Benchmarking the Robustness of Temporal Action Detection Models Against Temporal Corruptions

Runhao Zeng[1,2], Xiaoyong Chen[3], Jiaming Liang[3], Huisi Wu[3], Guangzhong Cao[3*], Yong Guo[4*]
[1]Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, China,
[2]Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing,
[3]Shenzhen University [4]South China University of Technology

Abstract

Temporal action detection (TAD) aims to locate action positions and recognize action categories in long-term untrimmed videos. Although many methods have achieved promising results, their robustness has not been thoroughly studied. In practice, we observe that temporal information in videos can be occasionally corrupted, such as missing or blurred frames. Interestingly, existing methods often incur a significant performance drop even if only one frame is affected. To formally evaluate the robustness, we establish two temporal corruption robustness benchmarks, namely THUMOS14-C and ActivityNet-v1.3-C. In this paper, we extensively analyze the robustness of seven leading TAD methods and obtain some interesting findings: 1) Existing methods are particularly vulnerable to temporal corruptions, and end-to-end methods are often more susceptible than those with a pre-trained feature extractor; 2) Vulnerability mainly comes from localization error rather than classification error; 3) When corruptions occur in the middle of an action instance, TAD models tend to yield the largest performance drop. Besides building a benchmark, we further develop a simple but effective robust training method to defend against temporal corruptions, through the Frame-Drop augmentation and Temporal-Robust Consistency loss. Remarkably, our approach not only improves robustness but also yields promising improvements on clean data. We believe that this study will serve as a benchmark for future research in robust video analysis. Source code and models are available at https://github.com/Alvin-Zeng/temporal-robustness-benchmark.

1. Introduction

Temporal action detection (TAD), an essential aspect of video understanding, seeks to pinpoint action locations and identify action categories in untrimmed videos. Despite the fruitful progress in this mission, the robustness of these methods against corruptions remains largely unexplored. If TAD models are very vulnerable to corruptions, it would become particularly problematic when applying them in various practical contexts, including autonomous driving, security monitoring, and robotics. To verify this, we conduct a preliminary experiment in which we introduce corruptions to a single frame within an action instance, simulating the phenomenon of "Black Frame" [17] that can occur during data transmission. Remarkably, as shown in Figure 1, the performance of existing TAD methods drops significantly, no matter what kind of features are used or whether the model is trained end-to-end. This outcome reveals that corrupting even a single frame in an action instance disrupts the temporal continuity of the video and damages the temporal information. The poor performance of existing models un-

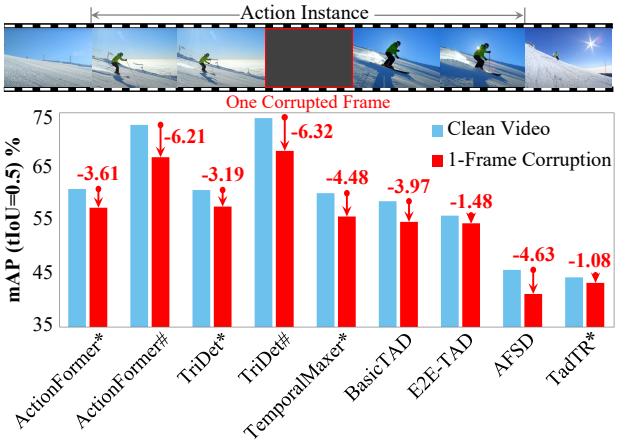

Figure 1. The mAP gap of temporal action detection methods when testing on clean and corrupted videos. * and # denote the video features extracted by I3D and VideoMAEv2, respectively. Other methods follow an end-to-end manner. Existing TAD methods incur a significant mAP drop of more than 1.08% even when **only one frame is corrupted** in an action instance on THUMOS14 dataset, highlighting a prevailing lack of robustness towards temporal corruptions.

*Corresponding author

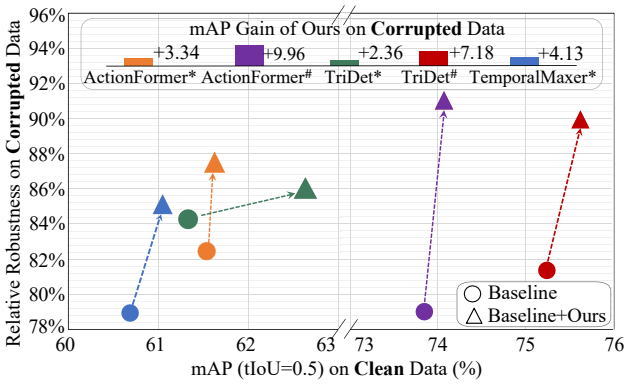

Figure 2. The gain of mAP and relative robustness brought by our proposed training strategy. * and # denote the video features extracted by I3D and VideoMAEv2, respectively. Our method enhances TAD models' robustness on corrupted videos and, surprisingly, boosts their performance on clean videos.

der these conditions suggests that TAD models generally exhibit weak ***temporal robustness***. As such, a comprehensive evaluation of temporal robustness becomes a cornerstone in advancing this field.

Robustness has been an active research topic yet the majority of studies have concentrated on images [27, 54]. Recently, in the video domain, [57, 79] present robustness benchmarks to evaluate action recognition models. By contrast, TAD not only requires action recognition; its key distinction lies in the need for temporal localization. Since these benchmarks do not take the change of temporal continuity into account, they are not directly applicable to assess TAD models. Thus, the design of an effective benchmark capable of evaluating temporal robustness remains an unexplored area. To this end, we propose two benchmark datasets, THUMOS14-C and ActivityNet-v1.3-C, that contain corruptions in the temporal domain. Specifically, we introduce 5 types of corruptions that are commonly observed in video acquisition and processing. To measure the severity of breaking the temporal continuity, we consider 3 levels by varying the number of frames to be corrupted in a video clip. We conduct in-depth experiments to analyze the robustness based on diverse leading TAD methods and obtain several interesting observations: 1) Existing TAD models demonstrate a notable vulnerability to temporal corruptions. Additionally, it has been observed that end-to-end TAD models are more susceptible to temporal corruptions compared to models that employ a fixed feature extractor. 2) The primary source of this vulnerability can be attributed to localization errors, as opposed to classification errors. 3) The vulnerability of TAD methods is most pronounced when corruption occurs at the center of an action instance. We believe these observations may suggest a potential avenue for future research towards robust TAD models.

Furthermore, we also develop a simple but effective method to improve the temporal robustness of TAD models. First, we propose a **FrameDrop** augmentation strategy, which randomly selects frames from adjacent actions and backgrounds of a video and introduces corruptions to break the temporal continuity. We highlight that training with such augmented data enables the model to locate and recognize actions against temporal corruptions. Second, we develop a **Temporal-Robust Consistency (TRC)** loss, which aligns the model's predictions on corrupted videos with those on clean videos. To increase the efficiency of this alignment, we propose an action-centric sampling strategy, selecting high-quality predictions that are temporally more relevant to the action instance for alignment. Interestingly, our experiments reveal that our robust training method not only increases robustness but also improves performance on clean data (see Figure 2). This study provides essential considerations for future model development.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to provide a comprehensive robustness analysis of temporal action detection (TAD) models. We believe that our new observations could be beneficial to developing robust TAD models for real-world deployment.
- We build two benchmark datasets and each involves 5 types of corruptions and 3 severities, resulting in 15 corruption types in total. We show that existing TAD methods are very vulnerable and often incur a significant performance drop on our benchmarks. Besides the performance on clean data, we highly recommend that researchers additionally evaluate their models in terms of temporal robustness in future research.
- We propose a simple but effective training method to improve temporal robustness. Interestingly, our method not only improves the robustness based on a diverse set of popular TAD models on corrupted videos but also obtains better performance on clean data in most cases.

2. Related Work

2.1. Temporal Action Detection

Temporal action detection approaches can be grouped into two categories: **Two-stage methods** primarily involves generating a set of proposals followed by their classification and boundary refinement [15, 62, 73, 84, 88]. To generate proposals, one can perform frame or segment-level classification and merge frames or segments of the same category [47, 49, 53, 63], while other methods use proposal generation methods [12, 40]. However, these methods heavily depend on the quality of the proposals, leading to the development of integrated approaches that combine proposal generation with classification and/or boundary regression [11, 31, 39, 68, 78], referred to **One-stage methods**. Notable contributions include the introduction of the anchor

mechanism for TAD by [39] and the exploration of anchor-free schemes [38, 60], further advanced the field by merging the advantages of both anchor-based and anchor-free methods [76]. Recently, transformer-based models, which have shown remarkable success in various vision tasks, have been adapted to TAD [85]. Other advances like Graph Convolution are also introduced to this task [75, 83] and end-to-end architectures have been explored in [41, 42, 77]. Despite their success, they focused on training and testing on a benchmark dataset with little distribution shift from training to testing samples, which poses challenges for real-world applications. This paper aims to investigate the robustness of TAD models and enhance their performance, particularly under conditions of corruptions.

2.2. Robustness of Neural Networks

Image Domain. Despite the outstanding performance of deep neural networks, they are not robust to image corruptions [29]. To address this, recent work explores recalibrating batch normalization statistics [8, 48, 58] or utilizing the frequency domain [56] to improve corruption robustness. However, data augmentation methods such as [18, 20, 28, 29, 55] represent the most prominent and successful line of work, ranging from simple Gaussian noise augmentation [55], over well-known schemes such as AutoAugment [18] to strategies specifically targeted towards corruption robustness such as AugMix [28] or DeepAugment [29]. Besides random corruptions, deep networks are susceptible to adversarial examples [21, 65]. While plenty of approaches for defending against adversarial examples have been proposed [1, 5, 10, 16, 22, 51, 64, 74, 81], adversarial training (AT) has become the de facto standard [44]. On the other hand, many efforts have been made to design or train a robust architecture to improve model robustness. Due to the success of Vision Transformers (ViTs) [19, 30, 70], many works seek to study and improve the robustness of ViTs [4, 7, 9, 24–26, 46, 52, 61, 89]. Interestingly, ViTs are often more robust than convolutional networks against corruptions [54, 66, 69] and adversarial attacks [6, 23, 37, 43, 45, 50, 59, 67]. Nevertheless, most of them mainly focus on the robustness issue in the image domain, leaving the robustness in the temporal domain (*e.g.*, inside videos) unexplored.

Video Domain. Data augmentation techniques have been shown effective in improving the robustness of video analysis models. Li *et al.* [36] utilized temporal cropping as video data augmentation. Mixup, cutmix, and cutout operations from the image domain were introduced to the video domain [34, 35, 82]. Isobe *et al.* [32] proposed the application of the same transformation across all frames in each mini-batch of video clips. In another line of research, Zhang *et al.* [86] used a Generative Adversarial Network (GAN) to create dynamic images that encapsulate motion information from video. Wu *et al.* [72] developed a generator to produce a frame encompassing all motion feature information. The robustness of video models against common corruptions has recently been analyzed [57, 79]. They benchmark the robustness of common convolutional- and transformer-based spatio-temporal architectures, against several corruptions in video acquisition and video processing. The corruptions they used can be generated across a set of continuous frames or depend solely on the content of a single frame. Their benchmarks apply corruptions to all frames of a trimmed video, aiming to measure the robustness of action recognition models. In this paper, we are particularly interested in examining the temporal detection robustness of TAD models and creating benchmarks by applying corruptions to a subset of the video frames to disrupt the temporal continuity.

3. Temporal Robustness Benchmark Creation

3.1. TAD Formulation and Notation

Temporal action detection (TAD) requires a machine to recognize the action instances and simultaneously identify their temporal positions in a video. Given an untrimmed video as $V = \{I_t\}_{t=1}^{T}$, where I_t denotes the frame at the time slot t, TAD predicts a set of action instances $\Phi_V = \{\phi_i = (t_{s_i}, t_{e_i}, k_i)\}_{i=1}^{N}$, where N is the number of action instances in V, t_{s_i}, t_{e_i} and k_i are the starting time, ending time, and category of the i-th action instance, respectively.

3.2. Temporal Corruptions in Videos

Our study begins with a scenario commonly encountered in daily life. We observed that during video playback, certain frames may suddenly experience interference and immediately disappear. For instance, an object might abruptly enter and then exit the frame during recording, or sudden changes in lighting might cause overexposure, which is then corrected by the camera. For humans, such interferences have minimal impact on our ability to locate actions in a video. However, our preliminary experiments (see Figure 1) demonstrate that even a single corrupted frame in an action sequence can significantly impair the temporal localization performance of TAD models.

Consequently, our research focuses on temporal corruptions that appear abruptly and vanish just as quickly during video recording—a common phenomenon in videos but one not previously addressed in research. Formally, given a video $V = \{I_1, I_2, \ldots, I_t, \ldots, I_T\}$, our approach to corruptions does not encompass all frames. Instead, we replace specific clean frames with corrupted ones, resulting in a corrupted video $V^c = \{I_1, I_2, \ldots, I_t^c, \ldots, I_T\}$ to disrupt the temporal continuity, where I_t^c represents the corrupted version of I_t. We study five categories of real-world corruptions, as depicted in Figure 3, including:

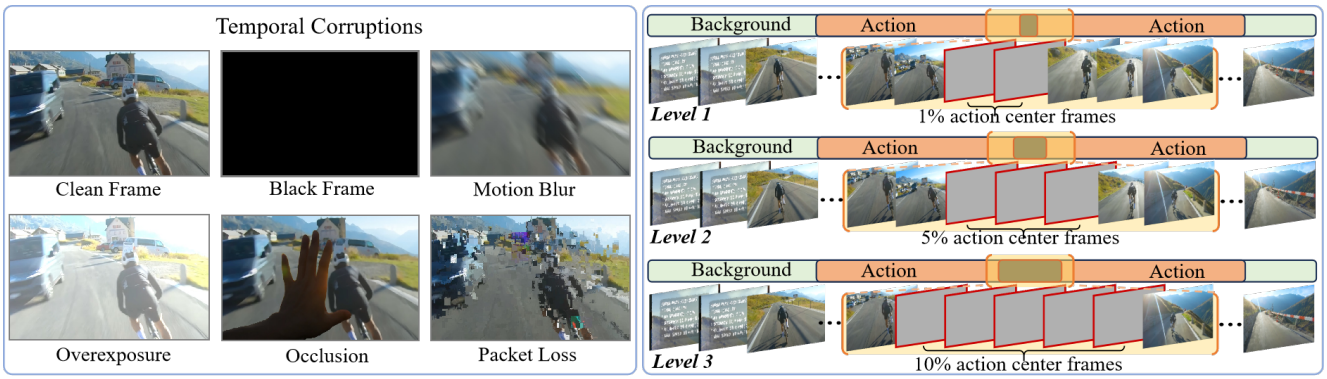

Figure 3. Our temporal robustness study introduces 5 types of temporal corruptions that are frequently encountered in real-world scenarios, including black frame [17], motion blur, overexposure, occlusion and packet loss [79]. Each type of corruptions has 3 levels of severity and each level refers to the $l\%$ ($l \in \{1, 5, 10\}$) action center frames being corrupted, eventually resulting in 15 distinct corruptions.

- **Black frame** [17]: caused by transferring tape-based content to digital files or temporary network disconnections during video streaming
- **Motion blur** [27]: occurs when the camera undergoes swift and rapid movements
- **Overexposure** [27]: due to fluctuations in daylight intensity, or sudden changes in photographic conditions
- **Occlusion**: resulting from the camera being accidentally blocked by another object while filming
- **Packet loss** [79]: arising from video transmission over imperfect channels in real-world settings

Discussion: Existing video robustness benchmarks apply corruptions across all frames, such an approach does not effectively validate temporal localization performance. With TAD models requiring both localization and recognition, it is challenging to ascertain whether issues arise in localization or recognition. By defining temporal corruptions where the majority of frames in a video remain clean, the impact on recognition is minimized. However, the few corrupted frames we introduce, although limited in number, directly disrupt the temporal continuity. Experiments have validated that our proposed corruption method effectively tests the localization capabilities of TAD models (see Section 4.2).

3.3. Benchmark Datasets

To benchmark the robustness of the TAD models against the temporal corruptions, we create two benchmark datasets, including THUMOS14-C and ActivityNe-v1.3-C.

THUMSO14-C. As a standard benchmark for action detection, THUMOS14 [33] contains a training set, known as the UCF-101 dataset, which consists of 13320 videos. The validation, testing, and background sets contain 1010, 1574, and 2500 untrimmed videos, respectively. The temporal action detection task of THUMOS14, which contains videos over 20 hours from 20 sports classes, is very challenging since each video has more than 15 action instances and its 71% frames are occupied by background items. In this study, we apply 5 distinct corruptions, each at 3 levels of severity, to the 213 annotated videos from the testing set. Specifically, we introduce corruptions to the central $l\%$ of frames in each action instance, where $l \in \{1, 5, 10\}$. Level 1 indicates a minimal temporal corruption that affects only 1% of the frames within a given action instance while Level 3 signifies a more substantial temporal corruption that affects 10% of the frames. We choose to corrupt the central frames since we empirically found that the robustness of TAD models degrades more significantly when the corrupted frame is located closer to the center of an action instance (see Section 4.3 for more results).

ActivityNet-v1.3-C. ActivityNet [13] is a popular benchmark for TAD on untrimmed videos. We create a benchmark on ActivityNet-v1.3, which contains approximately 10K training videos and 5K validation videos corresponding to 200 different activities. Each video has an average of 1.65 action instances. Similarly, we apply the proposed 5 corruptions with 3 severity levels to the validation set.

3.4. Robustness Metrics

We first introduce the standard mean Average Precision (mAP) that is widely used to evaluate TAD models. We further take the mAP on clean data into account and develop a new metric to measure how large the performance drop would be between the mAP on clean and corrupted data.

Mean Average Precision (mAP) is a commonly used evaluation metric for action detection performance. A predicted temporal bounding box is considered to be correct if its temporal IoU with the ground-truth instance is larger than a certain threshold and the predicted category is the same as this ground-truth instance. On THUMOS14-C, the tIOU thresholds are chosen from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ and we report mAP@tIoU=0.5 for comparisons; on ActivityNet-v1.3-C, the tIoU thresholds are from $\{0.5, 0.75, 0.95\}$, and we report the average mAP of the tIoU thresholds between 0.5 and 0.95 with the step of 0.05.

Table 1. Corruption robustness of TAD models on THUMOS14-C. * denotes end-to-end methods. Existing TAD models are particularly vulnerable to temporal corruptions, regardless of whether they are based on transformers or CNN.

Model	Feature	Clean mAP	Corrupted mAP	Relative Robustness
BasicTAD* [77]	SlowOnly	59.17	37.72 (21.45 ↓)	63.75
E2E-TAD* [41]	SlowFast	56.41	30.55 (25.86 ↓)	54.16
TemporalMaxer [68]	I3D	60.72	47.82 (12.90 ↓)	78.76
ActionFormer [85]	I3D	61.53	50.61 (10.92 ↓)	82.25
ActionFormer [85]	VideoMAEv2	73.84	58.33 (15.51 ↓)	78.99
AFSD* [38]	I3D	46.05	34.47 (11.58 ↓)	74.85
TriDet [60]	I3D	61.33	51.71 (9.62 ↓)	84.31
TriDet [60]	VideoMAEv2	75.16	61.10 (14.06 ↓)	81.29
TriDet [60]+Ours	VideoMAEv2	**75.60**	**68.28 (7.32↓)**	**90.31**

Relative robustness. We introduce a new metric, termed as relative robustness γ^r to measure the robustness of TAD models. We first calculate the mAP M_{clean} on the clean test set given a trained model g. Then, we test g on a corruption c at each of the severity levels s, and obtain mAP $M_{c,s}$. It should be noted that different models exhibit diverse performances on identical test videos, thus, an absolute drop in performance is also influenced by the model's performance on clean videos. Therefore, we determine relative performance drop as a measure of the model's robustness. Each severity level s and corruption c has its own relative robustness $\gamma^r_{c,s}$ computed as $\gamma^r_{c,s} = 1 - (M_{clean} - M_{c,s})/M_{clean}$. We average across all severity levels and corruptions to yield to yield γ^r of a TAD model.

4. Benchmarking Robustness of TAD Models

In our benchmark study, we train the TAD models with clean data and evaluate them on the corrupted data. It is a standard setting under the robust generalization study [80], which assumes that the model is unable to know the exact problem in the deployment in advance.

Model Variants. Our experiments evaluate seven popular TAD models, employing various architectural frameworks such as CNN, Transformer, and Graph Convolution. Regarding the detection heads, ActionFormer [85] and E2E-TAD [41] employs a Transformer architecture, while TriDet [60], AFSD [38], BasicTAD [77], and TemporalMaxer [68] are CNN-based models. VSGN [87] is constructed based on graph architectures. Note that E2E-TAD, BasicTAD and AFSD are trained in an end-to-end manner while others rely on pre-trained feature extractors. We exploit three different feature extractors, including I3D [14], VideoMAEv2 [71], and TSP [3]. Among these, the I3D model leverages 3D convolutions, whereas VideoMAEv2 adopts a Transformer-based approach, trained using a dual masking strategy. TSP, on the other hand, utilizes a ResNet-based backbone pre-trained on temporal sensitive tasks. In our experimental setup, we use the official implementations and pre-trained weights of these models.

Table 2. Corruption robustness of TAD models on ActivityNet-v1.3-C. * denotes end-to-end methods. TAD models remains highly susceptible to temporal corruptions, suggesting that the vulnerability is not specific to any particular dataset.

Model	Feature	Clean mAP	Corrupted mAP	Relative Robustness
VSGN [87]	I3D	31.85	30.08 (1.77 ↓)	94.44
TriDet [60]	TSP	36.66	15.18 (21.48 ↓)	41.41
ActionFormer [85]	TSP	36.50	27.79 (8.71 ↓)	76.12
ActionFormer [85]	VideoMAEv2	38.47	33.93 (4.54 ↓)	88.19
AFSD* [38]	I3D	32.49	29.56 (2.93 ↓)	90.98
AFSD* [38]+Ours	I3D	**32.86**	**30.78 (2.08 ↓)**	**93.68**

4.1. Existing TAD Models are Particularly Vulnerable to Temporal Corruptions

The robustness against temporal corruptions of TAD models on the THUMOS14-C dataset is presented in Table 1. All TAD models assessed in this study show susceptibility to temporal corruptions, evidenced by a reduction in detection mAP ranging from 9.62% to 25.86%. This vulnerability is observed regardless of the type of features employed or whether the model is end-to-end trained. When considering the relative robustness calculated as an average across five types of corruptions and three levels, the highest-performing model, Tridet, achieves only 84.31%, while the lowest, E2E-TAD, scores 54.16%. This indicates that the robustness of TAD models is influenced by both the model architecture and the input data characteristics. Furthermore, we observe that methods employing end-to-end training, such as BasicTAD, E2E-TAD, and AFSD, are more susceptible to temporal corruptions compared to those utilizing a fixed feature extractor approach (when using the same backbone). We also compare the temporal robustness of existing TAD models on ActivityNet-v1.3-C and report the results in Table 2. On this dataset, the performance of TAD models remains highly susceptible to temporal corruptions. The range of decrease in terms of mAP spans from 1.77% to 21.48%. This suggests that the limited temporal robustness of TAD models is not specific to any particular dataset. Please refer to the supplementary material for detailed performance on each type of corruption at every level.

4.2. Vulnerability Mainly Comes from Localization Error rather than Classification Error

We follow DETAD [2] to analyze the results predicted by the TriDet model. We categorize the causes leading to false positives predicted by the model into five types, including 1) Background Error: predict action as background, 2) Confusion Error: low-quality boundary with wrong category, 3) Localization Error: low-quality boundary with correct category, 4) Wrong Label Error: high-quality boundary with correct category, 5) Double Detection Error: two predictions match one action. Figure 4 illustrates the enhancement in the performance of the model when a certain type

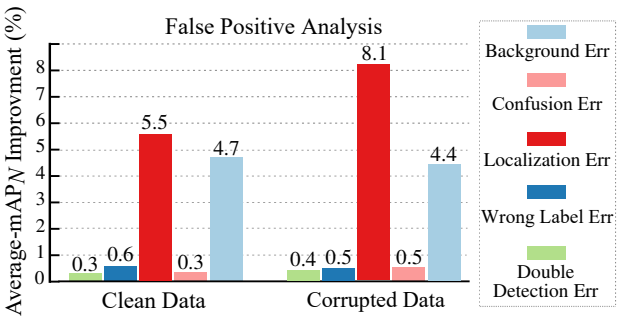

Figure 4. False positive profiling of the TriDet's predictions on THUMOS14-C. The Wrong Label (classification) Error remains relatively consistent, whereas the Localization Error increases significantly on corrupted data, revealing that vulnerability mainly comes from localization error rather than classification error.

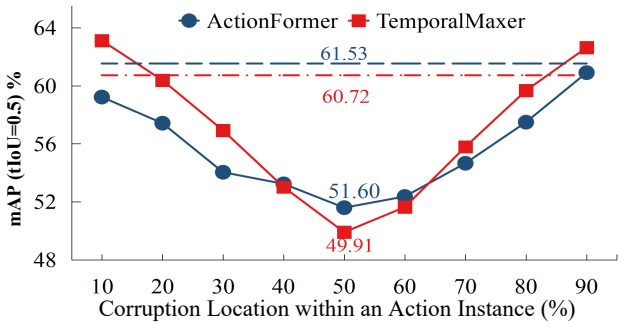

Figure 5. The performance of TAD models with varying corruption locations within an action instance on THUMOS14-C. The horizontal dashed lines refer to the model's performance on clean videos. As corruptions approach the center, its impact on the model becomes increasingly significant.

of error is eliminated. It is evident that the impact of Localization Error escalates from 5.5% to 8.1% upon the corrupted video, while the Wrong Label Error (0.6% v.s. 0.5%) remains relatively stable, indicating a minor influence on classification. This suggests that our proposed benchmark primarily assesses the robustness of the model's temporal localization in the face of temporal corruptions. Compared to other benchmarks designed for action recognition robustness [57, 79], ours is more suitable to examine the temporal robustness of TAD models. Please kindly refer to the supplementary material for more analysis of other TAD models.

4.3. Corrupting the Central Frames of Action Results in the Strongest Attack

To investigate the impact of corruption's location within action instances, we evaluate the robustness of ActionFormer and TemporalMaxer on the THUMOS14-C dataset. We sample five continuous frames at every 10%, 20%, ..., and 90% of each action instance and introduce black frame corruption. From Figure 5, when corruption is centered within an action instance, both models exhibit the most pronounced performance degradation. Interestingly, we also discover that the models' performance improved when corruptions are introduced near the boundaries of the action, surpassing the performance of clean data. We posit that this is due to the models interpreting the position of the black frame as the boundary of the action, thus improving the localization performance. Consequently, we opt to replace the frames at the center of each action instance with corrupted frames, in order to construct temporal corrupted datasets that maximally disrupt model performance.

5. Defending against Temporal Corruptions

With the above observations, we seek to improve existing TAD models' temporal robustness from two perspectives. First, in Section 5.1, we propose a **FrameDrop strategy** to simulate temporal corruptions. Our intuition is that during model training, introducing corruptions to the input video forces the model to better leverage the uncorrupted temporal context for localizing action and identifying action categories. Second, in Section 5.2, we develop a new **Temporal-Robust Consistency (TRC) loss** to improve the localization capability by guiding the model to predict temporal bounding boxes that are temporally related to actions.

5.1. FrameDrop Strategy

We propose to randomly drop frames from the input video to corrupt the temporal continuity. In particular, we divide the input video into multiple Action-Background (AB) pairs (*i.e.*, $V = \{P_i^{AB}\}_{i=1}^{N_p}$), where each pair is composed of adjacent action and background segments and N_p is the number of the resultant pairs. Then, within each AB pair $P^{AB} = \{ I_1^a, I_2^a, \ldots, I_{N_a}^a, I_1^b, I_2^b, \ldots, I_{N_b}^b\}$, where N_a and N_b refer to the number of frames in the action and background segment, we randomly select a single frame to drop (*i.e.*, replace with a black frame). The video, once subjected to FrameDrop, is then forwarded to a specific feature extractor (*e.g.*, I3D, VideoMAEv2) or directly fed into the TAD model (*e.g.*, the end-to-end method BasicTAD).

It is noteworthy that our method performs FrameDrop operations in both action and background segments. This prevents the model from memorizing that corruptions will occur in the action segments, thus circumventing a trivial solution. Further experiments are detailed in the supplementary materials. We empirically show that our FrameDrop is able to consistently improve the robustness of various TAD models against different corruptions that are unseen during training (see Section 6.2 for more results).

5.2. Temporal-Robust Consistency Loss

When both corruppted videos and clean videos processed through FrameDrop are fed into the TAD model, two sets of temporal bounding box $\hat{\Phi}^c = \{\phi_i^c = (t_{s_i}^c, t_{e_i}^c, k_i^c)\}_{i=1}^{N_c}$

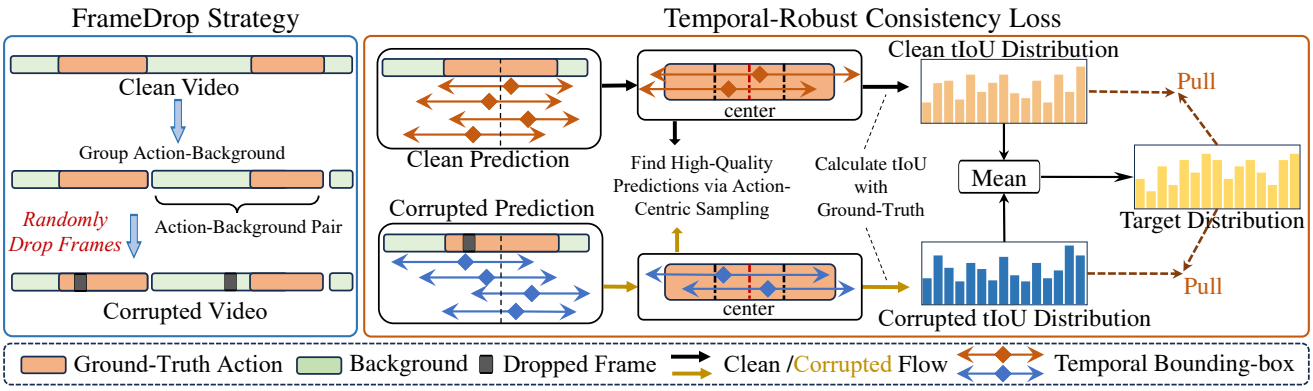

Figure 6. Our approach towards training a robust TAD model. We propose a FrameDrop strategy to interrupt the temporal continuity by first grouping the adjacent action and background instance into pairs and randomly replacing a clean frame within each pair with a black frame. Then, we perform action-centric sampling to find high-quality predictions from clean and corrupted videos and propose a temporal-robust consistency loss to align them in terms of their tIoU distributions w.r.t ground-truth actions.

and $\hat{\Phi}^d = \{\phi_i^d = (t_{s_i}^d, t_{e_i}^d, k_i^d)\}_{i=1}^{N_d}$ are predicted, where N_c and N_d are the number of predictions. A simple training approach would be to ensure consistency between these two prediction sets. However, this strategy presents two issues. The first is computational efficiency, as a single video can contain numerous predictions. More critically, the TAD method's primary focus is the localization of actions, hence our attention should be primarily focused on predictions temporally related to action instances. As a result, we propose an action-centric sampling strategy for selecting action-related predictions.

Without loss of generality, let a ground-truth action instance be $\phi = (t_s, t_e, k)$, where we omit i for simplicity. We compute the timestamp of its central frame by $t^* = \frac{t_e + t_s}{2}$ as a reference. Similarly, for the predicted temporal bounding boxes in $\hat{\Phi}^c$ and $\hat{\Phi}^d$, we could also calculate the timestamp of their central frame and obtain two central location sets $T^c = \{t_j^c\}_{j=1}^{N_c}$ and $T^d = \{t_j^d\}_{j=1}^{N_d}$, respectively. Then, we choose the top-K predictions from T^c (or T^d) with the minimal distance w.r.t t^* and calculate the temporal Intersection over Union (tIoU) between the selected K predictions and the ground truth (GT) action instance ϕ. Thus, we obtain two tIoU distributions - one corresponding to the corrupted input $p_c \in \mathbb{R}^K$, and the other to the clean input $p_d \in \mathbb{R}^K$. To align the predictions under these two circumstances, we average p_c and p_d to serve as the target tIoU distribution p_t, and then separately calculate the Kullback-Leibler (KL) divergence of the predicted distribution w.r.t p_t. In this way, we derive the temporal-robust consistency loss

$$L_{TRC} = \frac{1}{2}(\mathrm{KL}[p_t||p_c] + \mathrm{KL}[p_t||p_d]). \quad (1)$$

Note that we could instead compute the KL divergence between p_c and p_d or adopt the Mean Squared Error (MSE) as the loss function, but these do not perform as well. More ablated results of the loss function can be found in the supplementary material.

Table 3. Results of defending against temporal corruptions with the help of our proposed training strategy on THUMOS14-C. Our method consistently improves the robustness of various TAD models with different features.

Backbone(feature)	Clean mAP	Corrupted mAP	Relative Robutness
TemporalMaxer(I3D)	60.72	47.82	78.76
+ Ours	**61.04 (0.32 ↑)**	**51.95 (4.13 ↑)**	**85.10 (6.34 ↑)**
TriDet(I3D)	61.33	51.71	84.31
+ Ours	**62.63 (1.30 ↑)**	**54.07 (2.36 ↑)**	**86.32 (2.01 ↑)**
TriDet(VideoMAEv2)	75.16	61.10	81.29
+ Ours	**75.60 (0.44 ↑)**	**68.28 (7.18 ↑)**	**90.32 (9.03 ↑)**
ActionFormer(I3D)	61.53	50.61	82.25
+ Ours	**61.63 (0.10 ↑)**	**53.95 (3.34 ↑)**	**87.54 (5.29 ↑)**
ActionFormer(VideoMAEv2)	73.84	58.33	78.99
+ Ours	**74.06 (0.22 ↑)**	**68.29 (9.96 ↑)**	**92.21 (13.22 ↑)**

6. Improved Robustness and Further Analysis

6.1. Improved Robustness of TAD Models

In addressing the performance drop caused by temporal corruptions, we propose to enhance temporal robustness in training. Table 3 depicts the performance gain of different TAD models trained using our method and tested with varying corruptions on the THUMOS14-C dataset. The robustness of different models has markedly improved, even the model with the least improvement displays an absolute mAP rise of 2.36% (averaged on 3 levels of 5 corruptions). When using the videoMAEv2 feature, the increase in robustness of ActionFormer is substantial. Specifically, both the mAP and relative robustness have an increase of 9.96% and 13.22% respectively. This indicates that our method can enhance the temporal anti-interference ability of the detection head, allowing it to better pair with VideoMAE and thereby achieve satisfactory results on clean data whilst

Table 4. Results of defending against temporal corruptions with the help of our proposed training strategy on ActivityNet-v1.3-C. Despite the challenges of this large-scale dataset, our method still yields consistent robustness improvements.

Backbone(feature)	Clean mAP	Corrupted mAP	Relative Robustness
ActionFormer(TSP)	**36.50**	27.79	76.12
+ Ours	36.01 (0.49 ↓)	**28.41 (0.62 ↑)**	**78.91 (2.79 ↑)**
ActionFormer(VideoMAEv2)	**38.47**	33.93	88.19
+ Ours	38.44 (0.03 ↓)	**34.17 (0.24 ↑)**	**88.90 (0.71 ↑)**
TriDet(TSP)	**36.66**	15.18	41.42
+ Ours	36.42 (0.24 ↓)	**17.28 (2.10 ↑)**	**47.45 (6.03 ↑)**
AFSD(End-to-End)	32.49	29.56	90.98
+ Ours	**32.86 (0.37 ↑)**	**30.78 (1.22 ↑)**	**93.68 (2.70 ↑)**

Table 5. Ablation study on our proposed training strategy, measured by the performance of TriDet on THUMOS14-C. Using FrameDrop and TRC loss simultaneously yields improvements in robustness and even mAP gain on clean videos.

Feature	FrameDrop	TRC Loss	mAP (tIoU=0.5)		Relative
			Clean	Corrupted	Robustness
I3D			61.33	51.71	84.31
	✓		62.37	52.56	84.27
	✓	✓	**62.63**	**54.07**	**86.33**
VideoMAEv2			75.16	61.10	81.29
	✓		74.75	65.53	87.67
	✓	✓	**75.60**	**68.28**	**90.31**

maintaining commendable robustness. Table 4 displays the experimental results on the ActivityNet-v1.3-C dataset. It can be observed that, when trained with our method, the model is likewise capable of achieving consistent improvements in relative robustness across datasets with different levels of corruptions.

6.2. Further Investigation of Our Training Strategy

Effectiveness of our proposed training strategy. Our proposed training strategy comprises two components: the FrameDrop strategy and the Temporal-Robust Consistency (TRC) loss. We conduct experiments by gradually adding them to the baseline. According to Table 5, using only the FrameDrop strategy enhances the robustness of the TAD model across two types of features. However, when employing VideoMAEv2 features, the FrameDrop strategy alone does not improve the model's performance on clean videos. If both the FrameDrop strategy and TRC loss are utilized concurrently, not only is there a significant improvement in robustness, but the action detection performance on clean videos is also enhanced.

Generality of our proposed training strategy. The length of the corruptions is fixed at one frame in our proposed training method. Our experiments show improvements on both datasets with the corruptions of varying lengths, suggesting that our method is applicable to corruptions of differing lengths. We also conduct experiments by training TAD models under one corruption while testing them on distinct corruptions. Table 6 reveals that even if the corruptions are unseen during the training, our method still significantly enhances the model's robustness. This not only demonstrates the generality of our approach, unconstrained by any specific corruption but also indicates that our method does not simply train the model to memorize corruptions. Rather, it enhances the model's temporal robustness via our unique mechanism of disrupting temporal continuity and aligning localization distributions during the training.

Table 6. Performance comparison using different corruptions in the training, measured by the mAP of TriDet on THUMOS14-C. Our method is general and not limited to specific corruptions, and it consistently improves robustness on unseen corruptions.

Test \ Train	Clean	Motion Blur	Black Frame
Black Frame	43.47	43.61 (0.14 ↑)	56.44 (12.97 ↑)
Packet Loss	63.54	64.90 (1.36 ↑)	70.36 (6.82 ↑)
Overexposure	64.70	64.77 (0.07 ↑)	70.01 (5.31 ↑)
Motion Blur	70.36	73.62 (3.26 ↑)	74.84 (4.48 ↑)
Occlusion	63.44	67.24 (3.80 ↑)	69.76 (6.32 ↑)
Average	61.10	**62.83 (1.73 ↑)**	**68.28 (7.18 ↑)**

7. Conclusion

In this study, we have introduced a temporal robustness benchmark (THUMOS14-C and ActivityNet-v1.3-C), specifically designed for evaluating temporal action detection (TAD) methods. Unlike other video robustness benchmarks that apply corruptions to all frames of a video, we have designed a temporal corruption approach by corrupting a subset of frames within the video to disrupt its temporal continuity. We have conducted a robustness analysis on seven leading TAD methods, encompassing one-stage, two-stage, CNN-based and Transformer-based architectures. Our evaluation revealed that current TAD models are notably susceptible to temporal corruptions, with this vulnerability largely stemming from localization errors, rather than classification errors. We also observed that when corruption occurs in the middle of an action instance, TAD models tend to yield the largest performance drop. These observations might suggest a promising direction for future research towards robust TAD. Furthermore, we have proposed a simple yet effective strategy for training temporally robust TAD models. This approach not only enhanced robustness but also improved performance on clean data. Given its universality and significance, robustness in TAD emerges as a new research dimension to be systematically explored in future studies.

Acknowledgements. This work was partially supported by National Natural Science Foundation of China (NSFC) under Grants 62202311, 62273241, the Guangdong Basic and Applied Basic Research Foundation under Grants 2023A1515011512, 2024A1515011946, Excellent Science and Technology Creative Talent Training Program of Shenzhen Municipality under Grant RCBS20221008093224017, the Shenzhen Natural Science Foundation (the Stable Support Plan Program) under Grant 20220809180405001.

References

[1] Naveed Akhtar and Ajmal S. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 3

[2] 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. 5

[3] Humam Alwassel, Silvio Giancola, and Bernard Ghanem. Tsp: Temporally-sensitive pretraining of video encoders for localization tasks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3173–3183, 2021. 5

[4] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 26831–26843, 2021. 3

[5] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, and J. D. Tygar. Can machine learning be secure? In *Proceedings of the ACM on Asia Conference on Computer and Communications Security (AsiaCCS)*, pages 16–25, 2006. 3

[6] Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. Adversarial robustness comparison of vision transformer and mlp-mixer to cnns. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021. 3

[7] Philipp Benz, Chaoning Zhang, Soomin Ham, Adil Karjauv, and I Kweon. Robustness comparison of vision transformer and mlp-mixer to cnns. In *Proceedings of the CVPR 2021 Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges (AML-CV)*, 2021. 3

[8] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Revisiting batch normalization for improving corruption robustness. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 494–503, 2021. 3

[9] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10231–10241, 2021. 3

[10] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, pages 2154–2156, 2018. 3

[11] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017. 2

[12] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2911–2920, 2017. 2

[13] 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 (CVPR)*, pages 961–970, 2015. 4

[14] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017. 5

[15] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A. Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1130–1139, 2018. 2

[16] Ashutosh Chaubey, Nikhil Agrawal, Kavya Barnwal, Keerat Kaur Guliani, and Pramod Mehta. Universal adversarial perturbations: A survey. *arXiv preprint arXiv:2005.08087*, 2020. 3

[17] Sungwoo Choi, Moonsik Lee, Byunghee Jung, Kiok Ahn, Byungyong Ryu, Jaemyun Kim, and Oksam Chae. Automated content restoration system for file-based broadcasting environments. *SMPTE Motion Imaging Journal*, 124(8):39–46, 2015. 1, 4

[18] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 3

[20] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 3

[21] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 3

[22] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020. 3

[23] Jindong Gu, Volker Tresp, and Yao Qin. Evaluating model robustness to patch perturbations. In *ICML 2022 Shift Happens Workshop*, 2022. 3

[24] Yong Guo, David Stutz, and Bernt Schiele. Improving robustness of vision transformers by reducing sensitivity to patch corruptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4108–4118, 2023. 3

[25] Yong Guo, David Stutz, and Bernt Schiele. Robustifying token attention for vision transformers. In *Proceedings of the*

IEEE International Conference on Computer Vision (ICCV), pages 17557–17568, 2023.

[26] Xing Han, Tongzheng Ren, Tan Minh Nguyen, Khai Nguyen, Joydeep Ghosh, and Nhat Ho. Robustify transformers with robust kernel density estimation. *arXiv preprint arXiv:2210.05794*, 2022. 3

[27] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 2, 4

[28] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 3

[29] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8340–8349, 2021. 3

[30] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 11936–11945, 2021. 3

[31] Yupan Huang, Qi Dai, and Yutong Lu. Decoupling localization and classification in single shot temporal action detection. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1288–1293, 2019. 2

[32] Takashi Isobe, Jian Han, Fang Zhuz, Yali Liy, and Shengjin Wang. Intra-clip aggregation for video person reidentification. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 2336–2340, 2020. 3

[33] YG Jiang, J Liu, A Roshan Zamir, G Toderici, I Laptev, M Shah, and R Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2014. 4

[34] Taeoh Kim, Hyeongmin Lee, MyeongAh Cho, Ho Seong Lee, Dong Heon Cho, and Sangyoun Lee. Learning temporally invariant and localizable features via data augmentation for video recognition. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 386–403, 2020. 3

[35] Taeoh Kim, Jinhyung Kim, Minho Shim, Sangdoo Yun, Myunggu Kang, Dongyoon Wee, and Sangyoun Lee. Exploring temporally dynamic data augmentation for video recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 3

[36] Jie Li, Mingqiang Yang, Yupeng Liu, Yanyan Wang, Qinghe Zheng, and Deqiang Wang. Dynamic hand gesture recognition using multi-direction 3d convolutional neural networks. *Engineering Letters*, 27(3), 2019. 3

[37] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022. 3

[38] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3320–3329, 2021. 3, 5

[39] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 988–996, 2017. 2, 3

[40] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3889–3898, 2019. 2

[41] Xiaolong Liu, Song Bai, and Xiang Bai. An empirical study of end-to-end temporal action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20010–20019, 2022. 3, 5

[42] 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

[43] Giulio Lovisotto, Nicole Finnie, Mauricio Munoz, Chaithanya Kumar Mummadi, and Jan Hendrik Metzen. Give me your attention: Dot-product attention considered harmful for adversarial patch robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15234–15243, 2022. 3

[44] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 3

[45] Kaleel Mahmood, Rigel Mahmood, and Marten van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7838–7847, 2021. 3

[46] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12042–12051, 2022. 3

[47] Alberto Montes, Amaia Salvador, and Xavier Giro-i Nieto. Temporal activity detection in untrimmed videos with recurrent neural networks. *1st NIPS Workshop on Large Scale Computer Vision Systems (LSCVS)*, 2016. 2

[48] Zachary Nado, Shreyas Padhy, D. Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020. 3

[49] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Proposal-free temporal action detection via global segmentation mask learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 645–662, 2022. 2

[50] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 3

[51] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 3

[52] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pages 2071–2081, 2022. 3

[53] AJ Piergiovanni and Michael Ryoo. Temporal gaussian mixture layer for videos. In *Proceeding of the International Conference on Machine Learning (ICML)*, pages 5152–5161, 2019. 2

[54] Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. Understanding and improving robustness of vision transformers through patch-based negative augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 16276–16289, 2022. 2, 3

[55] E. Rusak, Lukas Schott, R. S. Zimmermann, Julian Bitterwolf, O. Bringmann, M. Bethge, and W. Brendel. A simple way to make neural networks robust against diverse image corruptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[56] Tonmoy Saikia, Cordelia Schmid, and Thomas Brox. Improving robustness against common corruptions with frequency biased models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10211–10220, 2021. 3

[57] Madeline Chantry Schiappa, Naman Biyani, Prudvi Kamtam, Shruti Vyas, Hamid Palangi, Vibhav Vineet, and Yogesh S Rawat. A large-scale robustness analysis of video action recognition models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14698–14708, 2023. 2, 3, 6

[58] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11539–11551, 2020. 3

[59] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021. 3

[60] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18857–18866, 2023. 3, 5

[61] Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. Robustness verification for transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 3

[62] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1049–1058, 2016. 2

[63] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-deconvolutional networks for precise temporal action localization in untrimmed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5734–5743, 2017. 2

[64] Samuel Henrique Silva and Peyman Najafirad. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*, 2020. 3

[65] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 3

[66] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan L. Yuille, Philip H. S. Torr, and Dacheng Tao. Robustart: Benchmarking robustness on architecture design and training techniques. *arXiv preprint arXiv:2109.05211*, 2021. 3

[67] Shiyu Tang, Siyuan Liang, Ruihao Gong, Aishan Liu, Xianglong Liu, and Dacheng Tao. Exploring the relationship between architecture and adversarially robust generalization. *arXiv preprint arXiv:2209.14105*, 2022. 3

[68] Tuan N Tang, Kwonyoung Kim, and Kwanghoon Sohn. Temporalmaxer: Maximize temporal context with only max pooling for temporal action localization. *arXiv preprint arXiv:2303.09055*, 2023. 2, 5

[69] Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, and Yugang Jiang. Deeper insights into vits robustness towards common corruptions. *arXiv preprint arXiv:2204.12143*, 2022. 3

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017. 3

[71] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023. 5

[72] Di Wu, Junjun Chen, Nabin Sharma, Shirui Pan, Guodong Long, and Michael Blumenstein. Adversarial action data augmentation for similar gesture action recognition. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. 3

[73] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5783–5792, 2017. 2

[74] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil K. Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020. 3

[75] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10156–10165, 2020. 3

[76] Le Yang, Houwen Peng, Dingwen Zhang, Jianlong Fu, and Junwei Han. Revisiting anchor mechanisms for temporal action localization. *IEEE Transactions on Image Processing*, 29:8535–8548, 2020. 3

[77] Min Yang, Guo Chen, Yin-Dong Zheng, Tong Lu, and Limin Wang. Basictad: an astounding rgb-only baseline for temporal action detection. *Computer Vision and Image Understanding*, 232:103692, 2023. 3, 5

[78] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2678–2687, 2016. 2

[79] Chenyu Yi, Siyuan Yang, Haoliang Li, Yap-peng Tan, and Alex Kot. Benchmarking the robustness of spatial-temporal models against corruptions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 2, 3, 4, 6

[80] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13255–13265, 2019. 5

[81] Xiaoyong Yuan, Pan He, Qile Zhu, Rajendra Rana Bhat, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019. 3

[82] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, and Jinhyung Kim. Videomix: Rethinking data augmentation for video classification. *arXiv preprint arXiv:2012.03457*, 2020. 3

[83] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7094–7103, 2019. 3

[84] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional module for temporal action localization in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6209–6223, 2021. 2

[85] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 492–510, 2022. 3, 5

[86] Yumeng Zhang, Gaoguo Jia, Li Chen, Mingrui Zhang, and Junhai Yong. Self-paced video data augmentation by generative adversarial networks with insufficient samples. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1652–1660, 2020. 3

[87] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 13658–13667, 2021. 5

[88] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2914–2923, 2017. 2

[89] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 27378–27394, 2022. 3