Boosting Semi-supervised Video Action Detection with Temporal Context
—*Supplementary Material*—

This supplementary material provides more implementation details (Section 1), additional ablation studies and an algorithm table of the global-local context fusion (Section 2) and more qualitative results (Section 3).

1. Implementation Details

GPU. Throughout the entire training and evaluation processes on both UCF101-24 [10] and JHMDB-21 [5] datasets, we utilized a single NVIDIA A100-80GB GPU.

Evaluation Protocol. Following previous studies [7, 14], for evaluation, we divided each video in the test set into several clips, each consisting of 8 frames. If there were not enough frames to form a clip due to misalignment with the video frame count, we zero-padded those missed frames to form the clip. After creating such clips for evaluation, we assessed our method on a clip-by-clip basis.

Hyper-parameter Selection. The update ratio β in Eq. (5) and τ in Eq. (8) in the paper is set to 0.995 and 0.1, respectively, following common practices in various semi-supervised frameworks [1, 2, 8] and [6, 9].

Training Epochs. For UCF101-24 [10], we trained the model for 250 epochs, while for JHMDB-21 [5], we trained it for 150 epochs in an end-to-end manner.

Training on AVA [4] with TubeR [13] and STMixer [12]. During training on AVA with TubeR, feature maps used for spatio-temporal semi-supervised learning are mainly obtained from the intermediate features of its backbone (CSN-50 [11]). For STMixer, we utilize 4-D feature from its video backbone (SlowFast [3] in the experiment).

2. Additional ablation studies and algorithm table of the global-local context fusion

Performance on corrupted clips. We found that the proposed temporal consistency learning significantly enhances performance on corrupted clips, where misaligned boundaries at the video's end cause temporal misalignment. We conducted an experiment comparing models trained with (denoted as **ST**) and without (denoted as **S**) temporal consistency losses on corrupted and normal clips from each dataset's test set. "Corrupted" clips have incomplete frames (e.g., some frames are missing and zero-padded), while "normal" clips have all frames intact. Corrupted clips constitute 6% for UCF101-24 and 37% for JHMDB-21 of all clips used for evaluating test set performance. Results in Table 1 show that the model trained with proposed temporal consistency losses achieves a smaller performance gap between corrupted and normal clips across all settings.

Dataset	Method	Clip status	f-mAP		v-mAP	
			0.2	0.5	0.2	0.5
UCF101-24	ST	normal.	93.0 / 79.9	81.0 / 69.4	98.2 / 84.7	84.0 / 72.9
		corrupted.	85.1 / 77.5	64.0 / 59.5	79.4 / 66.0	59.7 / 49.3
	S	normal.	91.7 / 77.8	78.5 / 67.6	97.9 / 81.8	80.4 / 68.2
		corrupted.	75.9 / 69.9	56.3 / 51.0	78.8 / 60.1	46.3 / 41.2
JHMDB-21	ST	normal.	99.4 / 38.9	88.1 / 38.0	99.5 / 39.0	89.7 / 37.7
		corrupted.	92.0 / 37.1	71.0 / 29.6	96.5 / 36.8	72.5 / 31.2
	S	normal.	99.3 / 36.3	83.6 / 34.8	99.0 / 38.0	86.9 / 35.5
		corrupted.	88.4 / 34.7	55.5 / 26.3	95.7 / 34.8	57.3 / 25.9

Table 1. Ablation studies of performance on *corrupted* and *normal* clips for each type of model, **S** and **ST** on UCF101-24 and JHMDB-21 test set. Performance is presented first without the action label, followed by its inclusion after the '/'.

Method	UCF101-24			JHMDB-21		
	f-mAP	v-mAP		f-mAP	v-mAP	
	0.5	0.2	0.5	0.5	0.2	0.5
Random	79.5 / 67.9	97.0 / **83.7**	82.1 / 71.3	81.2 / **35.1**	97.1 / **38.5**	81.2 / **35.4**
Fixed	**80.0 / 68.8**	**97.1** / 83.5	**82.5 / 71.5**	**81.8** / 34.9	**98.5** / 38.2	**83.3** / 35.3

Table 2. Ablation studies of the impact of the sampling method for shared frames on UCF101-24 and JHMDB-21 test sets.

m	UCF101-24			JHMDB-21		
	f-mAP	v-mAP		f-mAP	v-mAP	
	0.5	0.2	0.5	0.5	0.2	0.5
2	77.5 / 66.1	**97.3** / 81.2	81.1 / 70.0	81.1 / 34.2	97.1 / 37.4	79.9 / 34.8
4	80.0 / **68.8**	97.1 / **83.5**	**82.5 / 71.5**	**81.8** / 34.9	**98.5 / 38.2**	**83.3** / 35.3
6	**80.1** / 68.7	97.0 / 83.2	82.2 / 71.3	81.6 / **35.5**	97.4 / 38.1	81.0 / **35.8**

Table 3. Ablation studies of the impact of the number of the shared frames m on UCF101-24 and JHMDB-21 test sets.

T	UCF101-24				JHMDB-21			
	f-mAP		v-mAP		f-mAP		v-mAP	
	0.2	0.5	0.2	0.5	0.2	0.5	0.2	0.5
✓	**92.5 / 79.8**	**80.0 / 68.8**	**97.1 / 83.5**	**82.5 / 71.5**	**97.0 / 38.2**	**81.8 / 34.9**	**98.5 / 38.2**	**83.3 / 35.3**
✗	92.3 / 79.5	79.8 / 68.1	**97.1** / 83.1	82.3 / 70.9	96.4 / 36.7	80.1 / 33.6	98.1 / 36.9	79.7 / 34.2

Table 4. Ablation studies for constructing paths in time-ordered manner on UCF101-24 and JHMDB-21 test sets. **T** denotes time-ordered paths.

s	UCF101-24				JHMDB-21			
	f-mAP		v-mAP		f-mAP		v-mAP	
	0.2	0.5	0.2	0.5	0.2	0.5	0.2	0.5
3	92.5 / 79.8	80.0 / 68.8	97.1 / 83.5	82.5 / 71.5	97.0 / 38.2	81.8 / 34.9	98.5 / 38.2	83.3 / 35.3
5	92.4 / 80.1	80.5 / 69.0	97.0 / 83.7	**82.9** / 72.0	97.5 / 38.3	81.7 / 35.0	**99.0** / 38.4	83.2 / 35.8
7	**92.7 / 80.2**	**80.6 / 69.4**	**97.3 / 84.0**	**82.9 / 72.2**	**97.8 / 38.5**	**82.0 / 35.3**	**99.0 / 38.5**	**83.8 / 36.0**

Table 5. Ablation studies of the number of sampled paths s for the global-local context fusion on UCF101-24 and JHMDB-21 test sets.

Sampling method for shared frames. During semi-supervised learning, we kept the number of shared frames m fixed at 4, as explained in Section 4.1. We investigated the impact of two sampling methods for shared frames on the UCF101-24 and JHMDB-21 test sets: one with a random number of shared frames (2 to 7 out of 8) and the other with a fixed number of shared frames (4, as described in the paper). Results in Table 2 show performance improvement in 8 out of 12 cases with the fixed method.

Type	UCF101-24				JHMDB-21			
	f-mAP		v-mAP		f-mAP		v-mAP	
	0.2	0.5	0.2	0.5	0.2	0.5	0.2	0.5
B	91.8 / **79.9**	79.5 / 68.6	96.6 / 83.4	82.3 / **71.5**	96.7 / 38.1	80.5 / 33.3	**98.7 / 38.3**	78.9 / 32.1
S	91.3 / 79.6	79.1 / 67.9	96.9 / 83.1	81.8 / 70.9	94.9 / 35.7	78.8 / 32.4	97.7 / 37.9	77.3 / 33.1
T	**92.5** / 79.8	**80.0 / 68.8**	**97.1 / 83.5**	**82.5 / 71.5**	**97.0 / 38.2**	**81.8 / 34.9**	98.5 / 38.2	**83.3 / 35.3**

Table 6. Ablation studies of design choices for the global-local context fusion on the UCF101-24 and JHMDB-21 test sets. **B** denotes the fusion is applied to both output of the student and teacher networks, and **S** and **T** denote that the fusion is only applied to the output of the student and teacher network, respectively. Performance is reported without the action label, then with the label included (separated by '/'). The best and second best results are marked in **bold** and underline, respectively.

Impact on the number of shared frames m. The shared frames between two overlapped clips is important for semi-supervised learning in the temporal domain, as outlined in Eq.(7) and Eq.(8). We examined the impact of varying the number of shared frames m. Results, summarized in Table 3, indicate that $m = 4$ yielded the best performance in most cases.

Time-Ordered Paths for GLF. In global-local context fusion (GLF), we aggregate and propagate information in the temporal domain in a time-ordered manner. This strategy is essential as the temporal evolution of a video offers crucial cues for content understanding. We conducted ablation studies comparing paths constructed in a time-ordered manner to randomly constructed ones. Results in Table 4 demonstrate the superiority of this design choice.

Effect of the number of sampled paths s per target frame in GLF. In GLF, we randomly sample s paths per target frame for temporal dropout regularization with computational efficiency. We investigated the impact of varying the number of paths per target frame. Results in Table 5 demonstrate a general performance improvement with increasing sampled paths. Though more sampled paths result in better performance, we did not increase the number of paths further for computational efficiency; please note that three paths were enough to attain the state-of-the-art.

Applying Global-Local Context Fusion to Both Outputs of the Student and Teacher Networks. We apply the global-local context fusion to the output of the teacher network, as it provides pseudo-supervision to the student network. Ablation studies of this design choice are conducted, and the results are shown in Table 6, demonstrating that our method (**T**) achieved either the best or second-best performance for all evaluation settings.

Algorithm table for the global-local context fusion. We provide the algorithm table for the global-local context in Algorithm 1.

3. More qualitative results

In this section, we present additional qualitative results in Fig. 1 and Fig. 2 on the test sets of UCF101-24 [10] and JHMDB-21 [5], respectively. Moreover, we present qualitative results in Fig. 3 to demonstrate the effectiveness of the global-local context fusion (GLF) on the test set of UCF101-24, including a case of corrupted clip (**Cliff Diving**).

Algorithm 1 Global-Local Context Fusion

Input: Unlabeled video X, Pseudo localization map $M \in \{0,1\}^{n\times h\times w}$, Pixel-wise feature embedding $E \in \mathbb{R}^{n\times d\times h\times w}$
Output: Fused feature embeddings for each target frame
NewEmbs $\leftarrow$ [] $\quad\triangleright$ Empty feature list for each frame
for $j = 1$ to n **do**
  $\Omega_j \leftarrow$ {calculate all possible paths for v_j}
  $\Omega_j^* \leftarrow sample(\Omega_j, s)$ $\quad\triangleright$ Sample s paths from Ω_j
  Candid$_j$ $\leftarrow$ [] $\quad\triangleright$ Empty feature list for each path
  $i \leftarrow 0$
  for p in Ω_j^* **do**
    for $l = 0$ to length$(p) - 2$ **do**
      $v_s, v_t \leftarrow p[l], p[l+1]$
      $E_s \leftarrow \begin{cases} E_s & \text{if } l = 0, \\ E_{\text{prev}} & \text{otherwise} \end{cases}$
      $E_s^p \leftarrow g(E_s, M_s)$, $E_s^n \leftarrow g(E_s, \neg M_s)$
      $E'_{t,k} \leftarrow \begin{cases} f(E_s^p \oplus E_{t,k}) & \text{if } M_{t,k} = 1, \\ f(E_s^n \oplus E_{t,k}) & \text{otherwise} \end{cases}$
      $E_{\text{prev}} \leftarrow E'_t$
    end for
    Candid$_j[i] \leftarrow E'_t$
    $i \leftarrow i + 1$
  end for
  $E_j^* \leftarrow$ average(Candid$_j$) $\quad\triangleright$ Average over Candid$_j$ to get final fused feature
  NewEmbs$[j] \leftarrow E_j^*$
end for
Output NewEmbs

References

[1] Zexi Chen, Benjamin Dutton, Bharathkumar Ramachandra, Tianfu Wu, and Ranga Raju Vatsavai. Local clustering with mean teacher for semi-supervised learning. *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6243–6250, 2020. 1

[2] Mario Döbler, Robert A. Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7704–7714, June 2023. 1

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vi-*

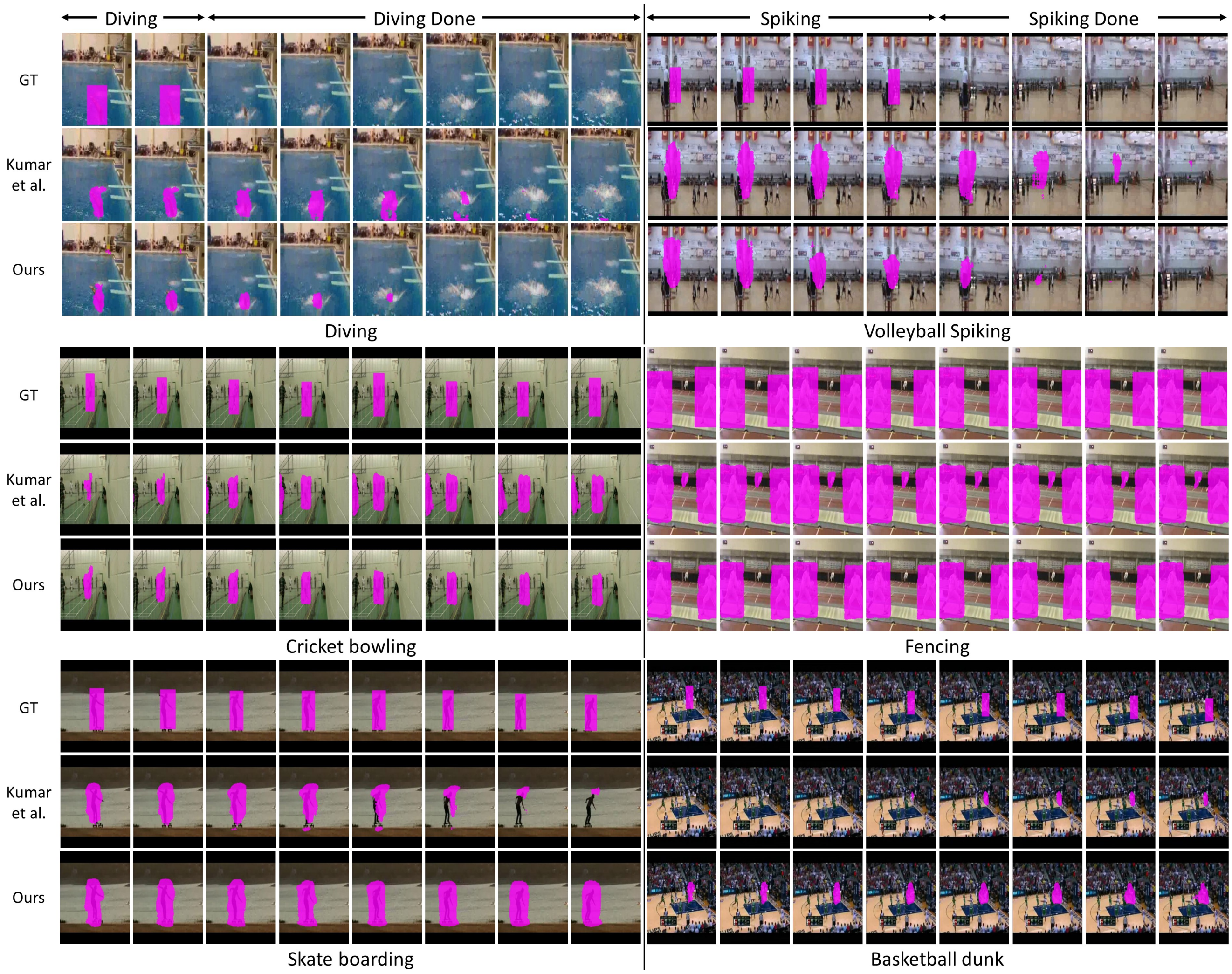

Figure 1. Qualitative results on *test* set of UCF101-24 [10] with ours and Kumar *et al.* [7]

sion, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pages 6201–6210. IEEE, 2019. 1

[4] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018. 1

[5] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *2013 IEEE International Conference on Computer Vision*, pages 3192–3199, 2013. 1, 2, 4

[6] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. 1

[7] Akash Kumar and Yogesh Singh Rawat. End-to-end semi-supervised learning for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14700–14710, 2022. 1, 3, 4

[8] Donghyeon Kwon and Suha Kwak. Semi-supervised semantic segmentation with error localization network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9957–9967, June 2022. 1

[9] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[10] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1, 2, 3, 5

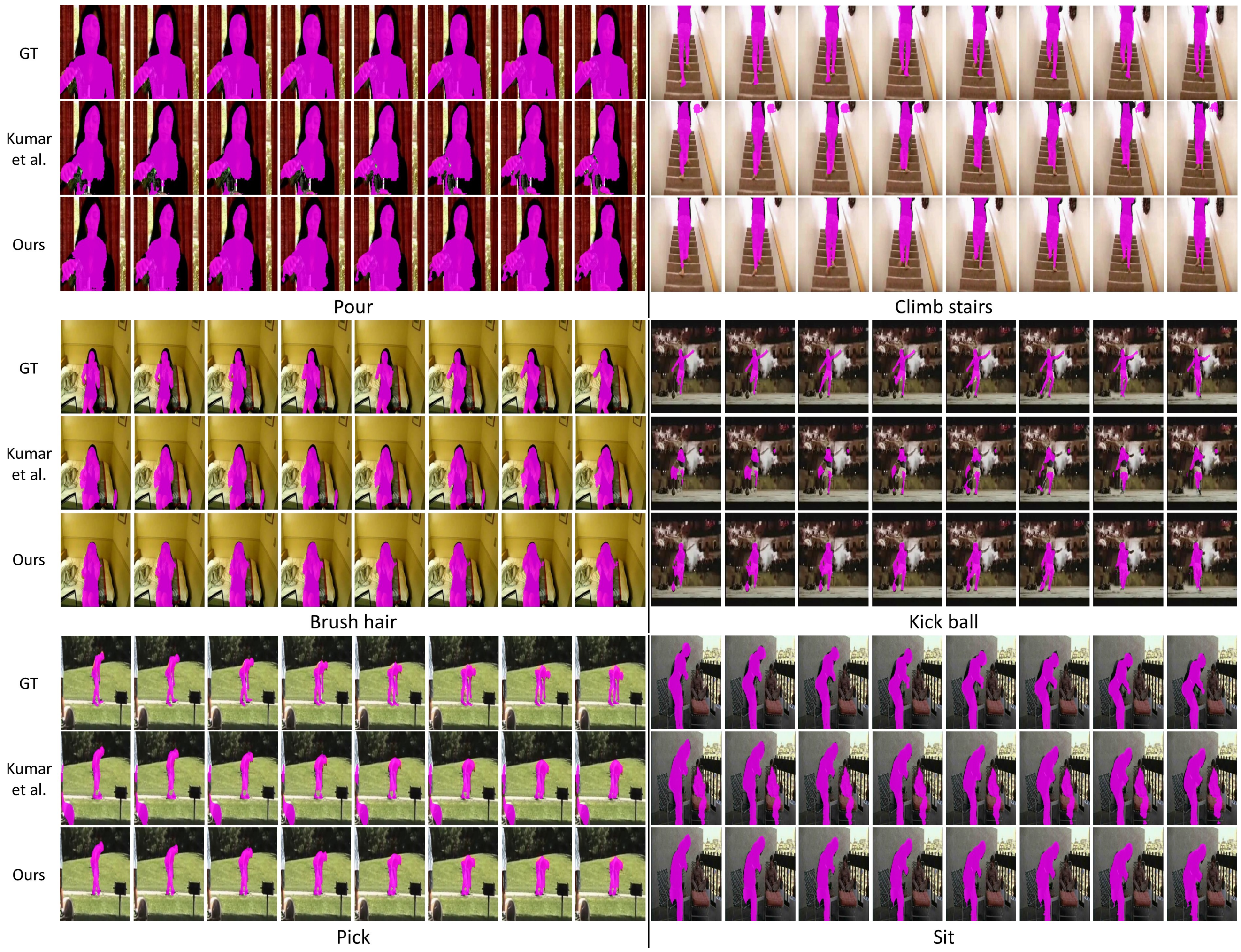

Figure 2. Qualitative results on *test* set of JHMDB-21 [5] with ours and Kumar *et al.* [7]

[11] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5551–5560, 2019. 1

[12] Tao Wu, Mengqi Cao, Ziteng Gao, Gangshan Wu, and Limin Wang. Stmixer: A one-stage sparse action detector. In *CVPR*, 2023. 1

[13] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13598–13607, 2022. 1

[14] Xian Zhong, Aoyu Yi, Wenxuan Liu, Wenxin Huang, Chengming Zou, and Zheng Wang. Background-weakening consistency regularization for semi-supervised video action detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

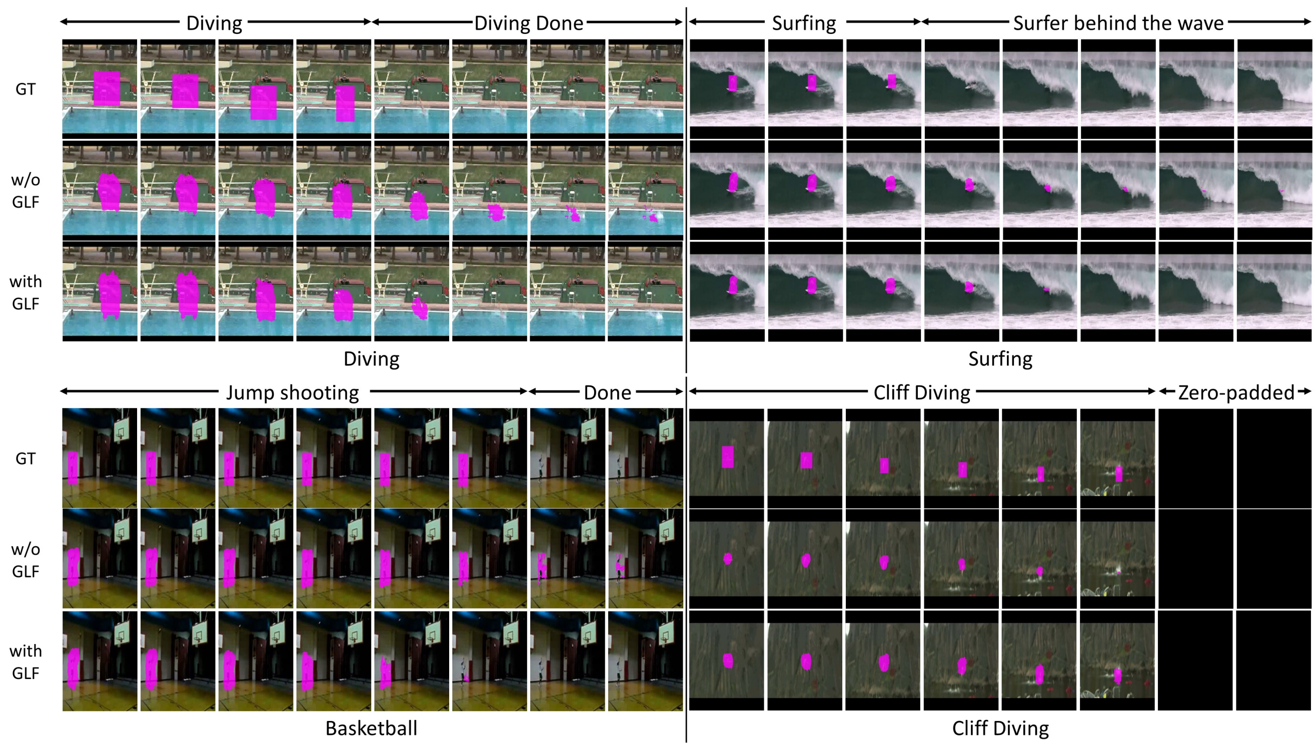

Figure 3. Qualitative results on *test* set of UCF101-24 [10] with GLF and without GLF