Learning from Temporal Gradient for Semi-supervised Action Recognition

Junfei Xiao[1] Longlong Jing[2] Lin Zhang[3] Ju He[1] Qi She[4]
Zongwei Zhou[1] Alan Yuille[1] Yingwei Li[1]

[1]Johns Hopkins University [2]The City University of New York
[3]Carnegie Mellon University [4]ByteDance

Abstract

*Semi-supervised video action recognition tends to enable deep neural networks to achieve remarkable performance even with very limited labeled data. However, existing methods are mainly transferred from current image-based methods (*e.g.*, FixMatch). Without specifically utilizing the temporal dynamics and inherent multimodal attributes, their results could be suboptimal. To better leverage the encoded temporal information in videos, we introduce temporal gradient as an additional modality for more attentive feature extraction in this paper. To be specific, our method explicitly distills the fine-grained motion representations from temporal gradient (TG) and imposes consistency across different modalities (*i.e.*, RGB and TG). The performance of semi-supervised action recognition is significantly improved without additional computation or parameters during inference. Our method achieves the state-of-the-art performance on three video action recognition benchmarks (*i.e.*, Kinetics-400, UCF-101, and HMDB-51) under several typical semi-supervised settings (*i.e.*, different ratios of labeled data).*

1. Introduction

As a fundamental task for video understanding, video action recognition has drawn much attention from the community and industry [2, 6, 10, 11, 47, 49, 55, 59]. Unlike image-related tasks, networks for video-related tasks are normally more easy to overfit due to the complexity of the tasks [25,47,48]. The common practice is to firstly pre-train the network on large-scale datasets (e.g., Kinetics [5] of up to $650,000$ video clips) and then finetune on downstream small datasets to obtain better performance [11, 17, 34, 35].

However, since annotating large-scale video datasets is time-consuming and expensive, training models on a large collected dataset with complete annotations is impeded. To utilize large-scale datasets with acceptable costs, some researchers have turned to design semi-supervised learning models which have good generalization ability with limited annotations [23, 39, 58, 62]. As pseudo-label based methods (e.g., FixMatch [40] and MixMatch [3]) have shown outstanding performance on semi-supervised image classification, most previous video-based methods are heavily built on them to utilize the unlabeled data. Although these preliminary attempts have obtained acceptable results, most methods [23, 62] are just taking video clips as 'images' in 3D without further consideration of the video properties.

Videos are significantly different from images, and the key differences are the temporal information span in multiple frames and the inherent multimodal property. The

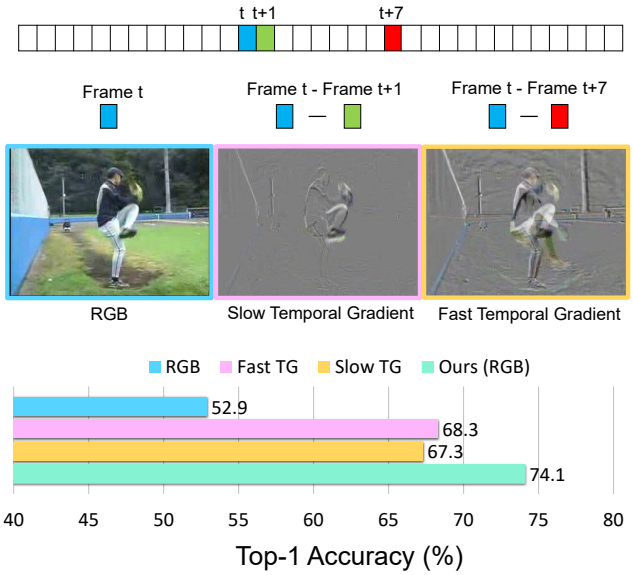

Figure 1. **Top:** A sketch diagram describing the formulation of different experimented modalities (*i.e.*, RGB, slow temporal gradient and fast temporal gradient). **Bottom:** A comparison of Top-1 accuracy with the baseline FixMatch [40] semi-supervised learning method. The chart compares the performance generated with different input modalities (*i.e.*, RGB, slow TG and fast TG). The remarkable performance with TG motivates us to figure out a way to efficiently utilize this fruitful modality. By learning from temporal gradient, our model is able to significantly outperform the models taking either temporal gradient or RGB frames as input.

temporal information refers to the motion signal between frames, and usually the features of contiguous frames from the same video change smoothly. The multimodal consistency refers to the features of multiple modalities in the same video clip should be consistent since they are encoding the same content. Without special designs to specifically focus on the temporal information and multimodal consistency, the potential of semi-supervised action recognition is not fully unleashed.

Some previous studies [53, 61] introduce temporal gradient[1] as an additional modality to better utilize the temporal information encoded in videos as it is rich in high-quality motion signals. Temporal gradient can be formulated as:

$$TG = x_t^{RGB} - x_{t+n}^{RGB}, \tag{1}$$

where x represents a video, t denotes the frame index and n denotes the interval for calculating temporal gradient.

Inspired by these studies, we made a trial with temporal gradient under the semi-supervised settings and found that a much better performance could be generated when the input frames in RGB are replaced with temporal gradient. As shown in Figure 1, the Top-1 accuracy of temporal gradient is $\sim$25% higher than using the RGB as input on UCF-101 dataset with only 20% of labeled data for training.

Why the temporal gradient is so much better than the RGB frames when the training data is very limited? We hypothesize that the key is in the detailed and fine-grained motion signals encoded in the temporal gradient. The gradient along the temporal dimension is color-invariant and explicitly encodes the representative motion information of the actions in video. This helps models generalize much easier when the labels are extremely limited. Therefore, in this paper, we propose to train a semi-supervised action recognition RGB based model to mimic both the fine-grained and high-level features from the temporal gradient.

We start from FixMatch [40], a typical pseudo-label based semi-supervised model, as the baseline framework. However, without any further constraints in the feature level, pseudo-label based methods perform poorly in the case of very limited labels, as many generated pseudo-labels are inaccurate. Therefore, we propose two constraints to help the model extract temporal information in video with multiple modalities and improve the consistency between the multimodal representations. To mimic the detailed and fine-grained motion signals from temporal gradient, we propose a knowledge distillation strategy with dense alignment in block-wise to help the student RGB model deeply and effectively learn from the teacher temporal gradient model. To learn the high-level features from temporal gradient, we further perform the contrastive learning between the features from RGB and temporal gradient sequences to enforce the high-order similarity. Given the two constraints at the feature level, our proposed model is able to achieve much better performance.

Compared with the existing methods, our model has these unique advantages: (1) our model distills the knowledge from temporal gradient to the RGB-based network during the training, and only the RGB model is required for inference. Therefore, there is no additional computation or parameters needed. (2) Our proposed model is extremely simple yet effective. We conducted experiments on multiple public action recognition benchmarks including UCF-101, HMDB-51, and Kinetics-400. Our proposed method significantly outperforms all the state-of-the-art methods by a large margin.

2. Related Work

Semi-supervised learning in images. The semi-supervised image classification task has been well studied and many methods have been proposed including Pseudo-Label [28], S4L [60], MeanTeacher [44], MixMatch [3], UDA [57], FixMatch [40], UPS [36], etc. The Pseudo-Label [28] is an early method which uses the confidence (softmax probabilities) of the unlabeled data as labels and to train the network jointly with a small ratio labeled data and much more unlabeled data. Many improved versions of Pseudo-Label have been proposed while the key is to improve the quality of the labels [36, 40]. Following a state-of-the-art method on image classification — FixMatch [40], many FixMatch-alike methods achieve the state-of-the-art performance on many other tasks including detection [51], segmentation [63], etc. Although these methods achieve remarkable performance on image-based tasks, some recent studies show that the performances are not satisfying when directly applying these methods to video semi-supervised tasks [23, 39].

Semi-supervised learning in videos. Although there have been a few semi-supervised video action recognition methods [23, 39, 58, 62] proposed recently, most of them directly apply the image-based methods to videos with less focus on the temporal dynamics of the videos. VideoSSL [23] made the first attempt to build a benchmark for the video semi-supervised learning task by training the network with ImageNet pre-trained models, which explicitly guides the model to learn the appearance information in each video. It also shows that the existing image-based methods such as Pseudo-Label [28], Mean-Teacher [44] have inferior performance on video semi-supervised benchmarks. TCL [39] is a recently proposed method which jointly optimizes the network by employing a self-supervised auxiliary task and a group contrastive learning. By using multimodal data, the MvPL [58] achieved the state-of-the-art performance by sharing the same model with different modalities as input (RGB, TG, and Optical Flow) and generating pseudo

[1] The difference between two RGB video frames with a short interval.

labels with the "confidence" of multiple modalities. Compared with these methods, our method specifically focusing on learning the temporal information from TG with our proposed constraints and significantly outperforms the state-of-the-art methods on multiple public benchmarks.

Multimodal Video Feature Learning. Videos could be viewed from different modalities while each modality encodes information from a unique perspective. For example, a video in general RGB couples both spatial and temporal information, the temporal gradient is color invariant which mainly encodes the difference between frames, and the optical flow explicitly encodes the motion information for each pixel. The features from different modalities are normally complementary to each other and therefore the feature fusions are normally performed for better performance. The pioneer work is the Two-Stream [13, 38] model which fuses features from both RGB video clips and optical flow clips. With the complementary information from different modalities, the multimodal network is able to achieve better performance [1, 13, 38, 52, 54]. However, there is additional computation and latency during inference. Unlike the normal multimodal feature fusion model, our model distills the motion-related representation from the temporal gradient to the base RGB model, while only the base model and RGB frames are needed during the inference stage. Moreover, our model outperforms the teacher model with only RGB as input during the inference.

Contrastive learning. Contrastive learning methods achieve remarkable performance on the self-supervised learning image classification [4, 7, 15, 18, 30, 46]. The key idea is that the representation can be learned on unlabeled data by minimizing the distance of features of positive pairs (two views of the same data sample) and maximizing the distance of features of negative pairs (two different data samples). Recently, many researchers proposed to use temporal contrastive learning for video self-supervised learning [12, 17, 21, 34, 35]. In this paper, to better utilize the unlabeled data for semi-supervised action recognition, we propose to use the cross-modal contrastive loss to enforce the consistency of features from RGB clips and temporal gradient clips. We demonstrate that the cross-modal contrastive method is very effective for the proposed semi-supervised learning.

3. Method

The objective of our method is to improve the performance on the semi-supervised action recognition task by introducing and utilizing an effective view of videos: Temporal Gradient. The overview of our proposed framework is shown in Figure 2, which consists of three main components: (1) the FixMatch framework with weak-strong augmentation strategy to generate better pseudo-labels for unlabeled data (2) cross-modal dense feature alignment between the features from RGB clips and TG clips for network to learn the fine-grained motion signals, and (3) cross-modal contrastive learning to learn high-level consistency feature across RGB and TG clips. The formulation for each component is introduced in the following subsections.

3.1. FixMatch

Considering a multi-class classification problem, we denote $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N_l}$ as the *labeled* training set, where $x_i \in \mathcal{R}^{T\times H\times W\times 3}$ is the i-th sampled video clip, y_i is the corresponding one-hot ground truth label, and N_l is the number of data points in the labeled set. Similarly, we denote $\mathcal{U} = \{x_j\}_{j=1}^{N_u}$ as the *unlabeled* set, where N_u is the number of data points in the unlabeled set. We use f_θ to denote a classification model with trainable parameters θ. We use $\alpha(\cdot)$ to represent the weak (standard) augmentation (*i.e.*, random horizontal flip, random scaling, and random crop in video action recognition), and $\mathcal{A}(\cdot)$ to represent the strong data augmentation strategies (i.e., Randaugment [9]).

The network f_θ is optimized with each video clip consisting of T frames as x_i. For a mini-batch of *labeled* data $\{(x_i, y_i)\}_{i=1}^{B_l}$, the network is optimized by minimizing the cross-entropy loss $\mathcal{L}_l$ as

$$\mathcal{L}_l = -\frac{1}{B_l}\sum_{i=1}^{B_l} y_i \log f_\theta(\alpha(x_i)), \tag{2}$$

where B_l is the number of labeled samples in each training batch.

For a mini-batch of *unlabeled* data $\{x_j\}_{j=1}^{B_u}$, FixMatch enforces the model to produce consistent predictions of the same unlabeled data sample but with different extent of augmentations. More specifically, pseudo labels $\hat{y}$ for the unlabeled data are usually generated via confidence thresholding as:

$$\mathcal{C} = \{x_j | \max f_\theta(\alpha(x_j)) \geq \gamma\}, \tag{3}$$

where γ denotes a pre-defined threshold and $\mathcal{C}$ is the confident example set from a mini-batch. The confident predictions $f_\theta(\alpha(x_j))$ in the set $\mathcal{C}$ are then transformed to one-hot labels $\hat{y}_j$ by taking *argmax* operation. Then a cross-entropy loss $\mathcal{L}_u$ will be optimized over the samples in $\mathcal{C}$ and its generated one-hot labels as:

$$\mathcal{L}_u = -\frac{1}{B_u}\sum_{x_j\in\mathcal{C}} \hat{y}_j \log f_\theta(\mathcal{A}(x_j)), \tag{4}$$

where B_u is the number of unlabeled samples in the training batch.

With the loss over both labeled data and unlabeled data, the entire FixMatch is optimized with the objective function as:

$$\mathcal{L}_{fm} = \mathcal{L}_l + \mathcal{L}_u. \tag{5}$$

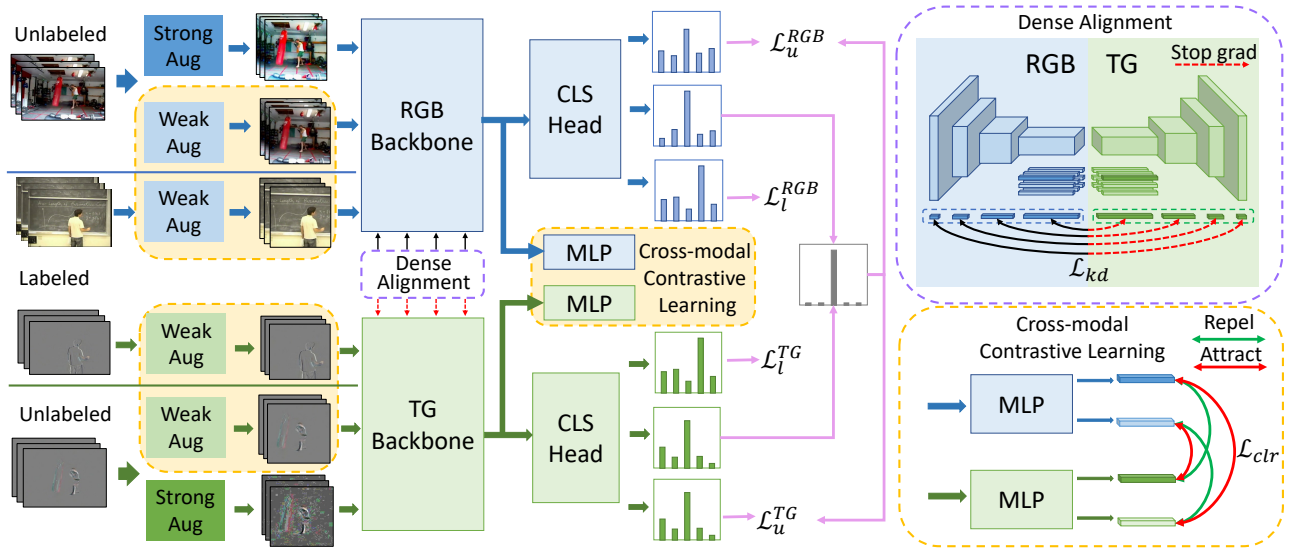

Figure 2. **An overview of our proposed framework.** Our method consists of two parallel models with different modalities (*i.e.*, RGB and TG) of video clips as input. The entire framework is jointly optimized with (1) two parallel FixMatch frameworks with pseudo-labeling, (2) cross-modal dense feature alignment, and (3) cross-modal contrastive learning.

3.2. Parallel Framework for Temporal Gradient

Temporal gradient (TG) ($\frac{\partial V}{\partial t}$) between two RGB frames in a video encodes the appearance change and corresponds to the temporal information that changes dynamically. Therefore, the response is accentuated by the moving objects, especially the boundaries. FixMatch [40] is originally designed for the image classification task and pays little attention to the temporal information of videos, therefore, we extend it to jointly train with RGB and TG to explicitly focus more on capturing the temporal information. To avoid additional computation and delays for processing temporal gradients during model inference on unseen videos, we propose to distill fine-grained motion signals from TG to RGB without introducing extra input or parameters for inference.

The RGB and temporal gradient information are complementary with each other. The RGB encodes spatial and temporal information in a general way, while the temporal gradient has a focus on the motion signals, as illustrated in Figure 1. Therefore, for each video clip, the predictions from both RGB network and TG network are averaged and then used to generate the pseudo labels. In this way, the fused pseudo-label generation is reformulated as:

$$\mathcal{C} = \{x_j | \max f_\theta(\frac{(\alpha(x_j^{RGB}) + \alpha(x_j^{TG}))}{2}) \geq \gamma\}. \quad (6)$$

Having access to both features from RGB and TG, the quality of the fused pseudo labels are more accurate than the predictions from each model alone, and a more detailed ablation study is provided in Section 4.5. The fused pseudo labels will be jointly used with unlabeled data to train both the TG and RGB model. For the temporal gradient model, the training objective is also a summation of Equation (2) and Equation (4) but for TG.

$$\mathcal{L}_{fm}^{TG} = \mathcal{L}_l^{TG} + \lambda_u \mathcal{L}_u^{TG}. \quad (7)$$

3.3. Cross-modal Dense Feature Alignment

To learn detailed fine-grained motions from temporal gradient, we propose to distill the knowledge from temporal gradient model to the RGB model. The similarities between the features from both temporal gradient and RGB clips are minimized by the cross-model dense feature alignment module as:

$$\min \left[\mathcal{D}\left(\mathcal{F}_i^{RGB}, \mathcal{F}_i^{TG}\right)\right], \quad (8)$$

where $\mathcal{F}_i^{RGB}, \mathcal{F}_i^{TG} \in \mathbb{R}^{C_i \times T_i \times H_i \times W_i}$ denote the output features of the i-th block in the RGB and TG models, and $\mathcal{D}$ represents a pairwise function evaluating the representation differences. There are many choices for $\mathcal{D}$ and we experiment with three different functions: L1, L2 and Cosine Similarity losses (shown in Equation (9), where $\|\cdot\|_1$ and $\|\cdot\|_2$ are ℓ_1/ℓ_2-norm). A more detailed discussion is provided in Section 4.5.

$$\begin{aligned} \mathcal{D}_{L1}\left(\mathcal{F}_1, \mathcal{F}_2\right) &= \|\mathcal{F}_1 - \mathcal{F}_2\|_1 \\ \mathcal{D}_{L2}\left(\mathcal{F}_1, \mathcal{F}_2\right) &= \|\mathcal{F}_1 - \mathcal{F}_2\|_2 \\ \mathcal{D}_{cos}\left(\mathcal{F}_1, \mathcal{F}_2\right) &= -\frac{\mathcal{F}_1}{\|\mathcal{F}_1\|_2} \cdot \frac{\mathcal{F}_2}{\|\mathcal{F}_2\|_2}. \end{aligned} \quad (9)$$

An key setting in our online knowledge distillation method is the stop-gradient ($stopgrad$) operation on the

temporal gradient side, which means the teacher model would not receive any gradient from the alignment loss. This helps the TG model avoid degeneration by the alignment with the RGB student model. As shown in Equation (10), the alignment loss term for learning fine-grained motion features is:

$$\mathcal{L}_{kd} = \left[\mathcal{D}\left(\mathcal{F}_i^{RGB}, stopgrad(\mathcal{F}_i^{TG})\right)\right]. \quad (10)$$

3.4. Cross-modal Contrastive Learning

The dense feature alignment explicitly enables the RGB network to mimic the fine-grained motion signals from temporal gradient. We hypothesize that the global high-level representations across different modalities are also valuable and crucial. Therefore, cross-modal contrastive learning is employed as another module to discover the mutual information that coexists in both TG and RGB clips. Following the principle of SimCLR [7] and CMC [46], we form the contrastive learning with positive pairs and negative pairs. Specifically, we consider the two modalities of the same video clip as a positive pair $\{k^+\}$ and the two modalities of different video clips as negative pairs $\{k^-\}$. The learning objective is to maximize the similarity of positive pairs and minimize the similarity of negative ones. We adopt InfoNCE loss [33] as the objective function over the features extracted from RGB and TG:

$$\mathcal{L}_{clr} = -\log \frac{\sum_{k\in\{k^+\}} \exp\left(\text{sim}(q,k)/\tau\right)}{\sum_{k\in\{k^+,k^-\}} \exp\left(\text{sim}(q,k)/\tau\right)}, \quad (11)$$

with τ being a temperature hyper-parameter for scaling. All embeddings are ℓ_2 normalized and dot product (cosine) similarity is used to compare them $\text{sim}(q,k) = q^\top k/\|q\|\|k\|$.

It is worth noting that this cross-modal contrastive learning directly uses all weakly augmented samples of the two modalities ($\alpha(x_i^{RGB/TG})$) in the FixMatch, including both labeled (the labels are not used) and unlabeled data. Therefore, there is no additional computation for the data loading and prepossessing.

Total Loss: Our entire model based is jointly trained with cross-entropy loss over labeled data,cross-entropy loss over the unlabeled data with pseudo-labels, the dense alignment over both labeled and unlabeled data, and the cross-modal contrastive loss over both labeled and unlabeled data. Overall, the final objective function of our method is:

$$\mathcal{L}_{total} = w_{fm}(\mathcal{L}_{fm}^{RGB} + \mathcal{L}_{fm}^{TG}) + w_{kd}\mathcal{L}_{kd} + w_{clr}\mathcal{L}_{clr}. \quad (12)$$

4. Experimental Results

4.1. Datasets and Evaluation

Datasets. Following previous state-of-the-art semi-supervised video action recognition methods [23, 58, 62], we evaluate our method on three public action recognition benchmarks: UCF-101 [41], HMDB-51 [27], and Kinetics-400 [26]. UCF-101 is a widely used dataset which consists of $13,320$ videos belonging to 101 classes. HMDB-51 is a smaller dataset which consists of $6,766$ videos with 51 classes. For UCF-101 and HMDB-51, we follow the data splits that released by VideoSSL [23]. The Kinetics-400 dataset is a large-scale dataset consisting of ∼235k training videos and ∼20k validation videos belonging to 400 classes. For Kinetics-400, we follow the most recent state-of-the-art method MvPL [58] by forming two balanced labeled subsets by randomly sampling 6 and 60 videos per class for 1% and 10% settings.

Evaluation. We report Top-1 accuracy for major comparisons, while Top-5 accuracy is also provided for some ablation studies.

4.2. Implementation Details

Network architecture. For a fair comparison with the state-of-the-art methods [23, 58], the FixMatch [40] framework is used as the backbone model while the 3D ResNet-18 [19, 49] is adopted as feature extractors for both RGB and TG (Section 3.2) modalities. For each feature extractor, two individual contrastive heads with 3-layer non-linear MLP architecture are added for the cross-modal contrastive learning (Section 3.4).

Video augmentations. There are two types of data augmentations: weak augmentation and strong augmentation. For the weak augmentation, the random horizontal flipping, random scaling, and random cropping following [62]. To be specific, given a video clip, we firstly resize the video making the short side be 256, and then a randomly resized crop operation is performed. The cropped clips are then resized to 224×224 pixels and flipped horizontally with a 50% probability. For strong augmentation, the RandAugment [9] is choose which randomly selects a small set of transformations from a large augmentation pools (e.g., rotation, color inversion, translation, contrast adjustment, etc.) for each sample and then perform the selected data augmentation over the samples. It is worth noting that both the teacher (TG) and student (RGB) share the same weak augmentation (*i.e.*, the inputs are identically cropped in the same area and both flipped or not). This provides a direct positional information matching, which plays a crucial role for the dense alignment in Section 3.2.

Training details. All experiments are done with an initial learning of 0.2 on 8 GPUs by following the settings in [11,58,62] using the cosine learning rate decaying scheduler [29] and also a linear warm-up strategy [14]. We use momentum of 0.9 and weight decay of 10^{-4}. Dropout [42] of 0.5 is used before the final classifier layer to reduce the over-fitting. Following [62], each mini-batch consists of 5 labeled data clips and 5 unlabeled data clips, while each

		Kinetics-400				UCF-101				HMDB-51	
		1%		10%		10%		20%		50%	
Alignment	Contrast	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
✗	✗	5.4	17.0	40.2	65.4	38.4	64.8	54.1	78.1	37.8	68.6
✓	✗	9.4	25.5	43.5	68.8	60.4	84.4	74.6	91.7	47.3	74.8
✗	✓	5.2	23.1	42.6	67.4	58.0	82.5	68.6	89.2	46.1	73.8
✓	✓	**9.8**	**26.0**	**43.8**	**69.2**	**62.4**	**84.9**	**76.1**	**92.1**	**48.4**	**75.9**

Table 1. **Effectiveness of the cross-modal alignment and contrastive learning.** The results are evaluated on the validation sets. The first row shows the results of the FixMatch baseline model without any proposed modules.

input clip consists of 8 frames with a sampling stride of 8, which covers 64 frames of the raw video. We consistently train our models with 180 and 360 epochs for all experiments on UCF-101 and HMDB-51, while 45 (1%) and 90 (10%) epochs are trained for Kinetics-400. More training details are provided in the supplementary material. For the pseudo-label threshold, we follow [58] which sets it to 0.3 for getting more training samples. For the loss weights, w_{fm} is set to 0.5 while w_{kd} and w_{clr} are set to 1.

Inference. Following the recent state-of-the-art methods [11, 58, 62], 10 clips are uniformly sampled for each video along its temporal axis and each clip is taken 3 crops of 256×256. A total of 3x10 crops are evaluated for each video.

4.3. Effectiveness of the Cross-modal Dense Alignment and Contrastive Learning

We begin with a direct comparison to investigate our hypothesis: *multimodal constraints on local and global features can serve as two compelling yet complementary extensions to existing semi-supervised methods* (FixMatch [40] as the baseline here). To this end, our dense alignment (Section 3.3) is devised to regularize the local features, and our contrastive loss (Section 3.4) is developed to distinguish global features.

For a fair comparison, we have ablated four experimental settings (detailed in Table 1): (1) none, (2) alignment-only, (3) contrast-only, and (4) both. Kinetics-400, UCF-101, and HMDB-51 with different labeled data ratios(*i.e.*, 1%, 10%, 20%, and 50%) are used to ensure the generalizability of the following observations. *First*, FixMatch (none) exhibits acceptable but worse performance than its three counterparts, suggesting that pseudo labeling only is inadequate when using very limited labeled data. *Second*, dense alignment significantly elevates the performance (more than contrast-only), indicating that the fine-grained motion signal across multimodal plays an essential role in semi-supervised action recognition. *Third*, introducing contrastive loss across RGB and TG modalities improves Top-1/Top-5 accuracy, revealing that global consistency in different modalities is advantageous. *Finally*, dense alignment and contrastive loss enforce the model learning from complementary perspectives because implementing both on top of FixMatch surpasses either one of them. We hope our discovery on multimodal constraints can shed new light on semi-supervised action recognition in video analysis.

4.4. Comparison with State-of-the-art Methods

To demonstrate the capability and potential of our proposed method, we compared with the most recent state-of-the-art methods for the semi-supervised action recognition task on public datasets including Kinetics-400, UCF-101 and HMDB-51. As shown in Table 2, we mainly compare with two types of methods including image-based methods [28, 44, 60] which were originally designed for image classification and then simply adopted to video tasks and video based methods [23, 39, 58, 62] which were specifically designed for video action recognition task.

Comparison with image-based methods. The first three rows in Table 2 show the results of image-based methods including *i.e.*, Pseudo-Label [28], MeanTeacher [44] and S4L [60]. In general, the results of all the three image-based methods across over the three datasets with all different labeled percentages are much lower than the results of all video-based based methods. This confirms that it is necessary to propose methods specifically designed based on the video temporal and multimodal attributes.

Comparison with video-based methods. The overall performance of the video-based performance are much higher. VideoSSL surpasses all the image-based methods by using Imagenet pre-trained model to guide the learning, and TCL [39] use self-supervised learning task as auxiliary task and the group contrastive for the video semi-supervised learning. Both the ActorCutMix [62] and MvPL [58] are adapted from FixMatch [40]. Benefited by our proposed cross-modal dense alignment and cross-modal contrastive, our method outperforms all these methods by a significant margin on **three datasets under all the experimental settings** (different ratio of labels).

4.5. Ablation Studies

To understand the impact of each part of the design in our method, we conduct extensive ablation studies on the UCF-101 dataset with 20% labeled setting.

Method	w/ ImageNet distillation	Backbone	Kinetics-400 1%	Kinetics-400 10%	UCF-101 5%	UCF-101 10%	UCF-101 20%	UCF-101 50%	HMDB-51 40%	HMDB-51 50%	HMDB-51 60%
Pseudo-Label [28] (ICMLW 2013)	✗	R3D-18	6.3	-	17.6	24.7	37.0	47.5	27.3	32.4	33.5
MeanTeacher [44] (NIPS 2017)	✗	R3D-18	6.8	19.5	17.5	25.6	36.3	45.8	27.2	30.4	32.2
S4L [60] (ICCV 2019)	✗	R3D-18	6.3	-	22.7	29.1	37.7	47.9	29.8	31.0	35.6
UPS [36] (ICLR 2021)	✗	R3D-18	-	-	-	-	39.4	50.2	-	-	-
VideoSSL [23] (WACV 2021)	✓	R3D-18	-	33.8	32.4	42.0	48.7	54.3	32.7	36.2	37.0
TCL [39] (CVPR 2021)	✗	R3D(TSM)-18	7.7	-	-	-	-	-	-	-	-
ActorCutMix [62] (arXiv 2021)	✗	R(2+1)D-34	-	-	27.0	40.2	51.7	59.9	32.9	38.2	38.9
MvPL* [58] (arXiv 2021)	✗	R3D-18	5.0	36.9	41.2	55.5	64.7	65.6	30.5	33.9	35.8
Ours	✗	R3D-18	**9.8**	**43.8**	**44.8**	**62.4**	**76.1**	**79.3**	**46.5**	**48.4**	**49.7**

* indicates the method is reimplemented by ourselves. The input modalities are RGB and TG.

Table 2. **Comparison with the state-of-the-arts methods.** The results are reported with Top-1 accuracy (%) on the validation sets. The best performance of each setting is in **bold**.

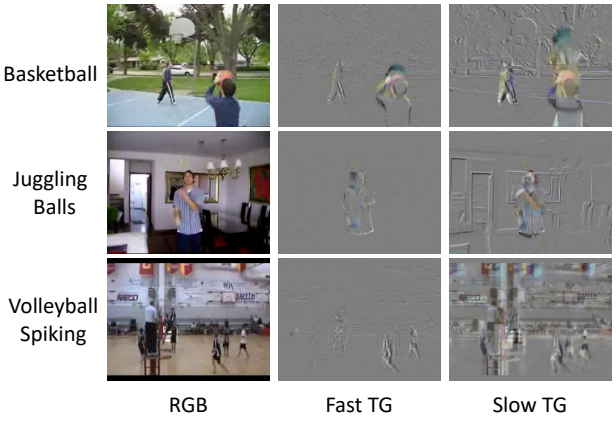

Figure 3. **Visualization of the slow and fast temporal gradient.** Slow temporal gradient contains a more noisy background of the shooting environment while fast temporal gradient focuses more on the activity-related moving objects.

Fast temporal gradient is better. Temporal gradient (TG) is calculated by differing two RGB frames and the stride of them could be small or large to generate either fast or slow TG. To delve deeper into the effect of different strides, we conduct experiments with fast TG (calculation stride = 1) and slow TG (calculation stride = 7), and the results are shown in Table 3a. The first group compare the performance with the baseline FixMatch framework with different modalities of data as input. The results confirm that both the slow TG and fast TG perform much better than RGB (more than 25% higher), and also demonstrate that the Fast TG is better than the slower TG for the semi-supervised setting. The second group of Table 3a compares the final performance of our model with different temporal gradients. When the pseudo-labels are generated by the fast TG, the model beats the performance with slow TG with a large margin (74.1% vs. 68.2%). To figure out the reason why the performance of fast TG is much higher than slower TG, we visualized the two types of temporal gradient for three video clips and the visualization are shown in Figure 3. The comparison shows that the slow TG has much noisy background information especially when the cameras have significant movement while the fast temporal gradient information focuses more on the boundary of the fast moving objects (*e.g.*, people, balls). Both the quantitative and qualitative results verify the advantages of the fast TG over the slower TG for semi-supervised action recognition.

The choice of alignment functions. As discussed in Section 3.3, there are many possible choices for the alignment loss function as long as it can effectively enforce the similarity between the two features. Here we studied the performance of three different alignment functions including L1, L2 and Cosine Similarity loss. As shown in Figure 3 (b), all the three loss functions in alignment achieve high performance while Cosine Similarity (74.6%) outperforms the other two functions (74.0% & 74.4%). A possible explanation is that L1 and L2 have more strict constraints on the scale of two representations, while the Cosine Similarity loss focuses on the vector orientation (*e.g.*, L1 and L2 losses of $\vec{v_1}$=(10,10,10) and $\vec{v_2}$=(1,1,1) are large while the Cosine Similarity loss is 0). Although TG is normalized to the 0-255 range during training, there is still a gap in the scales between the representations of RGB and TG. A strict constraint like L1 or L2 would have negative effects on the model for learning motion features.

Stop gradient in knowledge distillation. The stop-gradient operation on the TG side stated in Section 3.3 is one of the keys to the successful knowledge distillation with dense alignment. However, as the student RGB has much appearance information which TG does not have, directly training with the dense alignment strategy would make the teacher TG model degenerate greatly and hard to focus on extracting the fine-grained motion features. The stop-gradient avoids the fine-grained motion-related representations in TG model can be disturbed by the RGB model. As

Student	Teacher	Top-1
RGB	-	52.9
Slow TG	-	67.3
Fast TG	-	**68.3**
RGB	Slow TG	68.2
RGB	Fast TG	**74.1**

(a) **Fast temporal gradient is better.**

Align. Loss	Top-1	Top-5
-	54.1	78.1
L1	74.0	91.3
L2	74.4	91.4
Cosine	**74.6**	**91.7**

(b) **Dense alignment functions.**

Stopgrad	Top-1	Top-5
✗	60.0	84.4
✓	**74.6**	**91.7**

(c) **Stop Gradient in Knowledge Distillation.**

Pseudo-label Metric	Top-1	Top-5
RGB	73.6	91.0
TG	74.1	91.3
Self	72.8	91.6
Average	**74.6**	**91.7**

(d) **Metrics for the pseudo-labels.**

Aligned Block Index				Accuracy	
1st	2nd	3rd	4th	Top-1	Top-5
✗	✗	✗	✓	71.4	90.2
✗	✗[	✓	✓	74.0	91.4
✗	✓	✓	✓	74.4	**91.8**
✓	✓	✓	✓	**74.6**	91.7

(e) **Align them in block-wise.**

	Top-1	Top-5
Plain	71.1	90.0
+ LR warm-up	71.9	91.1
+ Sup. warm-up	74.1	91.0
+ PreciseBN	**74.6**	**91.7**

(f) **The crucial training tricks.**

Tempature τ	Top-1	Top-5
0.1	74.8	91.8
0.2	75.2	**92.4**
0.5	**76.1**	92.1
1.0	74.3	92.2

(g) **Ablation on contrastive temperature.**

Table 3. **Ablation studies** on UCF101 split-1 under 20% semi-supervised setting (only use 20% labeled data). The results are reported with Top-1 and Top-5 accuracy on the validation set. Backbone: 3D ResNet-18 [19, 49], each input clip consists of 8 frames sampled from a single video with the inter-frame interval of 8. Except for the study (a), all the other results are evaluated with PreciseBN. Except for the study (g), all the other experiments are without cross-modal contrastive learning for better comparisons.

shown in Table 3c, there is a 14.6% performance drop on Top-1 accuracy (60.0% vs. 74.6%) when stop gradient is taken off.

How to generate pseudo-labels? As there are two modalities of data in our model, there are multiple ways to generate pseudo-labels. In Table 3d, we compare the performance of four different ways: 1) use the prediction from RGB model as pseudo-labels, 2) choose the prediction from TG model as pseudo-labels, 3) each model uses the probabilities of its self-modality, and 4) use the fused results from both RGB and TG as pseudo-labels. As being estimated by more comprehensive information from both RGB and TG, the fused pseudo-labels are more reliable and achieve the best performance.

Dense alignment in block-wise. An intuitive question about our knowledge distillation framework is that which block or blocks should be densely aligned. Therefore, we conduct this ablation study by adding dense alignment to different positions (*i.e.*, blocks) and the results are shown in Table 3e. As the common practice of previous knowledge distillation methods [20, 43, 45] is to align the high-level features of the last layers. Therefore, we start to add the dense alignment module over the features from the last (4-th) block (ResNet basic block) and then experiment with more blocks. Their performances are consistently improved when more blocks are densely aligned and the best Top-1 accuracy is achieved with all blocks aligned. Compared with the baseline, our block-wise dense alignment strategy gains a considerable improvement of 20.5% (54.1% to 74.6%) which demonstrates that fine-grained motion signals are better at semi-supervised model generalization.

The crucial training tricks. Through the extensive experiments, we identified several training tricks which are essential to lead to the high performance. Table 3f shows the impact of learning rate warm-up [14], supervised warm-up [58] and PreciseBN [56]. All the three tricks could make a decent improvement, while the supervised warm-up (training with only the labeled data at the first several epochs) is the most effective one which gains an improvement of 2.7% (71.9% to 74.1%). This shows that the supervised warm-up can alleviate the cold-start issue that low-quality pseudo-labels would be generated at the beginning. The performance of the semi-supervised learning model could easily have large variations [32, 39, 51]. These three tricks can solidly improve the performance while in the meantime make the training more stable.

Contrastive temperature. An appropriate temperature is important to the good performance of contrastive learning [7], we ablate the contrastive loss temperature in Equation (11). As shown in Table 3g, a modest temperature (e.g., 0.2 or 0.5) could help the proposed cross-modal contrastive learning work better while a large (1.0) or small (0.1) temperature is not that optimal.

5. Conclusion

This paper has presented a novel semi-supervised learning method which introduces temporal gradient for the fruitful motion-related information and extra representation consistency crossing multiple modalities. Our proposed method uses the block-wise dense alignment strategy and cross-modal contrastive learning, without additional computation or delay during inference. Our method substantially outperforms all prior methods while achieving state-of-the-art performance on UCF-101, HMDB-51,

and Kinetics-400 datasets with all the experimented settings (different labeled ratios). In the future, we plan to study the effectiveness of temporal gradient on other video-based tasks and to explore more powerful modalities by automatically searching or generating.

References

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019. 3

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021. 1

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 3

[5] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 1

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3, 5, 8

[8] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 12

[9] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 3, 5, 12

[10] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition, 2020. 1

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1, 5, 6

[12] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021. 3

[13] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016. 3

[14] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 5, 8

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Pires, Zhaohan Guo, Mohammad Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Neural Information Processing Systems*, 2020. 3

[16] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, 2000. 12

[17] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020. 1, 3

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 8, 12

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 8

[21] Lianghua Huang, Yu Liu, Bin Wang, Pan Pan, Yinghui Xu, and Rong Jin. Self-supervised video representation learning by context and motion decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13886–13895, 2021. 3

[22] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 12

[23] Longlong Jing, Toufiq Parag, Zhe Wu, Yingli Tian, and Hongcheng Wang. Videossl: Semi-supervised learning for video classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1110–1119, 2021. 1, 2, 5, 6, 7

[24] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. imgaug. https://github.com/aleju/imgaug, 2020. Online; accessed 01-Feb-2020. 12

[25] Hirokatsu Kataoka, Tenga Wakamiya, Kensho Hara, and Yutaka Satoh. Would mega-scale datasets further enhance spatiotemporal 3d cnns? *arXiv preprint arXiv:2004.04968*, 2020. 1

[26] 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. 5, 12

[27] 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*, pages 2556–2563. IEEE, 2011. 5, 12

[28] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 2, 6, 7

[29] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 5

[30] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 3

[31] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML)*, pages 807–814, 2010. 12

[32] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in Neural Information Processing Systems*, 31:3235–3246, 2018. 8

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[34] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11205–11214, 2021. 1, 3

[35] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 1, 3

[36] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2020. 2, 7

[37] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 12

[38] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014. 3

[39] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10399, 2021. 1, 2, 6, 7, 8

[40] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 2, 4, 5, 6

[41] 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. 5, 12

[42] 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*, 15(1):1929–1958, 2014. 5

[43] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3d: Distilled 3d networks for video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 625–634, 2020. 8

[44] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1195–1204, 2017. 2, 6, 7

[45] Fida Mohammad Thoker and Juergen Gall. Cross-modal knowledge distillation for action recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 6–10. IEEE, 2019. 8

[46] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 3, 5

[47] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 1

[48] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri. Convnet architecture search for spatiotemporal feature learning. *arXiv preprint arXiv:1708.05038*, 2017. 1

[49] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 1, 5, 8, 12

[50] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 12

[51] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14615–14624, 2021. 2, 8

[52] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015. 3

[53] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 2

[54] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 3

[55] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 1

[56] Yuxin Wu and Justin Johnson. Rethinking" batch" in batchnorm. *arXiv preprint arXiv:2105.07576*, 2021. 8

[57] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[58] Bo Xiong, Haoqi Fan, Kristen Grauman, and Christoph Feichtenhofer. Multiview pseudo-labeling for semi-supervised learning from video. *arXiv preprint arXiv:2104.00682*, 2021. 1, 2, 5, 6, 7, 8, 12

[59] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 591–600, 2020. 1

[60] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1476–1485, 2019. 2, 6, 7

[61] Yue Zhao, Yuanjun Xiong, and Dahua Lin. Recognize actions by disentangling components of dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6566–6575, 2018. 2

[62] Yuliang Zou, Jinwoo Choi, Qitong Wang, and Jia-Bin Huang. Learning representational invariances for data-efficient action recognition. *arXiv preprint arXiv:2103.16565*, 2021. 1, 2, 5, 6, 7

[63] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. In *International Conference on Learning Representations*, 2020. 2

Appendix

This appendix covers the implementation details (§A), robustness evaluation with multiple types of corruptions (§B), the visualization of attention maps with Grad-CAM (§C) and t-SNE feature visualization (§D).

A. Additional Implementation Details

Network architecture. The details of the 3D ResNet-18 [19, 49] backbone architecture is illustrated in Table 4. This backbone is adopted as the feature extractor for both RGB and TG modalities. There are two heads following each backbone, one is for the general classification prediction with Softmax activation (Global Average Pooling + Dropout + FC) and the other is for the projection in the contrastive learning framework (3-layer non-linear MLP with BatchNorm [22] and ReLU [16, 31]).

Layer Name	Output Size	R3D-18
conv1	L ×56 ×56	$3 \times 7 \times 7, 64$, stride $1 \times 2 \times 2$
conv2_x	L ×56 ×56	$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 2$
conv3_x	$\frac{L}{2}$ ×28 ×28	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 2$
conv4_x	$\frac{L}{4}$ ×14 ×14	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 2$
conv5_x	$\frac{L}{8}$ ×7 ×7	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 2$

Table 4. **Backbone architecture.** Residual blocks are shown in brackets.

Video Augmentations. We implement our method with MMAction2 [8]. For weak augmentation, we use the *Resize*, *RandomResizedCrop*, and *Flip* in MMAction2. For strong augmentation, we use the RandAugment [9] implemented with imgaug [24] .

Temporal Gradient Normalization. Following Xiong *et al*. [58], we normalize the calculated temporal gradient for fitting the common 0-255 range by adding 255 and dividing by 2.

B. Robustness Against Input Corruptions

To verify the hypothesis that our method learns more motion-related features from the temporal gradient and is more robust to contrast and brightness variations, we evaluate the models with different corruptions (*i.e*., random contrast adjustment noise, random brightness adjustment noise and conversion to grayscale) during the testing stage. As shown in Table 5, our method is more robust than the baseline to all types of corruptions. It is worth noting that in the gray-scale corruption case (the inputs lose all color information), the performance of baseline drops 28.0% (51.8% relative) while ours only drops 14.6% (19.2% relative).

Corruptions	Baseline	Ours
No Corruption	54.1	76.1
Contrast Noise	53.1 (-1.0)	75.3 (-0.8)
Brightness Noise	52.2 (-1.9)	75.2 (-0.9)
Grayscale	26.1 (-28.0)	61.5 (-14.6)

Table 5. **Robustness evaluation with different corruptions.** The Contrast Noise and Brightness Noise are implemented with the *EnhanceContrast* and *EnhanceBrightness* of imgaug [24]. All results are reported in Top-1 accuracy. The models are trained with 20% labels (UCF101-20%).

C. Grad-CAM Attention Maps

To better demonstrate that our method focuses more on the motion-related information, we visualize the attention maps with Grad-CAM [37] of multiple videos of UCF-101 [41] validation set. As shown in Figure 4, our model's attention is more reasonable and focuses more on the acting humans and moving objects.

D. t-SNE Feature Visualization

We also visualize the high-level features with t-SNE [50] for showing a better latent representation space with our method. The visualization results covering the extracted features of the whole UCF-101 [41] validation set are shown in Figure 5. The features extracted with our method are more separable and easier to be classified in the latent representation space.

License of used assets: Kinetics-400 [26]: Creative Commons Attribution 4.0 International License; HMDB-51 [27]: Creative Commons Attribution 4.0 International License; UCF-101 [41]: https://www.crcv.ucf.edu/data/UCF101.php.

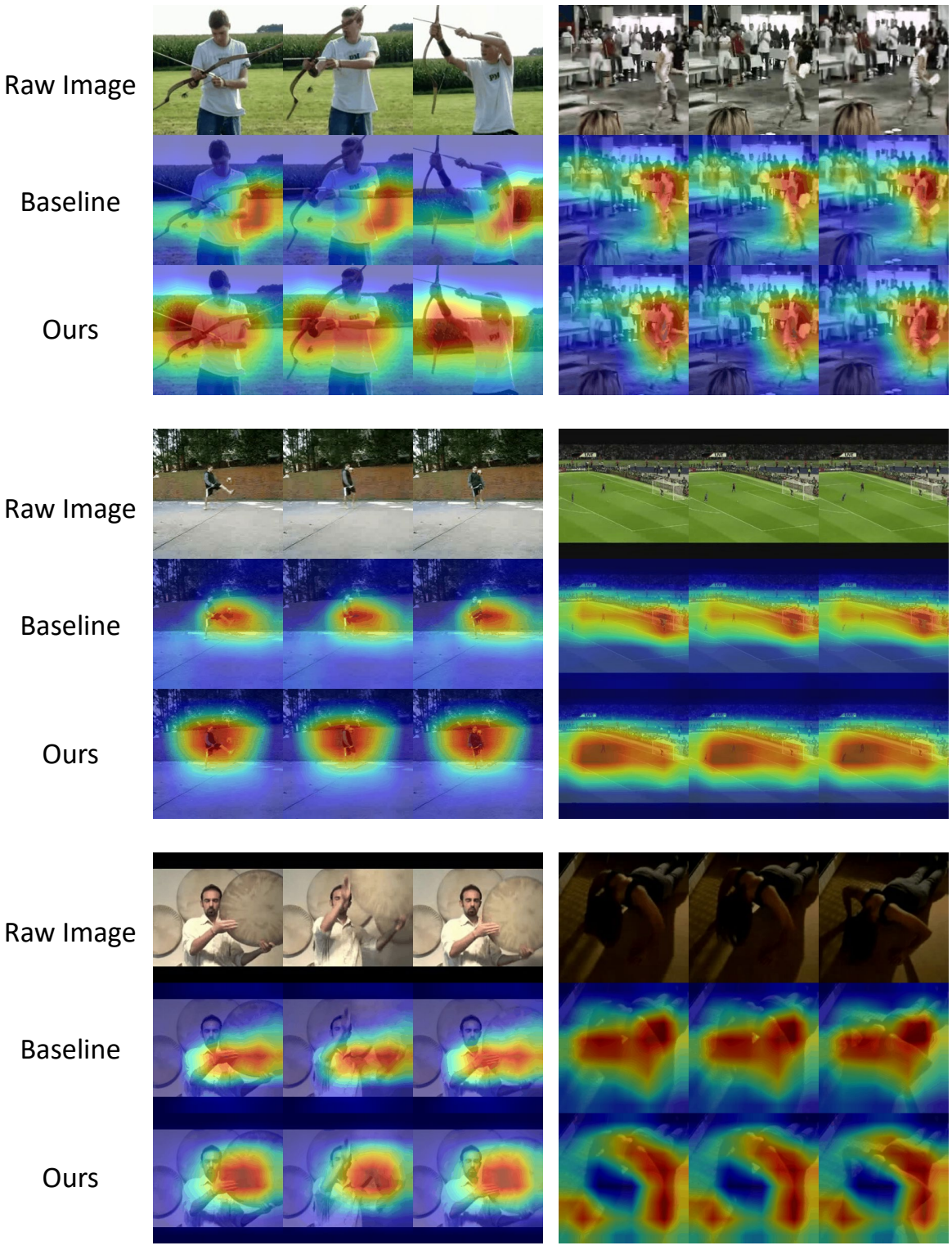

Figure 4. **Grad-CAM visualization of the attention maps.** The videos are sampled from the validation set of UCF-101. The models are trained with 20% labels (UCF101-20%).

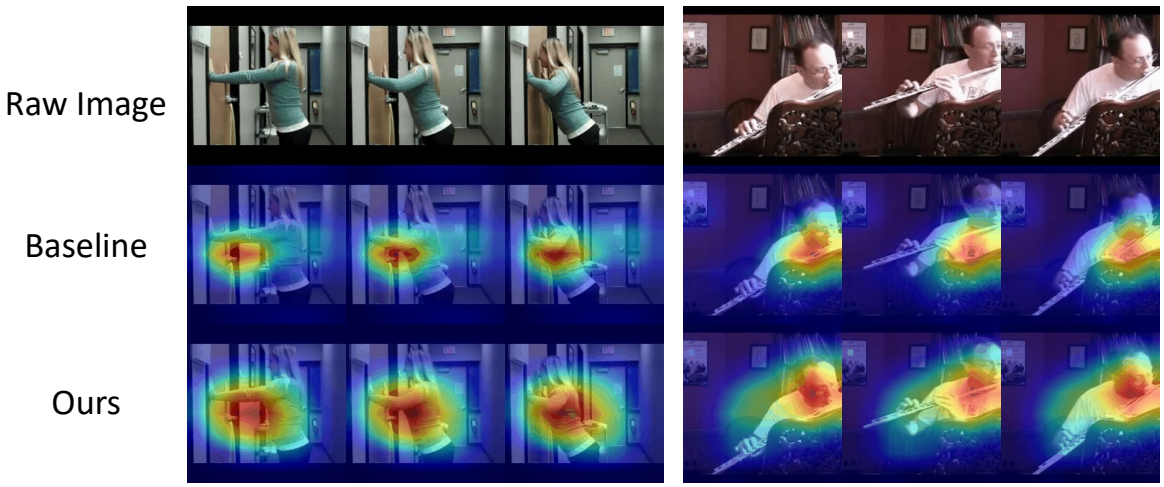

Figure 4. **Visualization of the Grad-CAM attention maps.** The videos are sampled from the validation set of UCF-101. The models are trained with 20% labels (UCF101-20%).

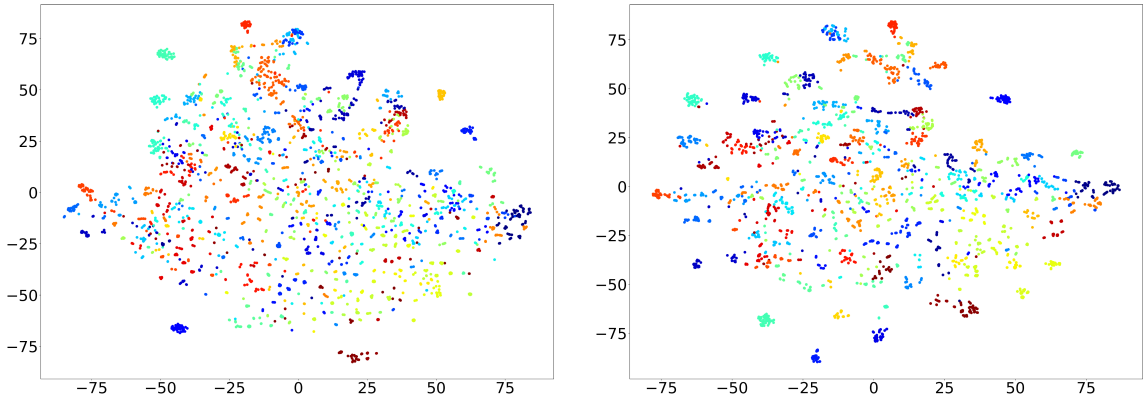

Figure 5. **The comparison of t-SNE visualizations of the baseline and our method.** The visualized features are globally averaged features extracted by the backbone. All the videos of the validation set of UCF-101 are evaluated. The models are trained with 20% labels (UCF101-20%).