Audio-Visual Glance Network for Efficient Video Recognition

Supplementary Material

1. Notation List

For the convenience of the reader, we listed the Table of Notation containing frequently used notations along with their definition in Table 1.

2. Datasets

ActivityNet-1.3 [6] contains 10,024 training videos and 4,926 validation videos sorted into 200 human action categories. The average duration is 117 seconds.

FCVID [13] contains 45,611 videos for training and 45,612 videos for validation, which are annotated into 239 classes. The average duration is 167 seconds.

Mini-Kinetics is a subset of the Kinetics [14] dataset. We establish it following [10, 24, 25, 44]. The dataset include 200 classes of videos, 121k for training and 10k for validation. The average duration is around 10 seconds [14].

3. Implementation Detail

3.1. Network Architecture

Encoders. For audio encoder f_{A} and f_{G} we use MobileNetV2[**?**] and for local visual encoder f_{L} we use ResNet-50[**?**]. We use a patch size of 128×128 for the input to f_L, thus the size of the patch extracted by the patch extraction network is also the same. To encode a single image, the f_{G} requires 0.33 GFLOPs and f_{L} requires 1.35 GFLOPs, meanwhile to encode the whole audio sequence f_{A} requires 0.68 GFLOPs.

AV-TeST. In our implementation we construct TF_{AV} using a multi-head attention transformer[33] with 256 encoder dimension size, 2 stacks, and 4 heads. As the input to the transformer is concatenated audio-visual feature, for each modality we embed them to $128 - d$ vectors with separate linear embedding layers. To reconstruct the visual token, we utilize a transformer with the same architecture and append a linear embedding layer at the end.

AESPA. In our implementation of AESPA module, we use the same transformer architecture for both audio and visual modality. To minimize the computational burden, we reduce the incoming channel of both audio and visual modality to 256. Then we use the reduced feature maps as input to the bottleneck fusion transformers. Each modality transformer consists of 4 stack of encoder with 4 heads. We use 4 bottleneck tokens to be appended to the modality tokens.

Training Details. To train the network, we use an SGD optimizer with cosine learning rate annealing and a momentum of 0.9. The L2 regularization co-efficient is set to 1e-4. The two encoders f_{G} and f_{L} are initialized using the ImageNet pre-trained models[1], while the rest of the network is trained from random initialization. The size of the mini-batch is set to 24. The initial learning rates of f_{G}, f_{A}, f_{L}, f_{C}, π, æ, and TF_{AV} are set to 0.001, 0.001, 0.002, 0.01, 2e-4, 2e-4, and 0.01. We use a masking ratio of 0.75 for L_{mask}, and for Gumbell-Softmax we use 5 as the temperature value

3.2. Patch Extraction Network.

We explain in detail the process inside the spatial patch extraction network. To enable end-to-end training, we adopt the differentiable solution proposed in [37] to obtain $\tilde{\boldsymbol{v}}_t$. Suppose that the size of the original frame $\boldsymbol{v}_t$ and the patch $\tilde{\boldsymbol{v}}_t$ is $H \times W$ and $P \times P$ $(P < H, W)$, respectively[2]. We assume that π outputs the continuous centre coordinates $(\tilde{x}_{\text{c}}^t, \tilde{y}_{\text{c}}^t)$ of $\tilde{v}_t$ using audio-enhanced global visual feature up to t^{th} $(\{e_1^{\text{GA}}, \ldots, e_t^{\text{GA}}\})$,

$$\begin{aligned} (\tilde{x}_{\text{c}}^t,\ \tilde{y}_{\text{c}}^t) &= \pi(\{e_1^{\text{GA}}, \ldots, \boldsymbol{e}_t^{\text{GA}}\}), \\ \tilde{x}_{\text{c}}^t \in [\frac{P}{2}, W - \frac{P}{2}], &\quad \tilde{y}_{\text{c}}^t \in [\frac{P}{2}, H - \frac{P}{2}], \end{aligned} \tag{1}$$

We refer to the coordinates of the top-left corner of the frame as $(0, 0)$, and Eq. (1) ensures that $\tilde{v}_t$ will never go outside of v_t.

The feed-forward process involves the bilinear interpolation method to enable backpropagation through $(\tilde{x}_{\text{c}}^t,\ \tilde{y}_{\text{c}}^t)$. As mentioned in the paper, the coordinates of a pixel in the patch $\tilde{v}_t$ can be expressed as the addition of $(\tilde{x}_{\text{c}}^t, \tilde{y}_{\text{c}}^t)$ and a fixed offset:

$$\begin{aligned} (\tilde{x}_{ij}^t,\ \tilde{y}_{ij}^t) &= (\tilde{x}_{\text{c}}^t,\ \tilde{y}_{\text{c}}^t) + o_{ij}, \\ o_{ij} &\in \left\{ -\frac{P}{2}, -\frac{P}{2} + 1, \ldots, \frac{P}{2} \right\}^2. \end{aligned} \tag{2}$$

[1] In most cases, we use the 224x224 ImageNet pre-trained models provided by PyTorch [**?**].

[2] In our implementation, the height/width/coordinates are correspondingly normalized using the linear projection $[0, H] \rightarrow [0, 1]$ and $[0, W] \rightarrow [0, 1]$. Here we use the original values for the ease of understanding.

Variables

Symbol	Definition
t	Frame or time index
a_t	Audio spectrogram clip at time
v_t	Input image frame at time t
y	label class
e_t^{A}	Audio feature at time t
e_t^{G}	Coarse/Global visual feature at time t
$z_{l,t}^{\mathrm{A}}$	AESPA audio vector at layer l at time t
$z_{l,t}^{\mathrm{G}}$	AESPA visual vector at layer l at time t
e_t^{GA}	Enhanced Coarse/Global visual feature at time t
e_t^{L}	Fine/Local visual feature at time t
e_t^{TF}	Audio-visual feature for AV-TeST input t
s_t	Frame saliency score at time t
$\tilde{e}_t^{\mathrm{A}}$	Transformed audio feature at time t
$(\tilde{x}_c^t, \tilde{y}_c^t)$	Center coordinates t
$\tilde{v}_t$	Visual patch at time t
$(\tilde{x}_{ij}^t, \tilde{y}_{ij}^t)$	Coordinates of pixel patch t
o_{ij}	Fixed offset for coordinate (i, j)
$\tilde{e}_t^{\mathrm{G}}$	AV-TeST embedded visual token (i, j)
$\hat{e}_t^{\mathrm{G}}$	Reconstructed AV-TeST embedded visual token
$p_t^{'}$	Softmax prediction of f_C^{AV} with feature only at time t
$\tilde{s}_t$	Pseudo-label saliency score
p_t	Class prediction

Functions

Symbol	Definition
f_A	Audio encoder
f_G	Global visual encoder
f_L	Local visual encoder
$\mathrm{TF}_{\mathrm{AV}}$	AV-TeST Transformer Network
FC_s	Saliency score prediction head
æ	Audio Enhanced Spatial Patch Attention (AESPA) module
$\mathrm{TF}_{\mathrm{A}}^l$	AESPA audio transformer at layer l
$\mathrm{TF}_{\mathrm{G}}^l$	AESPA visual transformer at layer l
π	Spatial patch extraction network
ψ	Fusion transformer
f_C^{AV}	Audio-visual classifier
f_C^A	Auxiliary audio prediction head
FC^{G}	Auxiliary frame-wise global visual prediction head
FC^{L}	Auxiliary frame-wise local visual prediction head
FC^{A}	Auxiliary frame-wise audio prediction head
f_C^{V}	Auxiliary visual prediction head

Hyperparameters

Symbol	Definition
T_G	Visual temporal glance limit
k	Number of selected frames for prediction
P	Patch size

Table 1. Table of Notation

$(\tilde{x}_{ij}^t, \tilde{y}_{ij}^t)$ denotes the corresponding horizontal and vertical coordinates in the original frame v_t to the i^{th} row and j^{th} column of $\tilde{v}_t$, while the offset o_{ij} is the vector from the patch center $(\tilde{x}_{\mathrm{c}}^t, \tilde{y}_{\mathrm{c}}^t)$ to this pixel. Given a fixed patch size, o_{ij} is a constant conditioned only on i, j, regardless of t or the inputs of π.

Since the values of $(\tilde{x}_{\mathrm{c}}^t, \tilde{y}_{\mathrm{c}}^t)$ are continuous, there does not exist a pixel of $\boldsymbol{v}_t$ exactly located at $(\tilde{x}_{ij}^t, \tilde{y}_{ij}^t)$ to directly get the pixel value. Hence, we utilize the four adjacent pixels of $(\tilde{x}_{ij}^t, \tilde{y}_{ij}^t)$ to obtain the pixel value using bilinear interpolation. We denote the four surrounding coordinates as $(\lfloor \tilde{x}_{ij}^t \rfloor, \lfloor \tilde{y}_{ij}^t \rfloor)$, $(\lfloor \tilde{x}_{ij}^t \rfloor + 1, \lfloor \tilde{y}_{ij}^t \rfloor)$, $(\lfloor \tilde{x}_{ij}^t \rfloor, \lfloor \tilde{y}_{ij}^t \rfloor + 1)$ and $(\lfloor \tilde{x}_{ij}^t \rfloor + 1, \lfloor \tilde{y}_{ij}^t \rfloor + 1)$, respectively, where $\lfloor \cdot \rfloor$ denotes the rounding-down operation. By assuming that the corresponding pixel values of these four pixels are $(m_{ij}^t)_{00}$, $(m_{ij}^t)_{01}$, $(m_{ij}^t)_{10}$, and $(m_{ij}^t)_{11}$, the pixel value at $(\tilde{x}_{ij}^t, \tilde{y}_{ij}^t)$ (referred to as $\tilde{m}_{ij}^t$) can be obtained via differentiable bilinear interpolation:

$$\begin{aligned}\tilde{m}_{ij}^t = &(m_{ij}^t)_{00}(\lfloor \tilde{x}_{ij}^t \rfloor - \tilde{x}_{ij}^t + 1)(\lfloor \tilde{y}_{ij}^t \rfloor - \tilde{y}_{ij}^t + 1) \\ &+ (m_{ij}^t)_{01}(\tilde{x}_{ij}^t - \lfloor \tilde{x}_{ij}^t \rfloor)(\lfloor \tilde{y}_{ij}^t \rfloor - \tilde{y}_{ij}^t + 1) \\ &+ (m_{ij}^t)_{10}(\lfloor \tilde{x}_{ij}^t \rfloor - \tilde{x}_{ij}^t + 1)(\tilde{y}_{ij}^t - \lfloor \tilde{y}_{ij}^t \rfloor) \\ &+ (m_{ij}^t)_{11}(\tilde{x}_{ij}^t - \lfloor \tilde{x}_{ij}^t \rfloor)(\tilde{y}_{ij}^t - \lfloor \tilde{y}_{ij}^t \rfloor).\end{aligned} \tag{3}$$

Consequently, we can obtain the image patch $\tilde{\boldsymbol{v}}_t$ by traversing all possible i, j in Eq. (3).

Assume we have the training loss $\mathcal{L}$, we can compute the gradient $\partial \mathcal{L} / \partial \tilde{m}_{ij}^t$ with standard back-propagation. Following the chain rule, we have

$$\frac{\partial \mathcal{L}}{\partial \tilde{x}_{\mathrm{c}}^t} = \sum_{i,j} \frac{\partial \mathcal{L}}{\partial \tilde{m}_{ij}^t} \frac{\partial \tilde{m}_{ij}^t}{\partial \tilde{x}_{\mathrm{c}}^t}, \quad \frac{\partial \mathcal{L}}{\partial \tilde{y}_{\mathrm{c}}^t} = \sum_{i,j} \frac{\partial \mathcal{L}}{\partial \tilde{m}_{ij}^t} \frac{\partial \tilde{m}_{ij}^t}{\partial \tilde{y}_{\mathrm{c}}^t}. \tag{4}$$

Combining Eq. (2) and Eq. (4), we can further derive

$$\frac{\partial \tilde{m}_{ij}^t}{\partial \tilde{x}_{\mathrm{c}}^t} = \frac{\partial \tilde{m}_{ij}^t}{\partial \tilde{x}_{ij}^t}, \quad \frac{\partial \tilde{m}_{ij}^t}{\partial \tilde{y}_{\mathrm{c}}^t} = \frac{\partial \tilde{m}_{ij}^t}{\partial \tilde{y}_{ij}^t}. \tag{5}$$

Given that $\tilde{x}_{\mathrm{c}}^t$ and $\tilde{y}_{\mathrm{c}}^t$ are the outputs of the network π, the back-propagation process is able to proceed in an ordinary way.

4. Qualitative Results

We present more qualitative results in image format in Fig. 1 and in video format. Our qualitative results show how the model is able to estimate the salient frames and prioritize them over the non-relevant ones, *e.g.* in (c) salient frames are the ones containing ice hockey-related actions and in

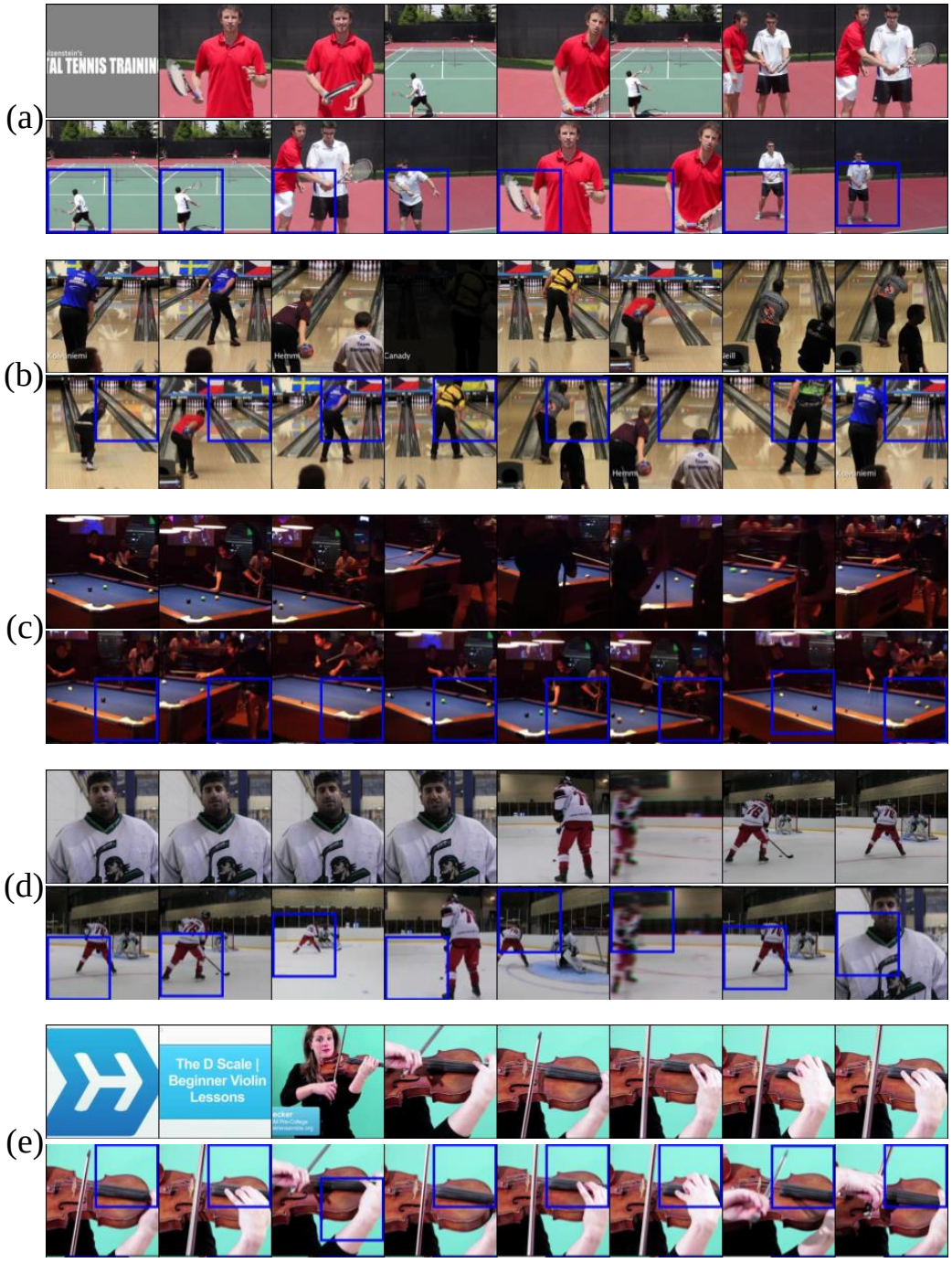

Figure 1. **Extended qualitative result** shows pair of the first 8 frames in original sequence and the Top-8 salient frames from classes (a) "`tennis serve`", (b) "`playing ten pins`", (c) "`playing pool`", (d) "`playing ice hockey`", and (e) "`playing violin`". We also provide qualitative results in video format to better comprehend the effect of the audio.

(c) and (a) frames with only text are non-salient. From the examples in video format, we observe how strong audio cues are present in the salient frames. For example, in "`playing ten pins`" class sample, the sound of the ball crashing the pins provide strong cues to estimate saliency.

References

[1] Mikhail Bortnikov, Adil Khan, Asad Masood Khattak, and Muhammad Ahmad. Accident recognition via 3d cnns for automated traffic monitoring in smart cities. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision*

Conference (CVC), Volume 2 1, pages 256–264. Springer, 2020.

[2] 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, Honolulu, HI, 2017. IEEE.

[3] Jiawei Chen and Chiu Man Ho. MM-ViT: Multi-Modal video transformer for compressed video action recognition. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Jan. 2022.

[4] Jun Chen, R Dinesh Jackson Samuel, and Parthasarathy Poovendran. LSTM with bio inspired algorithm for action recognition in sports videos. *Image Vis. Comput.*, 112:104214, Aug. 2021.

[5] Ishan Dave, Zacchaeus Scheffer, Akash Kumar, Sarah Shiraz, Yogesh Singh Rawat, and Mubarak Shah. Gabriellav2: Towards better generalization in surveillance videos for action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 122–132, 2022.

[6] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 1

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210, Seoul, Korea (South), 2019. IEEE.

[8] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2020*, pages 10457–10467, 2020.

[9] Rúben Geraldes, Artur Gonçalves, Tin Lai, Mathias Villerabel, Wenlong Deng, Ana Salta, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. UAV-Based situational awareness system using deep learning. *IEEE Access*, 7:122583–122594, 2019.

[10] Amir Ghodrati, Babak Ehteshami Bejnordi, and Amirhossein Habibian. FrameExit: Conditional early exiting for efficient video recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021. 1

[11] Shreyank N Gowda, Marcus Rohrbach, and Laura Sevilla-Lara. SMART frame selection for action recognition. *AAAI*, 35(2):1451–1459, May 2021.

[12] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[13] Yu-Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih-Fu Chang. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(2):352–364, Feb. 2018. 1

[14] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[15] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. EPIC-Fusion: Audio-Visual temporal binding for egocentric action recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5491–5500, Seoul, Korea (South), 2019. IEEE.

[16] Hanul Kim, Mihir Jain, Jun-Tae Lee, Sungrack Yun, and Fatih Porikli. Efficient action recognition via dynamic knowledge propagation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021.

[17] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. MoViNets: Mobile video networks for efficient video recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021.

[18] Bruno Korbar, Du Tran, and Lorenzo Torresani. SCSampler: Sampling salient clips from video for efficient action recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019.

[19] Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-Attentional Audio-Visual fusion for Weakly-Supervised action localization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[20] Sumin Lee, Sangmin Woo, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Modality mixer for multimodal action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3298–3307, 2023.

[21] Jintao Lin, Haodong Duan, Kai Chen, Dahua Lin, and Limin Wang. OCSampler: Compressing videos to one clip with single-step sampling. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022.

[22] 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. openaccess.thecvf.com, 2019.

[23] 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 *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11669–11676. AAAI Press, 2020.

[24] Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. AR-net: Adaptive frame resolution for efficient action recognition. In *ECCV*, Lecture notes in computer science, pages 86–104. Springer International Publishing, Cham, 2020. 1

[25] Yue Meng, Rameswar Panda, Chung-Ching Lin, Prasanna Sattigeri, Leonid Karlinsky, Kate Saenko, Aude Oliva, and Rogério Feris. AdaFuse: Adaptive temporal fusion network for efficient action recognition. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1

[26] Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Domain generalization through Audio-Visual relative norm alignment in first person action recognition. In *Proceedings of the*

IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1807–1818, 2022.

[27] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. AutoLoc: Weakly-supervised temporal action localization in untrimmed videos. In *Computer Vision – ECCV 2018*, Lecture notes in computer science, pages 162–179. Springer International Publishing, Cham, 2018.

[28] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015.

[29] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.

[30] Ximeng Sun, Rameswar Panda, Chun-Fu Richard Chen, Aude Oliva, Rogerio Feris, and Kate Saenko. Dynamic network quantization for efficient video inference. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021.

[31] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP, June 2022.

[32] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, June 2018.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv:1706.03762 [cs]*, Dec. 2017. 1

[34] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. UntrimmedNets for weakly supervised action recognition and detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.

[35] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(11):2740–2755, Nov. 2019.

[36] Yulin Wang, Zhaoxi Chen, Haojun Jiang, Shiji Song, Yizeng Han, and Gao Huang. Adaptive focus for efficient video recognition. In *ICCV*. IEEE, Oct. 2021.

[37] Yulin Wang, Yang Yue, Yuanze Lin, Haojun Jiang, Zihang Lai, Victor Kulikov, Nikita Orlov, Humphrey Shi, and Gao Huang. AdaFocus v2: End-to-End training of spatial dynamic networks for video recognition. In *CVPR*, pages 20062–20072, 2022. 1

[38] Yulin Wang, Yang Yue, Xinhong Xu, Ali Hassani, Victor Kulikov, Nikita Orlov, Shiji Song, Humphrey Shi, and Gao Huang. AdaFocusV3: On unified Spatial-Temporal dynamic video recognition. In *Computer Vision – ECCV 2022*, pages 226–243. Springer Nature Switzerland, 2022.

[39] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. *AAAI*, 37(3):2776–2784, June 2023.

[40] Wenhao Wu, Dongliang He, Tianwei Lin, Fu Li, Chuang Gan, and Errui Ding. MVFNet: Multi-View fusion network for efficient video recognition. *AAAI*, 35(4):2943–2951, May 2021.

[41] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6222–6231. IEEE, Oct. 2019.

[42] Wenhao Wu, Yuxiang Zhao, Yanwu Xu, Xiao Tan, Dongliang He, Zhikang Zou, Jin Ye, Yingying Li, Mingde Yao, Zichao Dong, and Yifeng Shi. DSANet: Dynamic segment aggregation network for Video-Level representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pages 1903–1911, New York, NY, USA, Oct. 2021. Association for Computing Machinery.

[43] Zuxuan Wu, Hengduo Li, Caiming Xiong, Yu-Gang Jiang, and Larry S Davis. A dynamic frame selection framework for fast video recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(4):1699–1711, Apr. 2022.

[44] Zuxuan Wu, Caiming Xiong, Yu-Gang Jiang, and Larry S Davis. LiteEval: A Coarse-to-Fine framework for resource efficient video recognition. In Hanna M Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7778–7787, 2019. 1

[45] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual SlowFast networks for video recognition. Technical Report arXiv:2001.08740, arXiv, Mar. 2020.

[46] Yeung, Russakovsky, Mori, and Fei-Fei. End-to-End learning of action detection from frame glimpses in videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 0, pages 2678–2687, June 2016.

[47] Yin-Dong Zheng, Zhaoyang Liu, Tong Lu, and Limin Wang. Dynamic sampling networks for efficient action recognition in videos. *IEEE Trans. Image Process.*, 29:7970–7983, 2020.

[48] Yuan Zhi, Zhan Tong, Limin Wang, and Gangshan Wu. MGSampler: An explainable sampling strategy for video action recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2021.