MULTI-SCALE COMPOSITIONAL CONSTRAINTS FOR REPRESENTATION LEARNING ON VIDEOS

Georgios Paraskevopoulos[†], *Chandrashekhar Lavania*[*], *Lovish Chum*[‡] , *Shiva Sundaram*[§]

AWS AI Labs

ABSTRACT

Combining simple concepts to form structured thoughts and decomposing complex concepts into their constituents is one key characteristic of human cognition. In this work we extract video representations by combining multi-scale processing with compositional constraints, i.e., we constrain the latent space created by the network so that coarse grained video features are composed from a set of fine-grained video features using simple functions. We integrate the proposed constraints in a state-of-the-art contrastive learning framework. In our ablations, we evaluate different formulations of the compositional constraints and composition functions. We evaluate the proposed approach for the downstream tasks of action detection in UCF-101, and video summarization in the SumMe dataset. We achieve significant improvements over the baseline, i.e., 3.9% and 6.3% relative improvements for UCF-101 and SumMe respectively, showcasing the importance of compositional video representations.

***Index Terms*—** Multimodal, Contrastive Learning, Audiovisual processing, Action Recognition, Compositionality

1. INTRODUCTION

Our perception of the world is multimodal. We acquire sensory inputs (e.g. sight, hearing) and combine them to help us understand and interact with our environment. Multimodal learning aims to create models that can leverage inputs from multiple sources, and combine them to create powerful representations for solving real-world tasks. Multimodal representations improve performance on several real world tasks, e.g. action recognition on videos [1, 2], sentiment analysis [3, 4], or speech recognition [5, 6]. In this work, our goal is to produce such useful multimodal representations for videos.

Structured thinking and the ability to formulate complex ideas from simpler ones is an important facet of human learning. For example, such thinking can be used for navigation through a neighborhood. The path to a place, say P, can be devised as route based on landmarks that are already known to the traveler e.g., pass the pharmacy, turn on the large tree on Oak St., and other similar options. Such compositional representation allows the traveler to efficiently memorize and devise new routes, give directions, and adapt to different situations (e.g. road closures, going by car or bike). In this example, the robustness of human learning yields from the ability to compose complex routes from salient segments.

Compositional learning can also be incorporated in training models that produce video representations. To this end, in this work we incorporate notions of multi-scale processing in a self-supervised instance discrimination setting, by introducing constraints that encourage compositional video representation learning across time-scales. Our notion of multi-scale compositionality focuses on composing the latent representations of the video clip from a set of fine-grained segment representations. An end to end training scheme is proposed, which integrates compositionality constraints in a state-of-the-art contrastive learning framework.

Our key contributions are: (I) We introduce end to end compositional learning for videos across time-scales, (II) we formulate constraints in a self-supervised setting and integrate them into a state-of-the-art contrastive framework [7], and (III) we present experiments on two different formulations of the proposed constraints, achieving state-of-the-art performance for action recognition on UCF-101 and unsupervised video summarization on SumMe when compared to models of similar scale in the literature.

2. RELATED WORK

Self-supervised multimodal learning: Recently, the self-supervised learning paradigm has opened avenues to building powerful data-driven models that yield state-of-the-art performance for challenging tasks in the fields of Natural Language [8], Image [9] and Speech Processing [10]. The power of self-supervised learning lies in obtaining the supervisory signals from data itself. Thus it can scale to large networks trained on massive amounts of data without the need of labeled data [11]. In the field of multimodal processing, and specifically video processing, self-supervised learning techniques range from Masked Modeling [12], to consistency and reconstruction losses [1], temporal re-ordering [13], clustering [14],

[†]National Technical University of Athens. Work done during an internship at AWS AI Labs. geopar@central.ntua.gr

[*]Correspondence to clavania@amazon.com

[‡]lchum@amazon.com

[§]sssundar@amazon.com

siamese networks [2], and contrastive learning [7, 15]. An interesting category of contrastive learning methods focus on instance discrimination [16], i.e., learning to distinguish clips of a video from other videos in the dataset. Qian et al. [17] propose spatiotemporal augmentations in the context of an instance discrimination framework. A co-training scheme is proposed in [18], where positive instances are mined using cross-modal similarity. AVID-CMA [7] uses a memory bank to store exponentially moving averaged features of video clips, and positive and negative instances are selected from the memory bank via cross-modal agreement. In this work, we focus on contrastive learning approaches on videos, without excluding possible future adaptations for other types of video representation frameworks.

Multi-scale processing: Multi-scale processing refers to the analysis of a source signal at different representation levels, by studying the signal at different time and/or space resolutions. Classical machine learning approaches that applied multi-scale processing in computer vision and speech primarily focused on the wavelet transformation [19] and scale-space theory [20]. Recently, multi-scale approaches have been proposed to improve performance of deep learning approaches in a variety of tasks. Wang et al. [21] propose a convolutional architecture to process hyper-spectral images at different dilation rates. Bian et al. [22] utilize coarse-grained and fine-grained image scales to produce a random walk transition matrix between image patches. Multi-scale transformers [23] extract hierarchical feature maps from images. They use pooling layers to modify the spatial scale of the image features at different stages (set of layers) of the architecture. In [24], a contrastive framework is proposed that utilizes global and local input representations. TS2vec [25] utilizes contrastive learning at successive pooling layers, obtaining hierarchical representations across time-scales. BraVe [1] is a self-supervised approach based on coarse to fine, and fine to coarse reconstruction of input signals.

Compositionality: The notion of compositionality in the context of neural networks is a broad term, encompassing a set of generalization properties in trained models [26]. We focus on one compositional property, i.e., localism, which explores how the representation of a complex notion can be derived as a function of its constituents, or simply how the whole can be composed from its parts. Compositionality has been explored in the context of language processing [27, 28] by introducing additional supervision signal to the attention maps of a sequence to sequence model. In the context of computer vision [29, 30] convolutional architectures have been proposed, that extract a global representation from salient objects of the image. These works highlight that compositionality can improve robustness, generalization and overall performance of trained networks.

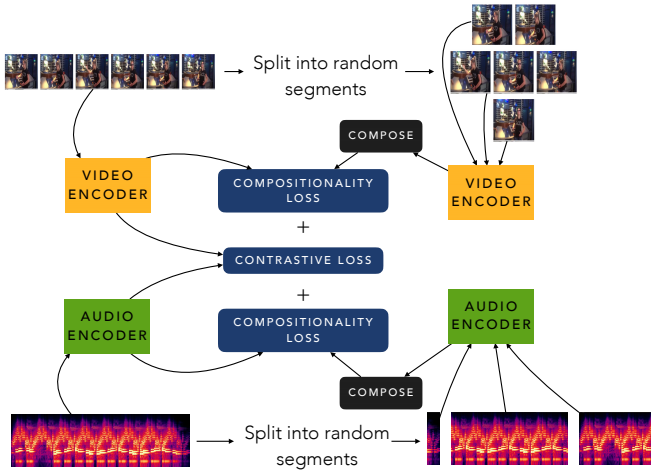

Fig. 1. Explicit Constraint (EC) framework with joint loss. Video encoders for the whole clip and the segmented clip are tied. Same for audio encoders.

3. MULTI-SCALE COMPOSITIONALITY

To define the notion of compositionality formally, let $\mathbf{h} = \phi(\mathbf{x}) \in \mathbb{R}^D$ be a latent vector of dimension D produced by the network ϕ given the input sequence $\mathbf{x}$, and $\tilde{\mathbf{h}}_i = \phi(\mathbf{x}_i) \in \mathbb{R}^D$ be a set of hidden states obtained using M arbitrary contiguous subsequences $\mathbf{x}_i$ of $\mathbf{x}$, where $1 \leq i \leq M$. We define compositionality by a composition function g as the constraint:

$$\mathbf{h} = g(\tilde{\mathbf{h}}), \tag{1}$$

where the i-th row of $\tilde{\mathbf{h}}$ is $\tilde{\mathbf{h}}_i$. Eq. (1) posits that the latent representations $\tilde{\mathbf{h}}_i$ of the segments can be composed to produce $\mathbf{h}$. In our notion of compositionality, g is preferably a *simple* function, i.e. to have few (if any) trainable parameters and to be a shallow mapping. The following composition functions are explored:

$$g_1(\tilde{\mathbf{h}}) = \sum_i \tilde{\mathbf{h}}_i \tag{2a}$$

$$g_2(\tilde{\mathbf{h}})_j = \begin{cases} \max_i \tilde{h}_{ij}, & |\max_i \tilde{h}_{ij}| \geq |\min_i \tilde{h}_{ij}| \\ \min_i \tilde{h}_{ij}, & \text{otherwise} \end{cases} \tag{2b}$$

$$g_3(\tilde{\mathbf{h}}) = \text{MHA}(\tilde{\mathbf{h}}), \tag{2c}$$

where $g_1(\tilde{\mathbf{h}}), g_2(\tilde{\mathbf{h}}), g_3(\tilde{\mathbf{h}}) \in \mathbb{R}^D$. g_1 is the sum of M segment representations, g_2 (absmax) produces a vector, where the j-th element ($1 \leq j \leq D$) is the j-th activation with the maximum absolute value across the M segments (maintaining the sign) and g_3 is the Multi-Head Attention (MHA) operation.

For this notion of compositionality, two input streams are needed per modality, the whole video clip and a time-segmented sequence of the input visual frames or spectrograms extracted from the audio stream. Let $s \in \mathbb{R}^{N \times C \times H \times W}$

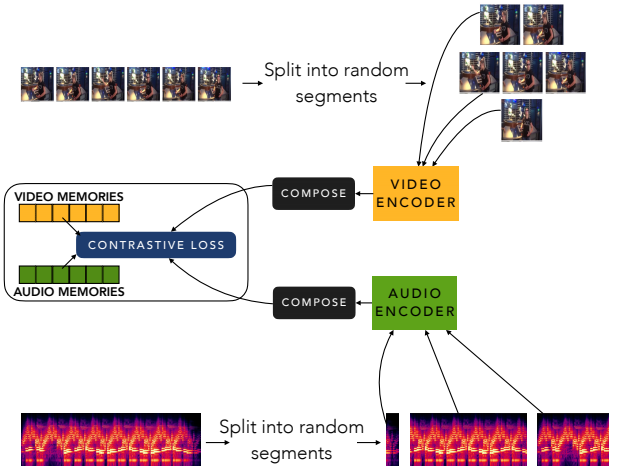

Fig. 2. Implicit Constraint (IC) framework.

be the input sequence of frames for the whole video clip, where N is the duration (number of frames), C the number of channels, H the frame height and W the frame width. The sequence s is split into M segments of durations $N_1, N_2, \ldots, N_M$, yielding a tensor of fine-grained clips $\tilde{\mathbf{s}}_i \in \mathbb{R}^{N_i \times C \times H \times W}, i \in \{1, 2, \ldots, M\}$. Durations N_i and the number of segments M are sampled from a uniform discrete distribution $\mathcal{U}\{1, N\}$ under the condition that $\sum_i N_i = N$ (no overlap between segments). The same procedure is followed for time-segmentation of spectrograms for audio.

3.1. Explicit Constraint

In Fig. 1 we see the Explicit Constraint (EC) framework, where compositionality is encouraged by introducing an explicit constraint as a joint loss. The frames and spectrograms for the whole video clip are passed through the video and audio encoders respectively to obtain the D-dimensional hidden representations $\mathbf{v}$ and $\mathbf{a}$. Similarly, the segments of frames and spectrograms are passed through the video and audio encoders, yielding the hidden representations $\tilde{\mathbf{v}}$ and $\tilde{\mathbf{a}}$. The compositionality loss L_c aims to minimized the distance of the coarse-grained representations $\mathbf{v}, \mathbf{a}$ and the composed fine-grained representations $g(\tilde{\mathbf{v}}), g(\tilde{\mathbf{a}})$, given the composition function g. For a vector $\mathbf{h}$ and the subsequences $\tilde{\mathbf{h}}$, the compositionality loss L_c is given in Eq. (3):

$$L_c(\mathbf{h}, \tilde{\mathbf{h}}) = \frac{1}{D} \sum_{j}^{D} (h_j - g(\tilde{\mathbf{h}})_j)^2 \tag{3}$$

L_c is the mean-squared error loss, and it encourages $\mathbf{h}$ to be equal to $g(\tilde{\mathbf{h}})$. The total loss is given by combining the self-supervised loss L_{ss} with the compositionality losses L_c for the audio and video modalities as shown in Eq. (4). In our experiments, the AVID-CMA loss [7] is used as L_{ss}.

$$L = L_{ss}(\mathbf{v}, \mathbf{a}) + \lambda(L_c(\mathbf{v}, \tilde{\mathbf{v}}) + L_c(\mathbf{a}, \tilde{\mathbf{a}})) \tag{4}$$

3.2. Implicit Constraint

One issue with the EC framework is that the constraint is added explicitly as an additional loss, which raises the need for careful tuning of the weight λ. This issue can be alleviated by directly integrating the compositionality as an implicit constraint in the contrastive loss. Fig. 2 depicts the Implicit Constraint (IC) framework. For this, we take advantage of the AVID-CMA framework [7]. The AVID-CMA loss performs instance discrimination by building a large memory bank that contains learned features for each sample in the pre-training dataset. The positive and negative samples are sampled from the memory bank, and then used for Noise Contrastive Estimation [31, 32]. The goal of the AVID-CMA loss is to bring the latent representations of a sample closer to stored features of samples that have high audio and video similarity with the original sample. The input frame and spectrogram sequences are split into a set of random segments and compose the encoded segments to obtain $g(\tilde{\mathbf{v}})$ and $g(\tilde{\mathbf{a}})$. The composed features are passed into the AVID-CMA loss, to be compared with the coarse-grained features stored in the memory bank. The resulting loss is:

$$L = L_{ss}(g(\tilde{\mathbf{v}}), g(\tilde{\mathbf{a}})) \tag{5}$$

4. EXPERIMENTAL SETUP

The proposed constraints are built on top of the AVID-CMA framework [7]. The starting point is an AVID-CMA checkpoint, provided by the original authors, pretrained on Audioset [33]. Training continues using the EC or the IC framework for 5 epochs using Adam optimizer with learning rate 10^{-4}, weight decay 10^{-5} and batch size 64. Hyper-parameter λ is 0.1 for the EC framework. The balanced split of Audioset, containing $\sim$18k videos is used for this step. Inputs are video clips of 1 second sampled at 16 fps and audio clips of 2 seconds sampled at 24 kHz. Video frames are resized to 224×224 pixels and random cropping, horizontal flip and color jitter augmentations are applied. Audio clips are downmixed to single channel, and then we extract spectrograms of size 200×257 (200 timesteps and 257 frequency bands). The video encoder is R(2+1)D-18, an 18-layer convolutional architecture [34], while the audio encoder is a 9-layer 2D convolutional network with batch normalization. The encoder outputs are max-pooled and projected into 128D feature vectors using a 3-layer fully connected network with 512 hidden states. L2 normalization is applied to the 128D feature vectors.

We evaluate our model on UCF-101 [35] for action recognition, which contains 13320 clips collected from YouTube. Clips are labeled across 101 diverse action categories (e.g.

Compositional Constraint	g	Acc@1
Implicit (IC)	sum	94.26
Implicit (IC)	absmax	**94.81**
Implicit (IC)	MHA	93.48
Explicit (EC)	sum	93.51
Explicit (EC)	absmax	94.47
Explicit (EC)	MHA	**95.06**
Baseline (AVID-CMA) [7]		91.5

Table 1. Action recognition ablations for the IC and the EC frameworks with different composition functions in Eq. 2 (Eq. 2a: sum, Eq. 2b: absmax, Eq. 2c: MHA).

Kayaking, Knitting). Three folds are provided for evaluation, with different train-test splits. For our evaluation, we attach a 1-layer, fully connected classification head on top of the video encoder and follow the full network finetuning protocol. Videos with 32 frames (2 seconds) are used. Predictions for 10 uniformly sampled snippets are extracted for each video, and then averaged to obtain video-level predictions, as in [7]. We report the top-1 classification accuracy, averaged over the 3 predefined folds.

The SumMe dataset [36] is used for unsupervised video summarization, which contains 25 videos with egocentric, static and moving cameras. Each video is accompanied by a set of user provided summaries. We use the framework of Zhou et al. [37] (unsupervised video summarization), by providing the features extracted by the video and / or the audio encoder using the absmax composition function . The common evaluation protocol is followed and the F-score metric is used for evaluation of the extracted summaries.

Model	Acc@1
AVID-CMA [7]	91.5
GDT [15]	92.5
XDC [14]	93.0
CrissCross [2]	92.4
BraVe $V \leftrightarrow A$ [1]	93.6
Ours (IC+absmax)	**94.81**
Ours (EC+MHA)	**95.06**

Table 2. Comparison with the state-of-the-art for action recognition on UCF-101

5. EXPERIMENTS & RESULTS

In Table 1 we present an ablation study on the effect of different composition functions and the loss type for action recognition on UCF-101. First, observe that including the compositionality constraints yields a significant improvement over the baseline with all configurations. The best results are highlighted with bold for the EC and the IC frameworks. For the EC framework, we observe that the best composition function is Multi-Head Attention (MHA), while for the IC framework, absmax function yields the best results. Interestingly, the MHA composition function yields the worst results for the IC framework. This indicates that, when the composition function induces additional trainable parameters, the EC framework is the best choice, while if a parameter-free composition function is used, the IC framework yields better results. One explanation for this is that, during back-propagation in the EC framework, the parameters of the composition function are trained only based on the L_c loss, limiting the catastrophic forgetting effect they could induce in the IC framework.

Model	F-score
Reinforce [37]	41.4
XDC [14]	41.4
AVID-CMA [7]	42.2
Cognizance [38] (visual)	42.7
Cognizance [38] (audio+visual)	43.5
Ours (IC+absmax) (visual)	**44.86**
Ours (IC+absmax) (audio+visual)	**46.65**

Table 3. Video summarization on the SumMe dataset

In Table 2 a comparison is presented with state-of-the-art self-supervised action recognition approaches in the literature. For fair comparison, we focus on approaches that use the same video encoder architecture, i.e., R(2+1)D-18, and the same pretraining dataset, i.e., Audioset. Both proposed IC+absmax and the EC+MHA frameworks yield significant improvements over the state-of-the-art. Finally, in Table 3 we present a comparison on unsupervised video summarization results using the features extracted from the proposed IC+absmax setting on the SumMe dataset. Incorporation of the proposed compositionality constraints outperforms other methods in the literature. The audio modality contributes further in improving the F-score.

6. CONCLUSIONS & FUTURE WORK

We explore constraints that can encourage the notion of multi-scale compositionality in video architectures trained using self-supervised learning. Two variants of these constraints are proposed, by introducing compositionality as a joint MSE loss (EC), or integrating the constraint implicitly in an instance discrimination contrastive learning framework (IC). Furthermore, we explore the effect of choosing different composition functions. Our ablations show that the EC framework is the best choice when the composition function introduces trainable parameters, while the IC framework is better for non-learnable composition functions. Our experiments show significant improvement for the action recognition and video summarization downstream tasks. In the future our approach can be enriched by exploring cross-modal compositionality. Furthermore, heuristics and fine-grained annotation can be used for video segmentation.

7. REFERENCES

[1] Recasens, A. et al., "Broaden your views for self-supervised video learning," in *Proc. CVPR*. IEEE/CVF, 2021, pp. 1255–1265.

[2] P. Sarkar and A. Etemad, "Self-supervised audio-visual representation learning with relaxed cross-modal temporal synchronicity," *CoRR arXiv:2111.05329*, 2021.

[3] G. Paraskevopoulos, E. Georgiou, and A. Potamianos, "Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis," in *Proc. ICASSP*. IEEE, 2022, pp. 4573–4577.

[4] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. ACMM*, 2020, pp. 1122–1131.

[5] Ramon S. et al., "How2: A large-scale dataset for multimodal language understanding," *In Proc. ViGIL workshop, NeurIps*, 2018.

[6] G. Paraskevopoulos, S. Parthasarathy, A. Khare, and S. Sundaram, "Multimodal and multiresolution speech recognition with transformers," in *Proc. 58th ACL*, 2020, pp. 2381–2387.

[7] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-visual instance discrimination with cross-modal agreement," in *Proc. CVPR*. IEEE/CVF, 2021, pp. 12475–12486.

[8] J. Devlin, MW. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, vol. 1, pp. 4171–4186.

[9] Chen, M. et al., "Generative pretraining from pixels," in *Proc. ICML*. PMLR, 2020, pp. 1691–1703.

[10] Hsu, WN. et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *Transactions on Audio, Speech and Language*, vol. 29, pp. 3451–3460, 2021.

[11] Han, X. et al., "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021.

[12] Sun, C. et al., "Videobert: A joint model for video and language representation learning," in *Proc. CVPR*. IEEE/CVF, 2019, pp. 7464–7473.

[13] Xu, D. et al., "Self-supervised spatiotemporal learning via video clip order prediction," in *Proc. CVPR*. IEEE/CVF, 2019, pp. 10334–10343.

[14] Alwassel, H. et al., "Self-supervised learning by cross-modal audio-video clustering," *Proc. NeurIps*, vol. 33, pp. 9758–9770, 2020.

[15] Patrick, M. et al., "Multi-modal self-supervision from generalized data transformations," *CoRR arXiv:2003.04298*, 2020.

[16] Z. Wu, Y. Xiong, SX. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. CVPR*. IEEE/CVF, 2018, pp. 3733–3742.

[17] Qian, R. et al., "Spatiotemporal contrastive video representation learning," in *Proc. CVPR*. IEEE/CVF, 2021, pp. 6964–6974.

[18] T. Han, W. Xie, and A. Zisserman, "Self-supervised co-training for video representation learning," *Proc. NeurIps*, vol. 33, pp. 5679–5690, 2020.

[19] G. Strang and T. Nguyen, *Wavelets and filter banks*, SIAM, 1996.

[20] T. Lindeberg, *Scale-space theory in computer vision*, vol. 256, Springer, 2013.

[21] Wang, X. et al., "A unified multiscale learning framework for hyperspectral image classification," *Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[22] Z. Bian, A. Jabri, AA Efros, and A. Owens, "Learning pixel trajectories with multiscale contrastive random walks," in *Proc. CVPR*. IEEE/CVF, 2022, pp. 6508–6519.

[23] Fan, H. et al., "Multiscale vision transformers," in *Proc. CVPR*. IEEE/CVF, 2021, pp. 6824–6835.

[24] Zeng, Z. et al., "Contrastive learning of global and local video representations," *Proc. NeurIps*, vol. 34, pp. 7025–7040, 2021.

[25] Yue, Z. et al., "Ts2vec: Towards universal representation of time series," in *Proc. AAAI*, 2022, vol. 36, pp. 8980–8987.

[26] D. Hupkes, V. Dankers, M. Mul, and E. Bruni, "Compositionality decomposed: How do neural networks generalise?," *JAIR*, vol. 67, pp. 757–795, 2020.

[27] Hupkes, D. et al., "Learning compositionally through attentive guidance," *CoRR arXiv:2003.04298*, 2018.

[28] Baan, J. et al., "On the realization of compositionality in neural networks," in *Proc. BlackboxNLP Workshop, ACL*, 2019, pp. 127–137.

[29] Kortylewski, A. et al., "Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion," *Int. J. Comput. Vision*, vol. 129, no. 3, pp. 736–760, 2021.

[30] Stone, A. et al., "Teaching compositionality to cnns," in *Proc. CVPR*. IEEE/CVF, 2017, pp. 5058–5067.

[31] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR arXiv:1807.03748*, 2018.

[32] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proc. 13th AISTATS*, 2010, pp. 297–304.

[33] Gemmeke, JF et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*. IEEE, 2017, pp. 776–780.

[34] Tran, D. et al., "A closer look at spatiotemporal convolutions for action recognition," in *Proc. CVPR*. IEEE/CVF, 2018, pp. 6450–6459.

[35] K. Soomro, AR. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *CoRR arXiv:1212.0402*, 2012.

[36] M. Gygli, H. Grabner, H. Riemenschneider, and LV. Gool, "Creating summaries from user videos," in *Proc. ECCV*. Springer, 2014, pp. 505–520.

[37] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proc. AAAI*, 2018, vol. 32.

[38] C. Lavania, S. Sundaram, S. Srinivasan, and K. Kirchhoff, "Enhancing contrastive learning with temporal cognizance for audio-visual representation generation," in *Proc. ICASSP*, 2022, pp. 4728–4732.