# CheckMATE: Efficient Video Summarization by Checking Mutually Averaged Temporal Encapsulation

Masud An-Nur Islam Fahim† , Jani Boutellier†
† University of Vaasa, Finland
{ masud.fahim, jani.boutellier}@uwasa.fi

## Abstract

*Video classification is a computationally demanding task at inference time, but especially at training time. The computation burden originates both from the number of training sequences needed, and from the high-volume data content of each sequence. On the model architecture side, video recognition is dominated by 3D ConvNets that are computationally much more demanding than their 2D counterparts. This paper proposes a simple yet efficient solution for large-scale learning from videos: the entire video clip is summarized into a single frame, which offers visual recognition performance comparable to the original video stream. The proposed video summarization algorithm distills the input video into a single frame in the feature space of the image classifier. After compressing the videos into individual frames, regular image classification training is performed for the purpose of action recognition. We validate the performance of our approach on UCF101 and HMDB51 datasets and observe results comparable to competing approaches that leverage expensive 3D ConvNets. In contrast, our approach uses only 2D image classification networks and does not require any pre-training on video datasets.*

## 1. Introduction

Video-based human action recognition is currently an intensively studied topic, where an action classifier is fed a video stream of arbitrary length with the objective of classifying the action taking place in the video [1, 4, 5, 25]. Most studies achieve this [1, 2, 4, 5, 13, 25] by relying on a CNN-based backbone. Since pure 2D image classifiers, commonly used for image recognition tasks, disregard the temporal context of videos [13], researchers are increasingly relying on 3D ConvNets, which however are very costly in terms of computation [8, 13].

In addition to their computation cost, 3D convolution operations also have challenges in capturing long spatiotemporal relationships [13]. Video transformers (e.g., [1]), on the other hand, often fail to encapsulate per-frame spatiotemporal features in shallow layers [13]. Also, training of both 3D ConvNets and transformers requires extensive tuning for achieving high classification performance [13].

This study addresses the issues mentioned above by *video summarization* that relaxes the video classification task into an image classification task; thus spatiotemporal feature extraction can be performed without resorting to 3D CNNs or video transformers. Moreover, since video sequences offer rich feature content and a lot of redundancy, video summarization enables eliminating temporal redundancy for the benefit of training efficiency.

In our video summarization approach, we segment the input videos into multiple clips and extract their temporal encapsulation with a pre-trained backbone network. Simultaneously, we initialize the summary frame by taking the average of a random video clip over time, followed by similar encapsulation of the initial summary image. Consequently, we perform a minimization process between the summary image and multi-clip encapsulation to update the summarized image. The updated image is then distilled by **Check**ing **M**utually **A**veraged **T**emporal **E**ncapsulation from all the clips, resulting in spatiotemporal summarization of the whole video into a single frame, for action classification. We have labeled our approach **CheckMATE**, and the contributions of this work can be expressed as follows:

- CheckMATE provides a simple, yet effective video summarization method that enables single image classification to act as a proxy task for action recognition.
- Differing from previous works, CheckMATE does not rely on customized regularizers or handcrafted features, which reduces computational complexity and mitigates possible feature biases.
- CheckMATE can be applied to 2D/3D action classifiers and delivers high classification performance using single-mode input, without additional supervision in action classification.

- To our best knowledge, CheckMATE is the first study to cover ultra-low resolution action classification with synthetic frames, performing better than previous dedicated studies.

## 2. Related work

**3D CNNs.** 3D ConvNets have had a significant influence to action detection and classification. Typically, these ConvNets process the video stream as short 3D patches, aiming to extract both spatial and temporal features. For efficient temporal modeling with 3D ConvNets, [10] uses a combination of long and short temporal sampling, [14] proposes temporal shifting, whereas [9] introduces temporal regulation. Although these techniques overcome problems related to the temporal fitting of clips within 3D convolutions, the computation cost issue of 3D ConvNets remains. Computation cost has been addressed in [5] by means of inflated 2D ConvNets. Further works on efficient 3D ConvNet based action recognition include [15, 16] and [20].

**Vision Transformers.** Recently, vision transformers [7] have influenced various recognition tasks after presenting competitive performance in image classification. Following ViT, several action recognition models have introduced transformer modules with customized attention operations: by patchifying input frames, ViT variants have used spatiotemporal attention [19, 31, 33, 37], factorized dot-product attention [1], multi-head attention [18], attention gates [36], cross-view attention [35], and divided space-time attention [2]. Relying on the transformer mechanism, [13] revisited ConvNets and unified convolution with self-attention to improve extraction of spatiotemporal representation.

**Single image action classification.** Another line of research investigates action recognition tasks through single image classification [3, 22, 24, 27, 28], where a training image represents a summary of the whole video through customized information aggregation. [3] identifies actions through intra-object relation per frame, [24] predicts missing temporal features, and [22] summarizes the input video in up to eight frames for action recognition. Similar video summarization works also utilize adaptive weighted operations [28] and adversarial distillation [27].

## 3. Proposed method

Before going into the technical description, we summarize the necessary notations below.

**Technical terms.** Say, a video sequence is comprised of $M$ clips such that $\mathcal{V} = \{v_n | n = 1, ..., M\}$, and each clip has $T_f$ frames. To establish a temporal encapsulation of a clip $v_n$ we use a 2D backbone network $\mathcal{B}_\theta$, whereas for action classification we use the 2D network $\mathcal{F}_\theta$. Moreover, $\mathcal{F}^*_\theta$ denotes the pre-trained network, and $\mathcal{F}_{\theta_l}$ refers to the feature embedding at layer $l$ for a given input. At the beginning

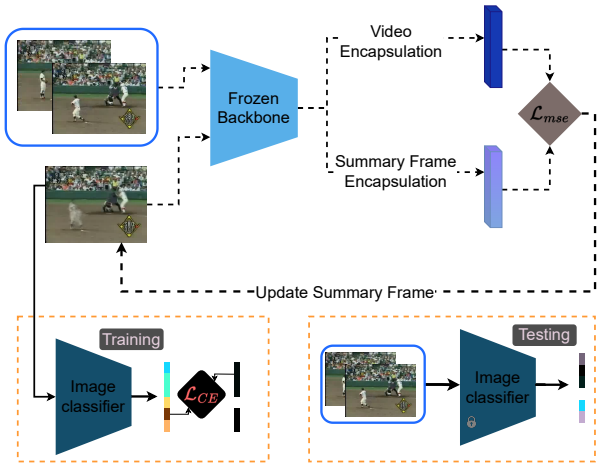


Figure 1. Flowchart of **CheckMATE**, which divides given videos into multiple clips and returns a summary frame for each video. The video summary dataset can then be applied to action classification.

of video summarization, we initialize the summary frame as $\mathcal{V}^s$. We denote the mean-square loss and the cross-entropy loss functions by $\mathcal{L}_{mse}$ and $\mathcal{L}_{ce}$, respectively.

**CheckMATE.** In the first step of the proposed CheckMATE approach, we initialize a candidate frame for video summarization: we extract a random clip $v_n$ from the video sequence $\mathcal{V}$ using a large frame count (e.g., $T_f = 64$) and average the clip into a single image, providing the initial summary image $\mathcal{V}^s$ such that $\mathcal{V}^s = \frac{1}{T_f} \sum_{i=1}^{T_f} v_{n_i}$. For $T > 1$ summary frames, we divide the input clips into $T$ groups by the order of the clips, followed by the averaging operation, and concatenate the resulting $T$ averaged frames into an initial summary clip with $T$ images. Here, $T_f$ denotes frame count for a given clip taken from the input video, and $T$ the number of summary frames.

The summary image could also be initialized by picking a random frame from the video sequence, or by initializing the summary image via random noise. However, for general video sequences, it is possible that a random frame might not contain action, whereas the use of random noise could complicate the optimization process due to visual fidelity constraints. The chosen approach of frame averaging avoids both of these pitfalls.

Next, from the segmented clips of $\mathcal{V}$, the temporal encapsulations are extracted through a pre-trained backbone network $\mathcal{B}_\theta$. Our study used ConvNext-small as the backbone. In temporal encapsulation, we concatenate feature embeddings from higher, middle, and lower layers of $\mathcal{B}_\theta$. Simultaneously, the backbone model $\mathcal{B}_\theta$ extracts temporal encapsulation from the initial summary image $\mathcal{V}^s$.

Consequently, we compare the average similarity of encapsulations extracted from the clips against the initial summary image — similarity matching is performed through $\mathcal{L}_{mse}$. After acquiring the loss, the frame gradient is calculated and the summary frame is updated, as shown in Fig. 1.

The aforementioned sequence of operations is repeated multiple times, comparing the temporal encapsulation similarity between the clips and the summary image. Over repetition, the summary frame converges to a state where the average temporal encapsulation for all the clips is equivalent (typically, $\mathcal{L}_{mse} \leq 0.01$) to the updated frame's encapsulation, with respect to the backbone model $\mathcal{B}_\theta$, which is a 2D ConvNet for both single and multiple summary frames.

For the given backbone $\mathcal{B}_\theta$, and any random video $\mathcal{V}$, we can symbolize our CheckMATE video summarization by

$$\min \frac{1}{|\mathcal{M}|} \sum_{\mathcal{M}} \mathcal{L}_{mse}(\mathcal{B}_\theta(\mathcal{V}^s), \mathcal{B}_\theta(v_n))$$

**High-low distillation.** We have applied summary frames generated by CheckMATE to ultra-low resolution action classification. In this task, the pre-trained high-resolution action classifier was used to distill the low-resolution implementation. The low-resolution summary images are passed to both high and low-resolution classification networks, and feature embeddings are extracted for minimizing feature similarity loss. Simultaneously, the low-resolution classifier is updated with regular cross-entropy loss. For any low-resolution clip $v_i$ from $\mathcal{S}$ clips, features from the low-resolution model $\mathcal{E}_1 = \mathcal{F}_{\theta_l}(v_i)$, and features from the pre-trained high-resolution model $\mathcal{E}_2 = \mathcal{F}^*_{\theta_l}(v_i)$, and label $y_i$, form the overall expression of our low-resolution action classification distillation as:

$$\min_{\mathcal{F}_\theta} \frac{1}{|\mathcal{S}|} \sum_{i} \mathcal{L}_{mse}(\mathcal{E}_1, \mathcal{E}_2) + \mathcal{L}_{ce}(\mathcal{F}_\theta(v_i), y_i)$$

**Inference.** In training, we extract the logits from the image classifier model by plugin summary frames and update the classifier. During validation, we feed unseen test videos to the optimized classifier for final prediction, similar to traditional action classification tasks. We do not need to extract summary frames for the test sequences, as summary frames work as the proxy for regular videos in action classification.

## 4. Experiments and results

**Datasets.** The CheckMATE approach was evaluated using the UCF101 [26] and HMDB51 [12] datasets that are commonly used for action recognition. UCF101 has 9.5K training videos for 101 action classes, whereas HMDB51 has 3.7K videos for 51 action classes.

**Implementation details.** In our study, we have used ConvNext [17] as our backbone for both video summarization and action classification. The ConvNext-small model [17] had only been pre-trained with the ImageNet-1k dataset (no ImageNet-21k pre-training), and no additional training with video datasets like Kinetics-400/600 was performed.

CheckMATE adopts random (crop, flip, brightness, hue, saturation, and contrast) image variation while summarizing the input stream into a single frame. Rotation, flip, and spatial dropout were used for image classifier training. Each video was segmented into ten clips, and gradient descent with 0.5 step size was used for video summarization. The image classification step used the Adam optimizer, 40 training epochs, and a batch size of 16 for each dataset. We used the Tensorflow-Keras framework on a single RTX 3090 GPU for all tasks.

Table 1. Comparison with state-of-the-art on UCF101 and HMDB51 *in full resolution* using single ($T = 1$) and double ($T = 2$) summary frames for CheckMATE. Notice that CheckMATE does not use 3D CNNs, Kinetics supervision (+Kinetics), or optical flow (+Flow).

| Method | +Flow | +Kinetics | U101 | H51 |
|---|---|---|---|---|
| IDT [30] | | | 86.4 | 61.7 |
| Two-stream [25] | ✓ | | 88.0 | 59.4 |
| TSN [32] | ✓ | | 94.2 | 69.4 |
| I3D [5] | | ✓ | 95.4 | 74.5 |
| S3D [34] | | ✓ | 96.8 | 75.9 |
| LGD-3D [21] | | ✓ | 97.0 | 75.7 |
| STM [11] | | ✓ | 96.2 | 72.2 |
| AVD [27] | ✓ | ✓ | 97.3 | 77.1 |
| I3D [5] | ✓ | ✓ | 97.9 | 80.2 |
| R(2+1)D [29] | ✓ | ✓ | 97.3 | 75.9 |
| LGD-3D [21] | ✓ | ✓ | 98.2 | 80.5 |
| IFS-3D [22] | | ✓ | 97.4 | 76.2 |
| IFS-3D+IFS-mot-3D [22] | | ✓ | 98.2 | 80.3 |
| **CheckMATE** (T = 1) | | | 80.7 | 66.8 |
| **CheckMATE** (T = 2) | | | 93.2 | 77.3 |

Table 2. Ultra-low resolution video action classification for the UCF-101 dataset. CheckMATE outperforms previous works with a clear margin in the 14×14 / 16×12 regime.

| Method | Input | Accuracy % |
|---|---|---|
| I3D [5] | 112×112 | 84.72 |
| SoSR [25] | 80×60 | 83.92 |
| Bicubic - I3D [5] | 14×14 | 14.14 |
| DVSR [6] | 14×14 | 68.17 |
| Prog. DVSR [6] | 14×14 | 70.55 |
| **CheckMATE (T = 2)** | **16×12** | **77.75** |
| Bicubic - I3D [5] | 28×28 | 66.72 |
| DVSR [6] | 28×28 | 82.37 |
| Prog. DVSR [6] | 28×28 | 82.87 |
| **CheckMATE (T = 2)** | **28×28** | **82.93** |

### 4.1. Performance evaluation

Two main evaluation results are presented: a) normal resolution action classification, and b) ultra-low resolution action classification. Currently, we only report results for UCF101 and HMDB51 datasets.

**Normal resolution action classification.** For normal resolution action classification, IFS [22] is regarded as the most relevant comparison to our work. From Table 1, it can

Table 3. Ultra-low resolution video action classification comparison for the HMDB-51 dataset. CheckMATE outperforms previous works with a clear margin.

| Method | Input | Accuracy % |
|---|---|---|
| I3D [5] | 112×112 | 52.61 |
| SoSR [25] | 80×60 | 54.77 |
| Bicubic - I3D [5] | 14×14 | 10.59 |
| Privacy-Preserv. [23] | 12×16 | 28.68 |
| F. Coupled [4] | 12×16 | 39.15 |
| DVSR [6] | 14×14 | 41.24 |
| Prog. DVSR [6] | 14×14 | 41.63 |
| **CheckMATE (T = 2)** | **16×12** | **47.68** |
| Bicubic - I3D [5] | 28×28 | 46.97 |
| Privacy-Preserv. [23] | 24×32 | 32.15 |
| DVSR [6] | 28×28 | 53.66 |
| Prog. DVSR [6] | 28×28 | 55.95 |
| **CheckMATE (T = 2)** | **28×28** | **59.33** |

be seen that IFS [22] equals or slightly surpasses the performance of other methods.

Even though IFS [22] proposes single-frame video summarization, it uses eight summarized frames and a 3D CNN to achieve state-of-the-art performance for UCF101 and HMDB51 datasets. Additionally, IFS uses Kinetics-400 fine-tuning for the results presented in Table 1. Similarly, [27] adopts multiple regularizers for video frame summarization, which necessitates manual tuning of hyperparameters for quality optimization. Furthermore, both [22, 27] are dependent on a 3D encoder for video summarization, followed by using a 3D classifier for action classification. On the contrary, CheckMATE offers performance comparable to [22, 27] without relying on a 3D encoder, 3D action classifier, multiple regularizers, extra hyperparameters, additional modalities, or Kinetics supervision.

For all of our results, better action classification performance was observed for our image classifier with double ($T = 2$) summary frames. This signals that further performance improvement could be achieved by incorporating further summary images. On the downside, multiple frame distillation would escalate the overall time cost for CheckMATE.

**Ultra-low resolution action classification.** For ultra-low resolution action classification, we have used 16x12 and 28x28 video resolution for training and evaluation, as present in Tables 2 and 3. During training, we distill the low-resolution model with the help of the pre-trained model with full-resolution synthetic frames. To reduce overfitting during model distillation, we have used high dropouts in dense layers of the low-resolution model. For low-resolution action classification tasks, previous works have often used super-resolution or multiple modalities (e.g. optical flow) [5, 6] for better performance. Surprisingly, our CheckMATE approach is independent of such operations and easily achieves better results without such additional information. Notice that we did not resynthesize new low-resolution summary images for this task, instead straight-forward downsampling was used.

Table 4. CheckMATE summary frame count vs. required time to complete three iterations. The iteration count is proportional to time and respective performance increase.

| Method | $T = 1$ | $T = 2$ | $T = 8$ |
|---|---|---|---|
| CheckMATE | **4.0 s** | **5.8 s** | **10.0 s** |

**CheckMATE complexity.** Large-scale dataset summarization is an extraordinarily time-consuming task [38], even more so if the number of classes is high. As Table 4 shows, this is not the case for CheckMATE. The proposed approach is computationally straightforward, as the distillation essentially summarizes each video into one or at most, eight frames. Unlike, e.g., I3D [5], CheckMATE is independent of 3D models, extra data modalities, tuning of multiple models, multistage training, or tedious hyperparameter tweaking; essentially, CheckMATE offers more efficient training while preserving scalability for massive datasets along with transferability to lower dimensional data. During action classification with the summarized dataset, only a simple 2D classifier is used, trained for 100 epochs. This makes our method efficient and affordable for large-scale video categorization task.

## 5. Conclusion

Our study explores video summarization by encapsulating video information into a single-frame representation. For this task, the proposed method relies on an off-the-shelf image classifier for feature space optimization to obtain the distilled frame for each video. Following this concept, a straightforward image classification task on summary frames provides a trained action classifier. The proposed method does not limit itself to single-frame condensation. Experimentally we have demonstrated the efficacy of multiple frames for action classification over a single frame. We have generated 2D image classifiers for action recognition applications with our distilled frames and obtained performance comparable to several state-of-the-art 3D CNN networks that are however much more expensive in terms of computation. The implementation of CheckMATE presented here relies only on ImageNet1k in terms of pre-training and does not use other video stream modalities such as the commonly used optical flow. In the future, we aim to conduct a more in-depth study where large 3D classifiers and richer image datasets will aid our video frame condensation method to validate its performance on large-scale action datasets like Kinetics 400/600.

## Acknowledgements

This work has been partially funded by the Academy of Finland project SPHERE-DNA.

## References


[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 1, 2

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 1, 2

[3] A Bobick and J Davis. The representation and recognition of action using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001. 2

[4] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4778–4787, 2017. 1, 4

[5] 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, 2, 3, 4

[6] Ugur Demir, Yogesh S Rawat, and Mubarak Shah. Tinyvirat: Low-resolution video action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7387–7394. IEEE, 2021. 3, 4

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[8] Quanfu Fan, Chun-Fu Chen, and Rameswar Panda. Can an image classifier suffice for action recognition? In *International Conference on Learning Representations*, 2022. 1

[9] Quanfu Fan, Chun-Fu Richard Chen, Hilde Kuehne, Marco Pistoia, and David Cox. More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[10] 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. 2

[11] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. Stm: Spatiotemporal and motion encoding for action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2000–2009, 2019. 3

[12] 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. 3

[13] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022. 1, 2

[14] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019. 2

[15] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. Teinet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11669–11676, 2020. 2

[16] Zhouyong Liu, Shun Luo, Wubin Li, Jingben Lu, Yufan Wu, Shilei Sun, Chunguo Li, and Luxi Yang. Convtransformer: A convolutional transformer network for video frame synthesis. *arXiv preprint arXiv:2011.10185*, 2020. 2

[17] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 3

[18] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 2

[19] Vittorio Mazzia, Simone Angarano, Francesco Salvetti, Federico Angelini, and Marcello Chiaberge. Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition*, 124:108487, 2022. 2

[20] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 2

[21] Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xinmei Tian, and Tao Mei. Learning spatio-temporal representation with local and global diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12056–12065, 2019. 3

[22] Zhaofan Qiu, Ting Yao, Yan Shu, Chong-Wah Ngo, and Tao Mei. Condensing a sequence to one informative frame for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16311–16320, 2021. 2, 3, 4

[23] Michael Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 4

[24] Marjaneh Safaei and Hassan Foroosh. Still image action recognition by predicting spatial-temporal pixel evolution. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 111–120. IEEE, 2019. 2

[25] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 1, 3, 4

[26] 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. 3

[27] Mohammad Tavakolian, Mohammad Sabokrou, and Abdenour Hadid. Avd: Adversarial video distillation. *arXiv preprint arXiv:1907.05640*, 2019. 2, 3, 4

[28] Mohammad Tavakolian, Hamed R Tavakoli, and Abdenour Hadid. Awsd: Adaptive weighted spatiotemporal distillation for video representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8020–8029, 2019. 2

[29] 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. 3

[30] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013. 3

[31] Jue Wang and Lorenzo Torresani. Deformable video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14053–14062, 2022. 2

[32] 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. 3

[33] Yuetian Weng, Zizheng Pan, Mingfei Han, Xiaojun Chang, and Bohan Zhuang. An efficient spatio-temporal pyramid transformer for action detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 358–375. Springer, 2022. 2

[34] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 3

[35] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3333–3343, 2022. 2

[36] Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14063–14073, 2022. 2

[37] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. *arXiv preprint arXiv:2110.08568*, 2021. 2

[38] Lei Zhang, Jie Zhang, Bowen Lei, Subhabrata Mukherjee, Xiang Pan, Bo Zhao, Caiwen Ding, Yao Li, and Dongkuan Xu. Accelerating dataset distillation via model augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11950–11959, 2023. 4