Patch-level Gaze Distribution Prediction for Gaze Following: Supplementary Material

Qiaomu Miao Minh Hoai Dimitris Samaras
Stony Brook University, Stony Brook, NY 11794, USA
{qiamiao, minhhoai, samaras}@cs.stonybrook.edu

In this supplementary material, we provide additional information for further understanding of our method:

- Section 1 provides the implementation details.
- Section 2 explains the detailed structure of the feature extraction module.
- Section 3 shows the results of different alternatives to the patch distribution creation method.
- Section 4 analyzed the level of consistency between the predicted heatmaps and patch distributions.
- Section 5 shows the comparison between our model and the VideoAtt model both with and without depth input.
- Section 6 shows some example failure cases of our model.

1. Implementation Details

In all our experiments, all input images are resized to 224×224. Both the scene backbone and head backbone are ResNet-50 [5] followed by an additional residual layer and average pooling layer for dimensionality reduction. The output feature dimensions from both backbones are $1024 \times 7 \times 7$. Same with the VideoAtt model [2], the head backbone is initialized with weights pretrained on the Eyediap dataset [4], and the scene backbone is initialized with pretrained weights on the Places dataset [12]. The encoder has two convolutional layers with kernel sizes of 1×1, which reduce the channels from 2048 to 512. Therefore, the number of patch tokens is $7 \times 7 + 1 = 50$. We set $\sigma = 3$ for generating the ground truth heatmap following the default setting of previous models [7, 2].

As the procedure used in VideoAtt [2] and DualAtt [3], the model is first trained on the GazeFollow dataset until convergence, and then finetuned on the VideoAttentionTarget dataset. Adam [6] was used to optimize the model with a learning rate of 2.5e-4, which is decreased with a decay factor of 0.2 at the 25th, 31st, and 40th epochs on the GazeFollow dataset. For finetuning on the VideoAttentionTarget dataset, we used a 5-frame sequence as one sample. The weights until the patch attention module are frozen, and the rest of the network is trained with a learning rate of 1e-4, with a decay factor of 0.5 at the 3rd and 6th epochs. The batch sizes for training on GazeFollow and VideoAttentionTarget are 80 and 16 respectively.

2. Structure of the Feature Extraction Module

The structure of the feature extraction module is shown in Figure 1. We leverage the feature extraction component of the VideoAtt model [2] for feature extraction, with some small modifications. The feature extraction module consists of two branches: a scene branch and a head branch. In the scene branch, the scene backbone $\mathcal{F}_s(\cdot)$, takes the scene image $I \in R^{3\times H_0\times W_0}$, the binary head position mask $P \in \{0,1\}^{H_0\times W_0}$, and a normalized depth map $D \in [0,1]^{H_0\times W_0}$ as input, and output the scene feature $f_s \in R^{C\times H\times W}$. We leveraged an additional depth map as input according to the insight from the DualAtt model [3] to incorporate scene depth information, which is computed with an off-the-shelf monocular depth estimation model [9]. The head backbone $\mathcal{F}_h(\cdot)$ takes the head crop of the person $H \in R^{3\times H_0\times W_0}$ as input, and output the head feature $f_h \in R^{C\times H\times W}$. The average pooled head feature is concatenated with the downsampled head mask and the depth map and fed into an attention layer to generate a spatial attention map $M_s \in R^{H\times W}$, which is multiplied with the scene feature and concatenated with the face feature:

$$f_{cat} = [M_s \otimes f_s, f_h], \tag{1}$$

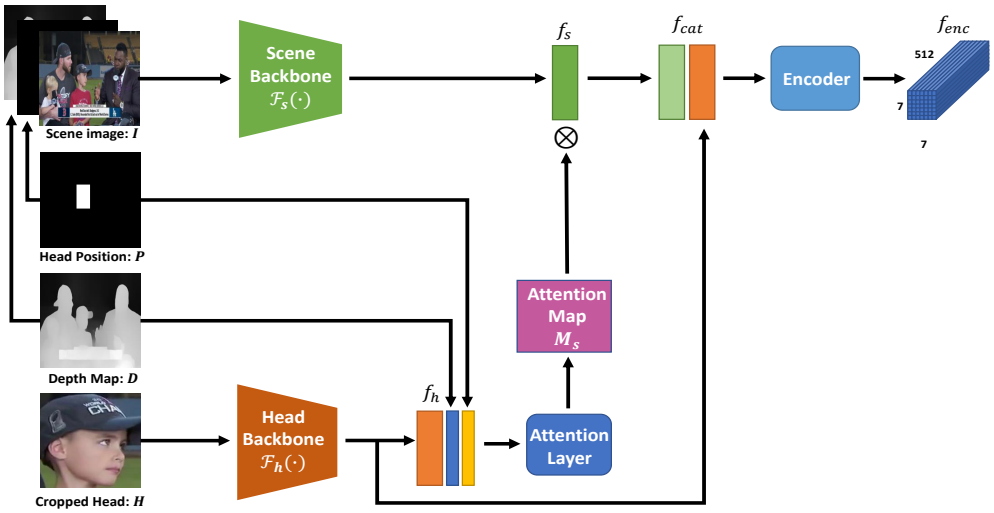

Figure 1: Structure of the feature extraction module. $\otimes$ indicates element-wise multiplication. Please see text for details.

where $[\cdot, \cdot]$ denotes concatenation operation, and $\otimes$ denotes element-wise multiplication on each channel in f_s. The Attention layer is a fully connected layer mapping from the dimension of the concatenated feature to the size of $H \times W$. The VideoAtt model uses two separate encoders for the heatmap prediction and in/out prediction branch after f_{cat}. In contrast, we used a single encoder with two convolutional layers to extract shared feature encoding $f_{enc} \in R^{C \times H \times W}$, as we expect the PDP task to benefit the learning of the shared feature by merging the two subtasks.

3. Ablations of the Ground Truth Patch Distribution Creation Method

In Table 1 we provide the results of ablations of the ground truth patch distribution creation method. First, we tested choosing different numbers of patches for the patch-level gaze distribution. As the number of patches corresponds to the number of tokens in f_{enc} after the feature extraction module, we only tested in the scale of 2 (4×4 and 14×14) for ease of implementation. Setting the patch number to 4×4 has a large drop in performance on GazeFollow, possibly because the overly coarse scale feature encodings make the target estimation in finer grain difficult. When using a patch number of 14×14, despite lower drop in performance on GazeFollow, the model performance on VideoAttentionTarget drops obviously. We infer that this is because when training on VideoAttentionTarget (which has lots of outside cases), the much larger number of inside tokens (4 times of 7×7) make the patch distribution prediction difficult when combined with the outside token.

Method	GazeFollow			VideoAttentionTarget		
	AUC ↑	Dist. ↓		*In frame*		*Out of frame*
		Avg.	Min.	AUC ↑	Dist ↓	AP ↑
Patch: 4×4	0.921	0.129	0.071	0.911	0.110	0.897
Patch: 14×14	0.930	0.127	0.067	0.908	0.113	0.892
MaxPool → AvgPool	0.931	0.124	0.066	0.913	0.112	0.903
One-hot	0.925	0.128	0.071	0.903	0.110	0.884
Ours	**0.934**	**0.123**	**0.065**	**0.917**	**0.109**	**0.908**

Table 1: Ablations of Patch Distribution Creation Method

We also trained the model using average pooling instead of max pooling to get the ground truth patch distribution. The model still shows good performance but with a slight drop in all metrics. Finally, replacing our PDP task with the one-hot patch classification as in [11] caused a significant drop in performance, showing little improvement compared to the VideoAtt_depth model. In Figure 2, we also visualized the distributions generated by different alternatives from one sampled

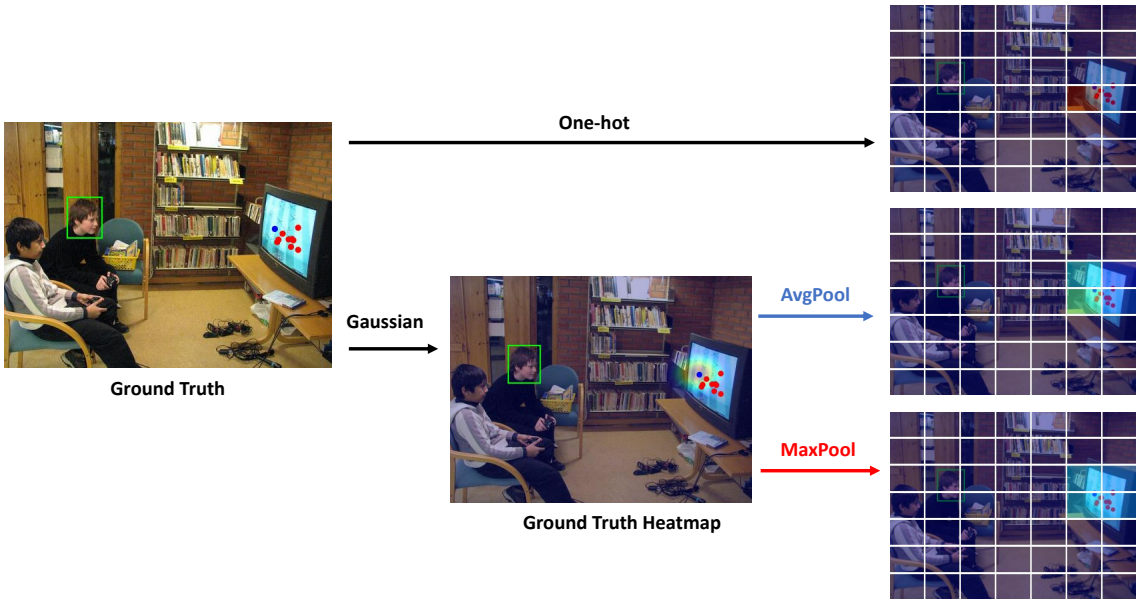

Figure 2: Visualizations of patch distributions generated in different ways from a sampled ground truth annotation (blue) on an example image in the GazeFollow test set. Annotations from other annotators are visualized in red. In the case of an annotation point being close to the patch boundary, the one-hot design simply regards the neighboring patches as unattended, while our max pooling method can generate much better distribution with high response in different patches, encouraging the model to predict multi-modal heatmaps in inference.

annotation (blue) on an example image in the GazeFollow test set. We assume in training, only the sampled annotation is available. It can be seen that the one-hot generation method assigns a hard label of 1 for the specific patch in which the annotated point is located, which is unimodal and cannot cover the adjacent patches if the target is located close to the patch boundaries. This limits the model's capability to predict multi-modal heatmaps. When discretizing the ground truth heatmap with average pooling, a smoother patch distribution can be obtained. However, the patch where the point is located still has a much higher response than the right neighboring patches, despite the right patches being very close to the point. Our method of max pooling generates the distribution that aligns well with the group-level human annotations by creating higher confidences in multiple patches, showing the best potential to make the model generate multi-modal outputs.

4. Consistency between the Heatmap and Patch Distribution Predictions

As mentioned in the main paper, we used PDP as a regularization method for heatmap prediction, and the predicted heatmap is used as the final output for the target prediction task if the target is located inside the image. However, after visualizing the outputs of our model as in Figure 6 in the main paper, we found high consistency between the predicted heatmaps and the patch distribution. To further investigate the level of consistency between the predictions, we created patch distributions from the predicted heatmaps using our method for creating the ground truth patch distribution from the ground truth heatmap. We call this created patch distribution as "Patch Distribution from Predicted Heatmap" (PDPH). We computed the similarities and distances between the patch distribution predicted from the patch prediction head and PDPH from the heatmap prediction head on the test set of GazeFollow and VideoAttentionTarget. We selected the Bhattacharyya coefficient [1] and the Jensen-Shannon (JS) divergence [8] as the evaluation metrics due to their suitability for computing similarity or distances between distributions. Note that we used JS divergence here instead of KL divergence due to its symmetrical property as we do not have 'ground truth' patch distribution here.

As shown in Table 2, the Bhattacharyya coefficients show very high value on both datasets while the values of JS divergence are very low. This high consistency between the outputs further demonstrates that our model design and the ground truth patch distribution creation method can regularize the heatmap prediction by acting on the common feature embedding before the prediction heads.

Dataset	Bhat. Coef. ↑	JS Div. ↓
GazeFollow	0.976	0.142
VideoAttentionTarget	0.971	0.158

Table 2: Analysis of Consistencies between Patch and Heatmap Predictions.

5. Model Performance with and without Depth

In Figure 3, we visualize the outputs of our model and VideoAtt model [2] with and without a depth map as input, to get a better understanding of the effect of the patch distribution prediction (PDP) method and the depth information. Our model can generate heatmap predictions better aligned with human annotations compared to the VideoAtt model, both with and without a depth map as input. This shows that the PDP task can regularize the heatmap regression task irrespective of depth information. However, without a depth map as input, it becomes more difficult for our model to predict the perfect target location. As shown in the figure, our model may predict heatmap confidence on objects inconsistent with human gaze in the depth channel (rows 1 and 4), or fail to predict on some potential gazed objects with confidence (rows 3 and 4). The depth map gives the model a much better understanding of the scene structure, making it easier for the model to infer the potential gazed objects.

6. Example Failure Cases

Figure 4 shows some example failure cases of our model on the GazeFollow [10] and VideoAttentionTarget [2] datasets. Our model sometimes predicts the gaze target incorrectly when the person has a subtle eye orientation that is inconsistent with the head pose, or predicts more confidently a person's head instead of the actual target (row 1 and row 3). This phenomenon may be attributed to the dataset statistics that the gaze target is located on a person's head in a large number of cases. In addition, our model only takes the cropped head without extracting the cropped eye region as input, which makes it easier to employ the model but sacrifices accuracy to some extent. We would like to make our model easier to be applied to most in-the-wild data, without using a complex pre-processing step to crop the eye images, as in the DualAtt model[3].

In some other circumstances, our model predicts multiple clusters due to the uncertainty of the input, but the predicted target determined from the maximum point in the heatmap is different from the annotation, or where most of the annotations lie, as shown in the 2nd and 4th row in Figure 4. Still, our model can predict heatmap response at some level in the annotated regions. Properly speaking, our model's predictions are not totally "wrong" in these cases and the predicted target determined from our heatmap still makes some sense. It is possible that the performance of the model increases if more annotations are obtained.

References

[1] Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

[2] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5406, 2020.

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11390–11399, 2021.

[4] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc 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*, pages 255–258, 2014.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[7] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer, 2018.

[8] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[9] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[10] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[11] Zehua Zhang, David J Crandall, Chen Yu, and Sven Bambach. From coarse attention to fine-grained gaze: A two-stage 3d fully convolutional network for predicting eye gaze in first person video. In *BMVC*, page 295, 2018.

[12] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. 2014.

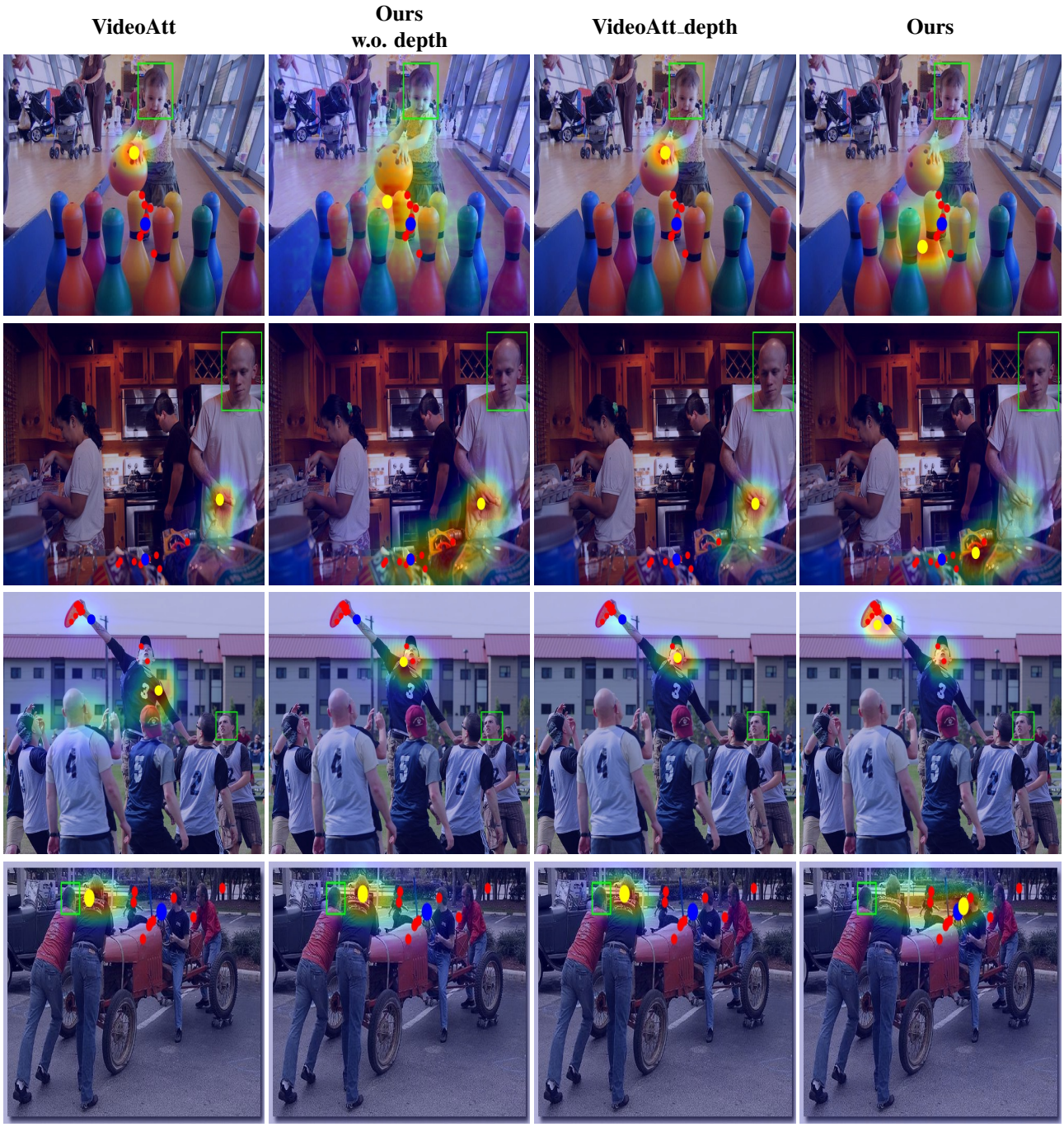

Figure 3: Visualizations of the output heatmaps of our model and the VideoAtt model with and without a depth map as input. The left two columns are the model predictions without using a depth map as input, and the right two columns are the model predictions using a depth map as input. Predicted targets are plotted in yellow and the ground truth annotations are plotted in red. The average annotation is plotted in blue.

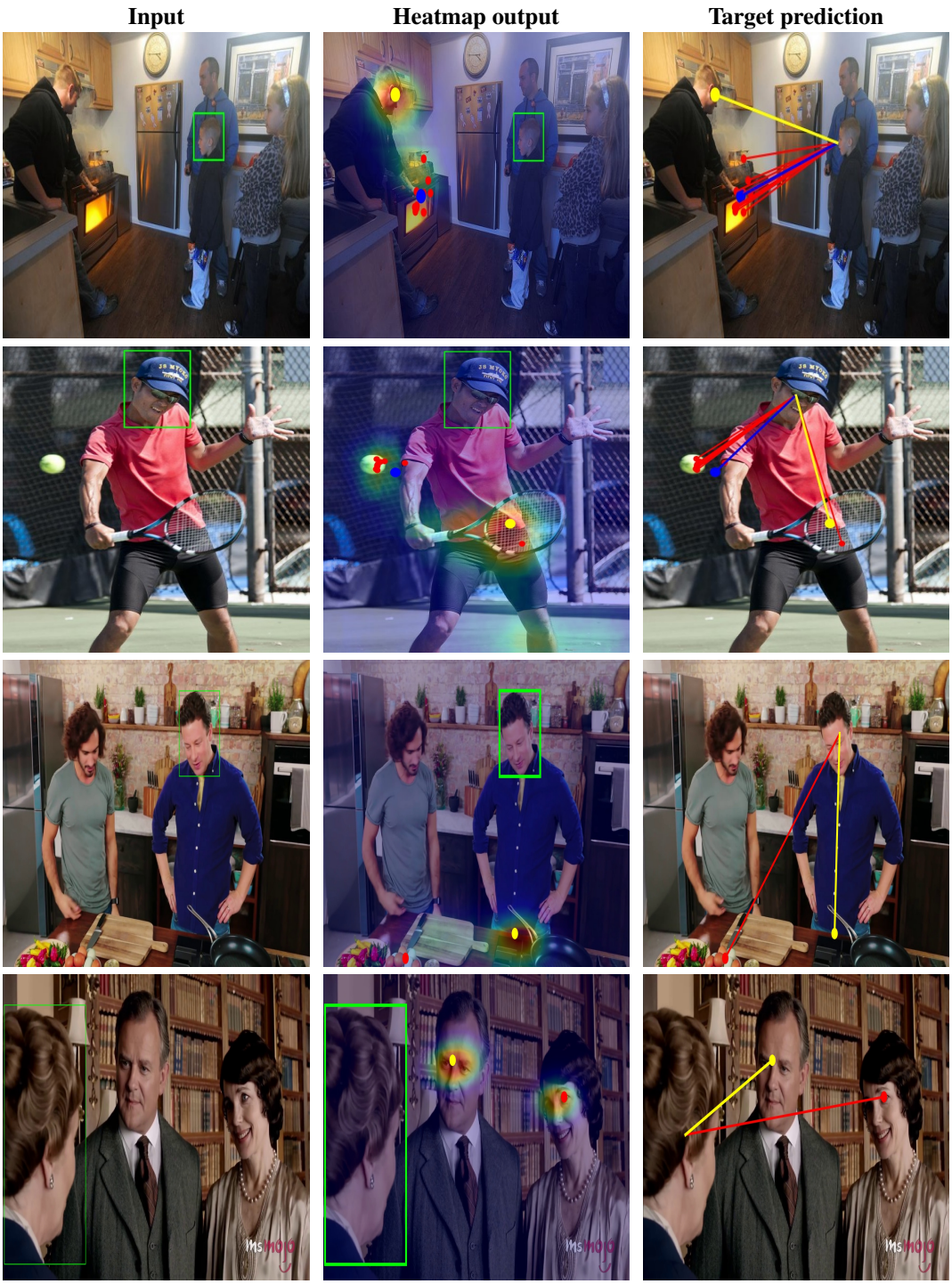

Figure 4: Example failure cases of our model on the GazeFollow (row 1 & 2) and VideoAttentionTarget (row 3 & 4) datasets. Predicted targets are plotted in yellow and the ground truth annotations are plotted in red. The average annotations for images in the GazeFollow dataset are plotted in blue