Benchmarking the Robustness of Temporal Action Detection Models Against Temporal Corruptions

Supplementary Material

In the supplementary material, we provide more details and more experimental results of our work. We organize the supplementary into the following sections.

- In Section A, we provide more ablated results of our Temporal-Robust Consistency (TRC) loss.
- In Section B, we include more analysis using DETAD on the ActivityNet-v1.3-C and THUMOS14-C datasets.
- In Section C, we show more examples of videos corrupted by our proposed temporal corruptions.
- In Section D, we present detailed results of TAD models under the five types of corruptions on THUMOS14-C and ActivityNet-v1.3-C.
- In Section E, we investigated the difference between generating black frames in action-background pairs and solely within actions.
- In Section F, we present experiments involving the addition of various types of corruptions in action segments.
- In Section G, we demonstrate the experimental results with a wider range of corruptions types.
- In Section H, we provide the experimental results of adding temporal corruptions to the MultiThumos dataset.

A. More Ablated Results of Our TRC Loss

Action-cetric sampling strategy in TRC loss. As discussed in Section 5.2, considering the characteristics of TAD, we select predictions that are more temporally aligned with the action instance to compute the TRC loss. Here, we design two variants: **1) Full-Video**: using all predictions without sampling, and **2) Full-Action**: using predictions whose center falls within the action instance. As can be observed from Table A, compared to the two variants, our proposed action-centric sampling method shows greater robustness on corrupted data and improvement on clean data while enjoying higher computational efficiency.

Table A. Comparison of different sampling strategies in TRC loss, measured by the performance of TriDet on THUMOS14-C. Our action-centric sampling leads to the best results on both clean and corrupted data.

Sampling Strategy	Clean mAP	Corrupted mAP
Without TRC	75.16	61.10
Full Video	74.21 (0.95 ↓)	67.21 (6.11 ↑)
Full Action	75.04 (0.12 ↓)	67.23 (6.13 ↑)
Action Center (Ours)	**75.60 (0.44 ↑)**	**68.28 (7.18 ↑)**

The Choice of Alignment Loss. Our approach to enhancing model robustness involves aligning predictions based on clean and corrupted videos. We compared Mean Square Error (MSE) and Kullback-Leibler (KL) divergence loss with our TRC loss. Models trained with different alignment losses are tested on clean and corrupted data. From Table B, all three types of losses can enhance the model's robustness on corrupted data, verifying the effectiveness of alignment. Notably, our proposed TRC loss not only enhances robustness but also improves the performance of clean data. Therefore, our method provides a new perspective that the performance and robustness of TAD methods can be simultaneously enhanced.

Table B. Comparison of different alignment loss, measured by the performance of TriDet on THUMOS14-C. All losses improve robustness while only our TRC loss enhances the performance on clean data.

Loss Function	Clean mAP	Corrupted mAP
Without Alignment	75.16	61.10
Mean Square Error	74.59 (0.57 ↓)	62.79 (1.69 ↑)
KL divergence	73.32 (1.84 ↓)	67.81 (6.71 ↑)
TRC (Ours)	**75.60 (0.44 ↑)**	**68.28 (7.18 ↑)**

B. More Analysis Using DETAD [1]

We present further analyses for more TAD models and more datasets using DETAD, a tool for diagnosing TAD models. Figure A depicts the results of analysis using DETAD on the predictions of TriDet on ActivityNet-v1.3-C, while Figure B illustrates the analysis of ActionFormer on THUMOS14-C. Evidently, the conclusions align consistently with those presented in the main paper across different models and datasets. Specifically, on our corrupted datasets, the results indicate a significant increase in localization errors, with minimal change in classification errors. This observation underscores the critical corruption introduced by our dataset, emphasizing its impact on temporal continuity (*i.e.*, localization) rather than compromising action recognition (*i.e.*, classification).

C. More Examples of Corrupted Videos

Figures C present more illustrative examples demonstrating the incorporation of five distinct types of corruptions into the actions depicted in the video. Although we only corrupt a small portion of frames within the action, and some of these corruptions may not appear significantly different from the surrounding clean frames—at least not to

Table C. The performance of TAD models concerning corruption robustness on THUMOS14-C, considering five distinct types of corruptions and three different levels, measured by mAP when the tIoU is set to 0.5.

Model	Feature	Clean Frame	Corruption Type														
			Black Frame			Packet Loss			Overexposure			Motion Blur			Occlusion		
			Level 1	Level 2	Level 3	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
BasicTAD	SlowOnly	59.17	46.82	26.07	16.23	56.17	45.64	41.53	54.78	33.63	21.95	53.85	36.98	28.51	48.66	32.41	22.54
E2E-TAD	SlowFast	56.41	29.23	15.54	12.22	40.32	32.07	30.23	50.03	24.33	14.13	45.33	41.51	39.30	38.86	25.73	19.46
TemporalMaxer	I3D	60.72	52.83	40.12	23.53	58.71	56.51	52.62	51.71	44.83	36.28	57.79	52.40	42.97	56.01	49.17	41.88
ActionFormer	I3D	61.53	54.89	44.53	29.43	58.96	57.39	55.13	54.14	47.84	41.01	58.89	54.09	46.36	57.99	52.07	46.43
ActionFormer	VideoMAEv2	73.84	62.83	41.19	22.34	68.01	61.91	56.43	67.90	62.09	58.46	71.13	68.88	66.30	64.57	54.56	48.30
AFSD	I3D	46.05	38.69	28.70	22.12	43.28	39.37	36.79	39.02	29.57	23.40	41.65	34.15	27.45	42.37	37.49	33.00
TriDet	I3D	61.33	55.61	46.74	33.08	59.94	57.95	55.90	54.63	48.94	43.35	59.30	54.20	47.65	58.56	52.90	46.88
TriDet	VideoMAEv2	75.16	64.39	42.42	23.60	68.99	64.26	57.39	69.83	64.04	60.24	72.53	70.72	67.81	69.32	62.92	58.07

Table D. The performance of TAD models concerning corruption robustness on ActivityNet-v1.3-C, considering five distinct types of corruptions and three different levels, measured by the average mAP of the tIoU thresholds between 0.5 and 0.95 with the step of 0.05.

Model	Feature	Clean Frame	Corruption Type														
			Black Frame			Packet Loss			Overexposure			Motion Blur			Occlusion		
			Level 1	Level 2	Level 3	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
VSGN	I3D	31.85	29.72	28.04	24.44	31.73	31.08	30.69	31.60	30.13	28.35	31.72	30.87	29.82	31.72	31.04	30.19
TriDet	TSP	36.66	22.23	10.69	6.66	20.58	13.91	12.82	20.70	13.21	11.79	20.67	13.94	12.77	20.59	14.16	13.04
ActionFormer	TSP	36.50	30.59	20.34	8.44	30.17	29.92	29.84	29.72	29.12	28.57	30.19	29.83	29.64	30.23	30.12	30.06
ActionFormer	VideoMAEv2	38.47	37.01	14.19	7.37	38.46	38.26	38.08	38.28	37.37	36.41	38.36	37.44	36.32	38.13	37.04	36.16
AFSD	I3D	32.49	30.19	24.73	19.36	32.13	31.22	30.43	31.28	29.07	27.69	32.08	30.74	29.50	32.21	31.52	31.20

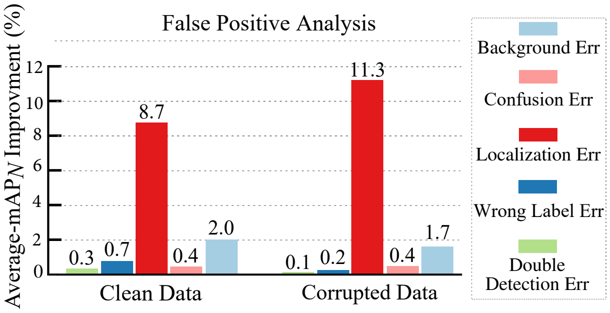

Figure A. False positive profiling of the TriDet's predictions on ActivityNet-v1.3-C. The Wrong Label (classification) Error remains relatively consistent, whereas the Localization Error increases significantly on corrupted data, revealing that vulnerability mainly comes from localization error rather than classification error.

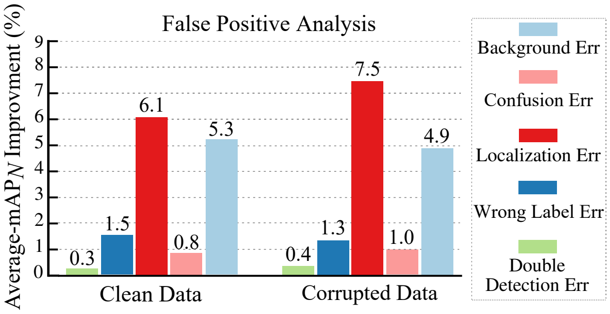

Figure B. False positive profiling of the ActionFormer's predictions on THUMOS14-C. The Localization Error of corrupted data is notably higher than that with clean data.

the extent of notably hindering human ability to locate actions—we discover through our experiments that even such subtle corruptions can substantially impair the localization capability of Temporal Action Detection (TAD) models. This finding indicates that the type of corruption we have introduced serves as an effective means to evaluate the temporal robustness of TAD models.

D. Performance of TAD models udner Each Type/Level of Temporal Corruptions

We present the detailed performance of different TAD models under the temporal corruptions of different types and levels in Tables C and D. It reads that existing TAD models are particularly vulnerable to temporal corruptions.

E. Comparisons on different black-frame locations for training.

As discussed in Section 5.1, our proposed FrameDrop Strategy generates black frames in action-background pairs. We also attempted to generate black frames solely in actions for model training. Employing these two methods on the THUMOS14-C dataset to train TemporalMaxer model, we obtained the results shown in Table E. It is evident that while the method of generating black frames solely in actions enhances the model's robustness, its performance on clean data diminishes. The reason may be that the model learns a bias—memorizing that corruptions is expected to occur within the action. Thus, we opt for corrupting the action-background pair in the FrameDrop Strategy.

Table E. Comparisons on different black-frame locations. It is clear that while the approach of exclusively generating black frames within actions improves the model's robustness, it leads to a decrease in performance on clean data.

THUMOS14-C	Train Test	Clean	Action	Action-background
TemporalMaxer [68]	Clean	60.72	59.28(**1.44** ↓)	**61.04** (**0.32** ↑)
	Corrupted	47.82	53.64(5.82 ↑)	51.95(4.13 ↑)

F. Multiple types of corruptions.

Building upon the addition of the five types of corruption mentioned in the main text, we also conducted two experiments. In these experiments, we randomly selected two corruption types and then: 1) applied both types to all middle frames (**spatial**) or 2) divided the middle frames temporally and applied one type to each half (**temporal**). We compared the results of TriDet and TemporalMaxer models on the THUMOS14-C dataset without adding corruptions (**clean**), adding only one type of corruptions (**ours**), and adding multiple types of corruptions simultaneously, as shown in Table F. The experimental results demonstrate that the simultaneous presence of multiple corruptions often degrades model robustness more significantly than the presence of only one type of corruptions.

Table F. Compared the performance of the TriDet and TemporalMaxer models on the THUMOS14-C dataset.

Corruption (# types)	Clean (0)	Ours (1)	Spatial (2)	Temporal (2)
TriDet [60]	61.33	51.71	43.31	50.85
TemporalMaxer [68]	60.72	47.82	40.84	49.43

G. More types of corruption.

In addition to the five types of corruptions mentioned in the main text, we also experimented with the effects of four additional types of corruptions, including:

- **Jittering**: caused by camera shake during filming
- **Different frame rate**: resulting from bandwidth limitations during network transmission
- **Slow-motion**: common shot types in videos
- **Time-lapse**: arising from video buffering issues and limitations in the processing power of playback devices

We tested TemporalMaxer model on the THUMOS14-C dataset, and the experimental results are shown in Table G. It can be seen that these four additional types of corruptions also significantly degrade the model's performance. This indicates that the degradation of model performance due to corruptions is independent of the corruptions type, suggesting that any corruptions in real-world scenarios could potentially lead to a decrease in model performance.

Table G. The average mAP of TemporalMaxer model with the addition of four types of corruptions at each corruptions level on the THUMOS14-C dataset. The experimental results indicate that these four types of corruptions also lead to a decrease in model performance.

Corruption	Clean	Jittering	Frame rate	Slow-motion	Time-lapse
TemporalMaxer [68]	60.72	50.16	40.05	56.03	45.25

H. More datasets for general evaluation.

In addition to conducting experiments on the commonly used THUMOS14 and ActivityNet-v1.3 datasets, we also attempted to construct MultiThumos-C using the same method as a benchmark for testing model robustness on the MultiThumos dataset. We tested this benchmark using the TemporalMaxer model and obtained the average mAP under five types of corruptions, as shown in Table H. The results indicate that adding corruptions to this dataset also significantly degrades model performance.

Table H. The average mAP of TemporalMaxer on the MultiThumos-C dataset for each type of corruptions. The results indicate that adding corruptions to the MultiThumos dataset also significantly degrades model performance.

tIoU	0.2	0.5	0.7	Average
Clean	44.30	30.54	15.72	27.55
Corrupted	41.24 (**3.06** ↓)	26.67 (**3.87** ↓)	12.65 (**3.07** ↓)	24.70 (**2.85** ↓)

References

[1] Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. Diagnosing error in temporal action detectors. In *Proceedings of the European conference on computer vision (ECCV)*, pages 256–272, 2018. 1

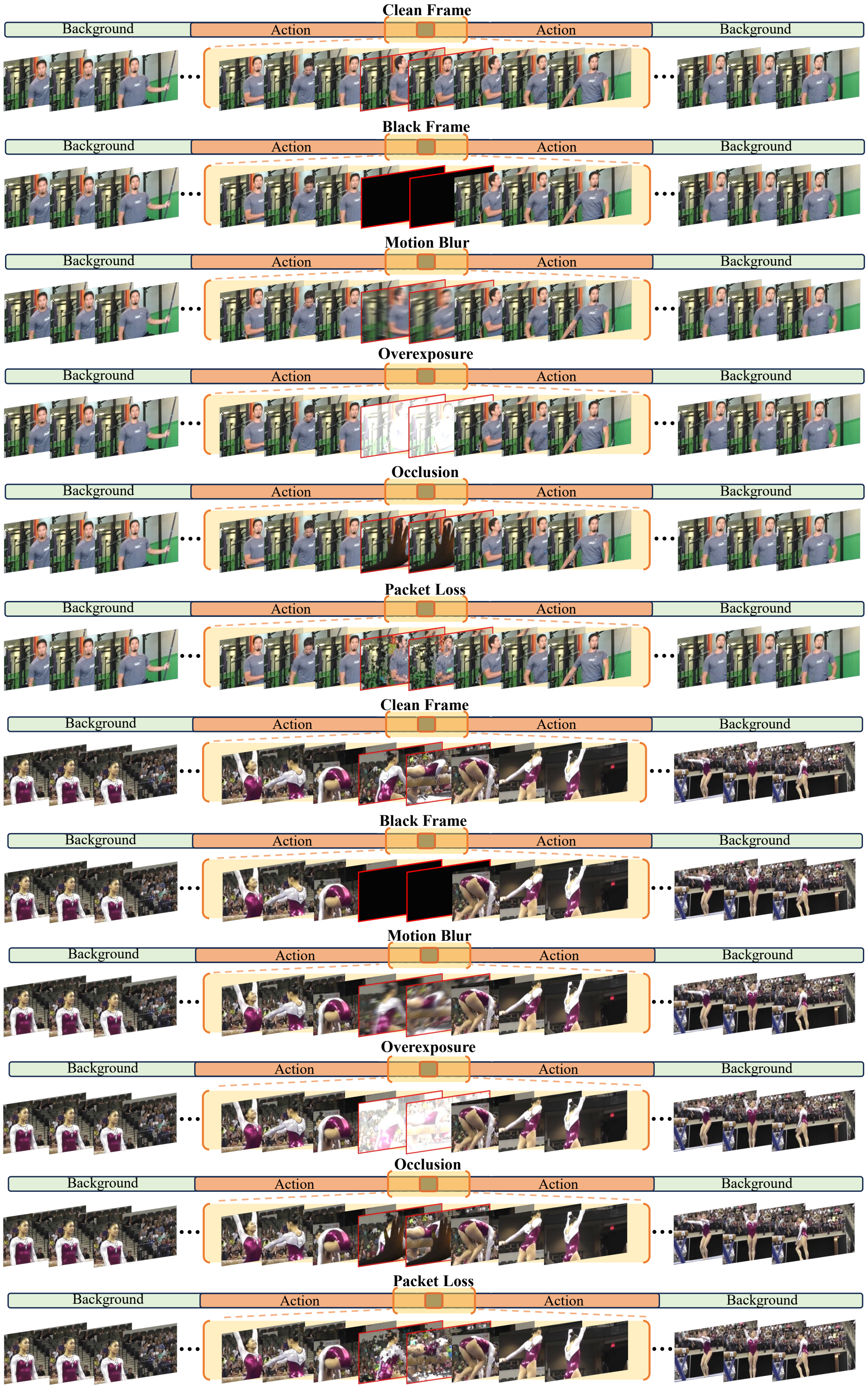

Figure C. More examples of our temporal corruptions dataset.