Interaction-aware Joint Attention Estimation Using People Attributes

Chihiro Nakatani[1] Hiroaki Kawashima[2] Norimichi Ukita[1]
[1] Toyota Technological Institute, Japan [2] University of Hyogo, Japan

Abstract

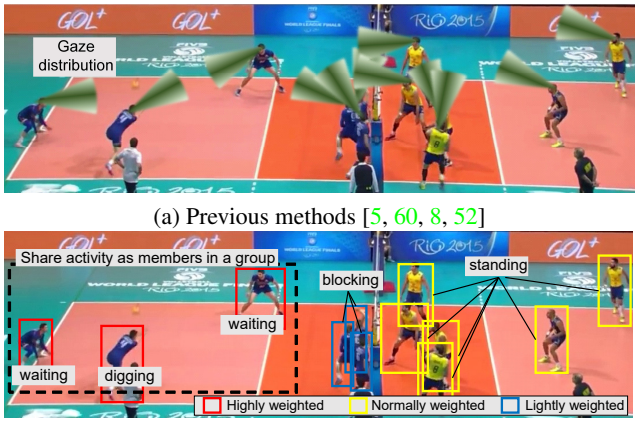
This paper proposes joint attention estimation in a single image. Different from related work in which only the gaze-related attributes of people are independently employed, (i) their locations and actions are also employed as contextual cues for weighting their attributes, and (ii) interactions among all of these attributes are explicitly modeled in our method. For the interaction modeling, we propose a novel Transformer-based attention network to encode joint attention as low-dimensional features. We introduce a specialized MLP head with positional embedding to the Transformer so that it predicts pixelwise confidence of joint attention for generating the confidence heatmap. This pixelwise prediction improves the heatmap accuracy by avoiding the ill-posed problem in which the high-dimensional heatmap is predicted from the low-dimensional features. The estimated joint attention is further improved by being integrated with general image-based attention estimation. Our method outperforms SOTA methods quantitatively in comparative experiments. Code: https://github.com/chihina/PJAE.

1. Introduction

Attention analysis enables various applications, such as customer's interest estimation [51], analyzing atypical gaze perception in autism spectrum disorder [21, 2], and human-robot interaction [47]. While attention is represented as a point, region, or object in the literature, we represent it as an attention point, AP, because a point can be used in any applications as an elemental representation. The confidence distribution of APs can be expressed in a heatmap image [5, 60, 8, 52] where each pixel value represents the confidence.

Attention estimation has two categories: single attention estimation [45, 46, 4, 3, 5, 9, 52, 18] and joint attention estimation [8, 60, 40, 41, 17]. While single attention estimation targets the attention of a person, attention shared by multiple people is detected by joint attention estimation.

In single attention estimation, an AP is estimated based on the gaze direction of a target person [42, 33, 10, 36, 26, 25, 14, 24] and the saliency map of a scene image [27, 38, 23, 56, 59, 24] in general. By simply aggregating (e.g., averaging) multi-people APs that are independently estimated by single attention estimation, a joint AP can be estimated [5, 52]. However, the APs of multiple people are not independent but jointly correlated in context.



(a) Previous methods [5, 60, 8, 52]

(b) Ours: gaze distributions are omitted for simple visualization

Figure 1. Difference between joint attention estimation methods. (a) Aggregating the gaze-related features of equally-weighted people without considering their interaction. (b) Aggregating the attributes of people weighted by their contextual attributes (i.e., locations and actions) via the interaction among all the attributes.

For joint attention estimation with such contextual correlation, in [8, 60], only the gaze-related attributes of people (e.g., "Gaze distribution" in Fig. 1 (a)) are employed. Such a straightforward approach has the problems below:

- **No contribution weights of people:** Some people share attention, but others do not. The latter people should not affect joint attention estimation, while all people are equally weighted in [8, 60].
- **No explicit interaction among people attributes:** As contextual cues related to joint attention, not only gazes [8, 60] but also other attributes of people, such as their locations and actions, are useful. Their interactions are also informative. For example, nearby people doing the same action may share the AP. Such interactions among people attributes are neglected in [8, 60].

These problems are resolved by the following novel contributions in this paper, as illustrated in Fig. 1 (b):

- **Activity awareness:** Each person's activity such as the location and action can be an important clue for joint attention estimation. For example, people sharing the AP tend to share their activities as group members. This assumption motivates us to focus on what and where each person is doing, namely the location and action of each person, to weight the contributions of people for joint attention estimation.

- **Interaction awareness:** Interactions among people attributes are explicitly modeled by our Position-embedded Joint Attention Transformer (PJAT), where a self-attention mechanism extracts the features of people sharing the AP.

- **Pixelwise joint attention heatmapping:** While the extracted joint-attention features are efficiently but sufficiently low-dimensional, it is ill-posed to estimate a high-dimensional heatmap image representing the AP confidences from such low-dimensional features. To avoid such an ill-posed estimation problem, we employ a network with image-coordinate embedding for estimating the AP confidence pixelwise.

2. Related Work

2.1. Single Attention Estimation

To understand a person's attention in a scene, appearance cues observed in the person's head, face, and eye images are important. Scene image features are also useful to extract saliency that attracts people's attention. Recasens *et al*. [45, 46] and Chong *et al*. [4] fuse CNN features extracted from a whole image and a cropped face image. Chong *et al*. [5] employ LSTM [16] for fusing these two kinds of features extracted in a video. Tu *et al*. [52] simultaneously estimate heads and their APs from a whole image.

Rather than the raw images of the head, face, and eyes used in the aforementioned methods, the gaze direction estimated from these images is more informative for identifying attention. Since the estimated gaze direction is not accurate enough, it is in general extended to a more noise-robust representation, such as a fan shape expressing the probabilistic distribution of the gaze direction [31, 9, 29]. As the scene features, features extracted from each object region can be more useful than those in the whole image [3].

2.2. Joint Attention Estimation

Joint attention estimation merges the APs of multiple people. For such estimation, gaze maps are superimposed to yield a social saliency field whose modes are regarded as multiple joint APs in [39]. In [40], the spatial relationship between multiple gaze directions and their attention is modeled via latent social charges inspired by Coulomb's law. As well as single attention estimation, joint attention estimation can be achieved by both raw head, face, and eye images [50] and gaze directions estimated in these images [8, 60]. For example, in [8], a fixed-size fan-shaped gaze map is drawn from the gaze direction of each person, and a CNN fuses the averaged gaze map of all people and the region proposal map of objects (i.e., saliency map) [61]. In [60], LSTM fuses the gaze maps observed in an image. Such CNN and LSTM might weight the gaze maps of people for joint attention estimation. However, the fixedly-shaped gaze maps cannot represent more flexible weights determined by interactions among people attributes (e.g., their locations, gaze directions, and actions). Such flexible weights are estimated by a self-attention mechanism in our method.

2.3. Location- and Action-, and their Interaction-Awareness

Our method is aware not only of the gaze directions of people but also of their other attributes such as the locations and actions, which are informative for a variety of tasks as follows. The locations of people provide a meaningful context for individual reasoning [12], person re-identification [53], people grouping [48], and trajectory prediction [49]. The action class of each person is informative for human pose estimation [11, 20], human motion synthesis [43, 35], action anticipation [1], human-human interaction estimation [57], human-object interaction estimation [32], and group activity recognition [28, 55, 44, 13, 37]. We can also simultaneously take into account the locations and the actions for further improving individual and group activity recognition [7], while these attributes are just implicitly encoded by general image feature extraction in [7].

While the effects of the locations and actions of people are validated for the above tasks, their effectiveness is unclear for joint attention estimation. For example, in [8, 60], the location is not directly used for attention estimation, while it is used for representing the gaze distribution, as shown in Fig. 1 (a). Our novelty in PJAT lies not in just verifying their effectiveness but in how to model interactions among these attributes (i.e., interaction awareness).

3. Proposed Method

The overview of the proposed method, consisting of three modules (α), (β), and (γ), is illustrated in Fig. 2. The attributes of each person are extracted from an image (Sec. 3.1). The extracted people attributes are fed into the Transformer encoder in (α) PJAT for interaction-aware joint attention estimation (Sec. 3.2 and Sec. 3.3). The estimated joint attention is integrated with the one estimated by (β) a general image-based network for further improvement in (γ) the fusion module (Sec. 3.4).

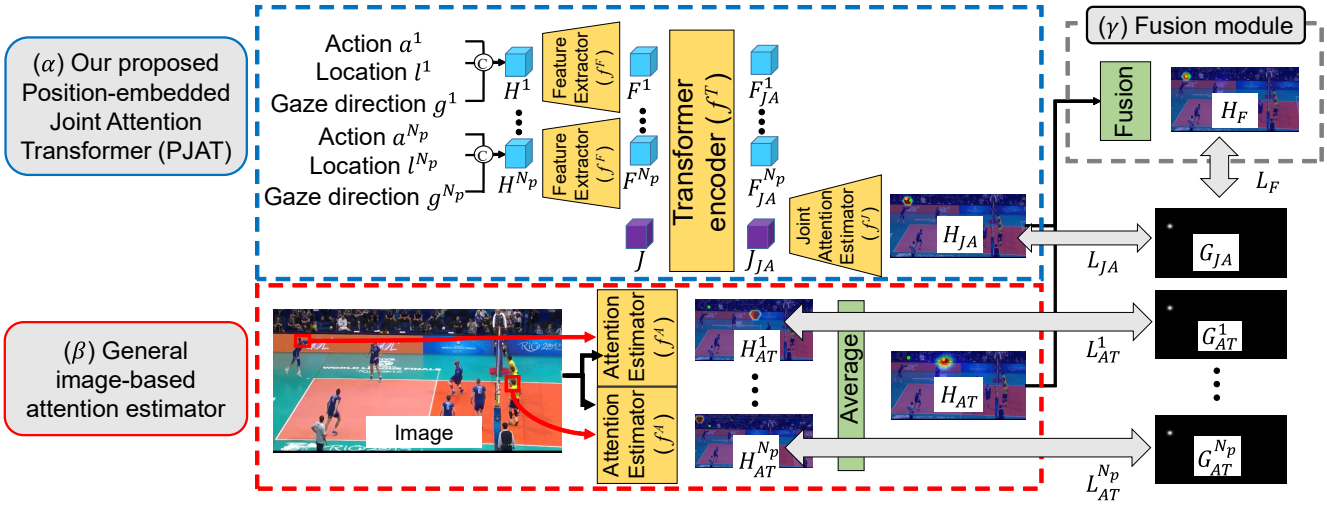


Figure 2. Overview of the proposed method consisting of the three modules. (α) Our proposed Position-embedded Joint Attention Transformer (PJAT) for joint attention estimation. In PJAT, a Transformer encoder models interactions among the attributes of people. (β) An image-based attention estimator. Any general attention estimation method can be fused with PJAT for improvement. (γ) Fusion module. The heatmaps obtained from (α) and (β), denoted respectively by H_{JA} and H_{AT}, are fused into the final heatmap, H_F.

3.1. Pre-processes for Person Attributes

The following pre-processes are used in our implementation, while these pre-processes are modularized so that they can be easily replaced with any SOTA methods.

Location detection. Our method employs the location of each person (denoted by $\boldsymbol{l}$) as one of the person attributes. While any location in the body can be regarded as the person's location, we use the head position as the person's location because the head is the edge point of a gaze line. For this head detection, the pretrained YOLOv5 [22] is fine-tuned with the head bounding boxes in each dataset.

Gaze direction estimation. The head bounding box of each person is fed into the gaze direction estimator. This estimator is a simple network consisting of VGG-16 for feature extraction followed by two fully-connected layers with output sizes of 64 and 2, in accordance with [60, 8]. The last output is a vector consisting of x and y directions (denoted by $\boldsymbol{g} = (g_x, g_y)$) whose norm is normalized to one.

Action recognition. As with the head, a full-body bounding box is detected by YOLOv5 [22]. This bounding box is fed into an action recognition network (ARG [55] in our experiments). Given N_a action classes, this network outputs a N_a-dimensional probability vector (denoted by $\boldsymbol{a}$) in which j-th component is the probability of j-th action class.

3.2. Transformer Encoder for Feature Interaction

The interaction among the features of people (i.e., location $\boldsymbol{l}^i$, gaze direction $\boldsymbol{g}^i$, and action $\boldsymbol{a}^i$, where i denotes the ID of each person), extracted by the pre-processes described in Sec. 3.1, is the main focus of this paper. We here employ Transformer to model such interactions. Transformer [54] has been proven in various fields to be powerful for modeling interactions of entities (e.g., spatial interaction among image patches split from an image for vision tasks [6, 34] and people interaction for group activity recognition [15, 13]). With the self-attention mechanism, Transformer successfully handles the interaction of multiple people. This mechanism is expected to play a crucial role in joint attention estimation because it can directly reason about who shares attention by using each person's location, gaze direction, and action. In addition, the characteristics of (i) accepting variable-length input and (ii) its permutation-invariant property are particularly important for our joint attention estimation problem, where the number and the order of detected people may change between images due to imperfect human detection.

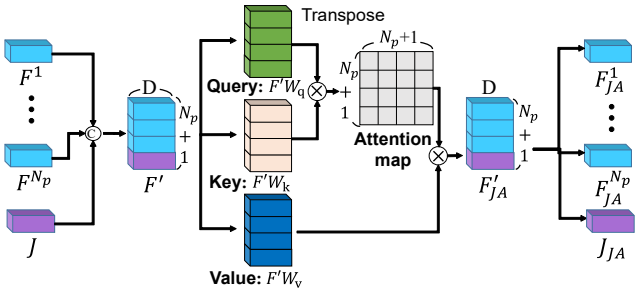


Figure 3. Self-attention network in the Transformer encoder. Self-attention models interactions among people attributes. The interactions are embedded into the joint attention feature $\boldsymbol{J}_{JA}$ using the features of individual people (i.e., $\boldsymbol{F}^1, \boldsymbol{F}^2, \cdots, \boldsymbol{F}^{N_p}$) and a learnable joint attention token (i.e., $\boldsymbol{J}$).

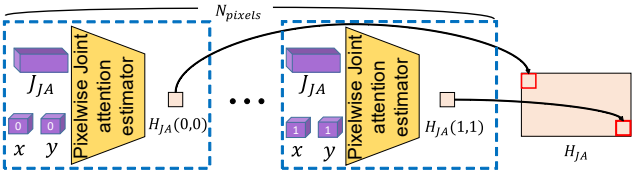


Figure 4. Pixelwise joint attention estimation network. Coordinates of each pixel (denoted by x and y) and extracted joint attention feature J_{JA} are fed into the network in which joint attention probability (denoted by $H_{JA}(x, y)$) is estimated.

The attributes of i-th person are concatenated to be $\boldsymbol{H}^i = (\boldsymbol{l}^i, \boldsymbol{a}^i, \boldsymbol{g}^i)$. $\boldsymbol{H}^i$ is fed into the feature extractor (f^F in Fig. 2) consisting of two fully-connected layers.

The features of all N_p people, each of which is a D-dimensional vector (denoted by $\boldsymbol{F}^i$), are fed into the Transformer encoder (f^T in Fig. 2). f^T consists of two transformer encoder layers with multi-head attention. Each transformer encoder is composed of the self-attention network, feed forward layer, layer normalization, and residual path, as with [54]. In the self-attention network shown in Fig. 3, all the features and a learnable token [6, 30] of joint attention (denoted by $\boldsymbol{J}$) are concatenated to be F'. F' is used to compute a query matrix $Q = F'W_q$, a key matrix $K = F'W_k$, and a value matrix $V = F'W_v$, where W_q, W_k, and W_v denote $D \times D$ learnable weight matrices. With Q, K, and V, F'_{JA} is defined to be $softmax\left(\frac{QK^T}{\sqrt{D}}\right)V$.

F'_{JA} is split into the feature vectors of N_p people (denoted by $\boldsymbol{F}^1_{JA}, \cdots, \boldsymbol{F}^{N_p}_{JA}$) and the joint attention feature $\boldsymbol{J}_{JA}$. This self-attention mechanism allows us to optimize $\boldsymbol{J}_{JA}$ by taking into account the mutual interaction between the people attributes (i.e., $\boldsymbol{l}^i$, $\boldsymbol{a}^i$, and $\boldsymbol{g}^i$ of all N_p people).

3.3. Pixelwise Joint Attention Estimator

With $\boldsymbol{J}_{JA}$, a low-dimensional latent vector, PJAT aims to generate a high-dimensional heatmap whose a value in each pixel (x, y) is the probability of joint attention in (x, y). While such a map can be directly estimated from $\boldsymbol{J}_{JA}$ using fully-connected layers [30], it is difficult to yield a map robustly due to its ill-posed nature. We therefore propose pixelwise estimation of a heatmap while maintaining the spatial relationship between pixels by positional embeddings.

The proposed joint attention estimator (f^J in Fig. 2) consists of three fully-connected layers followed by the Sigmoid activation at the last layer, as shown in Fig. 4. To estimate the probability of joint attention at (x, y), we feed the position (x, y) together with the encoded vector $\boldsymbol{J}_{JA}$ to the estimator by concatenating (x, y) with $\boldsymbol{J}_{JA}$. Its output is obtained between 0 and 1 via the Sigmoid activation, which is regarded as the probability value (denoted by $H_{JA}(x_i, y_i)$). We attach this specialized head to the Transformer encoder introduced in Sec. 3.2 to constitute PJAT.

3.4. Fusion with Image-based Attention Estimation

Scene features. As described in Sec. 3.2 and Sec. 3.3, PJAT estimates joint attention by focusing on interactions among people attributes. However, there are some specific situations where the people attributes alone cannot estimate joint attention well. For example, when the number of people present in a scene is small, precise estimation might be difficult due to their sparse gaze distributions. A larger number of people in a scene does not necessarily improve the estimation accuracy; for example, when the gaze directions are almost parallel, their intersection is sensitive to their noise. Here, using scene features extracted from each image may alleviate such difficulties. For example, visual saliency is demonstrated as useful appearance information in previous attention estimation methods [5, 52]. That is, the distribution of saliency captured by scene features helps localize the AP of each person in the gaze direction. While the AP of each person is independently estimated in single attention estimation [5, 52], the image-based scene features also benefit joint attention estimation [8] regardless of the number of people in a scene.

Fusion. Based on the discussion above, we design our method to fuse "the joint attention heatmaps H_{JA} estimated by PJAT" and "a map H_{AT} estimated by an image-based attention estimator (f^A in Fig. 2)" into the final joint attention heatmap H_F. We here employ DAVT [5], a SOTA attention estimation method using image-based scene features, as f^A. In DAVT, the average of individual attention heatmaps is regarded as H_{AT}. H_F is computed by the weighted fusion as follows: $H_F = W_{JA}H_{JA} + W_{AT}H_{AT}$, where W_{JA} and W_{AT} are weight coefficients. By training these coefficients for each dataset, our network fuses two heatmaps based on the scene properties.

Training objective. The overall network consisting of PJAT, DAVT, and the fusion module is trained with the heatmap estimation loss $L_{ALL} = L_{JA} + L_{AT} + L_F$, where L_{JA}, L_{AT}, and L_F denote the loss functions used for training PJAT, DAVT, and the fusion module, respectively. All the loss functions are based on the mean squared errors: $L_{JA} = \sum_n (H^n_{JA} - G^n_{JA})^2$, $L_{AT} = \frac{1}{N_p}\sum_i \sum_n (H^{i,n}_{AT} - G^{i,n}_{AT})^2$, and $L_F = \sum_n (H^n_F - G^n_{JA})^2$, where H^n_{JA}, H^n_F, and G^n_{JA} denote the n-th pixel value of the heatmaps estimated by PJAT, estimated by the fusion module, and given by the ground-truth, respectively. $H^{i,n}_{AT}$ and $G^{i,n}_{AT}$ denote the n-th pixel value of the i-th person's heatmaps estimated by the image-based attention estimator and given by the ground-truth, respectively. G_{JA} and G^i_{AT} are generated by drawing the 2D Gaussian distribution so that its center is located at the given ground-truth position of joint attention and i-th person's attention, respectively.

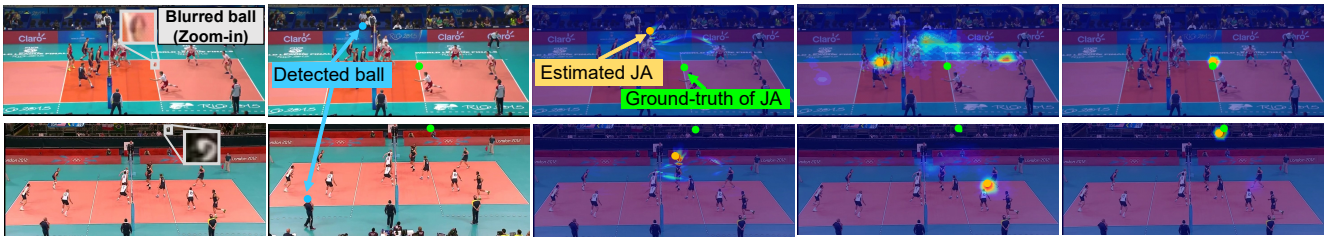


Figure 5. Visual comparison on the Volleyball dataset. Estimated joint attention heatmap is overlaid on the image. Green and yellow circles indicate the ground-truth and estimated joint APs, respectively.

Table 1. Experimental conditions on the Volleyball (Vol) and VideoCoAtt (Vid) datasets. People attributes given by prediction (Pr) and ground-truth (GT) are used in Ex.1 and Ex.2, respectively. The condition details are described in the supplementary material.

		Att	Body	Head	Action
Vol	Ex.1	$\boldsymbol{l}, \boldsymbol{g}, \boldsymbol{a}$	Pr	Pr in image	Pr
	Ex.2	$\boldsymbol{l}, \boldsymbol{g}, \boldsymbol{a}$	GT	Pr in GT body	GT
Vid	Ex.1	$\boldsymbol{l}, \boldsymbol{g}$	–	Pr in image	–
	Ex.2	$\boldsymbol{l}, \boldsymbol{g}$	–	GT	–

Figure 6. Visual comparison on the VideoCoAtt dataset. See the caption of Fig. 5 for details.

4. Experiments

4.1. Datasets and Evaluation Metrics

The proposed method is evaluated with the following two datasets. The Volleyball dataset, which includes many interactions among people, is mainly used to validate our contributions. Furthermore, the VideoCoAtt dataset is used to validate the generality of our method. While VideoCoAtt is used only in Sec. 4.3.2 with Table 3 and Fig. 6, detailed results are available in the supplementary material.

Volleyball dataset. This dataset [19] has 4,830 sequences. While each sequence has 41 frames, its center frame is annotated with the full-body bounding boxes of all players and their action classes, each of which is either of Waiting, Setting, Digging, Falling, Spiking, Jumping, Moving, Blocking, and Standing classes. In addition to these annotations, we newly provided the bounding box annotations of a ball whose center is regarded as the ground-truth of a joint AP. Since the ball is not observed in Left-winpoint and Right-winpoint sequences (662 sequences in total), these sequences are omitted. Consequently, the center frames in 4,168 sequences, consisting of 3,020 training and 1,148 test sequences [19], are used in our experiments.

VideoCoAtt dataset. This dataset [8] with 380 TV-show videos is for evaluating joint APs in more general scenes. Each frame is annotated with the bounding boxes of the ground-truth joint APs and people's heads. The number of APs differ between frames (i.e, 0, 1, and more APs), while it is fixed to be one in all frames in the Volleyball dataset. Since no action annotation is given to this dataset, only $\boldsymbol{l}^i$ and $\boldsymbol{g}^i$ compose H^i. DAVT as f^A was trained on GazeFollow [45] and VideoAttentionTarget [5] datasets.

Our experimental conditions on the aforementioned two datasets are shown in Table 1. In "Att," people attributes used in each dataset are shown. In each dataset, two types of experiments were conducted, namely those with the people attributes of prediction (Ex.1) and ground-truth (Ex.2). In the Volleyball dataset, an annotated full-body bounding box ("Body") is used only for head detection. Head bounding boxes are detected in a whole image and in ground-truth full-body bounding boxes in Ex.1 and Ex.2, respectively. In the VideoCoAtt dataset, annotated ground-truth head bounding boxes are used in Ex.2, while head bounding boxes are also detected in a whole image in Ex.1.

Evaluation metrics. The pixel with the max value in a heatmap image is regarded as a joint AP. All results are evaluated with the Euclidean distance between the pixels of the estimated joint AP and its ground-truth. The distances along x and y axes are also evaluated separately for detailed analysis because AP locations are biased about y axis, as mentioned in [8]. In addition, the detection rate is evaluated. Each detection is considered to be successful if the distance between the estimated and ground-truth joint APs is less than each threshold. In accordance with the diameter of joint attention in images (i.e., around 30 and 40 pixels in the Volleyball and VideoCoAtt datasets, respectively), "30, 60, and 90 pixels" and "40, 80, and 120 pixels" are selected as thresholds for Volleyball and VideoCoAtt, respectively.

Table 2. Quantitative comparison on the Volleyball dataset evaluated in the two experimental conditions mentioned in Sec. 4.1. Results obtained in ball detection, Ex.1, and Ex.2 are separated by double lines. Dist: the mean distance between the ground-truth and estimated joint APs. Thr: the threshold for the joint AP detection rate. The best result in each column is colored in red.

Method	Dist (x) ↓	Dist (y) ↓	Dist ↓	Thr=30 ↑	Thr=60 ↑	Thr=90 ↑
Ball detection [58]	147.6	66.5	174.8	54.3	56.0	58.5
ISA [8] (Ex.1)	53.1	35.5	70.1	60.7	69.7	75.9
DAVT [5] (Ex.1)	60.2	28.1	72.0	62.0	72.8	78.6
Ours (Ex.1)	44.1	25.2	56.0	64.5	76.8	83.0
ISA [8] (Ex.2)	36.7	24.7	48.7	46.0	79.1	92.8
DAVT [5] (Ex.2)	65.2	29.7	77.4	59.7	69.7	76.6
Ours (Ex.2)	9.3	4.7	11.4	96.3	98.9	99.6

Table 3. Quantitative comparison on the VideoCoAtt dataset. Accuracy and F-score: metrics for the joint AP prediction on the threshold given by validation data. AUC: area under the ROC curve for the joint AP prediction.

Method	Dist (x)	Dist (y)	Dist	Thr=40	Thr=80	Thr=120	Accuracy	F-score	AUC
ISA [8] (Ex.1)	108.5	85.7	152.7	8.5	24.9	48.9	0.41	0.19	0.41
DAVT [5] (Ex.1)	55.6	26.8	68.2	58.6	68.5	79.2	0.52	0.32	0.58
HGTD [52] (Ex.1)	112.5	65.7	142.7	20.4	32.9	46.3	0.18	0.28	0.50
Ours (Ex.1)	54.3	26.5	66.5	59.1	68.7	79.7	0.52	0.36	0.64
ISA [8] (Ex.2)	80.5	61.7	107.1	5.6	36.7	71.3	0.62	0.36	0.64
DAVT [5] (Ex.2)	35.7	21.1	46.6	72.9	80.7	89.2	0.61	0.30	0.57
HGTD [52] (Ex.2)	112.5	65.7	142.7	20.4	32.9	46.3	0.18	0.28	0.50
Ours (Ex.2)	34.4	21.0	45.0	74.3	82.5	89.6	0.57	0.37	0.65

For the VideoCoAtt, detection accuracy is also evaluated in accordance with [8, 5, 52]. However, more than using accuracy is needed because there is no joint AP in over 71% of test images. To relieve the class imbalance problem, F-score and Area Under the ROC Curve (AUC) are also used. If the max value in a heatmap is greater than a certain threshold, it is regarded that a joint AP is detected. The threshold which leads to the max F-score in validation data is used.

4.2. Training Details

For the Volleyball dataset, the overall network consisting of (α), (β), and (γ) in Fig. 2 is trained in an end-to-end manner after pretraining (α) and (β). For the VideoCoAtt dataset, only (γ) is trained after pretraining (α) and (β). The learning rates for the Volleyball and VideoCoAtt datasets were 0.001 and 0.00001, respectively.

4.3. Comparative Experiments

4.3.1 Volleyball Dataset

Our method is compared with SOTA methods [8, 5] on the Volleyball dataset. DAVT [5] is proposed as a single attention estimation, so joint attention is detected from the mean of the independently-estimated APs of all people. As the visual cue of a whole image, ISA [8] and DAVT [5] require a saliency map obtained by CenterNet [58] and a raw image, respectively. A head location is used for cropping a head image as an input feature in DAVT [5], while it is used only for computing the gaze direction in ISA [8]. In addition, our method is also compared with ball detection [58] because the ball is a strong cue for joint attention estimation in a ball game. As mentioned in Sec. 4.1, we evaluated these methods on two experimental conditions (i.e., using (Ex.1) predicted or (Ex.2) ground-truth people attributes).

Experimental results are shown in Table 2. Compared with all other methods, our method is better in all metrics. Visual results are shown in Fig. 5. While the joint attention is always on the ball in the Volleyball dataset, ball detection [58] often fails when a ball is visually unclear (i.e., blurred). Heatmaps estimated by SOTA methods [8, 5] tend to be erroneously blurred. On the other hand, our method can successfully estimate joint attention on the ball.

4.3.2 VideoCoAtt Dataset

For the evaluation with the VideoCoAtt dataset, HGTD [52] is also included as a comparative method. HGTD is trained with GazeFollow and VideoAttentionTarget as in the original paper since it requires a ground-truth head bounding box for its training. While HGTD is single attention estimation, joint attention is detected in a similar way as for DAVT.

Experimental results are shown in Table 3. Our method is better in all metrics except the accuracy of DAVT and ISA in Ex.1 and Ex.2, respectively. It is not surprising because accuracy is optimized for F-score by validation data, as described in Sec. 4.1.

Table 4. Ablation studies in Ex.1 on the Volleyball dataset. Ablated components about the people attributes and the network architectures are separated by double lines. Each metric is evaluated with two results, namely H_{JA} in branch (α) and H_F in fusion module (γ).

Method	Dist (α)	Dist (γ)	Thr=30 (α)	Thr=60 (α)	Thr=90 (α)	Thr=30 (γ)	Thr=60 (γ)	Thr=90 (γ)
Ours w/o $\boldsymbol{l}$	112.2	60.3	10.5	31.6	51.8	62.0	73.1	79.3
Ours w/o $\boldsymbol{g}$	138.9	70.8	22.8	40.7	53.2	60.5	72.0	78.3
Ours w/o $\boldsymbol{a}$	82.1	60.2	28.7	55.7	74.0	64.6	77.0	83.0
Ours w/o (α)	-	72.0	-	-	-	62.0	72.8	78.6
Ours w/o (β)	87.2	87.2	25.8	52.9	70.3	25.8	52.9	70.3
Ours	84.8	56.0	30.3	55.9	71.3	64.5	76.8	83.0

Table 5. Ablation studies in Ex.2 on the Volleyball dataset.

Method	Dist	Thr=30	Thr=60
Ours w/o $\boldsymbol{a}$	39.2	75.5	86.2
Ours w/o (α)	77.4	59.7	69.7
Ours w/o (β)	14.3	95.1	98.6
Ours	11.4	96.3	98.9

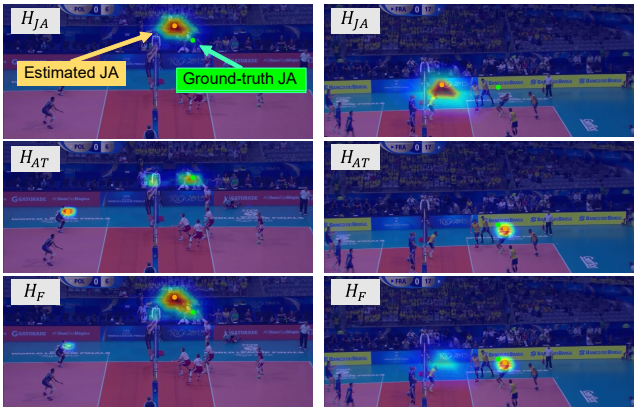


Figure 7. Visualization of H_{JA}, H_{AT}, and H_F obtained by modules (α), (β), and (γ) in Fig. 2. H_F is better than H_{JA} and H_{AT}.

Visual results are shown in Fig. 6. In [8], the blur is worse in the bottom example. In [5], the estimated AP is in the right person's face, while the ground-truth AP is in the left person's face in the bottom example. In contrast, our method can estimate the AP points more closely to their ground-truths in both examples.

4.4. Ablation Studies

The effect of each important component in our method is verified with the ablation studies shown in Tables 4 and 5 in which the results on the Volleyball dataset are shown; see the supplementary material for the VideoCoAtt dataset. We ablate either of $\boldsymbol{l}$, $\boldsymbol{g}$, and $\boldsymbol{a}$ (i.e., people attributes) by filling zero into ablated nodes in the first layer of the feature extractor network (Fig. 2). We also ablate either of network branches (α) and (β) shown in Fig. 2. For the experiments without branch (α) or (β), the output of each branch is regarded as the final joint attention estimation.

Table 6. Analysis of the negative impact caused by erroneous individual attributes, $\boldsymbol{l}$, $\boldsymbol{g}$, and $\boldsymbol{a}$, on the Volleyball dataset. GT and Pr denote the ground-truth and the prediction, respectively.

Inputs	Dist	Thr=30	Thr=60
($\boldsymbol{l}$=GT,$\boldsymbol{g}$=GT,$\boldsymbol{a}$=GT)	11.4	96.3	98.9
($\boldsymbol{l}$=GT,$\boldsymbol{g}$=Pr,$\boldsymbol{a}$=GT)	113.1	17.5	37.8
($\boldsymbol{l}$=GT,$\boldsymbol{g}$=GT,$\boldsymbol{a}$=Pr)	19.4	88.9	94.8
($\boldsymbol{l}$=GT,$\boldsymbol{g}$=Pr,$\boldsymbol{a}$=Pr)	116.2	16.7	37.5
($\boldsymbol{l}$=Pr,$\boldsymbol{g}$=Pr,$\boldsymbol{a}$=Pr)	150.8	12.4	26.8

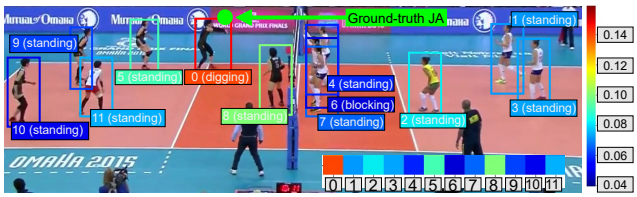


Figure 8. Visualized attention values, which come from the attention map shown in Fig. 3, learned by PJAT. The values for 12 people, which are in $s_{N_p+1,1}, \cdots, s_{N_p+1,N_p}$ where $s_{i,j}$ denotes the (i,j)-th entity of the attention map, are colored and shown on the bottom right side. The person ID ($\in 0, \cdots, 11$) is appended to this color map and each person's bounding box.

In Table 4, the best results for all metrics are obtained by "Ours" and "Ours w/o $\boldsymbol{a}$." The low performance of $\boldsymbol{a}$ can be attributed to the large recognition error of $\boldsymbol{a}$, where the accuracy of the action recognition is 53.3%. In fact, the use of the ground-truth of $\boldsymbol{a}$ in Ex.2 significantly improves performance, as shown in Table 5. Regarding $\boldsymbol{l}$ and $\boldsymbol{g}$, the results are improved in all metrics, as shown in Table 4. This is natural as (i) $\boldsymbol{g}$ is an essential gaze-related cue and (ii) head detection for estimating $\boldsymbol{l}$ is more reliable than the action recognition accuracy mentioned above (i.e., 53.3%).

While the contribution of branch (β) is larger in Ex.1, that of branch (α) is larger in Ex.2, as shown in Tables 4 and 5, respectively. The difference is also caused by prediction errors of $\boldsymbol{l}$, $\boldsymbol{g}$, and $\boldsymbol{a}$. The negative effect of such prediction errors is discussed in Sec. 4.5. In addition to the contribution of each branch, the combination of branches (α) and (β) is better than the results obtained by branch (α) or (β) independently. These results prove that each branch complementarily helps their estimation, as shown in Fig. 7.

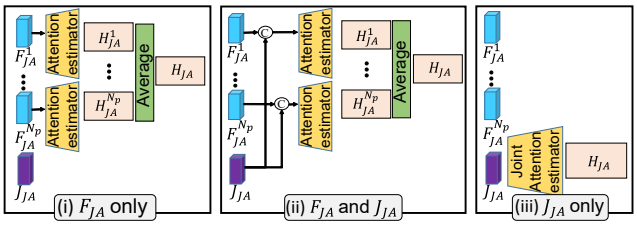


Figure 9. Three types of PJAT architectures. PJAT with "(iii) J_{JA} only" (the rightmost) is used as branch (α) in our method.

4.5. Detailed Analysis

Negative impact of erroneous individual attributes. The error in the person attributes ($\boldsymbol{l}$, $\boldsymbol{g}$, and $\boldsymbol{a}$) degrades joint attention estimation, as seen in Tables 4 and 5. To verify the negative impact of error in each individual attribute, we use the model trained in Ex.2, but use the possible combination of ground-truths and predictions of $\boldsymbol{l}$, $\boldsymbol{g}$, and $\boldsymbol{a}$ in inference (Table 6). Note that the three combinations with $\boldsymbol{l}$=Pr, i.e., ($\boldsymbol{l}$=Pr, $\boldsymbol{g}$=GT, $\boldsymbol{a}$=GT), ($\boldsymbol{l}$=Pr, $\boldsymbol{g}$=GT, $\boldsymbol{a}$=Pr), and ($\boldsymbol{l}$=Pr, $\boldsymbol{g}$=Pr, $\boldsymbol{a}$=GT), cannot be evaluated because the ground-truths of $\boldsymbol{g}$ and $\boldsymbol{a}$ can be used only when the location $\boldsymbol{l}$ is correct.

In comparison between ($\boldsymbol{l}$=GT, $\boldsymbol{g}$=Pr, $\boldsymbol{a}$=GT) and ($\boldsymbol{l}$=GT, $\boldsymbol{g}$=GT, $\boldsymbol{a}$=Pr), the error in $\boldsymbol{g}$ gives a much negative impact. This result is natural because the gaze direction $\boldsymbol{g}$ might be most important, while the positive effects of $\boldsymbol{l}$ and $\boldsymbol{a}$ are also validated in Tables 4 and 5. The significant performance gap between ($\boldsymbol{l}$=Pr, $\boldsymbol{g}$=Pr, $\boldsymbol{a}$=Pr) and ($\boldsymbol{l}$=GT, $\boldsymbol{g}$=GT, $\boldsymbol{a}$=GT) reveals that using different GT/prediction combinations in training and test phases degrades the performance. In fact, the result of Ex.1 shown in Table 2 is better than ($\boldsymbol{l}$=Pr, $\boldsymbol{g}$=Pr, $\boldsymbol{a}$=Pr) because the predicted attributes are used for both the training and test phases in Ex.1.

Attention values in self-attention. PJAT can weight the contributions of people by the self-attention mechanism, as shown in Fig. 8. It shows that the "digging" person (i.e., 0-th person) closest to the AP and people nearby this person (i.e., 5-th and 8-th people) are highly weighted. It can also be seen that the weights given to people doing the same action (e.g., "standing" of all people except 0-th and 6-th people) are different from each other. This is evidence that PJAT can learn complex people interactions so that the weight of each action is not fixed but changes depending on other attributes such as the location and gaze direction.

PJAT architecture comparison. Table 7 shows the comparison of different PJAT architectures in Fig. 9. To estimate H_{JA}, "F_{JA} only" takes the average of H^i_{JA}, the AP heatmaps estimated from $\boldsymbol{F}^i_{JA}$. In "F_{JA} and J_{JA}", $\boldsymbol{J}_{JA}$ is also used to estimate each H^i_{JA}. While the two methods aggregate estimated maps H^i_{JA} to compute H_{JA}, "J_{JA} only" estimates H_{JA} directly from $\boldsymbol{J}_{JA}$. "J_{JA} only" shows the best performance as it can directly utilize person-level contribution weights. However, the gap from "F_{JA} only" is small, which may reflect the characteristics of the Volleyball dataset, i.e., the large number of people in a scene yields robust estimation with simple averaging.

Table 7. Comparison of different heatmap generators in branch (α) in Ex.1 on the Volleyball dataset. Ours uses "(iii) J_{JA} only" for pixelwise estimation. See the supplementary material for Ex.2 and the results on VideoCoAtt.

Method	Dist (α)	Thr=30	Thr=60
(i) F_{JA} only	87.8	25.9	51.6
(ii) F_{JA} and J_{JA}	93.7	25.2	50.3
J_{JA} only for imagewise	126.7	10.7	28.4
(iii) J_{JA} only (Ours)	87.5	27.1	53.4

Table 8. Comparison of different fusion modules in Ex.1 on the Volleyball dataset. See the supplementary material for Ex.2 and the results on VideoCoAtt.

Fusion	Dist	Thr=30	Thr=60
CNN	94.6	44.1	68.8
Average	58.1	61.8	76.2
Weighted (Ours)	56.0	64.5	76.8

Pixelwise vs. imagewise. Pixelwise estimation with PJAT is compared with general imagewise heatmapping. As shown in Table 7, "Ours" outperforms "J_{JA} only for imagewise" because pixelwise estimation with positional information avoids an ill-posed problem mentioned in Sec. 1.

Fusion module comparison. Three fusion modules are compared in Ex.1. "CNN" fuses H_{JA} and H_{AT} by convolutional layers, where the details of the architecture are shown in the supplementary material. "Average" takes the average of H_{JA} and H_{AT} to compute H_F. In "Weighted," the weight coefficients for H_{JA} and H_{AT} (i.e., W_{JA} and W_{AT} in Sec. 3.4) are optimized. As shown in Table 8, "CNN" is worse than the others due to inefficient convolution for the sparse heatmaps, H_{JA} and H_{AT}. While the gap from "Average" is small, "Weighted" requires only a few parameters, leading to stable weight estimation and better results.

5. Concluding Remarks

We addressed joint attention estimation by modeling the interaction of people attributes as rich contextual cues. To relieve the difficulty in estimating a high-dimensional heatmap from a low-dimensional latent vector, we proposed the Position-embedded Joint Attention Transformer (PJAT). Our method achieves state-of-the-art results on two significantly-different datasets, which prove the wide applicability of our method. This paper focused on the image domain, as with [50, 52], to validate our key idea (i.e., activity- and interaction-aware joint attention heatmapping) more clearly, which is directly applicable also to videos, as with [5, 8]. On the other hand, the use of video-specific features is important future work.

References

[1] Mohammad Sadegh Ali Akbarian, Fatemehsadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Encouraging lstms to anticipate actions very early. In *ICCV*, 2017. 2

[2] Shi Chen and Qi Zhao. Attention-based autism spectrum disorder screening with privileged modality. In *ICCV*, 2019. 1

[3] Wenhe Chen, Hui Xu, Chao Zhu, Xiaoli Liu, Yinghua Lu, Caixia Zheng, and Jun Kong. Gaze estimation via the joint modeling of multiple cues. *IEEE Trans. Circuits Syst. Video Technol.*, 32(3):1390–1402, 2022. 1, 2

[4] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *ECCV*, 2018. 1, 2

[5] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. Detecting attended visual targets in video. In *CVPR*, 2020. 1, 2, 4, 5, 6, 7, 8

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 4

[7] Mahsa Ehsanpour, Alireza Abedin, Fatemeh Sadat Saleh, Javen Shi, Ian D. Reid, and Hamid Rezatofighi. Joint learning of social groups, individuals action and sub-group activities in videos. In *ECCV*, 2020. 2

[8] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *CVPR*, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[9] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *CVPR*, 2021. 1, 2

[10] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. RT-GENE: real-time eye gaze estimation in natural environments. In *ECCV*, 2018. 1

[11] Juergen Gall, Angela Yao, and Luc Van Gool. 2d action recognition serves 3d human pose estimation. In *ECCV*, 2010. 2

[12] Andrew C. Gallagher and Tsuhan Chen. Understanding images of groups of people. In *CVPR*, 2009. 2

[13] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees G. M. Snoek. Actor-transformers for group activity recognition. In *CVPR*, 2020. 2, 3

[14] Xin Geng, Xin Qian, Zeng-Wei Huo, and Yu Zhang. Head pose estimation based on multivariate label distribution. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(4):1974–1991, 2022. 1

[15] Mingfei Han, David Junhao Zhang, Yali Wang, Rui Yan, Lina Yao, Xiaojun Chang, and Yu Qiao. Dual-ai: Dual-path actor interaction learning for group activity recognition. In *CVPR*, 2022. 3

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 2

[17] Nora Horanyi, Linfang Zheng, Eunji Chong, Aleš Leonardis, and Hyung Jin Chang. Where are they looking in the 3d space? In *CVPRW*, 2023. 1

[18] Zhengxi Hu, Yuxue Yang, Xiaolin Zhai, Dingye Yang, Bohan Zhou, and Jingtai Liu. Gfie: A dataset and baseline for gaze-following from 2d to 3d in indoor environments. In *CVPR*, 2023. 1

[19] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, 2016. 5

[20] Umar Iqbal, Martin Garbade, and Juergen Gall. Pose for action - action for pose. In *FG*, 2017. 2

[21] Ming Jiang and Qi Zhao. Learning visual attention to identify people with autism spectrum disorder. In *ICCV*, 2017. 1

[22] Glenn Jocher et al. Yolov5. https://github.com/ultralytics/yolov5. 3

[23] Petr Kellnhofer, Adrià Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *ICCV*, 2019. 1

[24] Rakshit Kothari, Shalini De Mello, Umar Iqbal, Wonmin Byeon, Seonwook Park, and Jan Kautz. Weakly-supervised physically unconstrained gaze estimation. In *CVPR*, 2021. 1

[25] Felix Kuhnke and Jörn Ostermann. Deep head pose estimation using synthetic images and partial adversarial domain adaption for continuous label spaces. In *ICCV*, 2019. 1

[26] Donghoon Lee, Ming-Hsuan Yang, and Songhwai Oh. Fast and accurate head pose estimation via random projection forests. In *ICCV*, 2015. 1

[27] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Deep saliency with encoded low level distance map and high level features. In *CVPR*, 2016. 1

[28] Shuaicheng Li, Qianggang Cao, Lingbo Liu, Kunlin Yang, Shinan Liu, Jun Hou, and Shuai Yi. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In *ICCV*, 2021. 2

[29] Yunhao Li, Wei Shen, Zhongpai Gao, Yucheng Zhu, Guangtao Zhai, and Guodong Guo. Looking here or there? gaze following in 360-degree images. In *ICCV*, 2021. 2

[30] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *ICCV*, 2021. 4

[31] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *ACCV*, 2018. 2

[32] Xue Lin, Qi Zou, and Xixia Xu. Action-guided attention mining and relation reasoning network for human-object interaction detection. In *IJCAI*, 2020. 2

[33] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(3):1092–1099, 2021. 1

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3

[35] Qianhui Men, Hubert P. H. Shum, Edmond S. L. Ho, and Howard Leung. Gan-based reactive motion synthesis with class-aware discriminators for human-human interaction. *Comput. Graph.*, 102:634–645, 2022. 2

[36] L. R. D. Murthy and Pradipta Biswas. Appearance-based gaze estimation using attention and difference mechanism. In *CVPR*, 2021. 1

[37] Chihiro Nakatani, Kohei Sendo, and Norimichi Ukita. Group activity recognition using joint learning of individual action recognition and people grouping. In *MVA*, 2021. 2

[38] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency for object segmentation from image level labels. In *CVPR*, 2017. 1

[39] Hyun Soo Park, Eakta Jain, and Yaser Sheikh. 3d social saliency from head-mounted cameras. In *NIPS*, 2012. 2

[40] Hyun Soo Park, Eakta Jain, and Yaser Sheikh. Predicting primary gaze behavior using social saliency fields. In *ICCV*, 2013. 1, 2

[41] Hyun Soo Park and Jianbo Shi. Social saliency prediction. In *CVPR*, 2015. 1

[42] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *ECCV*, 2018. 1

[43] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer VAE. In *ICCV*, 2021. 2

[44] Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang. Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In *ECCV*, 2020. 2

[45] Adrià Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *NIPS*, 2015. 1, 2, 5

[46] Adrià Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *ICCV*, 2017. 1, 2

[47] Akanksha Saran, Srinjoy Majumdar, Elaine Schaertl Short, Andrea Thomaz, and Scott Niekum. Human gaze following for human-robot interaction. In *IROS*, 2018. 1

[48] Viktor Schmuck and Oya Çeliktutan. GROWL: group detection with link prediction. In *FG*, 2021. 2

[49] Shan Su, Jung Pyo Hong, Jianbo Shi, and Hyun Soo Park. Predicting behaviors of basketball players from first person videos. In *CVPR*, 2017. 2

[50] Ömer Sümer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Attention flow: End-to-end joint attention estimation. In *WACV*, 2020. 2, 8

[51] Hoi Ying Tsang, Melanie Tory, and Colin Swindells. eseetrack - visualizing sequential fixation patterns. *IEEE Trans. Vis. Comput. Graph.*, 16(6):953–962, 2010. 1

[52] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. End-to-end human-gaze-target detection with transformers. In *CVPR*, 2022. 1, 2, 4, 6, 8

[53] Norimichi Ukita, Yusuke Moriguchi, and Norihiro Hagita. People re-identification across non-overlapping cameras using group features. *Comput. Vis. Image Underst.*, 144:228–236, 2016. 2

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NIPS*, 2017. 3, 4

[55] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *CVPR*, 2019. 2, 3

[56] Zeng Yu, Yun-Zhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *ICCV*, 2019. 1

[57] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *CVPR*, 2021. 2

[58] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv*, abs/1904.07850, 2019. 5, 6

[59] Yijun Zhou and James Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. In *BMVC*, 2020. 1

[60] Ning Zhuang, Bingbing Ni, Yi Xu, Xiaokang Yang, Wenjun Zhang, Zefan Li, and Wen Gao. MUGGLE: multi-stream group gaze learning and estimation. *IEEE Trans. Circuits Syst. Video Technol.*, 30(10):3637–3650, 2020. 1, 2, 3

[61] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *ECCV*, 2014. 2