Cross-Domain Video Object Segmentation via Memory-Guided Transfer Learning and Lightweight Adaptation

Niklas Bryant
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
bryant432@oregonstate.edu

Karan D. Pandey
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
karandpandey966@uc.edu

Abstract

Video object segmentation is a critical capability in diverse domains ranging from autonomous driving and surgical robotics to environmental monitoring and augmented reality. Existing models achieve remarkable performance on curated benchmarks, yet they frequently degrade under domain shifts that arise from variations in illumination, camera viewpoint, object appearance, and scene context. The emergence of memory-augmented foundation architectures such as the Segment Anything Model series provides powerful spatiotemporal priors, but their vast parameter counts and memory footprints pose challenges for cross-domain adaptation in resource-constrained settings. This paper presents a systems-level examination of cross-domain video object segmentation through the lenses of memory-guided transfer learning and lightweight adaptation. We argue that effective deployment demands more than incremental algorithmic improvements; it requires a holistic rethinking of architecture selection, memory management, training regimes, and infrastructure orchestration. By analyzing the interplay between long-range temporal memory, parameter-efficient fine-tuning strategies, and edge-cloud resource partitioning, we identify fundamental trade-offs between segmentation fidelity, inference latency, energy consumption, and domain robustness. We further discuss governance considerations spanning fairness, data sovereignty, and sustainability, highlighting the societal responsibilities that accompany large-scale deployment of adaptive visual perception systems. Our analysis incorporates recent advances in memory-aware fine-tuning of video foundation models and situates them within a broader sociotechnical framework. We conclude by proposing evaluation protocols and policy guidelines that can steer the development of cross-domain video object segmentation toward equitable, efficient, and trustworthy outcomes.

Keywords

Video object segmentation; transfer learning; domain adaptation; memory-augmented networks; lightweight fine-tuning; edge computing.

1. Introduction

Video object segmentation, the task of delineating objects of interest across consecutive frames, has experienced transformative advances with the rise of large-scale pretrained models and memory-intensive architectures. Early solutions relied on online fine-tuning and

heuristic propagation, but the introduction of space-time memory networks enabled models to read and write frame-level embeddings over long temporal windows, substantially improving consistency and occlusion handling. More recently, foundation models such as the Segment Anything Model and its video extension, SAM2, have demonstrated that enormous pretraining on diverse data can endow segmentation models with remarkable zero-shot generalization to unseen object categories and scenes. These developments have shifted the research frontier from within-dataset accuracy toward cross-domain robustness, where models trained predominantly on natural web videos must operate reliably in specialized environments such as medical endoscopy, satellite imagery, or wildlife tracking. The distributional mismatch between source and target domains manifests in altered color statistics, object morphology, motion patterns, and background clutter, all of which challenge the inductive biases encoded in memory-based architectures.

The central thrust of this paper is that achieving robust cross-domain video object segmentation is not solely an algorithmic problem but a profoundly systemic one. The decisions about where to store memory states, how to update them during adaptation, and which model components to freeze or modify determine not only accuracy but also the feasibility of deployment on resource-constrained devices. The growing adoption of edge computing in autonomous vehicles, unmanned aerial systems, and wearable health monitors imposes strict latency and power budgets that are incompatible with full-model fine-tuning on target domain data. At the same time, privacy regulations and bandwidth limitations often prohibit continuous streaming of video to centralized cloud servers, necessitating on-device adaptation mechanisms that are both sample-efficient and computationally frugal. These constraints motivate the twin concepts we explore in depth: memory-guided transfer learning, wherein the rich spatiotemporal representations learned by large video segmentation models are selectively transferred and adapted, and lightweight adaptation, which comprises a family of parameter-efficient techniques that enable domain customization without retraining the entire model.

A growing body of work has begun to address these challenges by designing modular memory update rules, introducing adapter layers into memory read and write pathways, and leveraging low-rank decomposition to limit the number of trainable parameters. For instance, a memory-aware fine-tuning strategy for SAM2 was recently proposed that substantially reduces computational overhead while preserving segmentation quality on long video sequences [13]. Although such contributions advance the state of the art, a comprehensive systems perspective that ties together architectural design, training protocols, deployment infrastructure, fairness, and governance remains underdeveloped. This paper aims to fill that gap by offering a structured analysis organized around architectural trade-offs, adaptation mechanisms, infrastructure considerations, evaluation frameworks, and sociotechnical implications. We deliberately eschew mathematical formalisms and algorithmic pseudocode to foreground the conceptual and systemic dimensions that are essential for building responsible, scalable cross-domain video perception systems.

2. System Architecture and Cross-Domain Challenges

Contemporary video object segmentation architectures can be broadly understood as encoder-decoder pipelines augmented with external memory banks that store compact representations of previously processed frames. In a typical space-time memory design, a query frame encoder produces feature maps that are matched against keys derived from past frames, and the retrieved values are fused with the current decoder to refine segmentation masks. Models

such as XMem extended this paradigm with a multi-level memory hierarchy inspired by Atkinson-Shiffrin human memory models, introducing distinct working and long-term memory stores that improve handling of very long videos [3]. The Segment Anything Model 2 further unified image and video segmentation under a transformer-based framework, enabling promptable segmentation with spatiotemporal memory across heterogeneous domains [2]. These architectural advances have yielded unprecedented in-domain performance on benchmarks like DAVIS and YouTube-VOS [7]; however, they also embed strong assumptions about the visual world that can become brittle when confronted with domain shift.

Cross-domain challenges manifest at multiple architectural levels. At the input level, target domain videos may differ from training data in resolution, frame rate, and color space, forcing the encoder to extrapolate beyond its training manifold. At the memory level, the stored keys and values are computed by a source-domain encoder that may fail to capture discriminative features for novel object categories, leading to incorrect retrieval and mask propagation. Even small discrepancies in texture or shape can accumulate over time within the memory, causing the segmentation to drift or collapse entirely. Moreover, the computational demands of maintaining and querying a large memory bank grow linearly or superlinearly with sequence length, which conflicts with the strict latency constraints of edge deployment. Architectural decisions such as memory size, update frequency, and key-value dimensionality must therefore be carefully tuned for the target domain and hardware platform, yet exhaustive tuning is rarely feasible given the heterogeneity of real-world deployment scenarios.

The interplay between memory-guided architectures and domain adaptation introduces additional structural trade-offs. One strategy is to fine-tune the entire model, including the memory modules, on a small set of target-domain labeled frames. While this can yield high accuracy, it risks catastrophic forgetting of source-domain knowledge and is computationally prohibitive on edge accelerators. Alternatively, the memory module can be partially frozen, treating it as a reusable store of general spatiotemporal patterns that can be queried with domain-adapted encoders. This modular separation allows targeted adaptation of the encoder and decoder while preserving the stability of long-term memory representations, but it may limit the model's ability to learn domain-specific temporal dynamics. Identifying the optimal balance between plasticity of the encoder and stability of the memory is a system-level optimization that requires consideration of data availability, computational budget, and the severity of the domain shift.

3. Memory-Guided Transfer Learning: Design Principles and Trade-offs

Transfer learning for memory-augmented video object segmentation involves the selective reuse of pretrained components—most critically the memory bank and its associated read-write mechanisms—to accelerate adaptation to a new visual domain. The guiding principle is that the spatiotemporal correspondence patterns acquired during large-scale pretraining capture universal properties of object motion, occlusion, and reappearance that are largely invariant across domains. The challenge lies in preserving these general competencies while injecting domain-specific information that calibrates the memory representations to the target distribution. A naive fine-tuning approach updates all parameters jointly, which may overwrite valuable pretrained features, especially when target data are scarce. This is where lightweight transfer paradigms become essential: they restrict the capacity of the adaptation to avoid overfitting while maintaining the representational power of the backbone.

One promising direction involves inserting small adapter modules into the memory read and write pathways. These adapters, which consist of bottleneck layers with a fraction of the original parameter count, can be trained on domain-specific data while the original weights remain frozen. In the context of SAM2, adapter-based fine-tuning has been explored for medical image segmentation, but extending this to video requires careful handling of temporal consistency, as the adapters must learn to modulate attention scores across frames without introducing flicker. A parallel approach employs low-rank adaptation, where weight updates are factorized into low-rank matrices, dramatically reducing the number of trainable parameters. This technique has been successfully applied to large language models [10] and is beginning to find adoption in vision tasks. Applied to the query-key projection matrices within the memory attention layers of a video segmentation model, low-rank adaptation enables localized domain adjustments while preserving the global structure learned from pretraining. The choice between adapters, low-rank updates, and other parameter-efficient techniques such as prompt tuning or bias-only fine-tuning hinges on the degree of domain shift: larger distributional gaps may necessitate higher-rank updates or additional trainable layers, which in turn increase the risk of overfitting and the computational load.

The memory-aware fine-tuning approach introduced in [13] exemplifies how system-level efficiency can be co-optimized with segmentation accuracy. By selectively updating memory-related parameters and employing a training schedule that progressively unfreezes layers, the method achieves strong performance on long video sequences while reducing training time and memory consumption. This design underscores a broader lesson: the effectiveness of memory-guided transfer depends not merely on which parameters are updated, but on the temporal granularity and scheduling of those updates. For cross-domain scenarios, a staged adaptation process may first align the encoder with target statistics using domain-adversarial training or feature distribution matching [8], then gradually adapt the memory module using a limited set of annotated keyframes. Such a curriculum maintains model stability while closing the domain gap.

Trade-offs abound. Increasing the memory bank capacity improves long-range recall but inflates both storage requirements and query latency, which can be untenable on mobile GPUs. Conversely, aggressive memory compression through hashing or attention sparsification reduces resource usage but may discard fine-grained details essential for segmenting small or fast-moving objects. In cross-domain settings, these trade-offs interact with domain difficulty: for example, a surveillance domain with static cameras may benefit from a large but sparsely updated memory, whereas an endoscopic surgery domain with rapid tissue deformation and tool motion demands a compact but frequently refreshed memory. System designers must therefore treat memory configuration as a tunable policy that responds to domain characteristics and hardware constraints, rather than a fixed hyperparameter.

4. Lightweight Adaptation Mechanisms for Resource-Constrained Deployment

Lightweight adaptation encompasses a constellation of techniques aimed at tailoring large pretrained models to specific tasks with minimal computational and memory overhead. In video object segmentation, these methods are essential for bridging the gap between high-capacity foundation models and the realities of edge deployment. Parameter-efficient fine-tuning, knowledge distillation, network pruning, and quantization each offer distinct trade-offs in terms of accuracy retention, latency reduction, and energy efficiency, and their selection constitutes a critical system design decision.

Parameter-efficient fine-tuning, as previously discussed, freezes the bulk of the model and trains only a small set of additional or restructured parameters. Beyond adapters and low-rank adaptation, methods such as visual prompt tuning prepend a small number of learnable tokens to the input sequence, conditioning the frozen model on domain-specific cues. These tokens can be designed to modulate the memory readout process by biasing the attention mechanism toward domain-relevant features. In cross-domain video segmentation, prompt tokens learned on a few target-domain frames can be prepended to each query frame, effectively steering the model without modifying the core memory contents. While this approach is extremely lightweight, it may lack the expressiveness needed for severe distribution shifts, necessitating a hybrid strategy that combines prompts with low-rank updates of critical layers.

Knowledge distillation offers an alternative pathway: a large, fully fine-tuned teacher model adapted to the target domain can transfer its knowledge to a compact student model that inherits both the original pretrained representations and the domain-specific refinements [11]. The distillation process can be structured to focus on the consistency of memory retrievals, where the student learns to reproduce the teacher's attention patterns over the memory bank rather than merely imitating output masks. This memory-aware distillation aligns the internal temporal reasoning of the lightweight model with that of its powerful teacher, resulting in better long-sequence stability. However, the overhead of training a teacher model that itself requires domain-specific data and compute may be justified only when many student models are to be deployed across heterogeneous devices, as in a fleet of autonomous drones or distributed sensor networks.

Network pruning and quantization are orthogonal techniques that can be applied on top of parameter-efficient fine-tuning to further reduce the model footprint. Structured pruning removes entire filters or memory slots that contribute minimally to segmentation quality, while quantization reduces the bit-width of weights and activations. On specialized hardware such as field-programmable gate arrays or neural processing units, aggressively quantized integer operations can yield order-of-magnitude improvements in energy efficiency, a crucial factor for battery-powered devices operating in remote environments. However, quantization can amplify the domain sensitivity of the model, as low-precision arithmetic may degrade the subtle feature distinctions needed to differentiate objects under domain shift. Therefore, quantization-aware fine-tuning should be performed on target domain data, integrating adaptation and compression into a unified optimization loop.

A particularly attractive paradigm for cross-domain deployment is the concept of modular adapter banks, where a set of domain-specific adapters or low-rank matrices is pretrained offline and dynamically loaded at inference time based on context metadata such as geographic location, camera type, or time of day. This approach decouples adaptation from the base model, enabling rapid switching between domains without reloading the entire network. It also facilitates incremental learning, as new adapters can be added over the product lifespan without retraining or redistributing the core model. The governance implications of such modularity are significant, as it allows device manufacturers to distribute domain adaptation capabilities through over-the-air updates while maintaining a shared, auditable foundation model.

5. Infrastructure, Deployment, and Governance Considerations

The transition from laboratory prototypes to production-scale cross-domain video object segmentation systems introduces a host of infrastructural and governance challenges that extend well beyond model architecture. Reliable deployment demands a robust MLOps

pipeline capable of monitoring data drift, triggering retraining cycles, versioning models and adapters, and orchestrating inference across heterogeneous edge and cloud nodes. In many application contexts, raw video data cannot be exfiltrated from edge devices due to privacy regulations or bandwidth limitations, which necessitates federated or fully on-device adaptation strategies.

Federated learning frameworks enable collaborative model improvement without centralizing sensitive data. Applied to lightweight adaptation, a global base model can be distributed to edge devices, each of which fine-tunes a small set of adapter parameters on local domain data and shares only the adapter updates with a central server for aggregation [15]. This arrangement preserves data sovereignty while allowing the model to benefit from diverse deployment conditions. The communication overhead of transmitting adapter parameters is minimal compared to full model weights, making federated adaptation feasible over cellular or low-power wide-area networks. Differential privacy mechanisms can be integrated to provide formal guarantees against information leakage, aligning with regulatory frameworks such as the GDPR. These privacy-preserving techniques introduce a tension between model utility and protection strength, requiring careful calibration of the privacy budget.

Governance of adaptive segmentation systems also demands attention to fairness and bias. Video object segmentation models can exhibit disparate performance across demographic groups or environmental conditions, particularly when training data overrepresent specific geographic or cultural contexts. For instance, a model deployed in a smart city surveillance system might systematically fail to track pedestrians in certain clothing styles or under poor lighting conditions more common in underserved neighborhoods. Such failures can have cascading effects on downstream tasks such as anomaly detection, traffic management, or public safety analytics, amplifying existing societal inequities. The memory-guided nature of modern architectures can exacerbate these biases if the memory bank perpetuates and reinforces errors over time. Mitigation strategies include fairness-aware data collection, adversarial debiasing of memory representations, and continuous fairness auditing in production [16].

Explainability is another governance pillar. When a segmentation model misses an object in a safety-critical application like autonomous driving, stakeholders need to understand the root cause. Memory-augmented models offer a natural locus for explanation: visualizing which stored memory keys contributed most to a retrieval decision can provide interpretable evidence for model behavior [17]. Such explanations can support regulatory compliance, liability attribution, and user trust. Moreover, the modularity afforded by lightweight adapters can be exploited to isolate the contribution of a specific domain adapter, enabling targeted debugging and certification of domain-specific behaviors.

Sustainability considerations are increasingly central to the deployment of large-scale AI systems. The carbon footprint of training a video foundation model from scratch is substantial; however, the shift toward memory-guided transfer and lightweight adaptation dramatically reduces the environmental cost of domain customization. By reusing a single pretrained model across thousands of devices and retraining only small adapters, the aggregate energy consumption can be orders of magnitude lower than conventional retraining approaches [18]. System architects should adopt lifecycle assessment methodologies that account for both training and inference energy, and policies should incentivize the use of parameter-efficient adaptation through green AI certifications and procurement standards.

6. Evaluation Framework and Cross-Domain Validation

Evaluating cross-domain video object segmentation requires a departure from single-dataset metrics toward a multifaceted assessment that captures accuracy, robustness, efficiency, and fairness. The widely used Jaccard index and contour accuracy, while informative for within-dataset comparisons, do not expose the failure modes that arise under distribution shift. A comprehensive evaluation framework should measure performance on carefully constructed domain generalization splits where the target domain differs systematically in factors such as lighting, weather, camera motion, and object class composition. Cross-domain benchmarks that pair source datasets like YouTube-VOS with target datasets drawn from medical, underwater, aerial, or nighttime driving domains expose the limits of current models. Temporal stability metrics, which quantify mask consistency over time and resilience to occlusions, are especially important given the memory reliance of modern architectures.

System-level metrics must complement traditional segmentation scores. Inference latency, measured at various percentile thresholds, determines whether a model can meet the real-time constraints of a 30 frames-per-second video stream on a mobile processor. Memory footprint, both for the model parameters and the runtime memory bank, governs the feasibility of deployment on embedded platforms with limited RAM. Energy consumption per frame is a critical metric for battery-operated devices, and it should be reported alongside accuracy to enable Pareto-optimal trade-off analysis. Recent work on memory-aware fine-tuning for SAM2 demonstrated that efficiency gains can be achieved without sacrificing segmentation quality on long videos, but those results have yet to be validated systematically across multiple cross-domain pairs [13]. Extending such evaluations to diverse domains and quantifying the sensitivity of efficiency-accuracy trade-offs to domain shift severity is an urgent research need.

Fairness evaluation requires disaggregated performance reporting across subgroups defined by geographic region, sensor modality, and object attributes such as size, motion speed, and occlusion level. Statistical parity and equalized odds metrics adapted from the algorithmic fairness literature should be computed to detect systematic biases. Adversarial robustness must also be assessed, as segmentation models deployed in security-sensitive contexts can be targeted by perturbations designed to cause missegmentation or tracking failures. Cross-domain validation should incorporate adversarial testing with domain-relevant perturbations, such as simulated fog or sensor noise typical of the target environment.

Ablation studies that systematically vary the granularity of adaptation—from full fine-tuning to adapter-only to prompt-only—while measuring both segmentation quality and resource usage provide the empirical foundation for system design guidelines. Meta-analyses across multiple domain pairs can reveal whether certain adaptation techniques consistently outperform others or whether their relative merits are domain-dependent. The development of standardized benchmark suites and evaluation protocols promoted through community challenges would accelerate progress and facilitate fair comparisons, much as the DAVIS and YouTube-VOS benchmarks did for in-domain video object segmentation.

7. Cross-Sector Applications and Sociotechnical Implications

Cross-domain video object segmentation technologies are poised to transform a wide array of sectors, each bringing unique technical and sociotechnical demands. In autonomous driving, vehicles operate across diverse geographical regions, weather conditions, and traffic cultures, requiring segmentation models that can adapt quickly to new visual environments without compromising safety. A memory-guided system that retains representations of rare but critical objects, such as crossing wildlife or construction barriers, and adapts to local road textures

through lightweight fine-tuning, could significantly enhance perception robustness. However, the safety-critical nature of driving demands rigorous validation, transparent failure modes, and clear liability frameworks when segmentation errors contribute to accidents. Governance policies must mandate continuous monitoring and establish reporting standards for adaptive components.

In medical imaging, video object segmentation enables tasks such as surgical instrument tracking, cell lineage tracing, and endoscopic navigation. Cross-domain robustness is essential because imaging protocols, tissue appearance, and instrument designs vary widely across hospitals and device manufacturers. Lightweight adaptation offers a practical path to hospital-specific customization without sharing patient data, aligning with health privacy regulations. Nevertheless, the stakes of segmentation errors in medical contexts are extraordinarily high, necessitating clinician-in-the-loop verification systems and clear regulatory pathways for adaptive AI that evolves over time. The ability to explain memory retrieval decisions can help build clinician trust and align with medical device certification requirements.

Environmental monitoring and conservation represent another compelling application domain. Camera traps and aerial drones deployed in forests, oceans, or polar regions capture footage under highly variable conditions that differ starkly from internet video training data. Memory-guided models that can be fine-tuned on a handful of annotated frames from a specific ecosystem could automate species counting and behavioral analysis, providing ecologists with timely data to inform conservation policy. The sustainability of the AI system itself must be considered; edge devices operating in remote locations often rely on solar power and satellite backhaul, making energy-efficient inference and data-efficient adaptation environmental imperatives.

The sociotechnical implications of widespread cross-domain segmentation extend beyond individual applications. The ability to track objects persistently across diverse visual contexts raises privacy concerns, as individuals could be re-identified across non-overlapping camera views. Legal frameworks must evolve to govern the use of adaptive video understanding in public spaces, balancing societal benefits against the risk of mass surveillance. Moreover, the concentration of foundation model development in a few large technology companies could create dependencies and limit the capacity of smaller organizations and public-sector entities to tailor models to their needs. Open-source releases of base models and adaptation toolkits, coupled with transfer learning techniques that do not require large proprietary datasets, can democratize access and foster a more equitable innovation ecosystem.

8. Conclusion

This paper has examined cross-domain video object segmentation through a systems lens, arguing that the path toward robust, efficient, and equitable deployment lies at the intersection of memory-guided transfer learning and lightweight adaptation. We traced the architectural evolution of memory-augmented video segmentation models and identified the structural vulnerabilities that arise under domain shift. We then analyzed the design space of parameter-efficient transfer, highlighting the trade-offs that govern the selection and configuration of adapters, low-rank updates, knowledge distillation, and quantization. By integrating infrastructure considerations such as federated learning, MLOps pipelines, and edge-cloud partitioning, we demonstrated that technical choices about model architecture are inextricably linked to deployment feasibility, privacy preservation, and environmental sustainability. Governance challenges related to fairness, explainability, and regulation were shown to be equally critical and to demand proactive engagement from system designers. Recent advances

in memory-aware fine-tuning of foundation models, including the work referenced in [13], exemplify the kind of efficiency-oriented research that can anchor this systemic vision, but they must be extended to explicitly confront cross-domain diversity and sociotechnical accountability. Looking ahead, self-supervised domain adaptation, continual memory consolidation, and hardware-adaptive model compression represent promising research frontiers. By embracing a holistic perspective that spans algorithms, systems, and society, the research community can steer cross-domain video object segmentation toward outcomes that are not only accurate but also trustworthy, inclusive, and sustainable.

References

1. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R. (2023). Segment anything. arXiv preprint arXiv:2304.02643.

2. 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.

3. Cheng, H. K., & Schwing, A. G. (2022). XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. In European Conference on Computer Vision (ECCV).

4. Oh, S. W., Lee, J.-Y., Xu, N., & Kim, S. J. (2019). Video object segmentation using space-time memory networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

5. Yang, Z., Wei, Y., & Yang, Y. (2021). Associating objects with transformers for video object segmentation. In Advances in Neural Information Processing Systems (NeurIPS).

6. Wang, Z., Xu, J., Liu, L., Zhu, F., & Shao, L. (2021). Collaborative video object segmentation by multi-scale foreground-biased integration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(7), 2436–2450.

7. Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., ... & Huang, T. S. (2018). YouTube-VOS: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327.

8. Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In International Conference on Machine Learning (ICML).

9. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In International Conference on Machine Learning (ICML).

10. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

11. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

12. 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.

13. Li, G., Yuan, H., Chen, S., Hu, Q., Wang, J., & Jiang, K. (2026). MFT: Memory-Aware Fine-Tuning of SAM2 for Efficient Long-Sequence Video Object Segmentation. IEEE Signal Processing Letters.

14. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE, 107(8), 1738–1762.

15. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). Communication-efficient learning of deep networks from decentralized data. In International Conference on Artificial Intelligence and Statistics (AISTATS).

16. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency (FAT).

17. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

18. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).

19. Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A. A., & Darrell, T. (2018). CyCADA: Cycle-consistent adversarial domain adaptation. In International Conference on Machine Learning (ICML).

20. Dou, Q., Coelho de Castro, D., Kamnitsas, K., & Glocker, B. (2019). Domain generalization via model-agnostic learning of semantic features. In Advances in Neural Information Processing Systems (NeurIPS).