PROBING INTO OVERFITTING FOR VIDEO RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

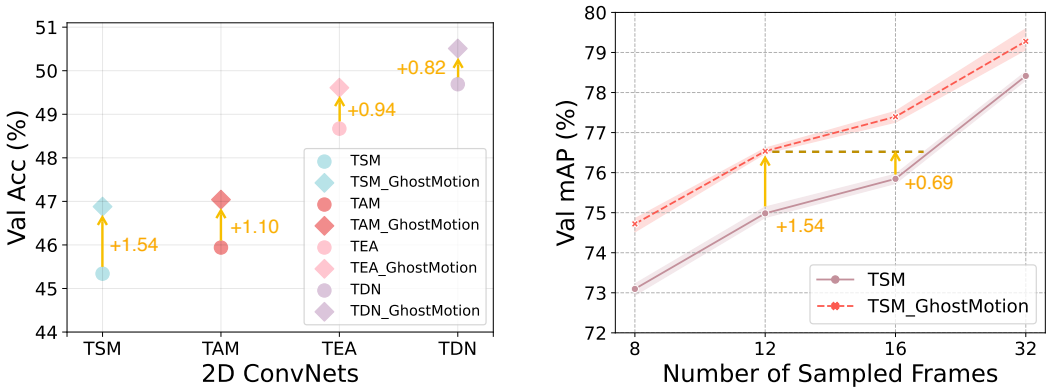
Video recognition methods based on 2D networks have thrived in recent years, leveraging advanced image classification techniques. However, overfitting is an even severe problem in 2D video recognition models as 1) the scale of video datasets is relatively small compared to image recognition datasets like ImageNet; 2) current pipeline treats background and semantic frames equally during optimization which aggravates overfitting. Based on these challenges, we design a video-specific data augmentation approach, named as *Ghost Motion* (GM), to alleviate overfitting. Specifically, GM shifts channels along temporal dimension to enable semantic motion information diffused into other frames which may be irrelevant originally, leading to improvement in frame-wise accuracy. In addition, for challenging video samples with significant temporal dependency (e.g., Something-Something), we further scale the logits during training to prevent overconfident predictions on background frames. Comprehensive empirical validation on various popular datasets shows that the proposed method can improve the generalization of existing methods and is compatible to other competing data augmentation approaches.

1 INTRODUCTION

Video recognition methods has evolved rapidly due to the increasing number of online videos and success of advanced deep neural networks. 2D based methods Wang et al. (2016); Lin et al. (2019); Li et al. (2020); Wang et al. (2021) arouse researchers' interests because of their efficiency in both computing and storage, even if 3D convolutional neural networks Feichtenhofer et al. (2019); Carreira & Zisserman (2017); Tran et al. (2015) exhibit strong performance in video recognition. However, 2D convolutional nerual networks still suffer from overfitting issue. For instance, on Something-Something V1 Goyal et al. (2017) which contains strong temporal dependency, the training and validation accuracies of TSM Lin et al. (2019) are 81.22% and 45.34%, respectively. Besides, its Expected Calibration Error (ECE) Guo et al. (2017) is 25.83% which means the model gives overconfident predictions and brings negative impact in practice.

There are two main reasons for overfitting in video: 1) The first is video recognition datasets always have fewer training samples compared with image classification (e.g., ImageNet Deng et al. (2009) with 1.2 million training samples compared to Kinetics Kay et al. (2017) with 240K videos). Furthermore, modeling video sample is harder than static images, which should require even more samples for temporal modeling. 2) The second reason is 2D based methods average the logits of all frames to vote for final decision. It causes the model to overfit certain frames which contain less semantic information (e.g., background scene). In this view, these frames can be regarded as noise in optimization because they do not provide motion information for temporal modeling.

To alleviate overfitting, many attempts have been proposed. To name a few, Dropout Srivastava et al. (2014) and Label Smoothing Szegedy et al. (2016) are widely used because of the regularization effects. Data augmentation Zhang et al. (2017); Yun et al. (2019); DeVries & Taylor (2017); Cubuk et al. (2020) is another research line to relieve overfitting. Although shown effectiveness in image-level tasks, it is hard to employ them on video tasks directly. The potential reason is these methods are not specifically designed for video data and some transformations will break the motion pattern resulting in detrimental temporal representations.



(a) Ghost Motion leads to consistent improvement on Something-Something V1 with different models.

(b) Ghost Motion continuously improves the performance with different sampled frames on ActivityNet.

Figure 1: The performance of Ghost Motion on different datasets. Ghost Motion introduces negligible cost to current competing methods and improves their generalization abilities.

In our work, we propose a data augmentation approach Ghost Motion (GM) which shifts channels along the temporal dimension to propagate motion information to adjacent frames. Specifically, all channels will be shifted for one step along the temporal axis leading to misaligned channel orders, and we interpolate between the original video and the misaligned one to form the new video, named as ghost video. It diffuses motion patterns from salient frames into other frames to enhance the overall representative capacity. In this way, the misclassification of semantic-irrelevant frames will be alleviated as they are enhanced by the shared motion information from salient frames. Although the shifting operation results in mismatched channel orders, we surprisingly find the disorder is beneficial to improve generalization as well. We elaborate this point in our analysis in Sec. 4.5.2. Ghost Motion effectively enlarges the input space without crashing the motion pattern and offers continuous variance in the input space. Moreover, we find utilizing a hyperparameter Temperature to scale the logits before Softmax can further mitigate overfitting and reduce the calibration error on especially challenging datasets such as Something-Something V1&V2.

Our proposed GM can be easily plugged into existing methods to improve their generalization abilities with a few line codes and brings almost zero costs. Shown in Fig. 1 (a), GM results in consistent improvement in various 2D models. In Fig 1 (b), GM continuously improves the performance of baseline model with different number of sampled frames. In addition, our method is compatible with existing image-level augmentations. We can jointly apply them to relieve overfitting and further boost the performance. The main contributions are summarized as follow:

- We propose a video recognition data augmentation method Ghost Motion (GM) which can *effectively improve the generalization of current video benchmark models* and is *compatible with existing image-level data augmentation approaches*.
- We find smoothing the logits can prevent overconfident predictions to alleviate overfitting for temporal-dependent datasets such as Something-Something.
- We conduct comprehensive experiments to *validate the strength of Ghost Motion on various datasets and methods*. Extensive ablation with detailed analysis illustrate the motivation and intuition about our GM strategy.

2 Related Work

2.1 Action Recognition

Action recognition has benefited a lot from the development in image recognition, especially the 2D based methods which share the same backbone with the models in image recognition. The main focus of these 2D methods lies in temporal modeling. For instance, TSN Wang et al. (2016) proposes to average the prediction of all the frames and laid a foundation for latter 2D models. While TSM Lin et al. (2019) is designed to exchange the information of adjacent frames by the shifting operation. Recently, people have paid more attention to multi-scale temporal modeling and many methods are proposed to not only model the motion variance locally, but capture the global temporal information as well Li et al. (2020) Wang et al. (2021). Another line of research focuses on using 3D-CNNs to

learn the spatial and temporal information jointly, such as C3D Tran et al. (2015), I3D Carreira & Zisserman (2017) and SlowFast Feichtenhofer et al. (2019). Though being effective, these methods usually cost great computations and are hard to be deployed on edge devices.

Ghost Motion is similar with temporal shift module (TSM) Lin et al. (2019) as both methods involves the shifting operation and enforces information exchange between neighboring frames. Nevertheless, the differences are: (1) TSM is designed as a temporal modeling module which is inserted densely in the network, so that the operations are made in the features space and the I/O cost is non-negligible; Ghost Motion is a data augmentation method which manipulates the input data, thus can be generalized to all models and cost trivial computations. (2) TSM shifts limited number of channels at every convolution block and the rest channels are kept in the original place; Ghost Motion shifts all channels initially which causes mismatch in the order of channels and the ghost video will be determined by the sampling coefficient, leading to a enlarged input space.

2.2 Mitigating Overfitting

Lots of effort has been made towards the study of regularization in the image domain to prevent overfitting. Dropout Srivastava et al. (2014), Weight Decay Krogh & Hertz (1991) and Label Smoothing Szegedy et al. (2016) are widely used in image recognition to reduce overfitting. Recently, RMS Kim et al. (2020) is proposed to vary the magnitude of low-frequency components randomly to regularize the 3D models. Besides these regularization methods, data augmentation has also proven to be an effective way to mitigate the effect of overfitting, especially for datasets which only have limited samples. Traditional augmentation methods including random flipping, resizing and cropping are frequently used to enforce invariance in deep networks He et al. (2016) Huang et al. (2017). Some studies have investigated to enlarge the input space by randomly occluding certain regions in images DeVries & Taylor (2017), blending two images Zhang et al. (2017) or replacing an image patch with the one in another image Yun et al. (2019). In recent years, automatic augmentation strategies Cubuk et al. (2018) Cubuk et al. (2020) are shown to be effective as it considers extensive augmentation types, such as color jittering, shear, translation, etc. However, there has been lack of studies of data augmentation methods tailored for action recognition.

Mixup Zhang et al. (2017) is a widely used data augmentation methods which encourages the model to behave linearly in-between training samples. Ghost Motion is related to mixup as both methods use interpolation to generate new training samples. Yet, we argue that these two methods enlarge the input space from two perspective and are complementary to each other. Mixup convex combine pairs of samples and their labels which can be understood as introducing more inter class-wise data. The success of mixup can be mostly attributed to the linear relationship between training samples and corresponding labels because of the regularization effects it brings Carratino et al. (2020). Ghost Motion does not touch the label of training sapmples and can be considered as doing interpolation between the original input space and the generated input space which introduces more smoothed intra class-wise data.

3 Methodology

3.1 Preliminaries

Given a video $x \in X$ containing T frames $x = \{f_1, f_2, ..., f_T\}$ with its one-hot class label $y \in Y$, we denote X as the input video space, $Y = [0, 1]^K$ and K is the number of classes. A 2D network F maps x to the vector $p = F(x)$ where $p = \{\rho_1, \rho_2, ..., \rho_T\}$ corresponds to the predicted logits of T frames. The final prediction P of video x is formed as:

$$P = \mathcal{G}(\rho_1, \rho_2, ..., \rho_T), \tag{1}$$

where $\mathcal{G}$ is the aggregation function and the most common adapted form is the average function. Incorporated with standard categorical cross-entropy loss, the final loss function is defined as:

$$\begin{aligned} \mathcal{L} &= -\sum_{k=1}^{K} y_k \log \frac{e^{P_k}}{\sum_{j=1}^{K} e^{P_j}} \\ &= -\sum_{k=1}^{K} y_k \log \frac{e^{\mathcal{G}(\rho_{k_1}, \rho_{k_2}, ..., \rho_{k_T})}}{\sum_{j=1}^{K} e^{\mathcal{G}(\rho_{j_1}, \rho_{j_2}, ..., \rho_{j_T})}}. \end{aligned} \tag{2}$$

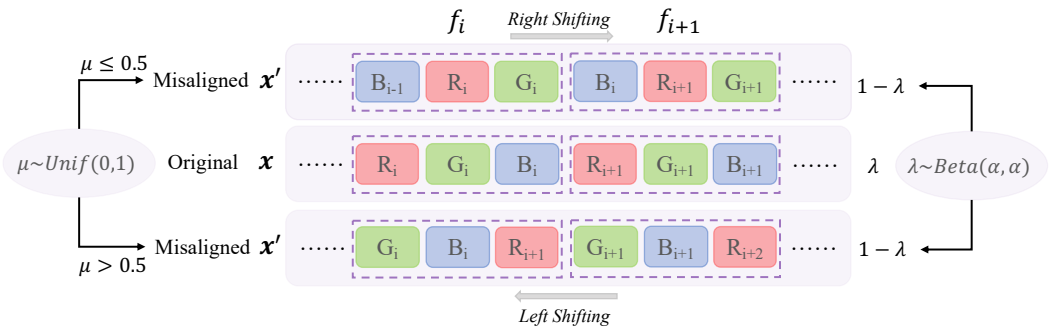


Figure 2: Illustration of Ghost Motion. Given a video $x = \{f_1, f_2, ..., f_T\}$, its channels will be shifted along the temporal dimension to generate a misaligned video x' and the shifting direction is decided by a random probability μ sampling from uniform distribution. Interpolation beween original video and misaligned video will be implemented to generate the Ghost Video. The coefficient λ is sampled from beta distribution to control the degree of interpolation.

As $y_k = 1$ and $y_h = 0$ for all $h \neq k$, minimizing this loss function equals maximizing the log-likelihood of the right label. The maximum can hardly be achieved with a finite P_k, but it can be approached if $P_k \gg P_h$ which means that the logit corresponding to the correct label is much greater than other logits. With this objective, it may result in overfitting as the model is encouraged to produce overconfident predictions and brings difficulties to generalization. This effect may be even disastrous in video recognition as:

$$P_k \gg P_h \rightarrow \mathcal{G}\left(\rho_{k_1}, \rho_{k_2}, ..., \rho_{k_T}\right) \gg \mathcal{G}\left(\rho_{h_1}, \rho_{h_2}, ..., \rho_{h_T}\right), \tag{3}$$

which can be achieved more easily if there is an imbalance distribution of $\{\rho_1, \rho_2, ..., \rho_T\}$ in each video. Unfortunately, motion-irrelevant frames exists in many video samples. The tendency of giving overconfident predictions may result in large logits of background frames which will overwhelm other frames' predictions and lead to wrong final decision.

3.2 Logits Smoothing

There are challenging datasets in video recognition such as Something-Something V1&V2 which involve significant temporal dependency and consists of abundant motion-irrevelant frames. This brings difficulties to optimizations and results in overfitting. To prevent the final predictions being overwhelmed by background frames, we propose to smooth the logits of each sample on these datasets. This procedure is similar with Agarwala et al. (2020) where it is applied in image recognition and the temperature is tuned delicately to improve the accuracy. However, our aim is to prevent overconfident predictions of background frames and relieve the overfitting problem. We introduce a hyperparameter temperature τ and assign it to each logits, so the standard cross-entropy loss can be redefined as:

$$\begin{aligned} \mathcal{L} &= -\sum_{k=1}^{K} y_k \log \frac{e^{\frac{P_k}{\tau}}}{\sum_{j=1}^{K} e^{\frac{P_j}{\tau}}} \\ &= -\sum_{k=1}^{K} y_k \log \frac{e^{\frac{\mathcal{G}\left(\rho_{k_1}, \rho_{k_2}, \cdots, \rho_{k_T}\right)}{\tau}}}{\sum_{j=1}^{K} e^{\frac{\mathcal{G}\left(\rho_{j_1}, \rho_{j_2}, \cdots, \rho_{j_T}\right)}{\tau}}}. \end{aligned} \tag{4}$$

As $\mathcal{G}$ is the average function, this operation equals smoothing the logits of every frame. When $\tau > 1$, the gap between $\frac{P_k}{\tau}$ and $\frac{P_h}{\tau}$ can be reduced which alleviates the overfitting problem. While temperature scaling is commonly used in model calibration Guo et al. (2017), it only applies temperature during testing for calibration which does not help to improve the generalization of models.

3.3 Ghost Motion

In this section, we introduce our proposed data augmentation method Ghost Motion (GM) which is specially designed for video recognition. Considering a RGB frame $f_i \in \mathbb{R}^{\left(C_i^R, C_i^G, C_i^B\right) \times H \times W}$ with 3 channels in total, we can denote the video sequence as $x \in \mathbb{R}^{T \times \left(C^R, C^G, C^B\right) \times H \times W}$.

Given a random probability μ, we generate a misaligned video $x' \in X'$ by shifting channels for one step along the temporal dimension with a random direction, so that we can represent the generated misaligned frame as:

$$\begin{cases} f_i{}' \in \mathbb{R}^{\left(C_i^G, C_i^B, C_{i+1}^R\right) \times H \times W}, & \mu \leq 0.5, \\ f_i{}' \in \mathbb{R}^{\left(C_{i-1}^B, C_i^R, C_i^G\right) \times H \times W}, & \mu > 0.5. \end{cases} \tag{5}$$

With this operation, we can enable information exchange in x'. The misaligned frame $f_i{}'$ is composed of three channels, with two channels from the original frame and one from adjacent frame which explicitly facilitate information interaction between neighboring frames. However, directly using the misaligned video as the input source may not be optimal because the generation follows a fixed pattern and the discrete variance may not be favorable. Therefore, we interpolate between the original video and misaligned video, and sample a coefficient λ from the Beta distribution $Beta\left(\alpha, \alpha\right)$ to control the strength. In our experiments, we set α to 1 so that λ is sampled from the uniform distribution $(0, 1)$ and we can denote the generated training sample $\tilde{x}$ and the corresponding label as:

$$\begin{cases} \tilde{x} = \lambda x + \left(1 - \lambda\right) x', \\ \tilde{y} = y. \end{cases} \tag{6}$$

We name $\tilde{x}$ ghost video as there exists artifacts in certain frames. For adjacent frames with great motion variance, the shifting operation will transfer semantic information to other frames, resulting in motion artifacts which looks like the RGB difference data. The reason is motion patterns locates in different spatial regions for neighboring frames and the overlay of channels from different frames will preserve the motion information of both. With this design, the generated data will enforce the model to focus more on the temporal variance guided by the motion artifacts which will be beneficial for temporal modeling. Besides, motion information may be diffused into background frames and increase their chance to be correctly classified to alleviate the imbalance in aggregation function $\mathcal{G}$. Though the shifting operation leads to disorder in channels, we find the color alteration is beneficial to further improve the generalization. Note that x' is produced by x, so ghost video $\tilde{x}$ shares the same label with original video sample x. We denote the generated channels of ghost frame $\tilde{f}_i$ as:

$$\begin{cases} \tilde{C}_i^R = \lambda C_i^R + \left(1 - \lambda\right) C_i^G, \\ \tilde{C}_i^G = \lambda C_i^G + \left(1 - \lambda\right) C_i^B, \\ \tilde{C}_i^B = \lambda C_i^B + \left(1 - \lambda\right) C_{i+1}^R, \end{cases} \mu \leq 0.5, \quad \begin{cases} \tilde{C}_i^R = \lambda C_i^R + \left(1 - \lambda\right) C_{i-1}^B, \\ \tilde{C}_i^G = \lambda C_i^G + \left(1 - \lambda\right) C_i^R, \\ \tilde{C}_i^B = \lambda C_i^B + \left(1 - \lambda\right) C_i^G, \end{cases} \mu > 0.5, \tag{7}$$

where $\tilde{C}_i^R, \tilde{C}_i^G, \tilde{C}_i^B$ stand for the RGB channels of the ghost frame $\tilde{f}_i$ respectively. Due to the introduced λ, the degree of temporal variance can be measured with a linear relation, leading to continuous variance in the input space.

4 Empirical Validation

In this section, we empirically evaluate the performance of the proposed method on various video recognition datasets. We first validate the effectiveness of Ghost Motion (GM) on different models and various datasets. Second, we present the comparisons of our method with competing data augmentation methods in image recognition and prove that GM is compatible with them. Further, we extend GM to other structures and depths of network. Finally, we conduct comprehensive experiments to explain Logits Smoothing, and present ablation results of other design choices of GM and visualizations which show GM can explicitly guide the model to focus on the motion variance.

4.1 Experimental Setup

Datasets. Our method is evaluated on six widely-used video datasets. For datasets which contain strong temporal dependency: (1) Something-Something V1&V2 Goyal et al. (2017) are two human action datasets which includes 98k and 194k videos respectively. For datasets which require less temporal modeling: (2) ActivityNet-v1.3 Caba Heilbron et al. (2015) contains 10,024 training videos and 4,926 validation videos with 200 categories and the average duration is 117 seconds; (3) UCF101 Soomro et al. (2012) dataset originally consists of 13,320 videos with 101 classes and we adopt the first training/testing split for training and evaluation; (4) HMDB51 Kuehne et al. (2011) dataset is made up of 6,766 videos with 51 classes and we use the original first training/testing split

Table 1: Evaluation of Ghost Motion (GM) on Something-Something datasets with different models. The best results are bold-faced. Note that we does not apply random flipping on these two datasets, so the reproduced results of TANet and TDN are lower than the number reported in their papers.

Method	Sth-Sth V1$^{\Theta}$			Sth-Sth V2$^{\Theta}$		
	Acc1.(%)	Acc5.(%)	Δ Acc1.(%)	Acc1.(%)	Acc5.(%)	Δ Acc1.(%)
TSN Wang et al. (2016)	17.21	43.17	+1.02	30.56	61.21	+1.23
TSN+GM	**18.23**	44.88		**31.79**	62.52	
TSM Lin et al. (2019)	45.34	74.45	+1.54	59.14	85.60	+0.69
TSM+GM	**46.88**	75.76		**59.83**	85.96	
TANet Liu et al. (2021)	45.94	75.56	+1.10	59.17	85.18	+0.87
TANet+GM	**47.04**	76.05		**60.04**	86.25	
TEA Li et al. (2020)	48.67	77.94	+0.94	61.11	87.20	+0.64
TEA+GM	**49.61**	78.03		**61.75**	87.04	
TDN Wang et al. (2021)	49.69	78.74	+0.82	63.10	88.12	+0.71
TDN+GM	**50.51**	79.06		**63.81**	88.59	

Table 2: Evaluation of Ghost Motion (GM) on UCF101, HMDB51, ActivityNet and Mini-Kinetics with TSM Lin et al. (2019) and TEA Li et al. (2020). The best results are bold-faced.

Method	UCF101		HMDB51		ActivityNet		Mini-Kinetics	
	Acc1.(%)	Δ Acc1.(%)	Acc1.(%)	Δ Acc1.(%)	mAP(%)	Δ mAP(%)	Acc1.(%)	Δ Acc1.(%)
TSM	79.57	+1.00	48.63	+2.48	73.10	+1.62	74.11	+0.71
TSM+GM	**80.57**		**51.11**		**74.72**		**74.82**	
TEA	80.10	+1.05	49.61	+3.00	73.09	+1.40	75.96	+0.55
TEA+GM	**81.15**		**52.61**		**74.49**		**76.51**	

for training and evaluation as well; (5) Kinetics400 Kay et al. (2017) is a large scale video recognition dataset which is categorized into 400 action classes. Mini-Kinetics is a subset of Kinetics400 with 200 categories in total and we select the classes following Meng et al. (2020).

Implementation details. We uniformly sample 8 frames for each video unless specified. During training, the training data is random cropped to 224×224 following Zolfaghari et al. (2018), and we did not perform random flipping on Something-Something V1&V2. At inference stage, all frames are center-cropped to 224×224 and we adopt one-crop one-clip per video during evaluation for efficiency. Similar with CutMix Yun et al. (2019), we add a hyperparamter ϱ which denotes the probability whether to implement GM for every batch samples. We set ϱ as 0.5 on Something-Something V1&V2 which are temporal-dependent and 0.2 in other datasets. Note that we only apply Logits Smoothing on Something-Something V1&V2 to calibrate the overconfident predictions and we use the symbol Θ to denote that this dataset is trained with Logits Smoothing. Due to limited space, more implementation details can be found in supplementary material.

4.2 Main Results

Results on different methods. We first empirically evaluate the performance of Ghost Motion (GM) on Something-Something V1 and V2 datasets using different models in Tab. 1. It can be observed that GM consistently increases the generalization ability of these methods on both datasets, including the state-of-the-art model TDN Wang et al. (2021) in 2D methods. The improvements on Sth-Sth V1 are generally larger than those on Sth-Sth V2 as V2 contains more training samples and models will suffer less from overfitting on this dataset. Specifically, the accuracy of TEA Li et al. (2020) can be improved to 49.61% on Sth-Sth V1 which is similar with our reproduced results of TDN Wang et al. (2021), adapting more complex structures and taking much longer training time.

Results on small scale datasets. We first validate Ghost Motion on UCF101, HMDB51 and ActivityNet datasets in Tab. 2. The improvements brought by GM are, quite obvious, greater than 1% as the scales of these datasets are relatively small and the overfitting phenomenon is more severe on them. GM can effectively alleviate overfitting on these datasets.

Results on large scale datasets. We further implement our method on Mini-Kinetics which is one of the large scale datasets in video recognition but GM still increases the generalization ability of the baseline models which validates the effectiveness of our method. Note that Mini-Kinetics has the same number of samples per class with Kinetics400 which means the overfitting phenomenon will not be more severe on this dataset and we conduct experiments on it for efficiency.

Table 3: Comparisons with competing data augmentation methods on different datasets using TSM Lin et al. (2019). The best results are bold-faced.

Method	UCF101		HMDB51		Sth-Sth V1$^{\Theta}$	
	Acc1.(%)	Δ Acc1.(%)	Acc1.(%)	Δ Acc1.(%)	Acc1.(%)	Δ Acc1.(%)
TSM	79.57	-	48.63	-	45.34	-
TSM+Cutout DeVries & Taylor (2017)	79.04	−0.53	48.69	+0.09	44.68	−0.66
TSM+CutMix Yun et al. (2019)	79.16	−0.41	48.17	−0.46	44.99	−0.35
TSM+Mixup Zhang et al. (2017)	79.83	+0.26	50.39	+1.76	46.03	+0.69
TSM+AugMix Hendrycks et al. (2019)	79.94	+0.40	50.59	+1.96	46.23	+0.89
TSM+RandAugment Cubuk et al. (2020)	82.13	+2.16	52.55	+3.92	47.20	+1.86
TSM+GM	80.57	+1.00	51.11	+2.48	46.88	+1.54
TSM+GM+Mixup	80.99	+1.42	51.44	+2.81	47.19	+1.85
TSM+GM+AugMix	81.10	+1.53	52.68	+4.05	46.93	+1.59
TSM+GM+RandAugment	**83.24**	**+3.67**	**53.79**	**+5.16**	**48.06**	**+2.72**

Table 4: Evaluation of Ghost Motion (GM) on Something-Something V1 with different depths.

Method	Sth-Sth V1$^{\Theta}$	
	Acc1.(%)	Δ Acc1.(%)
TSM(R18)	39.04	+0.76
TSM(R18)+GM	**39.80**	
TSM(R101)	45.85	+2.20
TSM(R101)+GM	**48.05**	

Table 5: Evaluation of Ghost Motion (GM) on Something-Something V1 and Mini-Kinetics with SlowOnly Feichtenhofer et al. (2019) and Uniformer Li et al. (2022).

Method	Sth-Sth V1$^{\Theta}$		Mini-Kinetics	
	Acc1.(%)	Δ Acc1.(%)	Acc1.(%)	Δ Acc1.(%)
SlowOnly	41.81	+1.19	65.29	+0.61
SlowOnly+GM	**43.00**		**65.90**	
Uniformer	41.74	+0.82	71.47	+1.00
Uniformer+GM	**42.56**		**72.47**	

4.3 Comparison with Other Approaches

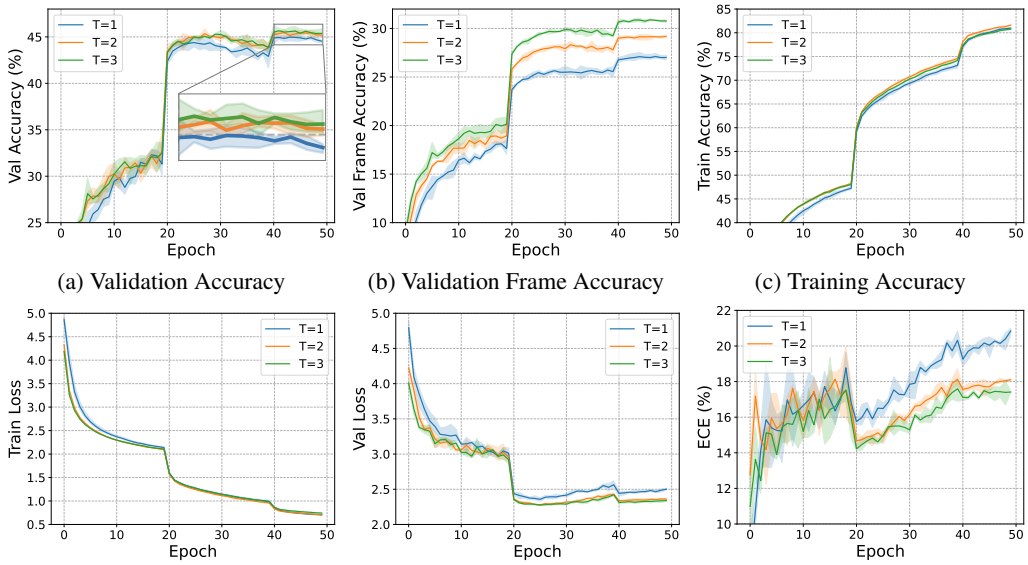
We compare our method with competing data augmentation methods on UCF101, HMDB51 and Something-Something V1 based on TSM in Tab. 3. It can be observed that popular data augmentation methods like Cutout DeVries & Taylor (2017), CutMix Yun et al. (2019) results in negative effects in video recognition. One possible explanation is that these methods break the motion pattern in video sequences and bring challenges to temporal modeling which is extremely essential in video recognition. Mixup Zhang et al. (2017) shows consistent improvement on these three datasets as it enforces a regularization effects which is helpful to deal with overfitting. AugMix Hendrycks et al. (2019) and RandAugment Cubuk et al. (2020) exhibits great generalization ability as both methods involve multiple augmentation operations and their magnitudes will be adaptively sampled.

Our method outperforms all approaches except for RandAugment, which is composed of various augmentation operations. However, note that GM is compatible with these methods as they only consider image-level transformations and GM mainly focuses on temporal-wise interactions to enlarge the input space. We further combine GM with Mixup, AugMix and RandAugment on these three datasets and the performance can be improved by 3.67%, 5.16% and 2.72% at most.

4.4 Further Exploration

Results on different depths of network. In previous experiments, all methods are built on ResNet-50 He et al. (2016) and we further conduct experiments on ResNet-18 and ResNet-101 to verify the effectiveness of our method when evaluated on models with different representation abilities. We can observe from Tab. 4 that GM consistently improves the performance of baseline method and the improvements are increasing (0.76%, 1.54%, 2.20%) with a deeper network. The explanation is that the overfitting phenomenon is more obvious on deeper network as it has stronger representation ability to fit the training data and GM can effectively alleviate this phenomenon which leads to greater improvements in accuracy compared with shallow networks.

Results on different structures of network. Although Logits Smoothing and Ghost Motion are inspired by the analysis in 2D-based methods, our method can be implemented on other video recognition architecture as it does not touch the structures inside the networks. Therefore, we implement our method on 3D-based method SlowFast Feichtenhofer et al. (2019) with the slow branch only and Transformer-based method Uniformer Li et al. (2022) to test the generalization capability of our method. The results in Tab. 5 show that GM leads to improvements in accuracy on both structures and datasets, indicating our method can be applied to other structures to improve their generalization capability as well. The possible explanation is that the data generated by GM has motion artifacts



(a) Validation Accuracy (b) Validation Frame Accuracy (c) Training Accuracy

(d) Training Loss (e) Validation Loss (f) Expected Calibration Error

Figure 3: Validation Accuracy, Validation Frame Accuracy, Training Accuracy, Validation Loss, Training Loss, Validation Loss and Expected Calibration Error (ECE) curves of TSM trained with Temperature $T = 1, 2, 3$ on Something-Something V1 dataset.

and the explicit pattern in data will enforce the model to focus more on the temporal variance which is beneficial for temporal modeling, regardless of what type of network it is. Note that we use one-crop one-clip per video during evaluation for efficiency in all experiments and we believe adopting the same procedure will not bring unfairness in comparisons.

4.5 Ablation and Analysis

4.5.1 What does logits smoothing bring?

In this part, we empirically evaluate the performance of introducing Logits Smoothing and try to explain what effects it brings. We train TSM Lin et al. (2019) with $T = 1, 2, 3$ on Something-Something V1 respectively and compare various metrics, where Frame Accuracy denotes the average accuracy of all sampled frames in the training set.

Shown in Fig. 3 (a)(b)(c), we first observe a disparity between training and validation accuracy which implies that overfitting is indeed a severe problem in video recognition. The model trained with higher temperature has clearly higher frame-wise accuracy which leads to advantages in accuracy of video samples as well. However, the improvement on frame accuracy is more obvious than the increase on video accuracy. It proves that there exists an imbalanced relationship between frames and the final prediction of video samples may be mislead by the motion-irrelevant frames which meets our motivation for designing Ghost Motion.

From Fig. 3 (d)(e)(f), one can see that TSM trained with different temperatures obtain similar training loss but have smaller validation loss with higher temperature. Expected Calibration Error (ECE) Guo et al. (2017) is proposed to measure the gap between accuracy and the confidence score. Training with higher temperature results in a smaller ECE which meets our expectation as overconfident models are risky in video recognition as we demonstrated in Sec. 3.2. Overall, logits smoothing enforces the model to less fit the training data and improves its generalization ability.

4.5.2 Why does Ghost Motion help?

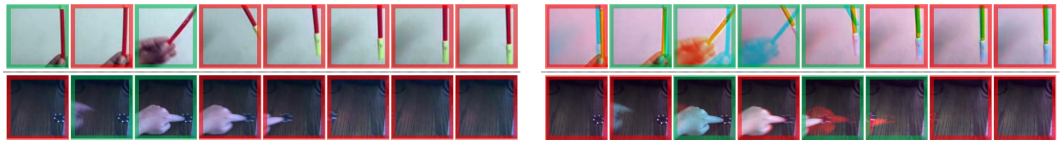
In this part, we first empirically compare different design choices of Ghost Motion (GM) in Tab. 6. At first, we explore other selections of hyperparameter α which control the beta distribution and the results show that $\alpha = 1$ results in the best performance. Besides, we compare GM with its variants which shift channels for two step and shift for single direction. The two-step shift obtains little worse results than one-step shifts and the possible reason is that two-step shift will lead to an offset in frame location. GM outperforms one-way shift obviously as it offers more variance in data

which is helpful to improve the generalization of baseline method. Moreover, we implement GM in the feature space following Verma et al. (2019) which applies Mixup Zhang et al. (2017) in the intermediate features. However, results show that this measure leads to negative effects compared with GM and we consider the misaligned channel order caused by GM in the feature space may be one of the possible reasons.

As implementing GM can lead to alteration of the video color, we compare GM with Color Jittering(CJ): randomly changing the brightness, contrast, saturation and hue of the video. We can observe an advantage in GM as it exchanges motion information between adjacent frames which is helpful for temporal modeling compared to Color Jittering which only adjusts the color. Another important design of GM is the interpolation between the original video and the misaligned video. Therefore, we compare GM with its variant which directly using the misaligned video as the input. The performance of GM without interpolation is lower than our method which demonstrates the superiority of this design as it offers a continuous distribution of the input data which enlarges the input space and is beneficial for classification.

Table 6: Ablation of design choices of Ghost Motion on ActivityNet. The best results are bold-faced and mAPs are averaged over multiple runs.

Method	Specification	mAP(%)
TSM Lin et al. (2019)	-	73.10±0.14
TSM+GM	$Beta\,(0.5, 0.5)$	74.35±0.09
TSM+GM	$Beta\,(2.0, 2.0)$	73.98±0.07
TSM+GM	Two-Step Shift	74.33±0.22
TSM+GM	One-Way Shift	73.91±0.03
TSM+GM	Manifold	73.22±0.15
TSM+CJ	Color Jittering	73.89±0.25
TSM+GM	w/o interpolation	74.03±0.05
TSM+GM	-	**74.72±0.04**



(a) Frame-wise predictions of TSM trained with standard protocol. Original frames are shown.

(b) Frame-wise predictions of TSM trained with Ghost Motion. Frames generated by GM are shown.

Figure 4: Visualization results on Something-Something V1 dataset. Frames annotated with red rectangles are wrongly classified and the green ones represent correctly categorized frames.

Moreover, we visualize the frame-wise predictions of TSM on Something-Something V1. Shown in Fig. 4, the left two samples are wrongly classified by the model trained with standard protocol as the final predictions are overwhelmed by motion-irrelevant frames. In contrast, the right samples are generated by GM and there exists motion artifacts in some frames as each frame exchanges parts of the channels with adjacent frames. The motion artifacts looks like the RGB difference between two consecutive frames which describe the appearance change and associate with motion salient regions. However, compared to RGB difference, the data generated by GM does not loss information on the static regions as we only shift parts of the channels rather than subtracting the similar parts. Therefore, the data generated by GM could be seen as a new modality which restore the good attributes of RGB and RGB difference. As a result, the model trained with GM can focus more on motion frames guided by the motion artifacts, which prevents the model from overfitting background frames towards the ground truth label and results in better generalization ability.

5 Conclusion

This paper probes into the overfitting problem in video recognition and proposes a special designed data augmentation method Ghost Motion (GM). Specifically, GM shifts channels along the temporal dimension to enforce motion-relevant information exchange among neighboring frames. This can enhance the learning process as motion information can be diffused into other frames which is beneficial for temporal modeling. Besides, we find the tendency of giving overconfident predictions is risky in 2D based methods as there is a high chance motion-irrelevant frames will overwhelm other frames' predictions. Therefore, we smooth the logits during training to prevent overconfident predictions on datasets with strong temporal dependency (e.g., Something-Something). Comprehensive experiments on popular datasets and methods demonstrate the effectiveness of GM and it can be further improved when combined with existing data augmentation approaches.

ETHICS STATEMENT

In our paper, we strictly follow the ICLR ethical research standards and laws. To the best of our knowledge, our work abides by the General Ethical Principles.

REPRODUCIBILITY STATEMENT

We adhere to ICLR reproducibility standards and ensure the reproducibility of our work. All datasets we employed are publicly available. As we have included comprehensive empirical validations on various methods and datasets in our work, we provide corresponding training setting for our method in the appendix. Moreover, we will provide the code to reviewers and area chairs using an anonymous link when the discussion forums is open.

REFERENCES

Atish Agarwala, Jeffrey Pennington, Yann Dauphin, and Sam Schoenholz. Temperature check: theory and practice for training models with softmax-cross-entropy losses. *arXiv preprint arXiv:2010.07344*, 2020.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

Luigi Carratino, Moustapha Cissé, Rodolphe Jenatton, and Jean-Philippe Vert. On mixup regularization. *arXiv preprint arXiv:2006.06049*, 2020.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019.

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 *ICCV*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Jinhyung Kim, Seunghwan Cha, Dongyoon Wee, Soonmin Bae, and Junmo Kim. Regularization on spatio-temporally smoothed feature for action recognition. In *CVPR*, 2020.

Anders Krogh and John Hertz. A simple weight decay can improve generalization. *NeurIPS*, 1991.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pp. 2556–2563. IEEE, 2011.

Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv preprint arXiv:2201.09450*, 2022.

Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *CVPR*, 2020.

Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019.

Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. Tam: Temporal adaptive module for video recognition. In *ICCV*, 2021.

Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. Ar-net: Adaptive frame resolution for efficient action recognition. In *ECCV*, 2020.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *ICML*, 2019.

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 *ECCV*, 2016.

Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *CVPR*, 2021.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *ECCV*, 2018.

A APPENDIX

A.1 DATA PRE-PROCESSING

We uniformly sample 8 frames for each video on all datasets unless specified. During training, the training data is randomly cropped to 224 $\times$ 224 following Lin et al. (2019), and we perform random flipping except for Something-Something Goyal et al. (2017) datasets. At inference stage, all frames are center-cropped to 224 $\times$ 224, obtaining a one-crop one-clip per video for evaluation unless specified.

A.2 IMPLEMENTATION DETAILS

All results reported in our paper are run on NVIDIA Tesla V100 GPUs with mixed precision. All the models are pretrained on ImageNet except for the experiment of Uniformer Li et al. (2022) on Something-Something V1 which utilize the pre-trained model on Kinetics400 following the original setting. We reimplement all 2D methods in our code and use the official implementations for Slowonly Feichtenhofer et al. (2019)[1] and Uniformer[2].

For Ghost Motion, α is set to 1 in all our experiments and we change the hyperparamter ϱ which denotes the probability whether to implement GM for every batch samples on different datasets. Note that we only apply Logits Smoothing to temporal-dependent datasets such as Something-Something V1&V2.

A.2.1 TEMPORAL SENSITIVE DATASETS

We set ϱ to 0.5 to enable more variance in training samples on Something-Something V1&V2 as the overfitting phenomenon is the most severe on these two datasets. For Logits Smoothing, the temperature is set to 3,2 for Something-Something V1&V2, respectively.

A.2.2 TEMPORAL INSENSITIVE DATASETS

We set ϱ to 0.2 for Ghost Motion on Mini-Kinetics Kay et al. (2017), ActivityNet Caba Heilbron et al. (2015), HMDB51 Kuehne et al. (2011) and UCF101 Soomro et al. (2012) and we did not apply Logits Smoothing on these datasets.

[1] https://github.com/facebookresearch/SlowFast

[2] https://github.com/Sense-X/UniFormer/tree/main/video$_classification$