Eyelid's Intrinsic Motion-aware Feature Learning for Real-time Eyeblink Detection in the Wild

Wenzheng Zeng, Yang Xiao, Guilei Hu, Zhiguo Cao, Sicheng Wei, Zhiwen Fang, Joey Tianyi Zhou, and Junsong Yuan, *Fellow, IEEE*

***Abstract*—Real-time eyeblink detection in the wild is a recently emerged challenging task that suffers from dramatic variations in face attribute, pose, illumination, camera view and distance, etc. One key issue is to well characterize eyelid's intrinsic motion (i.e., approaching and departure between upper and lower eyelid) robustly, under unconstrained conditions. Towards this, a novel eyelid's intrinsic motion-aware feature learning approach is proposed. Our proposition lies in 3 folds. First, the feature extractor is led to focus on informative eye region adaptively via introducing visual attention in a coarse-to-fine way, to guarantee robustness and fine-grained descriptive ability jointly. Then, 2 constraints are proposed to make feature learning be aware of eyelid's intrinsic motion. Particularly, one concerns the fact that the inter-frame feature divergence within eyeblink processes should be greater than non-eyeblink ones to better reveal eyelid's intrinsic motion. The other constraint minimizes the inter-frame feature divergence of non-eyeblink samples, to suppress motion clues due to head or camera movement, illumination change, etc. Meanwhile, concerning the high ambiguity between eyeblink and non-eyeblink samples, soft sample labels are acquired via self-knowledge distillation to conduct feature learning with finer supervision than the hard ones. The experiments verify that, our proposition is significantly superior to the state-of-the-art ones (i.e., advantage on F1-score over 7%) and with real-time running efficiency. It is also of strong generalization capacity towards constrained conditions. The source code is available at https://github.com/wenzhengzeng/blink_eyelid.**

***Index Terms*—Eyeblink detection in the wild, eyelid's intrinsic motion, visual attention, self-knowledge distillation**

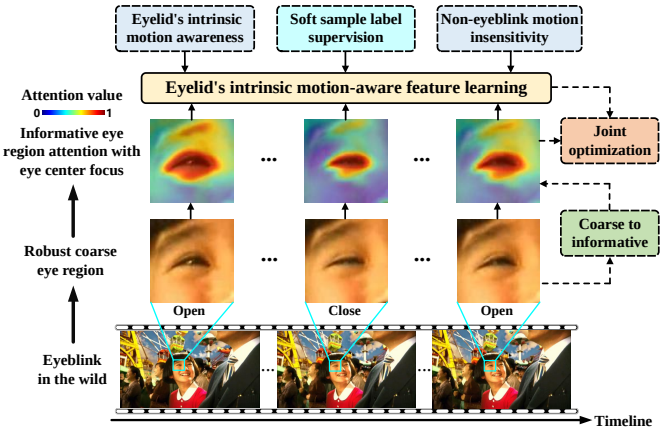


Fig. 1: The main research idea of our eyelid's intrinsic motion-aware feature learning approach. We propose a coarse-to-fine eye region discovery method to lead the feature extractor focus on the informative fine eye region with rich eyelid movement robustly. The constraints concerning eyelid's intrinsic motion awareness and non-eyeblink motion insensitivity are proposed for a better eyelid feature learning. We also introduce soft label supervision to alleviate the ambiguity between eyeblink and non-eyeblink samples.

I. Introduction

EYEBLINK detection in the wild is a recently emerged challenging research task [1], with a wide range of applications on drive fatigue detection [2], deception detection [3], face anti-spoofing [4], etc. Compared with eyeblink detection under constrained indoor conditions [4]–[7] where volunteers remain relatively still facing the camera under well-lit conditions, eyeblink detection in the wild places a stronger emphasis on the unconstrained nature of various practical scenarios. For example, as shown in Fig. 2, it may suffer from more dramatic variations in face attribute, pose, illumination, camera view and distance, as well as motion blurring and sample ambiguity. These actually lead to the unsatisfactory performance of the existing state-of-the-art eyeblink detection approaches [1], [6], [8], [9] towards "in the wild" cases.

Eyeblink can be essentially characterized by eyelid's intrinsic motion (i.e., approaching and departure between the upper and lower eyelid). However, well capturing this representative clue under unconstrained conditions is not trivial. The existing eyeblink detection approaches can be generally categorized into facial landmark-based [5], [6], [10] and region-based [1], [8], [11], [12] paradigms. The former one is good at revealing eyelid's fine status, while suffering from the potential landmark localization failure risk even with state-of-the-art manners [13]–[17]. On the other hand, region-based methods

Wenzheng Zeng, Yang Xiao, Zhiguo Cao and Sicheng Wei are with National Key Laboratory of Multispectral Information Intelligent Processing Technology, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China. E-mail: wenzhengzeng, Yang_Xiao, zgcao, sichengwei@hust.edu.cn.

Guilei Hu is with Shanghai Dingshan Information Technology Co., Ltd, Shanghai. E-mail: lishang19941020@gmail.com

Zhiwen Fang is with the School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China. He is also with the Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou 510515, China, the Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou 510515, China, and the Department of Rehabilitation Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou 510280, China. E-mail: fzw310@smu.edu.cn.

Joey Tianyi Zhou is with Centre for Frontier AI Research (CFAR), A*STAR, Singapore, and Institute of High Performance Computing (IHPC), A*STAR, Singapore. E-mail: zhouty@cfar.a-star.edu.sg.

Junsong Yuan is with the Computer Science and Engineering Department of University at Buffalo, the State University of New York, USA. E-mail: jsyuan@buffalo.edu.

Yang Xiao is the corresponding author of this paper.

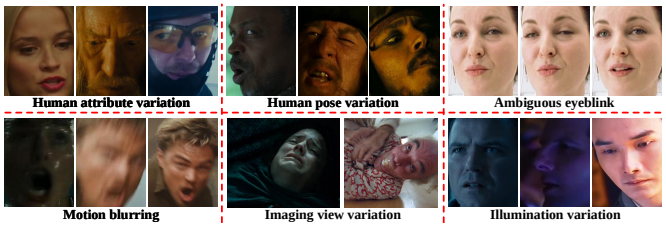


Fig. 2: The challenges towards eyeblink detection in the wild.

are generally of stronger robustness, but they still face the non-trivial issue of setting suitable local eye region size. That is, small region can well capture fine descriptive clues while suffering from unexpected missing coverage on eyes due to the unreliable landmark detection. Large region benefits robustness, but may involve more background information distracting for eyelid's status description. Eye region of box form is also not optimal for revealing eyelid's fine motion due to the rigidness. Last but not least, the existing approaches generally do not well concern the specific characteristics of eyelid's intrinsic motion to facilitate feature extraction.

To address the issues above, an eyelid's intrinsic motion-aware feature learning approach for eyeblink characterization in the wild is proposed by us. Particularly, we pay research efforts to answer 2 essential questions for effective eyeblink representation: *(1) how to lead feature extractor to focus on the informative fine eye region robustly?* and *(2) how to make feature learning be aware of eyelid's intrinsic motion?*

For the first question, a coarse-to-fine informative eye region discovery method is proposed via introducing visual attention mechanism to eyeblink detection task for the first time. Particularly, during the coarse stage a large coarse eye region is first extracted to ensure high coverage on eye. Compared with the existing region-based counterparts [1], [11], [12], our coarse eye region is of much larger size to better guarantee robustness. Within the consequent fine stage, adaptive informative eye region attention is imposed to the acquired coarse region, for extracting eyelid's fine motion feature and resisting background. During this, explicit constraints are designed to help the deformable attention map focus more on the center eye region. This can generally help to capture eyelid's fine motion across eye center during eyeblink.

Towards the second question, 2 constraints that concern eyelid's intrinsic motion awareness are proposed to guide eyeblink feature learning. The first one concerns the fact that the inter-frame feature divergence within eyeblink processes should be greater than the non-eyeblink ones, due to eyelid's intrinsic motion when eyeblink happens. The second one aims to resist non-eyeblink motion caused by head or camera movement, illumination change, facial expression variation, etc. The goal is achieved by minimizing the inter-frame feature divergence of non-eyeblink samples, to reveal non-eyeblink motion insensitivity. It is worth noting that, within our proposition the generation of attention maps and eyeblink feature learning are jointly optimized in an end-to-end learning manner to seek the optimal performance.

Additionally, concerning the high ambiguity between eyeblink and non-eyeblink samples, we further propose to use soft sample labels as the supervision to conduct feature learning for better revealing eyeblink's fine-grained characteristics instead of the hard ones. Particularly, the soft sample labels are self-generated via the teacher network under a self-knowledge distillation framework [18]. To our knowledge, we are the first to address the sample ambiguity problem for eyeblink detection. Overall, the main research idea of our proposition is shown in Fig. 1.

To verify the effectiveness and efficiency of our proposition, it is tested on one "in the wild" dataset (HUST-LEBW [1]) and one constrained dataset termed "Constrained+" built by us via combing 4 existing constrained eyeblink datasets [4], [5], [7], [19]. Actually, our method outperforms the state-of-the-art eyeblink detection approaches by a large margin (i.e., advantage on F1-score over 7%) towards "in the wild" cases and can also be well adapted to the constrained conditions, with real-time running efficiency (about 41 FPS). Besides, our model trained on Constrained+ can be directly applied to HUST-LEBW with promising performance. This verifies the strong generalization capacity of our proposition.

Overall, the main contributions of this paper include:

- An eyelid's intrinsic motion-aware feature learning approach is proposed for real-time eyeblink detection in the wild. Within it, informative eye region attention is introduced with eye center focus constraints;
- The constraints on eyelid's intrinsic motion awareness and non-eyeblink motion insensitivity are proposed to guide eyeblink feature learning;
- Self-generated soft sample label is applied to address the ambiguous categorization problem in eyeblink detection.

The remaining of this paper is organized as follows. Sec. II discusses the related work. Our eyeblink feature learning method is illustrated in Sec. III. Then, the ambiguous problem between eyeblink and non-eyeblink samples is addressed in Sec IV with soft sample label self-generation. The implementation details are given in Sec. V. Experiments are conducted in Sec. VI. Sec. VII concludes the whole paper.

II. Related Work

Here, the related works for eyeblink detection in the wild task on *dataset*, *eyeblink detection approach*, *visual attention* and *knowledge distillation* will be introduced respectively.

Eyeblink detection dataset. The existing datasets (e.g., ZJU [4], Eyeblink8 [7], Talking Face [19], RLDD [5], and mEBAL [8]) generally focus on constrained indoor cases. While, some "in the wild" applications (e.g., fatigue detection [2] and deception detection [3] in unconstrained environments) have not been well concerned. To fill this gap, HUST-LEBW [1] is recently proposed as the first "in the wild" dataset via capturing eyeblink samples from unconstrained movies. It suffers from large variations on human attributes, head pose, illumination, camera view and distance, motion blurring, and sample ambiguity, which is as in Fig. 2. These issues make real-time eyeblink detection on it indeed challenging. Essentially, one key challenge is to well characterize eyelid's intrinsic motion robustly, under unconstrained conditions.

Fig. 3: Failure cases of the state-of-the-art facial landmark detector [16] on HUST-LEBW dataset. The green and red points indicate the successful and failure cases respectively.

Eyeblink detection approach. The existing methods generally fall into landmark-based [5], [6], [10], [20] and region-based groups [1], [7], [8], [11], [12], [21], [22]. Using facial landmarks, the landmark-based methods can extract fine descriptive eyeblink features with promising performance, under the constrained indoor conditions. However, for "in the wild" cases, accurate landmark detection is actually not trivial even using the state-of-the-art manners [15]–[17]. Some failure cases on HUST-LEBW are given in Fig. 3. This leads to high eyeblink detection failure risk as revealed in [1]. To facilitate robustness, the region-based approaches [1], [12] generally choose to alleviate the dependence on facial landmarks via extracting eyeblink feature from the local rigid region around the detected eye center. Although the boosted robustness, region-based eyeblink features tend to sacrifice fine descriptive clues due to the relatively large receptive field in rigid box form. While shrinking the receptive field may increase failure risk since the eye center cannot always be detected accurately.

Our method is region-based. We propose a coarse-to-fine informative eye region discovery method to let the feature extractor focus on the representative eyeblink clues while still maintaining robustness. Meanwhile, the proposed eyelid's intrinsic motion-aware feature learning further leverages discriminative power. The deep insight is that, the feature extractor should be sensitive to eyelid's intrinsic motion and insensitive to non-eyeblink motion.

Attention for visual recognition. The attention mechanism helps the models to focus on key information. It first generates the weights of different features that represent their importance. Based on the generated weights, the model can reweight the features to highlight the most discriminative clues and suppress irrelevant information. In this way, the attention mechanism can facilitate the feature extraction quality and improve the performance and robustness of the model. Such an idea has been widely used in numerous visual recognition research fields such as image classification [23]–[27], object detection [23], [24], [27], semantic segmentation [27], action recognition [28]–[31], etc. In this work, we propose to lead the feature extractor to focus on informative eye region adaptively via introducing visual attention in a coarse-to-fine way. The generated attention adaptively focuses around the center eye region tightly. In this way, the eyelid's intrinsic motion on eye can be captured effectively for eyeblink characterization.

Knowledge distillation. To facilitate the student model's discriminative power, the research idea of knowledge distillation [32], [33] aims to transfer the discriminative information within the teacher model to it. The soft sample labels generated by teacher model are believed of containing privileged information on similarity among different categories [32]. Actually, under some unconstrained conditions, eyeblink and non-eyeblink samples are of high ambiguity that may confuse feature learning for eyeblink characterization. To address this, we use self-knowledge distillation [18], [34]–[36] to generate soft sample labels as the supervision to leverage eyeblink feature learning in the spirit of anti-ambiguity.

III. Eyelid's Intrinsic Motion-aware Feature Learning with Informative Eye Region Attention

The main technical pipeline of the proposed approach is shown in Fig. 4. Specifically, it runs in a coarse to informative way. Towards "in the wild" video clip, human faces and eye centers are first detected via InsightFace [17]. Then a coarse eye region of relatively large size is acquired around the eye center to ensure robustness. To facilitate discriminative power, an informative eye region attention map is generated by a learnable attention generator, which takes the whole face image as input and outputs the attention map that aligns with the resolution of the input face. The portion of the generated attention map that aligns with the previously extracted coarse eye region will be multiplied with this coarse eye region in a pixel-wise manner for informative eye region discovery. The attention-weighted coarse eye region will be inputted into a feature extractor for appearance and motion feature extraction, followed by temporal aggregation and a final classifier for eyeblink classification. Such a coarse-to-fine manner can let the feature extractor focus on the representative clues while still maintaining robustness. During training, the network including the attention generation and the feature extraction are end-to-end optimized by eyeblink classification loss, to seek optimal performance. To facilitate the focusing ability of the attention generator, eye center focus constraints are also designed to regulate its learning procedure. In order to make the feature extractor better aware of the intrinsic eyelid motion and be insensitive to non-eyeblink motion, 2 targeted constraints are imposed at feature level to facilitate feature learning. Besides, we also utilize soft eyeblink label supervision via self-knowledge distillation, which will be introduced in Sec. IV.

A. Coarse-to-fine Informative Eye Region Discovery

For region-based eyeblink detection approaches, how to set local eye region of suitable size around the detected eye center for eyeblink feature extraction is actually an essential issue. Specifically, a larger eye region can enhance robustness but may sacrifice discriminative power. In contrast, a smaller eye region can help extract fine features but may suffer from eye center localization error that leads to potential missing coverage on eye.

Towards eyeblink detection in the wild, we propose a coarse-to-fine informative eye region discovery approach that can capture eye region both robustly and discriminatively, as shown in Fig. 5. First, we argue that the local eye region should be set loosely around the detected eye center first to prioritize robustness, as the eye center cannot always be

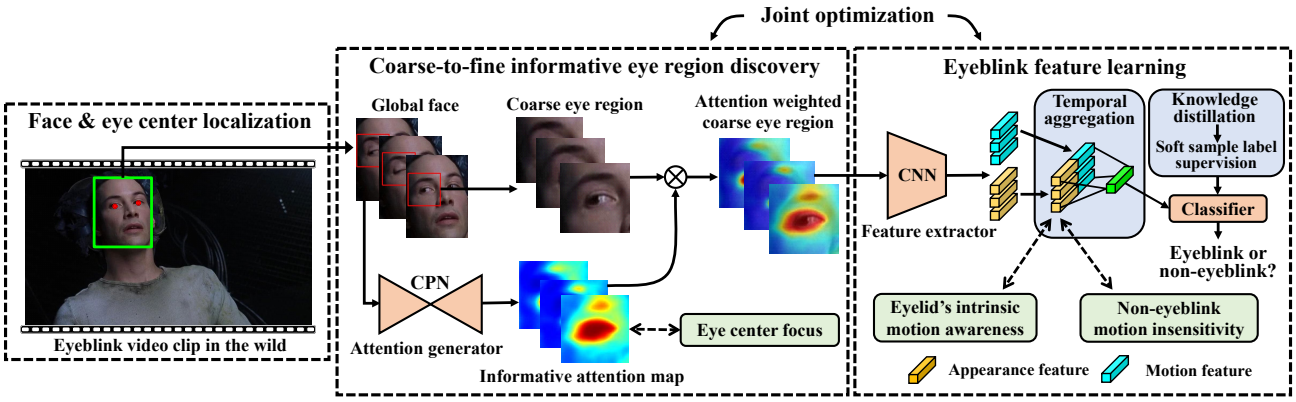


Fig. 4: The main technical pipeline of the proposed eyelid's intrinsic motion-aware feature learning approach for eyeblink characterization in the wild.

localized accurately under in-the-wild conditions, and directly cropping a tight region around the inaccurate eye center may lead to potential missing coverage on eye. Particularly, we first employ face and facial landmark detector (e.g., InsightFace [17] in our implementation) to acquire the face region and eye center position. After resizing the acquired face region to 256×192, the coarse eye region is loosely set around the eye center with a relatively large size of 100×100. To show the superiority of this eye region extraction strategy, some examples are shown in the middle row in Fig. 5 with the comparison to the previous approach [1] on the HUST-LEBW dataset under "in the wild" conditions. Note that we adopt the same face and eye center detector for ours and the compared one [1] for a fair comparison. It can be seen that the tight eye region extraction strategy in [1] (the top row in Fig. 5) cannot always cover eye regions robustly due to the high dependence on the eye center detector that might be unreliable under challenging in-the-wild scenarios. Under the same eye center localization accuracy, our loose eye region extraction strategy (the middle row in Fig. 5) can cover eyes more robustly under the challenging unconstrained conditions that correspond to the large variations in human attributes, face pose, and illumination. Essentially, this leads to a low failure rate towards eyeblink detection in the wild. Moreover, the proposed informative attention generation further captures the fine eye region with rich eyelid movement to boost the discriminative power (the bottom row in Fig. 5). Overall, the proposed coarse-to-fine informative eye region discovery strategy is capable of being both robust and discriminative. The attention generation will be illustrated as follows.

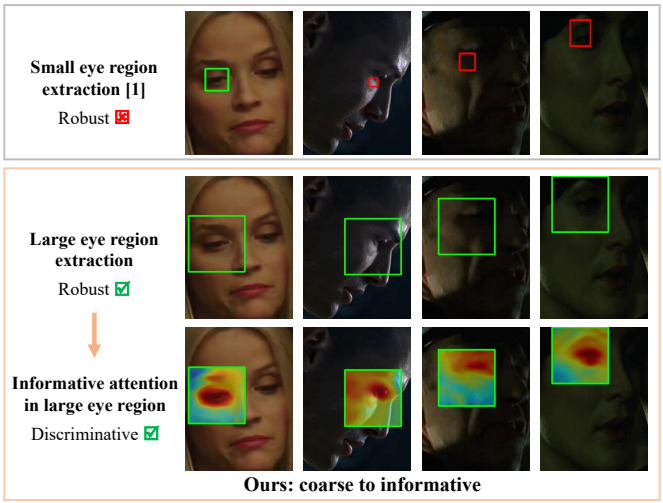


Fig. 5: The comparison between the state-of-the-art method [1] and our approach on eye region discovery on the HUST-LEBW dataset. The left eye in images is taken as an example. The green boxes indicate successful cases, and red ones reveal the failure cases that cannot fully cover the eye. The experiment is conducted under the same face and eye center localization results predicted by InsightFace [17] for a fair comparison.

B. Informative Eye Region Attention Generation with Eye Center Focus Constraints

As shown in the middle row in Fig. 5, although the acquired large coarse eye region helps to ensure robustness, it still involves much background not helpful for fine-grained eyeblink characterization. Thus, it is indeed necessary to further discover more fine and representative eye region for feature learning. We propose to address this by introducing adaptive attention for eye region discovery to better capture the eyelid's intrinsic motion (the bottom row in Fig. 5). Particularly, the adaptive eye region attention map is generated via a learnable attention generator followed by a Sigmoid normalization. We take the whole face image as the input of the attention generator in order to make it better aware of the global face context to facilitate robustness. The resolution of the output attention map is aligned with the input face, and the portion that aligns with the previously extracted coarse eye region will be multiplied with this coarse eye region pixel-wisely to reweight its information. The attention-weighted coarse eye region will be used for the subsequent feature learning. The attention generator will be jointly optimized with eyeblink feature learning to seek optimal performance.

It is worth noting that, imposing informative attention to coarse eye region holds two essential advantages. First, it can refine eyeblink feature learning to facilitate the fine repre-

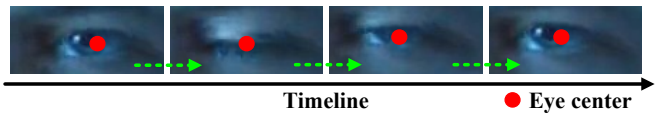


Fig. 6: The intuitive example of eyelid's intrinsic motion across eye center during eyeblink's occurrence.

sentative power via resisting background effect. Meanwhile, the attention map generated with global face context (instead of local coarse eye region) is insensitive to the unreliable eye center localization for stable attention output. This leads to more robust "in the wild" eyeblink feature extraction. Technically, any encoder-decoder liked network architecture for dense prediction can be used as the attention generator as long as the output of the decoder is a heatmap. Here we use cascaded pyramid network (CPN) [37] that is originally used for human pose estimation as our attention generator.

To boost the adaptive ability of the attention generator, we propose to introduce eye center focus constraints (ECFC) on the generated attention map. That is, the eye center is supposed to be of high attention value. This takes 2 main advantages. First, it enables the attention map to also play the role of eye center heatmap under facial landmark estimation framework, to ensure localization robustness. Secondly, it forces strong attention value to generally distribute around central eye region tightly without any shape supervision (e.g., ground-truth heatmap with Gaussian distribution [37]–[39]), which essentially helps to better capture eyelid's intrinsic motion across eye center when eyeblink happens as shown in Fig. 6. We can see that, eye center region indeed involves rich eyelid's motion information for effective eyeblink characterization. Thus, it should be focused on for feature extraction. To this end, the raw attention map $H(p)$ is first normalized via softmax operation as

$$\overrightarrow{H}(p) = \frac{\exp(\beta * H(p))}{\sum_{p' \in \Omega} \exp(\beta * H(p'))}, \tag{1}$$

where p and p' indicate pixel position; Ω is the pixel set within $H(p)$, and β is a scaling hyperparameter. Then, the centroid of the normalized attention map is calculated as

$$p_c = \sum_{p \in \Omega} \overrightarrow{H}(p) * p, \tag{2}$$

with the constraint that forces p_c to be close to the eye center in the spirit of eye center localization via minimizing

$$L_{ec} = \frac{1}{2} \|p_c - p_{ec}^*\|_2^2, \tag{3}$$

where p_{ec}^* denotes eye center's ground-truth position; $\|\cdot\|_2$ indicates $L2$ norm. Such a loss formulation is similar to a kind of centroid supervision by integral regression operation [40], in some heatmap-based facial landmark localization [41] or pose estimation methods [40], [42]. By this, we enable the attention map to also play the role of eye center probability map under landmark localization framework, in the spirit of eye localization. Meanwhile, p_c is also required to be of high attention via minimizing

$$L_{att} = 1 - Sigmoid(H(p_c)), \tag{4}$$

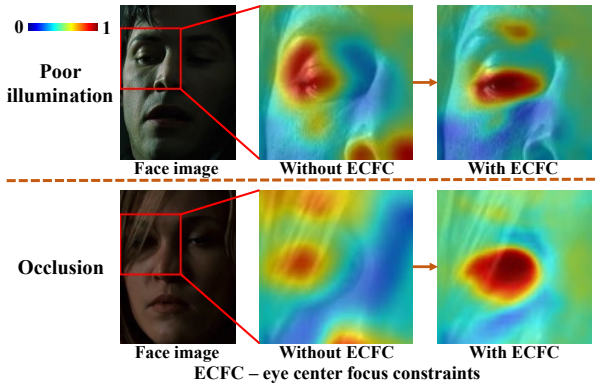


Fig. 7: The effect of eye center focus constraints towards informative eye region attention map generation under poor illumination and occlusion conditions. The red box indicates the pre-acquired local coarse eye region.

where $Sigmoid(\cdot)$ is Sigmoid function. Since Eqn. 2 can be regarded as soft-argmax operation [41]–[43] which can be closely equal to argmax with a large β value (i.e., 100 in our implementation), Eqn. 4 can be approximated as

$$L_{att} = 1 - \max(Sigmoid(H(p))). \tag{5}$$

Eqn. 3 and 5 generally lead the eye center to be of the maximum attention value of high magnitude. This actually reveals eye center focus constraints towards eyeblink characterization.

Some examples for demonstrating the effect of our proposed eye center focus constraints (ECFC) towards informative eye region attention generation are shown in Fig. 7, under the challenging conditions of poor illumination and occlusion. It can be clearly observed that, the introduction of eye center focus constraints indeed helps the attention map to focus more on the informative eye region accurately and stably, with effective background resistance. Accordingly, eye's fine status information can be well captured for eyeblink characterization.

C. Eyelid's Intrinsic Motion-aware Feature Learning

After being multiplied with the Sigmoid normalized attention map (i.e., $Sigmoid(H(p)) \in (0,1)$), the processed coarse eye region is sent into a shallow CNN model to learn the frame-wise appearance and motion features jointly for eyeblink characterization, which is leveraged by CNN's strong visual pattern fitting capacity and high running efficiency due to the shallow network structure. Specifically, the appearance feature A_t of the t-th frame is acquired via conducting global max-pooling on the last convolutional layer. For simplicity, the motion feature M_t is calculated as the appearance feature difference between 2 consecutive frames by

$$M_t = A_t - A_{t-1}, t \in [2, T], \tag{6}$$

where T is the sample frame length. Then, for each frame (except the first one) its 2-stream appearance-motion feature I_t will be the concatenation of A_t and M_t as $I_t = [A_t, M_t]$. The extracted frame-wise feature will be sent to a temproal pooling module (e.g., LSTM [1], [44] in our implementation)

Fig. 8: The examples of non-eyeblink motion under unconstrained in-the-wild conditions.

for temporal aggregation. Finally, a classifier (e.g., a fully-connected layer in our implementation) is used to make final eyeblink verification.

Since eyeblink can be essentially characterized via eyelid's intrinsic motion (i.e., approaching and departure between the upper and lower eyelid) as in Fig. 6, we propose that eyeblink feature learning should be aware of eyelid's intrinsic motion. This helps to better capture eyeblink's representative clue for enhancing discriminative power and generality. Meanwhile concerning non-eyeblink motion due to the unexpected human-camera movement, illumination change and face pose variation under unconstrained "in the wild" conditions as shown in Fig. 8, non-eyeblink motion insensitivity should also be met to make feature extractor focus more on eyelid's intrinsic motion for effective eyeblink characterization.

For eyelid's intrinsic motion awareness, we propose the constraint at feature-level. That is, the inter-frame feature divergence within eyeblink samples should be greater than non-eyeblink ones. Specifically, suppose there are N_1 eyeblink and N_2 non-eyeblink samples in a mini-batch during training, this is acquired by minimizing

$$L_{ma} = \max\left(0, \Delta + \frac{1}{N_2}\sum_{i=1}^{N_2} D_i^O - \frac{1}{N_1}\sum_{i=1}^{N_1} D_i^B\right), \quad (7)$$

where Δ is the expected margin between D_i^B and D_i^O. D_i^B and D_i^O indicate the inter-frame feature divergence of certain eyeblink and non-eyeblink sample given by

$$D_i^{idx} = \frac{1}{T-2}\sum_{t=3}^{T} \|I_t - I_{t-1}\|_2^2, idx \in \{B, O\}. \quad (8)$$

Then towards non-eyeblink motion insensitivity, it is acquired by imposing the constraint that forces the inter-frame feature divergence within non-eyeblink samples to be as small as possible, which is achieved by minimizing

$$L_{ni} = \frac{1}{N_2}\sum_{i=1}^{N_2} D_i^O. \quad (9)$$

Accordingly, the frame-wise eyeblink feature distribution within the non-eyeblink samples is asked to be uniform to resist the non-eyeblink motion information.

IV. Supervise Eyeblink Feature Learning with Soft Sample Label for Anti-Ambiguity

Towards effective eyeblink feature learning, we find that one essential challenge is the potential high ambiguity between the eyeblink and non-eyeblink samples. Specifically, the eyeblink and non-eyeblink procedure may share a very

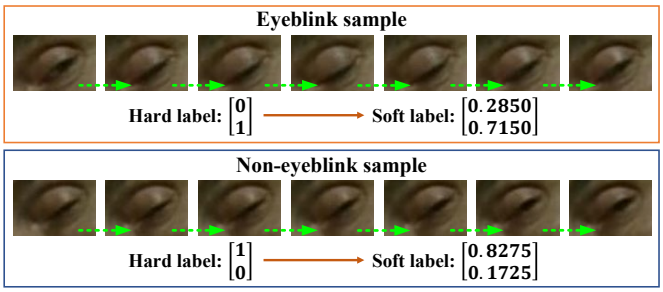


Fig. 9: The intuitive examples of ambiguous eyeblink and non-eyeblink samples within HUST-LEBW. Particularly, the positive eyeblink sample is of hard label $[0, 1]^T$ and soft label $[0.2850, 0.7150]^T$ generated via self-knowledge distillation. And the hard and soft sample label of the non-eyeblink sample is $[1, 0]^T$ and $[0.8275, 0.1725]^T$.

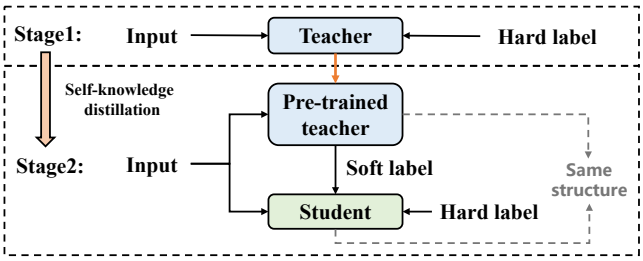


Fig. 10: The main idea of self-knowledge distillation for anti-ambiguity.

similar appearance (as shown in Fig. 9), even humans can not distinguish them confidently. Such a phenomenon may be due to the subtle eye status variation during eyeblink caused by personalized factors (e.g., small eyes or no complete closure of the eyelids during eyeblink). In those cases, it would be better to say that to what extent (probability) a sample is an eyeblink/non-eyeblink, rather than judge it by a "hard classification label" (i.e., 0 or 1). Thus, using the original hard sample labels (i.e., "1" for eyeblink, and "0" for non-eyeblink) as the supervision to conduct feature learning rigidly may confuse the classifier. As a consequence, eyeblink feature's fine-grained representative ability and generality cannot be well ensured. To address this, we propose to use the prior soft sample labels valued in $[0, 1]$ that indicates the probability of a sample being an eyeblink or a non-eyeblink to supervise feature learning, instead of using the hard ones (i.e., 0 or 1). In our opinion, this can facilitate the feature extractor to better capture eyelid's fine status variation during eyeblink with more reasonable supervision. We formulate this under a self-knowledge distillation framework [18]. The main idea is to transfer the prior knowledge about the similarity between classes within a pre-trained teacher network to a student network, to enable the student network to better handle sample ambiguity from the more reasonable supervision generated from the teacher as shown in Fig. 10.

Particularly, our whole network is trained with 2 stages. First, it is pre-trained as the teacher network via minimizing

$$L_T = L_{ce} + \lambda_1 L_e, \quad (10)$$

where L_{ce} is the cross-entropy loss and

$$L_e = L_{att} + \lambda_2 L_{ec} + \lambda_3 L_{ma} + \lambda_4 L_{ni}, \tag{11}$$

where L_{att} and L_{ec} are eye center focus constraints (ECFC) illustrated in Sec. III-B. L_{ma} and L_{ni} are the constraints of eyelid's intrinsic motion awareness and non-eyeblink motion insensitivity that are introduced in Sec. III-C. $\lambda_i, i \in [1, 4]$ are the hyperparameters to balance the effect of different losses.

With the soft sample labels generated via pre-trained teacher network as in Fig. 9, in the second phase a student network of the same structure will be then trained under the guidance of the teacher network using the loss function given by

$$L_S = L_{kd} + \lambda_1 L_e, \tag{12}$$

where L_{kd} is the self-knowledge distillation loss that leads the student to mimic its teacher's predicted soft labels as well as the original ground-truth hard labels as

$$L_{kd} = (1-\alpha)\, L_{ce}\left(h_{gt}, p_s\right) + \alpha D_{kl}\left(\widetilde{p}_t, \widetilde{p}_s\right), \tag{13}$$

where h_{gt} is sample's ground-truth hard label; p_s indicates the student's prediction; $\widetilde{p}_t$ and $\widetilde{p}_s$ denote the softened probability distribution [18] of the prediction yielded by the teacher and student network respectively; $D_{kl}\left(\widetilde{p}_t, \widetilde{p}_s\right)$ represents KL-divergence between $\widetilde{p}_t$ and $\widetilde{p}_s$; and α is the hyperparameter to balance the effect of $D_{kl}\left(\widetilde{p}_t, \widetilde{p}_s\right)$ and $L_{ce}\left(h_{gt}, p_s\right)$. After the second phase of training, the acquired student network will be finally used for eyeblink detection.

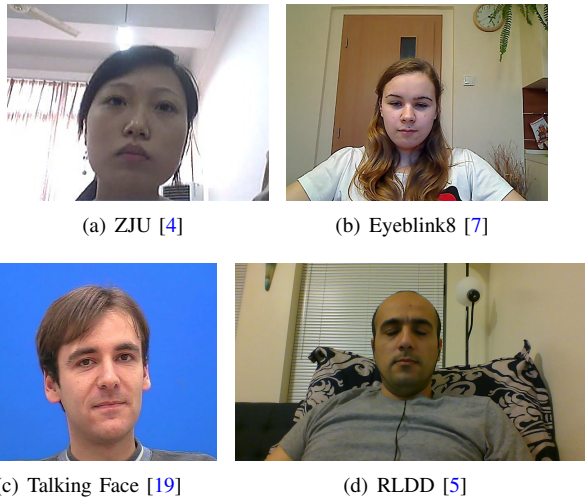
V. Implementation Details

The sample frames are resized to 256×256 for InsightFace. The architecture of the shallow CNN is: "Conv2D(5,3,24) - Conv2D(3,2,48) - Conv2D(3,2,80)", where Conv2D(k,s,c) denotes a 2D convolution layer with kernel size k, stride size s and c output channels. Each layer is followed by a batch normalization [45], a ReLU activation [46] and a maxpooling layer with a down-sampling ratio of 0.5. Δ in Eqn. 7 is set to 0.1. λ_1, λ_2, λ_3, and λ_4 in Eqn. 10 and 11 are set to 0.2, 0.0005, 0.5 and 0.25. α in Eqn. 13 is set to 0.1. Adam [47] is used as optimizer with a batch size of 8. The learning rate of the attention generation network begins with $5e^{-5}$, rises to $3e^{-4}$ after 6 epochs, then decays with a rate of 0.8 every 8 epochs. The learning rate for the other parts of the network is set as 4 times as that for the attention generation network. The whole training process terminates at 100 epochs.

VI. Experiments

A. Experimental Setup

The experiments are conducted on one "in the wild" dataset and one combined constrained dataset, to verify the effectiveness and generalization ability of our proposition. First, we conduct experiments on the HUST-LEBW dataset [1] to analyze the eyeblink detection capacity of the proposed method towards unconstrained in-the-wild cases (Sec. VI-B). As different methods adopt different face and landmark detectors, to eliminate such differences for a fairer comparison, we further make a comparison with the representative methods under the same face and landmark detection results (a unified face analysis toolbox InsightFace [17] is used for face and facial landmark detection). To further study the eyeblink detection capacity towards constrained cases, experiments are also conducted on a combined constrained dataset Constrained+ (Sec. VI-C). Moreover, we also evaluate the challenging cross-dataset setting (model is trained on Constrained+ and directly tested on HUST-LEBW) in Sec. VI-D. To explore the ability for eyeblink detection in untrimmed videos, we also test the proposed eyeblink detection method in untrimmed videos (Sec. VI-E). The real-time inference capacity of our eyeblink detection approach is also illustrated in Sec. VI-F. Then, the ablation studies of the proposed component are conducted in Sec. VI-G, followed by a parameter setting analysis in Sec. VI-H. Finally, a further discussion including attention visualization (Sec. VI-I) and case studies of success and failure (Sec. VI-J) is illustrated to reveal a deeper insight of the proposed method. We will briefly introduce the datasets and evaluation metrics used as follows.



(a) ZJU [4] (b) Eyeblink8 [7]

(c) Talking Face [19] (d) RLDD [5]

Fig. 11: The live eyeblink snapshots from the individual datasets within the Constrained+ dataset.

HUST-LEBW [1]. It is the first eyeblink in the wild dataset. Being different from the other datasets [4], [5], [7], [48], samples within HUST-LEBW are captured from the unconstrained movies instead of from the volunteers under the constrained indoor conditions. It involves 673 trimmed eyeblink samples (i.e., 381 positives, and 292 negatives) with 448 samples in the training set and 225 samples in the test set. It also provides a sub-set of 90 untrimmed videos for testing purposes. Some live eyeblink snapshots from HUST-LEBW are shown in Fig. 2, which reveals the critical challenges. More details respecting HUST-LEBW can be found in [1].

Constrained+. To verify the generalization ability of our proposition towards the constrained cases. We combine 4 existing constrained eyeblink datasets (i.e., ZJU [4], Eyeblink8 [7], Talking Face [19] and RLDD [5]) to form the Constrained+ and evaluate our method on it. Some live eyeblink snapshots from the individual datasets within Constrained+ are shown in Fig. 11. Specifically, we adopt an off-the-shelf face parsing

TABLE I: Performance comparison among the different methods on the HUST-LEBW dataset. * indicates that the method is trained on a larger dataset mEBAL [8].

Method	Eye	Recall	Precision	F1-score
Chau and Betke [49]	Left	1.64	**100.00**	3.23
	Right	0.00	0.00	0.00
Morris et al. [21]	Left	4.10	71.43	7.75
	Right	2.38	75.00	4.62
Drutarovsky and Fogelton [7]	Left	5.74	41.18	10.07
	Right	3.17	30.77	5.76
Tabrizi and Zoroofi [50]	Both	7.14	45.00	12.33
Soukupová and Cech [6]	Left	36.07	64.71	46.32
	Right	30.16	57.58	39.58
Hu et al. [1]	Left	54.10	89.19	67.35
	Right	44.44	76.17	56.28
Daza et al.* [8]	Left	**96.03**	60.80	74.46
	Right	79.50	73.48	76.37
Daza et al.* [9]	Both	93.39	75.33	83.39
Ours	Left	91.80	89.60	**90.69**
	Right	**91.27**	**92.74**	**92.00**

TABLE II: Performance comparison among the proposed method and the other methods using InsightFace [17] for face parsing as ours on the HUST-LEBW dataset.

Method	Eye	Recall	Precision	F1-score
Soukupová and Cech [6]	Left	36.07	64.71	46.32
	Right	30.16	57.58	39.58
Soukupová and Cech [6] + InsightFace [17]	Left	40.98	81.97	54.64
	Right	40.48	87.93	55.43
Hu et al. [1]	Left	54.10	89.19	67.35
	Right	44.44	76.17	56.28
Hu et al. [1] + InsightFace [17]	Left	68.85	73.04	70.89
	Right	76.98	79.51	78.23
Ours	Left	**91.80**	**89.60**	**90.69**
	Right	**91.27**	**92.74**	**92.00**

engine to automatically label human faces as well as facial landmarks. The frames with wrong prediction results will be ignored. The videos will be split into trimmed clips as in [1]. As the number of non-eyeblink clips is much more than the number of eyeblinks, we apply random sampling to balance the amount of the eyeblink and non-blink samples. Overall, the dataset contains 4,935 samples (i.e., 2,435 positives and 2,500 negatives). Particularly the training set involves 2,235 positives and 2,300 negatives, and the test set contains 200 positives and 200 negatives.

Evaluation metric. Following [1], Recall, Precision and F1-score are used to evaluate eyeblink detection on the trimmed samples. Meanwhile, average precision (AP) is applied to the untrimmed samples.

B. Comparison with State-of-the-art Methods on Eyeblink Detection in the Wild

HUST-LEBW: The proposed eyeblink detection in the wild approach is first compared with the state-of-the-art methods [1], [6]–[9], [21], [49], [50] towards the trimmed samples. The results are listed in Table I. It can be observed that:

TABLE III: Performance comparison among the different methods on the Constrained+ dataset.

Method	Eye	Recall	Precision	F1-score
Soukupová and Cech [6] + InsightFace [17]	Left	97.00	99.49	98.23
	Right	97.50	99.49	98.48
Hu et al. [1] + InsightFace [17]	Left	92.00	97.35	94.60
	Right	98.00	99.49	98.74
Ours	Left	**99.50**	**100.00**	**99.75**
	Right	**99.50**	**100.00**	**99.75**

• For both left and right eyes, our method significantly outperforms the others by a large margin with 7% at least on F1-score. Note that the most competitive method [9] is trained on a larger dataset mEBAL [8]. This essentially verifies the superiority of our proposition towards eyeblink detection in the wild;

• The performance of the other methods is actually not satisfactory enough (i.e., F1-score is less than 90%). This indeed reveals the challenges of eyeblink detection in the wild.

Comparison using the same state-of-the-art face parsing engine: For a more fair comparison, the proposed approach is also compared with others equipped with the state-of-the-art face engine (i.e., InsightFace [17]) as ours on HUST-LEBW. As the methods [8], [9] require frame-level annotations to be trained, they can not be trained on the HUST-LEBW dataset for a fair comparison. Thus, we choose to compare with the other 2 representative methods (i.e., one region-based method [1] and one landmark-based method [6]) in this experiment. The results are given in Table II. It can be summarized that:

• When using the same face parsing engine (i.e., InsightFace), our proposition still significantly outperforms the others. This indeed verifies the superiority of the proposed eyeblink feature extraction method;

• InsightFace facilitates the performance of previous methods [1], [6]. This verifies that, effective face parsing is actually an essential issue towards "in the wild" scenarios.

C. Comparison with State-of-the-art Methods on Eyeblink Detection under Constrained Conditions

Constrained+: Our method is also compared with [1] and [6]. We also equip them with InsightFace for face parsing to conduct a fair comparison. The results are given in Table III. We can see that:

• The proposed approach still outperforms others under constrained conditions, with an indeed high F1-score of 99.75%. This demonstrates the strong generalization capacity of our proposition for the different application scenarios;

• Both 2 methods for comparison can achieve promising results here. While compared with the results in Table II, their performance drop from Constrained+ to HUST-LEBW is huge (i.e., with over 20% F1-score drop). In contrast, this is not serious when using our method (i.e., about 9% F1-score drop). This first reveals the greater challenges of eyeblink detection in the wild over the constrained scenarios. And, this somewhat verifies that the proposed method has captured eyeblink's intrinsic features.

TABLE IV: Performance comparison among different methods under cross-dataset test setting from Constrained+ to HUST-LEBW.

Method	Eye	Recall	Precision	F1-score
Soukupová and Cech [6] + InsightFace [17]	Left	40.16	80.33	53.55
	Right	44.44	**77.77**	56.57
Hu et al. [1] + InsightFace [17]	Left	70.49	70.49	70.49
	Right	70.63	70.08	70.36
Ours	Left	**91.80**	**80.58**	**85.82**
	Right	**94.44**	76.77	**84.69**

TABLE V: Performance comparison among the different methods on untrimmed eyeblink samples within HUST-LEBW dataset.

Method	Eye	AP
Hu et al. [1]	Left	29.42
	Right	31.85
Ours	Left	**76.52**
	Right	**80.47**

D. Comparison with Cross-dataset Test Setting

To verify the generalization capacity of our proposition, it is also compared with [1] and [6] with a cross-dataset test. That is, they are trained on Constrained+ but test on HUST-LEBW. For a fair comparison, [1] and [6] are also equipped with InsightFace as ours for face parsing. The results are given in Table IV. It can be summarized that:

• For this challenging test setting, our approach still significantly outperforms the others with a high F1-score of about 85%. This indeed demonstrates the strong generality of the proposed region-based eyeblink feature extraction approach. That is, our proposition has captured eyeblink's intrinsic feature disregarding the application conditions;

• Landmark-based method [6] is inferior to region-based [1] and ours. This somewhat reveals the fact that the region-based paradigm is of stronger generalization capacity than the landmark-based counterpart.

E. Evaluation on Untrimmed Videos

To explore the eyeblink detection in the wild ability in untrimmed videos, our approach is also evaluated on untrimmed scenarios following the main technical paradigm and evaluation metrics in [1]. Specifically, the model is trained on the trimmed samples in HUST-LEBW, and directly inference on a sub-set of HUST-LEBW that consists of 90 untrimmed videos in a sliding window manner, for test only. The results are listed in Table V. It can be seen that:

• Our method still outperforms the existing method [1] by a large margin (i.e., at least over 47% on AP). This demonstrates the superiority of our method towards untrimmed cases that is closer to the practical applications;

• Towards the more challenging untrimmed case, the performance of our method is still not satisfactory enough (i.e., AP is only around 80%). One potential reason is the different characteristics between trimmed and untrimmed videos. Existing efforts including ours mainly focus on trimmed cases, so it is hard to get a precise eyeblink boundary by naively employing a rigid temporal sliding window. How to well address untrimmed cases is what we concern in future works.

TABLE VI: The average running time per frame of the proposed eyeblink detection method.

Procedure	Time (ms)
Face detection	10.33
Facial landmark detection	4.67
Attention generation	6.68
Feature extraction & eyeblink verification	2.20
Total	23.88

TABLE VII: Ablation studies on the proposed informative eye region attention and the eye center focus constraints (ECFC) on the HUST-LEBW dataset.

Eye	Attention	ECFC	Trimmed Recall	Trimmed Precision	Trimmed F1-score	Untrimmed AP
Left			75.41	79.31	77.31	45.03
	✓		85.25	85.95	85.60	70.88
	✓	✓	**89.34**	**86.51**	**87.90**	**75.73**
Right			73.02	85.98	78.97	39.69
	✓		83.33	85.37	84.34	76.38
	✓	✓	**88.10**	**87.40**	**87.75**	**77.64**
Average			74.22	82.65	78.14	42.36
	✓		84.29	85.66	84.97	73.63
	✓	✓	**88.72**	**86.96**	**87.83**	**76.69**

F. Real-time Running Capacity

The average running time per frame of our method is listed in Table VI, using a single NVIDIA 2080Ti GPU. It can be seen that the running speed of our approach is about 23.88ms per frame (i.e., around 41 FPS). Essentially, it meets the real-time running requirement. It is worth noting that, face detection is of the most time consumption. If it is replaced with the more efficient ones, the running efficiency of our method can be further enhanced.

G. Ablation Studies

In this session, we analyze the roles played by the proposed components by gradually adding them and evaluating their performance on the HUST-LEBW dataset in both trimmed and untrimmed settings.

Effectiveness of informative eye region attention: The results are listed in Table VII. To resist the effect of the other propositions, the baseline network here will be trained only using cross-entropy loss without any other proposed components. We gradually add the attention generator and the eye center focus constraints (ECFC) to the baseline to analyze their effects. From Table VII we can summarize that:

• Informative eye region attention plays an important role in our coarse-to-fine eye region discovery procedure. It can facilitate performance remarkably (i.e., 6.83% F1-score on average) towards eyeblink detection in the wild in trimmed setting, as the deformable attention can adaptively capture the fine region with rich eyelid movement. Moreover, the performance gain in untrimmed setting is more notable (i.e.,

TABLE VIII: Ablation studies on the constraints for eyelid motion-aware feature learning on the HUST-LEBW dataset.

Eye	L_{ni}	L_{ma}	Trimmed Recall	Trimmed Precision	Trimmed F1-score	Untrimmed AP
Left			89.34	86.51	87.90	75.73
	✓		88.52	88.52	88.52	71.83
		✓	90.16	88.00	89.07	64.06
	✓	✓	**90.98**	**89.51**	**90.24**	**76.15**
Right			88.10	87.40	87.75	77.64
	✓		89.68	88.28	88.98	80.27
		✓	89.68	90.40	90.04	**80.55**
	✓	✓	**91.27**	**90.55**	**90.90**	**80.55**
Average			88.72	86.96	87.83	76.69
	✓		89.10	88.40	88.75	76.05
		✓	89.92	89.20	89.56	72.31
	✓	✓	**91.13**	**90.03**	**90.57**	**78.35**

TABLE IX: Performance comparison of our method with and without prior soft supervision on HUST-LEBW dataset.

Eye	Soft supervision	Trimmed Recall	Trimmed Precision	Trimmed F1-score	Untrimmed AP
Left		90.98	89.51	90.24	76.15
	✓	**91.80**	**89.60**	**90.69**	**76.52**
Right		**91.27**	90.55	90.90	**80.55**
	✓	**91.27**	**92.74**	**92.00**	80.47
Average		91.13	90.03	90.57	78.35
	✓	**91.54**	**91.17**	**91.35**	**78.50**

TABLE X: Performance comparison with different λ values on the HUST-LEBW dataset.

λ_i	value	Left eye Recall	Left eye Precision	Left eye F1-score	Right eye Recall	Right eye Precision	Right eye F1-score
λ_1	0	85.25	85.95	85.60	83.33	85.37	84.34
	0.1	90.16	86.61	88.35	90.48	89.76	90.12
	0.2	90.98	**89.51**	**90.24**	**91.27**	90.55	**90.90**
	0.3	**93.44**	87.02	90.12	88.89	**91.06**	89.96
λ_2	0	90.98	86.72	88.80	87.30	90.16	88.71
	0.0001	**91.80**	86.82	89.24	**91.27**	89.15	90.20
	0.0005	90.98	**89.51**	**90.24**	**91.27**	**90.55**	**90.90**
	0.001	90.98	88.80	89.88	90.48	88.37	89.41
λ_3	0	88.52	88.52	88.52	89.68	88.28	88.98
	0.1	**90.98**	88.10	89.52	**92.06**	88.55	90.27
	0.5	**90.98**	**89.51**	**90.24**	91.27	**90.55**	**90.90**
	1	89.34	88.62	88.98	91.27	89.84	90.55
λ_4	0	90.16	88.00	89.07	89.68	90.40	90.04
	0.1	**91.80**	88.19	89.96	**91.27**	89.84	90.55
	0.25	90.98	**89.51**	**90.24**	**91.27**	**90.55**	**90.90**
	0.5	90.98	88.10	89.52	90.48	89.76	90.12

TABLE XI: Performance comparison with different Δ values on the HUST-LEBW dataset.

L_{ma}	Δ	Left eye Recall	Left eye Precision	Left eye F1-score	Right eye Recall	Right eye Precision	Right eye F1-score
×	-	88.52	88.52	88.52	89.68	88.28	88.98
✓	0	89.34	88.62	88.98	88.89	90.32	89.60
	0.1	90.98	**89.51**	**90.24**	**91.27**	**90.55**	**90.90**
	0.2	**92.62**	87.60	90.04	88.89	89.60	89.24

31.27% AP on average), which demonstrates that robust and accurate fine eye region capture is more crucial for untrimmed scenarios;

• The eye center focus constraints (ECFC) illustrated in Sec. III-B further boost the performance (i.e., 2.86% F1 score in trimmed scenarios and 3.06% F1 score in untrimmed scenarios on average), as such explicit constraints can enhance the focusing ability of the generated attention. The qualitative visualization analysis can be found in Fig. 7.

Effectiveness of constraints for eyelid motion-aware feature learning: We gradually add new components on the previously obtained model in Table VII to study the effect of the proposed constraints for eyelid's intrinsic motion awareness (L_{ma}) and non-eyeblink motion insensitivity (L_{ni}) during feature learning. The results are listed in Table VIII. It can be observed that:

• Applying either of them can boost the performance by a noticeable margin (i.e., at least 0.5% on F1-score) in trimmed scenarios, which verifies the effectiveness of the proposed constraints towards eyeblink detection in the wild;

• Actually, using either constraint alone cannot achieve the optimal performance, as it may lead to potential feature degradation issues. Specifically, solely employing the non-eyeblink motion insensitivity constraint (i.e., L_{ni}) may also reduce the extracted inter-frame feature differences for eyeblink samples, which is unfavorable for the characterization of eyelid movements. However, by adding the eyelid's intrinsic motion aware constraint (i.e., L_{ma}) at the same time, the feature extractor can remain sensitive to eyelid movements while being insensitive to non-eyeblink motions. On the other hand, if only eyelid's intrinsic motion aware constraint (i.e., L_{ma}) is applied, it might cause the feature extractor also be more sensitive on the non-eyeblink motions such as illumination and head pose change that are common in unconstrained in-the-wild scenarios. Further adding non-eyeblink motion insensitivity constraint can let the feature extractor be insensitive to those non-eyeblink motions here. Such a feature degradation issue is more notable for untrimmed setting (i.e., only using either constraint alone here will even decrease the performance). The experimental results show that when applying these two constraints together, the performance can be boosted by a large margin (i.e., 2.74% F1-score in trimmed setting and 1.66% AP in untrimmed setting on average). This essentially demonstrates the effectiveness of these constraints towards discriminative and robust eyeblink feature learning. The results also reveal the fact that eyelid's intrinsic motion is indeed one essential factor for eyeblink characterization in the wild.

Effectiveness of prior soft sample label supervision: The performance comparison of our method with and without it on the HUST-LEBW dataset is listed in Table IX. Actually, the self-generated prior soft label can give a more reasonable supervision and thus boost the performance consistently.

H. Parameter Setting Analysis

Analysis on λ: Here we investigate the settings of the parameters λ_i in Eqn. 10 and 11. The performance comparison among different λ_i values on the HUST-LEBW dataset is given in Tab X. Note that for the experiment within each λ_i, the other $\lambda_{j,j\neq i}$ are set to its optimal value and self-knowledge distillation is not included. It can be observed that:

• When λ_i>0, the performance improves consistently, especially for λ_1 that controls the overall effect of the proposed

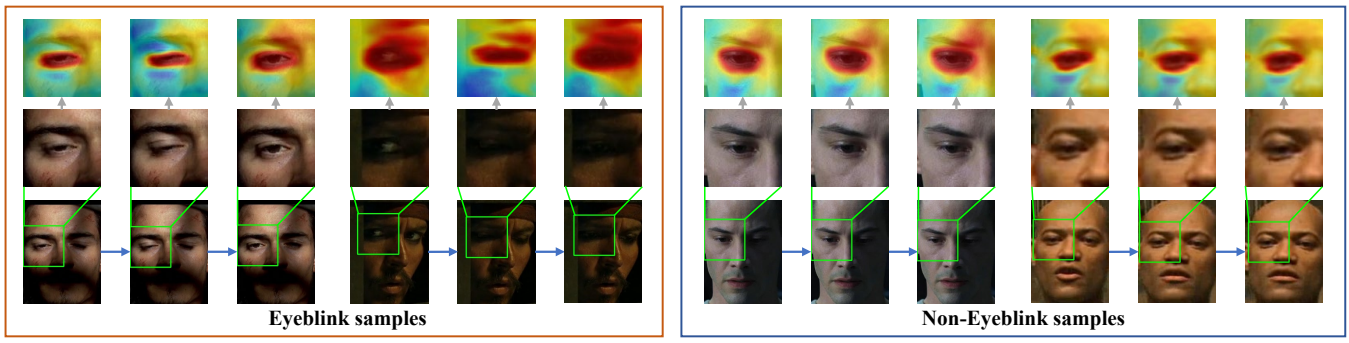
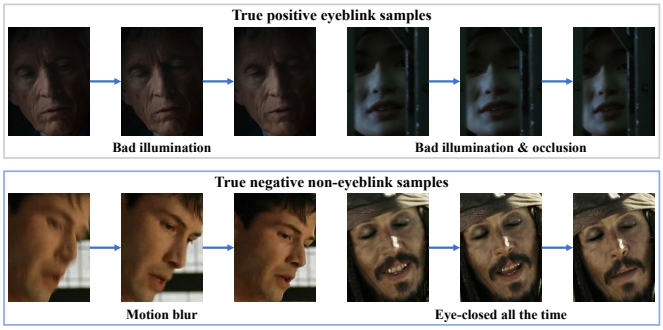
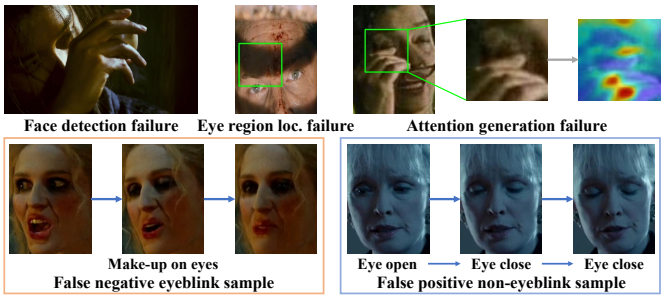


Fig. 12: Visual analysis on informative eye region attention.



(a) Successful cases



(b) Failure cases

Fig. 13: Case studies of success and failure.

constraints, which essentially demonstrates the effectiveness of the proposed constraints.

• The performance improvements are stable with different λ_i values as long as the model can be trained evenly. Generally, the optimal values are 0.2, 0.0005, 0.5, and 0.25 for λ_1, λ_2, λ_3, and λ_4 respectively.

Analysis on Δ: As in Eqn. 7, Δ is the expected margin between D_i^B and D_i^O. The performance comparison among different Δ values on the HUST-LEBW dataset is given in Tab XI. It can be observed that when $\Delta \geq 0$, the performance can be enhanced, which fits with our motivation that letting the inter-frame feature divergence within eyeblink samples be greater than non-eyeblink ones is beneficial for extracting eyelid's intrinsic motion and thus boosts the performance. Generally, 0.1 is optimal for Δ.

I. Visual Analysis on Informative Attention

We visualize the generated attention within our network in Fig. 12 to analyze the effect of informative eye region attention towards eyeblink characterization. It can be seen that within the eyeblink procedure, the attention can well capture eyelid's fine intrinsic motion for subsequent feature learning. For the non-eyeblink samples, the attention focuses on eye's open state consistently. Overall, the informative eye regions with rich eyelid movement have been captured and the effect of the background has been resisted remarkably. This intuitively verifies that the proposed approach helps to capture informative eye regions to better characterize eyelid motion.

J. Case Studies of Success and Failure

Here we further study the success and failure cases of the proposed method. From Fig. 13, we can see that the proposed method can work robustly under some challenging scenarios with bad illumination, severe occlusion, motion blur, etc. Notably, it can differentiate between eyeblink and eye-close, which is hard to meet by some frame-based methods [8], [9] that regard the eye-closed as eyeblink.

Nevertheless, it can also be observed that our method will fail to work in some cases. Firstly, it can not work if the face can not be detected by the face detector. Secondly, although the proposed coarse-to-fine informative eye region discovery approach alleviates the dependence of eyeblink detection on eye localization by landmark detector, it still can not handle the cases where the detected eye is far from its actual position, as the extracted coarse eye region can not cover the eye at this circumstance. Thirdly, the attention may fail to work due to serious occlusion on eyes. Besides, our method may not perform well when the eyes are heavily made up. Although it can distinguish between eyeblink and eye-closed, it may mistakenly recognize an "eye open, eye close, eye close" procedure as an eyeblink. We speculate that this is because there is a lack of samples of such rare non-eyeblink processes in the training set.

VII. Conclusions

Towards the challenging research task of real-time eyeblink detection in the wild, a novel eyelid's intrinsic motion-aware feature learning approach is proposed. Within it, the combination of large coarse eye region and informative eye region attention facilitates robust extraction of discriminative eyeblink features. The proposed constraints on attention generation and feature learning help to better capture eyeblink's

intrinsic feature. We also observe the ambiguity issue between eyeblink and non-eyeblink samples, and use self-generated soft supervision to address it. Experiments verify the superiority of our propositions. In the future, we will pay more attention to addressing the untrimmed cases and facilitating robustness.

ACKNOWLEDGMENT

This work is jointly supported by the National Natural Science Foundation of China (Grant No. 62271221, U1913602, 61702182, and 61876211), National Key R&D Program of China (No. 2018YFB1004600), Guangdong Basic and Applied Basic Research Foundation under Grant No. 2023A1515011260, and Science and Technology Program of Guangzhou under Grant No. 202201011672.

Joey Tianyi Zhou is supported by SERC Central Research Fund (Use-inspired Basic Research), Programmatic Grant No. A18A1b0045 from the Singapore government's Research, and Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain).

REFERENCES

[1] G. Hu, Y. Xiao, Z. Cao, L. Meng, Z. Fang, J. T. Zhou, and J. Yuan, "Towards real-time eyeblink detection in the wild: Dataset, theory and practices," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2194–2208, 2020. 1, 2, 3, 4, 5, 7, 8, 9

[2] L. M. Bergasa, J. Nuevo, M. A. Sotelo, R. Barea, and M. E. Lopez, "Real-time system for monitoring driver vigilance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 1, pp. 63–77, 2006. 1, 2

[3] K. Fukuda, "Eye blinks: new indices for the detection of deception," *International Journal of Psychophysiology*, vol. 40, no. 3, pp. 239–245, 2001. 1, 2

[4] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8. 1, 2, 7

[5] R. Ghoddoosian, M. Galib, and V. Athitsos, "A realistic dataset and baseline temporal model for early drowsiness detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 0–0. 1, 2, 3, 7

[6] T. Soukupová and J. Cech, "Real-time eye blink detection using facial landmarks," in *Proc. Computer Vision Winter Workshop (CVWW)*, 2016, pp. 1–8. 1, 3, 8, 9

[7] T. Drutarovsky and A. Fogelton, "Eye blink detection using variance of motion vectors," in *Proc. European Conference on Computer Vision (ECCV)*, 2014, pp. 436–448. 1, 2, 3, 7, 8

[8] R. Daza, A. Morales, J. Fierrez, and R. Tolosana, "Mebal: A multimodal database for eye blink detection and attention level estimation," in *Proc. International Conference on Multimodal Interaction (ICMI)*, 2020, pp. 32–36. 1, 2, 3, 8, 11

[9] R. Daza, D. DeAlcala, A. Morales, R. Tolosana, R. Cobos, and J. Fierrez, "Alebk: Feasibility study of attention level estimation via blink detection applied to e-learning," *arXiv preprint arXiv:2112.09165*, 2021. 1, 8, 11

[10] S. Al-gawwam and M. Benaissa, "Robust eye blink detection based on eye landmarks and savitzky–golay filtering," *Information*, vol. 9, no. 4, p. 93, 2018. 1, 3

[11] K. Cortacero, T. Fischer, and Y. Demiris, "Rt-bene: A dataset and baselines for real-time blink estimation in natural environments," in *Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019, pp. 0–0. 1, 2, 3

[12] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai created fake videos by detecting eye blinking," in *Proc. International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7. 1, 2, 3

[13] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 532–539. 1

[14] S. Wu, M. Kan, Z. He, S. Shan, and X. Chen, "Funnel-structured cascade for multi-view face detection with alignment-awareness," *Neurocomputing*, vol. 221, p. 138–145, Jan 2017. 1

[15] X. Wang, L. Bo, and L. Fuxin, "Adaptive wing loss for robust face alignment via heatmap regression," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6970–6980. 1, 3

[16] C. Lin, B. Zhu, Q. Wang, R. Liao, C. Qian, J. Lu, and J. Zhou, "Structure-coherent deep feature learning for robust face alignment," *IEEE Transactions on Image Processing*, 2021. 1, 3

[17] J. D. Jia Guo, "Insightface: 2d and 3d face analysis project," https://github.com/deepinsight/insightface, 2020. 1, 3, 4, 7, 8, 9

[18] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3903–3911. 2, 3, 6, 7

[19] "Talking face video," Face&Gesture Recognition Working Group, IST-2000-26434. [Online]. Available: http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html 2, 7

[20] S. Dari, N. Epple, and V. Protschky, "Unsupervised blink detection and driver drowsiness metrics on naturalistic driving data," in *Proc. IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020, pp. 1–6. 3

[21] T. Morris, P. Blenkhorn, and F. Zaidi, "Blink detection for real-time eye tracking," *Journal of Network and Computer Applications*, vol. 25, no. 2, pp. 129–143, 2002. 3, 8

[22] A. Królak and P. Strumiłło, "Eye-blink detection system for human–computer interaction," *Universal Access in the Information Society*, vol. 11, no. 4, pp. 409–419, 2012. 3

[23] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19. 3

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141. 3

[25] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4438–4446. 3

[26] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, and J. Jiao, "Selective sparse sampling for fine-grained image recognition," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6599–6608. 3

[27] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 713–13 722. 3

[28] Z. Tu, H. Li, D. Zhang, J. Dauwels, B. Li, and J. Yuan, "Action-stage emphasized spatiotemporal vlad for video action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2799–2812, 2019. 3

[29] X. Shu, J. Yang, R. Yan, and Y. Song, "Expansion-squeeze-excitation fusion network for elderly activity recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5281–5292, 2022. 3

[30] X. Shu, B. Xu, L. Zhang, and J. Tang, "Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[31] B. Xu, X. Shu, and Y. Song, "X-invariant contrastive augmentation and representation learning for semi-supervised skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 31, pp. 3852–3867, 2022. 3

[32] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015. 3

[33] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

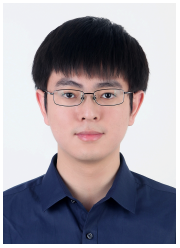
[34] K. Kim, B. Ji, D. Yoon, and S. Hwang, "Self-knowledge distillation with progressive refinement of targets," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6567–6576. 3

[35] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *Proc. International Conference on Machine Learning (ICML)*, 2018, pp. 1607–1616. 3

[36] C. Yang, L. Xie, C. Su, and A. L. Yuille, "Snapshot distillation: Teacher-student optimization in one generation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2859–2868. 3

[37] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7103–7112. 5

[38] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. European Conference on Computer Vision (ECCV)*, 2016, pp. 483–499. 5
[39] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 466–481. 5
[40] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 529–545. 5
[41] P. Chandran, D. Bradley, M. Gross, and T. Beeler, "Attention-driven cropping for very high resolution facial landmark detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5861–5870. 5
[42] U. Iqbal, T. Molchanov, Pavloand Breuel, J. Gall, and J. Kautz, "Hand pose estimation via latent 2.5d heatmap regression," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 125–143. 5
[43] O. Chapelle and M. Wu, "Gradient descent optimization of smoothed information retrieval metrics," *Information retrieval*, vol. 13, no. 3, pp. 216–235, 2010. 5
[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 5
[45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. International Conference on Machine Learning (ICML)*, 2015, pp. 448–456. 7
[46] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011, pp. 315–323. 7
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 7
[48] K. Radlak, M. Bozek, and B. Smolka, "Silesian deception database: Presentation and analysis," in *Proc. ACM Multimodal Deception Detection Workshop (MDDW)*, 2015, pp. 29–35. 7
[49] M. Chau and M. Betke, "Real time eye tracking and blink detection with usb cameras," Boston University Computer Science Department, Tech. Rep., 2005. 8
[50] P. R. Tabrizi and R. A. Zoroofi, "Open/closed eye analysis for drowsiness detection," in *Proc. IEEE Image Processing Theory, Tools and Applications Workshop (IPTAW)*, 2008, pp. 1–7. 8



Wenzheng Zeng received the B.S. degree from Huazhong University of Science and Technology, China, in 2021, where he is currently pursuing the M.S. degree with the School of Artificial Intelligence and Automation. His research interests include eyeblink detection, anomaly detection, and gaze estimation.

Yang Xiao received his B.S., M.S., and Ph.D. degrees from Huazhong University of Science and Technology, China. He is currently an associate professor in the School of Artificial Intelligence and Automation at Huazhong University of Science and Technology, China. Previously, he was ever the research fellow in the School of Computer Engineering and Institute of Media Innovation at Nanyang Technological University, Singapore. Dr. Xiao was a recipient of IEEE Innovation Spotlight Research Paper Award 2020, EurAgEng Outstanding Paper Award 2018, and the Best Paper Award at ICIRA 2018. His research interests involve computer vision, image processing and machine learning.

Guilei Hu received the B.S. degree in School of Automation in 2017, and the master degree in School of Artificial Intelligence and Automation in 2020, Huazhong University of Science and Technology, Wuhan, China. Currently, he works in Shanghai Dingshan Information Technology Co., Ltd, Shanghai. His main research interest is on ADAS (Advanced Driving Assistance System).

Zhiguo Cao is a professor of School of Artificial Intelligence and Automation in Huazhong University of Science and Technology. He received his B.S. and M.S. degrees in communication and information System from the University of Electronic Science and Technology of China, and his Ph.D. degree in Pattern Recognition and Intelligent System from Huazhong University of Science and Technology. His research interests spread across image understanding and analysis, depth information extraction, 3d video processing, motion detection and human action analysis. His research results, which have published dozens of papers at international journals and prominent conferences, have been applied to automatic observation system for crop growth in agricultural, for weather phenomenon in meteorology and for object recognition in video surveillance system based on computer vision.

Sicheng Wei received the B.S. degree from Huazhong University of Science and Technology, China, in 2023. His research interests include eyeblink detection and anomaly detection.

Zhiwen Fang received his B.S. and M.S. degrees in the Automation School of Beihang University, and his PhD degree from Huazhong University of Science and Technology, China, in 2004, 2008, and 2017, respectively. He was the research fellow in Institute of Media Innovation at Nanyang Technological University, and the research scientist with the Institute of High Performance Computing, Research Agency for Science, Technology, and Research, Singapore. He is currently an associate professor in the School of Biomedical Engineering at Southern Medical University, China. His research interests include medical image analysis, object detection, anomaly detection and machine learning. He also serves as the Associate Editor of IET Image Processing.

Joey Tianyi Zhou received the Ph.D. degree in computer science from Nanyang Technological University, Singapore, in 2015. He is currently a Senior Scientist with the Centre for Frontier AI Research (CFAR), Research Agency for Science, Technology, and Research, Singapore. Dr. Zhou was a recipient of the Best Poster Honorable Mention at ACML 2012, the Best Paper Award from the BeyondLabeler Workshop on IJCAI 2016, the Best Paper Nomination at ECCV 2016, and the NIPS 2017 Best Reviewer Award. He has served as an Associate Editor for IEEE TETCI, Access, and IET Image Processing.

Junsong Yuan is Professor and Director of Visual Computing Lab at Department of Computer Science and Engineering (CSE), State University of New York at Buffalo, USA. Before joining SUNY Buffalo, he was Associate Professor (2015-2018) and Nanyang Assistant Professor (2009-2015) at Nanyang Technological University (NTU), Singapore. He obtained his Ph.D. from Northwestern University in 2009, M.Eng. from National University of Singapore in 2005, and B.Eng. from Huazhong University of Science Technology (HUST) in 2002. He received Chancellor's Award for Excellence in Scholarship and Creative Activities from SUNY, Nanyang Assistant Professorship from NTU, Outstanding EECS Ph.D. Thesis award from Northwestern University, and Best Paper Award from IEEE Trans. on Multimedia. He serves as Senior Area Editor of Journal of Visual Communication and Image Representation (JVCI), Associate Editor of IEEE Trans. on Pattern Analysis and Machine Intelligence (T-PAMI), IEEE Trans. on Image Processing (T-IP), IEEE Trans. on Circuits and Systems for Video Technology (T-CSVT), and Machine Vision and Applications (MVA). He also serves as General/Program Co-chair of ICME and Area Chair for CVPR, ICCV, ECCV, ACM MM, etc. He was elected Faculty Senator at Both SUNY Buffalo and NTU. He is a Fellow of IEEE and IAPR.