Research Methodology Assignment 2

Submitted By:

Ajithkumar A K
[CB.AI.R4CEN24009]

November 17, 2024

Chapter 1

Abstract

Activity recognition in video surveillance plays a pivotal role in applications ranging from security monitoring to behavior analysis. However, traditional methods often face challenges in handling complex spatio-temporal dependencies and data scarcity. To address these limitations, Generative Adversarial Networks (GANs) and advanced video representation techniques have emerged as promising solutions. GANs can generate synthetic labeled data and augment training datasets, alleviating the challenge of limited labeled data in semi-supervised learning frameworks. Concurrently, video representation techniques, leveraging temporal dynamics and multimodal attributes, enable efficient feature extraction for capturing subtle motion patterns and context-specific cues. This survey reviews recent advancements in GAN-based approaches and video representation techniques, examining their impact on improving activity recognition for video surveillance. We highlight state-of-the-art methods, benchmark datasets, and the integration of GANs with temporal and multimodal frameworks. Finally, we discuss existing challenges and outline future research directions to further enhance the reliability and scalability of video surveillance systems.

Chapter 2

Introduction

Video surveillance has become an integral component of modern security systems, enabling automated monitoring and detection of activities in diverse environments. The task of activity recognition within video surveillance is pivotal for applications such as anomaly detection, crowd monitoring, and behavioral analysis. However, this domain presents several challenges, including the complexity of spatio-temporal dynamics, environmental variations, and the scarcity of labeled data. These challenges necessitate robust and scalable methods for improving activity recognition.

Generative Adversarial Networks (GANs) have emerged as a transformative technology, particularly in addressing data-related challenges. By generating realistic synthetic samples, GANs augment datasets and enhance model robustness, especially in semi-supervised and unsupervised learning scenarios. Moreover, GANs facilitate domain adaptation by bridging gaps between different environments, improving model performance in diverse surveillance settings.

In parallel, advancements in video representation techniques have significantly contributed to better understanding and encoding of spatio-temporal features. Techniques such as multimodal representation, temporal modeling, and fine-grained feature extraction enable systems to capture subtle motions and complex activities with higher accuracy. These methods leverage the rich information inherent in video data, ensuring that models effectively utilize spatial, temporal, and contextual cues.

This survey explores the intersection of GANs and video representation techniques in improving activity recognition for video surveillance. By reviewing state-of-the-art methods, benchmark datasets, and innovative architectures, this study aims to provide a comprehensive understanding of the current landscape and identify promising directions for future research.

Chapter 3

Literature review

Recent advancements in egocentric action recognition [8] have shown promise in enhancing human-robot interactions, particularly in industrial settings. This study leverages the MECCANO dataset, focusing on assembling actions using both RGB and Depth modalities. A novel framework based on the 3D Video SWIN Transformer effectively encodes these modalities, while a specialized training strategy employing an exponentially decaying focal loss addresses the long-tailed distribution of action classes. Late fusion integrates predictions from both modalities, significantly improving performance. The proposed method outperforms prior work and secured first place in the multimodal action recognition challenge at ICIAP 2023, showcasing its robustness and applicability.

Effectiveness of Transformers in Low-Labeled Video Recognition Transformers [14] have recently emerged as a powerful alternative to CNNs in computer vision, offering greater representational capacity due to their less restrictive inductive bias. Surprisingly, despite the common belief that transformers require vast amounts of data, they perform exceptionally well in low-labeled video recognition scenarios. An empirical study on video vision transformers demonstrated their superiority over CNNs, even outperforming semi-supervised CNN methods that leverage large-scale unlabeled datasets. Ablation studies highlighted the critical role of pretraining in achieving this performance. This research encourages future exploration of vision transformers across diverse datasets and architectures, potentially redefining video recognition paradigms.

SVFormer [18] addresses the challenge of semi-supervised action recognition by leveraging the power of transformers in video understanding. To handle unlabeled video data, it adopts a steady pseudo-labeling framework (EMA-Teacher). Recognizing the limitations of conventional augmentations for video data, the study introduces Tube TokenMix, a novel augmentation technique that mixes video clips using consistent temporal masks. Additionally, temporal warping augmentation handles temporal variations by altering frame durations. Experiments on Kinetics-400, UCF-101, and HMDB-51 reveal SVFormer's significant performance gains, outperforming state-of-the-art methods, particularly under low labeling rates, and setting a strong benchmark for future transformer-based research in video recognition.

OpenLDN [15] tackles the challenges of open-world semi-supervised learning, where labeled and unlabeled data do not necessarily share the same distribution. Unlike conventional semi-supervised learning (SSL) methods, OpenLDN identifies and clusters novel classes present in the unlabeled data. The framework employs a pairwise similarity loss to detect novel classes by leveraging labeled data through a bi-level optimization approach. By clustering novel samples and recognizing known classes, OpenLDN transforms the open-world SSL task into a standard SSL problem for improved performance. Extensive evaluations show that OpenLDN outperforms state-of-the-art methods on popular benchmarks, achieving high accuracy with an efficient training time.

This work [16] introduces a temporal contrastive learning framework to tackle the challenge of action recognition with limited labeled videos. By exploiting the temporal information in unlabeled videos, the method maximizes similarity between the representations of a video played at two different speeds while minimizing similarity across different videos at varying speeds. This novel manipulation of playback rates extracts supervisory signals from unlabeled data, effectively leveraging temporal dynamics. The approach demonstrates superior performance over state-of-the-art semi-supervised image recognition extensions and generalizes well even with out-of-domain data. Extensive experiments validate its robustness across diverse benchmarks and architectures.

This research [20] explores the impact of diverse data augmentation strategies to address the complexities of video-based action recognition. By incorporating photometric, geometric, temporal, and actor/scene augmentations, the study aims to instill representational invariances, improving model accuracy in both low-label and full-label settings. The proposed strategies are validated on benchmark datasets like Kinetics-100/400, Mini-Something-v2, UCF-101, and HMDB-51, demonstrating superior performance in semi-supervised and fully supervised frameworks. The findings emphasize the significance of temporally-coherent augmentations and their role in enhancing data efficiency, contributing valuable insights for advancing video action recognition techniques.

Learning from Temporal Gradient for Semi-Supervised Action Recognition This study [17] introduces a novel semi-supervised learning approach that leverages temporal gradients (TG) as an additional modality to enhance motion-related representation learning in video action recognition. By imposing cross-modal consistency between RGB and TG representations, the method effectively captures fine-grained motion dynamics, resulting in significant performance improvements. Employing block-wise dense alignment and cross-modal contrastive learning strategies, the approach achieves state-of-the-art results on benchmarks such as Kinetics-400, UCF-101, and HMDB-51 across various labeled data ratios. Importantly, the method requires no additional computational overhead during inference, making it efficient and scalable. Future directions include exploring the utility of TG in other video-based tasks and discovering new modalities for further enhancement.

SPAct: Self-supervised Privacy Preservation for Action Recognition [3] This paper presents a novel self-supervised framework, SPAct, for privacy-preserving action recognition in videos. Traditional approaches for mitigating privacy leakage require privacy labels alongside action labels, which is often impractical. SPAct addresses this by employing self-supervised learning to anonymize privacy-sensitive information without the need for privacy annotations. The framework consists of three key components: an anonymization function, a self-supervised privacy removal branch, and an action recognition branch. Using a minimax optimization strategy, SPAct balances action recognition and privacy preservation through a contrastive self-supervised loss. Extensive experiments demonstrate that SPAct not only achieves competitive performance in terms of action recognition while preserving privacy but also outperforms supervised methods in terms of generalization to novel action-privacy attributes. This work highlights the potential of self-supervised learning to tackle privacy concerns in video understanding.

TransRank: Self-supervised Video Representation Learning via Ranking-based Transformation Recognition [6] The paper introduces TransRank, a unified framework for self-supervised video representation learning that addresses the limitations of traditional transformation recognition approaches (RecogTrans) in pre-training. Unlike the commonly used instance discrimination methods, RecogTrans methods have struggled with noisy supervision signals. TransRank mitigates this issue by recognizing transformations relative to each other using a ranking-based formulation, providing more accurate supervision signals. The framework is highly flexible, allowing the use of various temporal or spatial transformations. Through extensive experiments, TransRank demonstrates superior performance, outperforming previous methods by 6.4% and 8.3% on UCF101 and HMDB51 action recognition tasks (Top1 accuracy), respectively, and improves video retrieval on UCF101 by 20.4% (R@1). These results validate the potential of RecogTrans for video self-supervised learning, and the authors highlight the promising direction of combining it with instance discrimination approaches to boost performance further. The authors will release their code and pre-trained models, providing a solid baseline for future research in the field.

Video Contrastive Learning with Global Context [9] This paper presents a novel method for video-level contrastive learning (VCLR) that addresses the limitation of existing methods, which primarily rely on short-range spatiotemporal salience for clip-level contrastive signals. VCLR overcomes this by using video segments to define positive pairs, allowing it to capture global context and become more robust to temporal content changes. Additionally, the method incorporates a temporal order regularization to enforce the inherent sequential structure of videos. Extensive experiments demonstrate that VCLR outperforms previous state-of-the-art methods on five video datasets for action classification, action localization, and video retrieval tasks. The framework is shown to generalize well, as it is compatible with different network architectures (2D and 3D CNNs) and self-supervised learning algorithms. The authors suggest that VCLR has broader applicability to various input types (e.g., short trimmed or long untrimmed videos) and can scale to future video datasets, positioning it as a strong baseline for video understanding research that utilizes global context.

Spatiotemporal Contrastive Video Representation Learning (CVRL) This paper [13] introduces a self-supervised Contrastive Video Representation Learning (CVRL) method that aims to learn spatiotemporal visual representations from unlabeled videos. CVRL employs a contrastive loss where augmented clips from the same video are pulled together in the embedding space, while clips from different videos are pushed apart. The authors emphasize the importance of both spatial and temporal information for effective video representation learning. To this end, they propose two novel data augmentation techniques: a temporally consistent spatial augmentation to maintain spatial variation across frames while preserving temporal consistency, and a sampling-based temporal augmentation to prevent excessive invariance on clips far apart in time. The performance of CVRL is evaluated on the Kinetics-600 dataset, where it achieves a top-1 accuracy of 70.4% using a 3D-ResNet-50 backbone, surpassing both ImageNet supervised pre-training by 15.7% and SimCLR unsupervised pre-training by 18.8%. With a larger backbone (R3D-152), the accuracy improves to 72.9%, significantly narrowing the gap between unsupervised and supervised video representation learning. This study demonstrates that CVRL effectively captures spatiotemporal cues and achieves state-of-the-art results on several video-related tasks, suggesting its potential for future applications involving larger unlabeled video datasets and multimodal integration.

Self-supervised Spatiotemporal Representation Learning by Exploiting Video Continuity This paper [10] introduces the Continuity Perception Network (CPNet), a novel self-supervised learning framework designed to exploit the continuity property of videos for representation learning. The authors argue that video continuity—spanning both local and long-range motion and context—has been underexplored in existing methods. CPNet addresses this gap by proposing three continuity-related pretext tasks: continuity justification, discontinuity localization, and missing section approximation. These tasks jointly guide a shared backbone to learn spatiotemporal features without supervision. By training on these tasks, CPNet enhances the model's ability to capture both fine- and coarse-grained motion and context representations. The proposed framework achieves superior performance on several downstream tasks, including action recognition, video retrieval, and action localization, surpassing previous methods. Furthermore, integrating CPNet's continuity-aware pretext tasks into existing self-supervised approaches significantly boosts performance. The work emphasizes the complementary nature of video continuity to other coarse-grained video properties, showcasing its potential to advance video representation learning.

TCLR: Temporal Contrastive Learning for Video Representation [2] In this work, the authors introduce Temporal Contrastive Learning (TCLR), a novel framework designed to enhance self-supervised video representation learning by explicitly encouraging temporal diversity. The framework includes two innovative losses: the local-local temporal contrastive loss, which distinguishes non-overlapping clips from the same video, and the global-local temporal contrastive loss, which discriminates between timesteps in the feature map of an input clip. These losses aim to improve the temporal structure of learned features, addressing a gap in prior contrastive learning methods for video. TCLR outperforms existing techniques on multiple downstream tasks, including action recognition, limited-label action classification, and nearest-neighbor video retrieval, across various video datasets and backbone architectures. Notably, the framework achieves significant improvements, such as a 5.1% increase in top-1 accuracy on UCF101 and a 5.4% increase on HMDB51 for action classification, along with a remarkable 11.7% boost in Top-1 Recall for video retrieval on UCF101. The study demonstrates the efficacy of temporal contrastive learning in enhancing the quality of learned video representations, showing its potential to advance video understanding tasks beyond standard instance discrimination.

ASCNet: Self-supervised Video Representation Learning with Appearance-Speed Consistency [7] The ASCNet framework introduces a self-supervised video representation learning approach that leverages appearance consistency within a video and speed consistency between videos with the same frame rate. The method learns robust video features by aligning clips in both appearance and speed embedding spaces, ensuring consistency across different time frames. ASCNet addresses challenges in existing methods that rely on contrastive loss and large batch sizes, which often include noisy data or require additional modalities. The framework introduces two key tasks: appearance consistency, which maximizes similarity between clips from the same video with varying playback speeds, and speed consistency, which ensures similarity between clips with the same playback speed but different appearance information. Extensive experiments show that ASCNet significantly outperforms existing approaches, achieving 90.8% accuracy on UCF-101 action recognition without the need for extra modalities or negative pairs. This work highlights the importance of consistency in learning video representations and demonstrates the potential of ASCNet for action recognition and video retrieval tasks.

Long Short View Feature Decomposition via Contrastive Video Representation Learning In this work, the authors introduce a novel approach for self-supervised video representation learning that decomposes video features into stationary and non-stationary components using contrastive learning. The core idea is to use long and short video views—long sequences and their shorter sub-sequences—to separately learn stationary features (which remain constant over time) and non-stationary features (which exhibit temporal variation). This decomposition enables a more targeted learning of video representations, with stationary features excelling in tasks requiring static information, such as action recognition, and non-stationary features benefiting tasks requiring temporal dynamics, such as action segmentation. The authors show substantial performance improvements in action recognition on UCF101 and action segmentation on the Breakfast dataset. This work underscores the importance of separating temporal attributes into distinct components, leading to better task-specific feature learning. The findings also highlight the potential of unsupervised learning for improving fine-grained video tasks like action segmentation, demonstrating the usefulness of decomposed features in various downstream video understanding tasks.

Controllable Augmentations for Video Representation Learning This paper proposes a novel framework for self-supervised video representation learning that addresses common limitations in existing contrastive learning methods, which often suffer from background bias and difficulties in capturing global temporal structures. The authors introduce controllable augmentations that allow for the generation of local clips and global video segments, enabling the learning of both detailed region-level correspondences and long-term temporal relations. The framework uses spatio-temporal region contrastive learning to align appearance and motion patterns accurately, minimizing low-level redundancies and enhancing generalization. Additionally, the inclusion of local-global temporal order dependency bridges the gap between clip-level and video-level representations, leading to improved temporal modeling. Experiments on action recognition and video retrieval tasks demonstrate that the proposed framework significantly outperforms existing methods by capturing more precise temporal dynamics and offering better generalization across various video benchmarks.

Time-Equivariant Contrastive Video Representation Learning This paper presents a novel approach to self-supervised contrastive video representation learning by introducing temporal equivariance to better capture video dynamics. Unlike existing methods, which often focus on learning invariance to temporal transformations, the authors argue that video representations should reflect temporal manipulations and preserve video dynamics. The proposed method encodes relative temporal transformations between augmented video clips and contrasts them using transformation vectors. Additionally, it includes a self-supervised classification task that categorizes clips as overlapping, ordered, or unordered to reinforce the learning of temporal equivariance. Experimental results on action recognition and video retrieval tasks demonstrate that the proposed method outperforms state-of-the-art approaches on datasets such as UCF101, HMDB51, and Diving48, showing significant improvements in capturing temporal dynamics.

Motion-aware Contrastive Video Representation Learning via Foreground-background Merging (FAME) [4] This paper addresses a key challenge in self-supervised video representation learning: the issue of background bias. Traditional contrastive learning methods, while effective in image domains, often struggle with video data as they tend to focus on the common static background rather than the motion information. This leads to weak generalization and poor performance in downstream tasks like action recognition. To overcome this, the authors propose a novel approach called Foreground-background

Merging (FAME), which deliberately merges the moving foreground of one video with the static background of another. Using frame difference and color statistics, they extract the moving foreground without requiring off-the-shelf detectors and shuffle background regions across videos. This method forces the model to focus more on motion patterns while mitigating the background bias. Extensive experiments on UCF101, HMDB51, and Diving48 datasets show that FAME significantly improves performance on various downstream tasks. The method's effectiveness lies in the semantic consistency between original and fused video clips, leading to more motion-aware representations. However, challenges remain, such as unstable foreground extraction and fixed foreground area ratios, suggesting potential areas for future improvement.

Self-Supervised Video Representation Learning with Meta-Contrastive Network (MCN) In this paper [11], the authors introduce a Meta-Contrastive Network (MCN) that combines contrastive learning and meta-learning to enhance self-supervised video representation learning. While existing methods primarily rely on contrastive loss to achieve instance-level discrimination, they often face limitations due to the lack of category information, which leads to the hard-positive problem. The authors propose a solution by integrating model-agnostic meta-learning (MAML) with a two-stage training process, comprising a contrastive branch and a meta branch. The contrastive branch uses NCE loss to optimize instance discrimination, while the meta branch employs binary classification loss to improve adaptation across tasks. This combination helps to better generalize across downstream tasks like video action recognition and video retrieval. Extensive evaluations on UCF101 and HMDB51 datasets show that MCN significantly outperforms state-of-the-art approaches, achieving Top-1 accuracies of 84.8% and 54.5% for video action recognition, and 52.5% and 23.7% for video retrieval. The paper marks the first integration of meta-learning into self-supervised video representation learning and demonstrates substantial improvements in generalization and task adaptation.

Self-Supervised Spatiotemporal Representation Learning by Exploiting Video Continuity In this paper [10], the authors propose a novel self-supervised learning framework called Continuity Perception Network (CPNet), which leverages the essential yet under-explored property of video continuity to improve self-supervised video representation learning. Unlike traditional methods that focus on temporal order or speed, CPNet introduces three continuity-related pretext tasks: continuity justification, discontinuity localization, and missing section approximation. These tasks are designed to jointly supervise a shared backbone network, encouraging the model to learn both local and long-range motion and contextual representations. The authors demonstrate that CPNet outperforms previous methods on various downstream tasks, such as action recognition, video retrieval, and action localization. Furthermore, they show that the continuity-based pretext tasks can be effectively integrated with other coarse-grained video properties to further enhance performance. The framework not only introduces innovative tasks for video learning but also complements existing self-supervised approaches, yielding significant performance gains.

VideoMoCo: Contrastive Video Representation Learning with Temporally Adversarial Examples In this paper [12], the authors propose VideoMoCo, an extension of the MoCo framework designed for self-supervised video representation learning. The key innovation of VideoMoCo lies in improving temporal feature representations by incorporating two main strategies. First, they introduce a generator that drops out frames from the video input during training. This adversarial learning process forces the discriminator (encoder) to learn to encode robust and consistent feature representations, regardless of frame omissions. By adaptively dropping frames during training iterations, the model is made more temporally robust. Second, the authors propose temporal decay to address the issue of memory queue degradation in the MoCo framework. This technique models how older keys in the memory queue contribute less to the contrastive loss as the encoder updates, allowing the model to focus more on recent samples. Together, these innovations empower VideoMoCo to learn video representations without needing manually designed pretext tasks. Extensive experiments on UCF101 and HMDB51 datasets demonstrate that VideoMoCo significantly improves MoCo's temporal robustness and outperforms state-of-the-art methods for self-supervised video representation learning.

Contrastive Spatio-Temporal Pretext Learning for Self-supervised Video Representation In this paper [19], the authors address the limitations of traditional contrastive learning methods for self-supervised video representation learning, which often overlook the intermediate states of learned representations. They propose a novel pretext task called spatio-temporal overlap rate (STOR) prediction. This task encourages the model to discriminate the degree of overlap between two video samples, capturing both spatial and temporal similarities. The authors observe that humans can naturally perceive the overlap rates of videos, and by training the model on this task, it learns to enhance its spatio-temporal representation. Furthermore, they introduce a joint optimization framework called CSTP (contrastive spatio-temporal pretext) that combines the STOR task with traditional contrastive learning. This framework boosts the overall spatio-temporal feature learning. Through extensive experiments, the authors demonstrate that the STOR prediction task significantly improves the performance of both action recognition and video retrieval tasks, achieving state-of-the-art results. The paper also provides valuable insights into the mutual influence of each component within the CSTP framework, offering new perspectives on self-supervised learning.

Dual Contrastive Learning for Spatio-temporal Representation This paper [5] proposes a dual contrastive learning method (DCLR) to address the challenge of background scene bias in self-supervised spatio-temporal representation learning. Traditional contrastive learning techniques often rely on sampling video clips to construct positive and negative pairs. However, the authors observe that this approach causes the model to prioritize background scenes, which are easier to discriminate than motion patterns. To overcome this, the paper introduces a novel formulation that decouples the input RGB video sequence into two complementary components: static scenes and dynamic motion. The model is trained to pull the RGB features closer to both the static scene and the aligned dynamic motion features, enabling it to better represent both types of information. The authors also employ activation maps to separate static and dynamic features, ensuring that each component is encoded effectively. Through extensive experiments on UCF-101, HMDB-51, and Diving-48 datasets, DCLR achieves state-of-the-art or competitive performance, showcasing the method's effectiveness in overcoming background bias and learning robust spatio-temporal features. While the authors acknowledge that the method's performance could be improved with larger model backbones and higher resolution inputs, their work presents a significant step forward in enhancing spatio-temporal representation learning.

RSPNet: Relative Speed Perception for Unsupervised Video Representation Learning In this paper, [1] the authors introduce RSPNet, an unsupervised video representation learning framework designed to learn both motion and appearance features from unlabeled videos. The challenge of this task lies in the complex spatial-temporal dynamics of videos and the absence of labeled data. Traditional methods of video representation learning, such as speed prediction, struggle with the imprecision of speed labels and may fail to adequately capture appearance features. The authors address these issues by leveraging relative playback speed between video clips as a stable and effective supervisory signal for motion learning. By focusing on relative speed, RSPNet can better capture motion patterns while maintaining stability in training. To ensure the model also learns appearance features, the authors propose an appearance-focused task, where the model perceives the appearance differences between video clips. The framework is trained jointly on these two tasks, significantly improving the learning of both motion and appearance features. Through extensive experiments, RSPNet achieves state-of-the-art performance on action recognition and video retrieval tasks, particularly achieving a remarkable 93.7% accuracy on the UCF101 dataset for action recognition without using any labeled data for pre-training. This work demonstrates the effectiveness of relative speed perception and dual-task learning in enhancing unsupervised video representation learning.

REFERENCES

[1] Peihao Chen, Deng Huang, Dongliang He, Xiang Long, Runhao Zeng, Shilei Wen, Mingkui Tan, and Chuang Gan. Rspnet: Relative speed perception for unsupervised video representation learning. In *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, volume 2A, 2021.

[2] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406, 2022.

[3] Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. Spact: Self-supervised privacy preservation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20164–20173, 2022.

[4] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9716–9726, 2022.

[5] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2022-June, 2022.

[6] Haodong Duan, Nanxuan Zhao, Kai Chen, and Dahua Lin. Transrank: Self-supervised video representation learning via ranking-based transformation recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3000–3010, 2022.

[7] Deng Huang, Wenhao Wu, Weiwen Hu, Xu Liu, Dongliang He, Zhihua Wu, Xiangmiao Wu, Mingkui Tan, and Errui Ding. Ascnet: Self-supervised video representation learning with appearance-speed consistency. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8096–8105, 2021.

[8] Jyoti Kini, Sarah Fleischer, Ishan Dave, and Mubarak Shah. Egocentric rgb+ depth action recognition in industry-like settings. *arXiv preprint arXiv:2309.13962*, 2023.

[9] Haofei Kuang, Yi Zhu, Zhi Zhang, Xinyu Li, Joseph Tighe, Sören Schwertfeger, Cyrill Stachniss, and Mu Li. Video contrastive learning with global context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3195–3204, 2021.

[10] Hanwen Liang, Niamul Quader, Zhixiang Chi, Lizhe Chen, Peng Dai, Juwei Lu, and Yang Wang. Self-supervised spatiotemporal representation learning by exploiting video continuity. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, volume 36, 2022.

[11] Yuanze Lin, Xun Guo, and Yan Lu. Self-supervised video representation learning with meta-contrastive network. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[12] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2021.

[13] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6964–6974, 2021.

[14] Farrukh Rahman, Ömer Mubarek, and Zsolt Kira. On the surprising effectiveness of transformers in low-labeled video recognition. *arXiv preprint arXiv:2209.07474*, 2022.

[15] Mamshad Nayeem Rizve, Navid Kardan, Salman H. Khan, Fahad Shahbaz Khan, and Mubarak Shah. Openldn: Learning to discover novel classes for open-world semi-supervised learning. *ArXiv*, abs/2207.02261, 2022.

[16] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogério Schmidt Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10384–10394, 2021.

[17] Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Loddon Yuille, and Yingwei Li. Learning from temporal gradient for semi-supervised action recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3252, 2021.

[18] Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Svformer: Semi-supervised video transformer for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18816–18826, 2023.

[19] Yujia Zhang, Lai Man Po, Xuyuan Xu, Mengyang Liu, Yexin Wang, Weifeng Ou, Yuzhi Zhao, and Wing Yin Yu. Contrastive spatio-temporal pretext learning for self-supervised video representation. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, volume 36, 2022.

[20] Yuliang Zou, Jinwoo Choi, Qitong Wang, and Jia-Bin Huang. Learning representational invariances for data-efficient action recognition. *ArXiv*, abs/2103.16565, 2021.