Improve Video Representation with Temporal Adversarial Augmentation

Jinhao Duan[1] , **Quanfu Fan**[2] , **Hao Cheng**[3] , **Xiaoshuang Shi**[4*] and **Kaidi Xu**[1*]
[1]Drexel University
[2]Amazon
[3]The Hong Kong University of Science and Technology (Guangzhou)
[4]University of Electronic Science and Technology of China
jd3734@drexel.edu, quanfu@amazon.com, hcheng046@connect.hkust-gz.edu.cn,
xsshi2013@gmail.com, kx46@drexel.edu

Abstract

Recent works reveal that adversarial augmentation benefits the generalization of neural networks (NNs) if used in an appropriate manner. In this paper, we introduce **T**emporal Adversarial **A**ugmentation (TA), a novel video augmentation technique that utilizes temporal attention. Unlike conventional adversarial augmentation, TA is specifically designed to shift the attention distributions of neural networks with respect to video clips by maximizing a temporal-related loss function. We demonstrate that TA will obtain diverse temporal views, which significantly affect the focus of neural networks. Training with these examples remedies the flaw of unbalanced temporal information perception and enhances the ability to defend against temporal shifts, ultimately leading to better generalization. To leverage TA, we propose **T**emporal Video **A**dversarial **F**ine-tuning (TAF) framework for improving video representations. TAF is a model-agnostic, generic, and interpretability-friendly training strategy. We evaluate TAF with four powerful models (TSM, GST, TAM, and TPN) over three challenging temporal-related benchmarks (Something-something V1&V2 and diving48). Experimental results demonstrate that TAF effectively improves the test accuracy of these models with notable margins without introducing additional parameters or computational costs. As a byproduct, TAF also improves the robustness under out-of-distribution (OOD) settings. Code is available at https://github.com/jinhaoduan/TAF.

1 Introduction

Deep learning has achieved significant successes in multiple domains [Yuan and Moghaddam, 2020; Shi *et al.*, 2020; Yuan *et al.*, 2023; Cao *et al.*, 2023]. However, adversarial examples have been widely recognized as a serious threat to neural networks (NNs) [Xu *et al.*, 2018; Xu *et al.*, 2020]. Imperceptible distortions created by advanced adversarial attack algorithms can easily manipulate the decision of well-trained neural networks. This issue would be more critical for security-sensitive scenarios, such as biological identification [Dong *et al.*, 2019] and autonomous [Wang *et al.*, 2022]. However, recent works also reveal that adversarial examples could benefit NNs if used in the appropriate manner. For instance, adversarial examples could be the special cases when perceiving the category decision boundaries [Tanay and Griffin, 2016]. Also, by regarding adversarial examples as special augmentations, jointly utilizing adversarial examples and natural examples during the training will ameliorate the generalization of NNs [Xie *et al.*, 2020; Chen *et al.*, 2021a].

Temporal modeling is the decisive procedure for video understanding tasks [Wang *et al.*, 2016]. Recently, various modules are proposed to capture temporal information. For example, equipping networks with temporal convolution operations [Lin *et al.*, 2019; Luo and Yuille, 2019; Carreira and Zisserman, 2017] and local/global attention mechanisms [Wang *et al.*, 2018; Fan *et al.*, 2019] are the most common practices. Although these methods make great progress on this issue, the main concern is that such strategies tend to achieve narrow and overly centered temporal attention. No mechanisms guarantee the surrounding temporal clues, which may contain valuable information, will also be fully explored equivalently. Figure 1 shows that current state-of-the-art model inferences only rely on a few frames, which ignores the following "**holding**" part and provides the wrong prediction. This property helps to converge training samples but might result in temporally over-fitting and hurts the generalization. Therefore, we may ask: do balanced temporal attention distributions help with the generalization?

Training with adversarial augmentation is a promising scheme to adaptively regularize the temporal attention distributions of NNs. On the one hand, considering that adversarially augmented examples share the same semantic contents as natural examples, training with these examples will keep the consistency and stability of learning targets. On the other hand, the constructed perturbations can largely affect the behavior of NNs, which provides an opportunity to concretely rectify models according to the needs. Motivated by that, we propose **T**emporal Adversarial **A**ugmentation (TA) to address the unbalanced perception of temporal informa-

[*] Equal corresponding author.

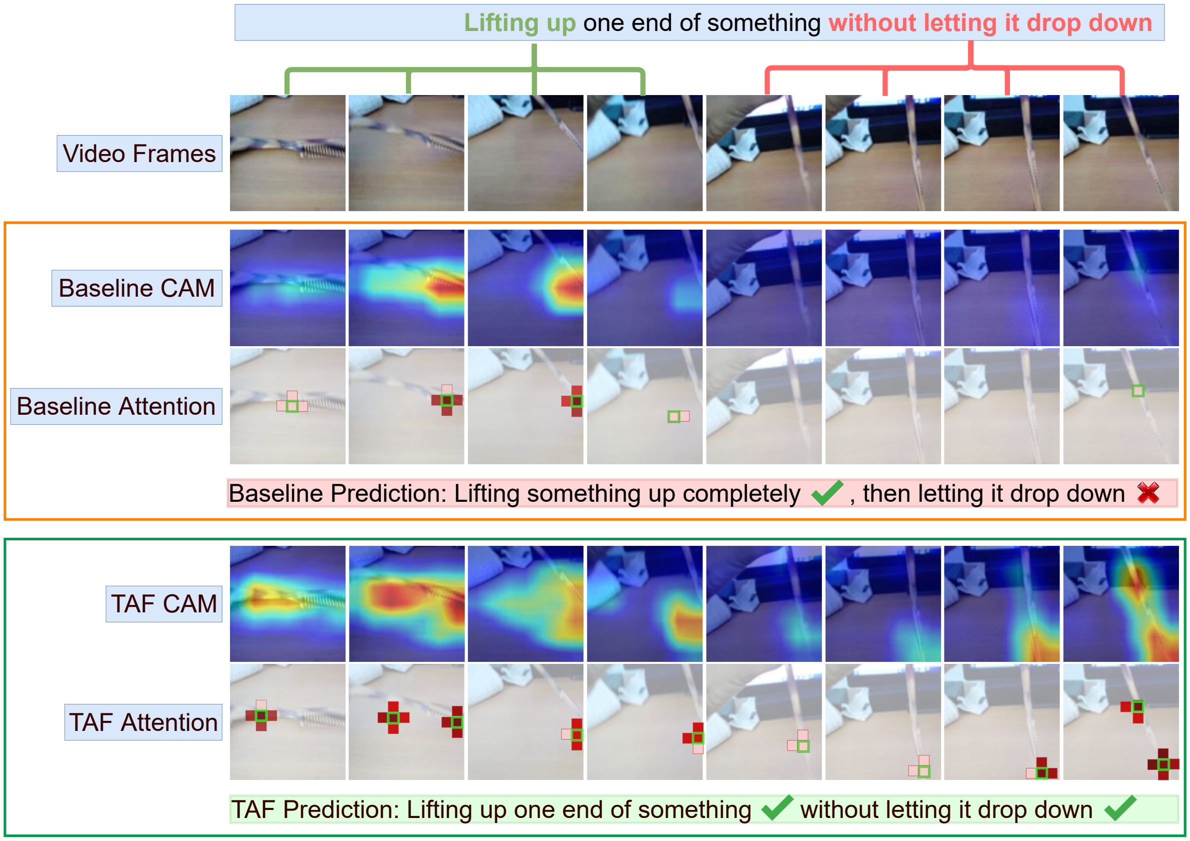
Lifting up one end of something without letting it drop down
Video Frames
Baseline CAM
Baseline Attention
Baseline Prediction: Lifting something up completely ✓ , then letting it drop down ✗
TAF CAM
TAF Attention
TAF Prediction: Lifting up one end of something ✓ without letting it drop down ✓

Figure 1: TAF enables models to capture global temporal cues by broadening and balancing the temporal attention distributions. As a result, the model is able to capture both the "lifting" and "holding" parts of the video, whereas the baseline model without TAF can only focus on the "lifting" process and ignores the "holding" parts.

tion. Different from conventional adversarial augmentation, TA is specifically designed to perturb the temporal attention distributions of NNs. Concretely, TA utilizes Class Activation Mapping (CAM)-based temporal loss function to represent the temporal attention distributions w.r.t. video clips, and disturb NNs' temporal views by maximizing the temporal loss function. In this way, videos augmented by TA will obtain diverse temporal attention (Figure 3). Training with temporally augmented examples will remedy the defect of unbalanced attention assignation. Our contributions can be summarized as the following:

- We introduce **T**emporal **A**dversarial **A**ugmentation (TA). TA changes the temporal distributions of video clips and provides more temporal views for video understanding models.
- We propose **T**emporal Video **A**dversarial **F**ine-tuning (TAF) framework to regularize the attention distribution of networks by utilizing temporal adversarial augmentation. TAF is a model-agnostic, generic, and interpretability-friendly training scheme. This is the first work to improve video understanding models by explicitly utilizing adversarial machine learning.
- TAF is performed on four powerful video understanding models, including TSM [Lin *et al.*, 2019], TAM [Fan *et al.*, 2019], GST [Luo and Yuille, 2019], and TPN [Yang *et al.*, 2020], and evaluated on three temporal related benchmarks (Somthing-something V1 & V2 and Diving48). Experimental results demonstrate that TAF can significantly boost test accuracy, without any additional parameters and computational costs.
- TAF is evaluated under the out-of-distribution [Hendrycks and Dietterich, 2018] settings and it effectively boosts robustness in defending naturally corrupted data with notable margins.

2 Related Works

2.1 Adversarial Machine Learning for Good

Enhancing NNs by taking advantage of adversarial machine learning is being popular. One of the main contributions in this field is revealing the relationship between robustness and generalization [Tsipras *et al.*, 2018; Su *et al.*, 2018]. [Tsipras *et al.*, 2018] theoretically proves that there is an inherent gap between robustness and generalization. Similar conclusions are proposed in [Su *et al.*, 2018] with empirical evidence. In applications, [Xie *et al.*, 2020; Chen *et al.*, 2021a] prove that by utilizing auxiliary batch normalization layers, adversarial examples could benefit the generalization and improve image

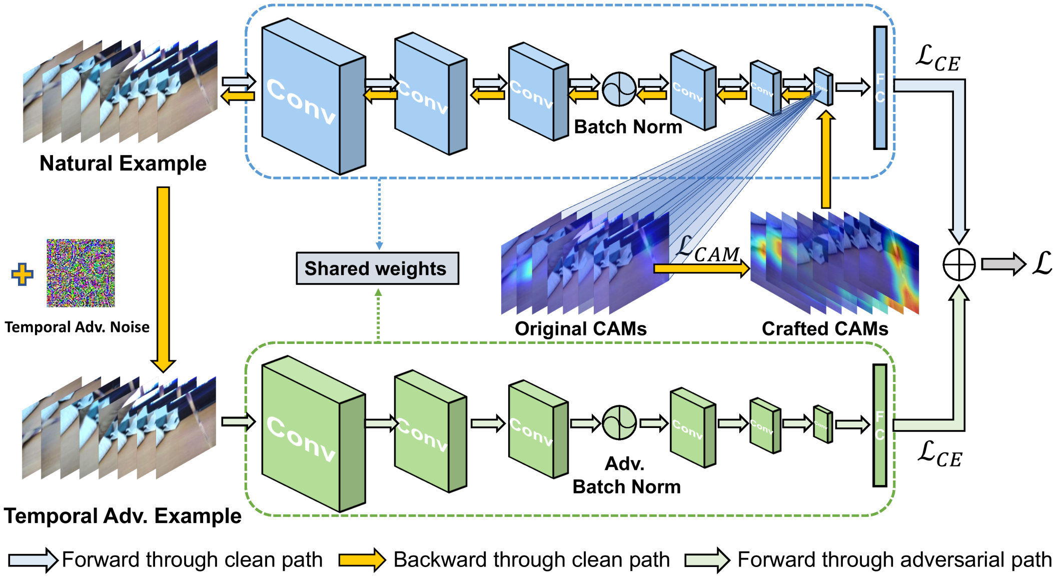

Figure 2: The overall pipeline of TAF. The upper network (clean path) and the lower network (adversarial path) share weights except for batch normalization layers.

classification tasks. Besides, adversarial examples show great potential in understanding and interpreting NNs [Ignatiev *et al.*, 2019; Boopathy *et al.*, 2020]. [Ignatiev *et al.*, 2019] reveals that adversarial examples and explanations can be connected by the general formal of hitting set duality. [Boopathy *et al.*, 2020] propose an interpretability-aware defensive mechanism to achieve both robust classification and robust interpretation.

2.2 Robustness of Video Understanding

Video understanding, as one of the most important vision tasks, has been promoted to a new level with the help of NNs. [Carreira and Zisserman, 2017; Xie *et al.*, 2018; Lin *et al.*, 2019; Wang *et al.*, 2016] propose to improve video representations with NNs constructed by different architectures (e.g., 2D/3D structures, recurrent architectures). [Arnab *et al.*, 2021; Wei *et al.*, 2022] prove that Transformer architectures benefit video understanding tasks a lot. Besides, boosting temporal modeling by incorporating NNs with additional attention mechanisms is also practical and popular [Wang *et al.*, 2018; Fan *et al.*, 2019]. In terms of robustness, various attack and defense algorithms are proposed in video scenarios. [Li *et al.*, 2018a; Chen *et al.*, 2021b] show that video understanding models can also be easily manipulated by imperceptible perturbations. Furthermore, [Li *et al.*, 2021] proves that video models can be attacked under different settings (e.g., black-box attack and transfer attack). On the other side, common defense techniques (e.g., adversarial training [Kinfu and Vidal, 2022], adversarial detection [Thakur and Li, 2022]) are also compatible with video understanding models.

2.3 Class Activation Mapping

Class Activation Mapping (CAM) [Zhou *et al.*, 2016] is one of the most popular approaches in understanding and interpreting the attention distributions of NNs. CAM measures the importance of channels by global average pooling and weighted fusion. After that, a series of gradient-based [Selvaraju *et al.*, 2017; Chattopadhay *et al.*, 2018] and gradient-free [Petsiuk *et al.*, 2021] CAM solvers are also introduced. Gradient-based CAM uses either first-order [Selvaraju *et al.*, 2017] or second-order [Chattopadhay *et al.*, 2018] gradient information to perceive the interested regions of NNs. Considering GradCAM [Selvaraju *et al.*, 2017] is compatible with various model structures, and it is more effective than gradient-free CAM, we choose it as the CAM producer in the rest of this paper.

3 Approach

In this section, we describe how to capture balanced temporal information with the proposed Temporal Video Adversarial Fine-tuning (TAF) framework. Firstly, we revisit the vanilla adversarial augmentation. Then, detailed illustrations of TA along with the CAM-based metrics are provided. At last, we outline the training and testing protocols of the TAF framework. The pipeline of TAF is shown in Figure 2.

3.1 Vanilla Adversarial Augmentation

Adversarial augmentation derives from adversarial perturbation, a kind of imperceptible noise that can easily disturb the predictions of well-trained NNs. [Xie *et al.*, 2020; Chen *et al.*, 2021a] show that adversarial perturbations can be seen as special augmentations to improve generalization and robustness in image-based vision tasks.

For a given model, $\mathcal{F}$, parameterized by weights θ and input $X \in \mathbb{R}^{C\times H\times W}$ with C channels and resolution $H \times W$, the adversarially augmented example, X', is defined as:

$$X' = X + \delta = X + \arg\max_{\delta\in[-\epsilon,\epsilon]} \mathcal{L}(X + \delta, y), \tag{1}$$

where δ is the adversarial noise solved by either single-step (e.g., FGSM [Goodfellow *et al.*, 2014]) or iterative (e.g., PGD [Kurakin *et al.*, 2018]) attack algorithms, ϵ is the attack budget, and $\mathcal{L}$ is the conventional classification loss (i.e., Cross-entropy Loss).

Vanilla adversarial augmentation is an effective technique for image tasks, but it is not well-suited for video scenarios. Video understanding models often suffer from serious overfitting issues, with over 40% overfitting gaps (i.e., top-1 training accuracy vs top-1 validation accuracy) observed on the Something-something V1 dataset. This severe overfitting suggests that a lot of generalization-irrelevant noise is introduced during the training. For neural networks, the loss function plays a crucial role in determining which features or information are absorbed. Therefore, using a classification loss alone may propagate these irrelevant noises back to the adversarial perturbation, which ultimately harms the generalization of neural networks.

3.2 Temporal Adversarial Augmentation

To address this issue, the proposed Temporal Adversarial Augmentation (TA) utilizes a CAM-based temporal loss function to leverage temporal attention-related information solely, which is one of the most fundamental and essential features of videos. Here we show how to incorporate temporal information into adversarial augmentation.

For a given model, $\mathcal{F}$, parameterized by weights θ and video volume $X \in \mathbb{R}^{T\times N_c\times H\times W}$ with T frames, N_c channels and $H \times W$ resolution, we first consider the CAMs of model $\mathcal{F}$ w.r.t. frame X_i,

$$X_i^{\mathcal{C}} = \mathcal{G}^{\mathcal{C}}(\theta, X_i, \hat{y}) = \frac{1}{HW}\sum^{H}\sum^{W}\frac{\partial \mathcal{L}_{ce}(\theta, X, \hat{y})}{\partial \mathcal{F}^{'}(X_i)}\mathcal{F}^{'}(X_i), \quad i = 1, \ldots, T, \tag{2}$$

where $\mathcal{F}^{'}$ refers to the subnetwork of $\mathcal{F}$ from the input layer to the final convolutional layer and $\mathcal{L}_{ce}(\cdot,\cdot,\cdot)$ is Cross-entropy loss. $X_i^{\mathcal{C}}$ represents the CAM of the i-th frame. To discriminate the importance of each frame, we normalize each CAM and define the overall CAM on video X as

$$\hat{X}_i^{\mathcal{C}} = \frac{X_i^{\mathcal{C}} - \min(X^{\mathcal{C}})}{\max(X^{\mathcal{C}})} \quad i = 1, \ldots, T, \tag{3}$$

where $\min(\cdot)$ and $\max(\cdot)$ refer to the minimal and maximal values among all CAMs and $X^{\mathcal{C}} = \{X_1^{\mathcal{C}}, \cdots, X_T^{\mathcal{C}}\}$.

We balance temporal attention distribution by amplifying those "unimportant" frames (i.e., frames with smaller CAM values). Concretely, we sort $\hat{X}_i^{\mathcal{C}}$ in ascending order according to the sum of CAM values at each frame:

$$\Sigma\hat{X}^{\mathcal{C}}_{\pi 1} < \Sigma\hat{X}^{\mathcal{C}}_{\pi 2} < \cdots < \Sigma\hat{X}^{\mathcal{C}}_{\pi N} < \cdots < \Sigma\hat{X}^{\mathcal{C}}_{\pi K} \tag{4}$$

The top N frames with the smallest CAM values are selected as the *non-key frames*. Then, the CAM-based loss is defined as

$$\mathcal{L}_{\mathcal{C}} = \frac{1}{N}\sum_{i}^{N}\hat{X}^{\mathcal{C}}_{\pi i}. \tag{5}$$

By maximizing $\mathcal{L}_{\mathcal{C}}$, global temporal attention will be reassigned to those unimportant frames, which balances the attention distribution.

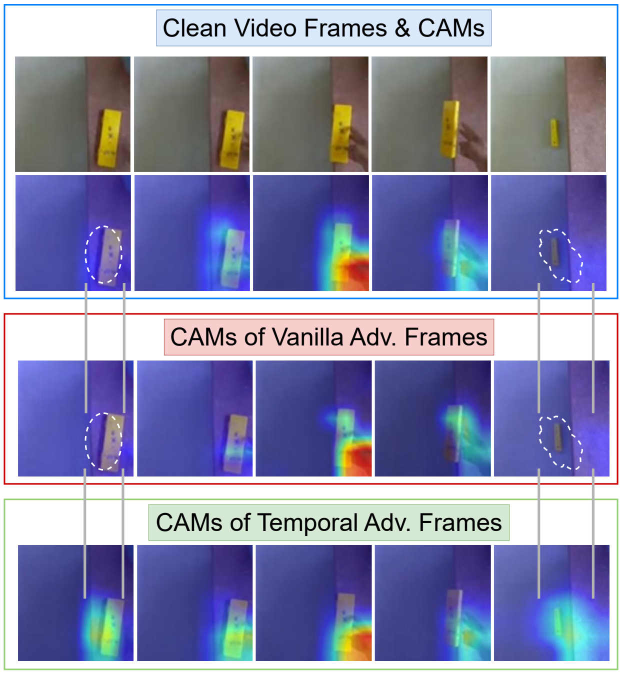
Clean Video Frames & CAMs
CAMs of Vanilla Adv. Frames
CAMs of Temporal Adv. Frames

Figure 3: Adversarially augmented examples and the corresponding CAMs. Vanilla adversarial augmentation does not substantially affect the temporal attention distribution, while our temporal adversarial augmentation changes it significantly. This highlights the benefits of TAF in terms of broadening and balancing the temporal attention distributions.

To generate the final temporal augmentation, we utilize the popular iterative gradient sign method (i.e. PGD [Kurakin *et al.*, 2018]) and update X in K iteration steps with the formulation

$$X_i^{k+1} = \Pi_{X\pm\epsilon}(X_i^k + \beta * sgn(\nabla_{X_i^K}\mathcal{L}_{\mathcal{C}}(\theta, X_i, \hat{y}))), \quad i = 1, \ldots, T, \;\; k = 1, \ldots, K, \tag{6}$$

where ϵ refers to the attack budget under ℓ_{inf} constraint and $\Pi_{X\pm\epsilon}(\cdot)$ projects tensor back to $[X - \epsilon, X + \epsilon]$. β represents the attack step size. In terms of the targeted label $\hat{y}$, we assign it based on the example's prediction. For correctly classified samples, models are expected to be robust against temporal shifting. Therefore, a random label is set to $\hat{y}$, which will generate diverse temporal views. However, for incorrectly classified samples, these are hard examples for the model and we choose to reduce the difficulties of these examples by correctly boosting the temporal attention [Lee *et al.*, 2021]. In this situation, we set $\hat{y}$ with the true label.

As shown in Figure 3, by shifting the temporal attention of models, temporally augmented video clips have diverse temporal views compared with conventional adversarial augmentation. Training with these examples can be seen as a regularization applied to the NNs for encouraging diverse inference, which results in better generalization to unseen samples.

3.3 Adversarial Fine-Tuning Framework for Video Understanding

TAF jointly utilizes both natural examples and adversarially augmented examples with the following optimization object:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}\left[\alpha * \mathcal{L}_{ce}(\theta, x, y) + (1-\alpha) * \mathcal{L}_{ce}(\theta, x^K, y)\right], \tag{7}$$

where $\mathcal{L}_{ce}$ is the Cross-entropy loss function, and x^K is the temporally augmented data based on Eqn. 6. α is used to control the contribution of loss items. Besides, inspired by [Xie *et al.*, 2020], we adopt additional normalization layers (i.e., Batch Normalization [Ioffe and Szegedy, 2015]) to deal with the distribution mismatching between adversarial examples and natural examples. We name the forward path with original normalization layers and additional normalization layers as *clean path* and *adversarial path*, respectively.

For training, the natural examples will be first passed through the clean path and get two outputs: clean logits and CAM feature maps. Then, temporally augmented examples are generated by performing Eq. 6 on CAM feature maps and back-propagating through the clean path to the input example. Next, pass the augmented examples through the adversarial path and get adversarial logits. Finally, optimizing both clean path and adversarial path with classification loss. For inference, we drop adversarial normalization layers and get the final prediction with the clean path. The pseudo-code of TAF is shown in Appendix A.

Dataset	Method	Backbone	frames	Top-1(%)	Top-5(%)
SSV1	TSM	resnet50	8	45.6	74.6
	TSM+TAF	resnet50	8	**46.9(+1.3)**	**75.0(+0.4)**
	TSM	resnet50	16	47.2	77.1
	TSM+TAF	resnet50	16	**47.9(+0.7)**	**77.7(+0.6)**
	GST	3d-resnet50	8	46.6	75.6
	GST+TAF	3d-resnet50	8	**47.6(+1.0)**	**76.4(+0.8)**
	GST	3d-resnet50	16	48.4	77.2
	GST+TAF	3d-resnet50	16	**48.8(+0.4)**	**77.5(+0.3)**
	GST-L	3d-resnet50	8	47.0	76.1
	GST-L+TAF	3d-resnet50	8	**47.7(+0.7)**	**76.3(+0.2)**
	TAM	resnet50	8	46.2	75.4
	TAM+TAF	resnet50	8	**46.6(+0.4)**	75.4
	TPN	resnet50	8	48.0	77.2
	TPN+TAF	resnet50	8	**49.1(+1.1)**	**78.0(+0.8)**
SSV2	TSM	resnet50	8	58.9	85.5
	TSM+TAF	resnet50	8	**59.8(+0.9)**	**86.0(+0.5)**
	TSM	resnet50	16	61.1	86.8
	TSM+TAF	resnet50	16	**62.0(+0.9)**	**87.3(+0.5)**
	GST	3d-resnet50	8	61.3	87.2
	GST+TAF	3d-resnet50	8	**61.7(+0.4)**	**87.4(+0.2)**
	TPN	resnet50	8	61.6	87.7
	TPN+TAF	resnet50	8	**62.1(+0.5)**	**88.3(+0.6)**
D48	TSM	resnet50	8	78.5	97.3
	TSM+TAF	resnet50	8	**79.1(+0.6)**	**97.6(+0.3)**
	GST	3d-resnet50	8	73.9	96.7
	GST+TAF	3d-resnet50	8	**74.7(+0.8)**	**96.9(+0.2)**
	TAM	resnet50	8	75.1	96.3
	TAM+TAF	resnet50	8	**75.8(+0.7)**	**97.1(+0.8)**
	TPN	resnet50	8	80.2	**98.4**
	TPN+TAF	resnet50	8	**80.9(+0.7)**	98.0

Table 1: Evaluations on Something-something V1 and V2 (SSV1 and SSV2), Diving48 (D48) benchmarks.

4 Experiments

In this section, we investigate the effectiveness of TAF through comprehensive experiments. Concretely, we first introduce our experiment settings, including datasets, baselines, and implementation details. The evaluations on multiple state-of-the-art models and challenging benchmarks are followed. Then, a series of ablation studies are conducted, including comparing with vanilla adversarial augmentation, impacts of α and attacking settings, relieving overfitting, and the discussion about computational costs. Moreover, we qualitatively analyze TAF by providing representative visualizations. At last, we examine the models' ability against naturally corrupted data [Hendrycks and Dietterich, 2018], i.e., out-of-distribution (OOD) robustness.

4.1 Datasets and Baselines

We evaluate TAF on three popular temporal datasets: Something-something V1&V2 [Goyal *et al.*, 2017], Diving48 [Li *et al.*, 2018b]. Something-something V1&V2 are large-scale challenging video understanding benchmarks consisting of 174 classes. Diving48 is a fine-grained temporal action recognition dataset with 48 dive classes. The detailed introduction of these datasets can be found in Appendix B

The reason why we choose these three datasets is that TAF aims to tackle temporal modeling issues. Both Something-something and Diving48 are the most challenging benchmarks in this field [Lin *et al.*, 2019]. Only action descriptions are reserved in these datasets, without introducing scene-related knowledge that enforces the model to learn the temporal information. In this way, TAF can be fairly and entirely evaluated.

Baselines. To evaluate the effectiveness of TAF, we conduct experiments on four powerful action recognition models: TSM [Lin *et al.*, 2019], GST [Luo and Yuille, 2019], TAM [Fan *et al.*, 2019] and TPN [Yang *et al.*, 2020]. These models cover the most representative action recognition methodologies: 2DConvNets-based recognition [Lin *et al.*, 2019; Fan *et al.*, 2019; Yang *et al.*, 2020], 3DConvNet-based recognition [Luo and Yuille, 2019] and attention-based recognition [Fan *et al.*, 2019; Yang *et al.*, 2020]. We train all the baseline models by strictly following the training protocols provided by their official codebases.

Fine-tuning. For fine-tuning, we load pre-trained weights and keep training 15 epochs with TAF. We conduct 3 trials for each experiment and report the mean results. The initial training settings (e.g., learning rate, batch size, dropout, etc.) are the same as the status when the pre-trained models are logged. The learning rates are decayed by a factor of 10 after 10 epochs. We set α as 0.7, and the number of attacked frames N as 8 or 16 according to the input temporal length. Note that considering most of the baseline models did not conduct experiments on Diving48, we adopt the same training settings as on the Something-something datasets.

Inference. For fairness and convenience, all the performances reported in this paper are evaluated on 1 center crop and 1 clip, with input resolution 224 × 224.

Method	Scratch	Fine-t.	CELoss	CAMLoss	Top-1	Δ
baseline	√				45.6	-
AdvProp [51]	√		√		44.5	-1.1
CE-based		√	√		46.1	+0.5
TAF		√		√	**46.9**	+1.3

(a) Scratch and Fine-t. stand for training from scratch and fine-tuning from pre-trained models. CELoss/CAMLoss refers to the objective function used to generate adversarial perturbations.

Method	Top-1 (%)	Top-5 (%)
baseline	45.6	74.6
$\alpha = 0.2$	46.3	75.1
$\alpha = 0.5$	46.7	75.1
$\alpha = 0.7$	**46.9**	75.0
$\alpha = 0.8$	46.5	74.8

(b) Impacts of applying different α when optimizing objectives.

Method	Top-1 (%)	Top-5 (%)
baseline	45.6	74.6
$\epsilon = 6$ $K = 1$	46.6	75.1
$\epsilon = 6$ $K = 3$	46.5	74.6
$\epsilon = 64$ $K = 1$	**46.9**	75.0
$\epsilon = 64$ $K = 3$	46.8	75.0

(c) Impacts of attacking settings. ϵ is scaled by 255. K refers to the number of attack steps.

Table 2: Ablation study on Something-something V1 benchmark.

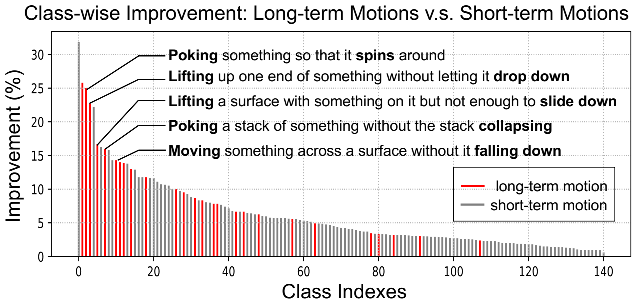

Figure 4: Class-wise improvements on the Something-something V1 dataset. TAF benefits more on the long-term motions.

4.2 Comparisons with State-of-the-Art Models

The performances of TAF on Something-something V1 & V2, and Diving48 are summarized in Table 1. It is shown that TAF effectively improves on TSM, TAM, TPN, and GST with 1.3%, 0.4%, 1.1%, and 1.0% top-1 accuracy on Something-something V1, respectively. Similar promotions are also observed on the V2 dataset. As for longer temporal inputs, we evaluate TAF in 16-frame settings and show that TAF augments TSM with 0.7% and 0.9% top-1 accuracy on V1 and V2, separately. The experiments of longer inputs on GST are also promising. For Diving48, TAF improves TSM by 0.6%, GST by 0.8%, TAM by 0.7%, and TPN by 0.7%.

To further investigate the impact of TAF, class-wise improvements v.s. class indexes are summarized in Figure 4. Since TAF benefits models by broadening the temporal attention distributions, we focus on the gains of motions requiring longer temporal knowledge, referred to as *long-term motions*. We define long-term motion as an action consisting of at least two motions or phenomenons, such as “**Lifting up** one end of something without letting it **drop down**”. Based on our definition, 35 categories are recognized as the long-term motions out of 174 classes from the Something-something benchmarks. Appendix C provides the complete list of long-term motions. The aggregation of red bars demonstrates that TAF substantially enhances the ability to capture long-term temporal cues: Over 10 long-term motions get improved largely (i.e., $\geq 10\%$).

4.3 Ablation Study

We conduct comprehensive experiments exploring the impacts of TAF. At first, we compare the proposed TAF with Cross-entropy based (CE-based) training strategy. Then, we study the effects of α and the performances of TAF under different attacking settings (e.g., epsilon ϵ, attack steps T). The training loss and testing loss are provided to justify the regularization of overfitting. Note that all ablation studies are performed on resnet50-TSM, with Something-something V1 as the dataset. Appendix D provides additional ablation studies and comparisons with other popular data augmentation methods.

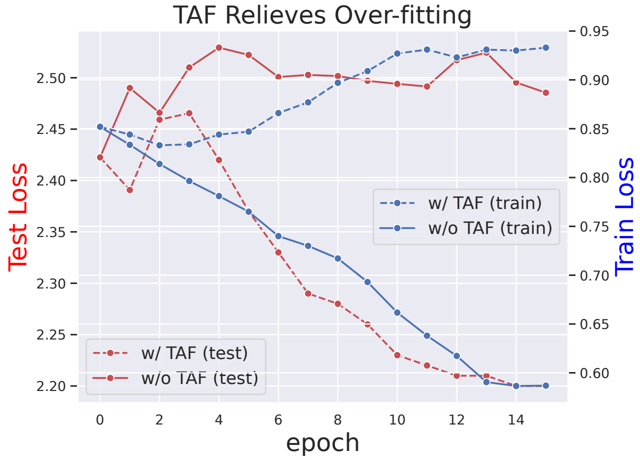

Figure 5: Training and testing loss curves of TSM w/ and w/o TAF on Something-something V1. The testing loss is decreased while the training loss is increased when TAF is utilized, which shows that TAF effectively relieves over-fitting.

Comparison with Vanilla Adversarial Augmentation

Results are placed in Table 2a. It is shown that although the CE-based method achieves a certain improvement (0.5%) over the baseline model, still TAF remarkably outperforms the CE-based method (0.8%) and baseline model (1.3%). As we mentioned in Sec. 3.1, directly taking advantage of conventional classification loss will introduce lots of irrelevant noises into the adversarial perturbations, especially in heavy-overfitting situations. Training with these examples benefits less on generalization. However, TAF utilizes a CAM-based loss function to filter all noise except temporal modeling knowledge. Therefore, training with temporally augmented examples is more effective and more suitable for video understanding tasks.

	Noise			Blur			Weather	
Model	Gauss.	Impulse	Speckle	Gauss.	Defocus	Zoom	Snow	Bright
TSM	19.9/41.5	16.7/35.7	20.1/41.4	20.0/41.5	19.3/40.8	18.9/40.3	11.0/25.3	17.2/36.2
TSM+TAF	**21.0/43.6**	**17.9/37.7**	**21.1/44.0**	**20.8/43.6**	**20.4/42.9**	**20.0/42.3**	**11.9/27.2**	**18.2/38.5**
TPN	24.7/48.8	17.3/37.6	24.4/48.8	25.3/49.6	24.7/48.7	24.3/48.2	13.6/31.0	19.5/40.1
TPN+TAF	**25.7/50.1**	**18.7/39.8**	**25.6/49.8**	**26.2/50.7**	**25.7/50.1**	**25.3/49.4**	**15.1/32.6**	**20.8/41.5**

Table 3: Evaluations of defending natural corruption. Performances are reported as Top-1(%)/Top-5(%).

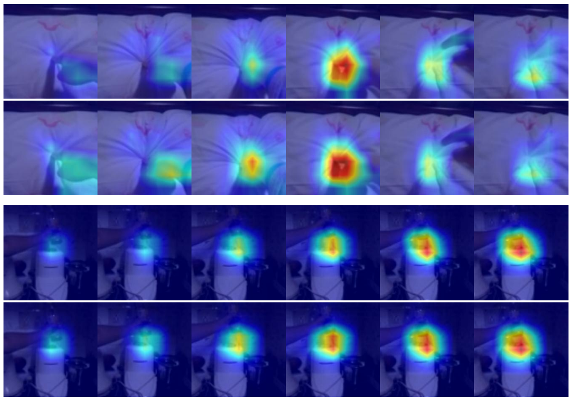

Figure 6: Two representative visualizations. For each group, row 1 refers to the CAM generated by the baseline model. Row 2 represents the CAM created by the model fine-tuned with TAF. TAF generates more balanced CAMs.

Impacts of α and Attack Settings

Impacts of α and different attack settings are shown in Table 2b and Table 2c. It shows that incorporating with a proper portion of temporally augmented examples can effectively boost TSM, and the best performance is achieved when α is set to 0.7. In terms of the impacts of attack settings, we investigate it by applying two schemes: small perturbation (i.e., $\epsilon = 6/255$, $\beta = 2/255$) and large perturbation (i.e., $\epsilon = 64/255$, $\beta = 32/255$) in Table 2c, along with single (i.e., $K = 1$) or multiply (i.e., $K = 3$) steps. Generally, larger ϵ allows injecting more temporal perturbations into natural examples and further achieves better temporal robustness.

Relieving Over-Fitting

We provide the training and testing curves of TSM w/ and w/o TAF in Figure 5. It is shown that the testing losses of TSM are significantly reduced (i.e., from approximately 2.4 to 2.2) while the training losses increase slightly during the initial fine-tuning phase when TAF is utilized. This demonstrates that TAF effectively alleviates the overfitting issue.

Computational Cost

For training costs, since we only fine-tune models 15 epochs with very limited attack steps, the computational overheads are marginal. For instance, TAF will bring 25% and 15% additional computational costs for TSM and TPN, respectively. For inference, TAF is only applied to the training stage, the inference costs will be identical to the baseline models.

4.4 Visualization

To qualitatively analyze the effect of TAF, a set of representative visualizations are provided in Figure 6 and Appendix E. Each group of visualizations includes a misclassified example and its corresponding CAM generated by the baseline model (**Row 1**), as well as the CAM generated by the model fine-tuned with TAF (**Row 2**). Generally, CAMs from TAF and the baseline model share similar trends, which is expected as TAF only fine-tunes the model for a limited number of epochs and is not expected to fundamentally change the underlying model. However, it is clear that the CAMs generated by TAF fine-tuned models are broader and uniform compared to the primitive results. This observation further verifies our hypothesis that fine-tuning models with TAF can regularize temporal modeling and achieve wider attention distributions.

4.5 Robustness Analysis

Spurred by [Hendrycks and Dietterich, 2018], we conduct experiments to evaluate the performance of TAF under (Out-of-distribution) OOD settings. Three types of noise (i.e., Gaussian, Impulse, Speckle), three types of blur (i.e., Gaussian, Defocus, Zoom), and two kinds of weather corruption (i.e., Snow and Bright) are adopted. Results are presented in Table 3. TPN fine-tuned with TAF outperforms the vanilla TPN by 1.5% on Weather corruptions and 1% on other noises or blurs. Similar results are also observed on TSM. It reflects that the proposed TAF not only benefits the generalization of NNs but also strengthens the robustness against natural perturbations from the physical world. Evaluations of other models in resisting natural corruption are provided in Appendix F.

5 Conclusions

In this work, we propose TAF, a Temporal Augmentation Framework, to regularize temporal attention distributions and improve generalization in video understanding tasks. TAF leverages specifically designed temporal adversarial augmentation during fine-tuning to enhance the performance of models. Our experiments on three challenging benchmarks using four powerful models demonstrate the improvements of TAF are multi-faceted: improving video representation, relieving over-fitting issues, and strengthing OOD robustness.

To the best of our knowledge, this is the first work to enhance video understanding tasks with the help of adversarial machine learning. We believe that we have established a novel and practical connection between the field of adversarial machine learning and the video understanding community.

Contribution Statement

Jinhao Duan performed all the experiments and wrote the paper. All other authors contributed to the idea discussion and wrote part of the paper. Part of the work was done while Quanfu Fan worked at MIT-IBM Watson AI lab.

References

[Arnab *et al.*, 2021] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[Boopathy *et al.*, 2020] Akhilan Boopathy, Sijia Liu, Gaoyuan Zhang, Cynthia Liu, Pin-Yu Chen, Shiyu Chang, and Luca Daniel. Proper network interpretability helps adversarial robustness in classification. In *International Conference on Machine Learning*, pages 1014–1023. PMLR, 2020.

[Cao *et al.*, 2023] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023.

[Carreira and Zisserman, 2017] 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*, pages 6299–6308, 2017.

[Chattopadhay *et al.*, 2018] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[Chen *et al.*, 2021a] Xiangning Chen, Cihang Xie, Mingxing Tan, Li Zhang, Cho-Jui Hsieh, and Boqing Gong. Robust and accurate object detection via adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16622–16631, 2021.

[Chen *et al.*, 2021b] Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, and Qi Tian. Appending adversarial frames for universal video attack. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3199–3208, 2021.

[Dong *et al.*, 2019] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019.

[Fan *et al.*, 2019] Quanfu Fan, Chun-Fu (Ricarhd) Chen, Hilde Kuehne, Marco Pistoia, and David Cox. More Is Less: Learning Efficient Video Representations by Temporal Aggregation Modules. In *Advances in Neural Information Processing Systems 33*. 2019.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[Goyal *et al.*, 2017] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[Hendrycks and Dietterich, 2018] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

[Ignatiev *et al.*, 2019] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On relating explanations and adversarial examples. *Advances in neural information processing systems*, 32, 2019.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[Kim *et al.*, 2020] 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 *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 386–403. Springer, 2020.

[Kinfu and Vidal, 2022] Kaleab A Kinfu and René Vidal. Analysis and extensions of adversarial training for video classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3416–3425, 2022.

[Kurakin *et al.*, 2018] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[Lee *et al.*, 2021] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4071–4080, 2021.

[Li *et al.*, 2018a] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy Chowdhury, and Ananthram Swami. Adversarial perturbations against real-time video classification systems. *arXiv preprint arXiv:1807.00458*, 2018.

[Li *et al.*, 2018b] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018.

[Li *et al.*, 2021] Shasha Li, Abhishek Aich, Shitong Zhu, Salman Asif, Chengyu Song, Amit Roy-Chowdhury, and Srikanth Krishnamurthy. Adversarial attacks on black box video classifiers: Leveraging the power of geometric transformations. *Advances in Neural Information Processing Systems*, 34:2085–2096, 2021.

[Lin *et al.*, 2019] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.

[Luo and Yuille, 2019] Chenxu Luo and Alan Yuille. Grouped spatial-temporal aggretation for efficient action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[Petsiuk *et al.*, 2021] Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11443–11452, 2021.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Shi *et al.*, 2020] Xiaoshuang Shi, Fuyong Xing, Kaidi Xu, Pingjun Chen, Yun Liang, Zhiyong Lu, and Zhenhua Guo. Loss-based attention for interpreting image-level prediction of convolutional neural networks. *IEEE Transactions on Image Processing*, 30:1662–1675, 2020.

[Su *et al.*, 2018] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[Tanay and Griffin, 2016] Thomas Tanay and Lewis Griffin. A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*, 2016.

[Thakur and Li, 2022] Nupur Thakur and Baoxin Li. Pat: Pseudo-adversarial training for detecting adversarial videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 131–138, 2022.

[Tsipras *et al.*, 2018] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2018.

[Wang *et al.*, 2016] 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 *European conference on computer vision*, pages 20–36. Springer, 2016.

[Wang *et al.*, 2018] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[Wang *et al.*, 2022] Ningfei Wang, Yunpeng Luo, Takami Sato, Kaidi Xu, and Qi Alfred Chen. Poster: On the system-level effectiveness of physical object-hiding adversarial attack in autonomous driving. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 3479–3481, 2022.

[Wei *et al.*, 2022] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.

[Xie *et al.*, 2018] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

[Xie *et al.*, 2020] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020.

[Xu *et al.*, 2018] Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin. Structured adversarial attack: Towards general implementation and better interpretability. In *International Conference on Learning Representations*, 2018.

[Xu *et al.*, 2020] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *European conference on computer vision*, pages 665–681. Springer, 2020.

[Yang *et al.*, 2020] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 591–600, 2020.

[Yuan and Moghaddam, 2020] Chenxi Yuan and Mohsen Moghaddam. Attribute-aware generative design with generative adversarial networks. *Ieee Access*, 8:190710–190721, 2020.

[Yuan *et al.*, 2023] Chenxi Yuan, Tucker Marion, and Mohsen Moghaddam. Dde-gan: Integrating a data-driven design evaluator into generative adversarial networks for desirable and diverse concept generation. *Journal of Mechanical Design*, 145(4):041407, 2023.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.