Learning an Augmented RGB Representation with Cross-Modal Knowledge Distillation for Action Detection Supplementary Material

Rui Dai[1,2], Srijan Das[3], François Bremond[1,2]
[1]Inria [2]Université Côte d'Azur [3]Stony Brook University
[1,2]{name.surname}@inria.fr [3]{name.surname}@stonybrook.edu

Overview
In this supplementary material, we provide more details regarding the model structure and experimental analysis to complement the descriptions from the main paper. In Section 1, we describe the network architecture in more details. In Section 2, we provide more qualitative results, e.g. T-SNE, boundary saliency figures. In Section 3, we analyze the sensitivity of the hyper-parameters involved in our proposed framework. In Section 4, we adapt and compare with additional Knowledge Distillation (KD) methods from other domains (i.e. model compression). Finally, in Section 5, we provide a complete state-of-the-art results on the evaluated datasets. For convenience, we use the same notation as in the main paper for this supplementary material.

1. Network Structure

In Section 3.1 of the main paper, we provide an overall description of the Seq2Seq KD network structure. Here, we describe the components of the network in details.

Overall structure: As shown in Fig 1 (1), we provide an overview of our pipeline for a single input stream. Firstly, every 16 frames of a snippet are encoded as a single feature vector representation by a visual encoder. The visual encoder varies depending on the input modality. The visual encoder is either I3D [1] for RGB and Optical Flow (OF), or AGCN [14] for 3D Poses. The channel size C_1 of the snippet representation is determined by the encoder: 1024 for I3D and 256 for AGCN. While stacking the features of all the snippets along time, we obtain the feature representation of a video. The encoded feature map of the video is fed to the temporal filter to model the temporal information. The output feature of the temporal filter is leveraged for both knowledge distillation and classification. The classifier predicts the logits for each snippet, followed by upsampling the logits to the same temporal resolution as the ground truth.

Temporal filter: Fig 1 (2) illustrates the structure of our default temporal filter: 5-layer SS-TCN [6]. SS-TCN is composed of a bottleneck layer followed by 5 dilated-layers. The bottleneck squeezes the channel size of the incoming feature map. In this work, we set the squeezed channel size C_2 to 256. The structure of the dilated-layer is provided in Fig 1 (3). Each layer is again composed of a dilated temporal convolutional layer, a ReLu activation, a bottleneck and a residual link. The dilation in the kernel increases the temporal reception field, in order to model longer temporal relations. Similar to [6], we set the dilation to 2^{i-1} for the i^{th} layer in our experiments. Note that the SS-TCN temporal filter can be replaced by its other variants, such as PDAN [3] and TGM [13]. The output features of the temporal filter are further used for knowledge distillation and classification.

Classifier: Similar to [13, 6], our classifier is a bottleneck layer with an activation (i.e. Sigmoid or Softmax) and trained with the cross entropy loss. C_3 represents the number of action classes in the dataset. For the densely-labeled datasets, following [13, 19], the classifier with sigmoid activation can be seen as a class-wise actionness detector (i.e. binary classifier). Thus, this setting enables us to process concurrent actions in densely-labeled videos. Following the same setting as in [15, 13], we obtain the action detection results with frame-based mAP (i.e. per-frame mAP). Different from densely-labeled videos, there is no concurrent action or dense action region in sparsely-labeled videos. Hence, similar to [11, 5], we add an additional class-label as background and apply Softmax activation to generate the class probabilities for every snippet. Thanks to the background probabilities, we can predict the action boundaries (i.e. proposal). For each proposal, we average the class probabilities of all the corresponding snippets and select the action class with the highest score as the predicted class label.

Two-stream network: While evaluating two-stream performance, we average the prediction logits of both streams before the evaluation to have the two-stream results.

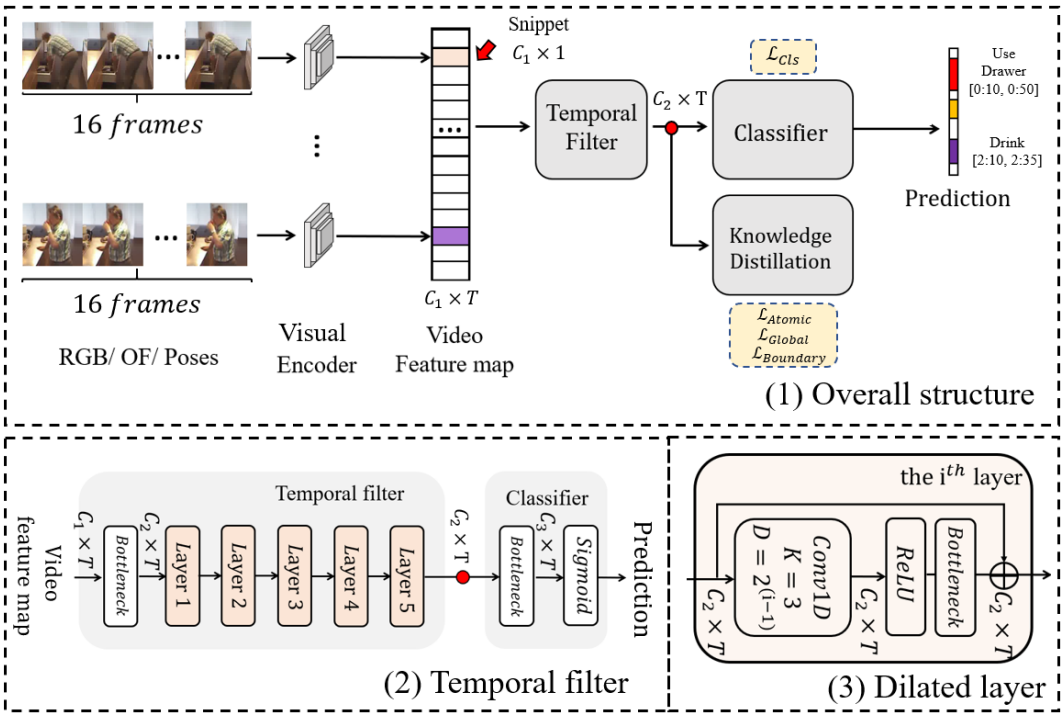

Figure 1. Network structure. (1) Overall structure of the processing pipeline, composed of three main components: the visual encoder, the temporal filter and the classifier. The location where the losses are applied is given in this figure. (2) We present the structure of our default temporal filter, the 5-layer SSTCN. (3) We provide the structure of a dilated layer in SSTCN.

2. More Qualitative Results

In this section, we provide more qualitative results of our proposed method.

Discriminative Power: As our proposed distillation framework predicts the classes of the snippets, we use a T-SNE plot [18] of the snippet features augmented by different distillation losses to reflect the discriminative power of these loss terms. In Fig. 2, we display the T-SNE plot of 12 actions of the vanilla-RGB stream and the RGB stream augmented by OF stream using different losses. We find that the vanilla-RGB stream often confuses the actions which have a similar appearance but different motion (e.g. *wear on or take off glasses*). In this figure, we also find that the proposed distillation losses $\mathcal{L}_{Atomic}$, $\mathcal{L}_{Global}$ and $\mathcal{L}_{Boundary}$ can help disambiguate part of these actions by infusing OF knowledge into RGB. As $\mathcal{L}_{Atomic}$ transfers the knowledge only from the corresponding snippet in the teacher stream, it can help to learn the regular atomic actions with salient motion, such as *sit down*. In addition, $\mathcal{L}_{Global}$ and $\mathcal{L}_{Boundary}$ can transfer the cross-snippet knowledge, more specifically, the global contextual information and the temporal evolution of the snippet features. As a result, sequence-level distillation ($\mathcal{L}_{Global} + \mathcal{L}_{Boundary}$) has larger discriminative power than atomic-level distillation ($\mathcal{L}_{Atomic}$) for longer and more complex actions. While applying all three loss terms, Augmented-RGB with strong discriminative power achieves the best performance, especially for ambiguous actions, such as *put something in pocket*. In Fig. 3, we also compare the T-SNE between the vanilla RGB and the Pose Augmented-RGB ($\mathcal{L}_{Total}$). We find that the vanilla RGB stream often mis-classifies similar actions with slightly different postures (e.g. *Jump up / hopping*). With the infusion of Poses into the RGB stream through our distillation strategies, the Augmented-RGB stream can now disambiguate these actions at inference time.

Boundary Saliency: As shown in Fig. 4, we display the boundary saliency plot of an example action instance. The curve represents the variation of the snippet features across time. We notice that both OF-Augmented-RGB and Pose-Augmented-RGB are more sensitive to the action boundaries than the vanilla-RGB, reflecting the effectiveness of $\mathcal{L}_{Boundary}$. As a result, the Augmented-RGB can better detect the action boundaries (see also Fig. 5 in main paper).

Covariance Matrix: We have provided in Fig. 4 of the main paper, the covariance matrix after a threshold of 0.5 for better visualization. We provide here, in Fig. 5, the covariance matrices (before threshold) of the vanilla RGB, vanilla OF, two-stream RGB+OF, and the OF Augmented-RGB networks. We observe that, with only the help of $\mathcal{L}_{Global}$, the global sequence statistics of Augmented-RGB are closer to the statistics of RGB+OF than vanilla RGB

and OF streams. Hence, the Augmented-RGB network can achieve similar performance than the one of the two-stream network.

3. Sensitivity of the hyper-parameters

In this section, we study the sensitivity of the hyper-parameters: α_i and $\mathcal{N}$.

Firstly, we discuss about the sensitivity of α_i, representing the weighting factors of the distillation losses. In Fig. 6, we provide the mAP performance with different values of α on Charades dataset. For each α_i, we fix the other hyper-parameters and fine-tune only the targeted α_i. In this figure, we reach a precision peak for $\{\alpha_1, \alpha_2, \alpha_3\}$=$\{300, 100, 5\}$, respectively.

Secondly, we discuss about $\mathcal{N}$, which represents the number of negative samples for each positive sample in contrastive learning. In previous work, as the training phase is conducted in an unsupervised manner, the increase of negative samples results in better performance for the image classification task [7, 16]. In this work, we train our network with the $\mathcal{L}_{atomic}$ in a supervised manner. Consequently, increasing the negative samples leads to similar convergence of the model (i.e. no significant improvement). Hence, in this work, our $\mathcal{N}$ is set to 1, meaning that one negative sample for each positive sample is sufficient.

The experiments show that our method is not sensitive to the value of these hyper-parameters. Therefore, we set the same value of the α_i and $\mathcal{N}$ for all the evaluated datasets.

4. Comparison with KD Methods in Model Compression

In the main paper (Sec.4.4), we have compared our approach with the cross-modal KD methods for action detection from the state-of-the-art. As mentioned in Section 2.2 of the main paper, among the KD methods, the model compression framework (i.e. cross-model) has the same input for teacher and student network, but the teacher network is more complex and more powerful. However, the cross-modal distillation framework has a similar architecture for teacher and student networks but different input modalities. For a fair comparison, we adapt the other KD methods from the other domain (i.e. model compression) towards conducting the cross-modal action detection task. We detail here five state-of-the-art methods of model compression: RKD-DA [12], SP [17], structural knowledge distillation (SKD) [10], Chen et al. [2] and CRD [16]. For RKD-DA and SP, the original works explore the relation of the image samples within a mini-batch. RKD-DA minimizes the distances between the batch samples while SP minimizes the similarity among the snippets. Here, we utilize these methods to explore the relations between the snippet features. Similar to SP, SKD transfers the similarity of the pixels at the feature-level for the semantic segmentation task, but adds two logit-level losses. As discussed in Sec.4.3 in main paper, logit-level distillation fails in the cross-modal case on many datasets. Thus, SKD gets even lower performance than SP on Charades and PKU-MMD. Most distillation frameworks for object-detection are designed for anchor-based architecture, which is very different in contrast to our Seq2Seq architecture for action detection (Sec.2.1 in main paper). For Chen et al. [2], as we do not have the regression module in our framework, we adapt only the $\mathcal{L}_{soft} + \mathcal{L}_{hint}$. For CRD, different from $\mathcal{L}_{atomic}$, we train the network in two steps for classification and distillation. Our experiments in Table 1 show that the proposed method outperforms all the methods (initially designed for model compression) for the task of action detection.

Table 1. Comparison with the model compression KD methods.

	Charades	PKU-MMD (0.1)
RKD-DA [12]	22.9	82.4
SP [17]	22.8	81.7
SKD [10]	22.6	81.6
Chen et al. [2]	23.4	82.2
CRD [16]	23.1	82.0
Ours ($\mathcal{L}_{Total}$)	**24.6**	**85.5**

5. Dataset Description & Complete State-of-the-Art Table

In this section, we describe the five datasets used to evaluate our method. Table 2 summarizes the properties of the datasets. Fig. 7 shows an example of a video with densely-labeled annotation and coarsely-labeled annotation, illustrating that the densely-labeled videos are more challenging [19].

Charades [15] was recorded by hundreds of people in their private homes. This dataset consists of 9848 videos across 157 actions. The actions are mainly object-based daily living actions performed at home. Each video is about 30 seconds containing complex co-occurring actions. In our experiments, we follow the original Charades settings for action detection [15] (i.e. Charades v1 localize evaluation). The performances are measured in terms of mAP by evaluating per-frame prediction.

PKU-MMD [9] covers a wide range of complex human actions with well annotated information. This dataset con-

Table 2. Dataset information.

	Annotation Type	Video Length	Video Type
TSU [4]	Dense	Long	ADL
PKU-MMD [9]	Sparse	Long	ADL
Charades [15]	Dense	Short	ADL
MultiTHUMOS [19]	Dense	Short	Sport
THUMOS14[8]	Sparse	Short	Sport

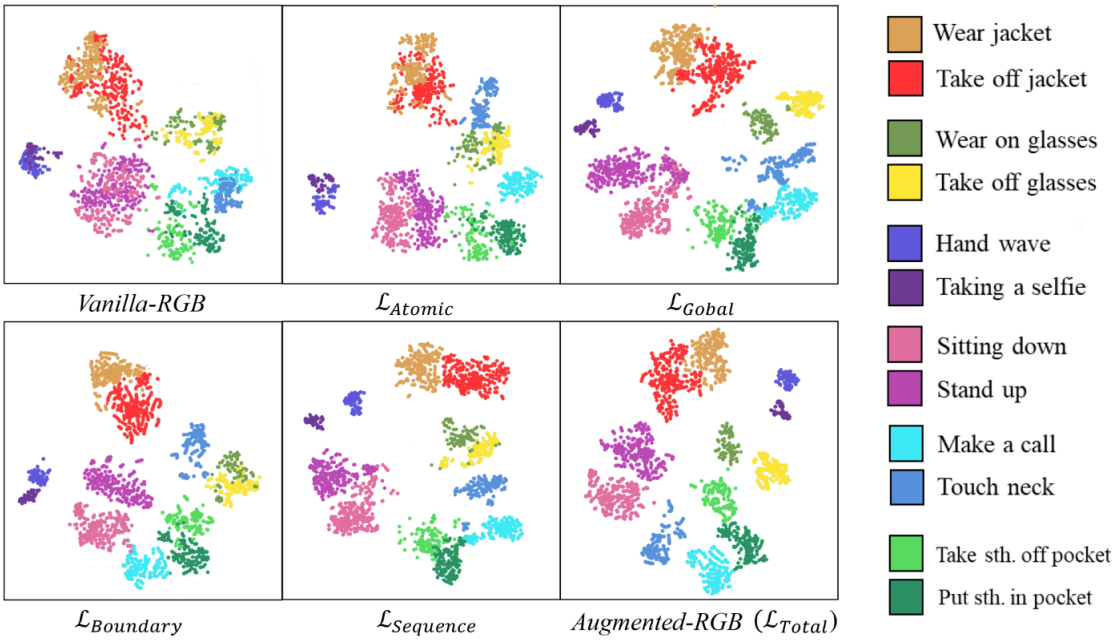

Figure 2. We present the T-SNE visualization of RGB stream augmented by OF in different settings. Note that: each point represents a snippet. The experiment is conducted on PKU-MMD dataset.

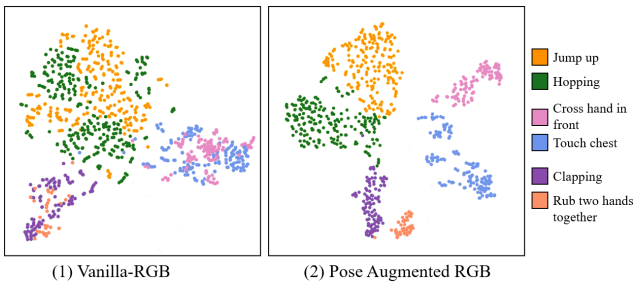

Figure 3. T-SNE for (1) vanilla-RGB stream and (2) Pose-augmented-RGB stream. The experiment is conducted on PKU-MMD dataset.

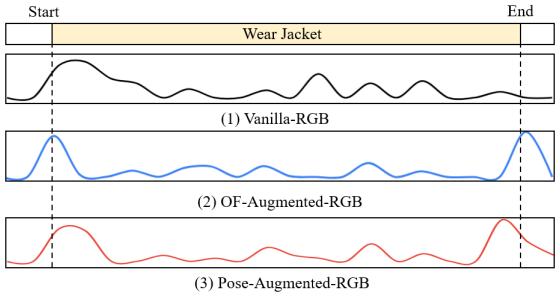

Figure 4. Boundary Saliency of (1) Vanilla-RGB, (2) OF-augmented-RGB, (3) Pose-augmented-RGB. The experiment is conducted on PKU-MMD dataset.

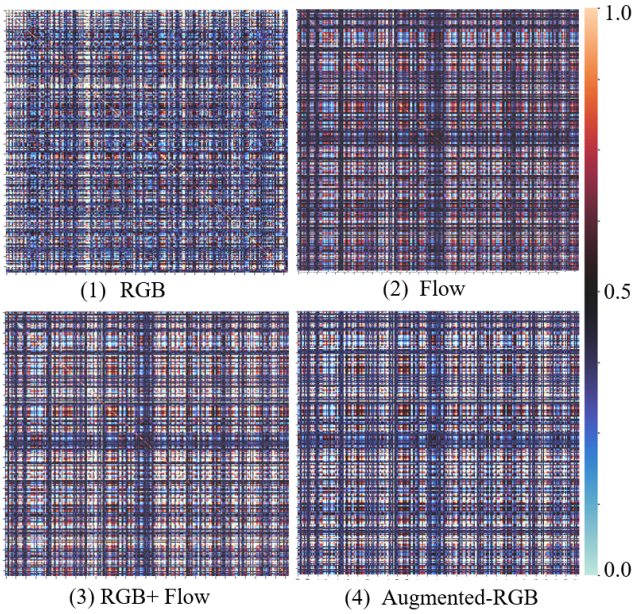

Figure 5. Covariance matrices of four networks on Charades: (1) Vanilla-RGB, (2) Vanilla-OF, (3) Two-stream-RGB+OF and (4) OF-Augmented-RGB.

tains 1076 long video sequences in 51 action categories, performed by 66 subjects. PKU-MMD provides multi-modality data sources, including RGB, depth, Infrared Radiation and Skeleton. Following the original paper, the performances are evaluated in terms of event-based mAP in Cross-Subject protocol (CS).

Toyota Smarthome Untrimmed (TSU) [4] is a real-world action detection dataset. This dataset consists of 536 long

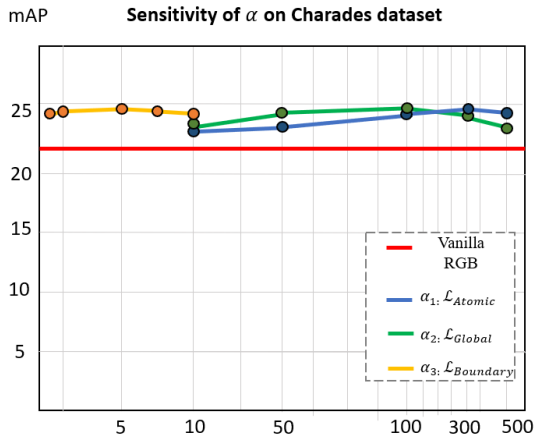

Figure 6. Sensitivity of α_i. The X-axis represents the value of the alpha. The Y-axis represents the mAP of the vanilla-RGB network compared to the Augmented-RGB network while fine-tuning the α_i.

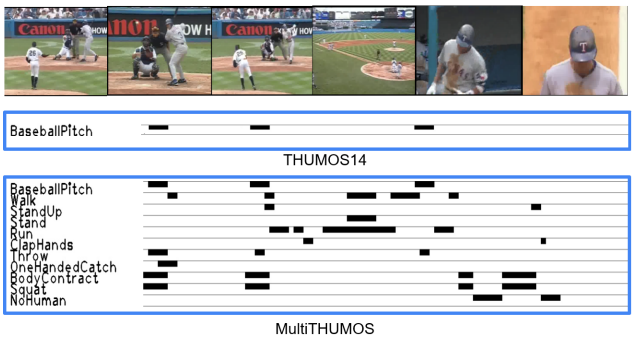

Figure 7. Example of a video with (1) Sparsely-labeled annotation: THUMOS14 and (2) Densely-labeled annotation: MultiTHUMOS.

videos (about 20 mins/video) recorded by 7 cameras with 51 densely annotated action classes. Besides long video duration, this dataset contains actions with high intra-class temporal variance. As a result, handling temporal information is critical to achieve good detection performance on this dataset. This dataset uses both frame-based and event-based mAP for evaluation.

THUMOS14 [8] and MultiTHUMOS [19]: We conducted our experiments on both THUMOS14 [8] and MultiTHUMOS [19] datasets, while using the more challenging MultiTHUMOS as the main dataset (see Fig. 7). MultiTHUMOS is an enhanced version of the THUMOS14 dataset, where videos are densely annotated. The dataset consists of 65 action classes, compared to 20 in THUMOS14, and contains on average 10.5 action classes per video and 1.5 labels per frame and up to 25 different action labels in each video. THUMOS14 and MultiTHUMOS consists of YouTube videos of various sport actions like baseball games, cliff diving.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017. 1

[2] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 742–751, 2017. 3

[3] Rui Dai, Srijan Das, Luca Minciullo, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Pdan: Pyramid dilated attention network for action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2970–2979, January 2021. 1

[4] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *arXiv preprint arXiv:2010.14982*, 2020. 3, 4

[5] Xiyang Dai, Bharat Singh, Joe Yue-Hei Ng, and Larry Davis. Tan: Temporal aggregation network for dense multi-label action recognition. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 151–160. IEEE, 2019. 1

[6] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3575–3584, 2019. 1

[7] 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

[8] Yu-Gang. Jiang, Jingen Liu, Amir Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014. 3, 5

[9] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for skeleton-based human action understanding. In *Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities*, VSCC '17, pages 1–8, New York, NY, USA, 2017. ACM. 3

[10] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2019. 3

[11] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1

[12] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pages 3967–3976, 2019. 3

[13] AJ Piergiovanni and Michael S Ryoo. Temporal gaussian mixture layer for videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[14] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019. 1

[15] Gunnar A Sigurdsson, Santosh Divvala, Ali Farhadi, and Abhinav Gupta. Asynchronous temporal fields for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 585–594, 2017. 1, 3

[16] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020. 3

[17] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1374, 2019. 3

[18] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 2

[19] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126(2-4):375–389, 2018. 1, 3, 5