Multi-Granularity Attention Fusion for Video-Based Person Re-Identification in Dynamic Urban Scenes

Geapak Shukla
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
shuklageapak@unr.edu

Abstract

Video-based person re-identification in dynamic urban environments constitutes a critical enabler of intelligent urban systems, yet it poses formidable challenges due to uncontrolled variations in illumination, viewpoint, background clutter, occlusion, and human motion. This paper presents a systems-oriented analysis of multi-granularity attention fusion as a robust architectural paradigm for video-based person re-identification. Moving beyond purely model-centric treatments, the discussion frames re-identification within a complex socio-technical infrastructure that spans edge computing, privacy governance, and long-term sustainability. We examine how attention mechanisms operating at multiple spatial, temporal, channel, and part-based granularities can be systematically fused to improve re-identification accuracy while accommodating the constraints of real-world deployment. The analysis foregrounds structural trade-offs among computational cost, latency, modularity, and model interpretability, and it situates attention fusion design within broader debates on algorithmic fairness, accountability, and urban data governance. By integrating conceptual investigation, cross-domain comparison, and forward-looking policy reflection, this paper offers an interdisciplinary perspective on the design and deployment of attentive person re-identification systems in urban surveillance networks. The article ultimately argues that sustainable urban re-identification requires a holistic fusion not only of visual features but also of engineering robustness, ethical safeguards, and adaptive operational architectures.

Keywords

person re-identification, multi-granularity attention, video analysis, urban surveillance, feature fusion, deep learning, system architecture, algorithmic fairness.

1. Introduction

Urban environments are increasingly instrumented with distributed camera networks that produce vast volumes of video data, creating a pressing need for automated systems capable of recognizing individuals across disjoint views. Person re-identification, the task of associating images or video tracks of the same identity captured by non-overlapping cameras, has therefore become a foundational component of intelligent transportation, public safety analytics, and smart city governance [1], [2]. Yet, dynamic urban scenes introduce extreme variability: pedestrians move under fluctuating lighting, partial occlusions are ubiquitous, clothing appearance shifts over time, and camera characteristics differ substantially across the network. These conditions render traditional identity matching brittle and underscore the demand for representation learning mechanisms that can selectively emphasize discriminative visual evidence while suppressing spurious signals. Multi-granularity attention fusion has emerged as a promising response to such complexity, enabling models to integrate fine-

grained local cues, mid-level structural patterns, and coarse global context in an end-to-end learnable framework [3], [4].

The central premise of this paper is that the design and deployment of multi-granularity attention mechanisms for video-based person re-identification must be understood not solely as a machine learning problem, but as a systems engineering challenge embedded within broader urban socio-technical infrastructures. Attention fusion architectures are not monolithic; they involve deep structural choices concerning where attention is applied, how features are aggregated across temporal sequences, and how multiple attention heads interact. These choices in turn shape the computational footprint, energy consumption, latency characteristics, and robustness of deployed systems. Moreover, the introduction of attention-based systems into civic surveillance raises urgent questions of fairness, transparency, and accountability that extend far beyond standard performance metrics. Thus, a system-level perspective—one that encompasses architecture, infrastructure, governance, and policy—is indispensable for advancing the field responsibly.

This article undertakes such an interdisciplinary analysis. It begins by characterizing the urban surveillance ecosystem and the specific re-identification challenges posed by dynamic scenes. It then unpacks the conceptual foundations of multi-granularity attention, discussing spatial attention, temporal attention, channel-wise recalibration, and part-based attention as complementary modalities. The core of the paper is dedicated to architectural fusion strategies, including early fusion, hierarchical fusion, and late gating, with a focus on their implications for deployment at scale. Subsequent sections examine infrastructure considerations such as edge-cloud partitioning, sustainability, and robustness under distribution shift, followed by a sustained engagement with governance and policy dimensions. Throughout, attention is directed to practical trade-offs and forward-looking design principles rather than to incremental model refinements. The goal is to provide a holistic account that bridges the gap between algorithmic innovation and responsible urban system integration.

2. The Urban Surveillance Ecosystem and Person Re-Identification Challenges

The modern urban surveillance ecosystem is a heterogeneous assemblage of Internet Protocol cameras, edge computing nodes, cloud analytics platforms, and human-in-the-loop monitoring centers, governed by an overlapping patchwork of municipal, regional, and sectoral regulations. In such systems, video-based person re-identification serves as a linking function: it associates fragmented tracklets across cameras to build continuous movement trajectories, enabling applications ranging from traffic flow analysis to forensic search. However, the very nature of the urban environment—its density, diversity, and constant flux—imposes a distinctive set of demands. Unlike controlled indoor settings with relatively stable lighting and limited occlusions, outdoor urban scenes exhibit large illumination swings between direct sunlight and deep shadow, drastic perspective changes due to cameras mounted at various heights and angles, and a high density of moving and static objects that frequently occlude pedestrians. Furthermore, individuals may alter their appearance by adding or removing outer garments, carrying bags, or changing accessories, a problem known as clothing variation that severely degrades the performance of models reliant on static appearance features [2], [5].

These conditions have motivated a shift from image-based re-identification, which assumes a single representative snapshot per identity, to video-based approaches that analyze sequences of frames, thereby capturing motion dynamics and temporal coherence [2], [6]. Video offers the advantage of multiple observations of the same person under varying conditions within a single camera view, which can be exploited to build richer and more invariant representations.

However, video also introduces its own challenges: sequences contain redundant information, moving backgrounds, and noisy frames, and the sheer volume of data demands efficient aggregation mechanisms. Temporal pooling strategies, recurrent networks, and spatiotemporal attention models have all been applied, yet the core difficulty remains how to identify the sparse set of frames, regions, and spatial frequencies that are truly identity-discriminative across cameras [7]. Multi-granularity attention represents a principled approach to this difficulty by decomposing the problem into distinct levels of visual abstraction and jointly learning what to attend to at each level.

3. Multi-Granularity Attention Mechanisms: Conceptual Foundations

Attention mechanisms in deep learning can be understood as learned gating functions that modulate the contribution of different parts of an input to a representation. In the context of person re-identification, attention is most useful when it operates at multiple granularities that correspond to the natural structure of human appearance and motion. We can distinguish four primary granularity levels: spatial attention that selects salient image regions such as the torso or head; temporal attention that weights frames within a video sequence according to their discriminative quality; channel attention that recalibrates the feature map dimensions to emphasize semantic attributes like color, texture, or shape; and part-based attention that aligns specific body parts across views, often informed by pose estimation or spatial transformer networks [6], [8]. These levels are not independent; a frame showing a partially occluded pedestrian, for instance, might require spatial attention to focus on the visible leg region while temporal attention downweights that frame relative to another where the full body is visible, all while channel attention enhances the texture channels that are consistent across the two views.

The theoretical appeal of multi-granularity attention lies in its capacity to resolve the inherent tension between discriminative power and invariance in person representations. A feature that is too localized may fail to match across different camera perspectives because a particular body part may not be visible; conversely, a feature that is too holistic may be contaminated by background clutter and non-identity-related variations. By hierarchically attending to cues at different spatial and temporal scales, the system can construct a representation that is both selective and robust. Recent work has demonstrated that explicitly decoupling feature-driven attention streams—such as separating the processing of static appearance attributes from motion and clothing-change patterns—can further enhance robustness in clothing-changing scenarios [12]. This conceptual decoupling aligns naturally with multi-granularity design, because different granularities can be assigned different roles in modeling stable versus transient identity characteristics.

From a systems perspective, the multi-granularity attention paradigm introduces important architectural degrees of freedom. Designers must decide whether attention modules are chained sequentially, with early spatial attention feeding into later temporal aggregation, or organized in parallel branches whose outputs are subsequently fused. They must also determine the granularity of temporal attention: operating directly on raw frames, on short video clips, or on clip-level features aggregated by three-dimensional convolutions. Each configuration embodies a distinct trade-off between computational cost and representational richness. Sequential designs tend to be more interpretable and easier to optimize, as each level receives a refined input from the previous stage, but they can suffer from error propagation. Parallel designs, on the other hand, can leverage complementary cues independently but require careful fusion strategies to avoid suboptimal interference. These design considerations

are often underrepresented in the re-identification literature, which tends to report standard performance benchmarks without evaluating system-level feasibility.

4. Architectural Design and Fusion Strategies in Video Re-Identification

The architectural integration of multi-granularity attention into a full video re-identification pipeline involves decisions at three levels: the base feature extractor, the attention modules themselves, and the fusion operations that combine attention-weighted features into a final identity descriptor. State-of-the-art systems typically employ a convolutional neural network backbone pretrained on large-scale image recognition as the base feature extractor, producing a sequence of spatial feature maps for each video frame [9]. These maps are then refined by spatial attention, often implemented as a small sub-network that predicts a mask highlighting informative regions. The resulting attended feature maps are aggregated over time by a temporal attention mechanism, which may output either a single sequence-level representation or a set of key-frame features. Finally, the aggregated representation is projected into a compact embedding space using loss functions that encourage intra-identity compactness and inter-identity separation, with the triplet loss and its variants being widely adopted for this purpose [5], [11].

Fusion strategies are particularly critical because they govern how information derived from different granularity levels is combined. Early fusion schemes inject attention-modulated features into a joint representation before the final classification or metric learning objective, while late fusion architectures maintain separate branches for each granularity stream and combine their outputs at the embedding level through concatenation, weighted summation, or learnable gating networks. Hierarchical fusion offers a middle ground, where features from coarser granularities guide the attention at finer granularities through top-down pathways. For example, a global scene-level attention map might modulate the region proposal process for part-based attention, ensuring that high-resolution part features are only extracted from pedestrian-occupied regions. This biologically inspired design, reminiscent of the coarse-to-fine processing in the primate visual system, has shown the potential to reduce both computation and false-positive attentional foci [10], [13].

The choice of fusion strategy has direct implications for deployment. Late fusion architectures are inherently modular, allowing individual attention branches to be updated, compressed, or even deployed on different physical devices without retraining the entire system. This modularity is highly desirable in urban surveillance contexts, where camera nodes may have heterogeneous computational capabilities and network bandwidth. Early fusion, by contrast, tends to produce more compact representations that are efficient for communication between nodes but is less flexible in the face of component-level changes. Moreover, the interaction between attention granularity and temporal processing must be carefully calibrated for real-time operation. A system that performs exhaustive multi-granularity attention on every frame of a high-resolution video stream is unlikely to meet the latency requirements of live tracking applications unless aggressive pruning and quantization are applied [14]. Thus, architecture design for video re-identification is as much a problem of resource management and adaptability as it is of accuracy maximization.

5. System-Level Deployment and Infrastructure Considerations

Deploying multi-granularity attention fusion models in operational urban surveillance networks raises a suite of infrastructure-level concerns that are seldom addressed in the experimental re-identification literature. The first concern is the partitioning of computation

between edge devices, on-premises servers, and cloud platforms. Camera-embedded processors are becoming increasingly capable of running lightweight neural networks, and recent advances in model distillation and neural architecture search have enabled spatial attention modules to be deployed directly on edge hardware. However, temporal attention across multiple frames and video-level multi-granularity fusion often require larger memory footprints and greater floating-point operations, making them candidates for execution on centralized servers. The design of an optimal edge-cloud split must balance latency, bandwidth consumption, power constraints, and the need for data minimization under privacy regulations [15]. A common architectural pattern is to perform spatial attention and person detection at the edge, transmitting only compact feature descriptors or attended body-part crops to a central server where temporal fusion and cross-camera matching occur. This approach reduces the transmission of raw video, aligning with privacy-by-design principles, but it imposes strict requirements on the invariance of edge-extracted features to compression and communication errors.

Sustainability is another crucial dimension. Large-scale urban re-identification infrastructures, if widely deployed, would entail continuous inference across thousands of cameras, with significant cumulative energy consumption. Multi-granularity attention models, particularly those with multiple parallel attention heads and dense feature map operations, can be computationally intensive. Quantifying the energy-delay product per query and exploring dynamic attention policies—where the granularity of attention is adaptively reduced during periods of low scene complexity—are necessary steps toward sustainable operation [16]. Furthermore, the lifecycle of such systems must account for hardware refresh cycles, software updates, and model retraining in response to evolving urban environments and population demographics. Without explicit planning for sustainability, the environmental and financial costs of at-scale re-identification may outweigh its societal benefits, especially in resource-constrained municipal contexts.

Robustness and resilience to distribution shift constitute a third infrastructure-level imperative. Urban scenes undergo both seasonal and long-term changes: foliage density, pedestrian clothing patterns, and even the physical configuration of camera mounts evolve. A multi-granularity attention model trained on a fixed dataset may suffer severe performance degradation when deployed years later unless its attention mechanisms can dynamically recalibrate. Online adaptation techniques, such as test-time attention recalibration using momentum-based batch normalization updates or episodic memory replay, represent important directions for maintaining reliability [17]. Yet these techniques introduce additional system complexity, requiring monitoring frameworks that can detect performance drift and trigger safe model adaptation procedures without catastrophic interference. From a systems engineering standpoint, robustness must therefore be designed as a first-class property encompassing not just model accuracy but also graceful degradation, fault tolerance, and continuous verification.

6. Governance, Fairness, and Policy Implications

The deployment of attentive person re-identification technologies in dynamic urban scenes carries profound governance and ethical implications that demand proactive scholarly engagement. Person re-identification acts as a de-anonymization mechanism in public spaces, fundamentally altering the relationship between citizens and urban authorities. In democratic societies, such systems must operate within transparent legal frameworks that define permissible use cases, data retention periods, access controls, and independent oversight

structures. However, the very characteristics that make multi-granularity attention powerful—its ability to selectively amplify subtle visual cues and integrate information across disjoint views—also amplify concerns related to profiling, disproportionate surveillance of marginalized communities, and chilling effects on public assembly. Algorithmic fairness is therefore a non-negotiable design requirement, not a post-hoc remediation step. Research has shown that re-identification models often exhibit differential accuracy across demographic groups, with error rates varying significantly by skin tone, gender presentation, and clothing style [18]. Multi-granularity attention mechanisms, if trained on imbalanced datasets, may learn to disproportionately focus on group-correlated features that further entrench biased outcomes.

Addressing these fairness challenges requires interventions at multiple levels of the system stack. At the data level, curated training datasets must be audited for demographic representativeness and potential proxy biases, and data collection must itself be governed by informed consent protocols where feasible. At the model level, fairness-aware attention training can be incorporated by adding regularization terms that penalize attention maps statistically associated with protected attributes, or by applying adversarial disentanglement to separate identity-relevant features from demographic cues [19]. From a systems governance perspective, fairness cannot be guaranteed solely through technical calibration; it also demands procedural mechanisms such as algorithmic impact assessments, community consultation processes, and accessible channels for contestation and redress. In the United Kingdom and the European Union, emerging regulatory instruments concerning artificial intelligence in law enforcement and public surveillance provide initial frameworks, but their translation into concrete technical standards for re-identification remains nascent and contested [20].

The policy implications extend to the international scale, as many urban surveillance technologies are developed in one jurisdiction and deployed in another with markedly different governance traditions. Multi-granularity attention models embedded in commercial off-the-shelf analytics products create complex accountability chains where the responsibility for biased or erroneous re-identification is diffused across hardware manufacturers, software developers, system integrators, and municipal operators. Establishing clear lines of responsibility and transparency mandates—such as model documentation requirements that describe the granularity of attention mechanisms, training data provenance, and known failure modes—is essential for building public trust and enabling effective oversight. The required reference [12] illustrates how advanced attention decoupling techniques can improve performance under challenging clothing-change conditions, a scenario highly relevant to urban settings where individuals may change appearance between camera encounters. Yet such technical progress must be accompanied by governance frameworks that ensure these capabilities are used to serve legitimate public interests without undermining fundamental rights.

Furthermore, the urban deployment of multi-granularity attention fusion systems intersects with broader smart city governance debates about data ownership, interoperability, and vendor lock-in. Cities that invest in proprietary re-identification analytics may find themselves dependent on a single vendor for updates, auditing, and maintenance, potentially ceding public authority over critical infrastructure. Open standards for feature representation and attention model exchange, analogous to open data formats in geographic information systems, could mitigate these risks by enabling competitive procurement and forensic

auditability. Such standards would need to balance the detailed specification of attention granularity parameters with the flexibility to accommodate rapidly evolving deep learning architectures. The interdisciplinary challenge thus lies in co-designing technical architectures and governance structures that are simultaneously high-performance, fair, interpretable, and institutionally sustainable.

7. Conclusion

Video-based person re-identification in dynamic urban scenes stands at the intersection of computer vision, systems engineering, and urban policy, demanding a deeply interdisciplinary analytic approach. Multi-granularity attention fusion has emerged as a compelling algorithmic paradigm, offering principled ways to combine spatial, temporal, channel, and part-based cues into identity representations that are robust to the complex variability of urban environments. However, the path from laboratory accuracy to responsible urban deployment is fraught with system-level challenges. This paper has argued that sustainable and trustworthy re-identification systems require careful attention to architectural modularity, edge-cloud partitioning, energy efficiency, and online robustness, as well as equally rigorous engagement with issues of fairness, accountability, and regulatory compliance. Multi-granularity attention is not a singular technical fix but a design space that can be navigated more or less wisely depending on the governance frameworks, infrastructure investments, and ethical commitments that surround its implementation. Future work should focus on co-designing attention architectures and institutional mechanisms that together enable urban surveillance systems to enhance public safety while upholding democratic values. Only through such holistic integration can the promise of intelligent urban analytics be realized without compromising the foundational principles of open and equitable public spaces.

References

1. Zheng, L., Yang, Y., & Hauptmann, A. G. (2016). Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984.

2. Ristani, E., Solera, F., Zou, R. S., Cucchiara, R., & Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In Computer Vision – ECCV 2016 Workshops (pp. 17-35). Springer.

3. Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). DeepReID: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 152-159).

4. Varior, R. R., Haloi, M., & Wang, G. (2016). Gated siamese convolutional neural network architecture for human re-identification. In European conference on computer vision (pp. 791-808). Springer.

5. Hermans, A., Beyer, L., & Leibe, B. (2017). In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737.

6. Wang, G., Yuan, Y., Chen, X., Li, J., & Zhou, X. (2018). Learning discriminative features with multiple granularities for person re-identification. In Proceedings of the 26th ACM international conference on Multimedia (pp. 274-282).

7. Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European conference on computer vision (ECCV) (pp. 480-496).

8. Chen, Y., Zhu, X., Zheng, W. S., & Lai, J. H. (2018). Person re-identification by camera correlation aware feature augmentation. IEEE transactions on pattern analysis and machine intelligence, 40(4), 961-974.

9. Liu, C. T., Wu, C. W., Wang, Y. C. F., & Chien, S. Y. (2019). Spatially and temporally efficient non-local attention network for video-based person re-identification. arXiv preprint arXiv:1908.01683.

10. Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., & Chen, X. (2019). VRSTC: occlusion-free video person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7183-7192).

11. Luo, H., Gu, Y., Liao, X., Lai, S., & Jiang, W. (2019). Bag of tricks and a strong baseline for deep person re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 0-0).

12. 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.

13. Wang, Q., Yuan, Y., & Li, X. (2019). Learning video representations for person re-identification with temporal pooling. IEEE Transactions on Circuits and Systems for Video Technology, 30(7), 2151-2161.

14. Liu, H., Jie, Z., Jayashree, K., Qi, M., Jiang, J., Yan, S., & Feng, J. (2017). Video-based person re-identification with accumulative motion context. IEEE transactions on circuits and systems for video technology, 28(10), 2788-2802.

15. Li, S., Bak, S., Carr, P., & Wang, X. (2018). Diversity regularized spatiotemporal attention for video-based person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 369-378).

16. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., & Tian, Q. (2017). Pose-driven deep convolutional model for person re-identification. In Proceedings of the IEEE international conference on computer vision (pp. 3960-3969).

17. Zheng, Z., Zheng, L., & Yang, Y. (2019). Pedestrian alignment network for large-scale person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 29(10), 3037-3045.

18. Xu, J., Zhao, R., Zhu, F., Wang, H., & Ouyang, W. (2018). Attention-aware compositional network for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2119-2128).

19. Song, C., Huang, Y., Ouyang, W., & Wang, L. (2018). Mask-guided contrastive attention model for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1179-1188).

20. Zhang, J., Wang, N., & Zhang, L. (2019). Multi-shot pedestrian re-identification via sequential decision making. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11078-11087).