Learning from Temporal Gradient for Semi-supervised Action Recognition

Junfei Xiao\textsuperscript{1} | Longlong Jing\textsuperscript{2} | Lin Zhang\textsuperscript{3} | Ju He\textsuperscript{1} | Qi She\textsuperscript{4}
Zongwei Zhou\textsuperscript{1} | Alan Yuille\textsuperscript{1} | Yingwei Li\textsuperscript{1}

\textsuperscript{1}Johns Hopkins University | \textsuperscript{2}The City University of New York | \textsuperscript{3}Carnegie Mellon University | \textsuperscript{4}ByteDance

\textbf{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). Code is made available at https://github.com/lambert-x/video-semisup.

1. Introduction

As a fundamental task for video understanding, video action recognition has drawn much attention from the community and industry [5, 9, 45, 47]. Unlike image-related tasks, networks for video-related tasks are normally easier to overfit due to the complexity of the tasks [23, 45, 46]. The common practice is to firstly pre-train the network on large-scale datasets (e.g., Kinetics [4] of up to 650,000 video clips) and then finetune on downstream small datasets to obtain better performance [9, 15, 32, 33].

However, since annotating large-scale video datasets is time-consuming and expensive, training models on a large dataset collected with complete annotations is impeded. To utilize large-scale datasets with acceptable costs, some researchers have turned to designing semi-supervised learning models which have good generalization ability with limited annotations [21, 37, 55, 58]. As pseudo-label based methods (e.g., FixMatch [38] and MixMatch [2]) 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 [21, 58] 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 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 expects the features extracted from the same video change smoothly. The multimodal consistency, the potential of semi-supervised action recognition is not fully unleashed. Without special designs to specifically focus on temporal information and multimodal consistency, the potential of semi-supervised action recognition is not fully unleashed.

Some previous studies [51, 57] introduce temporal gradient as an additional modality to better utilize the temporal information encoded in videos as it is rich in motion signals. The temporal gradient can be formulated as: \( TG = x_t^{RGB} - x_{t+n}^{RGB} \), 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 the UCF-101 dataset with only 20\% of labeled data for training.

Why is the temporal gradient so much better than the RGB frames when the training data are 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 the 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 [38], 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 leverage the detailed and fine-grained motion signals from temporal gradient, we propose a knowledge distillation strategy using block-wise dense alignment. It helps the student RGB model learn from the teacher temporal gradient model efficiently and effectively. To further improve high-level representation space across different modalities, we perform 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.

Unlike the existing methods, our model has two unique advantages. First, our model requires no additional computation or parameters for inference. In the training, we distill the knowledge from temporal gradient to the RGB-based network; in the testing, only the RGB model is required. Second, our model is 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 [26], S4L [56], MeanTeacher [42], MixMatch [2], UDA [54], FixMatch [38], UPS [34], etc. The Pseudo-Label [26] 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 [34, 38]. Following a state-of-the-art method on image classification—FixMatch [38], many FixMatch-alike methods achieve the state-of-the-art performance on many other tasks including detection [49], segmentation [59], 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 [21, 37].

Semi-supervised learning in videos. Although there have been a few semi-supervised video action recognition methods [21, 37, 55, 58] proposed, most of them directly apply the image-based methods to videos with less focus on the temporal dynamics of videos. VideoSSL [21] 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 (e.g. Pseudo-Label [26], Mean-Teacher [42]) have inferior performance on video semi-supervised benchmarks. TCL [37] is a recently proposed method that jointly optimizes the network by employing a self-supervised auxiliary task and a group contrastive learning. By using multimodal data, MvPL [55] achieved the state-of-the-art performance by sharing the same model with different input modalities (RGB, temporal gradient, and optical flow) and generating pseudo labels with the “confidence” of multiple modalities. Compared with these methods, our method specifically focuses on learning the
temporal information from Temporal Gradient 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 [11, 36] 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, 11, 36, 50, 52]. However, there are 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 have achieved remarkable performance on downstream image classification [3, 6, 13, 16, 28, 44]. The key idea is that the representation can be learned 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 [10, 15, 19, 32, 33]. 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 $x_i \in \mathbb{R}^{T \times H \times W \times 3}$ 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 $f_{\theta}$ to denote a classification model with trainable parameters $\theta$. 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 $A(\cdot)$ to represent the strong data augmentation strategies (i.e., Randaugment [8]).

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

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

where $B_l$ is the number of labeled samples in a 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 with different extent of augmentations. Specifically, pseudo labels $\hat{y}$ for the unlabeled data are usually generated via confidence thresholding as:

$$\mathcal{C} = \{x_j | \max_{\alpha(\cdot)} f_{\theta}(\alpha(x_j)) \geq \gamma\}, \quad (2)$$

where $\gamma$ 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 into one-hot labels $\hat{y}_j$ by taking the 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:

$$\mathcal{L}_u = -\frac{1}{B_u} \sum_{x_j \in \mathcal{C}} \hat{y}_j \log f_{\theta}(A(x_j)), \quad (3)$$

where $B_u$ is the number of unlabeled samples in a batch.

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

$$\mathcal{L}_{fm} = \mathcal{L}_l + \mathcal{L}_u. \quad (4)$$

#### 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 [38] 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:

$$
C = \{x_j | \max(f_{\theta_R}(\alpha(x_j^{\text{RGB}})) + f_{\theta_T}(\alpha(x_j^{\text{TG}})) \geq \gamma}\}. \quad (5)
$$

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 (1) and Equation (3) but for TG.

$$
L_{Tm}^{TG} = L_t^{TG} + \lambda_u L_u^{TG}. \quad (6)
$$

### 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-modal dense feature alignment module as:

$$
\min \left[ D \left( \mathcal{F}_i^{\text{RGB}}, \mathcal{F}_i^{\text{TG}} \right) \right], \quad (7)
$$

where $\mathcal{F}_i^{\text{RGB}}$, $\mathcal{F}_i^{\text{TG}} \in \mathbb{R}^{C \times T \times H \times W}$ denote the output features of the $i$-th block in the RGB and TG models, and $D$ represents a pairwise function evaluating the representation differences. There are many choices for $D$ and we experiment with three different functions: L1, L2 and Cosine Similarity losses (shown in Equation (8), where $\| \cdot \|_1$ and $\| \cdot \|_2$ are $l_1/l_2$-norm). A more detailed discussion is provided in Section 4.5.

$$
D_{L1}(F_1, F_2) = \|F_1 - F_2\|_1,
$$

$$
D_{L2}(F_1, F_2) = \|F_1 - F_2\|_2,
$$

$$
D_{\text{cos}}(F_1, F_2) = -\frac{F_1 \cdot F_2}{\|F_1\|_2 \cdot \|F_2\|_2}. \quad (8)
$$

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 (9), the alignment loss term for learning fine-grained
motion features is:
\[ \mathcal{L}_{kd} = \left[ D \left( x_i^{RGB}, \text{stopgrad}(x_i^{TG}) \right) \right]. \]  

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 [6] and CMC [44], 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 [31] as the objective function over the features extracted from RGB and TG:
\[ \mathcal{L}_{clr} = -\log \frac{\sum_{k \in \{k^+\}} \exp(\text{sim}(q, k)/\tau)}{\sum_{k \in \{k^+, k^-\}} \exp(\text{sim}(q, k)/\tau)}, \]
with \( \tau \) being a temperature hyper-parameter for scaling. All embeddings are \( \ell_2 \) 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 (\( x_i^{RGB/TG} \)) 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 preprocessing.

**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:
\[ \mathcal{L}_{total} = w_{fm}(\mathcal{L}^{RGB}_{fm} + \mathcal{L}^{TG}_{fm}) + w_{kd}\mathcal{L}_{kd} + w_{clr}\mathcal{L}_{clr}. \]

4. Experimental Results

4.1. Datasets and Evaluation

**Datasets.** Following previous state-of-the-art semi-supervised video action recognition methods [21, 55, 58], we evaluate our method on three public action recognition benchmarks: UCF-101 [39], HMDB-51 [25], and Kinetics-400 [24]. 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 [21]. 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 [55] 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 and Top-5 accuracy for some ablation studies.

4.2. Implementation Details

**Network architecture.** For a fair comparison with the state-of-the-art methods [21, 55], the FixMatch [38] framework is used as the backbone model while the 3D ResNet-18 [17, 47] 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 [58]. 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 [8] is chosen 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 performs 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 rate of 0.2 on 8 GPUs by following the settings in [9, 55, 58] using the cosine learning rate decaying scheduler [27] and also a linear warm-up strategy [12]. We use momentum of 0.9 and weight decay of 10^{-4}. Dropout [40] of 0.5 is used before the final classifier layer to reduce the over-fitting. Following [58], each mini-batch consists of 5 labeled data clips and 5 unlabeled data clips, while each 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 [55] 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 [9, 55, 58], 10 clips are uniformly sampled for each video along its temporal axis and each clip is taken 3 crops of $256 \times 256$. 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 examine our hypothesis: *multimodal constraints on local and global features can serve as two complementary extensions to existing semi-supervised methods* (FixMatch [38] as the baseline). 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 that our discovery on multimodal constraints can shed new light on semi-supervised action recognition in video analysis.

**Overfitting is alleviated.** Table 6 (Suppl.) presents a significant accuracy gap between the training and testing set, showing that FixMatch severely overfits to the training set. Our method effectively reduces the gap by imposing additional regularization on models with RGB as input.

### 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 [26, 42, 56] which were originally designed for image classification and then simply adopted to video tasks and video based methods [21, 37, 55, 58] 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 [26], MeanTeacher [42] and S4L [56]. In general, the results of all the three image-based methods across 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 [37] use self-supervised learning task as auxiliary task and the group contrastive for the video semi-supervised learning. Both the ActorCutMix [58] and MvPL [55] are adapted from FixMatch [38]. 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.

**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
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>w/ ImageNet distillation</th>
<th>Backbone</th>
<th>Kinetics-400 1%</th>
<th>Kinetics-400 10%</th>
<th>UCF-101 5%</th>
<th>UCF-101 10%</th>
<th>UCF-101 20%</th>
<th>UCF-101 50%</th>
<th>HMDB-51 40%</th>
<th>HMDB-51 50%</th>
<th>HMDB-51 60%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pseudo-Label [26] (ICMLW 2013)</td>
<td>✗</td>
<td>R3D-18</td>
<td>6.3</td>
<td>17.6</td>
<td>24.7</td>
<td>37.0</td>
<td>47.5</td>
<td>27.3</td>
<td>32.4</td>
<td>33.5</td>
<td></td>
</tr>
<tr>
<td>MeanTeacher [32] (NIPS 2017)</td>
<td>✗</td>
<td>R3D-18</td>
<td>6.8</td>
<td>19.5</td>
<td>25.6</td>
<td>36.3</td>
<td>45.8</td>
<td>27.2</td>
<td>30.4</td>
<td>32.2</td>
<td></td>
</tr>
<tr>
<td>S4L [56] (ICCV 2019)</td>
<td>✗</td>
<td>R3D-18</td>
<td>6.3</td>
<td>-</td>
<td>22.7</td>
<td>29.1</td>
<td>37.7</td>
<td>29.8</td>
<td>31.0</td>
<td>35.6</td>
<td></td>
</tr>
<tr>
<td>UPS [34] (ICLR 2021)</td>
<td>✗</td>
<td>R3D-18</td>
<td>-</td>
<td></td>
<td>-</td>
<td>39.4</td>
<td>50.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>VideoSSL [21] (WACV 2021)</td>
<td>✓</td>
<td>R3D-18</td>
<td>-</td>
<td>33.8</td>
<td>32.4</td>
<td>42.0</td>
<td>48.7</td>
<td>54.3</td>
<td>32.7</td>
<td>36.2</td>
<td>37.0</td>
</tr>
<tr>
<td>TCL [37] (CVPR 2021)</td>
<td>✗</td>
<td>R3D(TSM)-18</td>
<td>7.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ActorCutMix [58] (arXiv 2021)</td>
<td>✗</td>
<td>R(2+1)D-34</td>
<td>-</td>
<td>-</td>
<td>27.0</td>
<td>40.2</td>
<td>51.7</td>
<td>59.9</td>
<td>32.9</td>
<td>38.2</td>
<td>38.9</td>
</tr>
<tr>
<td>MvPL* [55] (arXiv 2021)</td>
<td>✗</td>
<td>R3D-18</td>
<td>5.0</td>
<td>36.9</td>
<td>41.2</td>
<td>55.5</td>
<td>64.7</td>
<td>65.6</td>
<td>30.5</td>
<td>33.9</td>
<td>35.8</td>
</tr>
<tr>
<td>Ours</td>
<td>✗</td>
<td>R3D-18</td>
<td>9.8</td>
<td>43.8</td>
<td>44.8</td>
<td>62.4</td>
<td>76.1</td>
<td>79.3</td>
<td>46.5</td>
<td>48.4</td>
<td>49.7</td>
</tr>
</tbody>
</table>

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

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.

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 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? There are multiple ways to generate pseudo-labels since our model takes two input
modalities. We compare the performance of four settings: 1) use the prediction from RGB model as pseudo-labels, 2) use the prediction from TG model as pseudo-labels, 3) each model uses the probabilities of its self-modality, and 4) fuse the results from both RGB and TG as pseudo-labels. Table 3d shows that the fused pseudo-labels are more reliable and achieve the best performance benefiting from comprehensive information of both RGB and TB.

**Dense alignment in block-wise.** An intuitive question about our knowledge distillation framework is 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 [18, 41, 43] 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.

**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 [12], supervised warm-up [55] and PreciseBN [53]. 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 74.6%). 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 [30, 37, 49]. 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 [6], we ablate the contrastive loss temperature in Equation (10). 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 automatically search or generate powerful modalities.

**Acknowledgments:** This work was supported by the National Science Foundation under Grant No. NSF-1763705.


This document contains the supplementary materials for “Learning from Temporal Gradient for Semi-supervised Action Recognition”. It covers the implementation details (§A), robustness evaluation with multiple types of corruptions (§B), the visualization of attention maps with Grad-CAM (§C), t-SNE feature visualization (§D) and an analysis of the overfitting issue (§E).

A. Additional Implementation Details

Network architecture. The details of the 3D ResNet-18 [17, 47] backbone architecture are 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 [20] and ReLU [14, 29]).

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

<table>
<thead>
<tr>
<th>Layer Name</th>
<th>Output Size</th>
<th>R3D-18</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv1</td>
<td>L x 56 x 56</td>
<td>3 x 7 x 7, 64, stride 1 x 1 x 2</td>
</tr>
<tr>
<td>conv2, x</td>
<td>L x 56 x 56</td>
<td>[3 x 3 x 3, 64] x 2</td>
</tr>
<tr>
<td>conv3, x</td>
<td>L x 28 x 28</td>
<td>[3 x 3 x 3, 128] x 2</td>
</tr>
<tr>
<td>conv4, x</td>
<td>L x 14 x 14</td>
<td>[3 x 3 x 3, 256] x 2</td>
</tr>
<tr>
<td>conv5, x</td>
<td>L x 7 x 7</td>
<td>[3 x 3 x 3, 512] x 2</td>
</tr>
</tbody>
</table>

Table 5. Robustness evaluation with different corruptions. The Contrast Noise and Brightness Noise are implemented with the EnhanceContrast and EnhanceBrightness of imgaug [22]. All results are reported in Top-1 accuracy. The models are trained with 20% labels (UCF101-20%).

<table>
<thead>
<tr>
<th>Corruptions</th>
<th>Baseline</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Corruption</td>
<td>54.1</td>
<td>76.1</td>
</tr>
<tr>
<td>Contrast Noise</td>
<td>53.1 (-1.0)</td>
<td>75.3 (-0.8)</td>
</tr>
<tr>
<td>Brightness Noise</td>
<td>52.2 (-1.9)</td>
<td>75.2 (-0.9)</td>
</tr>
<tr>
<td>Grayscale</td>
<td>26.1 (-28.0)</td>
<td>61.5 (-14.6)</td>
</tr>
</tbody>
</table>

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 [35] of multiple videos of UCF-101 [39] validation set. As shown in Figure 4, the attention of our model 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 [48] for showing a better latent representation space with our method. The visualization results covering the extracted features of the whole UCF-101 [39] 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.

E. Overfitting is Alleviated

Table 6 presents a significant accuracy gap between the training and testing set, showing that FixMatch severely overfits to the training set. Our method effectively reduces the gap by imposing additional regularization on models with RGB as input.

Table 6. Top-1 accuracy of the final models. The models are trained on UCF-101 with 20% labels.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-RGB</td>
<td>98.5</td>
<td>54.1</td>
<td>44.4</td>
</tr>
<tr>
<td>Ours-RGB (Student)</td>
<td>98.0</td>
<td>76.1</td>
<td>22.0</td>
</tr>
<tr>
<td>Baseline-TG</td>
<td>97.4</td>
<td>68.6</td>
<td>28.8</td>
</tr>
<tr>
<td>Ours-TG (Teacher)</td>
<td>96.5</td>
<td>75.3</td>
<td>21.2</td>
</tr>
</tbody>
</table>
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%).

**F. License of Used Assets**

Kinetics-400 [24]: Creative Commons Attribution 4.0 International License; HMDB-51 [25]: Creative Commons Attribution 4.0 International License; UCF-101 [39]: https://wwwcrcv.ucf.edu/data/UCF101.php.