Continual Learning Framework for Long-Sequence Video Understanding with Memory-Aware Visual Representations

Kailin Qin
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
kailin735@unr.edu

Rkshay Marayan
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
akshay1981@missouri.edu

Abstract

Continuous video streams generated by autonomous vehicles, surveillance networks, and interactive media platforms demand machine perception systems that sustain high performance over indefinitely long temporal horizons without catastrophic forgetting. Traditional architectures designed for short video clips struggle to reconcile memory efficiency with the preservation of long-range dependencies, while static representations fail to adapt to shifting visual statistics. This paper proposes a continual learning framework that integrates memory-aware visual representations for long-sequence video understanding. The framework treats memory as an architectural primitive rather than a passive storage buffer, coupling compressible episodic representations with dynamic parameter allocation strategies. We examine structural trade-offs among external memory banks, generative replay, and elastic weight consolidation in the context of streaming video. The discussion extends to system-level considerations including distributed training infrastructure, privacy-preserving federated adaptation across distributed cameras, energy-aware deployment on edge devices, and governance mechanisms for accountability in continuously evolving perception pipelines. Through a detailed analysis of architecture robustness, fairness under non-stationary class distributions, and policy implications for monitored public spaces, we argue that memory-awareness must be elevated from a component-level optimization to a first-class design principle. The paper articulates a holistic research agenda in which visual representations, learning dynamics, and operational infrastructure co-evolve to support trustworthy, sustainable continual video intelligence.

Keywords

continual learning, video understanding, memory-augmented networks, long-sequence modeling, system architecture, fairness, edge computing, governance.

1. Introduction

The ability to reason over long video sequences has moved from a niche capability to a foundational requirement across safety-critical and large-scale perceptual systems. Autonomous driving platforms ingest hours of multi-camera footage, urban surveillance networks accumulate petabytes of continuous streams, and assistive technologies for the

visually impaired demand persistent scene understanding that spans entire daily routines. Yet the predominant paradigm in video understanding research has centered on trimmed clips of a few seconds, optimized for action recognition benchmarks and short-term temporal modeling [1,2]. Extending these models to arbitrarily long sequences exposes structural weaknesses that cannot be remedied by merely increasing memory capacity or stacking recurrent layers. The core challenge lies at the intersection of representation learning and continual adaptation: a model must retain discriminative visual concepts acquired early in a sequence while rapidly absorbing novel objects, lighting shifts, or activity patterns without destructive interference.

Continual learning, the subfield concerned with learning from a non-stationary stream of data without forgetting previously acquired knowledge, provides a conceptual toolkit for addressing these challenges [3,4]. However, directly porting existing continual learning methods to video domains reveals a mismatch between the granularity of rehearsal, regularization, or architectural expansion mechanisms and the dense temporal structure of visual frames. Video data exhibit both fine-grained frame-to-frame redundancy and abrupt semantic shifts at scene boundaries, requiring memory mechanisms that operate at multiple temporal scales. A purely episodic memory that replays raw frames is storage-prohibitive and fails to capture compressed abstractions; a purely parametric approach that consolidates synaptic importance, such as elastic weight consolidation, may lose plasticity when faced with radically new visual environments.

This paper advances the thesis that long-sequence video understanding demands a tight coupling between continual learning algorithms and memory-aware visual representations. Rather than treating memory as an external remedial module, we argue for an integrated framework in which the representation itself encodes not only spatial and motion patterns but also a compressed trace of its own formation history. Such representations should be built upon architectures that disentangle what to remember from how to store it, enabling selective refreshment, compositional reuse, and graceful forgetting of transient distractors. The recent interest in memory-augmented transformers, memory networks, and segment-level attention mechanisms suggests that the field is converging toward this insight, though systematic design principles and deployment blueprints remain absent.

This work contributes a synthetic, system-level perspective on the design space, framed around structural trade-offs, infrastructure requirements, and societal implications. We connect representation-level design choices—external key-value memory, differentiable neural caches, memory-aware fine-tuning—with macro-level concerns including distributed edge-fog architectures, federated model updates, carbon-aware training regimes, and fairness auditing under distributional drift. The paper is organized as follows. Section 2 reviews related work on video architectures and continual learning. Section 3 presents the high-level architecture of a continual video understanding framework. Section 4 analyzes memory-aware visual representations. Section 5 examines continual learning strategies adapted to temporal streams. Section 6 discusses implementation infrastructure and scalability. Section 7 addresses evaluation paradigms and robustness. Section 8 explores governance, fairness, and policy implications. Section 9 considers deployment and sustainability, with a concluding synthesis in Section 10.

2. Background and Related Work

Video understanding has progressed through a succession of architectural innovations, from two-stream convolutional networks and 3D convolutions to factorized spatio-temporal attention mechanisms and pure transformer backbones [1,2,5]. Earlier works such as I3D

inflated 2D convolutional kernels into the temporal dimension, achieving strong results on benchmarks but struggling with sequences longer than a few dozen frames due to quadratic computation growth and loss of fine-grained temporal localization. The introduction of the SlowFast pathway decoupled spatial semantics from rapid motion, illustrating that multi-granular temporal processing is a powerful inductive bias that could be generalized to longer sequences if properly stabilized. Vision transformers further dissolved the frame-centric boundaries by operating on space-time patches, enabling global attention across frames, but the absence of explicit memory compression forced practitioners to subsample or truncate input streams, effectively discarding long-range context.

In parallel, continual learning research has produced a rich taxonomy of approaches. Regularization-based methods, exemplified by elastic weight consolidation, protect parameters deemed critical for previous tasks by penalizing large deviations [3]. Architectural methods expand network capacity or isolate task-specific sub-networks, while replay methods maintain an episodic buffer that is interleaved with new data to prevent catastrophic forgetting [4,6]. Memory-aware synapses extended the regularization paradigm by learning an importance weight for each synapse without requiring explicit task boundaries, a property highly desirable for streaming video where task demarcations are ambiguous [6]. Generative replay trains a separate model to produce pseudo-samples of past experiences, obviating the need to store raw data, although the fidelity of generated frames remains a limiting factor for high-resolution video.

Memory-augmented neural networks occupy a middle ground by equipping models with an explicit external memory that can be written to and read from via differentiable addressing mechanisms [7,8]. The original end-to-end memory networks demonstrated that read-write operations over a fixed-size memory matrix could support complex relational reasoning over text, a principle later extended to visual question answering and video question answering. Long Short-Term Memory (LSTM) units can be viewed as a primitive form of gated memory, but their capacity to store fine-grained visual details over hundreds of frames is limited by the fixed hidden state dimensionality and the high ratio of transient motion to stable scene structure [9]. More recently, the Segment Anything Model and its successors have shown that segment-level priors can drastically reduce the search space for object tracking, and memory-aware fine-tuning techniques that inject lightweight adapter modules have been proposed to adapt such large models to long video sequences efficiently [10,11]. These works treat memory not as a monolithic store but as a collection of object-centric and saliency-driven tokens that can be selectively retained, compressed, or discarded according to a learned policy. This line of research motivates the core proposition of this paper: that the visual representation should be co-designed with the memory management policy.

The landscape of long-sequence video analytics is further shaped by systems-oriented contributions that address resource allocation, distributed processing, and real-time constraints. Frameworks for elastic video analytics dynamically balance accuracy against latency by cascading lightweight models on edge nodes with heavier cloud-based detectors, and collaborative approaches enable multiple edge devices to share compressed representations without transmitting raw video, preserving privacy [12]. These system-level strategies have rarely been integrated with continual learning updates, leaving a gap between the agility of online adaptation and the stability demanded by safety-critical deployments.

3. System Architecture of the Continual Video Understanding Framework

A holistic framework for continual long-sequence video understanding must be architected as a modular yet tightly integrated processing pipeline that spans sensing, representation, learning, and decision-making layers. At the sensing layer, multi-camera streams are time-synchronized and preprocessed by adaptive frame selection policies that discard redundant information while preserving semantically salient transitions. Instead of feeding all frames at a fixed rate, a gating module estimates frame novelty based on motion vectors, attention map divergence, and semantic shift detectors, generating a variable-rate input schedule that reduces computational load while maintaining temporal coverage. This adaptive scheduling mechanism forms the first structural trade-off: aggressive subsampling saves energy and communication bandwidth but risks missing brief yet critical events such as a pedestrian suddenly entering a roadway.

The representation layer is responsible for converting selected frames into a memory-aware feature space. The backbone network extracts hierarchical feature maps and passes them through a multi-scale temporal aggregation module that fuses short-term motion cues with long-term context drawn from a persistent memory store. The memory store is partitioned into an episodic buffer that holds compressed representations of recent keyframes and a semantic index that captures prototypical objects, scenes, and activity templates encountered across the entire deployment history. Writes to the memory are governed by a learned controller that assesses the novelty, stability, and future utility of each candidate embedding. As new data arrive, the controller executes a selective overwrite policy that preserves high-utility representations while gradually forgetting spurious or outdated patterns. This design mirrors computational models of the hippocampal-neocortical memory consolidation process, where rapid episodic encoding in a short-term store is followed by slow integration into a more stable, semantically organized long-term store.

The continual learning layer orchestrates the interplay between new data and consolidated knowledge. It receives gradients from downstream task heads and determines the appropriate resource allocation strategy for knowledge integration. During stable regimes where visual statistics change slowly, the system employs a lightweight elastic consolidation mode that inflicts small penalty adjustments on weight updates. Upon detecting a distribution shift, triggered by a statistical change-point detector monitoring reconstruction error or prediction confidence decay, the framework shifts to a meta-learning-based rapid adaptation mode that dynamically reconfigures the sparse set of trainable parameters and temporarily expands the memory index to accommodate new concepts. This layer also manages a conflict resolution mechanism that resolves inconsistencies between the model’s evolving priors and newly observed evidence, essential when long-standing assumptions about scene geometry are invalidated by permanent environmental changes such as road construction. The decision-making layer consumes the updated representations to perform tasks such as object tracking, action anticipation, and anomaly detection, with uncertainty estimates calibrated to reflect both aleatoric noise and the epistemic confidence reduction that occurs after a distribution shift before full adaptation.

4. Memory-Aware Visual Representations: Design and Trade-offs

Memory-aware visual representations are distinguished from conventional feature embeddings by their dual role: they must serve as effective inputs for immediate perceptual tasks while simultaneously carrying a compressed history of their own formation that facilitates future retrieval, replay, or selective forgetting. Designing such representations involves navigating a multi-dimensional trade-off space that includes compression ratio,

reconstruction fidelity, temporal coverage, and representational entanglement. Excessive compression destroys fine-grained details needed for precise spatio-temporal localization, while insufficient compression causes the memory store to overflow, forcing costly eviction policies that may discard critical rarities.

One promising paradigm is the use of slot-based or object-centric representations that decompose a scene into a set of independent entity tokens, each with its own persistent memory slot. Object trackers can update the per-slot state incrementally, and the slots that do not receive updates for a configurable duration can be gradually faded or offloaded to a secondary slow storage tier. This object-centric factorization disentangles the memory footprint from the raw frame count, aligning the growth of the memory index with the number of distinct entities observed rather than the sequence duration. However, slot-based methods introduce a binding problem when dynamic occlusions or viewpoint changes cause identity switches, and resolving these ambiguities online without human supervision remains an open systems challenge. Recent work on memory-aware fine-tuning of large segmentation models for long-sequence video object segmentation provides a concrete instantiation: by injecting learnable memory tokens into the attention layers of a pretrained segmenter, the model can retain object masks across hundreds of frames with substantially lower computation than full fine-tuning, demonstrating how memory-awareness can be retrofitted into powerful foundation models without architectural disruption [11].

Another design axis concerns the degree to which the memory representation preserves temporal ordering. A purely bag-of-embeddings memory index is efficient for content-addressable retrieval but discards the sequential structure that carries predictive signals for activities. Structured memory formats, such as multi-scale feature pyramids with temporal positional encoding or memory graphs that connect temporally adjacent slots, preserve enough ordering to support sequence modeling while still enabling subsampled retrieval. The choice between structured and unstructured memory indexes thus depends on the downstream task profile: anomaly detection in manufacturing requires fine temporal ordering to spot subtle rhythm violations, whereas object re-identification in multi-camera networks benefits more from a robust content-based indexing that is invariant to temporal shifts.

The interplay between memory-awareness and model capacity also dictates the degree of representation reuse. In a continually deployed system, the same semantic concept may reappear under different illumination, occlusion, or sensor modalities. Representations designed with a disentangled style-content structure, where style variations are factored into separate modulation parameters, allow the core semantic signature to be reused while adapting to new conditions with minimal additional memory overhead. This reuse principle is especially significant for federated deployment topologies in which edge models trained on heterogeneous device cameras must exchange compressed representations without exposing raw data. A shared semantic codebook can be maintained across the federation, with each device contributing local adaptation vectors that align its idiosyncratic views to the common reference, reducing both communication cost and the risk of memorizing privacy-sensitive visual details.

5. Continual Learning Strategies for Dynamic Video Streams

Applying continual learning to video streams requires rethinking the unit of experience over which forgetting is measured and mitigated. Unlike image classification tasks with clear task boundaries, video perception unfolds continuously with soft transitions. Defining a task as a fixed-length clip is arbitrary and may split coherent activities, while treating the entire infinite

stream as a single task hides catastrophic interference that accumulates gradually. The framework must therefore operate on overlapping, multi-scale temporal windows that capture both short-range motion patterns and long-range scene evolution.

Elastic weight consolidation, when applied naively to streaming video, can suffer from an accumulating precision penalty that progressively freezes the network and prevents adaptation to genuine novelties after a long stable period. To counteract this gradual loss of plasticity, the framework combines elastic consolidation with an importance decay schedule that slowly reduces the penalty for older synapses if their associated concepts have not been reactivated within a relevance horizon. This dynamic consolidation scheme is complemented by a strategically curated episodic replay buffer that does not store raw frames but instead retains highly compressed feature prototypes, selected by a prioritization policy that balances temporal diversity with semantic rarity. Replayed prototypes are interleaved with incoming mini-batches during training, and the replay ratio is adjusted online according to a drift index that quantifies the mismatch between the current data distribution and the historical training distribution as estimated from the semantic index. This hybrid regularization-plus-replay strategy avoids the unbounded storage growth of pure replay while remaining more adaptable than pure regularization in the face of large environmental changes.

Meta-learning techniques offer a complementary mechanism for rapid context-based adaptation without extensive replay. By training the model to perform few-shot adaptation on synthetic video procedurally generated to simulate common distribution shifts—such as weather transitions, day-night cycles, or crowd density variations—the framework acquires the ability to quickly reconfigure its normalization statistics, attention masks, and routing parameters when a real shift is detected. The continual learning layer then treats each detected shift as a meta-task, solving it with a small number of gradient steps on a carefully assembled support set of recent memories. The overall stability-plasticity balance is governed not by a fixed hyperparameter but by a meta-controller that learns from deployment experience when to invoke fast adaptation, when to consolidate slowly, and when to expand the memory index structurally. This self-regulating architecture reduces the burden on human operators to tune learning rates and penalty coefficients for every new camera installation.

6. Implementation Infrastructure and Scalability Considerations

Realizing a continual video understanding framework at scale demands a refactoring of the typical split between training and inference infrastructure. Conventional pipelines train models offline on static datasets, freeze them, and deploy frozen checkpoints to edge devices. In a continual learning regime, training and inference are interleaved: the model must continuously consume live data, update its parameters, and serve predictions on the same stream without perceivable downtime. This necessitates a microservice-based architecture in which the representation, memory, and learning components are decoupled as independently scalable services that communicate through high-throughput message queues. The memory index, being the most state-intensive component, is maintained in a distributed in-memory database such as Redis or a custom key-value store optimized for high-dimensional vector search, with sharding across spatial regions or camera clusters to bound query latency. As the number of cameras scales, the memory store can be federated, with each edge node hosting a local cache of recent and highly salient representations, while a cloud-hosted global memory aggregates de-identified semantic prototypes across nodes for cross-site generalization.

The orchestration layer must support elastic resource allocation to handle variable frame rates and bursty event streams. A controller meshed with the adaptive frame selector monitors

computational load and network bandwidth, offloading heavy replay consolidation batches to idle GPU cycles in nearby fog nodes when the real-time inference load permits. Distributed continual learning introduces consistency challenges: if multiple cameras observe the same event from different viewpoints, their local memory updates may drift apart, causing conflicting representations for the same entity. A lightweight consensus protocol based on vector clocks and hashed representation fingerprints can detect divergence and trigger reconciliation, merging memory slots that are found to encode the same object tracked across camera boundaries. This reconciliation process must be carefully throttled to prevent cascading updates from saturating the communication fabric.

Privacy-preserving mechanisms are embedded into the infrastructure from the ground up. Raw video is never transmitted beyond the device unless explicit opt-in consent is provided or a security override is activated under predefined threat protocols. Instead, edge nodes exchange behaviorally anonymized representations and metadata. Federated averaging is extended to account for the non-stationary nature of video data by weighting local updates according to the temporal freshness and distributional novelty of each stream, ensuring that a camera observing a highly dynamic intersection contributes a proportionally larger update than one monitoring a static corridor. Differential privacy guarantees are applied to the gradients and memory representations transmitted to the central aggregator, with the privacy budget tracked over rolling windows to provide rigorous long-term protection.

7. Evaluation Paradigms and Structural Robustness

Evaluating a continually learning video system requires moving beyond static benchmark suites in which training and test sets are drawn from identical distributions. A robust evaluation must simulate the temporal dynamics of real deployment, including gradual domain shift, sudden concept drift, sensor degradation, and the open-world emergence of unknown object categories. We advocate for the construction of lifelong video benchmarks that span months of continuous capture, enriched with stream-specific metadata tags for weather, traffic volume, and maintenance events, enabling granular failure analysis. Standard online metrics such as frame-level accuracy, object-tracking identity preservation, and action classification precision should be complemented by plasticity-stability curves that trace accuracy on both recent and historical tasks over the deployment lifetime. Memory utilization efficiency should be reported in terms of the effective memory-to-concept ratio, capturing the number of distinguishable entities maintained per megabyte of storage.

Robustness evaluation must extend to adversarial settings in which an attacker attempts to poison the memory index by injecting subtly perturbed frames that induce erroneous consolidation or trigger premature forgetting. Probing the vulnerability of the episodic buffer to backdoor triggers that lie dormant until a specific pattern appears later in the stream reveals critical dependencies on the compression algorithm. Defenses such as anomaly-filtered memory writes, where candidate embeddings are screened by an ensemble of one-class anomaly detectors before being committed to the index, can mitigate such threats but may also reject rare legitimate events, a trade-off that must be calibrated per deployment context.

Structural robustness also encompasses graceful performance degradation under hardware resource constriction. When an edge device enters a low-power mode due to thermal throttling, the framework should smoothly reduce the frequency of memory consolidation operations, increase the sparsity of the episodic buffer, and optionally offload read-only queries to approximate nearest-neighbor backends, rather than suffering a hard failure. This requires careful design of heterogeneous compute fallback paths, for instance replacing exact

transformer attention lookup with locality-sensitive hashing when the GPU is unavailable. Architectural resilience testing through fault injection campaigns becomes an integral part of the development lifecycle, ensuring that the system degrades predictably and recovers without catastrophic memory corruption after abrupt restarts.

8. Governance, Fairness, and Policy Implications

When a continual learning video system is deployed in public spaces—transit hubs, smart city infrastructures, or retail environments—its evolution over time introduces governance challenges distinct from static classifier audits. Traditional fairness audits check for demographic parity at a fixed snapshot, but a continuously updating model may drift toward biased representations if the environmental data distribution exhibits systemic disparities. For example, a memory index that gradually over-represents pedestrians in business attire during weekday office hours might under-detect individuals in casual clothing during weekends, not because of intentional discrimination but due to sampling bias amplified by the memory consolidation policy that reinforces frequently accessed slots. Continuous fairness monitoring must therefore be mandated, with periodic statistical checks on recall parity across sensitive groups, flagged when confidence intervals drift beyond prescribed thresholds.

Explainability and contestability mechanisms are equally critical. Individuals captured in video streams have a right to understand what representations the system retains about them and to request deletion under data protection frameworks such as the General Data Protection Regulation. In a memory-augmented architecture, exact deletion of a person's influence from a distributed, frequently updated memory index is non-trivial. Mechanistic unlearning requires tracing all write operations linked to a specific data subject and culling the corresponding memory slots, potential cascading deletions of derived prototypes, and recomputing the global index. The framework must provision a tamper-proof audit log that records data provenance, consolidation events, and deletion requests, enabling external auditors to verify compliance without exposing proprietary model internals. This log must itself be secured against tampering, perhaps using a lightweight permissioned blockchain anchoring important state transitions.

Policy frameworks for continual perception systems should mandate transparency reports that detail the system's evolution: the rate of memory turnover, the introduction of new semantic categories, and the confidence calibration over drifting sub-populations. Certification processes could adapt safety assurance methods from medical device software engineering, where change impact analysis and regression testing are required for every significant update. In the long term, standards bodies may develop profiles for different risk tiers, with unconstrained public surveillance holding to the strictest requirements that include independent red-teaming, community oversight panels, and mandatory shut-off mechanisms when fairness violations remain uncorrected beyond a grace period. Researchers and system architects play a vital role in making these governance requirements technically tractable by designing memory management APIs that expose configurable fairness constraints, deletion interfaces, and drift audit reports as first-class primitives rather than afterthoughts.

9. Deployment and Sustainability in Real-World Systems

Deploying a continually learning video intelligence framework at city scale or across a large fleet of autonomous vehicles places substantial energy and maintenance burdens that cannot be externalized from the research design process. Continuous representation updates and memory consolidation steps consume electricity proportionally to the volume of incoming

video, which can reach tens of thousands of hours per day in a moderately sized metropolitan network. Sustainable deployment requires a carbon-aware scheduling policy that times energy-intensive global index recomputation to periods when the local energy grid has a high renewable mix, using day-ahead carbon intensity forecasts. Collaboration between edge nodes and regional data centers can shift replay batch processing to locations with lower marginal emissions, dynamically routing workloads while respecting jurisdictional data residency constraints.

Memory architecture choices directly influence hardware refresh cycles and e-waste production. Designs that require frequent writes to persistent flash storage accelerate its wear, shortening the lifespan of deployed sensor nodes. Frameworks optimized for in-memory random-access representations that minimize persistent storage pressure, combined with quantization of memory embeddings to low-bitwidth integers, can extend node lifetimes while retaining acceptable retrieval accuracy. The continual learning framework should also anticipate hardware generational turnover: when a new generation of more energy-efficient accelerators is installed, the memory index and model parameters must be seamlessly transferable without a costly full retraining. Modularity of the representation space, where the semantic codebook is stored in a hardware-agnostic format, facilitates this transition and reduces the carbon footprint of model migration.

Long-term maintenance staff must be considered, particularly in municipal deployments with limited AI expertise. The framework's operational dashboard should provide intuitive visualizations of memory health, forgetting trends, and fairness drift, with automated prescriptive actions that a non-specialist operator can approve or reject. The consolidation of governance and sustainability features into a unified operational interface lowers the barrier to responsible stewardship and prevents the system from becoming a black box that decays silently. As foundational video models continue to grow in size, the tension between scaling model capacity for improved accuracy and the imperative to minimize energy footprint will become more acute. The memory-aware principles discussed here offer a path to decoupling representational richness from raw parameter count, leveraging compressed reusable knowledge to achieve high competence under tight resource envelopes. Final deployment decisions must weight these architectural trade-offs against each application’s social value and risk profile, ensuring that the perpetual learning gaze remains aligned with societal benefit rather than unfettered surveillance.

10. Conclusion

This paper has articulated a continual learning framework for long-sequence video understanding that elevates memory-awareness from a utility function to a unifying architectural principle. By integrating compressible episodic buffers, dynamic consolidation schedules, and modular memory controllers, the framework addresses the core tension between stability and plasticity in open-world video streams. The system-level analysis spanned adaptive frame selection, distributed memory architectures, privacy-preserving federated updates, carbon-aware scheduling, and fairness-oriented governance. A central finding is that trade-offs historically confined to algorithmic design—compression ratio, temporal ordering, slot granularity—propagate into infrastructure costs, regulatory compliance, and long-term sustainability, demanding co-design across the full socio-technical stack. Future research should develop standardized lifelong video benchmarks that reflect deployment diversity, unlearnability APIs that facilitate data protection rights, and energy-proportional hardware-software co-design that aligns continual learning workloads with green

computing principles. As video perception infiltrates public and private life, the frameworks we build today will encode the operational, ethical, and environmental values of tomorrow's intelligent infrastructure.

References

1. Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the Kinetics dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6299–6308.

2. Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 6202–6211.

3. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521–3526.

4. Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. Neural Networks, 113, 54–71.

5. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 6836–6846.

6. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., & Tuytelaars, T. (2018). Memory aware synapses: Learning what (not) to forget. Proceedings of the European Conference on Computer Vision (ECCV), 144–161.

7. Sukhbaatar, S., Weston, J., Fergus, R., et al. (2015). End-to-end memory networks. Advances in Neural Information Processing Systems (NeurIPS), 2440–2448.

8. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T. P., & Wayne, G. (2019). Experience replay for continual learning. Advances in Neural Information Processing Systems (NeurIPS), 348–358.

9. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

10. Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., ... & Feichtenhofer, C. (2024). SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714.

11. S

12. Zhang, H., Ananthanarayanan, G., Bodik, P., Philipose, M., Bahl, P., & Freedman, M. J. (2017). Live video analytics at scale with approximation and delay-tolerance. Proceedings of the 14th USENIX Conference on Networked Systems Design and Implementation (NSDI), 377–392.

13. Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. Proceedings of the International Conference on Machine Learning (ICML), 1126–1135.

14. Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.

15. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

16. Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Proceedings of the Theory of Cryptography Conference (TCC), 265–284.

17. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

18. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. Proceedings of the International Conference on Learning Representations (ICLR).

19. Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). VideoBERT: A joint model for video and language representation learning. Proceedings of the IEEE International Conference on Computer Vision (ICCV).

20 Wu, C.-Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., & Girshick, R. (2020). Long-term feature banks for detailed video understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).