Self-Supervised Object-Person Joint Representation Learning for Large-Scale Visual Retrieval

Larry Richards
Department of Computer Science, George Mason University, Fairfax, VA, USA.
larryrichards74@gmu.edu

Sven D. Lane
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
lane668@missouri.edu

Abstract

The retrieval of visual content in large-scale, unconstrained environments increasingly demands representations that simultaneously encode objects and persons with high semantic fidelity. This paper presents a self-supervised learning framework that jointly learns embeddings for object and person instances from unlabeled image and video collections, targeting robust and scalable visual search. The proposed architecture combines a dual-stream vision transformer backbone with contrastive instance discrimination and a cross-modal clustering objective, avoiding costly manual annotations while preserving the complementary contextual cues shared between objects and individuals. From a systems perspective, we examine the structural trade-offs involved in distributed training, approximate nearest neighbor indexing, and privacy-aware deployment across heterogeneous infrastructure. A substantial portion of the discussion is devoted to governance and fairness issues arising from joint object-person representations, including the risk of reinforcing spurious correlations and profiling in surveillance contexts. Furthermore, we analyze the sustainability implications of training large-scale retrieval models and outline policy-aware deployment strategies that incorporate fairness auditing, domain adaptation resilience, and adversarial robustness. The study integrates insights from recent advances in self-supervised visual learning, person re-identification under clothing variation, and large-scale retrieval systems to propose an infrastructure-centric view of joint representation learning. By synthesizing architectural decisions with societal and operational requirements, this work contributes a multidimensional roadmap for building equitable, interpretable, and sustainable visual retrieval engines at global scale.

Keywords

Self-supervised learning; object-person joint representation; visual retrieval; large-scale systems; infrastructure governance; fairness; sustainability.

1. Introduction

The ubiquity of visual sensors in urban spaces, retail environments, and digital platforms has led to an explosion of unlabeled imagery that must be organized and searched at unprecedented scale. Conventional visual retrieval systems have predominantly treated object retrieval and person re-identification as separate problems, each with its own model architectures, loss functions, and evaluation protocols. However, real-world scenes exhibit a profound entanglement between objects and the individuals who interact with them, such as a

person carrying a distinctive bag, sitting on a particular piece of furniture, or driving a specific vehicle. Exploiting these co-occurrence patterns can significantly improve retrieval robustness, especially when person appearance is ambiguous due to occlusion, pose variation, or clothing changes. Joint object-person representation learning aims to embed both entity types into a shared metric space where relational proximity reflects not only visual similarity but also contextual association. Self-supervised learning offers a compelling path to train such models without the prohibitive expense of manual annotation, instead deriving supervisory signals from data augmentation invariance, temporal continuity in video, and cross-modal alignment between object and person tracks.

Recent breakthroughs in contrastive self-supervised learning have demonstrated that visual representations approaching or exceeding supervised counterparts can be obtained from large unlabeled image collections [1, 2]. Momentum-based memory banks and aggressive data augmentation regimes enable instance-level discrimination that groups semantically similar views while separating dissimilar ones. These methods, however, have largely been applied to object-centric datasets such as ImageNet, leaving the challenge of jointly modeling person and object categories underexplored. When extended to person re-identification, self-supervised pretraining has shown promise in learning generic identity features, yet the coupling with object context remains nascent. Large-scale retrieval systems built on top of such representations rely on efficient approximate nearest neighbor search to handle billion-scale galleries, where the quality of the embedding space directly determines recall under strict latency constraints [3]. The deployment of such systems in public settings further raises critical questions about fairness, privacy, and the environmental costs of training compute-intensive models.

This paper advances the interdisciplinary discourse by framing joint object-person representation learning as a systems problem, where architectural choices are inseparable from infrastructure provisioning, governance frameworks, and societal impact. We do not introduce a novel algorithmic contribution per se, but rather provide a comprehensive analysis of the design space, trade-offs, and responsibilities that accompany the development and deployment of these retrieval frameworks. The discussion is organized around four thematic pillars: architectural foundations that combine dual-stream encoding with cross-attention and self-supervised objectives; system-level infrastructure considerations for distributed training and online serving; governance, fairness, and policy implications that arise from the joint modeling of objects and persons; and operational robustness together with environmental sustainability. Throughout, we draw upon the latest literature, including advances in clothing-invariant person re-identification [5], to illustrate how component innovations must be integrated with a holistic socio-technical outlook.

2. Architectural Foundations and Structural Trade-offs

Designing a neural architecture that can simultaneously capture fine-grained identity cues for persons and discriminative category-level features for objects requires careful balancing between specificity and generalization. A typical dual-stream design employs two separate backbone networks, often vision transformers (ViTs), that process person and object image regions extracted by an upstream detector. The strengths of ViTs lie in their ability to model long-range dependencies through self-attention, which is crucial for associating a person's appearance with spatially distant objects that might carry contextual meaning [6]. The two streams can remain independent during initial feature extraction but are later fused through cross-attention modules that allow person features to attend to object token sequences and

vice versa. Joint supervision is then provided by a combination of contrastive loss operating on the fused embeddings and a clustering objective that encourages grouping of semantically congruent person-object pairs without explicit labels, drawing inspiration from self-supervised multi-modal methods [7].

One structural trade-off concerns the granularity of the object branch. Objects can range from generic categories (chair, car) to highly specific instances (a particular handbag model), and the level of granularity directly impacts the embedding space geometry. Instance-level object discrimination, analogous to person identity discrimination, risks overfitting to object-specific textures that are rare in the gallery, whereas coarse category-level representations may fail to provide useful contextual disambiguation. Hybrid designs that maintain a two-tier object representation, combining a categorical memory bank with an instance-based contrastive queue, have shown empirical advantages but introduce additional hyperparameter complexity. The additional memory footprint and computational load must be weighed against the downstream retrieval accuracy improvements, a trade-off that becomes pronounced when scaling to millions of gallery instances.

Fusion strategies present another architectural fork. Early fusion concatenates person and object tokens before passing them through a shared transformer encoder, promoting tight interplay but often leading to the dominance of one modality over the other. Late fusion, where separate embeddings are computed and then compared via a learned similarity metric, preserves modality-specific information but may miss subtle cross-modal interactions. Recent work on multimodal attention decoupling [5] suggests that separating feature-driven attention from fusion-driven attention can mitigate modality collapse and improve robustness in person re-identification under clothing changes. Adapting such decoupled attention mechanisms into the joint representation pipeline allows the system to maintain strong identity signals even when the visual appearance of a person is altered, while still leveraging object cues that remain stable. From a systems engineering perspective, the choice of fusion scheme also influences the ease of incremental updates: a late-fusion design permits retraining or finetuning of one modality without affecting the other, which is valuable in dynamic environments where object taxonomies may evolve but person privacy regulations forbid retraining on sensitive identity data.

3. System Design and Infrastructure for Large-Scale Retrieval

Translating a joint object-person representation into a deployable retrieval system requires addressing a chain of infrastructure challenges, from distributed training across GPU clusters to low-latency approximate search over embedding indices. Training self-supervised models on video or multi-view data at the scale of millions of tracks demands efficient data loading pipelines that can sample temporally adjacent frames and perform synchronized random augmentations on person and object crops with minimal CPU-GPU transfer bottlenecks. Frameworks built on gradient accumulation and mixed-precision training are standard, yet the dual-stream architecture with cross-attention amplifies memory consumption, often requiring model parallelism or sharding strategies that split attention heads across devices. The management of momentum encoders, which are central to many contrastive learning formulations, adds a further layer of complexity: the memory bank must be sharded consistently and updated with low staleness to maintain discriminability across worker replicas. System designers must navigate these coordination issues while maintaining enough flexibility to incorporate hyperparameter search for temperature scaling, queue size, and augmentation policy.

Once the embedding model is trained, the online retrieval subsystem must index gallery embeddings and perform nearest neighbor search under stringent serving requirements. Approximate nearest neighbor search libraries based on product quantization or graph-based methods, such as those developed for GPU-accelerated similarity search, have become indispensable [3]. Yet the joint representation introduces a subtle indexing challenge: because the embedding is jointly conditioned on both a person and an associated object, query-time composition of these two modalities must be handled efficiently. In a typical application, a query might consist of a person image and a separate object image that must be fused on the fly. This runtime fusion can be offloaded to lightweight neural layers that execute within the feature serving layer, but only if the underlying index stores separate person and object sub-embeddings rather than a single monolithic fused vector. Decoupled storage of sub-embeddings doubles the index size but enables flexible querying with missing modalities, such as person-only or object-only searches, while retaining the ability to combine them via late fusion during retrieval. The infrastructure cost of maintaining such a dual index must be justified by the operational gain in recall across diverse query types.

Deployment across heterogeneous edge and cloud environments further complicates the systems picture. In smart city or retail analytics scenarios, privacy regulations may require that raw images never leave the local device, necessitating on-device feature extraction with compressed, quantized models. The joint representation model must therefore be designed with neural architecture search or distillation loops that can produce edge-friendly variants without catastrophic loss of cross-modal alignment. Model quantization techniques, including integer-only inference and pruning of attention heads that contribute least to the cross-modal signal, have shown promise in reducing the computational budget while preserving retrieval accuracy. At the infrastructure orchestration layer, these edge-cloud interactions demand a control plane that can dynamically route queries, synchronize index updates, and enforce access control policies aligned with data governance rules.

4. Governance, Fairness, and Policy Implications

The fusion of object and person representations into a single embedding space introduces fairness risks that differ qualitatively from those present in unimodal retrieval. When an embedding space is trained to capture co-occurrence statistics between person appearances and object categories, it may inadvertently encode spurious associations that reflect societal biases present in the training data. For example, if certain types of personal belongings are disproportionately observed alongside specific demographic groups in the unlabeled corpus, the joint embedding can amplify these correlations, leading to retrieval results that reinforce stereotypes or enable discriminatory filtering. Fairness in visual retrieval demands not only demographic parity in person re-identification but also an audit of the object-person association landscape to detect and mitigate harmful correlations. Techniques such as adversarial debiasing of the joint embedding space and fairness-aware metric learning have been explored in the general machine learning fairness literature [9] and can be adapted to this multimodal setting.

The governance framework surrounding joint object-person retrieval must also confront the tension between public safety applications and civil liberties. Surveillance systems that can retrieve individuals based on the objects they carry or the vehicles they enter offer powerful investigative capabilities, but they simultaneously lower the barrier to mass tracking and profiling. Policymakers increasingly require impact assessments and transparency reports for AI systems deployed in public spaces, covering the data sources used for self-supervised

pretraining, the demographic composition of the training data, and the potential for misuse. From a technical design stance, embedding spaces can be instrumented to support fairness audits by exposing interpretable saliency maps that reveal which visual features — be they facial regions, clothing attributes, or object textures — dominate the similarity score. Explainability methods originally developed for convolutional networks [10] are being extended to transformer-based dual-stream architectures, enabling human reviewers to verify that retrieval decisions rest on legitimate contextual evidence rather than protected attributes.

Privacy-preserving infrastructure represents a further policy-driven constraint. Self-supervised training on large-scale unlabeled image collections often scrapes data from public cameras or social media, raising questions about consent and the right to be forgotten. One mitigation is federated learning, where model updates are computed locally on institution-controlled data and only aggregated gradients are shared, thus keeping raw images distributed. However, federated training of dual-stream models introduces additional challenges, such as maintaining cross-modal alignment when different institutions possess only partial views (for example, some nodes see mostly objects, others mostly persons). The governance protocol must therefore include data contribution agreements that specify modality coverage expectations and liability clauses for biased model outputs. As regulations such as the EU AI Act classify certain biometric retrieval systems as high-risk, the system architecture must incorporate logging, audit trails, and the ability to unlearn specific individuals or objects upon request, which in turn imposes constraints on how index shards are managed and how training checkpoints are stored.

5. Deployment Robustness and Sustainability in Operational Environments

Robustness of joint object-person retrieval systems extends beyond adversarial attacks on the embedding network to encompass domain shift, temporal decay of associations, and graceful handling of missing modalities. A model trained on daytime, clear-weather data may experience severe accuracy degradation when deployed in low-light or adverse weather conditions. Likewise, fashion trends and object designs evolve, causing the statistical linkage between person appearance and object presence to drift over time. Continuous domain adaptation strategies that leverage self-supervised signals from incoming query streams can help maintain performance, but these strategies must be carefully regulated to avoid silently incorporating fresh biases. Robustness against adversarial perturbations is equally critical, as malicious actors could manipulate object appearances or person attire to evade retrieval or force false matches. Adversarial training regimes developed for unimodal classifiers [15] must be extended to the joint embedding space, ensuring that small input perturbations do not cause a query embedding to cross decision boundaries in the nearest neighbor index.

The sustainability of large-scale retrieval systems has emerged as a pressing concern, given the considerable energy and carbon footprint of training modern transformer-based architectures on internet-scale data. Self-supervised learning often requires even longer training schedules and larger batch sizes than supervised learning, exacerbating energy consumption. A holistic sustainability analysis must account not only for training costs but also for the inference energy footprint of the deployed index, which grows with the gallery size and the dimensionality of the joint embedding. Techniques such as distillation from large dual-stream models into smaller, energy-efficient architectures, combined with elastic scaling of cloud resources based on query load, can reduce the average carbon intensity per query. Recent studies quantifying the environmental costs of deep learning highlight the importance of selecting training regions with clean energy, using carbon-aware job scheduling, and

reporting energy metrics alongside accuracy in research publications [17]. Institutional procurement policies for retrieval infrastructure should therefore mandate lifecycle carbon assessments, fostering an ecosystem where system architects optimize for both retrieval performance and environmental stewardship.

Beyond energy, the long-term sustainability of joint object-person retrieval depends on its alignment with societal values and legal frameworks. A representation that tightly couples persons and objects may accelerate certain investigative workflows, but if the resulting infrastructure erodes public trust or enables disproportionate surveillance, its societal license to operate will be revoked through regulatory action or public backlash. This prospect necessitates embedding policy compliance directly into the system lifecycle, from data collection and model design to serving and decommissioning. Forward-looking designs might include dynamic embedding privacy filters that obscure certain object associations when queries originate from contexts lacking proper authorization, or differential privacy guarantees that bound the information leakage of any given embedding. Such protective measures, while introducing additional latency and storage overhead, can future-proof the system against evolving regulatory landscapes and maintain long-term viability.

6. Conclusion

This paper has examined self-supervised joint object-person representation learning for large-scale visual retrieval through the lens of systems engineering, infrastructure design, and socio-technical governance. We have argued that architectural choices regarding dual-stream encoding, cross-modal fusion, and contrastive objective design are intimately linked with deployment constraints such as indexing efficiency, edge-cloud orchestration, and privacy-preserving computation. The coupling of objects and persons in a shared embedding space introduces unique fairness challenges, demanding fairness auditing, explainability instrumentation, and robust policy governance frameworks that extend beyond traditional unimodal retrieval systems. Robustness to domain shift, adversarial threats, and environmental sustainability were identified as integral to operational excellence and long-term public acceptability. By synthesizing insights from self-supervised learning, person re-identification under clothing variation, and large-scale similarity search, we have sketched a multidimensional roadmap that emphasizes the need for technical communities, policy makers, and infrastructure providers to collaborate in shaping retrieval ecosystems that are accurate, equitable, and sustainable. Continued interdisciplinary research and open dissemination of system design lessons will be essential as these retrieval technologies become foundational components of tomorrow's visual information infrastructure.

References

1. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning (ICML). PMLR.

2. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

3. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535-547.

4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In European Conference on Computer Vision (ECCV). Springer.

5. Ding, Y., Wang, X., Yuan, H., Qu, M., & Jian, X. (2025). Decoupling feature-driven and multimodal fusion attention for clothing-changing person re-identification. Artificial Intelligence Review, 58(8), 241.

6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR).

7. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (ICML). PMLR.

8. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. C. (2021). Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6), 2872-2893.

9. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

10. 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).

11. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

12. Wu, Y., & He, K. (2018). Group normalization. In Proceedings of the European Conference on Computer Vision (ECCV).

13. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579-2605.

14. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the International Conference on Machine Learning (ICML).

15. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations (ICLR).

16. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In IEEE Symposium on Security and Privacy (S&P).

17. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).

18. Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

19. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).

20. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning (ICML).