Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Self-Supervised Learning for Action Recognition: Trends, Models, and Applications

Mouwiya S. A. Al-Qaisieh[1], Mas Rina Mustaffa [1]

Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM)[1]

ARTICLE INFO

Received: 20 Dec 2024

Revised: 15 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Recent advances in self-supervised learning (SSL) have reshaped the landscape of human action recognition by reducing dependency on large-scale annotated datasets. This survey provides a comprehensive overview of state-of-the-art SSL techniques developed for understanding human actions in videos. We categorize methods into three primary paradigms: contrastive learning, masked video modeling, and multimodal or sensor-based approaches. Across each category, we discuss key innovations including motion-guided contrastive sampling, transformer-based masked autoencoders, and cross-modal alignment strategies that leverage audio, skeleton, or wearable sensor signals. Models such as VideoMAE, ST-MAE, XDC, and Actionlet-Contrastive represent significant milestones in capturing both spatial and temporal cues without supervision. Beyond model design, we identify major challenges facing current SSL systems, including generalization across domains, modeling long-horizon activities, and real-time deployment constraints. We also highlight underexplored areas such as explainability and unified evaluation protocols. To guide future work, we present a structured taxonomy, a comparative table of representative models, and a discussion of promising research directions including multimodal fusion, modality-agnostic learning, and hardware-aware training. This survey aims to equip researchers with a clear understanding of the evolving trends, persistent gaps, and opportunities that lie ahead in self-supervised action recognition.

Keywords: Component; Self-supervised learning; Human action recognition; Contrastive learning, Multimodal fusion

I. Introduction

Human action recognition (HAR) plays a central role in enabling intelligent systems to interpret human behaviors from video data. Applications span a wide spectrum, including surveillance, autonomous driving, healthcare monitoring, and interactive robotics. Traditionally, progress in this field has been driven by supervised deep learning models trained on large-scale, curated datasets such as UCF101 [1] and Kinetics [2]. While these benchmarks have enabled high-performing architectures, their reliance on dense human annotation introduces major limitations. Labeling video is costly, time-consuming, and often subjective especially when distinguishing between subtle, overlapping actions.

In recent years, self-supervised learning (SSL) has emerged as a compelling alternative. Rather than depending on labeled data, SSL derives supervisory signals from the data itself through the design of proxy tasks [3]. These tasks encourage models to learn semantically meaningful visual and temporal features by reconstructing masked inputs [4], predicting future frames [5], or distinguishing augmented clips [6]. This paradigm not only reduces the annotation burden but also tends to generalize better across datasets and tasks.

The growing interest in SSL for video understanding has led to the development of diverse model families. Masked modeling approaches such as VideoMAE [7] and ST-MAE [8] reconstruct missing patches in video sequences using transformer-based encoders, capturing both spatial and temporal context. In parallel, contrastive learning frameworks like SimCLR [6], MoCo [9], and BYOL [10] have been extended to video, leading to models like MIL-NCE [11] and MaCLR [12], which integrate motion

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

cues and temporal augmentations to enhance representation learning. Multimodal methods such as XDC [13], AVID [14], and MMV [15] leverage cross-modal alignment particularly between audio and video to provide richer, more grounded supervision. Even originally supervised architectures, such as TimeSformer [16] and ViViT [17], have demonstrated strong performance when adapted to self-supervised pretraining regimes.

Despite these successes, several challenges remain. Current SSL models often struggle with fine-grained action differentiation, long-range temporal reasoning, and generalization to real-world conditions [18], [19] and [20]. Moreover, the computational cost of pretraining and a lack of standardized evaluation protocols make it difficult to compare methods directly or deploy them efficiently in practice.

This survey presents a comprehensive review of self-supervised learning techniques for video-based human action recognition. We categorize current approaches into three major paradigms contrastive learning, masked video modeling, and multimodal representation learning and provide a critical comparison of their design strategies, capabilities, and limitations. Furthermore, we identify key research gaps and propose future directions centered on cross-modal fusion, efficient spatiotemporal pretraining, and deployment in dynamic environments.

By synthesizing recent developments and providing a unified perspective, this survey aims to guide researchers and practitioners working to advance action recognition without the constraints of manual labeling. The following section synthesizes existing surveys and organizes the evolving body of work on self-supervised action recognition into three dominant paradigms: contrastive learning, masked video modeling, and multimodal or sensor-based approaches.

II. Literature review

Human action recognition (HAR) has rapidly evolved from hand-crafted features to deep learning-based techniques powered by large-scale datasets like UCF101 [1] and Kinetics [2]. However, these supervised models require vast amounts of annotated data, which is often expensive, time-consuming, and limited in scalability across diverse domains.

In response to these limitations, self-supervised learning (SSL) has emerged as a promising alternative, enabling models to learn discriminative spatiotemporal features from raw video or sensor data. SSL has developed into several paradigms, including contrastive learning, masked modeling, multimodal representation, and sensor-based learning. This section surveys foundational developments and current research directions across these domains

A. Related Surveys and Background

Building upon the motivation introduced earlier, we begin by reviewing foundational survey works that have shaped the landscape of self-supervised learning in computer vision and video understanding

The foundations of SSL in vision were first mapped in early surveys such as [4] and [20], which covered general and video-specific SSL methods, respectively. More recent works like [18], [19], and [21] expanded this discussion to human activity recognition using wearable sensors, such as accelerometers and gyroscopes [22].

However, most existing reviews either overlook modern transformer-based models like VideoMAE [4] and ViViT [17], or fail to comprehensively integrate multimodal and masked modeling strategies [5]. Our survey fills this gap by bringing together cutting-edge techniques across all domains, including recent masked and contrastive methods for skeleton-based recognition.

B. Contrastive Learning for Video and Skeletons

Contrastive learning became foundational in SSL through frameworks like SimCLR [6], MoCo [9], and BYOL [10], which align similar representations (positives) while distinguishing different samples (negatives). These ideas transitioned to video with methods like TimeContrast [5] and CVRL [22], introducing temporal and motion-based augmentations.

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

MaCLR [12] advanced this further by encoding explicit motion cues as learning signals, while TCL [5] used transformation prediction. MIL-NCE [11] and XDC [13] showed that multimodal contrastive learning with audio-video alignment yields stronger representations. Part-aware methods [23] and actionlet-dependent contrastive learning refined these approaches by targeting discriminative body regions in skeleton-based data.

Despite their power, contrastive methods can be sensitive to negative sample quality, leading to growing interest in alternatives like masked modeling.

C. Masked Video Modeling

Inspired by the success of BERT in NLP, masked modeling pretrains models by hiding parts of input data and learning to reconstruct them. In video, this approach was pioneered by VideoMAE [4], which masks over 75% of input patches and uses lightweight decoders for spatiotemporal recovery [22].

ST-MAE [7] extended this with explicit temporal masking. Prompted masked modeling [21] introduced task-aware guidance to refine reconstruction targets. Recent models like SIGMA [24] and MVD [25] incorporate Sinkhorn matching and distillation to enhance semantic fidelity.

These methods often rely on transformer backbones such as ViViT [17] and TimeSformer [16], which are well-suited for handling long-range temporal dependencies.

D. Multimodal and Cross-Modal SSL

Real-world video understanding often benefits from combining vision with other modalities like audio or motion. AVID [26] and MMV [15] demonstrated the effectiveness of aligning sound and vision in self-supervised settings, while XDC [13] used uncurated videos to train cross-modal embeddings [27].

MIL-NCE [11] added flexibility by supporting arbitrary transformation pairs, allowing alignment between sight and sound or between different video views. These multimodal approaches improve robustness and generalization, especially in noisy environments. These self-supervised learning methods can be broadly categorized into three main paradigms: contrastive learning, masked modeling, and multimodal or sensor-based approaches.

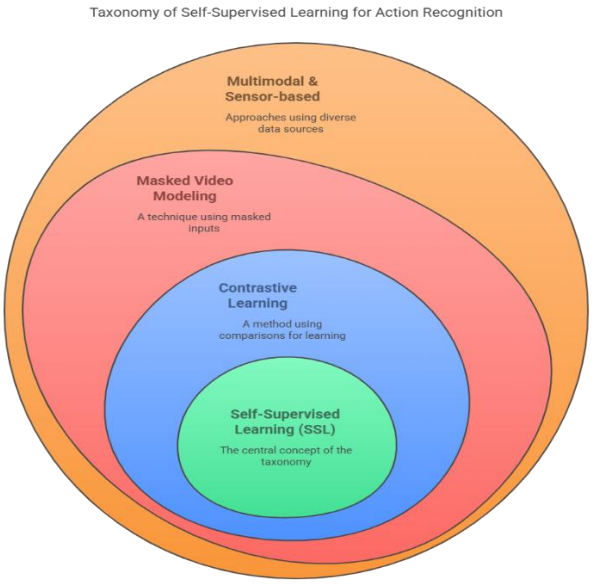

Fig. 1 Hierarchical taxonomy of self-supervised learning (SSL) approaches for human action recognition. Core SSL methods such as contrastive and masked modeling form the foundation, which is extended through multimodal and sensor-integrated paradigms that leverage diverse data modalities including audio, skeleton, and wearable sensors.

Fig.1 illustrates this taxonomy, highlighting how foundational techniques expand toward more complex multimodal systems.

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

E. SSL for Human Activity Recognition with Sensors

Sensor-based HAR applies SSL techniques to time-series data from wearable devices. This is especially relevant in healthcare, smart homes, and low-vision scenarios.

Surveys by Logacjov [18] and Yuan et al. [19] laid the groundwork for SSL in sensor streams. Models like Skeleton2ve [28] , SCD-Net [28] , and STARS [30] applied masked modeling, contrastive tuning, and disentanglement strategies to skeleton sequences for unsupervised action classification.

Improving upon augmentation strategies, 3s-AimCLR++ [31] showed that strong transformations can improve discriminability in self-supervised skeleton learning. These approaches demonstrate that the SSL paradigm is adaptable beyond visual inputs.

F. Emerging Trends and Open Challenges

While SSL has significantly advanced HAR, key challenges remain. These include the lack of benchmark consistency, limited explainability, and scalability to real-time tasks. Test-time adaptation strategies like ST2ST [32] aim to make models flexible to unseen data distributions without retraining.

Recent works like [33] advocate for unified SSL frameworks that bridge task families, promoting transferability between classification, segmentation, and retrieval. Meanwhile, modality-agnostic designs [34] and contextual contrastive schemes [35] continue to redefine the limits of self-supervised action recognition. Table I lists main models, their contributions, and types.

TABLE I SUMMARY OF KEY SELF-SUPERVISED MODELS FOR ACTION RECOGNITION, GROUPED BY LEARNING PARADIGM.

Model / Method	**Key Self-Supervised Models**	
	Type	***Key Contribution***
ViViT [17]	Transformer (Supervised)	Introduced a pure Transformer-based architecture for video classification by treating videos as a sequence of image patches, enabling scalable and parallel learning.
TimeSformer [16]	Transformer (Supervised)	Proposed space-time factorized attention to efficiently model temporal and spatial dynamics separately, reducing complexity.
VideoMAE [4]	Masked Modeling (SSL)	Pioneered masked video pretraining using high-ratio spatiotemporal masking and reconstruction with lightweight decoders, enabling data-efficient learning.
ST-MAE [7]	Masked Modeling (SSL)	Extended MAE to explicitly mask both temporal and spatial patches, improving temporal learning in self-supervised video modeling.
SimCLR [6]	Contrastive (Image SSL)	Introduced contrastive learning with strong augmentations, showing that instance discrimination can yield powerful representations.
MoCo [9]	Contrastive (Image SSL)	Used a dynamic memory bank and momentum encoder to stabilize contrastive learning, facilitating scalable training.
BYOL [5]	Contrastive (Image SSL)	Introduced a predictor-head and moving average target network to learn representations without negative samples.
CVRL [36]	Contrastive (Video SSL)	Applied contrastive learning to videos using temporal augmentations like cropping and shuffling to enforce temporal coherence.

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Model / Method	Key Self-Supervised Models	
	Type	***Key Contribution***
TCL [5]	Temporal Contrastive	Proposed a pretext task of temporal transformation classification to encode the temporal structure in videos.
XDC [13]	Multimodal Contrastive	Aligned visual and audio modalities from uncurated videos using contrastive objectives, improving robustness.
MIL-NCE [11]	Multimodal Contrastive	Introduced generalized cross-modal contrastive learning using arbitrary positive pairings across modalities.
MaCLR [12]	Motion-Aware Contrastive	Incorporated motion-aware sampling strategies for contrastive learning, improving motion-sensitive representation learning.
Part-Aware Contrastive [23]	Body-Region Contrastive	Focused on discriminative body parts in skeleton data to enhance self-supervised recognition of fine-grained actions.
Actionlet Contrastive [33]	Skeleton-based SSL	Modeled egocentric and skeleton joint groupings ("actionlets") to improve unsupervised action recognition in skeleton data.
MVD [25]	Motion-Focused Masked	Combined masked modeling with knowledge distillation and motion-specific representations to enhance temporal sensitivity.
Prompted Masked [21]	Masked Modeling + Prompt	Introduced learnable prompts to guide the reconstruction of motion-relevant parts in masked modeling.
MMV [15]	Multimodal SSL	Unified multiple modalities (vision, audio) through self-supervised learning with temporal synchronization.
SSL for Sensors (Luo) [18]	Sensor SSL	Surveyed self-supervised techniques applied to accelerometer and gyroscope data for wearable HAR.
SSL for Sensors (Li) [19]	Sensor SSL	Presented large-scale training on wearable datasets using self-supervised learning to improve transferability.
SSL Survey (Jing) [3]	General SSL Survey	Provided a comprehensive taxonomy of SSL techniques in computer vision across generative and contrastive methods.
SSL Survey (Liu) [20]	Video SSL Survey	Introduced a pure Transformer-based architecture for video classification by treating videos as a sequence of image patches, enabling scalable and parallel learning.

III. Self-supervised models for action recognition

Self-supervised learning (SSL) is now central to advancing video-based action recognition. Rather than relying on manually labeled datasets, SSL leverages proxy tasks to extract patterns from raw, unlabeled video. These approaches fall broadly into three families: contrastive learning, masked video modeling, and multimodal or sensor-based learning. This section highlights key models, underlying mechanisms, and how recent advances across these categories are pushing the state of the art forward [3], [20], [33], [34].

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A. Contrastive Learning Approaches

Contrastive learning builds representations by pulling together different views of the same data and pushing apart unrelated examples. Early breakthroughs like SimCLR [6], MoCo [9], and BYOL [10] from the image domain laid the foundation. In video, models such as TCL [5] and CVRL [36] adapted these ideas to spatiotemporal tasks using augmentations like frame permutations and temporal jittering.

To better capture motion, MaCLR [12] introduced motion-guided contrastive sampling. Part-Aware Contrastive Learning [23] directed attention to discriminative body parts, improving fine-grained recognition. Actionlet-Contrastive [35] extended this to skeleton sequences, representing small coordinated joint motions critical in subtle actions.

Cross-modal methods like MIL-NCE [11], XDC [13], and AVID [14] align visual and audio cues to guide learning. More recently, Actor-Aware Contrastive Learning [37] and Audio-Visual Synchronization Transformers [38] have improved modality fusion, particularly in cluttered or ambiguous environments. Hierarchical Motion Learning [39] also shows promise by modeling multi-scale temporal consistency.

While contrastive methods are conceptually elegant, they often require large memory banks and careful negative sampling. Recent trends like Meta Contrastive Pretraining [40] aim to stabilize training and increase robustness for real-world applications.

B. Masked Video Modeling

Masked modeling takes a different route teaching the model to reconstruct occluded or removed segments of video. Inspired by BERT and its variants in NLP, models like VideoMAE [4] and ST-MAE [7] have been pivotal in adapting this approach to spatiotemporal data.

Prompted Masked Modeling [21] adds learnable prompts to help the model generalize across domains. Masked Video Distillation (MVD) [25] and SIGMA [24] use teacher-student frameworks and transport-based alignments to enhance training signals. Other approaches like MGMAE [41] and ActionFormerSSL [42] incorporate temporal structure and localization capabilities, making them suitable for tasks beyond classification.

The field continues to explore ways to optimize masking strategies. Adaptive Masking [41] and Sample-Efficient Token Selection [41] improve reconstruction quality and training efficiency. SSL for Long-Horizon Actions [44] tackles sparsity and attention challenges over extended sequences.

These techniques integrate seamlessly with transformers such as ViViT [17] and TimeSformer [16], which excel at modeling long-range dependencies. Hybrid models like HybridSSL [45] and SSL Unification [20] combine masked and contrastive objectives, seeking to capitalize on both paradigms' strengths.

C. Multimodal and Sensor-Based Learning

In many scenarios, visual input alone is not enough. SSL techniques that leverage additional data like audio, skeleton joints, or wearable sensor streams have shown great potential.MMV [15] and AVID [26] train networks to align visual frames with audio tracks. These models are especially useful for recognizing similar visual gestures that sound distinct, such as "typing" versus "playing piano." Extensions like CMAE-V [46] further unify video and language representations under contrastive-masked frameworks.

For skeleton-based SSL, Skeleton2Vec [28], SCD-Net [47], STARS [30], and SSL-AimCLR++ [31] use structural priors from joint positions to model human movement with minimal labels. These methods are often applied to domains like sports analytics and rehabilitation monitoring.

In sensor-based HAR, MetaHAR [45], SSL-MotionSense [19], and HARFormer [48] apply SSL to IMU, accelerometer, or smartphone data. These systems achieve strong performance in real-world, privacy-sensitive environments. Efforts like Cross-Domain SSL [49] and Modality-Agnostic SSL [23] bridge modalities, allowing models trained on wearable data to transfer to video or vice versa [50]. To

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

better evaluate these models, Benchmarking SSL for HAR [40] has emerged as a critical step, while Test-Time Adaptation in SSL [51] enables real-world robustness when environments change post-training.

D. Summary

Across contrastive, masked, and multimodal strategies, self-supervised learning continues to reshape action recognition. Early models focused on spatial discrimination and proxy tasks, but newer methods increasingly leverage motion, cross-modal cues, and temporal structure to learn richer representations. The trend toward hybrid models, sensor integration, and benchmark unification indicates a maturing field poised for broader deployment. Table 1 above provides a side-by-side overview of the representative models referenced in this section.

IV. Open challenges and future directions

Despite the remarkable progress in self-supervised learning (SSL) for action recognition, several open challenges remain both technical and practical. This section outlines persistent limitations observed in current approaches and highlights promising research directions that may guide the next generation of models.

A. Robustness Across Real-World Variability

Many existing models focus on short-term clips with simple actions. However, real-life behaviors often unfold over longer temporal windows and involve subtle cues. Recognizing differences between "waving hello" and "signaling for help" requires understanding intent, motion duration, and context.

Models such as SSL for Long-Horizon Actions [44] and Temporal Cycle Learning [52] tackle this by incorporating long-range attention mechanisms and cyclic sequence modeling. Future research may benefit from memory-augmented transformers or hierarchical temporal reasoning architectures that bridge short- and long-term dependencies.

B. Fine-Grained and Long-Horizon Action Understanding

While models like VideoMAE [4] and ST-MAE [11] have improved our ability to capture short to mid-range motion, most SSL architectures still underperform when it comes to actions that unfold over longer sequences. Events such as “assembling a shelf” or “performing CPR” involve temporally extended dependencies and multi-step interactions. Addressing this may require multi-resolution architectures or memory-enhanced transformers that can model temporal abstractions beyond a few seconds.

C. Unified Multimodal Learning and Fusion

While several models incorporate audio, skeleton, or wearable sensor data such in [15], [29], and [46], most SSL methods still treat each modality in isolation. Yet, real-world perception is inherently multimodal [53].

There’s a need for unified pretraining frameworks that can handle asynchronous, missing, or noisy modalities. Works like CMAE-V [46] and Modality-Agnostic SSL [23] represent a shift toward this goal. Future research should explore modality-aware masking, dynamic fusion strategies, and transformer architectures capable of modality alignment even under weak or indirect supervision [54].

D. Evaluation Protocols and Benchmark Gaps

A key bottleneck in SSL research is the inconsistency in evaluation. Varying pretext tasks, backbone models, and downstream classifiers make it hard to compare methods fairly. Some works report top-1 accuracy, others report transfer learning metrics, and very few assess robustness or generalizability.

The introduction of meta-benchmarking efforts like Benchmarking SSL for HAR [40] is a valuable step. However, the community needs more standardized evaluation protocols, including:

- Generalization across domains (e.g., lab → wild)
- Few-shot or semi-supervised transfer performance

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- Efficiency metrics: pretraining time, memory footprint
- Such evaluations will help align SSL development with practical deployment needs.

E. Toward Real-Time and Edge Deployment

Many SSL models especially transformer-based ones like ViViT [17], ActionFormerSSL [42], or SIGMA [24] are compute-intensive. This limits their adoption on mobile or embedded platforms, where latency and power efficiency are crucial.

Recent works in sample-efficient masking [43] and lightweight temporal modeling [47] offer initial solutions.

Going forward, there's a growing need for hardware-aware SSL design, pruning strategies, and edge-optimized self-supervised pipelines.

F. Explainability and Trust in SSL Models

Unlike supervised models with labeled classes, SSL models learn latent spaces with no predefined meaning. This opacity makes them harder to interpret and trust especially in safety-critical domains like healthcare or security.

There is a gap in visualization tools, saliency-based explanation, and latent space probing techniques tailored for SSL. Models like FILS [34] and MetaHAR [47] begin to explore this, but further work is needed to integrate explainability into the design and training process itself.

G. Summary

The road ahead for SSL in action recognition is both exciting and challenging. As models become more generalizable, multimodal, and efficient, the key will be balancing complexity with practicality. To reach real-world deployment, future efforts must go beyond accuracy and tackle robustness, fairness, interpretability, and domain adaptation. Table II summarize key challenges and future directions in SSL.

By addressing these gaps, SSL can move from research labs into everyday applications empowering systems that understand human actions across contexts, devices, and cultures.

TABLE II KEY CHALLENGES IN SELF-SUPERVISED LEARNING FOR ACTION RECOGNITION, INCLUDING CONTRIBUTING FACTORS AND PROPOSED FUTURE RESEARCH DIRECTIONS TO IMPROVE GENERALIZATION, EFFICIENCY, AND INTERPRETABILITY.

Challenge	**Contributing Factors**	
	Description	***Future Directions***
Robustness in Real-World Environments	Sensitivity to occlusion, noise, lighting variations, and domain shifts [51]	Develop domain-adaptive SSL models, explore zero-shot learning techniques, and incorporate uncertainty modeling.
Long-Horizon and Fine-Grained Actions	Difficulty modeling extended temporal sequences and subtle inter-class variations [44]	Design memory-augmented transformers, use hierarchical attention, and explore multi-resolution temporal abstractions.
Unified Multimodal Learning	Separate treatment of modalities like audio, video, skeleton, and sensors [15], [23], [46]	Develop modality-agnostic encoders, dynamic fusion mechanisms, and unified pretraining pipelines for missing or noisy modalities.
Inconsistent Evaluation Benchmarks	Lack of standard protocols and inconsistent downstream tasks across papers [40]	Propose meta-benchmarking suites, standardized tasks (classification, retrieval),

Journal of Information Systems Engineering and Management
2025, 10(48s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Challenge	Contributing Factors	
	Description	***Future Directions***
		and unified performance metrics (efficiency, generalization).
Real-Time and Edge Deployment	High computation and memory requirements of transformer-based SSL models [43], [47]	Employ efficient token pruning, model distillation, hardware-aware neural architecture search (NAS), and edge-optimized SSL frameworks.
Explainability and Interpretability	SSL models often learn abstract latent features without human-understandable labels [34], [47]	Integrate saliency maps, attention visualization, and develop interpretable latent probes or post-hoc explanation modules.

V. Conclusion

Self-supervised learning has emerged as a transformative paradigm for video action recognition enabling models to learn from vast amounts of unlabeled data through pretext tasks that capture both spatial and temporal dynamics. As this survey has shown, SSL approaches have evolved from early contrastive techniques to sophisticated masked modeling and multimodal learning frameworks, with applications extending to skeleton sequences, wearable sensors, and real-world human activity understanding.

Through our structured review, we highlighted foundational models such as VideoMAE, TimeSformer, and ViViT, while also surfacing recent advances in motion-guided learning, prompt-based modeling, and hybrid SSL objectives. These methods demonstrate encouraging progress in reducing reliance on manual labels, improving generalization, and expanding to diverse modalities.

However, as powerful as these techniques are, they are not without limitations. Current challenges ranging from domain robustness and temporal reasoning to edge deployment and explainability continue to motivate ongoing research. The field is shifting from lab-constrained benchmarks toward real-world deployments, where SSL must not only perform well but also adapt, scale, and be interpretable.

Looking ahead, we see promising directions in multimodal fusion, unified SSL frameworks, and task-agnostic pretraining strategies. Standardized benchmarks, interpretability tools, and hardware-efficient designs will be essential to guide the community toward more robust and deployable solutions.

By synthesizing recent developments and surfacing open questions, this review aims to support researchers and practitioners working at the intersection of deep learning and human action understanding. As SSL continues to mature, it offers a foundation for truly intelligent systems that can learn from the world without relying on labels to make sense of it.

References

[1] K. Soomro, A. R. Zamir, and M. Shah, “UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild,” Dec. 03, 2012, *arXiv*: arXiv:1212.0402. doi: 10.48550/arXiv.1212.0402.

[2] W. Kay *et al.*, “The Kinetics Human Action Video Dataset,” May 19, 2017, *arXiv*: arXiv:1705.06950. doi: 10.48550/arXiv.1705.06950.

[3] X. Liu *et al.*, “Self-supervised Learning: Generative or Contrastive,” *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2021, doi: 10.1109/TKDE.2021.3090866.

[4] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 10078–10093, 2022.

[5] S. Jenni, G. Meishvili, and P. Favaro, "Video Representation Learning by Recognizing Temporal Transformations," in *Computer Vision – ECCV 2020*, vol. 12373, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., in Lecture Notes in Computer Science, vol. 12373. , Cham: Springer International Publishing, 2020, pp. 425–442. doi: 10.1007/978-3-030-58604-1_26.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PmLR, 2020, pp. 1597–1607. Accessed: Apr. 12, 2025. [Online]. Available: http://proceedings.mlr.press/v119/chen20j.html

[7] C. Feichtenhofer, Y. Li, and K. He, "Masked autoencoders as spatiotemporal learners," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 35946–35958, 2022.

[8] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660. Accessed: Apr. 12, 2025. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper

[9] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738. Accessed: Apr. 12, 2025. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2020/html/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.html

[10] J.-B. Grill *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 21271–21284, 2020.

[11] M. Patrick *et al.*, "Multi-modal self-supervision from generalized data transformations," 2020, Accessed: Apr. 12, 2025. [Online]. Available: https://openreview.net/forum?id=mgVbI13p96

[12] F. Xiao, J. Tighe, and D. Modolo, "MaCLR: Motion-Aware Contrastive Learning of Representations for Videos," in *Computer Vision – ECCV 2022*, vol. 13695, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., in Lecture Notes in Computer Science, vol. 13695. , Cham: Springer Nature Switzerland, 2022, pp. 353–370. doi: 10.1007/978-3-031-19833-5_21.

[13] H. Alamri, A. Bilic, M. Hu, A. Beedu, and I. Essa, "End-to-End Multimodal Representation Learning for Video Dialog," Oct. 26, 2022, *arXiv*: arXiv:2210.14512. doi: 10.48550/arXiv.2210.14512.

[14] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *Int. J. Comput. Vis.*, vol. 79, no. 3, pp. 299–318, Sep. 2008, doi: 10.1007/s11263-007-0122-4.

[15] J.-B. Alayrac *et al.*, "Self-supervised multimodal versatile networks," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 25–37, 2020.

[16] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *ICML*, 2021, p. 4. Accessed: Apr. 12, 2025. [Online]. Available: https://proceedings.mlr.press/v139/bertasius21a/bertasius21a-supp.pdf

[17] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846. Accessed: Apr. 12, 2025. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2021/html/Arnab_ViViT_A_Video_Vision_Transformer_ICCV_2021_paper.html?ref=https://githubhelp.com

[18]A. Logacjov, "Self-supervised Learning for Accelerometer-based Human Activity Recognition: A Survey," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 4, pp. 1–42, Nov. 2024, doi: 10.1145/3699767.

[19]H. Yuan *et al.*, "Self-supervised learning for human activity recognition using 700,000 person-days of wearable data," *NPJ Digit. Med.*, vol. 7, no. 1, p. 91, 2024.

[20] M. C. Schiappa, Y. S. Rawat, and M. Shah, "Self-Supervised Learning for Videos: A Survey," *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 1–37, Dec. 2023, doi: 10.1145/3577925.

[21]J. Zhang, L. Lin, and J. Liu, "Prompted Contrast with Masked Motion Modeling: Towards Versatile 3D Action Representation Learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, Ottawa ON Canada: ACM, Oct. 2023, pp. 7175–7183. doi: 10.1145/3581783.3611774.

[22] R. Zellers *et al.*, "Merlot: Multimodal neural script knowledge models," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 23634–23651, 2021.

[23] Y. Hua *et al.*, "Part Aware Contrastive Learning for Self-Supervised Action Recognition," May 11, 2023, *arXiv*: arXiv:2305.00666. doi: 10.48550/arXiv.2305.00666.

[24] M. Salehi, M. Dorkenwald, F. M. Thoker, E. Gavves, C. G. M. Snoek, and Y. M. Asano, "SIGMA: Sinkhorn-Guided Masked Video Modeling," in *Computer Vision – ECCV 2024*, vol. 15082, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., in Lecture Notes in Computer Science, vol. 15082. , Cham: Springer Nature Switzerland, 2025, pp. 293–312. doi: 10.1007/978-3-031-72691-0_17.

[25] R. Wang *et al.*, "Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6312–6322. Accessed: Apr. 13, 2025. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2023/html/Wang_Masked_Video_Distillation_Rethinking_Masked_Feature_Modeling_for_Self-Supervised_Video_CVPR_2023_paper.html

[26] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-visual instance discrimination with cross-modal agreement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12475–12486. Accessed: Apr. 13, 2025. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2021/html/Morgado_Audio-Visual_Instance_Discrimination_with_Cross-Modal_Agreement_CVPR_2021_paper.html

[27] H. Akbari *et al.*, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 24206–24221, 2021.

[28] R. Xu, L. Huang, M. Wang, J. Hu, and W. Deng, "Skeleton2vec: A Self-supervised Learning Framework with Contextualized Target Representations for Skeleton Sequence," Jan. 01, 2024, *arXiv*: arXiv:2401.00921. doi: 10.48550/arXiv.2401.00921.

[29] C. Wu *et al.*, "Scd-net: Spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, 2024, pp. 5949–5957. Accessed: Apr. 13, 2025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/28409

[30] S. Mehraban, M. J. Rajabi, and B. Taati, "STARS: Self-supervised Tuning for 3D Action Recognition in Skeleton Sequences," Jul. 15, 2024, *arXiv*: arXiv:2407.10935. doi: 10.48550/arXiv.2407.10935.

[31]T. Guo, M. Liu, H. Liu, G. Wang, and W. Li, "Improving self-supervised action recognition from extremely augmented skeleton sequences," *Pattern Recognit.*, vol. 150, p. 110333, 2024.

[32] M. A.-N. I. Fahim, M. Innat, and J. Boutellier, "ST2ST: Self-Supervised Test-time Adaptation for Video Action Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, 2024, pp. 1057–1066. Accessed: Apr. 13, 2025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024W/MAT/html/Fahim_ST2ST_Self-Supervised_Test-time_Adaptation_for_Video_Action_Recognition_CVPRW_2024_paper.html

[33] I. Dave *et al.*, "Unifying Video Self-Supervised Learning across Families of Tasks: A Survey," *Preprints*, 2024, Accessed: Apr. 13, 2025. [Online]. Available: https://www.preprints.org/frontend/manuscript/a4b3be9e108ca235884169651f940807/download_pub

[34] J. Zhang, L. Lin, S. Yang, and J. Liu, "Self-Supervised Skeleton-Based Action Representation Learning: A Benchmark and Beyond," Aug. 26, 2024, *arXiv*: arXiv:2406.02978. doi: 10.48550/arXiv.2406.02978.

[35] L. Lin, J. Zhang, and J. Liu, "Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2363–2372. Accessed: Apr. 13, 2025. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2023/html/Lin_Actionlet-Dependent_Contrastive_Learning_for_Unsupervised_Skeleton-Based_Action_Recognition_CVPR_2023_paper.html

[36] R. Qian *et al.*, "Spatiotemporal contrastive video representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6964–6974. Accessed: Apr. 13, 2025. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2021/html/Qian_Spatiotemporal_Contrastive_Video_Representation_Learning_CVPR_2021_paper.html

[37] M. Assefa, W. Jiang, K. Gedamu, G. Yilma, M. Ayalew, and M. Seid, "Actor-aware contrastive learning for semi-supervised action recognition," in *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, 2022, pp. 660–665. Accessed: Apr. 13, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10097985/

[38] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman, "Audio-Visual Synchronisation in the wild," Dec. 08, 2021, *arXiv*: arXiv:2112.04432. doi: 10.48550/arXiv.2112.04432.

[39] J. Zhang, L. Lin, and J. Liu, "Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 3427–3435. Accessed: Apr. 13, 2025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/25451

[40] S. Enomoto *et al.*, "Test-time adaptation meets image enhancement: Improving accuracy via uncertainty-aware logit switching," in *2024 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2024, pp. 1–8. Accessed: Apr. 13, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10650964/

[41] B. Huang, Z. Zhao, G. Zhang, Y. Qiao, and L. Wang, "Mgmae: Motion guided masking for video masked autoencoding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13493–13504. Accessed: Apr. 13, 2025. [Online]. Available: http://openaccess.thecvf.com/content/ICCV2023/html/Huang_MGMAE_Motion_Guided_Masking_for_Video_Masked_Autoencoding_ICCV_2023_paper.html

[42] C.-L. Zhang, J. Wu, and Y. Li, "ActionFormer: Localizing Moments of Actions with Transformers," in *Computer Vision – ECCV 2022*, vol. 13664, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., in Lecture Notes in Computer Science, vol. 13664. , Cham: Springer Nature Switzerland, 2022, pp. 492–510. doi: 10.1007/978-3-031-19772-7_29.

[43] Y. Feng, Y. Shi, F. Liu, and T. Yan, "Motion Guided Token Compression for Efficient Masked Video Modeling," Jan. 10, 2024, *arXiv*: arXiv:2402.18577. doi: 10.48550/arXiv.2402.18577.

[44] H. Lin *et al.*, "VEDIT: Latent Prediction Architecture For Procedural Video Representation Learning," *ArXiv Prepr. ArXiv241003478*, 2024, Accessed: Apr. 13, 2025. [Online]. Available: https://arxiv.org/abs/2410.03478

[45] L. Wang *et al.*, "Videomae v2: Scaling video masked autoencoders with dual masking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14549–14560. Accessed: Apr. 13, 2025. [Online]. Available: http://openaccess.thecvf.com/content/CVPR2023/html/Wang_VideoMAE_V2_Scaling_Video_Masked_Autoencoders_With_Dual_Masking_CVPR_2023_paper.html

[46] C.-Z. Lu, X. Jin, Z. Huang, Q. Hou, M.-M. Cheng, and J. Feng, "CMAE-V: Contrastive Masked Autoencoders for Video Action Recognition," Jan. 15, 2023, *arXiv*: arXiv:2301.06018. doi: 10.48550/arXiv.2301.06018.

[47] C. Li, D. Niu, B. Jiang, X. Zuo, and J. Yang, "Meta-HAR: Federated Representation Learning for Human Activity Recognition," in *Proceedings of the Web Conference 2021*, Ljubljana Slovenia: ACM, Apr. 2021, pp. 912–922. doi: 10.1145/3442381.3450006.

[48] S. Suh, V. F. Rey, and P. Lukowicz, "Tasked: transformer-based adversarial learning for human activity recognition using wearable sensors via self-knowledge distillation," *Knowl.-Based Syst.*, vol. 260, p. 110143, 2023.

[49] X. He, J. Wang, Q. Xia, G. Lu, Y. Tang, and H. Lu, "Cross-Domain Feature Semantic Calibration for Zero-Shot Sketch-Based Image Retrieval," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2024, pp. 1–6. Accessed: Apr. 13, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10687519/

[50] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2630–2640. Accessed: Apr. 13, 2025. [Online]. Available: http://openaccess.thecvf.com/content_ICCV_2019/html/Miech_HowTo100M_Learning_a_Text-Video_Embedding_by_Watching_Hundred_Million_Narrated_ICCV_2019_paper.html

[51] Y. He, A. Carass, L. Zuo, B. E. Dewey, and J. L. Prince, "Autoencoder based self-supervised test-time adaptation for medical image analysis," *Med. Image Anal.*, vol. 72, p. 102136, 2021.

[52] I. Hadji, K. G. Derpanis, and A. D. Jepson, "Representation Learning via Global Temporal Alignment and Cycle-Consistency," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 11063–11072. doi: 10.1109/CVPR46437.2021.01092.

[52] Ahmad, A. Y. B., Kumari, D. K., Shukla, A., Deepak, A., Chandnani, M., Pundir, S., & Shrivastava, A. (2024). Framework for Cloud Based Document Management System with Institutional Schema of Database. International Journal of Intelligent Systems and Applications in Engineering, 12(3s), 672-678.

[52] Ahmad, A. Y. Bani ahmad , (2019). Empirical Analysis on Accounting Information System Usage in Banking Sector in Jordan. Academy of Accounting and Financial Studies Journal, 23(5), 1-9.

[53] Alhawamdeh, H., Al-Saad, S. A., Almasarweh, M. S., Al-Hamad, A. A.-S. A., Bani Ahmad, A. Y. A. B., & Ayasrah, F. T. M. (2023). The Role of Energy Management Practices in Sustainable Tourism Development: A Case Study of Jerash, Jordan. International Journal of Energy Economics and Policy, 13(6), 321–333. https://doi.org/10.32479/ijeep.14724

[54] Allahham, M., & Ahmad, A. (2024). AI-induced anxiety in the assessment of factors influencing the adoption of mobile payment services in supply chain firms: A mental accounting perspective. International Journal of Data and Network Science, 8(1), 505-514.

[55] K. Daoud, D. . Alqudah, M. . Al-Qeed, B. A. . Al Qaied, and A. Y. A. B. . Ahmad, "The Relationship Between Mobile Marketing and Customer Perceptions in Jordanian Commercial Banks: The Electronic Quality as A Mediator Variable", ijmst, vol. 10, no. 2, pp. 1360-1371, Jun. 2023

[56] Kai, Z., Sharaf, M., Wei, S. Y., Al Shraah, A., Le, L. T., Bedekar, A. A., & Ahmad, A. Y. B. (2024). Exploring the asymmetric relationship between natural resources, fintech, remittance and environmental pollution for BRICS nations: New insights from MMQR approach. Resources Policy, 90, 104693

[57] Liang, P., Guo, Y., Nutakki, T. U. K., Agrawal, M. K., Muhammad, T., Ahmad, S. F., ... & Qin, M. (2024). Comprehensive assessment and sustainability improvement of a natural gas power plant utilizing an environmentally friendly combined cooling heating and power-desalination arrangement. Journal of Cleaner Production, 436, 140387.

[58] Liang, P., Guo, Y., Chauhdary, S. T., Agrawal, M. K., Ahmad, S. F., Ahmad, A. Y. A. B., ... & Ji, T. (2024). Sustainable development and multi-aspect analysis of a novel polygeneration system using biogas upgrading and LNG regasification processes, producing power, heating, fresh water and liquid CO2. Process Safety and Environmental Protection, 183, 417-436..

[59] Mohsin, H. J., Hani, L. Y. B., Atta, A. A. B., Al-Alawnh, N. A. K., Ahmad, A. B., & Samara, H. H. (2023). The impact of digital financial technologies on the development of entrepreneurship: evidence from commercial banks in the emerging markets. Corporate & Business Strategy Review, 4(2), 304-312.

[60] Ramadan, A., Alkhodary, D., Alnawaiseh, M., Jebreen, K., Morshed, A., & Ahmad, A. B. (2024). Managerial Competence and Inventory Management in SME Financial Performance: A Hungarian Perspective. Journal of Statistics Applications & Probability, 13(3), 859-870.

[61] Almestarihi, R., Ahmad, A. Y. A. B., Frangieh, R., Abu-AlSondos, I., Nser, K., & Ziani, A. (2024). Measuring the ROI of paid advertising campaigns in digital marketing and its effect on business profitability. Uncertain Supply Chain Management, 12(2), 1275-1284.

[62] Daoud, M. K., Al-Qeed, M., Al-Gasawneh, J. A., & Bani Ahmad, A. Y. (2023). The Role of Competitive Advantage Between Search Engine Optimization and Shaping the Mental Image of Private Jordanian University Students Using Google. International Journal of Sustainable Development & Planning, 18(8).

[67] Yahiya Ahmad Bani Ahmad (Ayassrah), Ahmad; Ahmad Mahmoud Bani Atta, Anas; Ali Alawawdeh, Hanan; Abdallah Aljundi, Nawaf; Morshed, Amer; and Amin Dahbour, Saleh (2023) "The Effect of System Quality and User Quality of Information Technology on Internal Audit Effectiveness in Jordan, And the Moderating Effect of Management Support," Applied Mathematics & Information Sciences: Vol. 17: Iss. 5, Article 12.

[68] C. Verma, V. P, N. Chaturvedi, U. U, A. Rai and A. Y. A. Bani Ahmad, "Artificial Intelligence in Marketing Management: Enhancing Customer Engagement and Personalization," 2025 International Conference on Pervasive Computational Technologies (ICPCT), Greater Noida, India, 2025, pp. 397-401, doi: 10.1109/ICPCT64145.2025.10940626.

[69] N. Parihar, P. Fernandes, S. Tyagi, A. Tyagi, M. Tiwari and A. Y. A. Bani Ahmad, "Using Machine Learning to Enhance Cybersecurity Threat Detection," 2025 International Conference on Pervasive Computational Technologies (ICPCT), Greater Noida, India, 2025, pp. 387-391, doi: 10.1109/ICPCT64145.2025.10939232.

[70] A. Y. A. Bani Ahmad, P. Sarkar, B. Goswami, P. R. Patil, K. Al-Said and N. Al Said, "A Framework for Evaluating the Effectiveness of Explainability Methods in Deep Learning," 2025 International Conference on Pervasive Computational Technologies (ICPCT), Greater Noida, India, 2025, pp. 426-430, doi: 10.1109/ICPCT64145.2025.10939073.

[71] Alhawamdeh, H., Abdel Muhsen Irsheid Alafeef, M., Abdel Mohsen Al-Afeef, M., Alkhawaldeh, B. Y., Nawasra, M., Al_Rawashdeh, H. A. A., ... & Al-Eitan, G. N. (2024). The relationship between marketing capabilities and financial performance: the moderating role of customer relationship management in Jordanian SMES. Cogent Business & Management, 11(1), 2297458.

[72] Alamad, T., Alrawashedh, N. H., Alhawamdeh, H., Harahsheh, A. A., Zraqat, O., Hussien, L. F., ... & Alkhawaldeh, B. Y. (2024). The Impact of Strategic Leadership on Strategic Performance in Higher Education Institutions: The Mediating Role of Change Management.

[73] Alhawamdeh, H., Alkhawaldeh, B. Y., Zraqat, O., & Alhawamdeh, A. M. (2024). Leveraging Business Intelligence in Organizational Innovation: A Leadership Perspective in Commercial Banks. International Journal of Academic Research in Accounting, Finance and Management Sciences, 14(1), 295-309.

[74] Alhawamdeh, A. M., Al-habash, M. A., Zraqat, O., Hussien, L. F., Taha, I. B., Alhawamdeh, H., & Alkhawaldeh, B. Y. (2023). The Effect of Religious and Ethnic Values on Executive Compensation in Jordanian Firms. KEPES, 21(3), 604-622.

[75] Alkhawaldeh, B. Y. S., Alhawamdeh, H., Almarshad, M., Fraihat, B. A. M., Abu-Alhija, S. M. M., Alhawamdeh, A. M., & Ismaeel, B. (2023). The effect of macroeconomic policy uncertainty on environmental quality in Jordan: Evidence from the novel dynamic simulations approach. Jordan Journal of Economic Sciences, 10(2), 116-131.

[76] Al-Afeef, M. A. M., Fraihat, B. A. M., Alhawamdeh, H., Hijazi, H. A., AL-Afeef, M. A., Nawasr, M., & Rabi, A. M. (2023). Factors affecting middle eastern countries' intention to use financial technology. International Journal of Data & Network Science, 7(3).

[77] Alhawamdeh, H., Al-Saad, S. A., Almasarweh, M. S., Al-Hamad, A. A. S., Ahmad, A. Y., & Ayasrah, F. T. M. (2023). The role of energy management practices in sustainable tourism development: a case study of Jerash, Jordan. International Journal of Energy Economics and Policy, 13(6), 321-333.

[78] Alkhawaldeh, B. Y., Alhawamdeh, H., Al_Shukri, K. S., Yousef, M., Shehadeh, A. Y. A., Abu-Samaha, A. M., & Alwreikat, A. A. (2023). The role of technological innovation on the effect of international strategic alliances on corporate competitiveness in Jordanian international business administration: Moderating and mediating analysis. Migration Letters, 20(6), 282-299.

[79] Alhawamdeh, H., Al-Eitan, G. N., Hamdan, M. N., Al-Hayek, Y. A. M., Zraqat, O., Alhawamdeh, A. M., & Alkhawaldeh, B. Y. (2023). The role of financial risk tolerance and financial advisor management in mediating the relationship between financial attitudes, financial knowledge, financial anxiety, and sustainable financial retirement planning. Journal of Namibian Studies: History Politics Culture, 33, 5071-5100.

[63] Alkhawaldeh, B., Alhawamdeh, H., Al-Afeef, M., Al-Smadi, A., Almarshad, M., Fraihat, B., ... & Alaa, A. (2023). The effect of financial technology on financial performance in Jordanian SMEs: The role of financial satisfaction. Uncertain Supply Chain Management, 11(3), 1019-1030.

[64] A lkhawaldeh, B. Y., Alhawamdeh, H., Al-Afeef, M. A. M., Abu-Alhija, S. M. M., Al_Rawashdeh, H. A. A., Mustafa, S. M. B., ... & Almarshad, M. (2023). Mediating effect of financial behaviour on the influence of financial literacy and financial technology on financial inclusion development in Jordanian MSMEs. Journal of Hunan University Natural Sciences, 50(3).

[65] Al-gharaibeh, S. M., Al-Zoubi, D. M., Hijazi, H. A., Al-Sakarneh, A., Alhawamdeh, H. M., & Al-Afee, M. (2021). The Relationship Between E-learning During Coronavirus Pandemic and Job Burnout among Faculty Members in Public and Private Universities in Jordan. International Journal of Academic Research in Business and Social Sciences, 11(11), 1983-2011.

Research Article

[66] Fraihat, B. A. M., Alhawamdeh, H., Alkhawaldeh, B. Y., Abozraiq, A. M., & Al Shaban, A. (2023). The effect of organizational structure on employee creativity: The moderating role of communication flow: A survey study. International Journal of Academic Reserach in Economics and Management Sciences, 12(2).

[52] Fraihat, B. A. M., Alhawamdeh, H., Alkhawaldeh, B. Y., Abozraiq, A. M., & Al Shaban, A. (2023). The effect of organizational structure on employee creativity: The moderating role of communication flow: A survey study. International Journal of Academic Reserach in Economics and Management Sciences, 12(2).

[52] Lehyeh, S. A., Alharafsheh, M., Hanandeh, R., Abuaddous, M., & Al-Hawamdeh, H. (2021). The effects of total quality management practices on strategic performance using the BSC methodology: the mediating role of knowledge sharing. Academy of Strategic Management Journal, 20(6), 1-12

[53] N. Parthasarathy, S. M. Eslami, J. Carreira, and O. Henaff, "Self-supervised video pretraining yields robust and more human-aligned visual representations," Adv. Neural Inf. Process. Syst., vol. 36, pp. 65743–65765, 2023.

[54] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 1728–1738. Accessed: Apr. 13, 2025. [Online]. Available: http://openaccess.thecvf.com/content/ICCV2021/html/Bain_Frozen_in_Time_A_Joint_Video_and_Image_Encoder_for_ICCV_2021_paper.html