveying high-level task information is fused with the feature map m_{3d} to construct the fusion feature map m_+, allowing the network to predict attentional objects using both the low-level human pose cue and the high-level task encoding information.

3.3. Data flow

Input is an image sequence $\{I_t|t=1,2,\ldots,T\}$ with T images, and the output is human attention $\{a_t|t=1,2,\ldots,T\}$ in all T images. For the convenience of expression, we omit the subscript t of all variables. That is, we use I to represent an image and a to represent the human attention in the image I. In the following, we detail the data flow of each module.

Encoder. The input of the encoder is the image I and human pose h. The image I is with the size of $3\times H\times W$ (3 channels, H pixels in height, and W pixels in width). Human pose h is represented by the human skeleton, which is an informative representation and has been widely used in various computer vision tasks [46,58,61]. We use the method proposed in [4] to extract human skeleton. To align the data format with I, we use a binary $1\times H\times W$ mask to represent h, where human skeleton pixels are set as "1" and other pixels are set as "0". I and h are concatenated together as a $4\times H\times W$ array to serve as the input of encoder, the encoder feature map m_{en} is defined as:

$$m_{en}=\mathcal{F}_{en}([I,h]) \tag{1}$$

where $\mathcal{F}_{en}(\cdot)$ is the encoder neural network, $[\cdot,\ \cdot]$ denotes the concatenation operation.

3D Convolution. Each image has an encoder feature map m_{en} with the size of $C_{en} \times H_{en} \times W_{en}$. Let T be the image number of the input image sequence. By concatenating each m_{en} together, we obtain a feature map M_{en} with the size of $T \times C_{en} \times H_{en} \times W_{en}$, which is taken as the input of 3D Convolution module. The output M_{3d} of 3D Convolution module is defined as:

$$M_{3d} = \mathcal{F}_{3d}(M_{en}) \tag{2}$$

where $\mathcal{F}_{3d}(\cdot)$ is the 3D convolution neural network. M_{3d} is with the size of $T \times C_{en} \times H_{en} \times W_{en}$. m_{3d} is extracted from M_{3d}, denoting the 3D convolution feature map for each image, and the size of m_{3d} is $C_{en} \times H_{en} \times W_{en}$.

Task Encoding. A task is usually composed of several sub-tasks. Let N_T be the number all possible tasks and N_S be the number of all possible sub-tasks, then there are totally $N_T \times N_S$ possible compositions. Task encoding ground truth Y_t is one among $N_T \times N_S$ possible compositions, so we use one-hot vector to represent Y_t.

y_t is a $N_T \times N_S$ vector, representing task encoding prediction. y_t is generated based on the task-driven feature map m_{td} by a linear function $\mathcal{F}_l(\cdot)$:

$$y_t = \mathcal{F}_l(m_{td}) \tag{3}$$

where m_{td} is computed by adding a 2D convolution neural network on the top of 3D convolution feature map m_{3d}, defined as:

$$m_{td} = \mathcal{F}_{2d}(m_{3d}) \tag{4}$$

where $\mathcal{F}_{2d}(\cdot)$ is the 2D convolution neural network. m_{td} has the same size with m_{3d}.

Decoder. The input of encoder is the fusion feature map m_+, which is computed by adding m_{3d} and m_{td} together in the element-wise manner:

$$m_+ = m_{3d} + m_{td} \tag{5}$$

The output m_{de} of the decoder has the same size with input image in width and height, defined as:

$$m_{de} = \mathcal{F}_{de}(m_+) \tag{6}$$

where $\mathcal{F}_{de}(\cdot)$ is the decoder neural network.

Attention Heat Map. Attention heat map y_a is computed as:

$$y_a = \sigma(m_{de}) \tag{7}$$

where σ is the sigmoid activation function, y_a is a probability map, and $y_a \in [0, 1]^{1 \times H \times W}$.

Human attention a is jointly inferred by y_a and attentional object candidates $\mathcal{O}$.

3.4. Loss function

The loss $\mathcal{L}$ consists of the local loss $\mathcal{L}_l$, global loss $\mathcal{L}_g$, and the task encoding loss $\mathcal{L}_t$:

$$\mathcal{L} = \lambda_1 \mathcal{L}_g + \lambda_2 \mathcal{L}_l + \lambda_3 \mathcal{L}_t \tag{8}$$

where λ_1, λ_2, and λ_3 are weights for the individual loss.

Let Y_a be the ground truth of attention heat map. $Y_a \in \{0, 1\}^{1 \times H \times W}$ is a binary map with attention region R^+ assigned with '1' and non-attention region R^- assigned with '0'.

The purpose of local loss $\mathcal{L}_l$ is to compute the element-wise loss between y_a defined in Eq. (7) and attention heat map ground truth Y_a. Let y_a^{ij} be the (i, j)th element of y_a, and Y_a^{ij} be the (i, j)th element of Y_a. One classic loss function is computing the cross entropy loss for every local (i, j)th element, and then add them up or compute their average. However, in many cases, the proportion of the attention region R^+ and the non-attention region R^- is imbalanced, leading to the poor performance of this kind of loss function. To weaken the effect of imbalance, inspired by the loss function used in [53], we use the average weighted cross entropy loss, defined as:

$$\mathcal{L}_l = -\frac{1}{W \times H} \sum_{i,j} ((1-\omega) Y_a^{ij} \log y_a^{ij} + \omega (1 - Y_a^{ij}) \log(1 - y_a^{ij})) \tag{9}$$

where ω is automatically computed, representing the area ratio of R^+ to Y_a.

The global loss $\mathcal{L}_g$ computes the overall loss between y_a and Y_a, which is targeted to compensate for the local loss to avoid its excessive dominance. Inspired by the Dice coefficient proposed in [11], we formulate the $\mathcal{L}_g$ as:

$$\mathcal{L}_g = 1 - \frac{2 \sum_{i,j} y_a^{ij} Y_a^{ij}}{\sum_{i,j} (y_a^{ij})^2 + \sum_{ij} (Y_a^{ij})^2} \tag{10}$$

$\mathcal{L}_g$ is essentially a measure of overlap between y_a and Y_a.

The task encoding loss $\mathcal{L}_t$ is defined as standard cross-entropy loss:

$$\mathcal{L}_t = CE(y_t, Y_t) \tag{11}$$

where CE is the cross-entropy computing function, y_t is the prediction of task encoding defined in Eq. (3), and Y_t is the ground truth of task encoding.

4. Learning and inference

Let W_n be all parameters involved in the neural network. Learning the optimal parameter W_n^* is equal to minimize the loss $\mathcal{L}$ defined in Eq. (8):

$$W_n^* = \arg\min_{W_n} \mathcal{L} \tag{12}$$

We use ADAM algorithm to learn the parameters, with the learning rate set as 0.0001.

The attentional object candidates are objects that are possible to be the attentional objects. We use an object detection model (RetinaNet [29]) to detect objects to be the attentional object candidates. Let $\mathcal{O}$ be N_o attentional object candidates:

$$\mathcal{O} = \{o_k | k = 1, 2, \ldots, N_o\} \tag{13}$$

The goal of inference is to estimate the score of each candidate being the attentional object, based on the estimation of attention heat map y_a defined in Eq. (7). The score S_{o_k} of kth object candidate o_k is computed as:

$$S_{o_k} = \frac{\sum_{(i,j) \in o_k} y_a^{ij}}{A_{o_k}} \tag{14}$$

where A_{o_k} is the area of o_k. The score factually indicates the ratio between the summation of attention heat map inside o_k and the area of o_k.

5. Implementation details

Our model is implemented with PyTorch. The input images are resized to $3 \times 224 \times 224$, that is, $H = W = 224$. The encoder neural network $\mathcal{F}_{en}$ defined in Eq. (1) is implemented by the ResNet18 [21], which encodes the input image as a $512 \times 7 \times 7$ feature map, that is, $C_{en} \times H_{en} \times W_{en} = 512 \times 7 \times 7$. The length of the input image sequence is set as $T = 7$, so the size of M_{en} in Eq. (2) is $7 \times 512 \times 7 \times 7$. $\mathcal{F}_{3d}(\cdot)$ defined in Eq. (2) is implemented with the "Conv3d" function in PyTorch. The linear function $\mathcal{F}_l$ defined in Eq. (3) is implemented with the fully connected layers, and 2D convolution neural network defined in Eq. (4) is composed of two convolution layers with the kernel size of 1×1. The decoder neural

Table 2
The statistics of the AttentionObject-VR dataset. Videos: video number, Images: image number, Attentional objects: attentional object annotation number, other objects: non-attention object annotation number.

–	Videos	Images	Attentional objects	Other objects
Train	596	100,951	117,643	1,330,431
Test	184	32,468	37,211	402,573
Total	780	133,419	154,854	1,733,004

network defined in Eq. (6) is implemented with five deconvolution layers.

λ_1, λ_2, and λ_3 in Eq. (8) are set as 1, 1, and 0.1, respectively. The reason why we set $\lambda_3 = 0.1$ is to adjust the task encoding loss to have the similar quantitative magnitude with the global loss and local loss. The batch size is set as 6. The RetinaNet model [29] is pretrained on the ImageNet dataset [10] and fine-tuned on our dataset.

6. A new VR dataset

Though there exists a large number of datasets for the studies of human gaze, visual attention, and human-object interaction, to our best knowledge, no publicly available dataset is targeted for inferring the task-driven inside-video human attention. Therefore, we collect a video dataset in VR (Virtual Reality) scenes using VRKitchen platform [19], we call it the "AttentionObject-VR" dataset. With the development of VR technique, the VR data has become extremely life-like as real data. In VR scenes, all objects are configured with accurate locations and sizes, allowing the automatic object annotations and large-scale data collection.

To collect the dataset, we build 8 different kitchen scenes using Unreal Engine 4 (UE4). In each scene, many furniture and objects are configured, objects can be divided into two categories: *tools* (e.g., knife, juicer, oven, etc.) and *ingredients* (e.g. bread, orange, tomato, etc.). A human can use a *tool* to change the state of an *ingredient*. For example, to do the task of making orange juice, a human uses a knife to cut an orange into halves and put them into a juicer to get juice. Some statistics of our dataset are summarized in Table 2 and some samples are shown in Fig. 3. The dataset has following characteristics:

Diverse and large. The dataset consists of 8 scenes, 10 tasks, 33 sub-tasks, and 4 humans. As shown in Fig. 3, different scenes vary significantly in the scene scale, furniture configuration, and object placement. For each scene, we collect videos from 3 different camera views to make the data more diverse. As shown in the Table 2, the images of different camera views notably differ from each other. The 10 tasks are: bake bread, cook soup, cut meat, fry steak, make coffee, make juice, make sandwich, microwave food, pour coke, and turn on light. The dataset consists of 133,419 images and 1,887,858 object annotations in total. Averagely, each video consists of 171 images. The video resolution resolution is 1280×720.

Well-organized. To make the dataset qualified for inferring human attentional objects, it is necessary to guarantee humans and attentional objects are inside images. Therefore, we remove the images and videos that do not satisfy this requirement. To divide the dataset into training set and testing set, the data collected in scene 7 and scene 8 are used for testing, and the data collected in other scenes are used for training.

Well-annotated. Fig. 4 shows an example of annotating a video. Given a video with a task label, it is segmented as several sub-tasks to guarantee that the attentional object in each sub-task is determinate. To accurately segment a task into several sub-tasks, three volunteers are asked to find the key frames in a video to segment sub-tasks. For most cases, the key-frame is not controversial. For controversial ones, the average key-frame is taken as the key-frame. For each frame, the location, size and class of both attentional objects and non-attentional objects are annotated. Averagely, one image has 1.16 attentional object annotations and 13 non-attentional object annotations. Benefiting from the rich annotations, the dataset can also be used for other studies like task/event recognition, video segmentation, and action recognition.

7. Experiments

In this section, we first introduce the three baseline methods and the metric to evaluate the methods, followed by the detailed description of the comparison experiments as well as the ablation experiments, finally, the extension experiment on a public real dataset is introduced.

7.1. Baselines

We study the problem of inferring the task-driven attentional objects of a human inside third-person view videos, to our best knowledge, there does not exist exactly same work with ours. The most related work is to estimate where a human is looking. Therefore, we select two state-of-the-art human face and head direction estimation methods as baselines. In addition, we design a classification method. We briefly describe the three baseline methods as follows.

PRNet [16] is a face alignment method that can estimate human face direction. It takes the raw image and human face as input, and the output is the dense (more than 40K) aligned face key points. These dense points are compared with a pretrained model to compute the camera matrix, which is further combined with 68 facial key points to estimate the human face direction.

Hopenet [41] is a head pose estimation method. It takes the raw image and human face as input, and the output is the three Euler angles that signal human head direction.

ResNet-BinCls is a binary classification method based on ResNet-18 [21]. It first detects the objects in an image, then a binary classifier estimates the score of each object being and not being the attentional object. To estimate the score of a candidate object, the human skeleton and the candidate object are represented as a binary $1 \times H \times W$ mask, which is concatenated with $3 \times H \times W$ raw image to serve as the input of the binary classifier. Same with our method, the RetinaNet model [29] and OpenPose model [4] are respectively used for attentional object candidate generation and human pose estimation.

7.2. Metric

Let n_2 be the total number of testing images and n_1 be the number of images in which human attentional objects are correctly estimated, we evaluate the performance of a method using the following defined accuracy:

$$acc = \frac{n_1}{n_2} \tag{15}$$

For PRNet [16] and Hopenet [41], the outputs are respectively the face and head direction. To evaluate whether the attentional objects in an image are correctly estimated, we propose to estimate whether the face/head direction line intersects with the ground truth bounding boxes of attentional objects. For ResNet-BinCls and our method, the output is the scores of the attentional object candidates. We first find the object with the highest score. Let p_o be the center point of the highest scored object and p_h be the center point of the human head. If the line, which starts from p_h and passes through p_o, intersects with the ground truth bounding boxes of attentional objects, this image is counted to be correctly estimated.

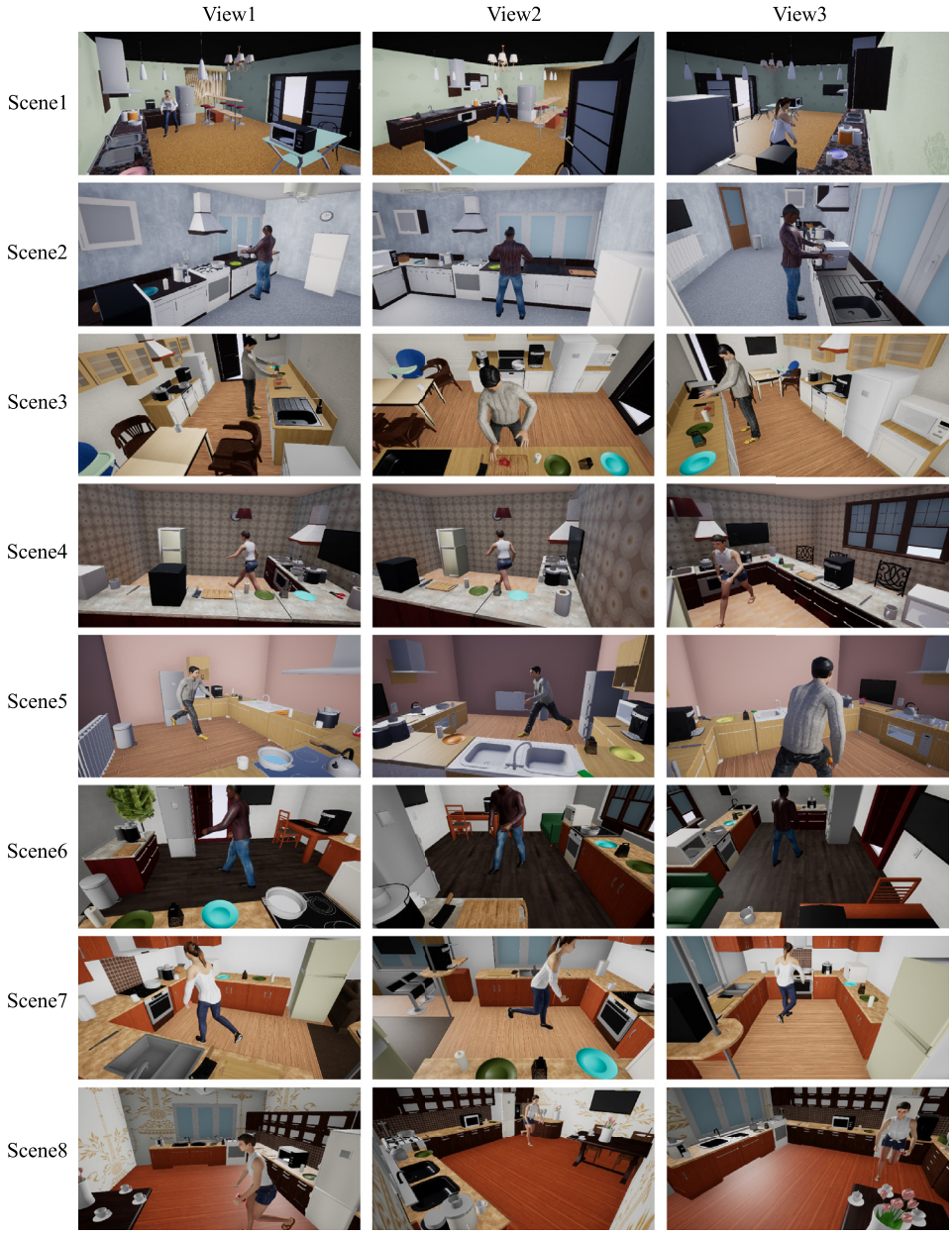

Fig. 3. Samples of the AttentionObject-VR dataset. The dataset is collected in eight scenes. In each scene, videos are captured from three different camera views. In this figure, each row shows three images from the three camera views at the same time in the same scene.

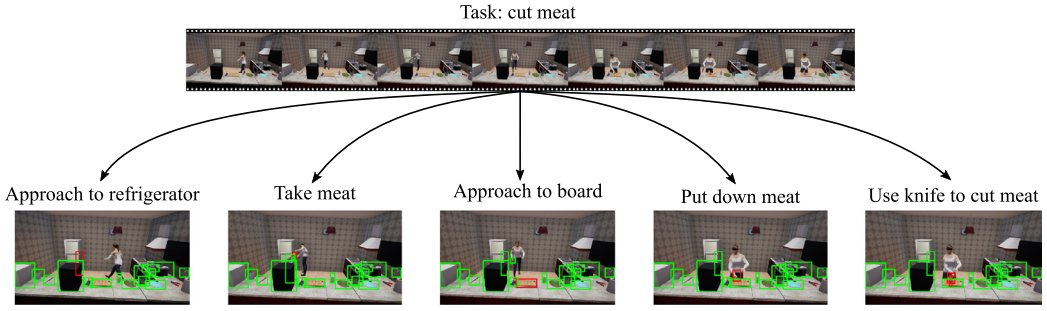

Fig. 4. An example of annotating a video. Given a video with the task label of "cut meat", the video is segmented as five sub-tasks ("approach to refrigerator", "take meat", "approach to board", "put down meat", and "use knife to cute meat"). In each sub-task, the attentional object (red bounding boxes) and other non-attentional objects (green bounding boxes) are annotated. To conclude, the annotations include task label, sub-task labels, attentional objects, and non-attentional objects. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Accuracies of different methods on the AttentionObject-VR dataset. "All" corresponds to the overall accuracy. T1 to T10 correspond to the accuracies on different tasks. T1: bake bread, T2: cook soup, T3: cut meat, T4: fly steak, T5: make coffee, T6: make juice, T7: make sandwich, T8: microwave food, T9: pour coke, and T10: turn on light. The last row corresponds to the accuracy of our method using the ground truth object annotations as attentional object candidates.

Methods	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	All
PRNet [16]	0.41	0.28	0.29	0.28	**0.26**	0.29	0.31	0.34	0.27	0.07	0.30
Hopenet [41]	0.54	0.36	0.36	0.37	0.17	0.37	0.39	0.39	0.29	0.00	0.35
ResNet-BinCls [21]	0.49	0.51	0.46	**0.55**	0.19	0.53	0.48	0.71	0.50	**0.48**	0.48
Our	**0.57**	**0.58**	**0.48**	0.53	0.12	**0.56**	**0.49**	**0.75**	**0.66**	0.40	**0.52**
Our*	0.72	0.65	0.68	0.54	0.70	0.68	0.68	0.79	0.82	0.45	0.69

Fig. 5. Samples of human attention heat map visualization in the task of make coffee. Red masks correspond to the regions with higher probability, while blue masks correspond to the regions with lower probability. Red bounding boxes are ground truth attentional object annotations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To provide the accurate head location for our method and the baselines, we first use the skeleton detector proposed in [4] to detect a human's five key points of nose, left eye, right eye, left ear, and right ear. Then, the average location of the available key points is taken as the head location. These five key points are distributed around human head, so the average location accurately indicates the location of human head. In addition, it rarely happens that no key point is detected, so the head location estimation is robust. To provide the accurate human face for PRNet [16] and Hopenet [41], instead of detecting faces on the whole image, we apply a face detector on a small region centered on the head location.

7.3. Quantitative results

Table 3 shows the human attention estimation accuracies of the baseline methods and our method. For each method, we compute the overall accuracy as well as the individual accuracy for each task. We can observe that our method achieves the highest overall accuracy and outperforms other methods on the majority of individual tasks, which benefits from our model that considers both the low-level visible human body cue and high-level invisible task encoding cue.

Attentional objects are usually the objects that the human's hands are operating on or the human's body is approaching to. Human skeleton conveys the motion and pose information of the whole body, thus significantly indicates the attentional objects. In addition, different from human facial cues, human skeleton is usually observable and easy to detect, contributing to robust performance when the detailed facial features are not available. Taking task encoding into consideration is also useful. Attention shifting procedure is a task-driven procedure of selecting attentional objects. Given a task, a human knows what to do now and what to do next. In this paper, we design a task encoding module. The module outputs the task-driven feature map and task encoding prediction. The task-driven feature map is fused with the low-level feature map, allowing the model to estimate human attention using both the low-level human body cue and the high-level task information. The task encoding prediction is compared with its ground truth to compute the loss, and the loss is used to update the parameters of neural network to involve the task guidance information.

From Table 3 we can observe that our method achieves a low accuracy on the task of make coffee (T5). In this task, the main attentional objects are the coffee machine and cup. However, the object detection model fails to detect the coffee machine in most cases, and the average precision for coffee machine detection is only 0.04. Actually, since the AttentionObject-VR dataset is collected in large and complex scenes and many objects are with small scales, the object detection on this dataset is not qualified. Object detection is not the focus of this paper, so we conduct another experiment using the ground truth object annotations as attentional object candidates, the results are shown in the last row (*Our**) in Table 3. We can observe that the performance improves by a large margin, the overall accuracy improves from 0.52 to 0.69. Especially for the task of make coffee, the accuracy improves from 0.12 to 0.70, proving that our attention heat map estimation is qualified. Fig. 5 shows some samples of attention heat map estimation in the task of make coffee, and we can observe the heat map concentrates on the ground truth bounding boxes of attentional objects. In the tasks of fly steak (T4) and turn on light (T10), our method behaves slightly worse than the ResNet-BinCls method. The main reason is, some attentional objects are with small sizes so that noisy objects near to the ground truth bounding boxes of attentional objects are easily to be falsely recognized as attentional objects.

7.4. Qualitative results

Fig. 6 shows some qualitative results of different methods in three typical scenarios. In easy scenarios, as shown in Fig. 6(a), the humans' facial cues are available and they are gazing at the attentional objects. Human pose also strongly indicates the attentional objects. Therefore, all methods correctly estimate the attentional objects.

However, in complex cases, as shown in Fig. 6(b), the humans are not facing to the cameras, so the detailed facial information is not available. As a result, the methods (PRNet [16] and Hopenet [41]) that heavily depend on facial features present poor performance, factually, when we analyze the experiment results, we found that the most failure of the PRNet [16] and Hopenet [41] happens in these situations. In contrast, our model takes the human skeleton as the low-level human pose cue, and the human skeleton is available and easy to detect even if the human face is partly or fully occluded. Therefore, our method presents better performance in these scenarios.

In some more challenging scenarios, as shown in Fig. 6(c), the human facial cue is difficult to extract, and human pose signals

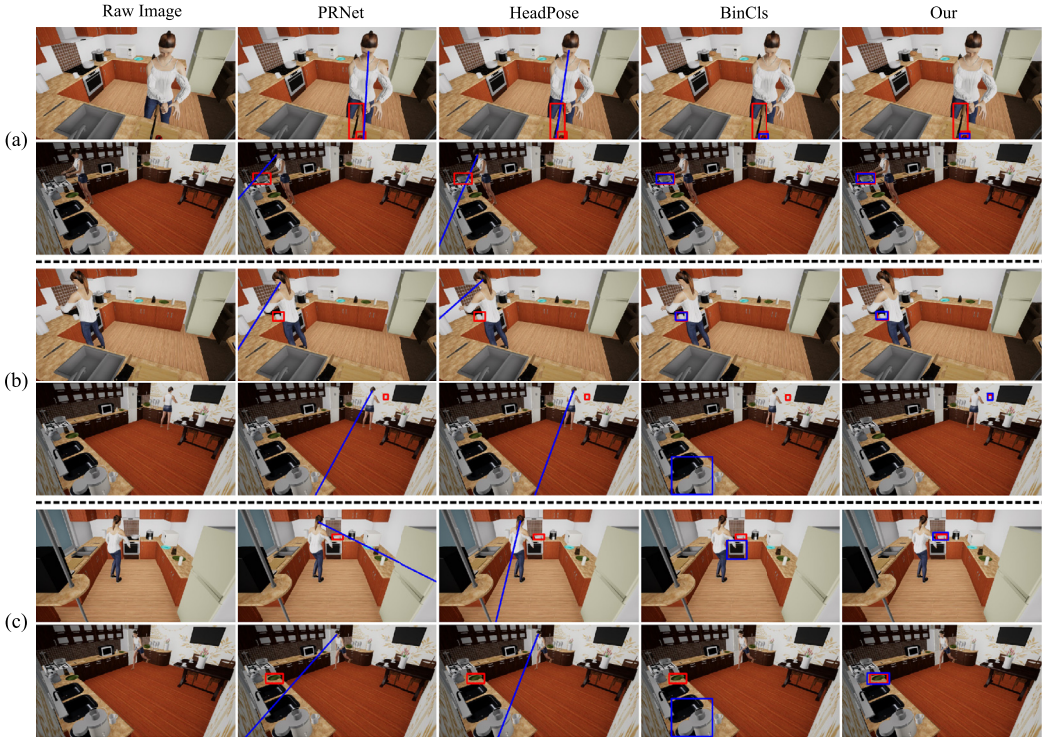

Fig. 6. Samples of qualitative results of different methods in three typical scenarios. (a) Human facial information is available and conveys the distinct cue to infer attentional objects. (b) Human facial information is not available, but the human pose provides the informative cue to infer attentional objects. (c) Human facial cue and human pose cue are not sufficient, and invisible high-level task information is needed to infer attentional objects. In this figure, the red bounding boxes represent the ground truth attentional object annotations, the blue lines represent the face and head directions estimated by the PRNet [16] model and Hopenet [41] model, and blue bounding boxes represent the attentional objects estimated by the ResNet-BinCls method [21] and our method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

multiple possible attentional objects. In these cases, the high-level task information is more important than the low-level cue. For example, as shown in the upper row images in Fig. 6(c), the human is facing a pot, a stove and a juicer. Assume that the task is cook soup, the attentional object is more likely to be the pot. Assume that the task is make juice, the attentional object is more likely to be the juicer. Assume that the task is bake bread, the attentional object is more likely to be the stove door.

7.5. Ablation studies

The purpose of ablation studies is to test the human attention estimation accuracies of our model with different model configurations. To this end, we design five experiments to test the effects of different loss compositions, encoder-decoder architectures, visual feature extraction methods, cue compositions and neural network configurations, respectively.

Experiment 1: loss compositions. In this paper, we propose three losses, local loss, global loss, and task encoding loss. In this experiment, we test the performance of our model with different loss compositions. We disable a loss by setting its weight as zero. As shown in Table 4, seven compositions are tested. From the table we can observe the highest accuracy is achieved when combing three losses, proving that every individual loss is useful. The accuracy of "global+task" is higher than that of "global" and the accuracy of "local+task" is higher than that of "local", proving the importance of the task encoding loss. The accuracy of "global+local" is same with that of "global" or "local", proving that the global constraint on the human attention heat map estimation and the local constraint on the human attention heat map estimation exhibit the similar effect.

Table 4
The human attention estimation accuracies of our model with different loss compositions. Local represents $\mathcal{L}_l$ defined in Eq. (9), global represents $\mathcal{L}_g$ defined in Eq. (10), and task represents $\mathcal{L}_t$ defined in Eq. (11).

Loss compositions	Accuracies
task	0.29
local	0.48
global	0.48
global+local	0.48
global+task	0.51
local+task	0.49
global+local+task	**0.52**

Experiment 2: encoder-decoder architectures. In this paper, we use the ResNet18 [21] neural network as the encoder and five deconvolution layers as the decoder. The ResNet18 neural network encodes a $3 \times 224 \times 224$ input image as a $512 \times 7 \times 7$ feature map. By removing the layers at the end of ResNet18 [21], the encoder outputs different sizes of the feature map. We test the performance our model when encoding the input image as different sizes of feature maps. As shown in Tab. 5, we test four feature map sizes, each size corresponds to one encoder-decoder architecture. From the table, we can observe that the $512 \times 7 \times 7$ feature map achieves the highest accuracy. Fig. 7 shows four samples of attention heat maps that are estimated by four different encoder-decoder architectures. From the figure we can also observe that the $512 \times 7 \times 7$ feature map achieves better attention heat maps. For the encoder-decoder

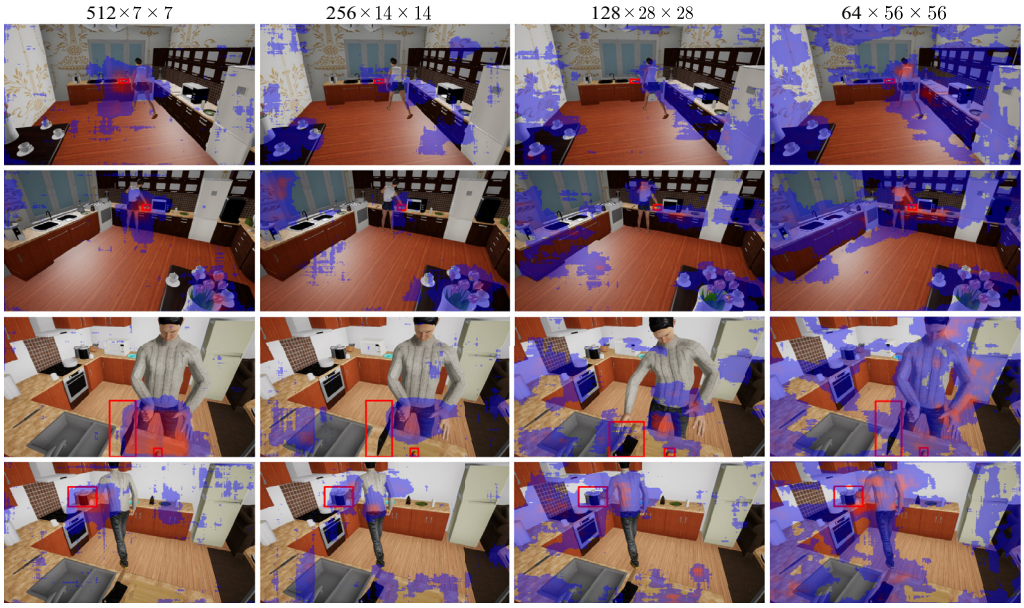

Fig. 7. Samples of attention heat maps that are estimated by four different encoder-decoder architectures. One architecture corresponds to one encoder feature map size, and the sizes are respectively $512 \times 7 \times 7$, $256 \times 14 \times 14$, $128 \times 28 \times 28$, and $64 \times 56 \times 56$. Red masks correspond to the regions with higher probability, while blue masks correspond to the regions with lower probability. Red bounding boxes are ground truth attentional object annotations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
The human attention estimation accuracies of our model with different encoder-decoder architectures.

Feature map sizes	Accuracies
$64 \times 56 \times 56$	0.43
$128 \times 28 \times 28$	0.46
$256 \times 14 \times 14$	0.47
$512 \times 7 \times 7$	**0.52**

Table 6
The human attention estimation accuracies of our model with different visual feature extraction neural networks.

Encoder types	Accuracies (pre-train)	Accuracies (no pre-train)
ResNet18	0.52	0.45
ResNet34	0.52	0.45
VGG16	0.52	0.48
VGG19	0.52	0.46

Table 7
The human attention estimation accuracies of our model with different cue compositions. image represents that the model only uses the raw image cue, image + task-cue represents that the model uses the raw image cue and the task encoding cue, image + skeleton-cue represents that the model uses the raw image cue and the human skeleton cue, and image + skeleton-cue + task-cue represents that the model uses the raw image cue, skeleton cue and task encoding cue.

Cue compositions	Accuracies
image	0.45
image + task-cue	0.50
image + skeleton-cue	0.49
image + task-cue + skeleton-cue	**0.52**

architecture, deeper encoder generates smaller feature map and exhibits better performance.

Experiment 3: feature extraction networks. The encoder in our model is actually a visual feature extraction neural network. In this experiment, we test the performance of our model with different feature extraction networks. As shown in Table 6, we test four widely used neural networks. For each network, we test the accuracy when it is pretrained on the ImageNet dataset [10] as well as the accuracy when it is not pretrained. From the table we can observe that pretrained networks present better performance than non-pretrained networks, and four pretrained networks achieve same accuracies.

Experiment 4: cue compositions. One main contribution of this paper is that we propose a model fusing both low-level human pose cue (represented by the human skeleton) and high-level task cue (represented by the task encoding). To analyze the effectiveness of individual cues, we conduct an ablation experiment to test the performance of our model with different cue compositions. The experiment results are summarized in the Table 7. We can observe that the model using all cues achieves the highest accuracy, proving that each individual cue is effective. The model additively using individual skeleton cue (or task encoding cue) achieves higher accuracy than the model only using the raw image, which also validates the effectiveness of human pose cue and task encoding cue.

Experiment 5: network configurations. Our model is composed of four neural network modules: encoder network, 3D convolution network, task encoding network, and decoder network. Encoder-decoder is the basic unit of our neural network model, 3D convolution network is a module to utilize the temporal information of input image sequence, and task encoding network is a module to utilize the task information. To analyze the effectiveness of individual network module, we conduct an ablation experiment to test the performance of our model with different network configurations. The experiment results are summarized in the Table 8. We can observe that the model achieves higher accuracy after adding the 3D convolution network and the model configured with all networks achieves the highest accuracy, which proves the effectiveness of individual networks.

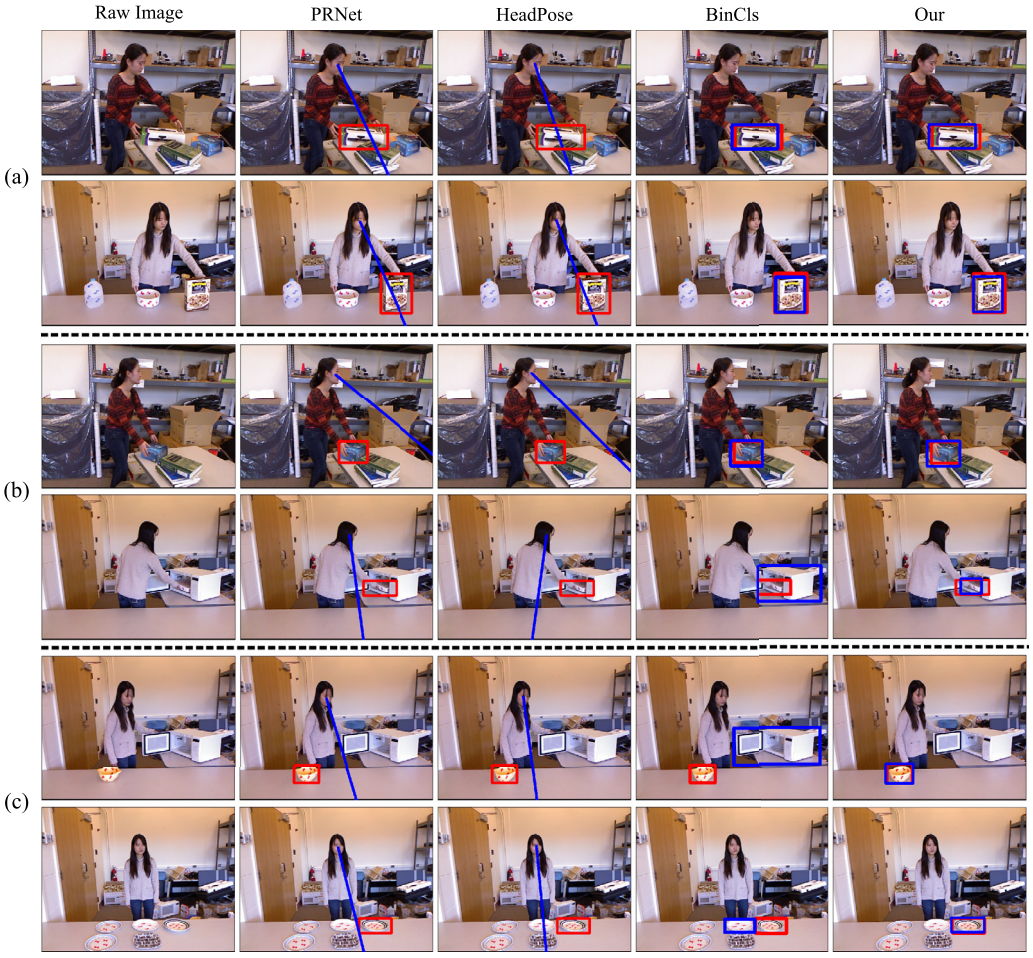

Fig. 8. Samples of qualitative results on the CAD120 dataset. In this figure, the red bounding boxes represent the ground truth attentional object annotations, the blue lines represent the face and head directions that are estimated by the PRNet [16] model and Hopenet [41] model, and blue bounding boxes represent the attentional objects that are estimated by the ResNet-BinCls method [21] and our method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 8
The human attention estimation accuracies of our model with different neural network configurations. "encoder-decoder" represents that the model is only configured with the basic encoder-decoder network, "encoder-decoder + 3DConv" represents that the model is configured with encoder-decoder network as well as the 3D convolution network, and "encoder-decoder + 3DConv + task-net" represents that the model is configured with encoder-decoder network, 3D convolution network and task encoding network.

Network configurations	Accuracies
encoder-decoder	0.45
encoder-decoder + 3DConv	0.49
encoder-decoder + 3DConv + task-net	**0.52**

7.6. Extension experiment on a real dataset

To further validate the robustness and effectiveness of our model, we re-annotate a publicly available dataset named CAD120 [26] and extensively evaluate our model on this real dataset. The CAD120 dataset [26] consists of 124 videos. In each video, a human is doing a task. There are totally 10 tasks, 10 sub-tasks, and 4 humans. The 10 tasks are: arranging objects, cleaning objects, making cereal, microwaving food, picking objects, stacking objects, taking food, taking medicine, unstacking objects, and having a meal.

Table 9
The statistics of the re-annotated CAD120 dataset used in our experiment. Videos: video number, Images: image number, Attentional objects: attentional object annotation number, other objects: non-attention object annotation number.

–	Videos	Images	Attentional objects	Other objects
Train	84	41,805	44,852	48,312
Test	28	13,499	14,258	16,739
Total	112	55,304	59,110	65,051

For each video, the task label and sub-task labels have been annotated. For each frame in a video, the objects are annotated. To make it eligible for inferring human attentonal objects, we annotate the attentional objects in all frames of all videos except for the videos with the task label of having a meal since the original annotations are not qualified. Following the setting in [26], we use the videos of three humans for training and the videos of one human for testing. The statistics of the re-annotated CAD120 dataset are summarized in the Table 9.

Table 10 shows human attention estimation accuracies of different methods on the CAD120 dataset, from which we can observe that our method achieves the highest accuracy in all tasks. Our method exhibits better performance on the CAD120 dataset than on the AttentionObject-VR dataset. One main reason is that the CAD120 dataset is a simple dataset collected in small-scale

Table 10
Accuracies of different methods on the CAD120 dataset. "All" corresponds to the overall accuracy. T1 to T9 correspond to the accuracies on different tasks. 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. The last row corresponds to the accuracy of our method using the ground truth object annotations as attentional object candidates.

Methods	T1	T2	T3	T4	T5	T6	T7	T8	T9	All
PRNet [16]	0.30	0.74	0.82	0.59	0.41	0.71	0.36	0.80	0.80	0.66
Hopenet [41]	0.19	0.80	0.67	0.73	0.31	0.51	0.48	0.55	0.57	0.59
ResNet-BinCls [21]	0.31	**0.97**	0.89	0.82	0.66	0.73	0.89	0.69	0.74	0.79
Our	**0.35**	**0.97**	**0.90**	**0.95**	**0.69**	**0.88**	**0.89**	**0.78**	**0.85**	**0.85**
Our*	0.62	0.96	0.91	0.93	1.0	0.88	0.99	0.82	0.85	0.89

scenes where human pose and motion are not complex as that in the AttentionObject-VR dataset. In addition, videos with the same task label share the similar camera view and object placement, while the AttentionObject-VR dataset is collected from different camera views and the appearance of scenes from different camera views significantly vary from each other. The last row (Our*) in the Table 10 shows the accuracies of our model using the ground truth attentional object annotations as the attentional object candidates. We can observe that the accuracy slightly improves from 0.85 (Our) to 0.89 (Our*). This gap is smaller than the gap between the accuracy of 0.52 (Our) and 0.69 (Our*) on the AttentionObject-VR dataset as shown in Table 3. The main reason is that object detection model presents better performance on the CAD120 dataset.

Fig. 8 shows some samples of qualitative results. In Fig. 8(a), the human is operating on and gazing at the attentional objects, so the attentional objects can be easily inferred by the human pose or human face/head direction. In Fig. 8(b), the human is not gazing at the attentional objects or the human face is not observable. Therefore, PRNet model [16] and Hopenet model [41] fail to infer the attentional objects. In Fig. 8(c), neither the human human gaze nor human pose reveals the attentional objects. Our method integrates the low-level human pose cue with the high-level task information, so exhibits better performance in this kind of situation.

8. Conclusion

This paper infers the attention of a human doing a task inside a third-person view video. Human attention is defined as the attentional objects coinciding with the on-going task. To solve the problem, we propose a neural network that fuses both low-level human pose cue and high-level task encoding cue. To validate the proposed method, we collect a new dataset and re-annotate a public dataset. A large number of experiments are conducted on these two datasets, and the experiment results show that our method is robust and effective.

For the problem, this paper is tackling inside-video human attention estimation, which is different from traditional human attention estimation targeting to infer the saliency regions that draw the attention of a human outside images or videos.

For the methodology, this paper not only considers the low-level human pose cue, but also involves the high-level task information that can not be observed from the image. This framework is more reasonable than the framework that only uses bottom-up information, and has strong theory supports from psychology and biology studies demonstrating that human attention is controlled in both bottom-up and top-down manner [9].

This paper may provide some inspirations to the related problems like human-object interaction, human gaze estimation, and human intention prediction. For example, involving invisible high-level information may improve the performance of human gaze estimation. However, since the high-level task information can not be observed in images, this paper only adopts a simple mechanism to incorporate the task information. In the future, we will explore the better way to make use of the invisible task information. In addition, we will extend this work by connecting human attention and human intention. Human attention is a strong signal to infer human intention. We plan to build a graph model to concurrently infer human attention and human intention.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NO. 61773312, 61790562, 61790563). This research is also supported by DARPA XAI grant N66001-17-2-4029, ONR MURI grant N00014-16-1-2007, and ARO grant W911NF-18-1-0296, USA.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2020.107314.

References

[1] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[2] A. Borji, M.N. Ahmadabadi, B.N. Araabi, M. Hamidi, Online learning of task–driven object-based visual attention control, Image Vis. Comput. 28 (7) (2010) 1130–1145.
[3] N. Bruce, J. Tsotsos, Saliency based on information maximization, in: Advances in Neural Information Processing Systems, 2006, pp. 155–162.
[4] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7291–7299.
[5] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, J. Deng, Learning to detect human-object interactions, in: IEEE Winter Conference on Applications of Computer Vision, IEEE, 2018, pp. 381–389.
[6] Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the devil in the details: delving deep into convolutional nets. arXiv:1405.3531.
[7] Z. Chen, Object-based attention: a tutorial review, Atten. Percept. Psychophys. 74 (5) (2012) 784–802.
[8] W.-L. Chou, S.-L. Yeh, Object-based attention occurs regardless of object awareness, Psychon. Bull. Rev. 19 (2) (2012) 225–231.
[9] M. Corbetta, G.L. Shulman, Control of goal-directed and stimulus-driven attention in the brain, Nat. Rev. Neurosci. 3 (3) (2002) 201.
[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[11] L.R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (3) (1945) 297–302.
[12] B. Du, Y. Wang, C. Wu, L. Zhang, Unsupervised scene change detection via latent Dirichlet allocation and multivariate alteration detection, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11 (12) (2018) 4676–4689.
[13] R. Edmondson, K. Light, A. Bodenhamer, P. Bosscher, L. Wilkinson, Enhanced operator perception through 3d vision and haptic feedback, Unmanned Systems Technology XIV, International Society for Optics and Photonics, 2012.
[14] M.S. El-Nasr, A. Vasilakos, C. Rao, J. Zupko, Dynamic intelligent lighting for directing visual attention in interactive 3-d scenes, IEEE Trans. Comput. Intell. AI Games 1 (2) (2009) 145–153.

[15] A. Fathi, Y. Li, J.M. Rehg, Learning to recognize daily actions using gaze, in: European Conference on Computer Vision, Springer, 2012, pp. 314–327.
[16] Y. Feng, F. Wu, X. Shao, Y. Wang, X. Zhou, Joint 3d face reconstruction and dense alignment with position map regression network, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 534–551.
[17] V. Fernández-Carbajales, M.Á. García, J.M. Martínez, Visual attention based on a joint perceptual space of color and brightness for improved video tracking, Pattern Recognit. 60 (2016) 571–584.
[18] K.A. Funes Mora, F. Monay, J.-M. Odobez, Eyediap: a database for the development and evaluation of gaze estimation algorithms from RGB and rgb-d cameras, in: Proceedings of the Symposium on Eye Tracking Research and Applications, ACM, 2014, pp. 255–258.
[19] X. Gao, R. Gong, T. Shu, X. Xie, S. Wang, S.-C. Zhu, VRKitchen: an Interactive 3D Environment for Learning Real Life Cooking Tasks, in: ICML workshop on Reinforcement Learning for Real Life, 2019.
[20] Gupta, S., Malik, J., 2015. Visual semantic role labeling. arXiv:1505.04474.
[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[22] X. Huang, C. Shen, X. Boix, Q. Zhao, Salicon: reducing the semantic gap in saliency prediction by adapting deep neural networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 262–270.
[23] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach.Intell. (11) (1998) 1254–1259.
[24] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: IEEE International Conference on Computer Vision, IEEE, 2009, pp. 2106–2113.
[25] G. Keren, A. Ben-David, M. Fridin, Kindergarten assistive robotics (KAR) as a tool for spatial cognition development in pre-school education, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp. 1084–1089.
[26] H.S. Koppula, R. Gupta, A. Saxena, Learning human activities and object affordances from RGB-D videos, Int. J. Robot. Res. 32 (8) (2013) 951–970.
[27] Kümmerer, M., Theis, L., Bethge, M., 2014. Deep gaze i: boosting saliency prediction with feature maps trained on imagenet. arXiv:1411.1045.
[28] Y. Li, X. Hou, C. Koch, J.M. Rehg, A.L. Yuille, The secrets of salient object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 280–287.
[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[30] G.-H. Liu, J.-Y. Yang, Z. Li, Content-based image retrieval using computational visual attention model, Pattern Recognit. 48 (8) (2015) 2554–2566.
[31] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2010) 353–367.
[32] Y. Liu, P. Wei, S.-C. Zhu, Jointly recognizing object fluents and tasks in egocentric videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2924–2932.
[33] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[34] A. Martínez, W. Teder-Sälejärvi, M. Vazquez, S. Molholm, J.J. Foxe, D.C. Javitt, F. Di Russo, M.S. Worden, S.A. Hillyard, Objects are highlighted by spatial attention, J. Cognit. Neurosci. 18 (2) (2006) 298–310.
[35] S. Park, A. Spurr, O. Hilliges, Deep pictorial gaze estimation, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 721–738.
[36] D. Parks, A. Borji, L. Itti, Augmented saliency model using automatic 3d head pose detection and learned gaze following in natural scenes, Vision Res. 116 (2015) 113–126.
[37] A. Pooresmaeili, P.R. Roelfsema, A growth-cone model for the spread of object-based attention during contour grouping, Curr. Biol. 24 (24) (2014) 2869–2877.
[38] S. Qi, W. Wang, B. Jia, J. Shen, S.-C. Zhu, Learning human-object interactions by graph parsing neural networks, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 401–417.
[39] A. Recasens, A. Khosla, C. Vondrick, A. Torralba, Where are they looking? in: Advances in Neural Information Processing Systems, 2015, pp. 199–207.
[40] A. Recasens, C. Vondrick, A. Khosla, A. Torralba, Following gaze in video, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1435–1443.
[41] N. Ruiz, E. Chong, J.M. Rehg, Fine-grained head pose estimation without keypoints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2074–2083.
[42] B.J. Scholl, Objects and attention: the state of the art, Cognition 80 (1–2) (2001) 1–46.
[43] A. Seemann, Joint Attention: New Developments in Psychology, Philosophy of Mind, and Social Neuroscience, MIT Press, 2011.
[44] B.A. Smith, Q. Yin, S.K. Feiner, S.K. Nayar, Gaze locking: passive eye contact detection for human-object interaction, in: Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, ACM, 2013, pp. 271–280.
[45] Y. Sugano, M. Fritz, X. Andreas Bulling, et al., It's written all over your face: Full-face appearance-based gaze estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 51–60.
[46] J. Sung, C. Ponce, B. Selman, A. Saxena, Unstructured human activity detection from RGBDimages, in: IEEE International Conference on Robotics and Automation, IEEE, 2012, pp. 842–849.
[47] M. Vincze, W. Zagler, L. Lammer, A. Weiss, A. Huber, D. Fischinger, T. Koertner, A. Schmid, C. Gisinger, Towards a robot for supporting older people to stay longer independent at home, in: 41st International Symposium on Robotics, 2014, pp. 1–7.
[48] K. Wang, Q. Ji, Real time eye gaze tracking with 3d deformable eye-face model, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1003–1011.
[49] Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., 2019a. Salient object detection in the deep learning era: an in-depth survey. arXiv:1904.09146.
[50] W. Wang, J. Shen, Deep visual attention prediction, IEEE Trans. Image Process. 27 (5) (2017) 2368–2378.
[51] W. Wang, J. Shen, X. Dong, A. Borji, R. Yang, Inferring salient objects from human fixations, IEEE Trans. Pattern Anal. Mach. Intell. (2019).
[52] W. Wang, J. Shen, L. Shao, Consistent video saliency using local gradient flow optimization and global refinement, IEEE Trans. Image Process. 24 (11) (2015) 4185–4196.
[53] W. Wang, J. Shen, L. Shao, Video salient object detection via fully convolutional networks, IEEE Trans. Image Process. 27 (1) (2017) 38–49.
[54] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, A. Borji, Revisiting video saliency prediction in the deep learning era, IEEE Trans. Pattern Anal. Mach. Intell. (2019).
[55] W. Wang, J. Shen, R. Yang, F. Porikli, Saliency-aware video object segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 40 (1) (2017) 20–33.
[56] P. Wei, D. Xie, N. Zheng, S.-C. Zhu, Inferring human attention by learning latent intentions., in: IJCAI, 2017, pp. 1297–1303.
[57] P. Wei, Y. Zhao, N. Zheng, S.-C. Zhu, Modeling 4d human-object interactions for event and object recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3272–3279.
[58] P. Wei, Y. Zhao, N. Zheng, S.-C. Zhu, Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2016) 1165–1179.
[59] E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, A. Bulling, A 3d morphable eye region model for gaze estimation, in: European Conference on Computer Vision, Springer, 2016, pp. 297–313.
[60] W. Xiong, L. Zhang, B. Du, D. Tao, Combining local and global: rich and robust feature pooling for visual recognition, Pattern Recognit. 62 (2017) 225–235.
[61] Xu, B., Li, J., Wong, Y., Kankanhalli, M. S., Zhao, Q., 2018. Interact as you intend: intention-driven human-object interaction detection. arXiv:1808.09796.
[62] C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang, Saliency detection via graph-based manifold ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3166–3173.
[63] X. Zhang, N. Mlynaryk, S. Japee, L.G. Ungerleider, Attentional selection of multiple objects in the human visual system, Neuroimage 163 (2017) 231–243.
[64] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, Appearance-based gaze estimation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4511–4520.
[65] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, Mpiigaze: real-world dataset and deep appearance-based gaze estimation, IEEE Trans. Pattern Anal. Mach. Intell. 41 (1) (2017) 162–175.

Zhixiong Nan is currently an assistant professor at Xi'an Jiaotong University. He received the Ph.D. degrees from Xi'an Jiaotong University in 2019. He has been a joint Ph.D. student at the University of California, Los Angeles (UCLA) from 2017 to 2019. His research interests include human attention estimation and traffic scene understanding.

Tianmin Shu received his Ph.D. degree from University of California, Los Angeles in 2019. He is currently a postdoctoral associate in the Department of Brain and Cognitive Sciences at Massachusetts Institute of Technology, where he is studying social perception and multi-agent systems. He is the recipient of the 2017 Cognitive Science Society Computational Modeling Prize in Perception/Action. His work has also been featured in multiple media outlets.

Ran Gong received his B.S. degree in computer science and engineering from University of California, Los Angeles in 2018. He is currently a master student in the Department of Computer Science at University of California, Los Angeles. His research interests include computer vision and machine learning. Now he focuses on building the Virtual-Reality platforms.

Shu Wang received the B.S. degree from Fudan University, Shanghai, China, in 2018. He was a visiting student at National University of Singapore in 2016. Currently, he is a first year Ph.D. student in the Department of Statistics at University of California, Los Angeles. His research interest resides at the intersection of Vision, Virtual Reality, and Logical Language.

Ping Wei received the B.E. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China. He is currently an associate professor with the Institute of Artificial Intelligence and Robotics at Xi'an Jiaotong University. He has been a postdoctoral researcher with Center for Vision, Cognition, Learning, and Autonomy (VCLA) at University of California, Los Angeles (UCLA) from 2016 to 2017. His research interests include computer vision, machine learning, and computational cognition. He serves as a coorganizer of the International Workshop on Vision Meets Cognition: Func-

tionality, Physics, Intents and Causality at CVPR 2017 and 2018, respectively. He is a member of IEEE.

Song-Chun Zhu received his Ph.D. degree from Harvard University. He is currently professor of Statistics and Computer Science at UCLA. His research interests include vision, statistical modeling, learning, cognition, situated dialogues, robot autonomy and AI. He received a number of honors, including the Helmholtz Test-of-time award in ICCV 2013, the Aggarwal prize from the IAPR in 2008, the David Marr Prize in 2003 with Z. Tu et al. for image parsing, twice Marr Prize honorary nominations with Y. Wu et al. in 1999 for texture modeling and 2007 for object modeling respectively. He received the Sloan Fellowship in 2001, a US NSF Career Award in 2001, and an US ONR Young Investigator Award in 2001. He is a Fellow of IEEE since 2011.

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