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]).

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 in brackets.

Video Augmentations. We implement our method using MMAction2[2] [7]. For weak augmentation, we use the *Resize*, *RandomResizedCrop*, and *Flip* in MMAction2. For strong augmentation, we use the RandAugment [8] implemented with imgaug [22] .

Temporal Gradient Normalization. Following Xiong *et al*. [55], we normalize the temporal gradient to fit 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 [22]. 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 [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.

	Training Acc.	Testing Acc.	Acc. Gap
Baseline-RGB	98.5	54.1	44.4
Ours-RGB (Student)	98.0	76.1	**22.0**
Baseline-TG	97.4	68.6	28.8
Ours-TG (Teacher)	96.5	75.3	**21.2**

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

[2]MMAction2: https://github.com/open-mmlab/mmaction2

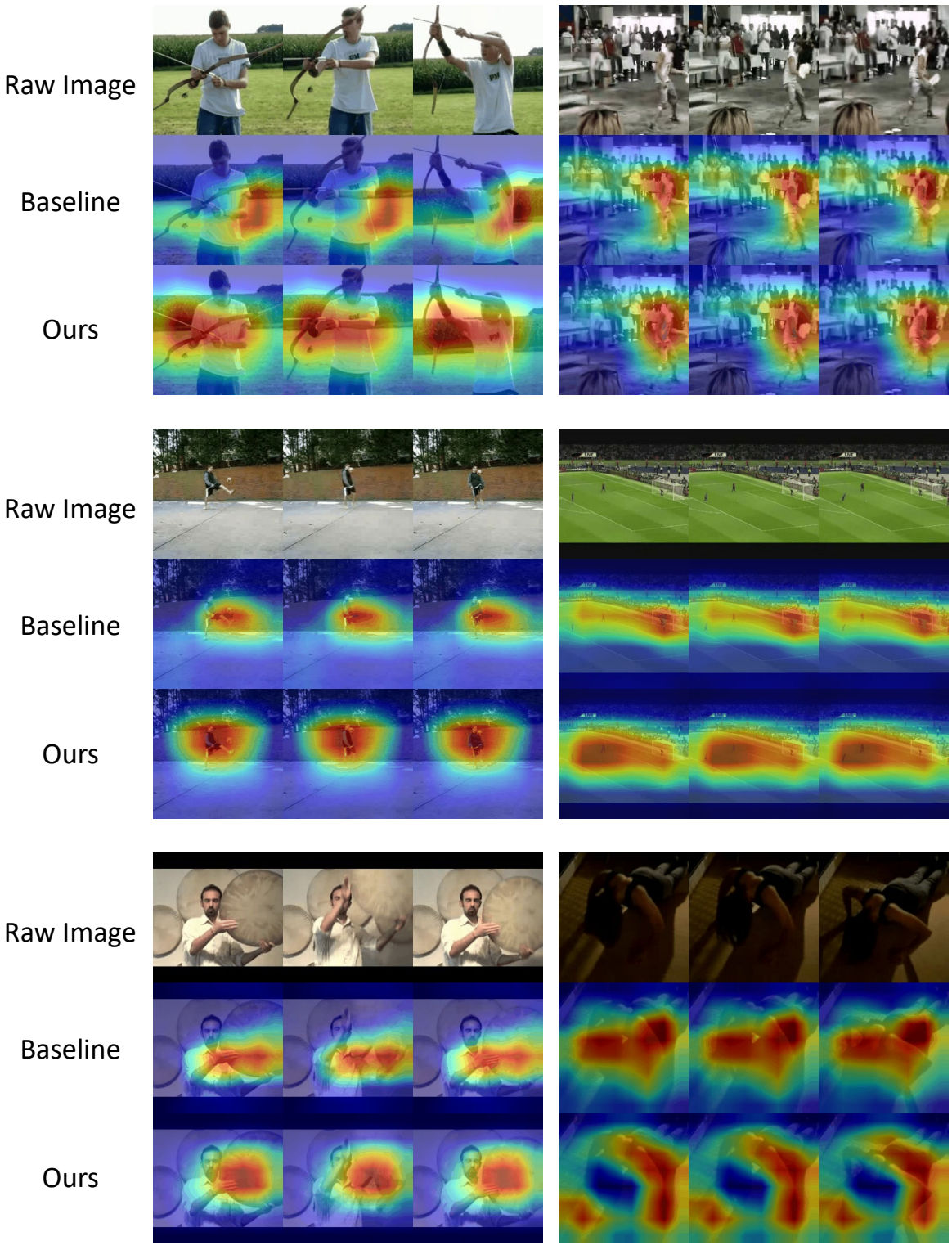

Figure 4. **Grad-CAM visualization of the attention maps.** The videos are sampled from the validation set of UCF-101. The models are trained with 20% labels (UCF101-20%).

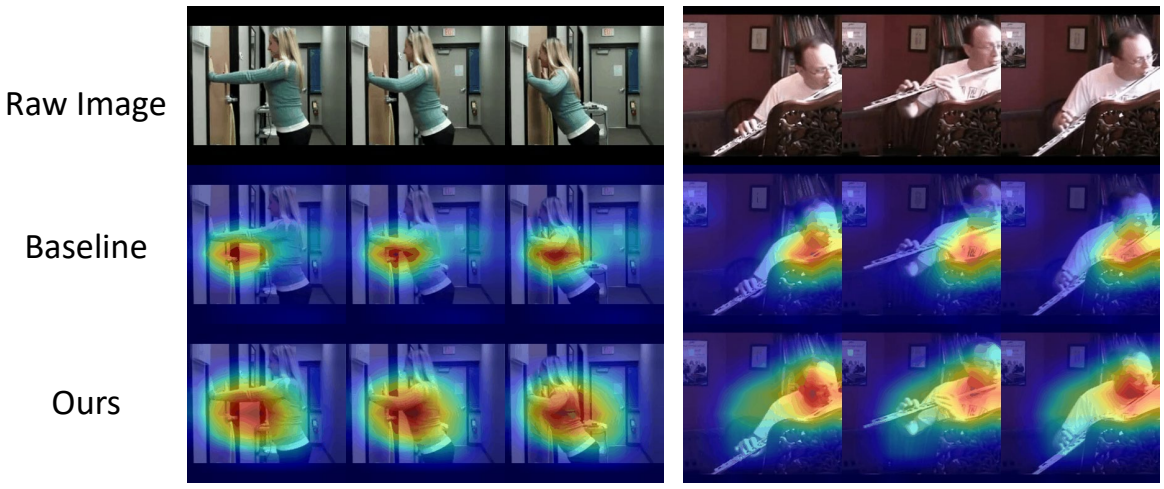

Figure 4. **Visualization of the Grad-CAM attention maps.** The videos are sampled from the validation set of UCF-101. The models are trained with 20% labels (UCF101-20%).

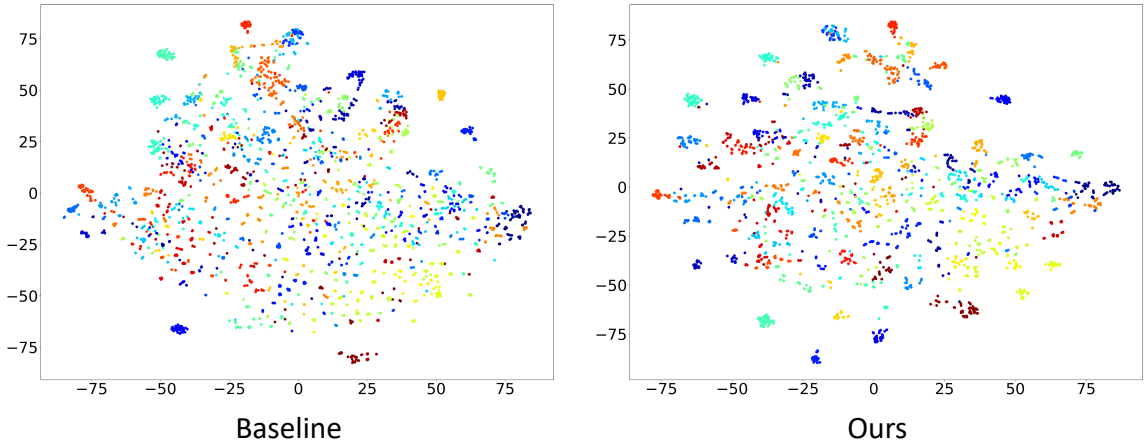

Figure 5. **The comparison of t-SNE visualizations of the baseline and our method.** The visualized features are globally averaged features extracted by the backbone. All the videos of the validation set of UCF-101 are evaluated. The models are trained with 20% labels (UCF101-20%).

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://www.crcv.ucf.edu/data/UCF101.php.