Self-Supervised Representation Learning from Videos of Audible Interactions

Himangi Mittal[1], Pedro Morgado[1], Unnat Jain[2], Abhinav Gupta[1]

[1] Carnegie Mellon University
[2] Meta AI Research
hmittal@andrew.cmu.edu, pmorgado@wisc.edu, unnatjain@gmail.com, abhinavg@cs.cmu.edu

Abstract. We propose a self-supervised algorithm to learn representations from egocentric video data. Given the uncurated nature of long-form continuous videos, learning effective representations require focusing on moments in time when interactions take place. To achieve this, we leverage audio signals to identify moments of likely interactions and also propose a novel self-supervised objective that learns from audible state changes caused by interactions. We validate these contributions on two large-scale egocentric datasets, EPIC-Kitchens-100 and Ego4D, and show improvements on downstream task of action recognition.

1 Introduction

Recent successes in self-supervised learning (SSL) [22, 4, 15, 12] has brought into question the need for human annotations. Current approaches learn from static images which lack temporal information and are unable to learn about state changes. Here, videos can be helpful to learn rich representations in self-supervised manner [26, 14, 16, 27, 34, 13, 9]. However, learning representations from videos can be quite challenging. The first challenge is choosing the right SSL loss. Approaches using only video modality [33, 25] or both audio-video modality [1, 20, 27, 23, 19] have attempted to learn representations but are not sensitive to state changes. The second challenge is that current video-based SSL approaches exploit single-action, curated video datasets, such as Kinetics [3]. This is in contrast to the predominantly *untrimmed*, real-world, egocentric data [5, 6, 24, 18, 30, 8, 11] which contain multiple, fine-grained actions [35, 29, 7, 17, 21]. Unlike action-centric datasets, learning from untrimmed videos is challenging as they contain long periods without interactions, which aren't useful for training.

Interactions like opening a fridge or placing a pan on stove create clear and consistent sound signatures due to the physical interaction between objects. To capture these right moments when actual interactions occur, audio can be used to train on interaction-rich portions of the untrimmed videos. Prior work on audio-visual correspondence (AVC) [28, 2, 20] still favors invariance which is not informative of the changes happening over time. To this end, we introduce RepLAI – **Rep**resentation **L**earning from **A**udible **I**nteractions, a self-supervised algorithm for representation learning from videos of audible interactions. RepLAI uses the audio signals to: (1) identify moments in time that are conducive to

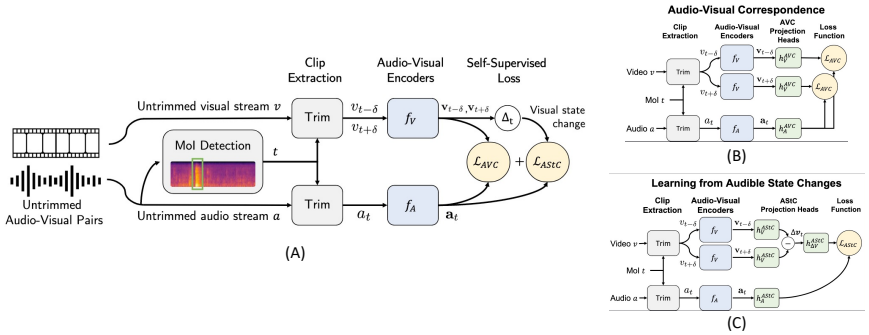

Fig. 1: Overview of RepLAI. RepLAI (A) seeks to learn audio and visual encoders (f_A and f_V) by detecting and training on moments of interaction (MoI) present in untrimmed videos and solving two tasks: audio-visual correspondence `AVC` (B) and audio identifiable state changes `AStC` (C).

self-supervised learning and (2) to learn representations that focus on the visual state changes caused by audible interactions.

2 RepLAI

Given a dataset $\mathcal{D} = \{(v_i, a_i)_{i=1}^N\}$ containing N long (untrimmed) audio-visual streams, our goal is to learn visual and audio encoders, denoted f_V and f_A. An overview of the proposed approach is depicted in Fig. 1. For each sample $(v, a) \in \mathcal{D}$, we search for moments of interaction (MoI) using the audio stream, and extract short audio and visual clips around these MoI. These trimmed clips are then encoded into a vectorized representation using f_V and f_A. The whole system is trained to optimize two self-supervised losses – audio-visual correspondence $\mathcal{L}_{\texttt{AVC}}$, and a novel objective that learns from audible state changes $\mathcal{L}_{\texttt{AStC}}$.

Audio-driven detection of moments of interaction. Audio signals are particularly informative of moments of interaction. To complete day-to-day activities, we physically interact with objects in our environments. These interactions typically produce distinct audio patterns - short bursts of energy that span all frequencies. To locate these patterns, we represent the audio as a log mel spectrogram where moments of interaction appear as *vertical edges*, and search for robust local maxima in the total energy (summed over all frequencies). To avoid weak local maxima, we ignore energy peaks with small prominence (lower than 1)[3]. We also ignore low prominence peaks found less than 50ms apart from other ones. Once detected, short clips around the moments of interaction are collected into a dataset $\mathcal{D}_{\texttt{MoI}}$, and used for training.

Learning from audible state changes Physical interactions often cause both state changes in the environment and distinct audio signals. To leverage this natural co-occurrence, we propose a self-supervised task that seeks to *associate the audio with changes in the visual state* during a moment of interaction. Specifically, the audible state change (`AStC`) loss seeks to (1) *increase* the probability of

[3] Peak prominence is defined as the size of the value range in a small window around the peak.

associating the audio with the visual change in the *forward* (*i.e.* correct) direction, (2) *decrease* the probability of associating the audio with the visual change in the *backward* (*i.e.* incorrect) direction

$$\mathcal{L}_{\texttt{AStC}} = \mathbb{E}_{v_t, a_t \in \mathcal{D}_{\texttt{MoI}}} \left[-\log\left(p^{\text{frwd}}(v_t, a_t)\right) - \log\left(1 - p^{\text{bkwd}}(v_t, a_t)\right) \right]. \quad (1)$$

The probabilities (p^{frwd}, p^{bkwd}) are computed from cross-modal similarities

$$p^{\text{frwd}}(v_t, a_t) = \sigma\left(\texttt{sim}\left(\Delta\mathbf{v}_t^{\text{frwd}}, \mathbf{a}_t\right)/\tau\right); \quad p^{\text{bkwd}}(v_t, a_t) = \sigma\left(\texttt{sim}\left(\Delta\mathbf{v}_t^{\text{bkwd}}, \mathbf{a}_t\right)/\tau\right), \quad (2)$$

where $\tau = 0.2$ is a temperature hyper-parameter, and σ denotes the sigmoid function. Audio representations ($\mathbf{a}_t$) are obtained by encoding the trimmed audio clips a_t via the audio encoder f_A (shared across all objectives) and an MLP projection $h_A^{\texttt{AStC}}$. State change representations ($\Delta\mathbf{v}_t^{\text{frwd}}$, $\Delta\mathbf{v}_t^{\text{bkwd}}$) are computed by considering two non-overlapping visual clips for each moment t at timestamps $t-\delta$ and $t+\delta$. Each clip is encoded via the visual encoder f_V and a projection MLP head $h_V^{\texttt{AStC}}$. Forward and backward state changes are then represented as

$$\Delta\mathbf{v}_t^{\text{frwd}} = h_V^{\texttt{AStC}} \circ f_V(v_{t+\delta}) - h_V^{\texttt{AStC}} \circ f_V(v_{t-\delta}); \quad \Delta\mathbf{v}_t^{\text{bkwd}} = h_V^{\texttt{AStC}} \circ f_V(v_{t-\delta}) - h_V^{\texttt{AStC}} \circ f_V(v_{t+\delta}). \quad (3)$$

Optimizing the loss of Eq. 1 requires the audio representation $\mathbf{a}_t$ to be aligned with the visual change $\Delta\mathbf{v}_t^{\text{frwd}}$ that took place, while different from the hypothetical backward state change $\Delta\mathbf{v}_t^{\text{bkwd}}$.

Learning from audio-visual correspondences *[28, 2, 20]* Audio-visual correspondence (`AVC`) seeks to align corresponding visual and audio clips into a common feature space. In particular, following [20, 31], audio-visual correspondence is established by minimizing a symmetric cross-modal InfoMax loss

$$\mathcal{L}_{\texttt{AVC}} = \mathbb{E}_{v_i, a_i \sim \mathcal{D}} \left[-\log \frac{e^{\texttt{sim}(\mathbf{v}_i, \mathbf{a}_i)/\tau}}{\sum_j e^{\texttt{sim}(\mathbf{v}_i, \mathbf{a}_j)/\tau}} - \log \frac{e^{\texttt{sim}(\mathbf{v}_i, \mathbf{a}_i)/\tau}}{\sum_j e^{\texttt{sim}(\mathbf{v}_j, \mathbf{a}_i)/\tau}} \right], \quad (4)$$

where $\tau = 0.07$ is a temperature hyper-parameter and `sim` the cosine similarity.

For readability, we absorbed the audio and visual projection MLP heads ($h_A^{\texttt{AVC}}$ and $h_V^{\texttt{AVC}}$) within $\texttt{sim}(\cdot, \cdot)$, and illustrate their usage in Fig. 1.

Training: Representations learned through `AVC` are biased towards static concepts while those learned through `AStC` are more sensitive to dynamic concepts. Since both are useful, representations are trained with $\mathcal{L} = \mathcal{L}_{\texttt{AVC}} + \mathcal{L}_{\texttt{AStC}}$.

3 Experiments

Experimental setup. Following prior work [20], we use an R(2+1)D video encoder [32] with depth 18 and a 10-layer 2D CNN as the audio encoder. Two video clips of duration 0.5s are extracted around MoI at 16 FPS, and separated by a 0.2s gap. We extract audio clips of 2s at 44.1kHz, downsample them to 16kHz and convert the mono signal to a log mel spectrogram with 80 frequency bands and 128 temporal frames. We train and evaluate on EPIC-Kitchens-100 [6] and Ego4D [11] using the Forecasting and Hand-Object interaction subset. Models are trained with stochastic gradient descent for 100 epochs with a batch size of

Method	$\mathcal{L}_{\texttt{AVC}}$	$\mathcal{L}_{\texttt{AStC}}$	MoI Sampling	AVID Init [20]	EPIC-Kitchens-100 Top1 Acc ↑ Verb	EPIC-Kitchens-100 Top1 Acc ↑ Noun	EPIC-Kitchens-100 Top5 Acc ↑ Verb	EPIC-Kitchens-100 Top5 Acc ↑ Noun	Ego4d Top1 Acc ↑ Verb	Ego4d Top1 Acc ↑ Noun
(1) Random					20.38	4.96	64.75	19.83	17.4	7.7
(2) AVID [20]				✓	26.62	9.00	69.79	25.50	18.3	10.7
(3) RepLAI w/o AVC		✓	✓	✓	29.92	10.46	70.58	29.00	20.3	12.4
(4) RepLAI w/o AStC	✓		✓	✓	29.29	9.67	73.33	29.54	21.1	13.5
(5) RepLAI w/o MoI	✓	✓		✓	28.71	8.33	73.17	27.29	19.8	11.2
(6) RepLAI (scratch)	✓	✓	✓		25.75	8.12	71.25	27.29	22.2	14.1
(7) RepLAI	✓	✓	✓	✓	**31.71**	**11.25**	**73.54**	**30.54**	**22.5**	**14.7**

Table 1: Action recognition on EPIC-Kitchens-100 and Ego4D.

128, a learning rate of 0.005 and a momentum of 0.9. For Ego4D, we use a batch size of 512 with a learning rate of 0.05. After self-supervised pre-training, the models are evaluated on action recognition (verb and noun) by training a linear classifier on a small annotated dataset.

Baselines and ablations. We consider various baselines as well as ablated versions of RepLAI. *Random* represents an untrained (randomly initialized) model, *AVID* [20] is a model pre-trained on 2M audio-visual pairs from AudioSet [10] with only audio-visual correspondence. The full method *RepLAI* is initialized by AVID weights. Other ablations include our method trained without AVID initialization (*RepLAI from scratch*), trained with only AVC (*RepLAI w/o AStC*), only state change loss (*RepLAI w/o AVC*), and trained on random moments in time (*RepLAI w/o MoI*).

Discussion of Results

RepLAI enhances large-scale AVC driven approaches. Comparing *AVID*, *RepLAI (Scratch)* and *RepLAI* in Tab. 1, it is clear that RepLAI enhances large-scale AVID pre-training by significant margins on all the downstream tasks. We also see that pre-training can be useful when training on the smaller EPIC-Kitchens-100, but less so on Ego-4D as the latter already contains a large diversity of scenes.

Detecting moments of interaction (MoI) helps representation learning. To establish the benefits of our audio-driven MoI detection algorithm, we compared *RepLAI* with *RepLAI w/o MoI*. As can be seen in Tab. 1, sampling clips around MoI leads to significantly better representations.

AVC and AStC are complementary. Comparing rows (3) and (4) to the full model in row (7) of Tab. 1 shows that both terms, AVC and AStC, are complementary to each other, as AVC focuses on learning visual representations of the sounding object present in the video, while AStC helps the model to differentiate between visual representations of before and after state change interactions.

4 Conclusion

We propose an audio-driven self-supervised method for learning representations of egocentric videos. By learning to focus on moments of interaction (MoI), strong representations can be learned for untrimmed datasets. Moreover, by learning to focus on the changes in the state of an environment caused by agents interacting with the world, state-aware representations can be learned which are particularly useful for egocentric downstream tasks.

References

1. Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., Tran, D.: Self-supervised learning by cross-modal audio-video clustering. Advances in Neural Information Processing Systems **33**, 9758–9770 (2020)
2. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 609–617 (2017)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the International Conference on Machine Learning. pp. 1597–1607. PMLR (2020)
5. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: European Conference on Computer Vision (ECCV) (2018)
6. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision. arXiv preprint arXiv:2006.13256 (2020)
7. Doughty, H., Snoek, C.G.: How do you do it? fine-grained action understanding with pseudo-adverbs. arXiv preprint arXiv:2203.12344 (2022)
8. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: A first-person perspective. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1226–1233. IEEE (2012)
9. Feichtenhofer, C., Fan, H., Xiong, B., Girshick, R., He, K.: A large-scale study on unsupervised spatiotemporal representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3299–3309 (2021)
10. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE (2017)
11. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR (2022)
12. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Pires, B., Guo, Z., Azar, M., et al.: Bootstrap your own latent: A new approach to self-supervised learning. In: Advances in Neural Information Processing Systems (2020)
13. Han, T., Xie, W., Zisserman, A.: Memory-augmented dense predictive coding for video representation learning. In: European conference on computer vision. pp. 312–329. Springer (2020)
14. Han, T., Xie, W., Zisserman, A.: Self-supervised co-training for video representation learning. Advances in Neural Information Processing Systems **33**, 5679–5690 (2020)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
16. Hu, K., Shao, J., Liu, Y., Raj, B., Savvides, M., Shen, Z.: Contrast and order representations for video self-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7939–7949 (2021)

17. Kazakos, E., Huh, J., Nagrani, A., Zisserman, A., Damen, D.: With a little help from my temporal context: Multimodal egocentric action recognition. arXiv preprint arXiv:2111.01024 (2021)
18. Lai, K., Bo, L., Fox, D.: Unsupervised feature learning for 3d scene labeling. In: 2014 IEEE International Conference on Robotics and Automation (ICRA). pp. 3050–3057. IEEE (2014)
19. Morgado, P., Misra, I., Vasconcelos, N.: Robust audio-visual instance discrimination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12934–12945 (2021)
20. Morgado, P., Vasconcelos, N., Misra, I.: Audio-visual instance discrimination with cross-modal agreement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12475–12486 (2021)
21. Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 122–132 (2020)
22. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
23. Patrick, M., Asano, Y.M., Kuznetsova, P., Fong, R., Henriques, J.F., Zweig, G., Vedaldi, A.: Multi-modal self-supervision from generalized data transformations. arXiv preprint arXiv:2003.04298 (2020)
24. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 2847–2854. IEEE (2012)
25. Purushwalkam, S., Ye, T., Gupta, S., Gupta, A.: Aligning videos in space and time. arXiv preprint arXiv:2007.04515 (2020)
26. Qian, R., Meng, T., Gong, B., Yang, M.H., Wang, H., Belongie, S., Cui, Y.: Spatiotemporal contrastive video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6964–6974 (2021)
27. Recasens, A., Luc, P., Alayrac, J.B., Wang, L., Strub, F., Tallec, C., Malinowski, M., Pătrăucean, V., Altché, F., Valko, M., et al.: Broaden your views for self-supervised video learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1255–1265 (2021)
28. de Sa, V.R.: Learning classification with unlabeled data. In: Advances in Neural Information Processing Systems. pp. 112–119. Citeseer (1994)
29. Singh, B., Marks, T.K., Jones, M., Tuzel, O., Shao, M.: A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1961–1970 (2016)
30. Su, Y.C., Grauman, K.: Detecting engagement in egocentric video. In: European Conference on Computer Vision. pp. 454–471. Springer (2016)
31. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: Proceedings of the European Conference on Computer Vision. pp. 776–794. Springer (2020)
32. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
33. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2794–2802 (2015)
34. Zeng, Z., McDuff, D., Song, Y., et al.: Contrastive learning of global and local video representations. Advances in Neural Information Processing Systems **34**, 7025–7040 (2021)

35. Zhang, C., Gupta, A., Zisserman, A.: Temporal query networks for fine-grained video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4486–4496 (2021)