Fast Adversarial Training With Weak-to-Strong Spatial-Temporal Consistency in the Frequency Domain on Videos

Songping Wang, Hanqing Liu, Yueming Lyu, *Member, IEEE*, Xiantao Hu, Ziwen He, Wei Wang, *Member, IEEE*, Caifeng Shan, *Senior Member, IEEE*, and Liang Wang, *Fellow, IEEE*

***Abstract*—Adversarial Training (AT) has been shown to significantly enhance adversarial robustness via a min-max optimization approach. However, its effectiveness in video recognition tasks is hampered by two main challenges. First, fast adversarial training for video models remains largely unexplored, which severely impedes its practical applications. Specifically, most video adversarial training methods are computationally costly, with long training times and high expenses. Second, existing methods struggle with the trade-off between clean accuracy and adversarial robustness. To address these challenges, we introduce Video Fast Adversarial Training with Weak-to-Strong consistency (VFAT-WS), the first fast adversarial training method for video data. Specifically, VFAT-WS incorporates the following key designs: First, it integrates a straightforward yet effective temporal frequency augmentation (TF-AUG), and its spatial-temporal enhanced form STF-AUG, along with Fast Gradient Sign Method (FGSM) to boost training efficiency and robustness. Second, it devises a weak-to-strong spatial-temporal consistency regularization, which seamlessly integrates the simple TF-AUG and the more complex STF-AUG. Leveraging the consistency regularization, it steers the learning process from simple to complex augmentations. Both of them work together to achieve a better trade-off between clean accuracy and robustness. Extensive experiments on UCF-101 and HMDB-51 with both CNN and Transformer-based models demonstrate that VFAT-WS achieves great improvements in adversarial robustness and corruption robustness, while accelerating training by nearly 490%.**

***Index Terms*—Adversarial training, video recognition models, single-step adversarial attack, training efficiency.**

I. Introduction

DEEP Neural Networks (DNNs) have achieved significant success in various tasks [1], [2], [3], [4]. However, recent research indicates that DNNs are susceptible to carefully designed input samples with minor perturbations, which cause incorrect predictions [5], [6], [7], [8]. These input samples are collectively known as adversarial examples, which present significant challenges to security-critical applications [9], [10], [11], [12]. Video recognition represents an important computer vision subfield. However, existing video recognition models are typically built on deep neural networks, which have inherent vulnerabilities that can undermine the robustness of these models [13], [14], [15], [16]. Therefore, enhancing the adversarial robustness of video recognition models is particularly necessary for certain safety-critical tasks to ensure their secure operation.

To counter the threat of adversarial examples, a vast amount of research has been dedicated to developing various adversarial defense strategies [17], [18], [19]. Among these, adversarial training has been identified as one of the most effective methods for enhancing model robustness against adversarial threats. Recent scholarly work has been vigorously pursuing the development of an enhanced form of adversarial training [20], [21], [22]. Unfortunately, the field of video fast adversarial training is unexplored. Current methods suffer from high computational costs and long training time, which are major pain points in the industry and severely limit their application. What's worse, compared to images, videos have higher dimensions, which further exacerbate the challenge of reducing the time cost in adversarial training. In addition, these methods cannot effectively balance clean accuracy and adversarial robustness. These issues greatly hamper reliable application in safety-critical video recognition tasks.

When exploring the aforementioned issues, we have the following key observations: 1) Currently, existing adversarial training methods for videos involve multi-step iterative attacks, which lead to extensive training time. While replacing these iterative attacks with single-step attacks can improve training

Received 21 April 2025; revised 3 October 2025; accepted 17 December 2025. Date of publication 22 December 2025; date of current version 7 January 2026. This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project under Grant 2025ZD0123504, in part by the National Natural Science Foundation of China under Grant 62502200 and Grant 62402228, in part by Jiangsu Provincial Science and Technology Major Project under Grant BG2024042, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20251203 and Grant BK20240699. The associate editor coordinating the review of this article and approving it for publication was Dr. Roberto Caldelli. *(Corresponding authors: Yueming Lyu; Caifeng Shan.)*

Songping Wang, Yueming Lyu, and Caifeng Shan are with the School of Intelligence Science and Technology, Nanjing University, Suzhou 215163, China (e-mail: theone@buaa.edu.cn; ymlv@nju.edu.cn; caifeng.shan@gmail.com).

Hanqing Liu is with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: hqliu@buaa.edu.cn).

Xiantao Hu is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: huxiantao481@gmail.com).

Ziwen He is with the Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: ziwen.he@nuist.edu.cn).

Wei Wang and Liang Wang are with the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China (e-mail: wwang@nlpr.ia.ac.cn; wangliang@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TIFS.2025.3647231

efficiency, it severely compromises adversarial robustness. 2) An interesting study has demonstrated that deep learning models often leverage high-frequency details, which are imperceptible to the human eye, to achieve better performance [23]. However, we observe that adversarial noise frequently concentrates in these high-frequency details, while semantic information is primarily associated with low-frequency information [24]. This sensitivity to high-frequency details poses a significant threat to the model's adversarial robustness. 3) Human visual systems typically maintain a consistent perception of the same object despite minor variations. In contrast, DNNs often exhibit inconsistent perception of samples with weak augmentations compared to those with strong augmentations [25], with adversarial examples also being considered a special form of augmentation.

Based on the above insights, we have the following three key designs for video data: 1) We carefully design an efficient and straightforward TF-AUG to reduce the model's reliance on high-frequency details and encourage focus on low-frequency information. Integrating TF-AUG with FGSM, we propose **VFAT-W** (**V**ideo **F**ast **A**dversarial **T**raining with **W**eak temporal frequency augmentation), which accelerates adversarial training and enhances the model's adversarial robustness for video data. 2) Building on TF-AUG, we introduce STF-AUG that adds spatial augmentations to explore a more extensive perturbation space. Integrating STF-AUG with FGSM, **VFAT-S** (**V**ideo **F**ast **A**dversarial **T**raining with **S**trong spatiotemporal frequency augmentation) is proposed, which achieves better performance than VFAT-W by exploring videos' spatial-temporal characteristics more deeply. 3) We propose a weak-to-strong spatial-temporal consistency regulation to encourage consistent predictions for the same video with different perturbations, enhancing the model's adversarial robustness and corruption robustness.

Building on these designs, we integrate weak TF-AUG with strong STF-AUG via consistency regularization, enabling the model to focus on data's low-frequency information instead of overfitting to specific input perturbations. Based on this, we propose the enhanced VFAT-WS. VFAT-WS empowers the model with stronger consistency perception, thus enhancing overall performance and generalization capability. As shown in Figure 1, VFAT-S and VFAT-WS accelerate video adversarial training: VFAT-S has the fastest training speed, while VFAT-WS balances training speed and robust accuracy better. Our major contributions can be summarized as follows:

- We introduce VFAT-W and VFAT-S, the first fast adversarial training frameworks for videos, which achieve a better balance among the triple objectives of robustness, accuracy, and efficiency.
- Simple yet effective video augmentation techniques (TF-AUG and STF-AUG) are proposed to facilitate robust learning by suppressing the model's focus on high-frequency noise and emphasizing low-frequency spatiotemporal patterns.
- We design a novel weak-to-strong spatial-temporal consistency regularization that aligns predictions across augmentation and perturbation strengths, guiding the model toward stable representations.

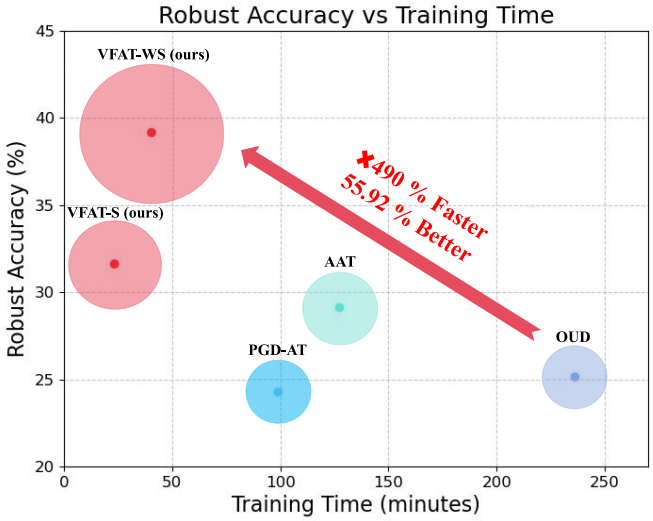

Fig. 1. AutoAttack accuracy and training time for various video adversarial training methods that utilize a 3D Pre-activation ResNet-18 architecture on the UCF-101 dataset are presented. The x-axis represents training time (where lower values signify higher efficiency), while the y-axis represents robust accuracy (where higher values signify greater robustness).

- Experiments show VFAT-WS achieves better robustness (+ 9.77% over prior art) and nearly 490% speedup, with consistent gains across various architectures.

The remainder of this paper is organized as follows. Section II reviews the related work in video attack and defense. Section III presents the proposed VFAT-WS framework, including the TF-AUG and STF-AUG augmentation strategies and the weak-to-strong consistency regularization. Experimental results and analysis are presented in Section IV. Finally, Section V concludes the whole paper.

II. Related Work

A. Adversarial Attack on Videos

Recent research indicates that video recognition models are vulnerable to adversarial attacks. Wei et al. [26] introduce 3D sparse perturbations into videos to generate adversarial examples under a white-box setting. Jiang et al. [27] expand the natural evolution strategy from images to videos to efficiently estimate adversarial gradients. Wei et al. [28] use a heuristic search on a subset of frames, employing an optimization-based approach to find suitable and minimal noise for selective key frames. However, this attack requires a high number of queries. To further streamline the process and enhance attack efficiency, Yan and Wei [29] propose an efficient reinforcement learning-based approach for key frame selection. Inspired by distillation techniques, they carefully design rewards to guide the agent in learning to select better key frames while maintaining a high attack success rate. To further eliminate temporal and spatial redundancy in videos, Wei et al. [13] design a novel video adversarial spatial-temporal focus (AstFocus) attack, which attacks key frames and key regions simultaneously from both inter-frames and intra-frames in the video. Deng et al. [30] propose a dual-branch neural network model to generate sparse adversarial video examples, achieving faster

and more effective attacks with minimal pixel perturbation, promoting robustness in industrial applications. Gao et al. [31] propose the ReToMe-VA framework, which generates imperceptible and highly transferable adversarial videos by optimizing perturbations in diffusion models' latent space and merging tokens across frames.

In addition, there have been relevant studies in the field of skeletal action recognition. BASAR [32] reveals that adversarial examples in skeletal action recognition under black-box scenarios widely exist both on and off the manifold. It verifies the effectiveness of these examples through perceptual research, thereby promoting the development of more robust classifiers. The BEAT [33] transforms vulnerable black-box classifiers into robust models via full Bayesian processing and adversarial example sampling based on the natural motion manifold. It is applicable to various classifiers, datasets, and attack scenarios. By introducing data manifolds, Diao et al. [34] uncover the widespread existence of adversarial examples in skeletal action recognition. It leverages MMAT to achieve efficient defense while identifying model vulnerabilities. TASAR [35] enhances the smoothness of pre-trained models and disrupts motion dynamics through dual Bayesian optimization, addressing the issue of weak adversarial transferability in S-HAR (Skeleton-based Human Activity Recognition) and establishing the first large-scale robust benchmark. However, these methods focus on skeletal action recognition, whereas this paper focuses on action recognition and is committed to improving its robustness for safety-critical applications.

B. Adversarial Defense on Videos

Research into video defense mechanisms is notably lacking, thereby exacerbating the grave security threats associated with various video attacks on security-critical video tasks. AdvIT [36] detects adversarial frames through temporal consistency but does not provide defense against adversaries. Spatial and temporal defenses [37] introduce a similar detector along with different defense strategies. However, these defenses are only tested in a black-box setting, leaving their resistance to stronger white-box attacks unclear. OUDefend [38] proposes an over-and-under complete restoration network for defending against adversarial videos. AAT [39] combines curriculum-style and adaptive adversarial training to enhance the robustness of video recognition models against variable attack budgets and types. However, these approaches also incur significant extended training durations and elevated training expenses. This is primarily attributed to the fact that the complexity of multi-step iterative video attacks, with their substantial computational demands, hinders the practical deployment of standard adversarial training approaches. Adversarial training with FGSM attack can alleviate the problem of slow training speed, but it may lead to catastrophic overfitting [40]. Moreover, since video data is more complex and does not take into account the spatial-temporal characteristics of video, it limits the robustness of the trained model. In order to solve the problems, We propose the first Video Fast Adversarial Training, which achieves better accuracy and robustness than traditional video adversarial training [38] with a shorter training time.

III. The Proposed Method

In this section, we first review the concept of adversarial training on videos. Then, we introduce a simple yet effective temporal frequency augmentation, TF-AUG and its enhanced vision, STF-AUG. Finally, we systematically describe VFAT-WS with weak-to-strong spatial-temporal consistency and give more details. The whole flowchart of VFAT-WS is shown in Figure 2.

A. Preliminaries: Adversarial Training on Videos

Compared to images, videos have an additional temporal dimension. A video input can be represented as $X \in R^{T\times W\times H\times C}$. The symbols T, W, H, C denote the number of video frames, frame width, frame height, and the number of video channels, respectively. Let Y denote the ground-truth and F_θ represent the video recognition model parameterized by θ. Adversarial training is described as a min-max optimization process, in which the inner part is to maximize the loss to generate adversarial examples, and the outer part is to input adversarial examples to minimize the loss. Adversarial training for video data is defined as follows:

$$\delta = Proj\Big(\alpha \cdot sign\big(\nabla_x J(F_\theta(X), Y)\big)\Big), \tag{1}$$

$$\min_\theta \max_{\delta\in\Delta} J(\mathrm{F}_\theta(X+\delta), Y), \tag{2}$$

where δ is the adversarial noise generated through inner-loop loss maximization, $Proj(\cdot)$ ensures that the updated adversarial noise remains within the effective range, α is the learning rate, $sign(\cdot)$ is the sign function, and $J(\cdot)$ is the cross-entropy loss function. Let θ_t be the parameters of the model at the t-th iteration. Our method generates adversarial noise through internal loss maximization of FGSM. The external minimization loss is consistent with standard AT:

$$\delta = Proj\Big(\delta_0 + \alpha \cdot sign\big(\nabla_\delta J(F_\theta(X+\delta_0), Y)\big)\Big), \tag{3}$$

$$\theta_{t+1} = \theta_t - \beta \cdot \nabla_{\theta_t} J\left(\left(F_{\theta_t}(X+\delta), Y\right)\right), \tag{4}$$

where δ_0 is a random uniform initialization noise, β is an appropriate learning rate.

B. Preliminaries: Frequency Components in Videos

To clarify the concepts of low-frequency information and high-frequency details, we follow previous work [41] and provide an operational definition based on Gaussian filtering. Low-frequency information refers to slowly varying, semantically rich structures in videos, such as target shapes, global motion, and smooth regions. This type of information can be preserved through Gaussian filtering:

$$X_{LF} = GaussianBlur(X, k), \tag{5}$$

where X_{LF} is low-frequency information of video frames, $GaussianBlur(\cdot)$ is a Gaussian filtering operation performed on the video frames, k represents the intensity of the filtering. **High-frequency details** corresponds to rapid changes in

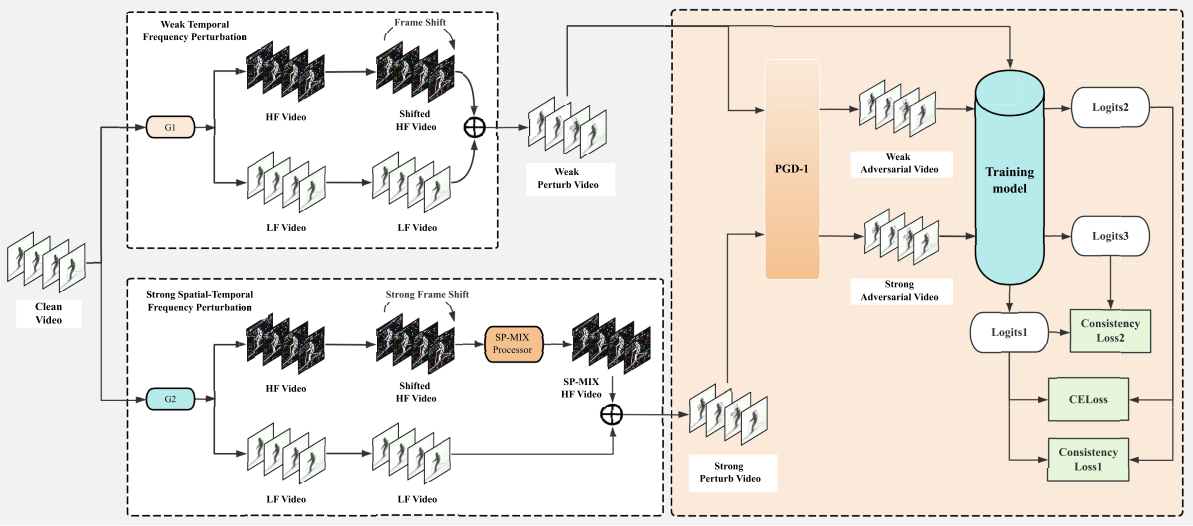

Fig. 2. The pipeline of the proposed VFAT-WS. VFAT-WS first employs Gaussian filtering (G1 and G2) to decompose a video into low-frequency and high-frequency components. A two-branch strategy is then adopted to generate perturbed videos: In the weak perturbation branch, TF-AUG performs temporal shifting on high-frequency frames, which are subsequently fused with low-frequency information; In the strong perturbation branch, STF-AUG builds on TF-AUG by further integrating spatio-temporal mixing operations, expanding the perturbation space. Adversarial examples in both branches are generated using FGSM with non-zero initialization. Finally, the model takes inputs from both branches and conducts joint optimization via weak-to-strong spatio-temporal consistency regularization, thereby enhancing the robustness-accuracy-efficiency triad.

videos, including edges, textures, and fine details, captured as residuals:

$$X_{HF} = X - X_{LF}, \tag{6}$$

where X_{HF} is high-frequency details of video frames. This decomposition separates semantic content (dominated by low-frequency components) from perceptual details (dominated by high-frequency components), forming the foundation of our spatiotemporal frequency augmentation design.

C. TF-AUG: Temporal Frequency Augmentation

Research has revealed a notable distinction in how humans and DNNs process information. Humans predominantly concentrate on low-frequency components, while DNNs exhibit heightened sensitivity to high-frequency details [41]. Since adversarial noise predominantly resides in these high-frequency details, diminishing the model's responsiveness to such noise emerges as a crucial strategy for bolstering adversarial robustness.

Inspired by these observations, we propose a simple and effective Temporal Frequency Augmentation (TF-AUG) in Figure 2. Here we use Gaussian filters to filter out the high-frequency details and low-frequency information of the video respectively. Then, in order to further enhance the diversity of high-frequency noise, we sequentially shift the corresponding high-frequency video frames by n frames, and merge them with the low-frequency information of the original video. The formulas are as follows:

$$Perturb(X, N, k) = Shift(X_{HF}, N) + X_{LF}, \tag{7}$$

where $Shift(X, N)$ denotes the N frame sequential shift operation on video frames, $Perturb(\cdot)$ represents the operation of temporal enhancement of high-frequency noise. After obtaining the frames with temporal enhancement of high-frequency details, we first obtain the initialization noise through uniform sampling. In order to improve the efficiency of adversarial training, we use FGSM to attack the video with temporal enhancement of high-frequency details. The formulas are as follows:

$$\delta_0 = Uniform(-\epsilon, \epsilon), \quad X_P = Perturb(X, N, k), \tag{8}$$

$$\delta^\star = Proj\left(\alpha \cdot sign\left(\nabla_x J(F_\theta(X_P + \delta_0), Y)\right)\right), \tag{9}$$

where $Uniform(\cdot)$ represents uniform sampling, $\epsilon = 8/255$ denotes the perturbation amplitude range for sampling. δ_0 represents initialization noise, and $\delta^\star$ represents the adversarial noise updated through FGSM attack.

To enhance the model's consistency under different perturbations, we introduce a consistency loss L_{TC}. Specifically, we leverage the cross-entropy loss L_{CE} and the temporal consistency loss L_{TC} to jointly update the model weights θ, which can be defined as follows:

$$L_{CE} = J\Big(F_\theta(Cat(X_p + \delta^\star, X_p), Cat(Y, Y))\Big), \tag{10}$$

$$L_{TC} = JSD\Big(F_\theta(X_p), F_\theta(X_p + \delta^\star)\Big), \tag{11}$$

where $Cat(\cdot)$ represents concat along the 0-th dimension, $JSD(\cdot)$ represents Jenson-Shannon divergence. Therefore, the weight updated in each iteration of the model can be expressed as:

$$\theta_{t+1} = \theta_t - \beta \cdot \nabla_{\theta_t}(\theta_t, \lambda * L_{TC} + (1 - \lambda) * L_{CE}), \tag{12}$$

where λ is a parameter to balance two losses.

TF-AUG enhances the model's robustness by manipulating the high-frequency details in video frames. Specifically, it performs displacement operations on high-frequency details to continuously expand the high-frequency perturbation space between different frames. By diversifying the high-frequency information across frames, TF-AUG suppresses the model's

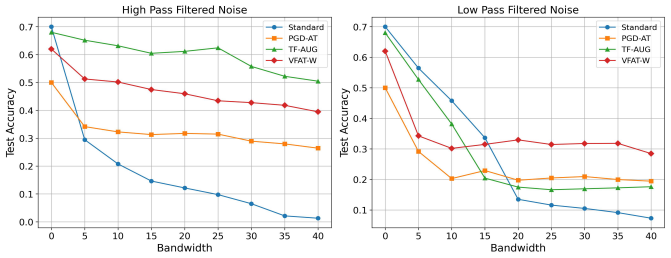

Fig. 3. The robustness of models trained with different strategies under additive noise and varying frequency distributions. We sample Gaussian noise, apply low-pass/high-pass filters and add it to videos. Results show that TF-AUG, VFAT-W, and PGD-AT improve tolerance to high-frequency noise, with TF-AUG showing the best performance. For low-frequency noise, TF-AUG and VFAT-W initially drop faster but demonstrate stronger robustness as bandwidth increases compared to standard training.

learning of associations between high-frequency details and video labels. This reduces the model's reliance on high-frequency information and forces it to focus more on stable low-frequency information, which is more critical for video semantics. As a result, the model is less disturbed by adversarial perturbations embedded in high-frequency details, thereby improving its overall robustness.

To further elucidate the mechanism of the proposed method, we conduct related experiments by adding low-pass or high-pass noise to models trained with different strategies. This is done to verify their robustness against high-frequency and low-frequency disturbances, as well as their focus on high-frequency details and low-frequency information. As shown in the left part of Figure 3, compared with standard training and PGD-AT, TF-AUG and VFAT-W effectively reduce the sensitivity to high-frequency details and enhance the model's robustness to high-frequency noise. In the right part of the figure, compared with standard training, TF-AUG and VFAT-W experience a sharper drop in accuracy in the initial stage with a smaller bandwidth. This indicates that they force the model to focus more on low-frequency information, thereby increasing its sensitivity to such information. Additionally, as the bandwidth of the low-pass filter increases, TF-AUG and VFAT-W exhibit greater robustness than standard training, highlighting their advantages in enhancing robustness. VFAT-W achieves a better balance of training speed and robustness against both high-frequency and low-frequency noise through the collaborative efforts of the TF-AUG and FGSM adversarial training strategies.

Moreover, Figure 4 shows that our strategies assist the model in enhancing its tolerance to strong adversarial noise existing in high-frequency details while maintaining its focus on semantically relevant low-frequency information.

D. Weak-to-Strong Spatial-Temporal Consistency

We introduce a weak-to-strong perturbation pipeline (Figure 2) to fully broaden the perturbation space. The weak-to-Strong Spatial-Temporal Consistency is designed to ensure consistency between predictions of videos with weak temporal perturbations and those with strong spatial-temporal adversarial perturbations. This consistency regularization leverages the model's higher accuracy on weakly perturbed data to guide predictions on strongly perturbed videos. By incorporating additional information and mitigating confirmation bias, our method progressively guides the model from simple to complex perturbation learning while enhancing perceptual consistency and robustness.

Fig. 4. From top to bottom: Grad-CAM visualizations [42] of PGD-1, VFAT-S, and VFAT-WS are presented, demonstrating their responses to corrupted inputs under the scrutiny of AutoAttack. Among them, PGD-1 represents the use of FGSM as the adversarial training strategy. The 3D Pre-activation ResNet-18 serves as the underlying architecture for these evaluations.

TABLE I
SPATIAL-TEMPORAL MIX OPERATIONS

Name	Subtle MixUp	Temporal CutMix	Temporal CutMixUp	3D CutMix	3D CutMixUp
Spatial				✓	✓
Temporal	✓	✓	✓	✓	✓

The effectiveness stems from its alignment with video's spatiotemporal structure and frequency dynamics: weak augmentation (TF-AUG) applies only temporal perturbations for slight semantic enhancement, and strong augmentation (STF-AUG) to add spatial perturbations for expanding the spatiotemporal perturbation space. These progressive designs help form consistent representations and suppress the model's focus on unstable high-frequency details.

Specifically, we combine the weak TF-AUG pipeline and the strong STF-AUG pipeline to form a new framework. The TF-AUG is introduced in the previous section, while the strong STF-AUG is an enhanced version of the weak TF-AUG. It further explores the spatial information of high-frequency details and increases the degree of temporal enhancement, including the size of the convolution kernels and the offset of high-frequency frame details. Specifically, STF-AUG utilizes Spatial-Temporal Mix operations we use in Table I to process high-frequency frame details, thereby further expanding the perturbation source. The formulas are as follows:

$$Perturb\_W(X) = Perturb(X, N_1, k_1), \tag{13}$$

$$Perturb\_S(X) = STMix(Perturb(X, N_2, k_2), \gamma), \tag{14}$$

where $Perturb\_W(\cdot)$ represents the weak perturbation operation, $Perturb\_S(\cdot)$ represents the strong perturbation operation, $N_2 > N_1 \geq 0$, $k_2 > k_1 > 0$, $STMix(\cdot)$ represents Spatial-Temporal Mix operation, γ is the mixing coefficient. Then the two generated video frames enhanced by strong and

weak perturbations operation are introduced into uniformly sampled initialization noise δ_0, and the sampling range of the disturbance size is $(-\epsilon, \epsilon)$. In order to save time of maximizing loss within adversarial training, we still only use FGSM to attack video frames with enhanced strong and weak perturbations operation, which can be expressed as:

$$\delta_W = Proj\Big(\alpha \cdot sign\big(\nabla_{\delta_0} J(F_\theta(Perturb\_W(X) + \delta_0), Y)\big)\Big), \tag{15}$$

$$\delta_S = Proj\Big(\alpha \cdot sign\big(\nabla_{\delta_0'} J(F_\theta(Perturb\_S(X) + \delta_0'), Y)\big)\Big), \tag{16}$$

where δ_S and δ_W respectively correspond to the adversarial noise obtained by performing FGSM attack after strong and weak perturbation operation. Then the final loss L_{Total} that we externally minimize consists of cross entropy loss L_{CE}, weak consistency loss L_{TCW} and strong consistency loss L_{TCS}. The formulas are as follows:

$$L_{CE} = \frac{1}{N}\sum_{i=1}^{N} J\Big(F_\theta(Cat(X_{WP_i} + \delta_{W_i}, X_{WP_i})), Cat(Y_i, Y_i)\Big), \tag{17}$$

$$L_{TCW} = \frac{1}{N}\sum_{i=1}^{N} JSD\Big(F_\theta(X_{WP_i} + \delta_{W_i}), F_\theta(X_{WP_i})\Big), \tag{18}$$

$$L_{TCS} = \frac{1}{N}\sum_{i=1}^{N} JSD\Big(F_\theta(X_{SP_i} + \delta_{S_i}), F_\theta(X_{WP_i})\Big), \tag{19}$$

where N represents the batch size of multiple input videos, X_{WP_i} represents the i-th weak perturbation video by weak perturbation operation and X_{SP_i} represents the i-th strong perturbation video by strong perturbation operation. Then we set two hyperparameters to control the weight between different losses and get the final loss, which can be expressed as:

$$L_{Total} = \lambda * (L_{TCW} + \mu * L_{TCS}) + (1 - \lambda) * L_{CE}, \tag{20}$$

where μ is a hyperparameter that adjusts L_{TCW} and L_{TCS}, λ is a hyperparameter that adjusts L_{CE} and L_{TC} loss. Eq. (20) defines the overall objective of VFAT-WS. The cross-entropy loss $\mathcal{L}_{\text{CE}}$ (Eq. (17)) ensures basic classification accuracy under perturbations. The weak consistency loss $\mathcal{L}_{\text{TCW}}$ (Eq. (18)) enforces prediction consistency between a weakly augmented video and its adversarial version, providing a stable reference for learning. The strong consistency loss $\mathcal{L}_{\text{TCS}}$ (Eq. (19)) aligns the prediction of the strongly perturbed adversarial sample with the weakly augmented one, anchoring the model to semantic content under heavy perturbations. By balancing these terms with weights λ and μ, VFAT-WS enables a progressive, robust learning process. This weak-to-strong consistency design effectively enhances the robustness-accuracy-efficiency triad, as validated by extensive experiments.

Through weak-to-strong consistency regularization, the model becomes more adaptable to varying noise levels. Our proposed method effectively enhances the model's adversarial robustness and consistency. Specifically, our method ensures that the model's output for a sample enhanced by a weak adversarial perturbation remains consistent with the output for the same sample enhanced by a strong adversarial perturbation.

Algorithm 1 VFAT-WS Algorithm

Input: Input video (X, Y), model parameters θ, perturbation budget ϵ, weak temporal augmentation $\mathcal{T}_{\text{TF-AUG}}$, strong spatiotemporal augmentation $\mathcal{T}_{\text{STF-AUG}}$, learning rate β, total training epochs T.

Output: Robust model parameters $\theta^\star$

while $t < T$ **do**
  Apply weak TF-AUG $X_{wt} \leftarrow \mathcal{T}_{\text{TF-AUG}}(X)$▷ Eq. (13)
  Apply strong STF-AUG $X_{st} \leftarrow \mathcal{T}_{\text{STF-AUG}}(X)$ ▷ Eq. (14)
  Initialize perturbation $\delta_{wt}, \delta_{st} \leftarrow$ Uniform$(-\epsilon, \epsilon)$
  Generate weak adversarial video $\hat{X}_{wt} \leftarrow X_{wt} + \delta_{wt}$ ▷ Eq. (15)
  Generate strong adversarial video $\hat{X}_{st} \leftarrow X_{st} + \delta_{st}$ ▷ Eq. (16)
  Compute cross entropy loss L_{CE}▷ Eq. (17)
  Compute weak consistency loss L_{TCW}▷ Eq. (18)
  Compute strong consistency loss L_{TCS}▷ Eq. (19)
  Compute total loss L_{Total}▷ Eq. (20)
  Update parameters $\theta_{t+1} \leftarrow \theta_t - \beta \cdot \nabla_{\theta_t} L_{total}$
end while
return $\theta^\star$

The VFAT-WS framework we introduced continues to achieve favorable effects of adversarial defense even in the presence of previously unseen strong attacks, while maintaining high accuracy on clean data and high training efficiency. The whole algorithm is summarized in Algorithm 1.

IV. Experiments

In this section, we conduct a comprehensive evaluation of our proposed methods. First, we detail the experimental setup. Next, we compare our methods with state-of-the-art adversarial training techniques under two prominent adaptive attacks AutoAttack [43] and PGD [44], using varying attack budgets. We also assess performance against nine distinct types of video adversarial attacks. Furthermore, we perform in-depth experiments on the hyperparameters of our method and conduct ablation studies on different modules to highlight their contributions. Finally, we present additional results to validate the effectiveness of our approach.

A. Experimental Setup

1) Datasets and Recognition Models: We conduct our experiments using the widely used UCF-101 dataset [45], which consists of 13,320 videos spanning 101 action classes and HMDB-51 [46], which consists of 7,000 videos across 51 action categories for evaluation. Following [26], we resize dimensions of video frames to 112×112 and uniformly sample each video. Following [47], we adopt classic 3D ResNet-18 [48] and 3D Pre-activation ResNet-18 [48] as target models. To further validate the generality of our method, we also conduct experiments on the Video Swin Transformer [49].

2) Baselines: We compare our method with standard training and other advanced adversarial video training methods, including PGD-AT [44], OUD [38], and AAT [39]. To further expand the evaluation scope, we extended several mainstream methods from the image domain to video data for comprehensive evaluation, including FAST-AT [47], ATAS [50] and

FGSM-PKG [51]. For a fair comparison, we use the PGD or FGSM attack as the inner attack in all methods. For our approach, we use two variants: (1) VFAT-S: represents our proposed STF-AUG in collaboration with the FGSM adversarial training method (2) VFAT-WS: indicates our final method.

3) Implementation Details: Following the experimental settings of the classic fast adversarial training method [47], we use an SGD optimizer with a weight decay of 5e-4, a batch size of 60, and a momentum of 0.9. The maximum learning rates are set to 0.2 for 3D ResNet, and 0.09 for 3D Pre-activation ResNet and Video Swin Transformer (to balance training speed and stability). Cyclic learning rates are employed to aid convergence and reduce training time. Models are trained for 40 epochs. Following common range for video adversarial training [39], we set the perturbation strength to $\epsilon = 8/255$ to ensure the appropriateness of the perturbation magnitude. For data preprocessing, we follow the general conventions of experimental settings in the video recognition field: we use the HMDB51 and UCF101 data processing pipelines; for frame sampling, we adopt a uniform sampling method; for data augmentation, we use lightweight video augmentation methods such as Group Random Horizontal Flip and Group Scale, which are consistent with the conventions in the video recognition. All hyperparameters are kept as consistent as possible across all models and datasets to ensure fair and unbiased comparisons. Experiments are conducted on NVIDIA RTX A6000 GPUs with 48 GB memory. The experimental results in all tables are run five times with different random seeds to calculate the standard deviation and ensure stability.

B. Main Experiment

1) Quantitative Results With Variable Attack Budget: To demonstrate the effectiveness of our proposed method, we conduct a comprehensive evaluation comparing our approach with several baselines on the UCF-101 and HMDB-51 datasets using multiple models, including 3D ResNet-18, 3D Pre-Activation ResNet-18, and Video Swin Transformer.We employ AutoAttack and PGD as attack methods with budgets of $\epsilon = \{10/255, 12/255, 14/255, 16/255\}$. Moreover, RA-PGD represents the robust accuracy (RA) under PGD attacks. As shown in Table II and Table III, our methods, VFAT-S and VFAT-WS, achieve significant improvements in both clean accuracy and adversarial robustness. In terms of clean accuracy, VFAT-S and VFAT-WS consistently outperform other methods. For example, on UCF-101 with 3D ResNet-18, VFAT-S achieves 45.27% accuracy, surpassing PGD-AT (42.53%) and OUD (42.42%). With Video Swin Transformer, VFAT-S reaches 63.66%, higher than PGD-AT (40.22%) and OUD (39.12%). On HMDB-51 with 3D ResNet-18, VFAT-WS attains 36.90% accuracy, outperforming PGD-AT (15.72%) and OUD (20.81%).

Regarding robustness, VFAT-S and VFAT-WS achieve the highest accuracy under all attack strengths. For instance, on HMDB-51 with 3D ResNet-18, VFAT-WS achieves 25.39% accuracy under RA-AutoAttack at $\epsilon = 16/255$, higher than PGD-AT (10.63%) and OUD (11.73%). Under RA-PGD, VFAT-WS achieves 28.49% accuracy, surpassing PGD-AT (24.43%) and OUD (23.39%). Additionally, VFAT-S and VFAT-WS have effectively shorter training times. For example, on UCF-101 with 3D ResNet-18, VFAT-S trains in 24 minutes, 325% faster than PGD-AT (102 minutes) and 854% faster than OUD (229 minutes). VFAT-WS trains in 40 minutes, 472% faster than OUD. Our methods thus enhance robustness and reduce training time while maintaining high clean accuracy, demonstrating their efficiency and effectiveness.

2) Quantitative Results Against Multiple Adversarial Video Attacks: To thoroughly evaluate our method, we conduct comprehensive tests using a diverse set of video attack methods, as shown in Table IV and Table V. Specifically, we test against nine unseen types of video attacks, including ROA [52], SPA [53], AF [54], Frame Border, Frame saliency (one-shot), Frame saliency (iterative) [55], SparseAdv [26], Masked PGD [39], and TT attack [56]. As shown in Table IV, VFAT-WS achieves the best average performance of 41.46% against various strong video attacks on Video Swin Transformer, outperforming OUD by 9.85% and AAT by 14.64%. Additionally, in Table V, VFAT-WS achieves the best average performance of 32.2% against various attacks on 3D Pre-Activation ResNet-18, surpassing OUD by 12.88% and AAT by 7.13%. To further evaluate robustness, we conduct comparative tests using corruption benchmarks [57]. The results on UCF-101 with 3D Pre-activation ResNet-18, shown in Figure 5, demonstrate that VFAT-WS exhibits higher corruption robustness compared to adversarial training with FGSM (PGD-1) and OUD, further confirming the effectiveness of our proposed method.

3) Comparison of Positioning Between VFAT-S and VFAT-WS: While both VFAT-S and VFAT-WS are fast adversarial training methods, they serve different priorities. VFAT-S focuses on **speed**: it uses STF-AUG with FGSM, without consistency regularization. On UCF-101 (3D ResNet-18), it trains in just 24 minutes—over 4× faster than PGD-AT (102 min)—making it ideal for edge deployment and real-time applications. Though its robustness (34.29% under RA-AutoAttack) is slightly lower than VFAT-WS, it still outperforms traditional methods. In contrast, VFAT-WS prioritizes **robustness** by introducing weak-to-strong spatio-temporal consistency (L_{TCW}, L_{TCS}) between TF-AUG and STF-AUG. It achieves 35.71% robust accuracy under RA-AutoAttack and 41.05% against unseen attacks (e.g., TT), with training in 40 minutes—still 5.7× faster than OUD (229 min). This makes it suitable for safety-critical tasks like surveillance. Thus, VFAT-S and VFAT-WS are **complementary**: one optimizes for speed, the other for robustness, together covering diverse real-world needs. Experimental results validate this design rationale.

C. Hyperparameter Studies

In this subsection, we investigate the impact of hyperparameters on the validation set of UCF-101 using 3D ResNet-18. We focus on how these hyperparameters affect clean accuracy and adversarial robustness under PGD and AutoAttack with an attack budget of $\epsilon = 16/255$.

Figure 6 .(A) shows the parameter tuning results for VFAT-WS with different values of λ, defined in Eq. (12). The y-axis represents accuracy in percentage, while the x-axis

TABLE II

ACCURACY(%) OF DIFFERENT ADVERSARIAL TRAINING METHODS UNDER DIFFERENT PERTURBATION STRENGTH ϵ AGAINST RA-AUTOATTACK AND RA-PGD WITH DIFFERENT MODELS AS THE BACKBONE ON UCF-101

Dataset: UCF-101 & Model: 3D ResNet-18										
Method	CLEAN(%)	RA-AutoAttack(%)				RA-PGD(%)				Time
		(ϵ = 10/255)	(ϵ = 12/255)	(ϵ = 14/255)	(ϵ = 16/255)	(ϵ = 10/255)	(ϵ = 12/255)	(ϵ = 14/255)	(ϵ = 16/255)	(min)
Standard	51.10±0.20	2.418±0.16	1.758±0.15	1.209±0.15	0.769±0.12	1.209±0.15	0.659±0.10	0.110±0.16	0.110±0.16	–
PGD-AT	42.53±0.29	30.44±0.34	28.24±0.45	26.48±0.33	24.51±0.34	39.78±0.26	39.56±0.26	38.79±0.37	38.24±0.34	102
OUD	42.42±0.22	30.99±0.34	29.45±0.25	27.80±0.26	26.15±0.44	38.79±0.25	37.69±0.26	36.92±0.29	36.70±0.26	229
AAT	41.10±0.31	32.09±0.26	30.44±0.29	27.69±0.40	26.15±0.30	38.79±0.28	38.24±0.39	37.36±0.54	36.26±0.34	130
FAST-AT	38.46±0.25	27.91±0.38	25.05±0.40	23.41±0.26	20.88±0.53	33.30±0.25	32.64±0.34	31.21±0.33	30.55±0.41	21
ATAS	40.56±0.31	28.93±0.30	25.91±0.34	23.36±0.38	21.00±0.34	35.60±0.26	34.11±0.36	32.60±0.33	31.11±0.39	28
FGSM-PGK	44.09±0.29	31.44±0.31	28.98±0.34	27.02±0.29	25.04±0.26	38.22±0.28	37.76±0.38	37.29±0.33	36.98±0.32	31
VFAT-S	**45.27**±0.27	34.29±0.36	31.87±0.33	29.89±0.38	28.57±0.41	41.54±0.25	40.0±0.30	39.01±0.32	38.46±0.30	24
VFAT-WS	44.51±0.25	**35.71**±0.34	**34.84**±0.37	**32.86**±0.25	**31.32**±0.36	**42.42**±0.26	**41.98**±0.26	**41.10**±0.25	**40.99**±0.32	40
Dataset: UCF-101 & Model: 3D Pre-Activation ResNet-18										
Method	CLEAN(%)	RA-AutoAttack(%)				RA-PGD(%)				Time
		(ϵ = 10/255)	(ϵ = 12/255)	(ϵ = 14/255)	(ϵ = 16/255)	(ϵ = 10/255)	(ϵ = 12/255)	(ϵ = 14/255)	(ϵ = 16/255)	(min)
Standard	50.44±0.23	6.154±0.24	3.626±0.27	2.418±0.10	1.978±0.19	5.604±0.19	4.176±0.15	3.187±0.16	2.527±0.15	–
PGD-AT	40.00±0.20	24.29±0.37	21.54±0.29	19.12±0.38	16.70±0.34	36.15±0.29	35.49±0.31	33.85±0.32	33.08±0.26	99
OUD	41.43±0.25	25.16±0.33	23.08±0.36	20.66±0.46	18.13±0.36	40.00±0.26	38.90±0.29	38.79±0.30	38.68±0.33	236
AAT	40.55±0.34	29.12±0.41	26.26±0.32	23.41±0.36	21.65±0.34	38.35±0.28	37.91±0.29	36.81±0.36	36.26±0.42	127
FAST-AT	36.70±0.33	18.24±0.40	15.49±0.25	12.86±0.20	10.44±0.41	32.53±0.26	30.11±0.43	28.35±0.47	27.14±0.42	20
ATAS	37.00±0.34	22.18±0.33	20.31±0.38	17.36±0.34	14.36±0.34	33.82±0.30	32.51±0.33	31.78±0.38	30.18±0.37	27
FGSM-PGK	41.42±0.25	24.67±0.28	23.24±0.25	20.33±0.36	18.24±0.40	37.42±0.34	35.07±0.26	34.80±0.34	33.20±0.25	30
VFAT-S	**44.73**±0.25	31.65±0.46	29.23±0.42	27.03±0.33	25.71±0.35	40.55±0.34	39.34±0.41	37.58±0.26	37.14±0.29	23
VFAT-WS	44.18±0.25	**39.23**±0.29	**38.46**±0.26	**38.02**±0.29	**37.58**±0.34	**44.29**±0.26	**44.18**±0.33	**43.96**±0.30	**43.4**±0.31	40
Dataset: UCF-101 & Model: Video Swin Transformer										
Method	CLEAN(%)	RA-AutoAttack(%)				RA-PGD(%)				Time
		(ϵ = 10/255)	(ϵ = 12/255)	(ϵ = 14/255)	(ϵ = 16/255)	(ϵ = 10/255)	(ϵ = 12/255)	(ϵ = 14/255)	(ϵ = 16/255)	(min)
Standard	59.89±0.53	3.30±0.14	3.08±0.29	2.42±0.27	1.98±0.34	12.75±0.46	11.65±0.53	10.11±0.24	7.58±0.38	–
PGD-AT	40.22±0.50	27.14±0.50	25.82±0.45	23.52±0.30	22.09±0.34	43.74±0.72	42.31±0.35	40.00±0.56	39.34±0.45	372
OUD	39.12±0.48	23.74±0.49	23.30±0.46	20.99±0.46	20.77±0.37	43.08±0.54	42.31±0.75	39.89±0.42	38.13±0.66	518
AAT	42.42±0.67	27.03±0.39	25.82±0.69	23.30±0.81	20.55±0.58	42.20±0.65	39.67±0.38	37.80±0.72	36.37±0.26	411
FAST-AT	40.11±0.77	25.38±0.39	23.08±0.47	22.53±0.64	21.21±0.48	38.02±0.51	36.81±0.41	35.93±0.32	34.62±0.65	60
ATAS	39.02±0.32	26.62±0.61	25.11±0.47	23.07±0.56	22.40±0.36	39.69±0.49	38.38±0.50	37.62±0.66	35.80±0.71	79
FGSM-PGK	44.89±0.87	27.18±0.64	25.64±0.75	24.67±0.65	23.64±0.42	42.20±0.71	40.11±0.77	38.36±0.55	37.07±0.74	88
VFAT-S	**63.66**±0.42	33.96±0.38	28.90±0.73	26.70±0.68	23.74±0.47	48.46±0.35	43.96±0.76	40.44±0.70	36.26±0.62	72
VFAT-WS	61.65±0.47	**46.15**±0.59	**42.42**±0.54	**39.23**±0.43	**37.25**±0.47	**49.01**±0.64	**45.38**±0.67	**44.73**±0.39	**41.21**±0.59	115

indicates varying values of λ. A well-calibrated λ helps the model capture patterns from both weakly and strongly perturbed instances. We find that λ impacts clean accuracy and adversarial robustness. Increasing λ generally improves overall performance, highlighting the positive effect of the weak-to-strong consistency loss. The optimal trade-off between clean accuracy and adversarial robustness is achieved when $\lambda = 0.8$, which we use for subsequent experiments.

Similarly, Figure 6.(B) explores the impact of μ, which balances L_{TCW} and L_{TCS} in Eq. (12). We set $\mu = 0.8$ for the best performance. Figures 6.(C) and 6.(D) examine the effects of the offset frame count N and Gaussian blur kernel size k. Figure 6.(C) shows that increasing N improves both clean accuracy and robustness against PGD and AutoAttack, with optimal performance at $N = 6$. Figure 6.(D) indicates that when $k = 3$, Gaussian filtering not only effectively suppresses high-frequency adversarial noise but also retains the low-frequency semantic features of videos to the greatest extent, achieving the optimal balance between robustness and accuracy. However, when the kernel size is too large, performance degradation occurs due to over-smoothing. Thus, We set $N = 6$ and $k = 3$ to balance clean accuracy and robustness.

D. Ablation Study

1) Ablation Study On Different Modules Of VFAT-WS: All ablation experiments are conducted using 3D ResNet-18 on the UCF-101 dataset, employing AutoAttack with

TABLE III

ACCURACY(%) OF DIFFERENT ADVERSARIAL TRAINING METHODS UNDER DIFFERENT PERTURBATION STRENGTH ϵ AGAINST RA-AUTOATTACK AND RA-PGD WITH DIFFERENT MODELS AS THE BACKBONE ON HMDB-51

Dataset: HMDB-51 & Model: 3D ResNet-18										
Method	CLEAN(%)	RA-AutoAttack(%)				RA-PGD(%)				Time
		(ϵ = 10/255)	(ϵ = 12/255)	(ϵ = 14/255)	(ϵ = 16/255)	(ϵ = 10/255)	(ϵ = 12/255)	(ϵ = 14/255)	(ϵ = 16/255)	(min)
Standard	37.93±0.36	3.100±0.17	2.14±0.25	1.770±0.24	1.330±0.22	7.090±0.34	6.270±0.30	6.200±0.22	5.310±0.28	–
PGD-AT	15.72±0.37	11.81±0.43	11.51±0.35	11.14±0.39	10.63±0.49	26.42±0.48	25.31±0.66	25.17±0.60	24.43±0.57	61
OUD	20.81±0.52	13.43±0.29	12.47±0.31	12.47±0.39	11.73±0.31	25.83±0.60	24.87±0.66	24.21±0.38	23.39±0.47	137
AAT	18.08±0.41	11.68±0.51	12.23±0.48	12.09±0.32	11.64±0.58	27.44±0.43	27.66±0.54	26.03±0.45	26.62±0.54	77
FAST-AT	12.62±0.58	8.410±0.44	8.340±0.38	7.900±0.40	7.750±0.43	24.32±0.39	24.06±0.58	23.36±0.39	22.79±0.43	12
ATAS	15.04±0.44	9.990±0.38	9.640±0.39	9.100±0.32	9.010±0.42	24.50±0.36	24.43±0.45	24.13±0.59	23.59±0.61	16
FGSM-PGK	17.58±0.42	12.52±0.43	12.16±0.37	11.72±0.47	10.16±0.39	26.84±0.74	26.12±0.51	25.79±0.55	24.91±0.55	20
VFAT-S	35.18±0.74	21.99±0.62	20.81±0.55	19.93±0.51	18.60±0.54	30.63±0.57	29.89±0.75	27.82±0.71	26.35±0.64	14
VFAT-WS	**36.90**±0.43	**29.15**±0.46	**28.19**±0.40	**27.45**±0.68	**25.39**±0.46	**31.59**±0.38	**30.70**±0.29	**29.37**±0.53	**28.49**±0.48	24
Dataset: HMDB-51 & Model: 3D Pre-Activation ResNet-18										
Method	CLEAN(%)	RA-AutoAttack(%)				RA-PGD(%)				Time
		(ϵ = 10/255)	(ϵ = 12/255)	(ϵ = 14/255)	(ϵ = 16/255)	(ϵ = 10/255)	(ϵ = 12/255)	(ϵ = 14/255)	(ϵ = 16/255)	(min)
Standard	37.12±0.30	1.480±0.19	1.110±0.16	0.810±0.21	0.740±0.16	5.680±0.28	4.800±0.29	4.430±0.32	3.990±0.29	–
PGD-AT	15.42±0.56	11.59±0.47	11.66±0.51	11.59±0.54	11.44±0.45	27.31±0.80	27.08±0.58	27.31±0.47	26.64±0.54	59
OUD	11.07±0.60	8.190±0.47	7.900±0.48	7.750±0.40	7.380±0.36	27.60±0.62	27.23±0.61	27.08±0.62	26.79±0.72	138
AAT	15.20±0.71	11.37±0.44	11.29±0.57	10.55±0.41	10.55±0.55	27.45±0.69	27.37±0.48	27.15±0.67	26.53±0.56	75
FAST-AT	14.75±0.47	10.04±0.32	9.990±0.39	9.910±0.44	9.790±0.36	24.28±0.50	23.91±0.55	23.76±0.73	23.47±0.50	12
ATAS	14.83±0.60	10.70±0.43	10.63±0.49	10.41±0.42	9.820±0.46	26.16±0.85	25.64±0.58	25.61±0.51	25.15±0.78	16
FGSM-PGK	15.00±0.42	12.55±0.33	12.28±0.34	12.07±0.32	12.00±0.31	27.90±0.41	27.60±0.46	27.45±0.45	26.85±0.42	19
VFAT-S	35.72±0.66	27.23±0.67	25.17±0.50	23.99±0.52	22.95±0.48	30.18±0.43	28.63±0.36	26.86±0.70	26.27±0.20	14
VFAT-WS	**36.68**±0.68	**28.27**±0.46	**27.16**±0.36	**26.57**±0.41	**25.46**±0.44	**30.41**±0.29	**29.30**±0.41	**28.63**±0.50	**27.68**±0.41	24
Dataset: HMDB-51 & Model: Video Swin Transformer										
Method	CLEAN(%)	RA-AutoAttack(%)				RA-PGD(%)				Time
		(ϵ = 10/255)	(ϵ = 12/255)	(ϵ = 14/255)	(ϵ = 16/255)	(ϵ = 10/255)	(ϵ = 12/255)	(ϵ = 14/255)	(ϵ = 16/255)	(min)
Standard	49.59±0.51	1.550±0.18	1.330±0.26	1.110±0.25	1.030±0.23	7.530±0.40	6.130±0.46	5.460±0.29	4.580±0.29	–
PGD-AT	18.52±0.58	11.81±0.44	11.51±0.37	11.22±0.34	11.00±0.32	32.18±0.30	31.66±0.64	31.00±0.58	30.48±0.56	220
OUD	21.70±0.32	14.54±0.57	13.73±0.36	12.55±0.31	12.47±0.60	32.40±0.60	31.73±0.46	31.07±0.66	30.11±0.69	316
AAT	22.73±0.43	14.24±0.38	12.10±0.34	11.88±0.47	10.48±0.30	36.53±0.57	34.83±0.54	33.87±0.59	32.62±0.68	244
FAST-AT	18.55±0.38	11.39±0.32	10.28±0.41	9.690±0.46	9.400±0.37	30.18±0.52	29.23±0.36	29.08±0.50	27.75±0.53	36
ATAS	19.49±0.51	11.85±0.34	11.66±0.37	11.33±0.39	10.78±0.27	31.22±0.75	30.58±0.58	29.21±0.51	28.70±0.43	47
FGSM-PGK	22.69±0.55	13.34±0.29	13.15±0.27	12.72±0.25	12.19±0.27	33.76±0.51	32.30±0.57	31.42±0.67	30.73±0.39	53
VFAT-S	47.23±0.50	34.91±0.32	33.14±0.54	30.85±0.65	29.37±0.48	37.56±0.50	35.87±0.60	33.65±0.46	33.21±0.36	43
VFAT-WS	**48.04**±0.44	**35.72**±0.42	**34.61**±0.49	**32.69**±0.45	**31.00**±0.48	**39.26**±0.46	**37.27**±0.53	**35.87**±0.43	**34.54**±0.61	69

ϵ = 16/255 as the attack method to evaluate robust accuracy. As shown in Figure 7.(A), the baseline Clean method achieves a robust accuracy of about 1%, while TF-AUG improves this to around 4.7%, indicating its positive contribution to model robustness. PGD-1 adversarial training further increases robust accuracy to approximately 20%. Combining TF-AUG and PGD-1 in VFAT-W boosts robust accuracy to around 27%. Building on this, VFAT-WS achieves the highest robust accuracy of about 32%, demonstrating that STF-AUG and consistency regularization enhance the model's ability to handle adversarial attacks. However, VFAT-WSC shows a slight decrease to around 30%, suggesting that weakly perturbed samples are more effective for consistency constraints than clean samples, as they better guide the learning of complex augmentation patterns and enhance robustness.

2) Ablation Study on Different Consistency Constraint Losses: The experimental results in Figure 7.(B) provide an ablation analysis of robust accuracy using different types of losses (Cosine similarity, KL divergence, and JS divergence) as the consistency constraint loss. The results show that the JSD method achieves the highest robust accuracy of approximately 32%, indicating its advantage in capturing distributional differences. KL achieves a robust accuracy of around 30%, which, although lower than JSD, is still superior to the Cosine similarity. In contrast, the Cosine similarity has the lowest robust accuracy of about 28%, likely because Cosine focuses only on the directional similarity of distributions and fails to

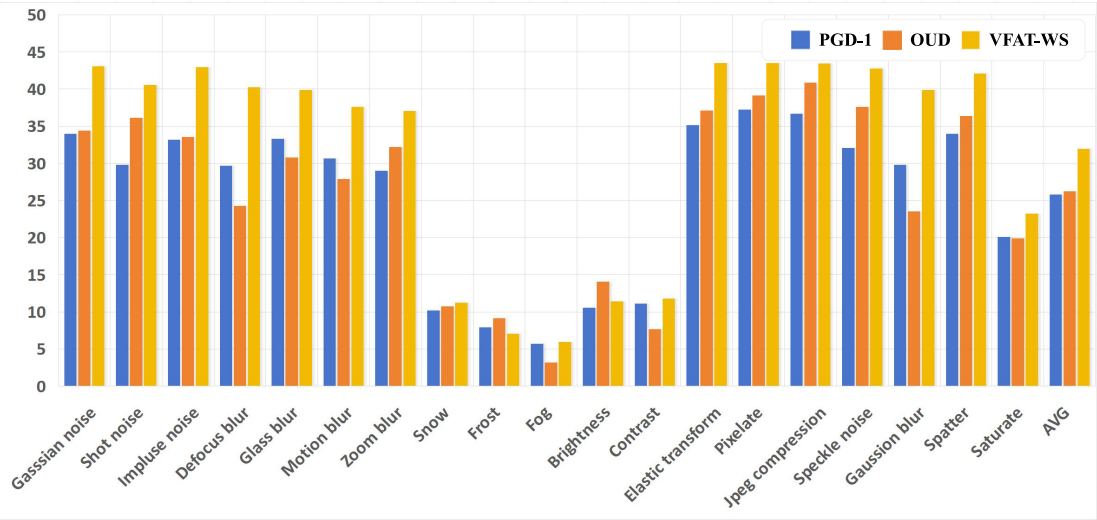

Fig. 5. Accuracy (%) on each corruption type of UCF-101 dataset where the x-axis labels denote different corruption types. Experiments are evaluated on 3D Pre-activation ResNet-18 trained under PGD-1, OUD, VFAT-WS, respectively.

TABLE IV

ACCURACY(%) OF DIFFERENT ADVERSARIAL TRAINING METHODS UNDER DIFFERENT VIDEO ATTACK METHODS WITH DIFFERENT MODELS AS THE BACKBONE ON UCF-101

Method	ROA	SPA	AF	Frame Border	Frame saliency (one shot)	Frame saliency (iterative)	SparseAdv	Masked PGD	TT	Average
Dataset: UCF-101 & Model: 3D ResNet-18										
Standard	15.27±0.28	0.220±0.16	1.099±0.26	0.110±0.13	12.20±0.12	12.53±0.17	1.650±0.13	3.960±0.15	5.600±0.28	5.849±0.19
PGD-AT	35.38±0.29	27.36±0.27	24.40±0.26	35.38±0.22	35.60±0.27	35.82±0.32	29.56±0.22	35.93±0.21	33.74±0.28	32.58±0.26
OUD	37.25±0.24	23.96±0.33	17.80±0.26	36.26±0.23	34.84±0.27	35.71±0.29	27.58±0.21	34.62±0.27	31.00±0.39	31.56±0.28
AAT	36.70±0.25	28.79±0.23	22.86±0.21	35.60±0.23	32.64±0.33	32.75±0.30	25.82±0.25	33.96±0.24	36.00±0.29	31.68±0.26
FAST-AT	32.53±0.27	23.30±0.22	16.90±0.30	27.03±0.22	29.01±0.33	29.67±0.30	23.52±0.25	29.23±0.24	27.64±0.26	26.53±0.27
ATAS	34.10±0.25	26.89±0.23	17.99±0.21	31.02±0.23	31.18±0.27	31.47±0.24	26.65±0.22	32.25±0.22	29.64±0.25	29.02±0.24
FGSM-PGK	36.32±0.28	28.53±0.25	24.83±0.23	35.17±0.23	35.33±0.33	35.78±0.30	29.84±0.25	35.51±0.24	32.76±0.28	32.67±0.27
VFAT-S	**41.98**±0.28	29.56±0.27	28.79±0.24	38.57±0.29	36.70±0.30	36.81±0.31	**34.29**±0.27	**39.34**±0.25	36.59±0.42	35.84±0.29
VFAT-WS	39.23±0.28	**32.75**±0.30	**28.79**±0.25	**39.12**±0.22	**37.69**±0.29	**37.80**±0.26	34.18±0.20	39.23±0.27	**40.33**±0.33	**36.57**±0.27
Dataset: UCF-101 & Model: 3D Pre-Activation ResNet-18										
Standard	18.68±0.12	1.978±0.14	0.000±0.17	2.198±0.10	11.76±0.14	12.64±0.14	6.923±0.10	7.912±0.14	4.730±0.22	7.424±0.14
PGD-AT	34.84±0.29	17.47±0.40	11.65±0.24	28.35±0.28	27.25±0.27	27.14±0.24	18.35±0.20	27.58±0.27	32.86±0.33	25.05±0.28
OUD	33.96±0.33	22.20±0.25	23.41±0.27	31.43±0.26	33.63±0.23	33.41±0.25	29.56±0.24	37.47±0.25	36.48±0.34	31.28±0.27
AAT	36.26±0.29	25.27±0.25	26.48±0.23	34.07±0.23	32.97±0.33	32.75±0.39	25.60±0.23	33.63±0.25	36.80±0.38	31.54±0.29
FAST-AT	32.31±0.34	16.15±0.30	10.66±0.31	23.41±0.23	21.43±0.29	22.09±0.31	16.37±0.22	22.86±0.22	28.13±0.37	21.49±0.29
ATAS	33.28±0.29	17.46±0.25	18.41±0.23	26.64±0.23	24.81±0.33	25.66±0.39	19.69±0.23	24.80±0.25	30.49±0.37	24.58±0.29
FGSM-PGK	35.01±0.24	21.91±0.30	20.32±0.24	30.47±0.22	29.12±0.25	29.61±0.24	24.23±0.20	30.20±0.25	32.69±0.36	28.17±0.26
VFAT-S	39.23±0.24	26.59±0.23	20.88±0.28	35.93±0.23	34.29±0.25	33.41±0.36	28.02±0.30	36.26±0.22	32.40±0.51	37.03±0.29
VFAT-WS	**40.06**±0.32	**42.92**±0.27	**37.78**±0.32	**40.55**±0.30	**42.09**±0.26	**37.88**±0.30	**41.63**±0.21	**41.83**±0.25	**44.73**±0.33	**41.05**±0.28
Dataset: UCF-101 & Model: Video Swin Transformer										
Standard	7.360±0.35	0.660±0.19	5.390±0.18	40.11±0.32	19.34±0.43	20.66±0.47	5.390±0.24	4.180±0.18	10.11±0.68	12.58±0.34
PGD-AT	31.10±0.39	29.56±0.37	25.49±0.38	49.45±0.30	31.10±0.29	31.21±0.45	24.07±0.39	26.81±0.32	38.57±0.40	31.93±0.36
OUD	31.54±0.75	26.37±0.37	18.79±0.30	51.21±0.27	27.80±0.53	27.80±0.32	20.55±0.51	22.64±0.33	37.82±0.67	29.38±0.45
AAT	27.80±0.44	19.34±0.34	15.27±0.26	49.56±0.25	26.81±0.65	26.59±0.39	20.11±0.35	20.22±0.28	35.71±0.57	26.82±0.39
FAST-AT	24.35±0.36	22.92±0.27	19.08±0.34	43.58±0.58	25.01±0.44	26.44±0.31	18.97±0.36	21.05±0.34	31.38±0.51	25.86±0.39
ATAS	26.96±0.29	25.84±0.26	22.71±0.27	45.82±0.43	28.80±0.47	29.75±0.78	20.39±0.38	24.06±0.38	35.64±0.86	28.89±0.45
FGSM-PGK	33.57±0.55	27.02±0.23	26.92±0.29	51.91±0.37	31.55±0.43	31.64±0.76	22.36±0.36	27.22±0.28	37.48±0.92	32.19±0.47
VFAT-S	42.53±0.32	28.46±0.26	42.42±0.26	62.20±0.32	33.52±0.48	32.53±0.60	21.43±0.53	29.45±0.31	38.24±0.82	36.75±0.43
VFAT-WS	**39.23**±0.42	**41.32**±0.44	**35.38**±0.31	**60.55**±0.38	**39.01**±0.44	**36.26**±0.29	**42.75**±0.24	**37.36**±0.32	**41.32**±0.75	**41.46**±0.40

fully capture the overall differences between distributions. JSD is symmetric, which enables bidirectional consistency between weak perturbations and strong perturbations, and is more conducive to maintaining semantic alignment under complex

TABLE V
ACCURACY(%) OF DIFFERENT ADVERSARIAL TRAINING METHODS UNDER DIFFERENT VIDEO ATTACK METHODS WITH DIFFERENT MODELS AS THE BACKBONE ON HMDB-51

Dataset: HMDB-51 & Model: 3D ResNet-18										
Method	ROA	SPA	AF	Frame Border	Frame saliency (one shot)	Frame saliency (iterative)	SparseAdv	Masked PGD	TT	Average
Standard	9.080±0.26	1.480±0.19	3.910±0.15	6.940±0.15	13.86±0.32	11.08±0.30	10.32±0.26	7.090±0.22	7.600±0.30	7.930±0.23
PGD-AT	27.23±0.33	24.27±0.38	26.48±0.45	16.31±0.28	15.20±0.33	14.46±0.23	11.14±0.26	22.79±0.28	26.42±0.51	20.48±0.34
OUD	25.90±0.28	24.14±0.27	24.67±0.42	15.06±0.30	16.61±0.45	16.75±0.39	10.26±0.24	23.48±0.28	26.57±0.45	20.38±0.34
AAT	28.55±0.41	24.63±0.24	25.99±0.27	16.87±0.32	16.24±0.43	17.09±0.29	15.97±0.22	24.59±0.31	27.80±0.53	21.97±0.33
FAST-AT	23.54±0.55	23.24±0.26	25.26±0.31	15.94±0.26	12.25±0.32	12.03±0.31	9.889±0.25	20.87±0.30	23.40±0.39	18.49±0.32
ATAS	24.96±0.38	24.59±0.29	25.76±0.27	15.61±0.29	14.34±0.29	14.13±0.27	10.73±0.32	23.68±0.39	23.55±0.36	19.71±0.32
FGSM-PGK	26.48±0.36	25.10±0.21	27.23±0.26	16.85±0.29	16.84±0.36	17.13±0.31	14.27±0.27	24.02±0.23	28.90±0.42	21.87±0.30
VFAT-S	32.40±0.60	25.38±0.27	29.06±0.39	17.34±0.32	29.37±0.43	27.97±0.44	16.38±0.27	26.93±0.26	32.92±0.60	26.42±0.39
VFAT-WS	**34.10**±0.56	**32.54**±0.43	**30.61**±0.47	**23.54**±0.28	**31.29**±0.55	**33.06**±0.38	**27.82**±0.23	**29.95**±0.29	**33.21**±0.41	**30.68**±0.40
Dataset: HMDB-51 & Model: 3D Pre-Activation ResNet-18										
Method	ROA	SPA	AF	Frame Border	Frame saliency (one shot)	Frame saliency (iterative)	SparseAdv	Masked PGD	TT	Average
Standard	6.860±0.28	0.520±0.13	1.850±0.19	3.910±0.13	14.24±0.59	14.43±0.21	12.18±0.22	3.250±0.15	5.540±0.22	6.970±0.24
PGD-AT	28.71±0.42	28.57±0.45	31.88±0.42	29.59±0.34	15.79±0.27	16.16±0.27	16.61±0.26	27.38±0.47	26.72±0.48	24.60±0.35
OUD	24.65±0.35	21.70±0.33	22.88±0.36	20.96±0.53	11.81±0.30	11.59±0.34	11.72±0.44	21.85±0.42	26.72±0.61	19.32±0.41
AAT	27.82±0.29	29.82±0.25	30.69±0.23	29.96±0.23	15.87±0.33	14.54±0.39	20.37±0.23	27.90±0.25	28.63±0.38	25.07±0.29
FAST-AT	22.88±0.34	23.99±0.30	26.64±0.31	24.21±0.23	14.98±0.29	15.65±0.31	15.28±0.22	24.50±0.22	23.76±0.37	21.32±0.29
ATAS	23.56±0.23	26.25±0.34	27.06±0.55	25.21±0.45	10.11±0.26	10.10±0.32	11.87±0.26	23.56±0.25	25.46±0.25	18.13±0.32
FGSM-PGK	28.92±0.50	29.07±0.22	32.21±0.24	28.30±0.22	17.71±0.44	17.12±0.37	18.65±0.28	27.68±0.29	28.27±0.48	25.33±0.34
VFAT-S	31.81±0.28	34.13±0.36	35.31±0.31	32.15±0.27	27.31±0.28	27.68±0.40	22.07±0.31	30.81±0.30	31.66±0.32	30.33±0.35
VFAT-WS	**33.43**±0.38	**35.02**±0.30	**35.76**±0.37	**33.40**±0.30	**30.41**±0.44	**30.55**±0.55	**26.13**±0.27	**32.88**±0.30	**32.18**±0.53	**32.20**±0.38
Dataset: HMDB-51 & Model: Video Swin Transformer										
Method	ROA	SPA	AF	Frame Border	Frame saliency (one shot)	Frame saliency (iterative)	SparseAdv	Masked PGD	TT	Average
Standard	19.70±0.30	0.070±0.12	3.690±0.11	25.76±0.14	16.90±0.31	16.83±0.21	0.890±0.18	4.130±0.13	4.350±0.40	10.26±0.24
PGD-AT	32.25±0.39	10.63±0.33	19.34±0.26	34.61±0.30	14.10±0.31	13.95±0.62	9.82±0.28	15.35±0.26	30.04±0.39	20.01±0.35
OUD	33.87±0.33	11.22±0.28	19.78±0.26	36.68±0.23	15.35±0.44	13.80±0.38	11.07±0.31	15.20±0.24	29.67±0.51	20.74±0.33
AAT	36.97±0.29	8.340±0.27	18.60±0.24	41.33±0.26	13.36±0.37	12.40±0.32	7.900±0.29	7.090±0.31	31.22±0.60	19.69±0.33
FAST-AT	27.82±0.37	8.480±0.27	17.04±0.22	31.87±0.23	13.42±0.60	13.65±0.29	10.18±0.33	15.12±0.34	26.34±0.49	18.21±0.35
ATAS	31.37±0.27	10.21±0.39	18.90±0.24	32.78±0.26	14.52±0.35	14.50±0.34	10.35±0.33	15.57±0.23	27.84±0.32	19.56±0.30
FGSM-PGK	34.06±0.52	11.36±0.27	20.99±0.24	38.49±0.31	17.70±0.51	17.52±0.35	14.13±0.24	17.25±0.23	31.60±0.62	22.57±0.37
VFAT-S	41.25±0.31	25.46±0.27	27.23±0.24	45.76±0.24	30.85±0.25	32.10±0.42	31.00±0.26	26.64±0.29	33.06±0.68	32.59±0.33
VFAT-WS	**44.21**±0.36	**27.16**±0.40	**29.89**±0.24	**48.04**±0.32	**33.51**±0.29	**32.99**±0.62	**33.06**±0.34	**28.56**±0.26	**34.02**±0.38	**34.60**±0.36

augmentations. Therefore, we use JSD as the consistency constraint loss.

3) Ablation Study On Different Spatial-Temporal Mix Operations: Figure 7.(C) illustrates the impact of different spatial-temporal augmentation operations (SMU, TCM, TMU, CM, and CMU) on robust accuracy. The experimental results show that CMU achieves the highest robust accuracy of approximately 32%, indicating its advantage in enhancing model robustness. CM achieves a robust accuracy of around 31%, slightly lower than CMU. In contrast, SMU, TCM, and TMU, which employ only temporal augmentations, have relatively close robust accuracies, all around 29%, suggesting that these augmentations alone have limited effectiveness in improving robustness. This indicates that CM and CMU, by further exploiting both temporal and spatial information, are more effective in enhancing the model's adversarial robustness. Therefore, we use CMU as the Spatial-Temporal Mix operation.

4) Ablation Study on Different VFAT Versions: We investigate the effectiveness of various training frameworks that incorporate different spatial-temporal perturbation mechanisms. The results are presented in Figure 8. Specifically, VFAT-W and VFAT-WR represent sequential and random high-frequency frame shifts, respectively, as distinct weak perturbation strategies. VFAT-S employs STF-AUG paired with FGSM attack as its training approach. VFAT-WSS builds on our VFAT-WS method by adding Jensen-Shannon divergence during the FGSM operation to increase the gap between weakly and strongly perturbed videos.

As shown in the results, VFAT-WR's random frame shifts lead to excessive offsets, causing a decline in both clean and adversarial accuracy under AutoAttack. In contrast, VFAT-WS achieves similar clean accuracy to VFAT-W and VFAT-S while significantly enhancing adversarial robustness across various attacks. Additionally, VFAT-WSS underperforms compared to VFAT-WS, indicating that incorporating Jensen-Shannon

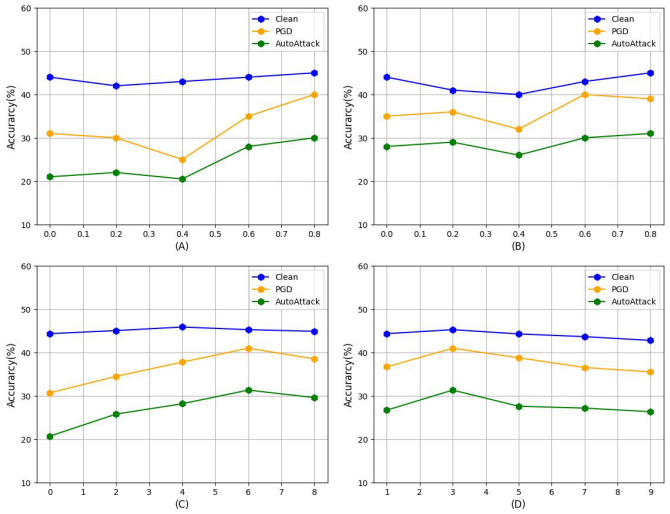

Fig. 6. (A) The parameter tuning results with different λ. (B) The parameter tuning results with different μ. (C) The parameter tuning results with different N. (D) The parameter tuning results with different k.

TABLE VI

EFFECTS OF SPATIAL AND TEMPORAL AUGMENTATION UNDER AUTOATTACK ON UCF-101 USING A 3D RESNET-18

Method	VFAT-W	VFAT-WT	VFAT-WV	VFAT-WS
Accuracy	27.11±0.46	30.42±0.28	29.21±0.32	31.32±0.36

divergence (JSD) into FGSM does not improve adversarial robustness. This is because the long-term regularization objective of JSD conflicts with the objective of FGSM, which aims to minimize the probability of the correct class. Overall, VFAT-WS effectively balances clean accuracy and robustness. It outperforms VFAT-W by 7.47% and VFAT-S by 1.87% under PGD attack, and by 3.85% and 2.75% under AutoAttack, respectively.

5) Ablation Study on STF-AUG: To analyze the contribution degrees of spatial augmentation and temporal augmentation in STF-AUG, we conduct further fine-grained experiments. As shown in Table VI, VFAT-W represents the baseline by introducing TF-AUG. VFAT-WT replaces the STF-AUG in VFAT-WS with only temporal augmentation, and VFAT-WV replaces it with only spatial augmentation. The experiments are conducted on the UCF-101 dataset (based on the 3D ResNet-18 model), and the robust accuracy of different methods is evaluated using AutoAttack. The results show that VFAT-WT (30.42%) outperforms VFAT-WV (29.21%) by 1.21%, indicating that using only temporal augmentation achieves higher performance than using only spatial augmentation. This gap clearly demonstrates that, within STF-AUG, the temporal dimension contributes more to improving model robustness than the spatial dimension. Finally, VFAT-WS, which fully explores the potential of both temporal augmentation and spatial augmentation, achieves the highest robust accuracy of 31.32%, outperforming the baseline (+ 10.44%) and all other single or dual-component variants, demonstrating the overall effectiveness of our proposed TF-AUG and STF-AUG.

TABLE VII

RUNNING-TIME PROPORTION OF VFAT-S AND VFAT-WS MODULES ON UCF-101 USING A 3D RESNET-18

Method	TF-AUG	STF-AUG	Inner-Attack	Outer-Train
VFAT-S	0.06%	0.54%	43.97%	26.05%
VFAT-WS	0.19%	0.45%	50.61%	26.77%

TABLE VIII

PERFORMANCE COMPARISON BETWEEN TRADES AND VFAT-WS ON UCF-101 USING A 3D RESNET-18

Method	CLEAN (%)	PGD (%)	AutoAttack (%)	Time (m)
PGD-AT	42.53±0.29	38.24±0.34	24.51±0.34	102
TRADES	42.78±0.31	40.33±0.32	25.18±0.30	110
VFAT-WS	**44.51**±0.25	**40.99**±0.32	**31.32**±0.36	40

E. Analysis of Running-Time Proportion

To verify the efficiency of TF-AUG and STF-AUG, we analyze their running-time proportions in VFAT-S and VFAT-WS. As shown in Table VII, the overhead of the augmentation operations is extremely low: the time consumption of TF-AUG accounts for only 0.06% (in VFAT-S) and 0.19% (in VFAT-WS), while that of STF-AUG is merely 0.54% and 0.45% respectively. In contrast, the combined computing time of adversarial sample generation (Inner-Attack) and model update (Outer-Train) accounts for 70% (in VFAT-S) and over 77% (in VFAT-WS).

The results indicate that the training bottleneck lies mainly in adversarial sample construction rather than frequency-domain augmentation. This fully verifies the efficiency of the proposed spatiotemporal augmentation methods.

F. Robustness-Accuracy-Efficiency Trade-off Analysis

Different from the traditional "robustness-accuracy trade-off", VFAT-WS focuses more on balancing the triple of robustness, accuracy, and efficiency. To verify this, we compare it with the classic TRADES [58] (which focuses on the "robustness-accuracy trade-off") on the UCF-101 dataset. As shown in Table VIII, VFAT-WS outperforms TRADES in several key metrics. In terms of adversarial robustness, its AutoAttack accuracy reaches 31.32%, which is significantly higher than TRADES' 25.18%; it also performs better in clean accuracy (44.51% vs. 42.78%), indicating that the method effectively preserves the understanding of original semantics while improving robustness.

This advantage stems from the core design of VFAT-WS. Unlike TRADES, which explicitly balances natural and adversarial losses through regularization terms, VFAT-WS expands the perturbation space in the spatiotemporal frequency domain via TF-AUG and STF-AUG, and introduces "weak-strong" consistency regularization to force the model to output consistent predictions for different augmented versions of the same video. This mechanism urges the model to focus on stable low-frequency semantics between frames (such as motion trajectories and action structures) and avoid relying on high-frequency details that are vulnerable to attacks.

Similar to contrastive learning, this process guides the model to gradually adapt from weakly augmented samples to

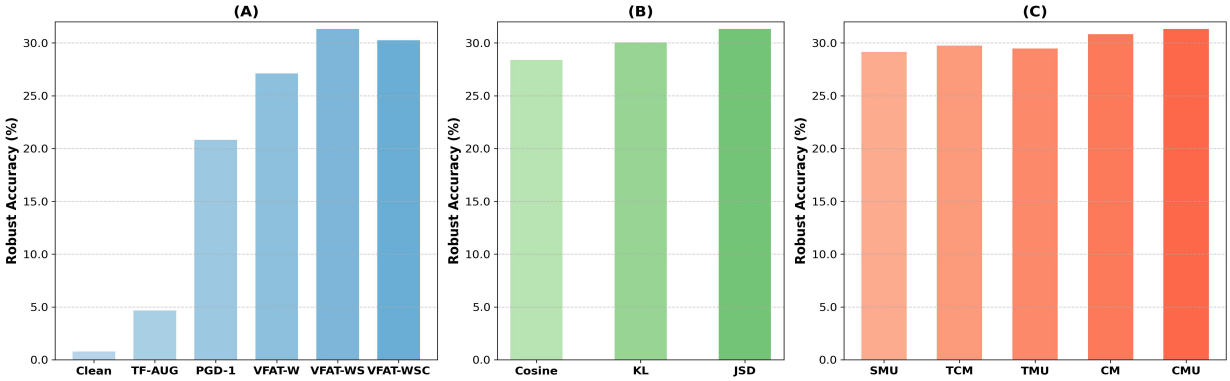

Fig. 7. Ablation study of different modules. Figure (A) shows the effect of different training strategies: "Clean" refers to training with clean samples, "PGD-1" represents adversarial training using FGSM, and "VFAT-WSC" denotes adversarial training with consistency constraints where noisy samples in VFAT-WS are replaced with clean samples. Figure (B) presents an ablation study using different types of consistency losses. "Cosine" represents cosine loss, "KL" stands for KL divergence, and "JSD" indicates JS divergence. Figure (C) displays the ablation study of various augmentations mentioned in Table 1. "SMU" stands for Subtle MixUp, "TCM" for Temporal CutMix, "TMU" for Temporal CutMixUp, "CM" for 3D CutMix, and "CMU" for 3D CutMixUp.

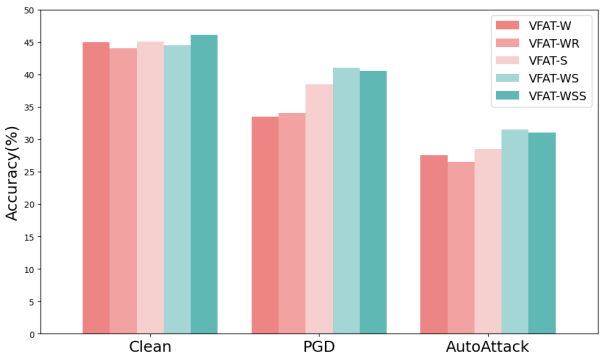

Fig. 8. The accuracy rates (%) of different VFAT versions. All experiments were conducted on the UCF-101 dataset using a 3D ResNet-18 as the backbone, with results averaged over three runs to obtain the final outcomes.

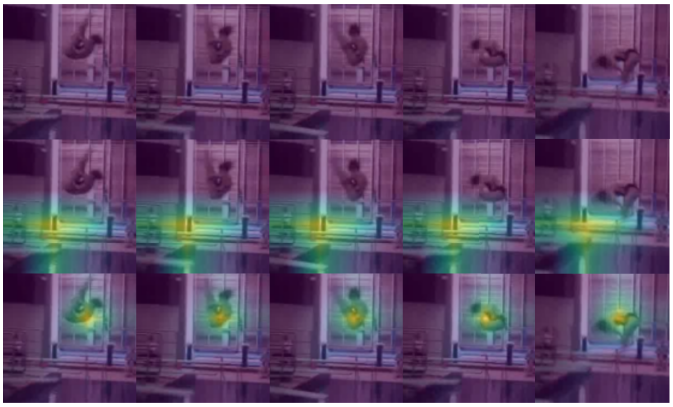

Fig. 9. From top to bottom: Grad-CAM visualizations of PGD-1, VFAT-S, and VFAT-WS are presented, demonstrating their responses to corrupted inputs under the scrutiny of AutoAttack. The 3D Pre-activation ResNet-18 serves as the underlying architecture for these evaluations.

stronger perturbations, achieving the coordinated improvement of robustness and clean accuracy. In addition, VFAT-WS only takes 40 minutes for training, which is much shorter than TRADES' 110 minutes—this is due to the high efficiency of frequency-domain augmentation and the avoidance of multi-step optimization (e.g., PGD). VFAT-WS achieves a better "robustness-accuracy-efficiency" triple balance and demonstrates greater potential for practical applications.

G. More Results

1) Qualitative Results: We provide GradCam visualizations of different methods against AutoAttack with $\epsilon = 16/255$ in Figure 9. The model trained with the FGSM (PGD-1) method fails to withstand the strong perturbations of AutoAttack, resulting in misclassifications. In contrast, VFAT-S mitigates the model's reliance on unstable high-frequency details to some extent. VFAT-WS further improves performance by focusing the model on semantically relevant low-frequency information, enabling accurate predictions under strong attack disturbances. By optimizing the model's attention to low-frequency information, VFAT-WS promotes the learning of more stable feature representations and enhances adversarial robustness.

2) Visualization Of Proposed Perturbations: The visualization of our proposed spatial-temporal frequency perturbations is shown in Figure 10. It can be observed that the texture of the Spatial-Temporal Mix enhanced frequency perturbation is more complex and variable compared to the temporal shifted frequency perturbation. TF-AUG broadens the perturbation space by swapping high-frequency details and low-frequency information between frames of the same video, reducing the model's reliance on high-frequency details. In contrast, the STF-AUG proposed in our research adopts a far more adaptable and versatile strategy. It facilitates the exchange of high-frequency details and low-frequency information not merely among frames but also with external videos. Through this approach, it penetrates deeper and explores the perturbation space in a more comprehensive manner. These perturbations, which play a crucial role in guiding the model to boost its robustness, scarcely modify the semantic content of the original video (in this instance, centered around skiing). Additionally, STF-AUG impels the model to place greater emphasis on low-frequency information. This type of information is typically more closely associated with the video's label data, thus enhancing the model's robustness.

3) Reducing Robust Overfitting With STF-AUG And Our Consistency Regulation: In this section, we elucidate the pivotal role of STF-AUG and our consistency regulation in

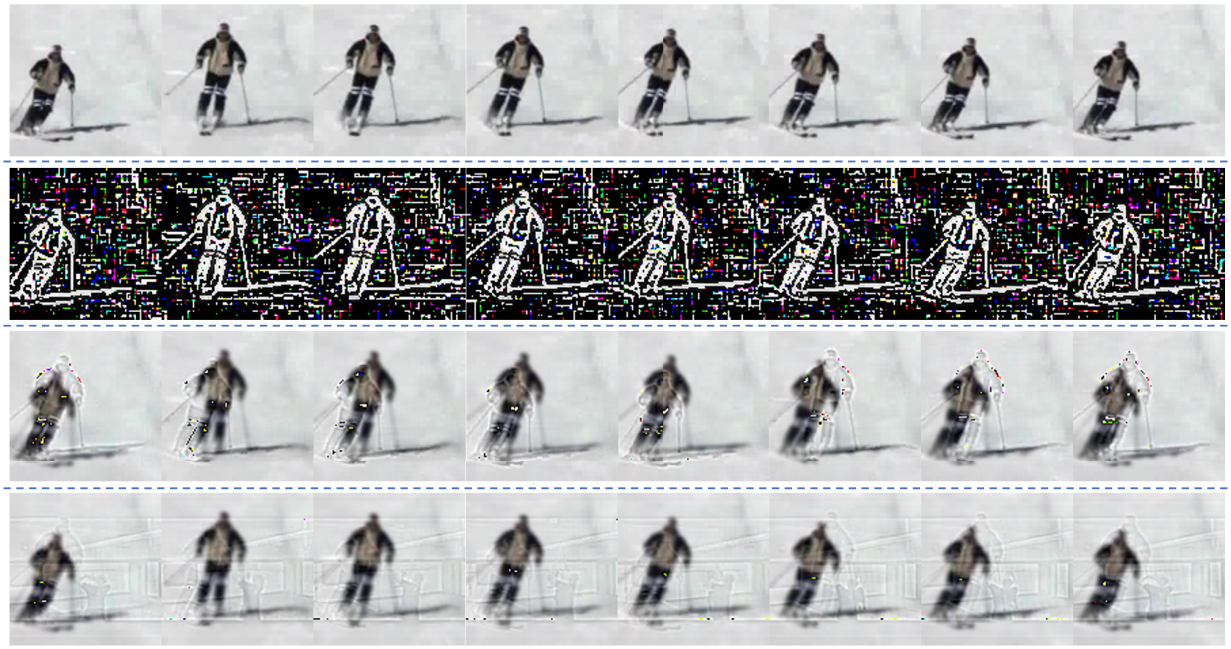

Fig. 10. Visualization of spatial-temporal frequency perturbations. The first row depicts clean video frames, the second row shows high-frequency of video frames, the third row represents video frames enhanced by TF-AUG, and the last row depicts video frames enhanced by STF-AUG.

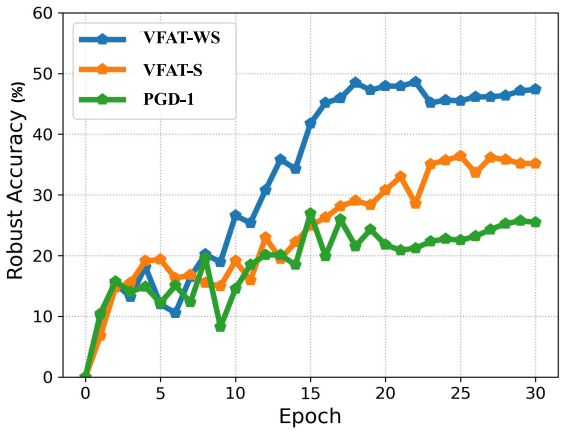

Fig. 11. Difference between the training processes of PGD-1, VFAT-S and VFAT-WS under FGSM attack with $\epsilon = 16/255$.

mitigating robust overfitting. The experiment assesses robust accuracy (%) under FGSM attack across various adversarial training methods. We evaluate the performance on PreAct-ResNet-18, which is trained on the UCF-101 dataset, employing an l_∞ with $\epsilon = 16/255$ for FGSM attack.

The experimental results in Figure 11 show that although PGD-1 alleviates catastrophic overfitting through random initialization, appropriate step size, and cyclic learning, it still exhibits severe robust accuracy fluctuations in the early stages of training. This indicates its over-reliance on high-frequency details and overfitting to a single perturbation path, which severely limits the improvement of its adversarial robustness. Compared with PGD-1, VFAT-S mitigates the robust accuracy fluctuations in the initial stage and shows a more significant upward trend. This demonstrates that with the help of STF-AUG, VFAT-S shifts the model's focus to more stable low-frequency information and expands the perturbation path.

In contrast, VFAT-WS effectively avoids catastrophic overfitting and further improves robustness through the following three mechanisms: 1) Diversified spatiotemporal frequency-domain perturbations: On the basis of PGD-1's random initialization, TF-AUG and STF-AUG introduce inter-frame temporal shifts and 3D spatial mixing. This effectively expands the perturbation space, covers high-frequency details and spatiotemporal dynamics, and breaks through the limitation of PGD-1's single path. 2) Weak-strong consistency regularization: Through the L_{TCW} and L_{TCS} losses, the model is forced to maintain consistent predictions for weakly perturbed and strongly perturbed samples of the same video. This enhances perturbation invariance and alleviates overfitting to specific patterns. 3) Efficient convergence mechanism: By combining cyclic learning rate and spatiotemporal consistency constraints, VFAT-WS approaches convergence in only 20 epochs (fewer than PGD-1), which shortens the training cycle and reduces the risk of overfitting. The synergistic effect of the above mechanisms enables VFAT-WS to maintain efficient training while effectively avoiding catastrophic overfitting and improving robust performance

V. Conclusion

In this paper, we propose VFAT-WS, the first fast adversarial training framework specifically designed for video data. VFAT-WS introduces two key components: (1) TF-AUG and its spatial-temporal extension (STF-AUG), combined with FGSM attack to enhance both training efficiency and adversarial robustness; (2) a weak-to-strong spatial-temporal consistency regularization that progressively guides the model from simpler (TF-AUG) to more complex (STF-AUG) augmentations, thereby improving generalization. Collectively, these mechanisms establish an improved balance among clean accuracy, adversarial robustness and efficiency. Extensive experiments

across different models on UCF-101 and HMDB-51, involving diverse seen and unseen attacks, demonstrate the effectiveness of our proposed VFAT-WS. It achieves great improvements in adversarial robustness and corruption robustness, along with a remarkable 490% speed enhancement, effectively defending against a wide range of strong attacks.

REFERENCES

[1] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3239–3259, Jun. 2022.

[2] X. Jia, H. Yan, Y. Wu, X. Wei, X. Cao, and Y. Zhang, "An effective and robust detector for logo detection," 2021, *arXiv:2108.00422*.

[3] Y. Nie, J. Hou, X. Han, and M. Niesner, "RfD-Net: Point scene understanding by semantic instance reconstruction," in *Proc. CVPR*, 2021, pp. 4606–4616.

[4] W. Jiang et al., "Moderating the generalization of score-based generative model," 2024, *arXiv:2412.07229*.

[5] Y. Dong et al., "Boosting adversarial attacks with momentum," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[7] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, Mar. 2016, pp. 372–387.

[8] S. Wang, X. Yue, Y. Lyu, and C. Shan, "Exploring adversarial transferability between Kolmogorov–Arnold networks," 2025, *arXiv:2503.06276*.

[9] L. Kong et al., "Multi-modal data-efficient 3D scene understanding for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 5, pp. 3748–3765, May 2025.

[10] X. Hu et al., "Transformer tracking via frequency fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 2, pp. 1020–1031, Feb. 2024.

[11] X. Hu, B. Zhong, Q. Liang, S. Zhang, N. Li, and X. Li, "Toward modalities correlation for RGB-T tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 10, pp. 9102–9111, Oct. 2024.

[12] H. Yang, Y. Zhou, J. Wu, H. Liu, L. Yang, and C. Lv, "Human-guided continual learning for personalized decision-making of autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 4, pp. 5435–5447, Apr. 2025.

[13] X. Wei, S. Wang, and H. Yan, "Efficient robustness assessment via adversarial spatial–temporal focus on videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10898–10912, Sep. 2023.

[14] S. Wang, H. Liu, and H. Zhao, "Public-domain locator for boosting attack transferability on videos," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2024, pp. 1–6.

[15] K. Chen, Z. Wei, J. Chen, Z. Wu, and Y.-G. Jiang, "Attacking video recognition models with bullet-screen comments," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 1, pp. 312–320.

[16] K. Jiang et al., "Efficient decision-based black-box patch attacks on video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2023, pp. 4356–4366.

[17] H. Xu et al., "Adversarial attacks and defenses in images, graphs and text: A review," *Int. J. Autom. Comput.*, vol. 17, no. 2, pp. 151–178, Apr. 2020.

[18] J. Zhang et al., "Towards efficient data free black-box adversarial attack," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 15115–15125.

[19] Y. Zhang, G. Zhang, P. Khanduri, M. Hong, S. Chang, and S. Liu, "Revisiting and advancing fast adversarial training through the lens of bi-level optimization," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 26693–26712.

[20] Z. Wei, Y. Wang, Y. Guo, and Y. Wang, "CFA: Class-wise calibrated fair adversarial training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8193–8201.

[21] J. Dong, S.-M. Moosavi-Dezfooli, J. Lai, and X. Xie, "The enemy of my enemy is my friend: Exploring inverse adversaries for improving adversarial training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 24678–24687.

[22] Z. Wang, X. Li, H. Zhu, and C. Xie, "Revisiting adversarial training at scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 24675–24685.

[23] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8684–8694.

[24] D. Yin, R. G. Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer, "A Fourier perspective on model robustness in computer vision," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1–11.

[25] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann, "Data augmentation can improve robustness," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 29935–29948.

[26] X. Wei, J. Zhu, S. Yuan, and H. Su, "Sparse adversarial perturbations for videos," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 8973–8980.

[27] L. Jiang, X. Ma, S. Chen, J. Bailey, and Y.-G. Jiang, "Black-box adversarial attacks on video recognition models," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 864–872.

[28] Z. Wei et al., "Heuristic black-box adversarial attacks on video recognition models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 12338–12345.

[29] H. Yan and X. Wei, "Efficient sparse attacks on videos using reinforcement learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2326–2334.

[30] W. Deng, C. Yang, K. Huang, Y. Liu, W. Gui, and J. Luo, "Sparse adversarial video attack based on dual-branch neural network on industrial artificial intelligence of things," *IEEE Trans. Ind. Informat.*, vol. 20, no. 7, pp. 9385–9392, Jul. 2024.

[31] Z. Gao et al., "ReToMe-VA: Recursive token merging for video diffusion-based unrestricted adversarial attack," in *Proc. 32nd ACM Int. Conf. Multimedia*, 2024, pp. 4485–4494.

[32] Y. Diao, T. Shao, Y.-L. Yang, K. Zhou, and H. Wang, "BASAR: Black-box attack on skeletal action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Roslyn, NY, USA: Black, Jun. 2021, pp. 7593–7603.

[33] H. Wang, Y. Diao, Z. Tan, and G. Guo, "Defending black-box skeleton-based human activity classifiers," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2546–2554.

[34] Y. Diao et al., "Understanding the vulnerability of skeleton-based human activity recognition via black-box attack," *Pattern Recognit.*, vol. 153, Sep. 2024, Art. no. 110564.

[35] Y. Diao et al., "TASAR: Transfer-based attack on skeletal action recognition," 2024, *arXiv:2409.02483*.

[36] C. Xiao et al., "AdvIT: Adversarial frames identifier based on temporal consistency in videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Feb. 2019, pp. 3967–3976.

[37] X. Jia, X. Wei, and X. Cao, "Identifying and resisting adversarial videos using temporal consistency," 2019, *arXiv:1909.04837*.

[38] S.-Y. Lo, J. M. J. Valanarasu, and V. M. Patel, "Overcomplete representations against adversarial videos," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1939–1943.

[39] K. A. Kinfu and R. Vidal, "Analysis and extensions of adversarial training for video classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 3415–3424.

[40] H. Kim, W. Lee, and J. Lee, "Understanding catastrophic overfitting in single-step adversarial training," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 9, pp. 8119–8127.

[41] M. K. Yucel, R. G. Cinbis, and P. Duygulu, "HybridAugment++: Unified frequency spectra perturbations for model robustness," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 5695–5705. [Online]. Available: https://api.semanticscholar.org/CorpusID:260125802

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[43] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2206–2216.

[44] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[45] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[46] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[47] E. Wong, L. Rice, and J. Zico Kolter, "Fast is better than free: Revisiting adversarial training," 2020, *arXiv:2001.03994*.

[48] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.
[49] Z. Liu et al., "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 3192–3201.
[50] Z. Huang et al., "Fast adversarial training with adaptive step size," *IEEE Trans. Image Process.*, vol. 32, pp. 6102–6114, 2023.
[51] X. Jia et al., "Improving fast adversarial training with prior-guided knowledge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 9, pp. 6367–6383, Sep. 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:257913654
[52] T. Wu, L. Tong, and Y. Vorobeychik, "Defending against physically realizable attacks on image classification," 2019, *arXiv:1909.09552*.
[53] S.-Y. Lo and V. M. Patel, "Defending against multiple and unforeseen adversarial videos," *IEEE Trans. Image Process.*, vol. 31, pp. 962–973, 2022.
[54] M. Zajac, K. Zołna, N. Rostamzadeh, and P. O. Pinheiro, "Adversarial framing for image and video classification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 10077–10078.
[55] N. Inkawhich, M. Inkawhich, Y. Chen, and H. Li, "Adversarial attacks for optical flow-based action recognition classifiers," 2018, *arXiv:1811.11875*.
[56] Z. Wei, J. Chen, Z. Wu, and Y.-G. Jiang, "Boosting the transferability of video adversarial examples via temporal translation," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, pp. 2659–2667.
[57] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," 2019, *arXiv:1903.12261*.
[58] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7472–7482.

Songping Wang received the B.Eng. degree from Hefei University of Technology and the M.Eng. degree from Beihang University. He is currently pursuing the Ph.D. degree with Nanjing University. His research interests include adversarial robustness in deep learning and machine learning.

Hanqing Liu is currently pursuing the Ph.D. degree with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include adversarial machine learning and the robustness of vision-language models.

Yueming Lyu (Member, IEEE) received the B.Eng. degree from Nanjing University of Aeronautics and Astronautics and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2019 and 2024, respectively. Since 2024, she has been with the School of Intelligence Science and Technology, Nanjing University, where she is currently an Assistant Professor. Her current research interests include computer vision, content generation, and AI safety.

Xiantao Hu received the M.S. degree from Guangxi Normal University, Guilin, China. He is currently pursuing the Ph.D. degree with Nanjing University of Science and Technology, Nanjing, China. His research interests include computer vision and machine learning.

Ziwen He received the Ph.D. degree from the Center for Research on Intelligent Perception and Computing (CRIPAC), State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, China, in 2023. He is currently a Lecturer with the School of Computer and Software and the Engineering Research Center of Digital Forensics, Nanjing University of Information Science and Technology, Nanjing. His current research interests include adversarial example and computer vision.

Wei Wang (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2012. He is currently an Associate Professor with the New Laboratory of Pattern Recognition (NLPR), CASIA. His research interests include artificial intelligence safety and multimedia forensics.

Caifeng Shan (Senior Member, IEEE) received the B.Eng. degree from the University of Science and Technology of China, the M.Eng. degree from the Institute of Automation, Chinese Academy of Sciences, and the Ph.D. degree from the Queen Mary University of London. He has co-authored more than 150 papers and 80 patent applications. His research interests include computer vision, pattern recognition, medical image analysis, and related applications. He served as an Associate Editor for journals, including IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Liang Wang (Fellow, IEEE) received the B.Eng. and M.Eng. degrees from Anhui University in 1997 and 2000, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2004. From 2004 to 2010, he was a Research Assistant with Imperial College London, U.K., and Monash University, Australia, a Research Fellow with The University of Melbourne, Australia, and a Lecturer with the University of Bath, U.K. He is currently a Full Professor with the Hundred Talents Program, National Laboratory of Pattern Recognition, CASIA. He has widely published in highly ranked international journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and IEEE TRANSACTIONS ON IMAGE PROCESSING and leading international conferences, such as CVPR, ICCV, and ECCV. His research interests include machine learning, pattern recognition, and computer vision. He is an IAPR Fellow. He served as an Associate Editor for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, and *Pattern Recognition*.