Real-time Multi-person Eyeblink Detection in the Wild for Untrimmed Video

Wenzheng Zeng[1], Yang Xiao[1†], Sicheng Wei[1], Jinfang Gan[1], Xintao Zhang[1], Zhiguo Cao[1], Zhiwen Fang[2,3] and Joey Tianyi Zhou[4,5]

[1]Key Laboratory of Image Processing and Intelligent Control, Ministry of Education, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China

[2]School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China

[3]Department of Rehabilitation Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou 510280, China

[4]Centre for Frontier AI Research, Agency for Science, Technology and Research (A*STAR), Singapore

[5]Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore

{wenzhengzeng, Yang_Xiao, sichengwei, jinfangan, u202115202, zgcao}@hust.edu.cn, fzw310@smu.edu.cn, zhouty@cfar.a-star.edu.sg

https://github.com/wenzhengzeng/MPEblink

Abstract

Real-time eyeblink detection in the wild can widely serve for fatigue detection, face anti-spoofing, emotion analysis, etc. The existing research efforts generally focus on single-person cases towards trimmed video. However, multi-person scenario within untrimmed videos is also important for practical applications, which has not been well concerned yet. To address this, we shed light on this research field for the first time with essential contributions on dataset, theory, and practices. In particular, a large-scale dataset termed MPEblink that involves 686 untrimmed videos with 8748 eyeblink events is proposed under multi-person conditions. The samples are captured from unconstrained films to reveal "in the wild" characteristics. Meanwhile, a real-time multi-person eyeblink detection method is also proposed. Being different from the existing counterparts, our proposition runs in a one-stage spatio-temporal way with end-to-end learning capacity. Specifically, it simultaneously addresses the sub-tasks of face detection, face tracking, and human instance-level eyeblink detection. This paradigm holds 2 main advantages: (1) eyeblink features can be facilitated via the face's global context (e.g., head pose and illumination condition) with joint optimization and interaction, and (2) addressing these sub-tasks in parallel instead of sequential manner can save time remarkably to meet the real-time running requirement. Experiments on MPEblink verify the essential challenges of real-time multi-person eyeblink detection in the wild for untrimmed video. Our method also outperforms existing approaches by large margins and with a high inference speed.

†Yang Xiao is corresponding author (Yang_Xiao@hust.edu.cn).

Figure 1. The illustration on multi-person eyeblink detection. This task aims at being aware of the presence of people and detecting their eyeblink activities at instance level.

1. Introduction

Real-time eyeblink detection in the wild is a recently emerged challenging research task [19] that can widely serve for fatigue detection [2], face anti-spoofing [34], affective analysis [7], etc. Although remarkable progress has been made [9, 10, 19], the existing methods generally focus on single-person cases within trimmed videos. Multi-person scenario within untrimmed videos has not been well concerned yet. However, detecting long-term eyeblink behaviors at multi-instance level is more preferred for some practical application scenarios. For example, it can be used to estimate attendees' attention level and emotional state change during social interaction [9, 33]. Thus, effective and real-time multi-person eyeblink detection in the wild for untrimmed video is indeed required.

To this end, we shed the light on this research problem with essential contributions on dataset, theory, and practices. First, a challenging labeled multi-person eyeblink de-

tection benchmark termed MPEblink is built under in-the-wild conditions. It consists of 686 untrimmed long videos captured from unconstrained movies to reveal the "in the wild" characteristics. The contained scenarios are realistic and diverse, such as social interactions and group activities. To our knowledge, MPEblink is the first multi-person eyeblink detection dataset that focuses on in-the-wild long videos. Fig. 1 illustrates a sample video with ground truth annotations within it. Different from the existing eyeblink detection benchmarks [10, 12, 19], the proposed benchmark aims at being aware of all the attendees along the whole video and detecting their eyeblinks at instance level. In summary, MPEblink is featured with multi-instance, unconstrained, and untrimmed, which makes it more challenging and realistic than previous formulations.

To perform eyeblink detection, previous methods [9, 10, 12, 19, 40] generally take a sequential approach containing face detection, face tracking, and classification on the pre-extracted local eye clues within a temporal window. Whereas such a pipeline seems reasonable, it has several critical drawbacks. First, the use of isolated components may lead to sub-optimal results as they are not jointly optimized and it is inefficient due to the inability to reuse the features of each stage. Second, the eyeblink features only contain the information from the pre-extracted local eye clues (i.e., a small part of face), lacking the useful information from global face context such as head pose and illumination condition that are crucial for detecting eyeblinks in the wild. Moreover, the pre-extracted local eye clues are also unreliable due to the localization challenge towards unconstrained scenarios. Third, the sequential approach leads to computational cost being sensitive to the subject amount, which is hard to meet the real-time running requirement.

To tackle these issues, we propose a one-stage multi-person eyeblink detection framework called InstBlink, which can simultaneously detect human faces, track them over time, and do eyeblink detection at instance level. InstBlink takes inspiration from the existing query-based methods [41, 45, 47, 48] and models the spatio-temporal face as well as eyeblink representations at instance level within each query. The insight is that the features can be effectively shared among these sub-tasks and the eyeblink features can be facilitated via face's global contexts (e.g., head pose and illumination condition) with joint optimization and interaction, especially in unconstrained in-the-wild cases. Experiments verify the superiority of InstBlink in both effectiveness and efficiency, while also highlighting the critical challenges of real-time multi-person eyeblink detection in the wild for untrimmed video.

The main contributions of this work lie in 3 folders:

- To our knowledge, it is the first time that instance-level multi-person eyeblink detection in untrimmed videos is formally defined and explored;
- We introduce an unconstrained multi-person eyeblink detection dataset MPEblink that contains 686 untrimmed videos with 8748 eyeblink events, featured with more realistic and challenging.
- We propose a one-stage multi-person eyeblink detection method that can jointly perform face detection, tracking, and instance-level eyeblink detection. Such a task-joint paradigm can benefit the sub-tasks uniformly.

2. Related Work

Eyeblink detection dataset. Existing eyeblink detection datasets [1, 10, 12, 13, 34, 36] generally focus on constrained indoor cases with a consistent environmental setup. HUST-LEBW [19] extends the scenarios to the unconstrained outdoor cases. From then on, the community starts to pay more attention to eyeblink detection in the wild. Nevertheless, HUST-LEBW mainly focuses on single-person scenarios towards trimmed videos, which limits the application of methods in more realistic scenarios such as group activities and social interactions. To address such limitation, we introduce a new dataset termed MPEblink. The proposed dataset is featured with multi-person, unconstrained, and untrimmed, thus being more realistic and challenging.

Eyeblink detection method. Existing methods generally judge eyeblinks from the pre-extracted local eye clues (e.g., local eye region [6, 9, 10, 19, 25] or landmarks around eyes [8, 14, 35, 40]). To obtain the local eye clues, the existing works generally run in a sequential way including face detection, face tracking, and landmark detection. Finally, the eyeblink result is classified from the pre-extracted local eye clues within a temporal window. Such a pipeline is inefficient due to the isolated optimization among different sub-tasks. Meanwhile, the eyeblink features only contain the information from the pre-extracted local eye clues, lacking useful global information such as head pose and illumination, and the pre-extracted eye clues are also unreliable due to the localization challenge toward unconstrained cases. Besides, due to the multi-stage characteristic, existing methods do not scale well with the number of people in the scene to meet the real-time running requirement. In this paper, we propose a one-stage eyeblink detection framework that can simultaneously detect human faces, track them along the time, and do eyeblink detection at instance level. Within it, eyeblink features can be facilitated via global face context, and the features can be effectively shared among the sub-task to achieve high inference speed.

Spatio-temporal action detection. Spatio-temporal action detection [15, 20, 22, 26, 27, 39, 48] aims at simultaneously obtaining the spatial and temporal location of an action, which is defined as an action tube [22]. The task only focuses on the isolated action tubes without instance awareness and thus can not make an instance-level analysis along the whole video. Different from that, our proposed multi-

Figure 2. The snapshots within MPEblink. Our dataset is featured with more realistic and diverse scenarios.

person eyeblink detection task aims at simultaneously being aware of the presence of people and analyzing their eyeblink behaviors at instance level.

Query-based methods. Query-based methods are impacting many computer vision tasks. After DETR [4], the pioneer that introduces the query-based framework to address object detection, numerous efforts have been subsequently paid to facilitate visual recognition under the query concept in the fields of object detection [24, 32, 41, 49], pose estimation [38], action detection [30, 48], and segmentation [5, 43, 45, 47]. Inspired by these, we build the first one-stage multi-person eyeblink detection framework for untrimmed video. We propose a unified solution to simultaneously model the instance-level face and eyeblink representations in the whole video.

3. The MPEblink Benchmark

Existing eyeblink detection datasets generally focus on single-person scenarios, and are also limited in aspects of constrained conditions or trimmed short videos. To explore unconstrained eyeblink detection under multi-person and untrimmed scenarios, we construct a large-scale multi-person eyeblink detection dataset termed MPEblink to shed the light on this research topic that has not been well studied before. The distinguishing characteristics of MPEblink lie in 3 aspects: multi-person, unconstrained, and untrimmed long video, which makes it more realistic and challenging than previous datasets.

3.1. Task Formulation

We think that a good multi-person eyeblink detection algorithm should be able to (1) detect and track human instances' faces reliably to ensure the instance-level analysis ability along the whole video, and (2) detect eyeblink boundaries accurately within each human instance to ensure the precise awareness of their eyeblink behaviors.

Formally, for an untrimmed video with T frames and N_{gt} people in it, towards the j-th person, let $\hat{l}^j \in \mathbb{R}^{T\times 4}$ denote its face bounding boxes across the video. $\hat{c}^j \in \mathbb{R}^T$ is given as the face labels to reflect the face existence of this instance in each frame (i.e., $\hat{c}_t^j \in \{1, 0\}$, where $\hat{c}_t^j = 0$ indicates that the face is not visible due to serious occlusion or departure at time t). $\hat{l}_t^j$ is also set to $\varnothing$ when the person is not visible. We use $\hat{l}^j$ and $\hat{c}^j$ to represent the instance-level information of each person. Suppose person j blinks $\hat{K}^j$ times in the video. Let $\hat{B}_{\hat{k}}^j = \left[\hat{s}_{\hat{k}}^j, \hat{e}_{\hat{k}}^j\right]$ denote the $\hat{k}$-th eyeblink event within this person, where $\hat{s}_k^j$ and $\hat{e}_k^j$ denote its starting and ending time. Suppose a multi-person eyeblink detection algorithm produces H human instance hypotheses. Within each human instance hypothesis i, it needs to produce a sequence of predicted face bounding boxes l^i, face classification scores c^i, and K^i eyeblink proposals $B_k^i = \left[s_k^i, e_k^i\right]$. Better performance can be acquired when predictions get closer to ground truths.

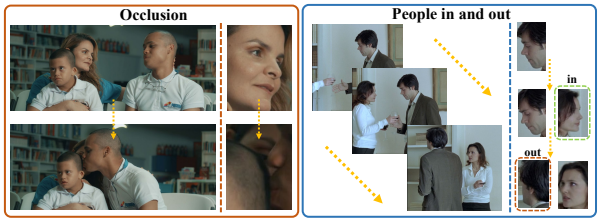

Figure 3. Occlusion and people in & out cases during interaction among instances.

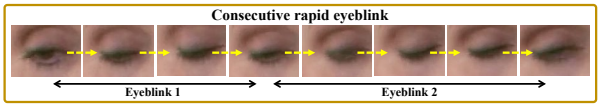

Figure 4. An example of rapid and consecutive eyeblinks in untrimmed video.

3.2. Data Collection

We collect 686 untrimmed video clips with various lengths (i.e., 7.1-85.9s) from 86 different unconstrained movies that involve the "in the wild" characteristics. Some snapshots of the acquired samples are shown in Fig. 2. It can be observed that the samples are with variational indoor and outdoor scenarios for different themes. Thus, the

Table 1. High-level statistics of MPEblink dataset and existing eyeblink detection datasets.

	Talking Face [1]	ZJU [34]	Eyeblink8 [12]	Silesian5 [36]	Researcher's Night [13]	mEBAL [10]	HUST-LEBW [19]	MPEblink (Ours)
Num. of videos	1	80	8	5	107	6000	673	686
Video length	3.3m	4.1-4.7s	2.7-8.9m	1.4-2.7m	0.3-3.4m	0.6s	0.5s	7.1-85.9s
Unconstrained	×	×	×	×	×	×	✓	✓
Instances/video	1	1	1	1	1	1	1	1-8
Blink count	61	255	353	300	1867	3000	381	8748

acquired samples from these movies are much more realistic and closer to practical applications. Meanwhile, due to the high divergence of the selected movie data source, the captured multi-person eyeblink samples are also of great challenges for effectively detecting eyeblinks. For instance, the pose, illumination condition, expression, and face size among instances differ a lot, even within the same video. Moreover, because of the various types of interaction among instances, there may exist severe occlusion or people in and out scenarios as shown in Fig. 3, making it harder for both instance localization and their eyeblink event detection. Besides, different from the existing unconstrained counterpart [19] that mainly focuses on trimmed short videos, the samples from MPEblink are untrimmed. As a result, we not only need to recognize eyeblink but also need to locate the starting and ending time of each eyeblink event, which is more challenging especially when rapid and consecutive eyeblinks occur. An example is shown in Fig. 4. It can be observed that the appearance difference among eye statuses is small when consecutive rapid eyeblinks happen, making it hard to distinguish the boundary of each eyeblink. Considering the above challenges together, it can be summarized that unconstrained multi-person eyeblink detection in untrimmed videos is indeed a challenging research task.

3.3. Data Annotation

For human instances in each video, we labeled their face bounding boxes exhaustively across the whole video. To facilitate the research of landmark-based methods and include a broader range of applications, 68 facial landmark positions [44] are also annotated. Technically, this part of the annotation is under a semi-supervised manner: we use the state-of-the-art face analysis engine InsightFace [21] to achieve face bounding box [16] and landmark [44] detection. To track instances across frames, a matching strategy that considers bounding box IoU and similarity among deep face features [11] is designed. Besides, human annotators carefully check the annotation quality and correct the wrong bounding boxes and tracking results generated by the algorithm. For eyeblink events within each instance, human annotators carefully define their start and end frame. Finally, 8748 eyeblink events are labeled.

3.4. Data Statistics

The dataset statistics comparison among MPEblink and the other eyeblink detection datasets are given in Table 1. Actually, MPEblink provides the largest number of eyeblink events (i.e., 8748) within 686 untrimmed videos. The biggest difference from the previous datasets is that the videos in our dataset contain various number of human instances (i.e., 1-8) with exhaustive annotation. The videos are captured under unconstrained in-the-wild conditions with high diversity as shown in Fig. 2. Compared with the existing in-the-wild counterpart [19] that generally focuses on trimmed short videos, the length of videos from our dataset is much longer and more diverse. In summary, MPEblink is featured with more realistic and challenging and thus has broader application values.

3.5. Evaluation Metrics

Existing metrics for eyeblink detection only focus on single-person cases without multi-instance awareness. To address this limitation we introduce 2 new metrics termed Inst-AP and Blink-AP to give attention to both instance awareness quality and eyeblink detection quality.

Inst-AP. It aims to evaluate the instance detection and tracking ability of methods, which is the basis for analyzing the eyeblink behaviors of each instance. We modify the standard evaluation metric video-AP in spatio-temporal action detection [22] to fit our task. Different from the original implementation, the 3D IoU is calculated between each instance proposal (i.e., a sequence of face bounding boxes) and ground truth across the whole video, rather than the action tube only. As a result, the proposed metric can reveal the algorithm's ability for detection and tracking instances, while video-AP can just reveal the localization ability of the isolated action tube without instance-level awareness. We report the averaged Inst-AP under the IoU from 50%-95% with step 5%. During calculating Inst-AP, each obtained true positive (TP) prediction has matched with one ground-truth instance, and thus, we can calculate the eyeblink detection accuracy of these TP predictions using the proposed Blink-AP. The TP predictions under IoU of 50% are used to calculate the subsequent Blink-AP.

Blink-AP. This metric is used to reflect the model's eyeblink detection ability within each instance. We tailor the standard AP metric in temporal action detection [3] into our task. Only the matched (i.e., true positive) predictions when calculating Inst-AP are taken into consideration in this metric as the blink accuracy of these predictions is rarely affected by the face detection and tracking accuracy. We report the Blink-AP under the temporal IoU of 50% and 75%.

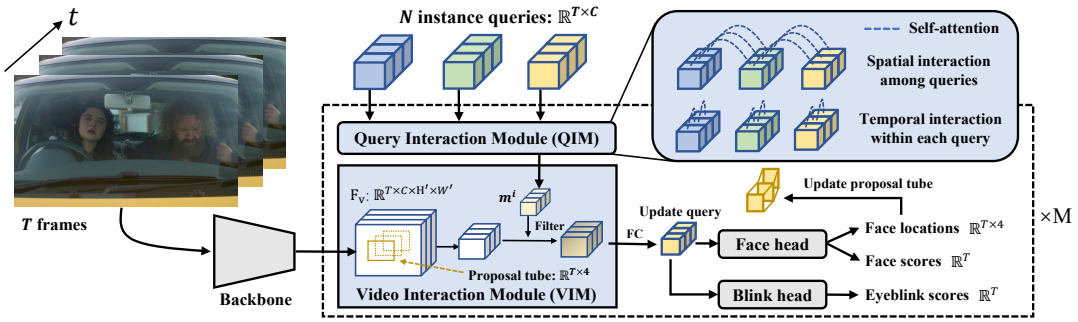

Figure 5. Overview of the InstBlink framework.

4. Method

In this section, we present our InstBlink that takes a video clip as input and directly outputs the face positions of each instance and their eyeblink intervals in the whole video clip. The InstBlink design takes inspiration from the existing query-based methods [4, 41, 45, 47, 48] but formulates the architecture to model both face and eyeblink representations at instance level along the video clip. The overall architecture is illustrated in Fig. 5.

Given a video clip $I \in \mathbb{R}^{T\times 3\times H\times W}$ where T denotes the number of frames and $H \times W$ is the spatial size of each frame, InstBlink first applies a backbone network to extract video feature $F_v \in \mathbb{R}^{T\times C\times H'\times W'}$ where C is the channel number and $H' \times W'$ is the spatial size of the feature. Afterwards, a query-based architecture iterates M times, which consists of 3 components: the Query Interaction Module (QIM), Video Interaction Module (VIM), and task-specific heads (i.e., face head and blink head). At the end of each interaction, the queries will be updated and the instance-level face and eyeblink predictions will be output by the task-specific heads. During inference, the output from the last iteration will be used as the final prediction results.

4.1. Instance Query

Within InstBlink, the spatio-temporal instance queries $\{q^i\}_{i=1}^N$ are responsible for characterizing every human instance's joint face and eyeblink features in a video. Each query contains T embeddings (i.e., $q^i \in \mathbb{R}^{T\times C}$), where C is the feature dimension. Each embedding generally focuses on the instance's face and eyeblink representations in the corresponding frame. Each query is also paired with a proposal tube $p^i \in \mathbb{R}^{T\times 4}$ (i.e., spatio-temporal bounding boxes across the time), which aims at indicating the face location of i-th instance across the entire video clip. At the first iteration of each complete forward propagation, q^i and p^i are initialized by copying 2 learnable parameters $\bar{q}^i \in \mathbb{R}^{1\times C}$ and $\bar{p}^i \in \mathbb{R}^{1\times 4}$ T times along the temporal dimension.

4.2. Query Interaction Module (QIM)

QIM targets at (1) enhancing the association between the specifc query and its corresponding human instance and (2) modeling spatio-temporal face and eyeblink representations of the associated instance. Specifically, we first adopt a spatial self-attention layer to allow spatial interaction among queries within each frame:

$$\{\boldsymbol{q}_t^i\}_{i=1}^N = \mathrm{MHSA}\left(\{\boldsymbol{q}_t^i\}_{i=1}^N\right), \quad t \in [1, T], \tag{1}$$

where MHSA is the multi-head self-attention [42]. Spatial interaction can build strong communication among queries to better model the instance features under complex circumstances such as occlusion due to human interactions. Then, temporal self-attention is used within each query along the temporal dimension to realize temporal interaction:

$$\{\boldsymbol{q}_t^i\}_{t=1}^T = \mathrm{MHSA}\left(\{\boldsymbol{q}_t^i\}_{t=1}^T\right), \quad i \in [1, N]. \tag{2}$$

Applying temporal interaction within each query allows the embeddings from different frames to communicate with each other to facilitate instance tracking and model temporal eyeblink representations of the corresponding instance.

4.3. Video Interaction Module (VIM)

VIM aims at collecting the face and eyeblink information of the target instance from the video feature. Particularly, the dynamic filters [41] m^i are first generated from each query embeddings q^i. Then, the filters will filter an RoI feature by dynamic convolution to extract highly related features for both face and eyeblink. The RoI feature is obtained by applying RoI align [17] on the video feature using the proposal tube p^i. After obtaining the filtered feature, a linear projection is applied to form the updated query feature $\tilde{q}^i$. The updated query feature will be used to make both face and eyeblink predictions by the task-specific heads.

4.4. Task-specific Heads

The instance-level face and eyeblink predictions can be obtained simultaneously by applying task-specific heads on the query features. The heads are shared among queries.

Face head. Given the updated query feature $\tilde{q}^i$, a Multilayer Perceptron (MLP) layer with Sigmoid normalization is used to indicate the existence of human faces as

$$c^i = Sigmoid\left(MLP_c\left(\tilde{q}^i\right)\right), \tag{3}$$

Figure 6. An example of the blink merging procedure, which takes the frame-level eyeblink scores b^i and outputs the interval-level eyeblink results B^i. The threshold is set to 0.3 for demonstration.

where $c^i \in \mathbb{R}^T$ denotes the face classification scores across frames. Face localization is achieved in a similar way as

$$l^i = MLP_l\left(\tilde{q}^i\right), \tag{4}$$

where $l^i \in \mathbb{R}^{T\times 4}$ denotes the face bounding boxes across frames. It will also be used to renew the proposal tube p^i.

Blink head. Eyeblink prediction towards $\tilde{q}^i$ is achieved by an MLP layer with Sigmoid normalization as

$$b^i = Sigmoid\left(MLP_b\left(\tilde{q}^i\right)\right), \tag{5}$$

where $b^i \in \mathbb{R}^T$ is the frame-level eyeblink scores. Each element b_t^i in it indicates the possibility that the t-th frame is inside an eyeblink event. During inference, we simply merge the adjacent frame-level eyeblink predictions based on a threshold (e.g., 0.3 in our implementation) to form the final interval-level eyeblink predictions $B_k^i = \left[s_k^i, e_k^i\right]$ where $k \in K$ is the k-th eyeblink interval within the total K eyeblink predictions. An intuitive example of the merging procedure is shown in Fig. 6.

Different from the existing works that only extract features from the local eye region, our blink head also considers useful global information such as global face appearance, head pose, and illumination condition to aid the implicit eye localization and eyeblink characterization. The reason is that the blink head directly filters global features from the query embeddings where the global face context is already stored. During training, the eyeblink clues will not only flow back to the RoI feature, but also to the instance query to enhance the eyeblink-related feature representation ability of queries. Moreover, because of the shared feature among blink head and face head, a kind of interaction and synergy is built where the information flow from blink head can also benefit the face representations to boost the face detection and tracking ability as verified in Sec. 5.3.

As we have N queries, we finally obtain the predictions from them in parallel:

$$\left\{y^i\right\}_{i=1}^N = \left\{\left(c^i, l^i, b^i\right)\right\}_{i=1}^N. \tag{6}$$

4.5. Training

To optimize InstBlink, one-to-one assignment between instance-level predictions $\left\{y^i\right\}_{i=1}^N$ and ground truths is first conducted based on their instance-level face tracklet similarity. Then, the loss is calculated concerning face and eyeblink items jointly for optimization. As in Sec. 3.1, the ground truth annotations can be summarized as

$$\left\{\hat{y}^j\right\}_{j=1}^{N_{gt}} = \left\{\left(\hat{c}^j, \hat{l}^j, \hat{b}^j\right)\right\}_{j=1}^{N_{gt}}, \tag{7}$$

where $\hat{c}^j \in \mathbb{R}^T$, $\hat{l}^j \in \mathbb{R}^{T\times 4}$, and $\hat{b}^j \in \mathbb{R}^T$ indicate the frame-level face existence, face position (i.e., bounding box) and eyeblink existence (i.e., $\hat{b}_t^j = 1$ indicates frame t is inside an eyeblink event and conversely for $\hat{b}_t^j = 0$, which can be derived from $\hat{B}^j$ using an inverse process in Fig. 6) respectively. We use Hungarian algorithm [23] to perform instance-level bipartite matching between predictions and ground truths. The matching cost is given as

$$\begin{aligned}\mathcal{L}_{\text{Hung}}\left(y^i, \hat{y}^j\right) = \sum_{t=1}^{T}\Big[&\mathcal{L}_{cls}\left(c_t^i, \hat{c}_t{}^j\right) \\ &+1_{\left\{\hat{c}_t^j \neq 0\right\}}\left(\mathcal{L}_{\text{box}}\left(l_t^i, \hat{l}_t^j\right)\right)\Big],\end{aligned} \tag{8}$$

where $\mathcal{L}_{cls}$ indicates the focal loss [29] for face existence classification. $\mathcal{L}_{box}$ is the combination of L1 loss and GIoU loss [37] for face localization. The loss weights are the same as [41, 47]. After obtaining the optimal assignment $\hat{\sigma}$ between predictions and ground truths, the network is optimized by the loss as

$$\mathcal{L} = \sum_{j=1}^{N_{gt}}\left(\mathcal{L}_{face}\left(\hat{y}^j, \hat{y}^{\hat{\sigma}(j)}\right) + \lambda\mathcal{L}_{blink}\left(\hat{y}^j, \hat{y}^{\hat{\sigma}(j)}\right)\right), \tag{9}$$

where $y^{\hat{\sigma}(j)}$ is the prediction that has been matched with the ground truth instance $\hat{y}^j$. $\mathcal{L}_{face}$ is the face tracklet detection loss that shares the same form as $\mathcal{L}_{Hung}$ (i.e., $\mathcal{L}_{face}\left(\hat{y}^j, \hat{y}^{\hat{\sigma}(j)}\right) = \mathcal{L}_{Hung}\left(\hat{y}^j, \hat{y}^{\hat{\sigma}(j)}\right)$), and $\mathcal{L}_{blink}$ denotes instance-level eyeblink detection loss as $\mathcal{L}_{blink}\left(\hat{y}^j, \hat{y}^{\hat{\sigma}(j)}\right) = \sum_{t=1}^{T}\mathcal{L}_{cls}\left(\hat{b}_t^j, \hat{b}_t^{\hat{\sigma}(j)}\right)$ with $\lambda = 5$. For unmatched predictions, only $\sum_{t=1}^{T}\mathcal{L}_{cls}\left(c_t^i, \hat{c}_t^i = 0\right)$ is used to supervise their face existence classification output c^i to be close to 0 (i.e., no face).

5. Experiment

Dataset and evaluation metrics. We first conduct experiments on the proposed MPEblink dataset. We split the dataset into 423 training videos, 128 validation videos, and 135 test videos. The videos from the same movie will only appear in one set. We train the network on the training set, validate it on the validation set and report the performance on the test set. We report Inst-AP to reveal the instance localization ability and Blink-AP to reveal the eyeblink detection ability within instances, as introduced in Sec. 3.5. Meanwhile, experiments are also conducted on HUST-LEBW for single-person and trimmed cases. Recall, Precision, and F1 score are reported following [19].

Table 2. The main results on the MPEblink dataset.

Type	Method	Blink-AP$_{50}$	Blink-AP$_{75}$	Inst-AP
Landmark	Soukupová and Cech [40]	0.50	0.05	56.70
	Blink detection+ [35]	0.62	0.08	
Region	Hu et al. [19]	2.68	0.04	
	Daza et al.	5.85	0.88	
	InstBlink (Ours)	**27.19**	**7.16**	**67.89**

Implementation details. We use ResNet-50-FPN [18, 28] backbone. The network is pre-trained on YouTube-VIS [46] under [47] for a general instance representation ability. The query number N and iteration time M are set to 50 and 4 respectively. AdamW [31] optimizer with a batch size of 8 is used to train the model. The initial learning rate is set to 2.5e-5 for the backbone and 2.5e-4 for the other parts. At the training stage, for the sake of memory efficiency, we use the frame rate of 12 FPS for image sampling and the input clip length is set to 11 that is longer than most eyeblink events (i.e., 0.2-0.4s). The frames are also resized to 640×360 before sending into the network. The whole training procedure lasts for 10,000 iterations and the learning rate is multiplied by 0.1 at iteration 6000 and 9000. During test, the original frame rate is used and the input clip length is set to 36 with a stride of 18. The predictions within adjacent clips are linked via concerning face bounding box IoU.

5.1. Benchmark Results on MPEblink Dataset

Baselines. Multi-person eyeblink detection in untrimmed videos is a new task that has not been well concerned before. Therefore, we tailor the existing eyeblink detection approaches with a unified instance detection and tracking method. Specifically, we use the state-of-the-art face analysis toolbox InsightFace [21] to achieve face [16] and landmark [44] detection. To track each instance, we link the face bounding boxes from the adjacent frames based on their similarity on box IoU. After obtaining the instance tracklets (i.e., a sequence of face bounding boxes and landmarks), we employ 4 representative eyeblink detection method [9, 19, 35, 40] within each instance tracklets. Such a sequential pipeline is commonly used in the existing approaches under the single-person assumption.

Main results. From Table 2, it can be summarized that:

• All of the multi-person eyeblink detection algorithms cannot achieve a satisfactory performance (i.e., Blink-AP$_{50}$ lower than 30% and Blink-AP$_{75}$ lower than 10%). This indicates that eyeblink detection under multi-person, untrimmed, and unconstrained scenarios is indeed challenging and has not been well solved yet.

• For Blink-AP, InstBlink significantly outperforms the others by large margins (i.e., 21% at least of Blink-AP$_{50}$ and 6% at least of Blink-AP$_{75}$), which verifies the superiority of the proposed framework. We argue that one essential reason is that our framework can model a better long-term temporal eyeblink representation than the frame-based method [9] and the sliding window based methods [19, 35, 40]. Moreover, under the proposed framework, eyeblink features can be facilitated via face's global context (e.g., head pose and illumination condition) with joint optimization and interaction, while previous works that utilize a sequential manner cannot. From Table 2, it can also be summarized that the landmark-based methods [35, 40] perform poorer than the region-based counterparts. We think that one essential reason is that landmark detection is unreliable under unconstrained conditions. A similar conclusion has already been made in [19] but the property of multi-person and long video in MPEblink make it worse for landmark-based eyeblink detection methods. Our method is region-based, but localizes eye region in an implicit way where global face context is included and thus becomes more robust.

• InstBlink also outperforms others on Inst-AP. We think that is because our framework can better model the long-term spatio-temporal instance representations, while the counterparts achieve tracking under a tracking-by-detection framework, which contains limited spatio-temporal modeling and may suffer from heavy occlusion.

Inference speed analysis. The result is listed in Table 3, assuming the 4 compared methods use InsightFace for face & landmark detection and InstBlink inferences within a clip length of 36. It can be seen that InstBlink is also of high inference speed (i.e., 112 FPS for network forwarding) while the real-time capacity of other methods is not superior. As the number of instances increases, their running time also increases while our method is not sensitive to the number of people because of the one-stage inference property.

5.2. Benchmark Results on HUST-LEBW Dataset

Experiments are also conducted on the HUST-LEBW dataset to explore the generality of InstBlink towards single-person and trimmed in-the-wild cases. The results are given in Table 4. For models trained on the HUST-LEBW dataset, our method still outperforms others by a large margin (i.e., 3.78% on F1 score), even with limited training data (less than 450 trimmed samples) to train our one-stage model for multiple tasks (face detection, tracking, and eyeblink detection). Meanwhile, the model trained on MPEblink can obtain 83.45% F1 score on HUST-LEBW. This indeed verifies the strong generalization ability of InstBlink and MPEblink towards eyeblink detection task.

5.3. Ablation Study

Spatial and temporal modeling in QIM. From Table 5, we can see that (1) without applying temporal interaction, the performance of Blink-AP drops from 27.19% to 4.58%. This indicates that temporal modeling is critical for eyeblink detection in untrimmed videos (i.e., only using appearance features can not accurately localize eyeblinks). It can also be observed that temporal interaction facilitates

Table 3. The inference speed comparison on a single NVIDIA 3090 GPU, assuming the compared methods use InsightFace for face detection (time consumption T=9.3ms including pre-processing) and landmark detection. #faces denotes the face amount in the scene.

Method	InstBlink (Ours)	Soukupová and Cech [40]	Blink detection+ [35]	Hu et al. [19]	Daza et al. [9]
Time/image (ms)	**8.9** + 2.6 for data processing	T(=9.3)+5.4×#faces	T+5.4×#faces	T+5.7×#faces	T+9.1×#faces

Table 4. Performance comparison on the HUST-LEBW dataset, where InstBlink_cross indicates InstBlink trained on MPEblink and tested on HUST-LEBW.

Training set	Method	Eye	Recall	Precision	F1
HUST-LEBW [19]	Soukupová and Cech [40]	Left	36.07	64.71	46.32
		Right	30.16	57.58	39.58
	Hu et al. [19]	Left	54.10	**89.19**	67.35
		Right	44.44	76.71	56.28
	Blink detection+ [35]	Both	58.99	80.05	67.90
	InstBlink (Ours)	Both	**97.64**	56.62	71.68
mEBAL [10]	Daza et al. [10]	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
MPEblink	InstBlink_cross (Ours)	Both	91.34	76.82	**83.45**

Table 5. Ablation studies on QIM and VIM.

Method	Blink-AP$_{50}$	Blink-AP$_{75}$	Inst-AP
w/o QIM	3.20	0.39	58.93
w/o temporal interaction in QIM	4.58	0.55	63.61
w/o spatial interaction in QIM	26.65	6.18	62.45
w/o filter operation in VIM	22.27	4.59	65.01
Full model	**27.19**	**7.16**	**67.89**

Table 6. The performance comparison with and without blink head on the MPEblink dataset.

Blink head	Inst-AP	Inst-AP$_{50}$	Inst-AP$_{75}$
×	65.86	81.73	71.23
✓	**67.89**	**84.51**	**73.76**

the performance of Inst-AP. We argue that this is because it can also build instance-level association among the features across frames to facilitate the instance tracking ability. (2) Spatial interaction can boost Inst-AP by a large margin (i.e., 5.44%) and slightly boost Blink-AP, as it may build strong communication among queries to better model the instance features under complex circumstances such as occlusion due to human interactions.

Filter operation in VIM. As in Table 5, compared with directly using RoI features to update query, using filtered RoI features can boost 4.92% on Blink-AP$_{50}$ and 2.88% on Inst-AP. We speculate that the filter operation can activate task-specific RoI features and resist background. Thus, finer face and eyeblink clues can be collected to update queries.

Multi-task learning mechanism. Here we study the effect of the blink head from a multi-task learning perspective. As shown in Table 6, adding blink head can also boost 2.03% on Inst-AP, which demonstrates that features for face detection and tracking can also benefit from eyeblink clues.

5.4. Qualitative Analysis

We visualize the prediction results of InstBlink in Fig. 7. The examples show that our model can aware the attendees robustly and detect their eyeblinks under different facial appearance and head pose. Nevertheless, it can not distinguish the rapid consecutive eyeblinks, which also reveals the challenge of eyeblink detection in untrimmed videos.

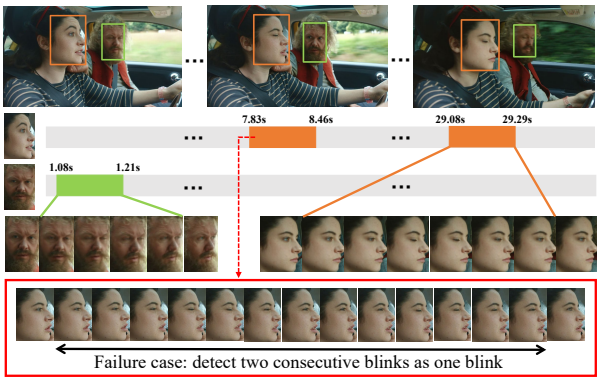

Figure 7. Qualitative examples of the predictions of InstBlink.

6. Conclusions and Limitations

In this work, we shed the light on a new research task termed multi-person eyeblink detection in the wild for untrimmed video. We formally define the task and introduce a multi-person eyeblink detection dataset named MPEblink. To perform multi-person eyeblink detection in untrimmed videos, we introduce a one-stage multi-person eyeblink detection framework InstBlink. Experiment results demonstrate the superiority of InstBlink in both effectiveness and efficiency. However, our SOTA performance is still unsatisfactory (i.e., Blink-AP$_{50}$ $<$ 30%). This verifies the challenges of unconstrained multi-person eyeblink detection in long videos. Besides, although the proposed dataset focuses on multi-person, unconstrained, and untrimmed scenarios that are more realistic and have a broader application value, there are lack of crowd scenes (e.g., $>$10 instances) within it. In the future, we will pay more attention to the crowd scenes and enrich the dataset.

Acknowledgment

This work is jointly supported by the National Natural Science Foundation of China (Grant No. 62271221 and U1913602), and Natural Science Foundation of Guangdong Province (Grant No. 2023A1515011260). Joey Tianyi Zhou is funded by the SERC (Science and Engineering Research Council) Central Research Fund (Use-Inspired Basic Research), and the Singapore Government's Research, and Innovation and Enterprise 2020 Plan (Advanced Manufacturing and Engineering Domain) under programmatic Grant A18A1b0045.

References

[1] Talking face video. Face&Gesture Recognition Working Group, IST-2000-26434. 2, 4

[2] Luis Miguel Bergasa, Jesús Nuevo, Miguel A Sotelo, Rafael Barea, and María Elena Lopez. Real-time system for monitoring driver vigilance. *IEEE Transactions on Intelligent Transportation Systems*, 7(1):63–77, 2006. 1

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015. 4

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. European conference on computer vision (ECCV)*, pages 213–229. Springer, 2020. 3, 5

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proc IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 3

[6] Kévin Cortacero, Tobias Fischer, and Yiannis Demiris. Rt-bene: A dataset and baselines for real-time blink estimation in natural environments. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 0–0, 2019. 2

[7] Antonio AV Cruz, Denny M Garcia, Carolina T Pinto, and Sheila P Cechetti. Spontaneous eyeblink activity. *The ocular surface*, 9(1):29–41, 2011. 1

[8] Simone Dari, Nico Epple, and Valentin Protschky. Unsupervised blink detection and driver drowsiness metrics on naturalistic driving data. In *Proc. IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6, 2020. 2

[9] Roberto Daza, Daniel DeAlcala, Aythami Morales, Ruben Tolosana, Ruth Cobos, and Julian Fierrez. Alebk: Feasibility study of attention level estimation via blink detection applied to e-learning. *arXiv preprint arXiv:2112.09165*, 2021. 1, 2, 7, 8

[10] Roberto Daza, Aythami Morales, Julian Fierrez, and Ruben Tolosana. Mebal: A multimodal database for eye blink detection and attention level estimation. In *Proc. International Conference on Multimodal Interaction (ICMI)*, pages 32–36, 2020. 1, 2, 4, 8

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proc. IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4690–4699, 2019. 4

[12] Tomas Drutarovsky and Andrej Fogelton. Eye blink detection using variance of motion vectors. In *Proc. European Conference on Computer Vision (ECCV)*, pages 436–448, 2014. 2, 4

[13] Andrej Fogelton and Wanda Benesova. Eye blink detection based on motion vectors analysis. *Computer Vision and Image Understanding*, 148:23–33, 2016. 2, 4

[14] Reza Ghoddoosian, Marnim Galib, and Vassilis Athitsos. A realistic dataset and baseline temporal model for early drowsiness detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops(CVPRW)*, pages 0–0, 2019. 2

[15] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6047–6056, 2018. 2

[16] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. In *Proc. International Conference on Learning Representations (ICLR)*, 2022. 4, 7

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. of the IEEE international conference on computer vision (ICCV)*, pages 2961–2969, 2017. 5

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 7

[19] 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*, 15:2194–2208, 2020. 1, 2, 4, 6, 7, 8

[20] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proc. IEEE international conference on computer vision (ICCV)*, pages 3192–3199, 2013. 2

[21] Jiankang Deng. Jia Guo. Insightface: 2d and 3d face analysis project. https://github.com/deepinsight/insightface, 2020. 4, 7

[22] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 4405–4413, 2017. 2, 4

[23] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[24] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13619–13627, 2022. 3

[25] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *Proc. International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018. 2

[26] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 13536–13545, 2021. 2

[27] Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *Proc. European Conference on Computer Vision (ECCV)*, pages 68–84. Springer, 2020. 2

[28] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2117–2125, 2017. 7

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. IEEE international conference on computer vision (ICCV)*, pages 2980–2988, 2017. 6

[30] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022. 3

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[32] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 3651–3660, 2021. 3

[33] Tamami Nakano. Information processing in the human brain revealed by eyeblink. *Brain and nerve= Shinkei kenkyu no shinpo*, 66(1):7–14, 2014. 1

[34] Gang Pan, Lin Sun, Z. Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. pages 1–8, 2007. 1, 2, 4

[35] Tran Thanh Phuong, Lam Thanh Hien, Ngo Duc Vinh, et al. An eye blink detection technique in video surveillance based on eye aspect ratio. In *Proc. 2022 24th International Conference on Advanced Communication Technology (ICACT)*, pages 534–538. IEEE, 2022. 2, 7, 8

[36] Krystian Radlak, Maciej Bozek, and Bogdan Smolka. Silesian deception database: Presentation and analysis. In *Proc. ACM Multimodal Deception Detection Workshop (MDDW)*, pages 29–35, 2015. 2, 4

[37] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 658–666, 2019. 6

[38] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11069–11078, 2022. 3

[39] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[40] Tereza Soukupová and Jan Cech. Real-time eye blink detection using facial landmarks. In *Proc. Computer Vision Winter Workshop (CVWW)*, pages 1–8, 2016. 2, 7, 8

[41] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proc. IEEE conference on computer vision and pattern recognition (CVPR)*, pages 14454–14463, 2021. 2, 3, 5, 6

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[43] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8741–8750, 2021. 3

[44] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proc. IEEE international conference on computer vision (ICCV)*, pages 3681–3691, 2021. 4, 7

[45] Jialian Wu, Sudhir Yarram, Hui Liang, Tian Lan, Junsong Yuan, Jayan Eledath, and Gerard Medioni. Efficient video instance segmentation via tracklet query and proposal. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 959–968, 2022. 2, 3, 5

[46] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 5188–5197, 2019. 7

[47] Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Wenyu Liu, Xun Zhao, and Ying Shan. Temporally efficient vision transformer for video instance segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2885–2895, 2022. 2, 3, 5, 6, 7

[48] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, et al. Tuber: Tubelet transformer for video action detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13598–13607, 2022. 2, 3, 5

[49] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *Proc. International Conference on Learning Representations (ICLR)*, 2021. 3