Inferring Shared Attention in Social Scene Videos

Lifeng Fan*[1], Yixin Chen*[1], Ping Wei†[2,1], Wenguan Wang[3,1], and Song-Chun Zhu[1]

[1] Center for Vision, Cognition, Learning and Autonomy, UCLA, Los Angeles, USA
[2] Xi'an Jiaotong University, Xi' an, China
[3] Beijing Institute of Technology, Beijing, China
{lfan,ethanchen}@g.ucla.edu, pingwei@xjtu.edu.cn, wenguanwang.ai@gmail.com, sczhu@stat.ucla.edu

Abstract

This paper addresses a new problem of inferring shared attention in third-person social scene videos. Shared attention is a phenomenon that two or more individuals simultaneously look at a common target in social scenes. Perceiving and identifying shared attention in videos plays crucial roles in social activities and social scene understanding. We propose a spatial-temporal neural network to detect shared attention intervals in videos and predict shared attention locations in frames. In each video frame, human gaze directions and potential target boxes are two key features for spatially detecting shared attention in the social scene. In temporal domain, a convolutional Long Short-Term Memory network utilizes the temporal continuity and transition constraints to optimize the predicted shared attention heatmap. We collect a new dataset VideoCoAtt[1] from public TV show videos, containing 380 complex video sequences with more than 492,000 frames that include diverse social scenes for shared attention study. Experiments on this dataset show that our model can effectively infer shared attention in videos. We also empirically verify the effectiveness of different components in our model.

1. Introduction

Shared attention is defined as the attention focus shared by two or more individuals on one object or human [5]. Shared attention differs from joint attention in a subtle way and in the literature the two terms are used interchangeably [5]. Shared attention is everywhere in our daily life and we can observe it every now and then in almost all social interactions. Imagine in a party, usually humans can easily recognize a group of people with shared attention and what exactly is their shared attention in the group at present. They can join the group and form shared attention with them naturally and instantly. However, patients with autism may feel it difficult to interact with people around them since they lack the ability to build shared attention with others [3]. Fig. 1 shows some examples of shared attention in social scenes and how shared attention shifts temporally as well as who are currently involved in the shared attention.

Figure 1. **Shared attention is everywhere in our daily life.** Shared attention is a crucial first step towards social interaction, the primary basis of social intelligence and a precursor of Theory of Mind [7].

Research in developmental psychology clearly states that the development of skills to understand, manipulate and coordinate attentional behavior plays a pivotal role for imitation, social cognition and the development of language [8, 15, 35]. And among the complicated cognitive functions of human minds, the ability to form, recognize and understand shared attention is pretty crucial in human social interactions [14, 15, 18]. All human communication, even including linguistic communication, is only possible when the people involved in such communications have built a

*Lifeng Fan and Yixin Chen contributed equally.
†Ping Wei is the corresponding author.
[1]This dataset is available at: http://www.stat.ucla.edu/~lifengfan/shared_attention.

common conceptual ground consisting of shared attention, shared experience, common cultural knowledge, *etc*. [34]. Overall, shared attention is a crucial first step towards social interaction, as well as the primary basis of social intelligence and a precursor of theory of mind [14, 15, 18], language learning [15, 16, 17], the ability of imitation [13] and so on. The study of shared attention is important because it helps a computer vision system to better understand and interpret human activities in images or videos. Robotics equipped with the ability to detect and understand human shared attention can also be more intelligent when interacting with humans.

Despite the importance of this topic, works on shared attention are quite limited in the computer vision community. Some previous works address the problem by using special input data, such as first-person videos taken by multiple head-mounted cameras [2, 22, 23, 31]. Some limited shared attention to the field of Human Robot Interaction [4, 11, 19, 20, 30, 32]. Few works studied shared attention in human social interaction based on third-person social scene videos.

In order to be clarified in our paper with the concept of shared attention, we formulate our problem as follows: shared attention is the gaze focus shared by two or more individuals on one object or human; given a video clip, the task is to detect which frames contain shared attention and where is the shared attention in those frames. To tackle this problem, we collect a new dataset VideoCoAtt and build a deep spatial-temporal neural network with four modules: gaze estimation module, region proposal module, spatial detection module and temporal optimization module. The intuitions for building such a deep neural network architecture are as follows: 1) Firstly, gaze direction, which can be utilized to learn external environment state and internal mental state, is a key feature for shared attention detection. The strongest and most direct indication of human gaze direction is the closeup image patch of human head. We need to detect human heads in videos and predict gaze directions for each detected head. 2) Secondly, gaze direction is of course important, but still not the whole story. Shared attention is more than gaze intersection. According to our definition, there must be an object or human body part as the carrier of shared attention, which means the shared attention detection task is object-driven. Thus, bounding box proposals of object or human body parts, such as laptop, human face, *etc*., is another key feature for our task. We didn't use saliency models (like [21, 37]) because shared attention is more influenced by social group interaction instead of visual importance, and people engaged in shared attention are not free-viewing and may not look at the most salient object in the environment. We use a generic object proposal generation method to generate all potential bounding boxes independent of their categories. 3) Shared attention may last for a while before termination. Temporal information is a good constraint to make the detection results more accurate and robust. The input to our model is just a video clip without any other additional annotation, and the output is a shared attention heatmap for each video frame and the final shared attention prediction results can also be inferred based on the shared attention heatmap.

This paper makes **three major contributions**:

1) It addresses a new problem - inferring shared attention in third-person social scene videos. To the best of our knowledge, this is the first work to deal with such problem in computer vision community.
2) It proposes a spatial-temporal network to address the problem of inferring shared attention in videos. The proposed model explicitly leverages human gaze direction, target region candidates, and temporal inter-frame constraints for identifying shared attention.
3) It presents a large-scale dataset covering diverse social scenes with full annotations, VideoCoAtt, and benchmark results on the dataset for shared attention study.

2. Related Work

The problem of inferring shared attention from third-person videos is closely related to the following works:

Gaze Prediction: Recasens *et al*. proposed a deep learning based model for gaze prediction in images [25] and contributed a dataset called GazeFollow. Given head location, their method extracts head pose and gaze orientation, follows the gaze of the person and identifies the object being looked at in the image. Then they further extended their work to gaze prediction in videos [26] and contributed another new dataset VideoGaze. Given a video clip and the annotations of head and eye location, their model combines gaze pathway, saliency pathway and transformation pathway to predict where a person is looking even when the object being looked at is in a different frame. These works only focus on predicting single-person gaze, while do not consider the task of inferring attention shared by multiple persons in social activities.

Shared attention in Social Interaction: There are some inspiring studies of shared attention in human social interaction. Park *et al*. presented a method to construct a 3D social saliency field and locate multiple gaze concurrences that occur in a social scene from videos taken by head-mounted cameras [22]. After that, they proposed a method to predict social saliency from images or videos captured by multiple first-person view cameras [23]. These works directly study social saliency, which by their definition represents the likelihood of shared attention in a social group. Besides, they also use shared attention as a constraint to predict social behavior in first-person videos, such as individuals' future movements and future gaze directions in a social group. The predicted behaviors reflect an individual physical space that

Figure 2. **Example frames from VideoCoAtt Dataset,** where the shared attentions are annotated by red rectangles and red points. Different groups of people involved in different shared attentions are annotated by rectangles in different colors. Best viewed in color.

affords to take the next actions while conforming to social behaviors by engaging to shared attention [31]. Generally, these work well explored and illustrated shared attention detection and application in social activities. However, they only focus on first-person videos without generalizing to ordinary third-person videos.

Shared attention in HRI: The field of Human-Robot Interaction (HRI) strives to enable easy, intuitive interactions between people and robots, which requires natural communication [1]. Many of the difficulties encountered in human-robot interaction and the communication between autonomous robots could be traced back to unsolved issues related to shared attention [10]. There are many works that try to realize gaze-following and shared attention between robot and human in HRI with or without external evaluation [4, 11, 19, 20, 30, 32]. The key points of these work are inferring human gaze direction and then forming shared attention between robot and human by making the robot head turn to that direction. Our work is beneficial to improve the implementation of shared attention in HRI because robots can further detect, understand and learn to join in the ongoing shared attention in the environment.

3. VideoCoAtt Dataset

In this section we describe our proposed VideoCoAtt dataset, which is specifically designed for studying shared attention in social scenes. Some example frames with annotations are presented in Fig. 2.

Dataset Collection. The following principles drive the collection of our dataset:

• *Natural social interaction.* Shared attention usually occur in daily life naturally. If we deliberately shoot videos for the purpose of shared attention study, then the social interactions performed by the volunteers may seem unnatural and not convincing. Instead, TV show is a good choice because social interactions in TV shows appear to be relatively more natural. As summarized in Table 1, there are some TV show datasets available in the computer vision community, *e.g.*, HMDB [12], TVHI [24], *etc*. However, they are designed for different purposes, like action recognition, human interaction understanding, *etc*., and none of them offer annotations of shared attention. Differently, the proposed VideoCoAtt dataset is carefully collected for studying shared attention in human social activities. The videos are sourced from 20 different TV shows on Youtube.

• *Large scale and high quality.* Both scale and quality are essential to build a long-lifespan benchmark. We carefully collect 380 RGB video sequences from 20 different TV shows or movies. Each video sequence lasts for various time, from around 20s to more than 1 minute with a frame rate of 25 fps. In total, there are 492,100 frames at the spatial resolution of 320×480.

• *Diversity and generality.* The videos in the VideoCoAtt dataset cover different countries and cultures, such as American, Chinese, Indian, European, *etc*. The appearances of actors/actresses, the costume and props vary a lot. There are also diverse scenario settings in VideoCoAtt, including living room, kitchen, restaurant, Cafe, office, outdoor, *etc*. See Table 2 for detailed statistics and Fig. 2 for example frames. Moreover, the number of shared attentions per frame and the number of involved people per shared attention can vary in different frames and videos, as can be seen from the sample frames in Fig. 2 and the statistics in Table 3. This generality in VideoCoAtt dataset is beneficial for the trained model to deal with multiple cases as in real life. Fig. 3 shows the shared attention location distribution averaged over the whole dataset. It appears that shared attention in our dataset tends to lie near the top part of the image frame, as is consistent with previously analyzed eye tracking datasets [39, 33, 9].

Dataset	Year	Format	Size	Annotation	Goal	Shared Attention	Data Source
HMDB [12]	2011	Video	7,000 clips, 51 action categories	Human action	Action recognition	-	Digitized movies, YouTube
TVHI [24]	2012	Video	300 video clips, 30 to 600 frames per clip	Upper body bbx, discrete head orientations, interaction label	Human interaction learning in TV shows	-	23 different TV shows
MPII-MD [28]	2015	Video	94 videos, 68,337 clips	Video description	Automatic video description	-	British Amazon, Hollywood2
GazeFollow [25]	2015	Image	122,143 images, 130,339 people	Eye loc. and gaze loc.	Gaze following in images	-	Actions 40, MS COCO, SUN, PASCAL, *etc.*
VideoGaze [26]	2017	Video	140 movies, 6 frames per movie	Eye loc., head bbx, gaze loc.	Gaze following in videos	-	MovieQA
Sitcom Affordance [38]	2017	Image	11,449 indoor scenes, 28,882 human poses	Human pose	Affordance prediction	-	7 sitcoms
VideoCoAtt (Ours)	2018	Video	380 videos, 492,100 frames	Shared attention bbx, involved head bbx.	Shared attention detection in videos	✓	20 different TV shows

Table 1. **Comparison of several related datasets.** Our dataset is large, diverse and specially designed for shared attention study.

Culture Distribution		Scenario Setting Distribution			
American	44.1 %	Living Room	29.4 %	Dining Room	4.7 %
Chinese	40.7 %	Kitchen	14.3 %	Office	4.7 %
Indian	9.1 %	Restaurant	7.0 %	Bathroom	2.3 %
European	4.1 %	Bedroom	6.8 %	Outdoor	16.4 %
(Others)	2.0 %	Cafe	5.8 %	(Others)	8.6 %

Table 2. **Distributions of culture and scenario settings** in VideoCoAtt dataset.

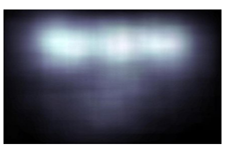


Figure 3. Illustration of shared attention location averaged over VideoCoAtt dataset.

VideoCoAtt	#shared attentions per frame				
	0		1		≥2
#frames	349,468		139,348		3,284
VideoCoAtt	#people involved per S.A.				
	2	3	4	5	≥6
#S.A.	86,988	34,105	16,396	4,955	3,661

Table 3. Statistics regarding to the number of shared attentions per frame and the number of people involved per shared attention in VideoCoAtt dataset.

Dataset Annotation. We manually annotate all the video frames using the online tool Vatic [36]. For each frame, we mark whether there is shared attention in the scene. If there is on-going shared attention in the scene, we mark all the shared attentions with bounding boxes. Only those shared attentions within the view of the scene will be annotated; those out of view or occluded will not be counted as shared attention. Furthermore, for each shared attention, we annotated all the heads that are currently engaged in the certain shared attention using bounding boxes and attributes related to the shared attention numbering.

Dataset Splitting. We split our VideoCoAtt dataset into three parts for training, validation and testing respectively. There are 181 videos (250,030 frames) in the training set, 90 videos (128,260 frames) in the validation set and 109 videos (113,810 frames) in the testing set. To avoid overfitting caused by similarities in human appearances and scenario settings, we split our videos by different sources. Videos for training, validation and testing come from different TV shows, which we believe is necessary and will require a strong generalization ability of our shared attention model.

4. Our Model

Shared attention usually locates at the objects or human body parts gazed by two or more people simultaneously. Obviously, human gaze and target objects in the context environment are essential for inferring shared attention in social scene videos. Thus our shared attention detection model comprises of four modules: 1) the gaze estimation module (§ 4.1) that extracts individual gaze directions to generate a gaze heatmap for the whole scene; 2) the region proposal module (§ 4.1) that extracts region proposals from the context environment; 3) the spatial detection module (§ 4.2) that combines the gaze heatmap and the region proposal map to detect shared attention in spatial space; and 4) the temporal optimization module (§ 4.2) that utilizes inter-frame correlation to optimize the predicted shared attention heatmap in temporal space. An illustration of our whole model architecture is presented in Fig. 4.

4.1. Gaze and Region Proposal Modules

Gaze Estimation Module. Suppose for an input frame I_t in a video sequence $\{I_t\}_{t=1,...,T}$, our head detector outputs a set of head locations $q_{t,i} = (x^{min}_{t,i}, y^{min}_{t,i}, x^{max}_{t,i}, y^{max}_{t,i})$, $i = 1, 2, \ldots, n$, where n could be zero when no head is detected in frame I_t (see the red rectangles in Fig. 5 (a) and (c)). The corresponding closeup image patch for head location $q_{t,i}$ is cropped out from I_t and denoted as $w^h_{t,i}, i = 1, 2, \ldots, n$. We then use a batch of neural network layers $\Psi(\cdot)$ to regress a gaze direction $d_{t,i} \in [-1, 1]^2$ (yellow arrows in Fig. 5 (a) and (c)) for the input image patch $w^h_{t,i}$:

$$d_{t,i} \triangleq (d^x_{t,i}, d^y_{t,i}) = \Psi(w^h_{t,i}). \quad (1)$$

We use a Gaussian distribution to model the variation of a gaze ray with respect to the predicted primary gaze direction $d_{t,i}$, and the probability distribution is

$$P(\theta_{t,i}|d_{t,i}) \propto \frac{1}{\sigma} \exp\{-\frac{\theta^2_{t,i}}{2\sigma^2}\}, \quad (2)$$

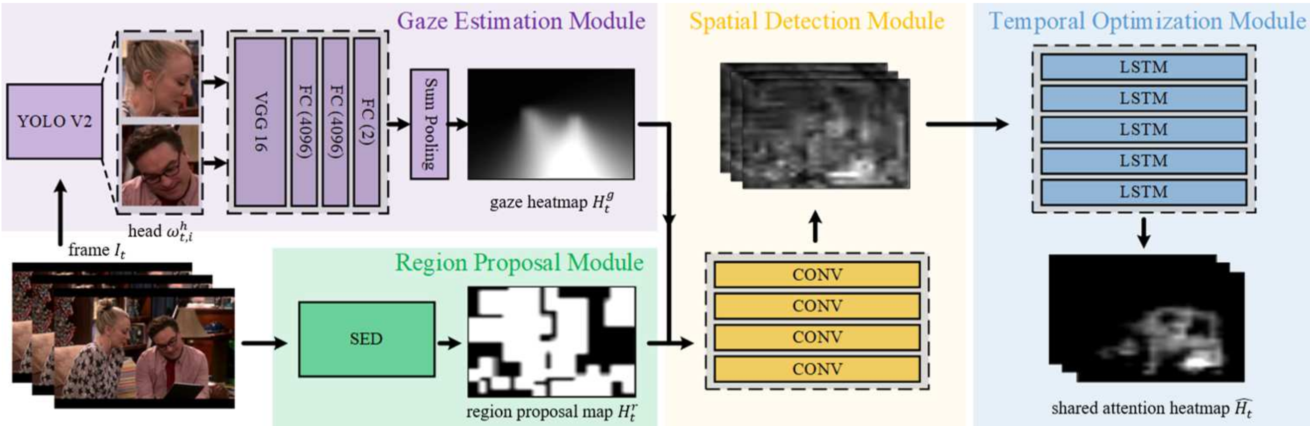


Figure 4. **Illustration of our model architecture.** The gaze estimation module and the region proposal module extract two key features of individuals and the scene context from raw input videos. The subsequent spatial detection module integrates the outputs from the two base modules to perform shared attention detection on a single frame. The temporal optimization module utilizes temporal constraints to optimize the predicted shared attention heatmap.

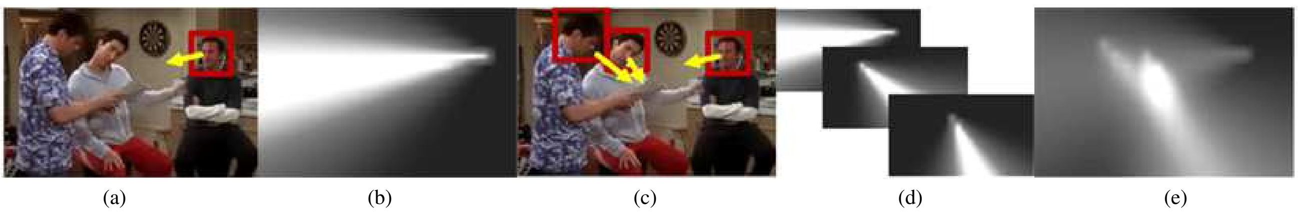


Figure 5. **Illustration of gaze heatmap H_t^g generation procedure.** With detected head position $q_{t,i}$ (red rectangles in (a)(c)) and corresponding predicted gaze direction $d_{t,i}$ (yellow arrows in (a)(c)), we first generate individual gaze heatmap $H_{t,i}^g$ in (b) and (d), and then get the final gaze heatmap H_t^g in (e) via sum-pooling all the gaze heatmaps in (d).

where $\theta_{t,i}$ is the angle between a gaze ray and the predicted primary gaze direction $d_{t,i}$. With detected head position $q_{t,i}$ and corresponding predicted gaze direction $d_{t,i}$, we compute $\theta_{t,i}$ for each grid in the image and then use Eq. 2 to get the probability for this grid to be gazed at by head $q_{t,i}$. After a gaze heatmap $H_{t,i}^g$ (see Fig. 5 (b) and (d)) for each head position $q_{t,i}$ is prepared, we generate the final gaze heatmap H_t^g (Fig. 5 (e)) of size $M \times N$ via Sum-Pooling $\{H_{t,i}^g\}_i$:

$$H_t^g = \sum\nolimits_{i=1}^{n} H_{t,i}^g = \sum\nolimits_{i=1}^{n} \phi(\Psi(w_{t,i}^h), q_{t,i}), \qquad (3)$$

where $\phi(\cdot)$ indicates the gaze heatmap generator based on Eq. 2. More illustrations about the gaze heatmap generation procedure are shown in Fig. 5.

Region Proposal Module. To exploit context information, we use a region proposal module $Z(\cdot)$ to generate a binary region proposal map H_t^r of size $M \times N$ for input image I_t:

$$H_t^r = Z(I_t). \qquad (4)$$

This module is implemented by Structured Edge Detector (SED) [40] to get region bounding boxes $\{b_{t,i}, i = 1, 2, \ldots, m\}$ for each frame I_t and then setting all the pixel values within the bbx proposals to 1 and all other pixel values outside to 0.

4.2. Spatio-temporal Shared Attention Network

The output feature maps of the gaze estimation module and the region proposal module are then fed to the subsequent spatial detection module and temporal optimization module for shared attention detection.

Spatial Detection Module. Shared attention detection is firstly conducted in a frame-by-frame style. We apply a spatial detection module $F(\cdot)$ that consists of several convolutional layers to combine the gaze heatmap H_t^g and region proposal map H_t^r for intra-frame shared attention detection:

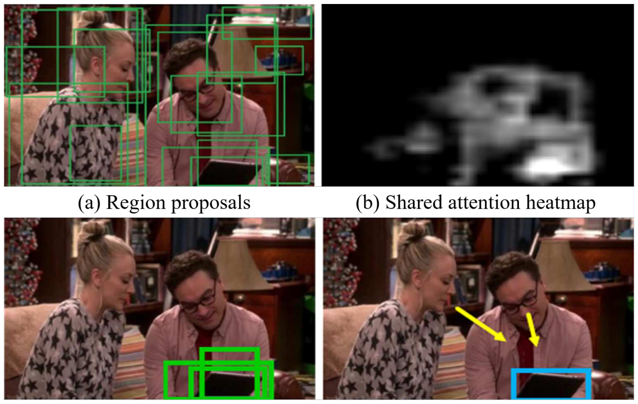
$$\tilde{H}_t = F(H_t^g, H_t^r), \qquad (5)$$

where $\tilde{H}_t$ indicates the intermediate shared attention heatmap output from the spatial detection module.

Temporal Optimization Module. To further exploit the temporal inter-frame constraints in videos, we add a temporal optimization module $LSTM(\cdot)$ that consists of several convolutional Long Short-Term Memory (convLSTM) network [29] layers to optimize the output shared attention heatmap $\tilde{H}_t$:

$$\{\hat{H}_t\}_t = LSTM(\{\tilde{H}_t\}_t), \qquad (6)$$

where $\hat{H}_t$ denotes the eventual shared attention heatmap.



(a) Region proposals (b) Shared attention heatmap

(c) Results after thresholding (d) Final results after NMS

Figure 6. **Illustration of inference process.** Given (a) proposal bounding boxes and (b) shared attention heatmap, we first compute the score for each bounding box by accumulating all the confidence values inside the bounding box. (c) Then we select the bounding boxes with score higher than a certain threshold. (d) NMS is applied for generating final shared attention prediction.

4.3. Learning and Inference

For the loss function, we apply the Mean Squared Error (MSE) between the predicted shared attention heatmap $\hat{H}_t$ and the ground truth shared attention binary map H_t:

$$L(\hat{H}_t, H_t) = \frac{1}{M \cdot N} \parallel \hat{H}_t - H_t \parallel^2, \tag{7}$$

where both $\hat{H}_t$ and H_t are of size $M \times N$.

The inference is possible given the predicted shared attention heatmap $\hat{H}_t$, based on which we can compute the cumulative score for each region proposal bounding box $b_{t,i}$. We only keep those proposal bounding boxes with a score higher than a threshold. Then we conduct a Non-Maximum Suppression (NMS) [6] and treat the remaining bounding boxes as our final shared attention prediction for frame I_t. See Fig. 6 for more detailed illustration.

Since there may be no shared attention or more than one shared attention in a scene, our model is designed to support multimodal predictions instead of regressing a single shared attention location.

4.4. Implementation Details

We implement our model using Keras with Tensorflow as backend. For the gaze estimation module, we first fine-tuned YOLO V2 darknet [27] on our own training set. The re-trained YOLO V2 is applied as a head detector to generate human head image patches $\{\omega^h_{t,i}\}$ for the following gaze direction estimation. We apply the VGG16 network to regress gaze direction, and replace the last fully connected (fc) layer (1000) with a new fc layer of size 2. Then the tanh activation is used for generating a unit gaze direction vector and the gaze direction regression network is fine-tuned on our training set with mean-squared-error loss. To generate the gaze heatmap, we assume that the gaze cone projected from each head is subject to a gaussian distribution with standard deviation $\sigma = 0.5$. For the region proposal module, we use the Structured Edge Detection Toolbox [40] to generate the bounding box proposals for each frame.

The outputs of the gaze estimation module and the region proposal module are of size 28×28. We concatenate the gaze heatmap H^g_t and the region proposal map H^r_t as the input to the spatial detection module, which contains the first three convolutional layers with kernel size 3×3 and output channel size 16, 16, 8 respectively, and the last one convolutional layer with kernel size 1×1, output channel size 1 and sigmoid activation. The output of spatial detection module is a 28×28 shared attention heatmap $\tilde{H}_t$ for each frame. The subsequent temporal optimization module consists of five convLSTM layers. The filter sizes are 40, 40, 40, 40 and 1 respectively. The kernel size is 3×3 for the first four convLSTM layers and 1×1 for the last convLSTM layer. The final convLSTM layer uses sigmoid as activation function.

5. Experiments

5.1. Experimental Setup

We train and evaluate our model on disjoint training, validation and testing sets from VideoCoAtt in our experiments, as described in §3. The ground truth annotations of shared attention bounding boxes and relevant human faces' bounding boxes are only used in training. For testing, the input to our model only includes the raw videos without any additional annotation.

Evaluation Metrics. We use several metrics to compare our model predicted shared attentions with the ground truth shared attention annotations across the testing videos. For the shared attention interval detection task, the percentage of frames with right shared attention existence prediction over all the video frames is applied as a metric ***Prediction Accuracy***. For the shared attention location prediction task, we use the region proposal bounding boxes and shared attention heatmap to generate a ***ROC Curve***, reflecting the precision and recall when predicting shared attention bounding boxes under different score thresholds. ***AUC*** refers to the area under the ROC curve (higher is better). Then given a certain score threshold, the L^2 ***Distance*** (measured in pixel) is the Euclidean distance between the predicted shared attention bbx and the annotated ground truth.

Baseline Methods. We compare our approach against several baselines ranging from simple (Random, Fixed Bias) to more complex (Gaze Follow, Gaze+Saliency, Gaze+Saliency+LSTM) as described below. ***Random***: A weak baseline that draws a Gaussian heatmap with random mean and variance. ***Fixed Bias***: As visible in Fig. 3, there

Model	Prediction Acc.	L^2 Dist.
Raw Img.	52.3 %	188
Only Gaze	64.0 %	108
Only RP	58.0 %	110
Gaze+RP	68.5 %	74
Gaze+RP+Img.	54.0 %	72
Fixed Bias	52.4 %	122
Random	50.8 %	286
Gaze Follow [25]	58.7 %	102
Gaze+Saliency[21]	59.4 %	83
Gaze+Saliency[21]+LSTM	66.2 %	71
Ours (Gaze+RP+LSTM)	**71.4 %**	**62**

Table 4. **Quantitative evaluation results** with Prediction Accuracy and L_2 Distance. As seen, our full model achieves the best performance over the test set of the VideoCoAtt dataset.

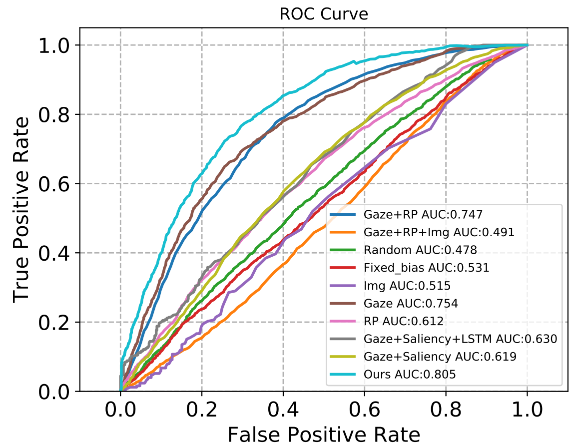


Figure 7. **Quantitative evaluation results** with ROC Curve, computed over the test set of the VideoCoAtt dataset.

exists shared attention location bias in the TV shows. We use a fixed-biased heatmap subject to a 2D Gaussian Distribution with mean and variance learned from our dataset as a baseline to model such bias. ***Gaze Follow***: We apply the gaze following model in [25] to detect all the people's gaze fixations and gaze concurrences in a frame as a baseline. ***Gaze+Saliency*** and ***Gaze+Saliency +LSTM***: We replace our region proposal module with a top-performance saliency model [21], and consider two baselines with and without the temporal optimization module respectively.

Ablation Study. To better understand the importance of each module in our proposed model architecture, we also studied the model performance after removing some modules. ***Raw Img.***: We first only use raw image as input to train an end-to-end model, which means we only keep the spatial detection module. ***Only Gaze***: Then we try to augment the model by adding gaze estimation module to spatial detection module. ***Only RP***: We also tested the architecture with only region proposal module and spatial detection module. ***Gaze+RP***: We add both gaze estimation and region proposal modules before spatial detection module. ***Gaze+RP+Img.***: This is a variation of our model that uses gaze, region proposal and raw image feature as input to spatial detection module without using temporal optimization module.

5.2. Results and Analysis

Quantitative results. Table 4 shows the comparison of our model with baseline methods and several ablation models by two evaluation metrics *Prediction Accuracy* and L^2 *Distance*. Our model achieves the best performance in both the shared attention interval detection task (Prediction Acc.: 71.4%) and the shared attention location prediction task (L^2 Dist.: 62).

Among all the baseline models, the second best model is *Gaze+Saliency+LSTM* with a Prediction Acc. of 66.2% and a L^2 Dist. of 71. The replacement of region proposal module with a saliency model impairs our model performance because the shared attention of people in a social interaction may not be the most visually salient object in the scene, but more influenced by the on-going interaction. The performance of the Gaze Follow baseline in detecting shared attention is mediocre, which is mainly because that shared attention of a social group is goal-driven and object-related, not just the concurrence of human gazes.

Among all the ablation models, *Gaze+RP* shows a overall best performance (Prediction Acc.: 68.5% and L^2 Dist.: 74), but is still inferior to our full model with all the four modules. And overall *Only Gaze* performs better than *Only RP*, indicating the gaze estimation module plays a more important role than the region proposal module in shared attention detection, which is consistent with our intuitions. The simplest model without any module design *Raw Img.* performs worst. The ablation study shows that each of the four modules proposed by our model (§ 4) is important and necessary for shared attention detection in videos.

Fig. 7 shows the ROC Curve and AUC comparison results among our full model, baseline models and ablation models. Our model has the best precision and recall performance and the largest AUC value than all the other models. *Gaze+RP* and *Gaze* also perform significantly better than the remaining models. The result further confirms the significance and effectiveness of our model architecture design. The gaze direction feature and the region proposal feature as well as the temporal constraints indispensably help our model to gain great performance improvements in the task of inferring shared attention in social scene videos.

Qualitative results. Fig. 8 exhibits an internal visualization of shared attention detection results by our full model on some example frames. The *Gaze Heatmap* roughly features the attention of each individual in the social scene and is not enough to accurately feature shared attention. The *Region Proposal Map* gives some potential shared attention

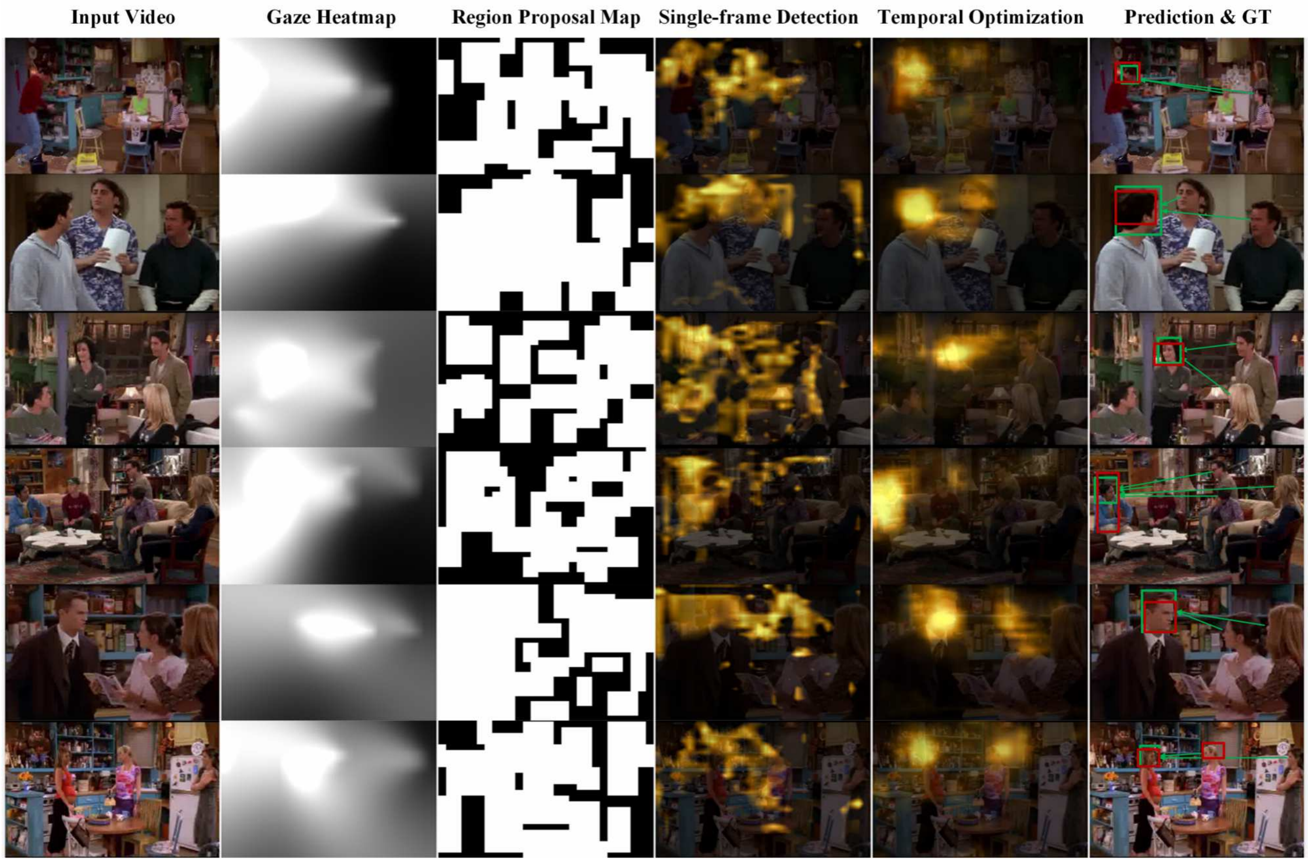


Figure 8. **Shared attention detection results on example frames.** With the input video frames, we show the outputs of the gaze estimation module and the region proposal module in the second and third columns. The *Single-frame Detection* column shows the shared attention heatmap $\tilde{H}_t$ trained on a single frame. The *Temporal Optimization* column shows the eventually optimized shared attention heatmap $\hat{H}_t$. Our final prediction results (red rectangles) and the ground truth annotations (green rectangles) are presented in the last column.

proposals and provides the important spatial constraints. *Single-frame Detection* combines the *Gaze Heatmap* and the *Region Proposal Map* to generate a preliminary shared attention heatmap, which still has too much noises. After the *Temporal Optimization* by convLSTM, the shared attention heatmap is much clearer and can provide more accurate shared attention distribution information. The final column in Fig. 8 compares our eventual shared attention prediction results (depicted in red rectangles) with the ground truth shared attention annotations (depicted in green rectangles). As shown, there are good predictions that can exactly locate the shared attention in the social scenes, like the prediction in the first example. However, there are also some false alarms existing. For example, The scene in the last row actually has only one shared attention, but our model gives two predictions located near the two human faces. This is an interesting failure example since whether the third person on the right side is looking at the person on the left side or the person in the middle is somehow ambiguous for our model to distinguish. That's why the shared attention heatmap gets two peaks for this example. But similar situation in the fifth scene is successfully solved by our model.

6. Conclusion

This paper addresses a new problem of inferring shared attention in third-person social scene videos. Although shared attention is common in daily life and important for social interactions, relevant studies are quite limited in the computer vision community. We propose a dataset VideoCoAtt and a model to detect shared attention in videos. Our model combines individual gaze features and context region proposal features from the raw video inputs. Based on the two bottom features, our model learns to spatially detect and temporally optimize shared attention in videos. Although we get some reasonable results in the experiments, we are still far from completely solving this problem. We hope our dataset and model will serve as important resources to facilitate future studies related to this topic.

Acknowledgement. The authors would like to thank Xiaofeng Gao and Tianmin Shu from UCLA Statistics Department for helpful support and insightful suggestions. This work was supported by ONR MURI project N00014-16-1-2007, DARPA XAI Award N66001-17-2-4029, and NSF IIS 1423305.

References

[1] H. Admoni and B. Scassellati. Social eye gaze in human-robot interaction: A review. *J. Hum.-Robot Interact.*, 6(1), 2017. 3

[2] I. Arev, H. S. Park, Y. Sheikh, J. K. Hodgins, and A. Shamir. Automatic editing of footage from multiple social cameras. *ACM Trans. Graph.*, 33(4), 2014. 2

[3] S. Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind.* MIT Press, 1995. 1

[4] R. R. da Silva, C. A. Policastro, and R. A. F. Romero. Relational reinforcement learning applied to shared attention. In *IJCNN*, pages 2943–2949, 2009. 2, 3

[5] N. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6):581 – 604, 2000. 1

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–1645, 2010. 6

[7] M. S. Gazzaniga. *The new cognitive neurosciences: second edition*. The MIT Press, 1999. 1

[8] P. Hobson. *The cradle of thought: challenging the origins of thinking*. MacMillan: London, UK., 2002. 1

[9] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009. 3

[10] F. Kaplan and V. Hafner. The challenges of joint attention. *Interaction Studies*, 7(2):135–169, 2006. 3

[11] H. Kim, H. Jasso, G. Dek, and J. Triesch. A robotic model of the development of gaze following. pages 238 – 243, 2008. 2, 3

[12] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011. 3, 4

[13] M. Kumashiro, H. Ishibashi, Y. Uchiyama, S. Itakura, A. Murata, and A. Iriki. Natural imitation induced by joint attention in japanese monkeys. *International Journal of Psychophysiology*, 50(1):81 – 99, 2003. 2

[14] C. Moore and V. Corkum. Social understanding at the end of the first year of life. *Developmental Review*, 14(4):349 – 372, 1994. 1, 2

[15] C. Moore and P. J. Dunham. *Joint Attention: Its Origins and Role in Development.* Lawrence Erlbaum Associates, 1995. 1, 2

[16] M. Morales, P. Mundy, and J. Rojas. Following the direction of gaze and language development in 6-month-olds. *Infant Behavior and Development*, 21(2):373 – 377, 1998. 2

[17] P. Mundy and A. Gomes. Individual differences in joint attention skill development in the second year. *Infant Behavior and Development*, 21(3):469 – 482, 1998. 2

[18] Y. Nagai. *Understanding the Development of Joint Attention from a Viewpoint of Cognitive Developmental Robotics*. PhD thesis, Osaka University, 2004. 1, 2

[19] Y. Nagai. The role of motion information in learning human-robot joint attention. In *ICRA*, 2005. 2, 3

[20] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. A constructive model for the development of joint attention. *Connect. Sci.*, 15:211–229, 2003. 2, 3

[21] J. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, and N. E. OConnor. Shallow and deep convolutional networks for saliency prediction. In *CVPR*, 2016. 2, 7

[22] H. S. Park, E. Jain, and Y. Sheikh. 3d gaze concurrences from head-mounted cameras. In *NIPS*, 2012. 2

[23] H. S. Park and J. Shi. Social saliency prediction. In *CVPR*, 2015. 2

[24] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman. Structured learning of human interactions in tv shows. *IEEE TPAMI*, 34(12):2441–2453, 2012. 3, 4

[25] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba. Where are they looking? In *NIPS*, 2015. 2, 4, 7

[26] A. Recasens, C. Vondrick, A. Khosla, and A. Torralba. Following gaze in video. In *ICCV*, 2017. 2, 4

[27] J. Redmon and A. Farhadi. YOLO9000: Better, faster, stronger. In *CVPR*, 2017. 6

[28] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In *CVPR*, 2015. 4

[29] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NIPS*. 2015. 5

[30] A. P. Shon, D. B. Grimes, C. L. Baker, M. W. Hoffman, S. Zhou, and R. P. N. Rao. Probabilistic gaze imitation and saliency learning in a robotic head. In *ICRA*, 2005. 2, 3

[31] S. Su, J. P. Hong, J. Shi, and H. S. Park. Social behavior prediction from first person videos. *arXiv preprint arXiv:1611.09464*, 2016. 2, 3

[32] H. Sumioka, K. Hosoda, Y. Yoshikawa, and M. Asada. Acquisition of joint attention through natural interaction utilizing motion cues. *Advanced Robotics*, 21(9):983–999, 2007. 2, 3

[33] B. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4–4, 2007. 3

[34] M. Tomasello. *Origins of Human Communication*. The MIT Press, 2008. 2

[35] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675691, 2005. 1

[36] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *IJCV*, 101(1):184–204, 2013. 4

[37] W. Wang and J. Shen. Deep visual attention prediction. *IEEE TIP*, 27(5):2368–2378, 2018. 2

[38] X. Wang, R. Girdhar, and A. Gupta. Binge watching: Scaling affordance learning from sitcoms. In *CVPR*, 2017. 4

[39] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32–32, 2008. 3

[40] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 5, 6