

Ishan Rajendrakumar Dave

1995	Born in Rajkot, India
2013-2017	B.Tech, S.V. National Institute of Technology, Surat, India
2017-2018	Research Assistant, Indian Institute of Technology, Indore
2018-2019	Project Assistant, Indian Institute of Science, Bangalore
2022-2023	Research Scientist Intern, Adobe Research, San Jose, CA
2024	AI/ ML Intern, Apple Inc., Cupertino, CA
2019-2024	Ph.D., University of Central Florida

UCF | Center for Research in Computer Vision

UNIVERSITY OF CENTRAL FLORIDA

FINAL ORAL EXAMINATION

OF

Ishan Rajendrakumar Dave

B.Tech, Sardar Vallabhbhai National Institute of Technology, India, 2017

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

(COMPUTER SCIENCE)

30 October 2024, 1 P.M.
HEC 101

DISSERTATION COMMITTEE

Professor Mubarak Shah, *Chair*, *shah@crcv.ucf.edu*
Professor Chen Chen, *chen.chen@crcv.ucf.edu*
Professor Ser Nam Lim, *sernam@ucf.edu*
Professor Sean Pang, *pang@ucf.edu*

DISSERTATION RESEARCH IMPACT

Video understanding involves tasks such as action recognition, video retrieval, and human pose propagation, which are crucial for applications like surveillance, surgical video analysis, sports analysis, and content recommendation. Progress in this domain has been largely driven by advancements in deep learning, relying on large-scale labeled datasets. However, the process of labeling these video datasets requires a tremendous amount of human effort and time. Another significant issue in video understanding is the leakage of private attributes, such as skin color, race, gender, and clothing information, which is undesirable when recognizing a person's actions.

This dissertation addresses these fundamental challenges in video understanding through several contributions. We introduce contrastive learning-based and temporal pretext task-based self-supervised frameworks to learn from unlabeled video datasets. To leverage smaller labeled datasets, we also propose semi-supervised learning frameworks for both coarse-grained and fine-grained action recognition. Additionally, we tackle privacy preservation in action recognition by introducing a self-supervised privacy removal objective.

SELECTED PUBLICATIONS (h-index: 13, total citation: 600)

1. **TCLR: Temporal Contrastive Learning for Video Representation**, IR Dave, R Gupta, MN Rizve, and M Shah, in *Computer Vision and Image Understanding (CVIU),* 2022.
2. **SPAct: Self-supervised Privacy Preservation for Action Recognition**, IR Dave, C Chen, and M Shah, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 2022.
3. **TimeBalance: Temporally-Invariant and Temporally-Distinctive Video Representations for Semi-Supervised Action Recognition**, IR Dave, MN Rizve, C Chen, and M Shah, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 2023.
4. **EventTransAct: A Video Transformer-based Framework for Event-camera based Action Recognition**, T de Blegiers*, IR Dave*, A Yousaf, and M Shah, in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* 2023. (*=equal contribution)
5. **TeD-SPAD: Temporal Distinctiveness for Self-supervised Privacy-preservation for Video Anomaly Detection**, J Fioresi, IR Dave, and M Shah, in *IEEE/CVF International Conference on Computer Vision (ICCV),* 2023.
6. **No More Shortcuts: Realizing the Potential of Temporal Self-Supervision**, IR Dave, S Jenni, and M Shah, in *AAAI Conference on Artificial Intelligence,* 2024.
7. **CodaMal: Contrastive Domain Adaptation for Malaria Detection in Low-Cost Microscopes**, IR Dave, T de Blegiers, C Chen, and M Shah, in *IEEE International Conference on Image Processing (ICIP),* 2024.
8. **FinePseudo: Improving Pseudo-Labelling through Temporal-Alignability for Semi-Supervised Fine-Grained Action Recognition**, IR Dave, MN Rizve, and M Shah, in *European Conference on Computer Vision (ECCV),* 2024.
9. **Sync from the Sea: Retrieving Alignable Videos from Large-Scale Datasets**, IR Dave, F Caba, M Shah, and S Jenni, in *European Conference on Computer Vision (ECCV),* 2024.

DISSERTATION

TOWARDS LABEL EFFICIENCY AND PRIVACY PRESERVATION IN VIDEO UNDERSTANDING

The progress in video understanding has been largely driven by large-scale labeled datasets. However, video annotation presents significant challenges due to its time-consuming and expensive nature. This limitation underscores the importance of developing methods that can learn effectively from unlabeled or limited-labeled data, which makes self-supervised learning (SSL) and semi-supervised learning particularly relevant for video understanding. Another significant challenge in video understanding is privacy preservation, as methods often inadvertently leak private information, presenting a growing concern in the field. In this dissertation, we present methods to improve the label efficiency by employing self-supervised and semi-supervised methods, and a self-supervised method designed to mitigate privacy leakage in action recognition task.

Our first contribution is Temporal Contrastive Learning framework for Video Representation (TCLR). Unlike prior contrastive self-supervised learning methods which aim to learn temporal similarity between different clips of the same video, TCLR encourages the learning differences rather than similarities in clips from the same video. TCLR consists of two novel losses to improve upon existing contrastive self-supervised video representations, contrasting temporal segments of the same video at two different temporal aggregation steps: clip level and temporal pooling level.

While TCLR offers an effective solution for video-level downstream tasks, it does not encourage framewise video representation for addressing low-level temporal correspondence-based downstream tasks. Next, “No More Shortcuts” (NS) to promote a more effective framewise video representation. We first eliminate learning shortcuts present in existing temporal pretext tasks by introducing framewise spatial jittering and proposing more challenging frame-level temporal pretext tasks. Our approach results in state-of-the-art performance across a wide range of downstream tasks, encompassing both high-level semantic and low-level temporal correspondence tasks.

Above Video SSL approaches, TCLR and NS, focus only on learning from unlabeled videos, in practice, however some labeled data often exists. Our next focus is on semi-supervised action recognition, where we have a small set of labeled videos with a large pool of unlabeled videos. Using the observations from the SSL representations, we leverage the unlabeled videos using the complementary strengths of temporally-invariant and temporally-distinctive contrastive self-supervised video representations. Our proposed semi-supervised method “TimeBalance” introduces a student-teacher framework that dynamically combines the knowledge of two self-supervised teachers based on the nature of the unlabeled video using the proposed reweighting strategy.

Even though TimeBalance performs well for coarse-grained actions, it struggles with fine-grained actions. To address this, next we propose “FinePseudo” framework, which leverages temporal alignability to learn phase-aware distances. It also introduces collaborative pseudo-labeling between video-level and alignability encoder, refining the pseudo-labeling process for fine-grained actions.

Although the above mentioned video representations are useful for various downstream applications, they often leak a considerable amount of private information present in the videos. To mitigate the privacy leaks in videos, we propose SPAct, a self-supervised framework that removes private information from input videos without requiring privacy labels. SPAct exhibits competitive performance compared to supervised methods and introduces new evaluation protocols to assess the generalization capability of the anonymization across novel action and privacy attributes.