Supplementary: GateHUB: Gated History Unit with Background Suppression for Online Action Detection

Junwen Chen*‡ Gaurav Mittal*† Ye Yu† Yu Kong‡ Mei Chen†
†Microsoft ‡Rochester Institute of Technology
{gaurav.mittal, yu.ye, mei.chen}@microsoft.com {jc1088,yu.kong}@rit.edu

In this document, we provide additional analysis of our method GateHUB both quantitatively and qualitatively. We include details and analysis that were ready at the time of submission but could not be included in the paper due to the space constraints. Besides this document, we also include few videos demonstrating the task of online action detection.

1. Implementation Details

To extract features from TSN [12], we take the average of RGB features of 6 consecutive frames at 24 FPS to represent each frame at 4 FPS. Similarly, we stack optical flow maps of 5 frames preceding each frame along channel dimension at 24 FPS to obtain optical flow features for each frame at 4 FPS. Since TimeSformer [1] requires an input of 96 RGB frames, we uniformly sample 96 frames from the time duration set for past frames, t_{ps}, and future frames, t_f, for Future-augmented History (FaH) as input to TimeSformer and use the output corresponding to the last frame as the feature for a history frame.

2. Additional Quantitative Analysis

2.1. Future-augmented History (FaH) on I3D

Method	mAP (%)
WOAD [8]	67.1
w/o FaH	68.1
w/ FaH	**69.1**

Table 1. Ablation study for Future-augmented History (FaH) using I3D [3]. With FaH significantly outperforms both WOAD and without FaH. Without using FaH also outperforms WOAD showing that the other novel aspects of our method (*i.e*. GHU and background suppression objective) are also instrumental in improving online action detection even with I3D.

In Table 3c of the main paper, we show an ablation study on the impact of using Future-augmented History (FaH) with TimeSformer [1] feature backbone. We use TimeSformer as it supports multi-frame input and is therefore compatible with FaH. To further highlight the applicability and benefit of FaH on different spatio-temporal feature backbones supporting multi-frame input, we also conduct an ablation study with I3D [3]. Similar to TSN [12], I3D is a commonly used feature backbone for online action detection in prior art [8]. Table 1 shows the results for the ablation on FaH using I3D. We conducted the experiments on THUMOS'14. For comparison, we also provide the accuracy achieved by the existing method, WOAD [8], that uses I3D for THUMOS'14. From the table, we can observe that using I3D with FaH in GateHUB ('w/ FaH') significantly outperforms both WOAD and using I3D without FaH in GateHUB ('w/o FaH'). This highlights the significance of the proposed FaH module to make the history frames more informative using their *future*, *i.e*. subsequently observed frames. This, in turn, improves the history encoding and accuracy of current frame prediction. This also highlights that FaH can be successfully applied to improve accuracy on other spatio-temporal feature backbones that support multi-frame input. Moreover, we can observe that even without using FaH with I3D in GateHUB ('w/o FaH'), the accuracy is still 1% better than WOAD. This shows that other novel aspects of GateHUB, *i.e*. using Gated History Unit (GHU) to enhance or suppress history frames based on how informative they are to current frame prediction and using background suppression to apply separate emphasis on low-confident action and background frame predictions, are also instrumental in improving the accuracy regardless of the feature backbone.

2.2. Additional comparison on TSN pretrained on ActivityNet

As shown in Table 1 of the main paper, we compare GateHUB with existing state-of-the-art (SoTA) methods on the standard setting of using RGB and optical flow features from TSN [12] pretrained on Kinetics-400 [3]. Earlier approaches [15] often compare with the setting of using the

*Authors with equal contribution.
This work was done as Junwen Chen's internship project at Microsoft.

Method	Portion of Action									
	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%
IDN [5]	81.7	81.9	83.1	82.9	83.2	83.2	83.2	83.0	83.3	86.6
PKD [17]	82.1	83.5	86.1	87.2	88.3	88.4	89.0	88.7	88.9	87.7
OadTR [14]	81.2	84.9	87.4	87.7	88.2	89.9	88.9	88.8	87.6	86.7
LSTR [16]	84.4	85.6	87.2	87.8	88.8	89.4	89.6	89.9	90.0	90.1
GateHUB (Ours)	**84.5**	**87.6**	**89.5**	**90.0**	**90.2**	**91.0**	**91.3**	**91.3**	**91.3**	**90.7**

Table 2. Evaluation on TVSeries by dividing all action occurrences into ten equal parts (*i.e. portions of action*) and computing a separate mcAP(%) for each portion of action. GateHUB outperforms all existing methods for all portions of action considered. This shows that GateHUB is more accurate in predicting for the current frame irrespective of whether it is the start, middle or end of an action occurrence.

same TSN backbone but pretrained on ActivityNet [2]. So in addition to results on TSN and TimeSformer pretrained on Kinetics in the main paper, we compare GateHUB with SoTA methods on THUMOS'14 on this setting of using RGB and optical flow features from TSN pretrained on ActivityNet. We present the results in Table 3. From the table, we can observe that GateHUB is able to significantly outperform all existing methods by at least 3.8% on this setting. Compared to TSN pretrained on Kinetics, this setting gives consistently lower accuracy for all methods. Table 4 further shows the results on TVSeries. We can again observe that GateHUB is able to outperform all existing methods. This further validates that GateHUB can outperform all existing methods on multiple benchmark datasets using multiple different input feature representations.

Method	mAP (%)
CDC [11]	44.4
RED [7]	45.3
TRN [15]	47.2
FATS [9]	51.6
IDN [5]	50.0
LAP [10]	53.3
TFN [6]	55.7
OadTR [14]	58.3
LSTR [16]	65.3
GateHUB (Ours)	**69.1**

Table 3. Online action detection results on THUMOS'14 comparing GateHUB with SoTA methods on mAP (%) when the RGB and optical flow-based features are extracted from TSN pretrained on ActivityNet. We can see that GateHUB significantly outperforms all existing methods.

2.3. Evaluation on TVSeries for different portions of action

Following prior art [4, 14–16], we also evaluate the accuracy of online action detection on TVSeries when only a certain portion of the action occurrences is considered. The objective of this evaluation is to assess how well a method performs at different stages of an ongoing action. Following prior art, we divide each action occurrence into ten equal parts (*i.e. portions of action*). We then compute a separate mcAP for each portion of action over all action occurrences. For example, mcAP for $20-30\%$ *portion of action* refers to mcAP computed using only frames of an action occurrence lying between 20% and 30% of the total duration of that action occurrence. We tabulate the results across all portions of action in Table 2. From the table, we can observe that our method outperforms all existing methods for all the different portions of action considered. This shows that irrespective of whether it is the start, middle or end of an action occurrence, GateHUB is able to predict the action for the current frame more accurately than all existing methods.

Method	mcAP (%)
RED [7]	79.2
FATS [9]	81.7
TRN [15]	83.7
IDN [5]	84.7
TFN [6]	85.0
LAP [10]	85.3
OadTR [14]	85.4
LSTR [16]	88.1
GateHUB (Ours)	**88.4**

Table 4. Online action detection results on TVSeries comparing GateHUB with SoTA methods on mcAP (%) when the RGB and optical flow-based features are extracted from TSN pretrained on Activity Net. We can see that GateHUB outperforms all existing methods.

2.4. Effect of History and Present duration

History Duration (s)	Present Duration (s)			
	1	2	4	8
64	69.2	68.4	67.9	67.8
128	68.3	67.9	68.5	65.8
256	69.3	**70.7**	69.9	67.4
512	69.1	68.6	68.4	68.0

Table 5. Ablation study showing mAP (%) for different durations (in seconds) of history and present frames sampled at 4 FPS using RGB and optical flow features from TSN on THUMOS'14.

We show an analysis where we experiment with different durations of history and present frames in GateHUB in Table 5. We test on THUMOS'14 using RGB and optical flow features from TSN [12]. For each setting, the frames are sampled at 4 FPS. We consider a duration for history ranging from 64s to 512s and for present ranging from 1s to 8s (duration doubling for each subsequent setting). From the table, we can observe that we can get the best accuracy using history and present of duration 256s and 2s respectively. We can also observe that the model gets the worst performance when the duration of present is 8s for any given duration of history. This suggests that the present should constitute a very small set of most recently observed frames preceding the current frame. This allows to effectively model the most immediate observable context around the current frame which is important for accurate prediction for the current frame.

2.5. Action Anticipation Result

Following LSTR [16], we evaluate GateHUB on action anticipation task. Specifically, we anticipate the future up to 2 seconds at 4FPS by adding 8 learnable tokens to the most recent history frames $[h_t]_{t=-t_{pr}-1}^{0}$. Table 6 shows that GateHUB significantly outperforms the existing methods by 4.1% and 1.2% on THUMOS and TVSeries respectively, using the ActivityNet pretrained features.

Method	mAP (%)	mcAP (%)
RED [7]	37.5	75.1
TRN [15]	38.9	75.7
TTM [13]	40.9	77.9
LAP [10]	42.6	78.7
OadTR [14]	45.9	77.8
LSTR [16]	50.1	80.8
GateHUB (Ours)	**54.2**	**82.0**

Table 6. Results of online action anticipation using ActivityNet features on THUMOS'14 and TVSeries in terms of mAP and mcAP, respectively.

2.6. Analysis of gating scores G vs $Q_iK_i^T/\sqrt{dk}$

$Q_iK_i^T/\sqrt{d_k}$	0.22 ± 0.06	0.09 ± 0.04	0.10 ± 0.02	0.05 ± 0.01
G	-2.49	-1.70	0.92	0.98
$Q_iK_i^T/\sqrt{d_k} + G$	-2.27 ± 0.06	-1.61 ± 0.04	1.02 ± 0.02	1.03 ± 0.01

Figure 1. Analysis of gating scores G vs $Q_iK_i^T/\sqrt{dk}$ for the some of the most suppressed and enhanced history frames

We compare the value of G with $Q_iK_i^T/\sqrt{dk}$ to assess whether the gating scores G are indeed able to calibrate the attention weights. Statistically, on obtaining G and $Q_iK_i^T/\sqrt{dk}$ across all history frames, we find that G lies in $[-9.5, 1.0)$ and $Q_iK_i^T/\sqrt{dk}$ lies in $[-0.1, 0.6]$. So, G is large/small enough to change relative order of attention scores. Further, Fig 1 further provides G and range of $Q_iK_i^T/\sqrt{dk}$ and $Q_iK_i^T/\sqrt{dk} + G$ (Eqn 3, main paper) for some frames in Fig 3 of main paper. We can see that G is able to calibrate $Q_iK_i^T/\sqrt{dk}$ so that informative and uninformative history frames are correctly enhanced and suppressed respectively.

3. Additional Qualitative Analysis

We provide additional qualitative assessment by visualizing GateHUB's current frame prediction with and without GHU in Fig. 2. There are six video examples from THUMOS'14. For each example, we show the video frames at the top, then the ground truth (blue denoting the action occurrences), followed by current frame predictions using GateHUB with GHU (red) and without GHU (brown) where the confidence in the range $[0, 1]$ on y-axis denotes the probability of predicting the correct action. At the bottom of each example, we present the most suppressed and the most enhanced frames determined by GHU. From the visualization, we can observe that when using GHU (red), our method is able to significantly reduce false positives for the background frames (particularly that closely resemble action frames such as the frames closely following the second *Diving* and *Javelin Throw* action occurrence in Fig. 2). At the same time, without GHU (brown), the model is prone to high number of false positives (as can be been in the *Diving* example after the first action occurrence for frames showing swimming pool). In addition to reducing false positives, using GHU is also able to boost the confidence of true positives (as can been seen from *Shotput* example in Fig. 2).

Below the visualization of current frame prediction for each video in Fig. 2, we also visualize examples of the *most suppressed* and the *most enhanced* history frames in that video when ordered as per the gating scores G learned by GHU in 'w/ GHU' as per Eqn. 2 (main paper). From all the examples, we can observe that a significant number of the *most suppressed frames* contain athletes as they are walking in the field either to begin preparing for the action, leave after finishing the action or stopping to respond to the interviewer. In all these scenarios, we cannot draw any meaningful context about what and when the action will begin or end. As a result, GHU helps to suppress such history frames that are highly uninformative for current frame prediction. At the same time, a significant number of *most enhanced frames* are the frames where either the action is in progress or the athlete is close to commencing the action. Both these scenarios provide informative context in inferring what and when the action will take place. As as result, GHU enhances these frames that are highly informative for current frame prediction. We can also observe that occasionally few in-

Figure 2. Visualizing the current frame prediction for six videos from THUMOS'14 (separated by dotted lines). For each video, the first row shows the video frames, then ground truth (blue denoting action occurrence) followed by the plot for current frame prediction comparing GateHUB with GHU ('w/ GHU') (red) and without GHU ('w/o GHU') (brown). Below the plots for each video, we also highlight examples of the most suppressed and the most enhanced frames in the corresponding video as ordered based on the gating score G in GateHUB with GHU. 'w/ GHU' is able to reduce false positives observed in 'w/o GHU' where background frames closely resemble action frames.

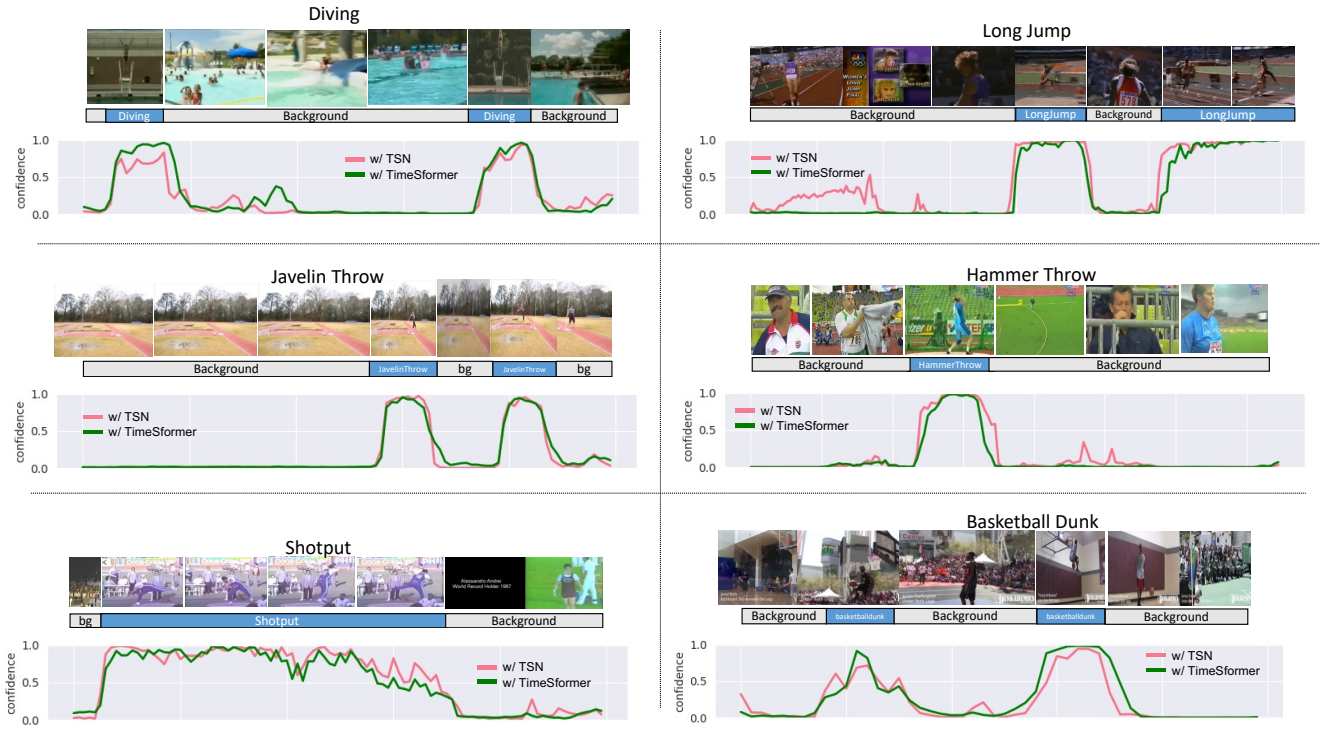

Figure 3. Visualizing the current frame prediction for six videos from THUMOS'14 (separated by dotted lines). For each video, the first row shows the video frames, then ground truth (blue denoting action occurrence) followed by the plot for current frame prediction comparing GateHUB using RGB features from TSN (red) and TimeSformer (green).

formative frames (such as the second and fourth frame in the 'most suppressed frames' for *Basketball Dunk*) get suppressed by GHU. This is likely due to the action being far-away in the frame making it difficult for the model to accurately assess the informative-ness of the frame. Spatially localizing small and far-away objects and their corresponding motion is still an open challenge. Capturing more fine-grained higher resolution features could potentially mitigate the problem.

In Fig. 3, we further visualize and compare GateHUB's current frame prediction using RGB features from TSN (red) and TimeSformer (green). We can observe that GateHUB using RGB features from TimeSformer (green) is more effective in reducing false positives while improving the confidence score for true positives (as can be seen from false positive reduction in *Long Jump* and true positive enhancement in *Basketball Dunk* in Fig. 3). This is potentially due to the comparatively limited feature representation capacity of the TSN feature backbone to extract sufficiently discriminative features when the frames include slow motion, motion blur, or small/far-away objects. In comparison, using the Timesformer feature backbone considerably mitigates false positives (green). This shows that with stronger feature representation, GateHUB can be more effective in reducing false positives while increasing true positives thereby improving current frame prediction. Also worth noting is that the confidence to correctly predict the *Shotput* action reduces for all methods in both Fig. 2 and Fig. 3 toward the end of the action occurrence. Similar to background suppression, we can explore putting separate emphasis on accurately predicting the frames near the boundary to mitigate such low-confident near-boundary predictions.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *ICML*, 2021. 1

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 2

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In *CVPR*, 2017. 1

[4] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In *ECCV*, 2016. 2

[5] Hyunjun Eun, Jinyoung Moon, Jongyoul Park, Chanho Jung, and Changick Kim. Learning to discriminate information for online action detection. In *CVPR*, 2020. 2

[6] Hyunjun Eun, Jinyoung Moon, Jongyoul Park, Chanho Jung, and Changick Kim. Temporal filtering networks for online action detection. *Pattern Recognition*, 2021. 2

[7] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. RED: Reinforced encoder-decoder networks for action anticipation. In *BMVC*, 2017. 2, 3

[8] Mingfei Gao, Yingbo Zhou, Ran Xu, Richard Socher, and Caiming Xiong. WOAD: Weakly supervised online action detection in untrimmed videos. In *CVPR*, 2021. 1

[9] Young Hwi Kim, Seonghyeon Nam, and Seon Joo Kim. Temporally smooth online action detection using cycle-consistent future anticipation. *Pattern Recognition*, 2021. 2

[10] Sanqing Qu, Guang Chen, Dan Xu, Jinhu Dong, Fan Lu, and Alois Knoll. LAP-Net: Adaptive features sampling via learning action progression for online action detection. *arXiv:2011.07915*, 2020. 2, 3

[11] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *CVPR*, 2017. 2

[12] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1, 3

[13] Wen Wang, Xiaojiang Peng, Yanzhou Su, Yu Qiao, and Jian Cheng. Ttpp: Temporal transformer with progressive prediction for efficient action anticipation. *Neurocomputing*, 438:270–279, 2021. 3

[14] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. Oadtr: Online action detection with transformers. *ICCV*, 2021. 2, 3

[15] Mingze Xu, Mingfei Gao, Yi-Ting Chen, Larry S Davis, and David J Crandall. Temporal recurrent networks for online action detection. In *ICCV*, 2019. 1, 2, 3

[16] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *NeurIPS*, 2021. 2, 3

[17] Peisen Zhao, Jiajie Wang, Lingxi Xie, Ya Zhang, Yanfeng Wang, and Qi Tian. Privileged knowledge distillation for online action detection. *arXiv:2011.09158*, 2020. 2