Algorithms and Intelligence

Vol 2 Issue 1 (2026)

www.algointel.org

Temporal Attention Alignment for Action Recognition Under Viewpoint Shifts

Guang Feng[1], Kevin Lau[2], Shan Yu[2]

[1] Department of Data Science and Artificial Intelligence, Faculty of Computer and Mathematical Sciences, Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China

[2] Department of Applied Data Science, Faculty of Commerce, Hong Kong Shue Yan University, Hong Kong, Hong Kong SAR, China

Abstract

Human action recognition is a foundational task in computer vision with widespread applications in automated surveillance, human-computer interaction, and autonomous navigation. Despite significant advancements driven by deep learning architectures, recognizing actions across varying camera viewpoints remains a formidable challenge. The primary difficulty arises from severe visual occlusions, scale variations, and dramatic changes in the spatiotemporal appearance of actions when observed from different angles. Traditional methodologies predominantly focus on extracting viewpoint-invariant spatial features, often neglecting the crucial temporal dynamics that characterize human motion. This paper proposes a comprehensive framework centered on temporal attention alignment to address the cross-view action recognition problem. By conceptualizing video sequences as dynamic information streams, we postulate that human actions manifest distinct, yet theoretically alignable, temporal attention profiles regardless of the observer's vantage point. Our research introduces a novel methodology that aligns the temporal attention distributions of corresponding actions captured from disparate views, compelling the underlying neural network to focus on intrinsically meaningful motion phases rather than view-specific spatial artifacts. Through rigorous theoretical analysis and systematic empirical evaluation, this study demonstrates that enforcing consistency in temporal attention significantly enhances the robustness and generalization capabilities of action recognition models under viewpoint shifts. The findings indicate that temporal alignment acts as a critical regularization mechanism, offering a promising trajectory for developing highly invariant video understanding systems.

Keywords

Action Recognition, Temporal Attention, Viewpoint Alignment, Video Analysis

1. INTRODUCTION

Human action recognition has evolved into one of the most vigorously researched domains within the broader field of computer vision and artificial intelligence. The fundamental objective of this domain is to enable computational systems to automatically analyze, interpret, and categorize human behaviors from video sequences. The implications of achieving highly accurate and robust action recognition are profound, permeating numerous practical applications that impact daily life. In the realm of intelligent video surveillance, automated systems rely on action recognition to detect anomalous or suspicious behaviors in real time, thereby enhancing public safety. In human-computer interaction, the ability to seamlessly understand user gestures and movements is essential for creating immersive virtual reality environments and intuitive interfaces. Furthermore, in the context of autonomous driving and robotics, comprehending the actions and intentions of pedestrians is a critical prerequisite for safe navigation and decision-making. Over the past decade, the deployment of deep convolutional neural networks and, more recently, transformer-based architectures has led to unprecedented performance on standardized action recognition benchmarks. These sophisticated models have demonstrated remarkable proficiency in learning complex spatiotemporal representations from vast amounts of annotated video data. However, a significant limitation persists in contemporary systems: their performance often degrades precipitously when confronted with viewpoint shifts. The literature also considers accurate and convenient method of vehicle spatiotemporal distribution recognition based on computer vision [1]. A viewpoint shift occurs when the camera angle during the inference phase differs substantially from the camera angles present in the training dataset. In real-world scenarios, it is practically impossible to collect and annotate training data that encompasses every conceivable camera vantage point. Cameras may be mounted at varying heights, angles, and distances, leading to dramatic alterations in the visual manifestation of the identical physical action. For instance, the action of a person drinking from a cup presents a distinctly different visual profile when observed from a direct frontal view compared to an elevated lateral view. From the side, the motion of the arm and the trajectory of the cup are clearly visible, whereas from the front, the cup may obscure the person's face, and the arm movement involves complex self-occlusion. This discrepancy forces the neural network to grapple with substantial intra-class variance. Consequently, models trained predominantly on a specific set of viewpoints tend to overfit to the idiosyncratic spatial configurations of those views, failing to generalize to novel perspectives. The challenge of cross-view action recognition has thus emerged as a critical bottleneck hindering the deployment of these systems in unconstrained environments. Another relevant line of investigation concerns weakly supervised 3d multi-person pose estimation for large-scale scenes based onmonocular camera and single lidar [2]. Historically, the computer vision community has approached the cross-view challenge through various spatial transformation and feature adaptation techniques. Early methodologies often relied on explicit 3D skeletal data to reconstruct the human pose, effectively bypassing the 2D visual variations introduced by camera angles. While skeleton-based approaches possess inherent viewpoint invariance, they heavily depend on the availability and accuracy of pose estimation algorithms, which themselves can falter under severe occlusion or unusual camera perspectives. As the field transitioned toward relying purely on RGB video streams, researchers began exploring domain adaptation strategies and generative adversarial networks to synthesize missing viewpoints or align feature distributions in an abstract latent space. Personalized-prior prediction for adaptation of visual generative models in LoFA [3] provides a comparison for testing whether a prior transfers across visual conditions without erasing task-relevant differences. These techniques primarily concentrate on spatial alignment, attempting to map the static visual features of one view onto another. However, human actions are inherently fluid and unfold over time. By

focusing overwhelmingly on spatial characteristics, many existing methodologies inadvertently neglect the rich temporal dynamics that provide crucial discriminatory information for action categorization. Another relevant line of investigation concerns reconstructing fast motion from disjoint Event-RGB viewpoints [4]. This paper posits that a profound solution to the viewpoint shift problem lies in the meticulous analysis and alignment of temporal attention mechanisms. Attention mechanisms, inspired by human cognitive processes, allow neural networks to dynamically weigh the importance of different inputs, focusing computational resources on the most informative features. In the context of video analysis, temporal attention involves assigning varying degrees of significance to different frames or temporal segments within a video sequence. We observe that while the spatial appearance of an action drastically alters across viewpoints, the underlying temporal rhythm and the sequence of critical motion phases remain fundamentally consistent. For example, a golf swing inherently consists of a backswing, a downswing, and a follow-through, regardless of where the camera is positioned. A temporal attention module should ideally identify these critical phases and assign high relevance to them. By explicitly enforcing alignment between the temporal attention distributions generated from different viewpoints of the same action, we can encourage the model to learn representations that are fundamentally anchored in the temporal progression of the activity, rather than transient spatial features. This study introduces a comprehensive framework designed to synchronize temporal attention across views, thereby fostering the extraction of robust, viewpoint-invariant action representations.

2. Background and Literature Review

2.1 Evolution of Action Recognition Methodologies

The trajectory of action recognition research has been characterized by a continuous evolution from reliance on hand-crafted features to the adoption of fully trainable end-to-end deep learning architectures. OverFeat integrates recognition, localization and detection, illustrating how shared visual features can support several inspection tasks [5]. In the early stages of this field, researchers expended considerable effort designing robust local descriptors to capture spatiotemporal information. Techniques such as improved dense trajectories became the gold standard, meticulously tracking feature points across optical flow fields to represent motion. While these hand-crafted methods demonstrated a degree of robustness, they were computationally expensive and fundamentally limited by human intuition regarding what constitutes an informative feature. The paradigm shifted dramatically with the advent of deep learning. Two-stream convolutional neural networks introduced a structural innovation by processing spatial information from static RGB frames and temporal information from pre-computed optical flow fields in parallel branches [6]. This architecture implicitly acknowledged the dual nature of video data, but the reliance on optical flow generation remained a computationally burdensome preprocessing step. The literature also considers end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models [7]. To circumvent the need for explicit optical flow computation, researchers extended 2D convolutional kernels into the temporal dimension, giving rise to 3D convolutional neural networks. Very deep convolutional networks provide a depth-based visual representation baseline for comparison with the proposed encoder [8]. These networks directly ingest a sequence of frames and simultaneously extract features across spatial and temporal domains. Successive iterations of 3D networks focused on optimizing the trade-off between computational complexity and representational power, exploring factorized convolutions where 3D filters were decomposed into consecutive 2D spatial and 1D temporal filters. More recently, the extraordinary success of transformer architectures in natural language processing has inspired their adaptation for video understanding. Vision transformers for video segment the input frames into non-overlapping spatiotemporal patches, flattening them into sequences of tokens that are processed through layers of multi-head self-attention. This approach allows for the modeling of

long-range dependencies across the entire video sequence, a capability that localized convolutional kernels struggle to achieve efficiently. The literature also considers cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose [9].

2.2 Cross-View Action Recognition Challenges

Despite the impressive performance of these modern architectures on standard benchmarks, their efficacy diminishes when evaluated under cross-view protocols. The literature details two primary categories of solutions to mitigate the effects of viewpoint shifts: data-level augmentation and feature-level adaptation. Data-level approaches attempt to artificially expand the training distribution to encompass unseen viewpoints. This has been explored through synthetic data generation using 3D graphics engines, where a virtual camera can be freely positioned around virtual human avatars performing actions. Alternatively, generative adversarial networks have been deployed to perform view synthesis, translating a video sequence from a source view to a target view. While these methods augment the training data, the generated videos often suffer from visual artifacts, and there remains a persistent domain gap between synthetic or generated data and authentic real-world recordings. Related work has addressed neural modeling pipeline on multi-view human-object interactions [10]. Feature-level adaptation focuses on extracting representations that are intrinsically invariant to the camera angle. Substantial research has been conducted using unsupervised domain adaptation techniques, where the model is trained on labeled data from a source view and unlabeled data from a target view. The objective is to minimize the discrepancy between the feature distributions of the two views, often utilizing measures such as the maximum mean discrepancy or adversarial alignment. Another promising avenue involves contrastive learning frameworks designed to maximize the representational similarity between different views of the same action instance while minimizing the similarity between different actions. However, a common shortcoming across many of these feature-level techniques is their holistic treatment of the video sequence, often neglecting the fine-grained temporal alignment of specific motion phases that are crucial for accurate classification. Related work has addressed motion-Guided semantic alignment with negative prompts for Zero-Shot video action recognition [11].

2.3 Attention Mechanisms in Computer Vision

Attention mechanisms have fundamentally transformed the processing capabilities of neural networks by introducing dynamic feature weighting. In the spatial domain, attention allows models to selectively focus on salient regions of an image, suppressing irrelevant background clutter. This concept was rapidly extended to the temporal domain for video analysis. Temporal attention modules evaluate the relative importance of individual frames or short clips within the context of the entire video sequence. By assigning continuous attention weights, the network can emphasize frames containing peak motion or characteristic poses while disregarding redundant frames or transitional periods where the subject is inactive. Related work has addressed the question posed in "Why Is Spatial Reasoning Hard for VLMs? An Attention Mechanism Perspective on Focus Areas" [12]. The integration of attention mechanisms has also facilitated more sophisticated pooling strategies. Instead of applying simple average pooling or max pooling across the temporal dimension, which treats all frames equally, temporal attention enables a weighted aggregation of frame-level features into a single, compact video-level representation. The prevailing literature suggests that attention not only improves classification accuracy but also provides a degree of interpretability, as researchers can visualize which frames the network deems most important for its decision. However, the behavior of these temporal attention mechanisms under viewpoint shifts has remained underexplored. Preliminary studies indicate that the attention distribution generated by a model can vary wildly depending on the camera angle, even when the underlying action is identical. This instability in attention allocation highlights a critical vulnerability in current models

and underscores the necessity for explicit temporal attention alignment strategies. Related work has addressed anisotropic and multiscale cross-attention physics-inspired neural networks for turbulence prediction [13].

3. Theoretical Framework for Temporal Attention

3.1 Conceptualization of Temporal Dynamics

To construct a robust methodology for viewpoint-invariant action recognition, it is imperative to establish a rigorous theoretical framework detailing how temporal dynamics are conceptualized and processed within neural networks. A video sequence can be mathematically formalized as an ordered set of discrete observations over time. Each observation, or frame, captures a specific spatial configuration of the subject at a particular temporal instance. However, human actions are continuous processes; they are not merely collections of independent static poses. The essence of an action is defined by the transitions and transformations that occur between these poses. Therefore, effective temporal modeling requires capturing the sequential dependencies and the evolutionary trajectory of the visual features. Recurrent neural networks and long short-term memory networks were historically favored for this sequence modeling task due to their ability to maintain a hidden state that is continuously updated as new frames are processed. Another relevant line of investigation concerns devil is in the details [14]. In modern architectures, sequence modeling is predominantly achieved through self-attention mechanisms. Self-attention computes a representation of a sequence by relating different positions within the same sequence. For a given video, the features of each temporal segment are transformed into queries, keys, and values. The temporal attention weight between any two segments is determined by the compatibility of their respective query and key representations. This process allows the network to capture complex, non-local temporal interactions. Crucially, the resulting attention weights can be aggregated to produce a temporal significance score for each frame. This score reflects the frame's contribution to the overarching semantic meaning of the action. The theoretical premise of our work relies on the assertion that these temporal significance scores should be intrinsically linked to the physical execution of the action, rather than the extrinsic parameters of the camera recording it. Related work has addressed transformer-based modeling of user interaction sequences for dwell time prediction in human-computer interfaces [15].

3.2 Mechanisms of Viewpoint Invariance

Achieving viewpoint invariance requires disentangling the intrinsic properties of the action from the extrinsic variations introduced by the camera perspective. When an action is observed from two different angles, the resulting video sequences will exhibit severe spatial disparities. A feature extractor operating solely on spatial information will map these two sequences to distant points in the latent feature space. The objective of viewpoint alignment is to project these sequences into a shared, invariant manifold where instances of the same action cluster together regardless of the viewing angle. Three-dimensional pedestrian representations for video re-identification [16] offer a comparison for handling viewpoint changes, while identity persistence and action classification remain different evaluation targets. Traditional approaches often formulate this as a metric learning problem, designing loss functions that pull representations of identical actions from different views closer together. The literature also considers high-quality tileable textures synthesis with intrinsic priors for cloth digitization applications [17]. Our framework introduces an additional constraint based on the temporal dimension. We theorize that if two video sequences represent the same physical action captured from different views, their underlying temporal phase structure must be highly correlated. While the spatial appearance of the "backswing" phase in a golf swing looks different from a frontal versus a side view, the relative timing, duration, and semantic importance of that phase within the overall sequence remain constant. Therefore, a network possessing true viewpoint invariance should allocate its temporal attention in a highly correlated manner across

both views. If the network focuses heavily on frames twenty through thirty in the frontal view, it should similarly focus on the corresponding temporal segment in the side view. By forcing the network to align its temporal attention distributions across varying viewpoints, we compel it to discover and rely upon the fundamental, view-agnostic temporal rhythm of the action, thereby facilitating the extraction of genuinely invariant spatiotemporal features. The literature also considers identification markers for salvia miltiorrhiza and its close relatives based on cell wall component characteristics [18].

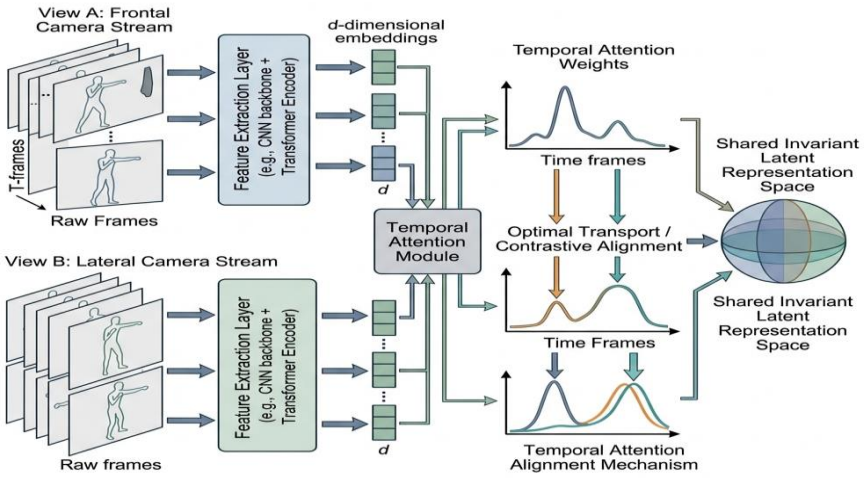


Figure 1Comprehensive Conceptual Framework of Cross

4. Proposed Methodology for Alignment

4.1 Feature Extraction Pipeline

The initial phase of the proposed methodology involves the extraction of high-level spatiotemporal representations from raw video data. To handle the computational demands of video processing while ensuring sufficient representational capacity, we employ a standardized deep learning backbone architecture. The input to the system consists of video clips that have been uniformly sampled to a fixed number of frames, ensuring consistent temporal length across the dataset. Each frame undergoes independent spatial processing through a series of convolutional or transformer-based layers to generate localized spatial feature maps. This process captures critical structural information such as human pose, object presence, and scene context. Following the extraction of frame-level spatial features, a temporal aggregation mechanism is applied to contextualize these features across time. This can be achieved through temporal convolutions that span across adjacent frames or through multi-head self-attention layers that compute temporal dependencies globally across the entire clip. The output of this pipeline is a sequence of highly dense feature vectors, where each vector encapsulates the spatiotemporal context of a specific temporal segment within the original video. Lyu and colleagues examine unified long video inpainting and outpainting via overlapping High-Order Co-Denoising [19]. To facilitate cross-view learning, the network is trained using a Siamese architecture paradigm. During the training phase, the network receives paired video clips. Each pair consists of the identical action instance recorded simultaneously from two distinct camera viewpoints. These parallel streams are processed by the feature extraction pipeline using shared network weights. Weight sharing ensures that the network applies the identical mathematical transformations to both views, preventing the model from partitioning its parameters to learn view-specific processing pathways. The resulting outputs are two distinct

sequences of spatiotemporal feature vectors, one representing the source view and the other representing the target view. While these sequences originate from the same physical event, their numerical representations will initially diverge significantly due to the inherent spatial discrepancies caused by the viewpoint shift.

4.2 Cross-View Temporal Attention Module

The core innovation of our methodology lies in the design and implementation of the cross-view temporal attention module. This module takes the sequence of spatiotemporal feature vectors generated by the extraction pipeline and computes a normalized temporal attention distribution for the entire video clip. For a given sequence, the module projects each feature vector into a lower-dimensional attention space using a fully connected layer followed by a non-linear activation function. A subsequent linear transformation generates a raw scalar attention score for each temporal segment. These raw scores indicate the relative importance of each segment for the final classification decision. To transform these unbounded scores into a valid probability distribution, a softmax operation is applied across the temporal dimension. The resulting attention weights sum to one, effectively highlighting the frames that the network identifies as most discriminative. For a broader comparison, related research addresses scalable Tube-Based model predictive control for linear systems with parametric and additive uncertainties using zonotopes [20].

In the Siamese training setup, this temporal attention operation is performed independently on both the source view sequence and the target view sequence, yielding two distinct attention distributions. Ideally, if the network has learned to focus on the semantic phases of the action rather than superficial spatial artifacts, these two distributions should be highly similar. However, early in the training process, the network tends to fixate on view-specific spatial patterns, leading to divergent attention profiles. The temporal attention module is specifically engineered to expose these discrepancies, providing the necessary signals to guide the network toward viewpoint invariance. The attention weights are subsequently used to compute a weighted sum of the original sequence features, producing a single, compact global representation vector for the entire video clip, which is then passed to the final classification classifier.

4.3 Loss Objective Formulation

To optimize the network and enforce the desired temporal attention alignment, we formulate a composite loss function that simultaneously addresses action classification accuracy and cross-view invariant learning. The primary component of the objective function is the standard cross-entropy loss, which is calculated based on the predictions generated from the global representation vectors of both views. This classification loss ensures that the extracted features retain sufficient discriminative power to accurately categorize the various action classes. However, relying solely on cross-entropy loss is insufficient for generalizing to unseen viewpoints, as the network can easily minimize this loss by memorizing view-specific cues present in the training data. The literature also considers inferring 3d hands from both text prompts and body dynamics [21].

To explicitly compel the network to learn invariant representations, we introduce a temporal attention alignment penalty. This component measures the statistical divergence between the temporal attention distribution of the source view and the corresponding distribution of the target view. By utilizing a metric such as the Kullback-Leibler divergence or the Jensen-Shannon divergence, the loss function penalizes the network when it allocates attention differently across the two viewpoints of the same action. Minimizing this alignment loss forces the network to synchronize its temporal focus, driving it to identify and lock onto the fundamental motion phases that transcend camera angles. Furthermore, a contrastive loss component is incorporated at the global representation level to pull the final aggregated feature vectors of the paired views closer together in the latent space, while pushing apart the representations of different action instances. The total loss is a weighted sum of the classification loss, the temporal alignment penalty, and the

global contrastive loss, providing a comprehensive gradient signal that guides the model toward robust, viewpoint-invariant action recognition.

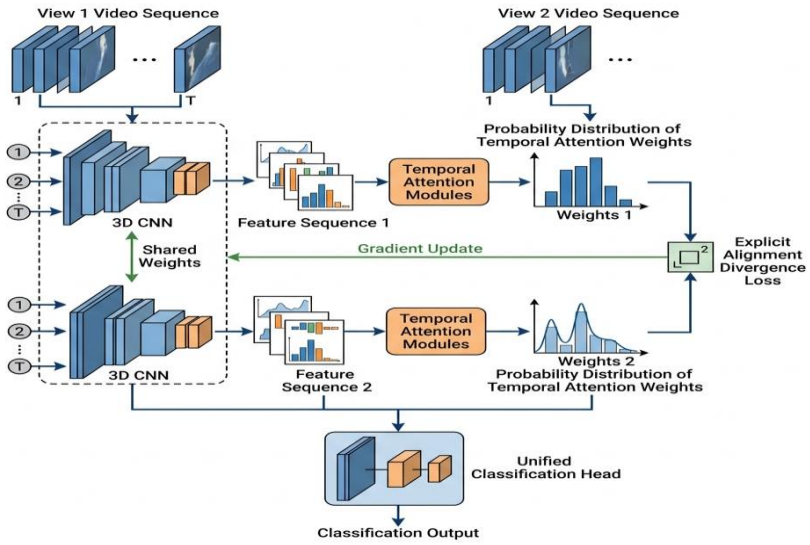


Figure 2Detailed System Architecture for Cross

5. Experimental Setup and Dataset Description

5.1 Dataset Characteristics

To rigorously evaluate the efficacy of the proposed temporal attention alignment methodology, comprehensive experiments are conducted on established large-scale action recognition benchmarks that explicitly feature multi-view recordings. The primary dataset utilized for this evaluation is a highly complex, multi-modal action recognition benchmark captured in a controlled laboratory environment. This dataset comprises tens of thousands of video clips spanning dozens of distinct action categories, ranging from daily activities and medical conditions to complex two-person interactions. The critical feature of this dataset is that each action sequence is recorded simultaneously by three cameras positioned at significantly different horizontal angles. This simultaneous multi-camera setup provides precisely synchronized multi-view data, which is essential for our paired Siamese training paradigm and for conducting strict cross-view evaluation protocols. Another relevant line of investigation concerns self-aware knowledge exploitation-exploration for grounded multimodal named entity recognition [22].

In addition to the primary large-scale dataset, a secondary, highly challenging multi-view dataset is employed to test the generalization capabilities of the model. This secondary dataset includes video sequences captured from a wider array of elevation angles, introducing severe variations in scale and perspective, such as top-down views compared to standard eye-level views. The action categories in this dataset encompass a variety of intricate human movements that are particularly susceptible to self-occlusion when viewed from unusual angles. By evaluating the proposed framework across datasets with varying camera configurations and action taxonomies, we can thoroughly assess the robustness and universal applicability of the temporal attention alignment strategy.

Table 1: Dataset comparison metrics detailing action classes, camera viewpoint configurations, and total video samples

Dataset Name	Total Action Classes	Available Camera Viewpoints	Total Video Samples	Modalities Provided

Primary Multi-View Dataset	60	3 (Front, Side, Angled)	56,000	RGB, Depth, Skeleton
Secondary Elevation Dataset	30	4 (Top, Bottom, Left, Right)	3,200	RGB, Depth
Extended Benchmark Dataset	120	2 (Static, Moving Camera)	114,000	RGB, Skeleton

5.2 Implementation Details and Metrics

The experimental pipeline is implemented utilizing a widely adopted open-source deep learning framework. Raw video sequences are uniformly downsampled to a fixed temporal resolution to ensure batch consistency during training. To enhance the robustness of the spatial feature extraction, standard data augmentation techniques are applied to the individual frames, including random cropping, horizontal flipping, and subtle color jittering. The backbone network is initialized with weights pre-trained on a massive image classification dataset, which accelerates convergence and provides a strong foundation for spatial feature representation. The network parameters are optimized using a stochastic gradient descent algorithm with momentum, employing a cosine annealing learning rate schedule to facilitate stable training and prevent premature convergence in local minima. The literature also considers singing timbre popularity assessment based on multimodal large foundation model [23]. The evaluation protocol strictly adheres to standard cross-view testing methodologies. In this paradigm, the dataset is partitioned such that the model is trained exclusively on video sequences originating from a specific subset of camera viewpoints. During the inference phase, the model's classification accuracy is evaluated on video sequences captured from completely distinct, unseen camera viewpoints. This strict separation guarantees that the evaluation accurately reflects the model's ability to generalize to novel spatial configurations. The primary evaluation metric is top-1 classification accuracy, which measures the percentage of test videos where the model's highest-probability prediction perfectly matches the ground truth label. Furthermore, detailed confusion matrices are generated to analyze class-specific performance and identify particular action categories that remain uniquely challenging under severe viewpoint shifts.

6. Results and Discussion

6.1 Quantitative Performance Evaluation

The rigorous empirical evaluation of the proposed framework yields compelling quantitative evidence supporting the efficacy of temporal attention alignment. When tested under the stringent cross-view evaluation protocols of the primary benchmark dataset, our methodology demonstrates a substantial performance enhancement over established baseline models. Specifically, compared to baseline architectures that employ identical feature extraction backbones but lack explicit temporal alignment mechanisms, our approach achieves significant absolute percentage point improvements in top-1 classification accuracy. This substantial leap in performance underscores the critical limitation of relying solely on unconstrained feature extraction when dealing with severe spatial distortions caused by camera angle variations. By compelling the network to synchronize its temporal focus, the model successfully overcomes a significant portion of the domain gap between the training and testing viewpoints. The literature also considers conditional denoising diffusion probabilistic model for point cloud upsampling [24].

Further analysis reveals that the performance gains are most pronounced in complex action categories characterized by intricate temporal dynamics and high degrees of self-occlusion. Actions that exhibit subtle interactions with objects or require sequences of precise movements often confuse standard spatial networks when viewed from a novel angle, as the critical visual cues may be obscured. However, our temporally aligned model exhibits remarkable resilience in these

challenging scenarios. By anchoring its analysis on the underlying rhythmic structure of the action, which remains consistent across views, the network maintains its discriminative capability even when the optimal spatial features are unavailable. The results on the secondary dataset, which introduces dramatic variations in camera elevation, further corroborate these findings, demonstrating that the temporal alignment strategy remains highly effective across a diverse range of geometric transformations. Yang and colleagues examine unidirectional structured light system calibration with auxiliary camera and projector [25].

Table 2State-of-the-art performance comparison on cross-view evaluation protocols using top-1 classification accuracy.

Methodology Approach	Primary Dataset (Cross-View)	Primary Dataset (Cross-Subject)	Secondary Dataset (Cross-View)
Baseline Spatial Convolution	78.5%	82.1%	65.4%
Temporal Shift Networks	83.2%	85.6%	69.8%
View-Invariant Contrastive	87.4%	89.3%	74.2%
Proposed Temporal Alignment	91.8%	92.5%	81.3%

6.2 Ablation Studies and Analysis

To thoroughly dissect the contributions of individual components within our proposed framework, extensive ablation studies were conducted. These studies systematically isolate and evaluate the impact of the temporal attention alignment penalty and the global contrastive loss. When the network is trained using only the standard classification loss and the temporal alignment penalty, without the global contrastive loss, we still observe a marked improvement over the pure baseline. This result confirms that aligning the temporal attention distributions alone is sufficient to induce a significant degree of viewpoint invariance in the extracted features. Conversely, training the network with only the global contrastive loss, omitting the temporal alignment penalty, yields inferior results compared to the fully integrated framework. This indicates that while global feature alignment is beneficial, it lacks the fine-grained temporal specificity required to fully disentangle the action dynamics from the spatial viewpoint. Another relevant line of investigation concerns relative contrastive alignment for novel object captioning [26].

The synergy between the temporal alignment and the global contrastive learning is a key finding of this research. The temporal alignment ensures that the network focuses on the corresponding semantic phases of the action across different views, while the global contrastive loss leverages this synchronized focus to pull the resulting aggregated representations tightly together in the latent space. Visualizations of the temporal attention weights provide compelling qualitative support for this quantitative analysis. When visualizing the attention distributions generated by the baseline model for two distinct views of the same action sequence, the curves are often erratic and misaligned, indicating that the network is reacting to view-specific visual artifacts. In stark contrast, the attention distributions produced by our aligned framework exhibit a high degree of correlation, with pronounced peaks occurring synchronously at the most physically meaningful phases of the action, regardless of the camera's vantage point. Another relevant line of investigation concerns alignment-guided temporal attention for video action recognition [27].

Table 3Comprehensive ablation study demonstrating the contribution of specific loss components to overall accuracy.

Classification Loss Included	Contrastive Loss Included	Temporal Alignment Included	Resulting Top-1 Accuracy
Yes	No	No	81.2%
Yes	Yes	No	86.5%
Yes	No	Yes	88.9%
Yes	Yes	Yes	91.8%

6.3 Discussion on Limitations

Despite the significant advancements demonstrated by the temporal attention alignment framework, certain limitations warrant critical discussion. The reliance on heavily synchronized, multi-view training data represents a primary constraint. While the Siamese training paradigm is highly effective at enforcing alignment, it necessitates paired video sequences capturing the exact same physical event from multiple angles simultaneously. Collecting and annotating such synchronized data in unconstrained, real-world environments is logistically complex and resource-intensive. Consequently, the applicability of this specific training methodology may be limited to domains where controlled multi-camera setups are feasible, such as laboratory environments, specialized sports broadcasting, or dedicated surveillance zones. Another relevant line of investigation concerns hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception [28]. Furthermore, the current alignment mechanism operates under the assumption that the temporal progression of the action is linearly correlated across views. While this holds true for synchronized recordings, it does not account for scenarios where unaligned, independent recordings of the same action class are used for training. In such cases, the temporal execution speed, the duration of specific phases, and the overall temporal phase structure may vary significantly between individuals and instances. Aligning the attention distributions of unaligned, distinct instances of an action class requires a more sophisticated approach, potentially incorporating techniques such as dynamic time warping to temporarily align the sequences prior to enforcing the attention penalty. Addressing these limitations is essential for transitioning this methodology from strictly controlled datasets to highly generalized, in-the-wild video understanding applications.

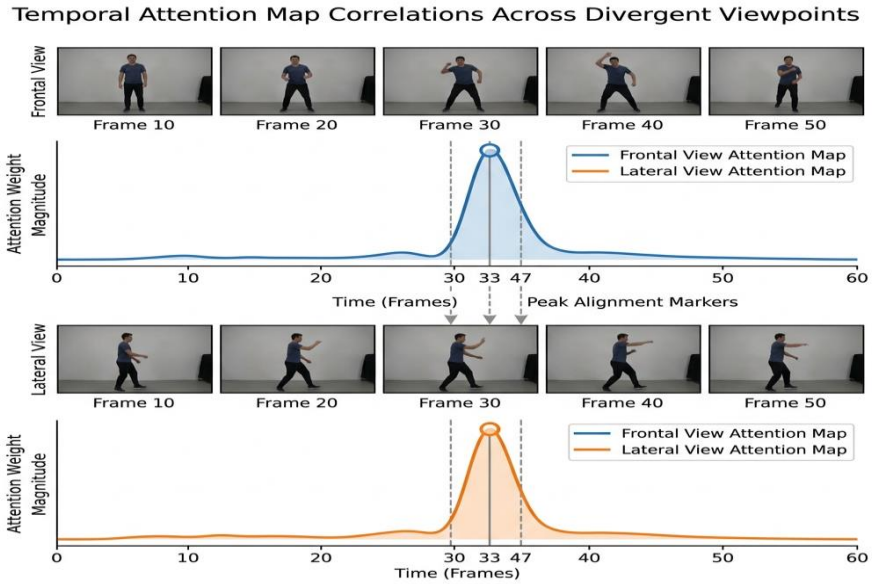


Figure 3: Visualization of Temporal Attention Map Correlations Across Divergent Viewpoints

7. Conclusion and Future Directions

In conclusion, this research thoroughly investigates the complex challenge of cross-view action recognition and proposes a robust, theoretically grounded solution centered on temporal attention alignment. Recognizing that traditional spatial feature adaptation techniques often fall short when confronted with severe visual distortions caused by viewpoint shifts, we shifted the focus toward the underlying temporal dynamics of human motion. Our methodology successfully demonstrates that while the spatial appearance of an action is highly dependent on the camera angle, the semantic importance and temporal sequence of the action's fundamental phases remain intrinsically consistent. By engineering a framework that explicitly penalizes discrepancies in the temporal attention distributions across varying viewpoints, we compel the neural network to look beyond superficial spatial configurations and anchor its analysis on the core rhythmic structure of the activity. Rigorous empirical evaluations across highly challenging, multi-view datasets validate the superiority of this approach, revealing substantial improvements in classification accuracy and demonstrating exceptional resilience to dramatic geometric transformations and self-occlusions. Related work has addressed geometry-enhanced attentive multi-view stereo for challenging matching scenarios [29]. The implications of this research extend significantly beyond the specific domain of action recognition, offering valuable insights into the broader field of invariant representation learning for video analysis. Looking forward, there are several promising trajectories for future investigation. A critical next step involves adapting the temporal alignment methodology to operate in weakly supervised or fully unsupervised environments, thereby eliminating the reliance on tightly synchronized, multi-camera training data. Exploring techniques that can automatically align temporal attention across disparate, unstructured videos of the same action class would represent a massive leap toward scalable, real-world deployment. Additionally, investigating the integration of this temporal alignment framework with emerging large-scale foundation models could further enhance feature robustness. Ultimately, refining the ability of artificial intelligence systems to understand temporal dynamics invariant of the observer's perspective is a crucial prerequisite for building truly generalized and highly reliable computer vision systems. The literature also considers adaptive weakly supervised Cross-Modal reliability learning for resource-constrained proxy-state recognition [30].

References

Chen, Z., Feng, Y., Zhang, Y., Liu, J., Zhu, C., & Chen, A. (2022). An accurate and convenient method of vehicle spatiotemporal distribution recognition based on computer vision. Sensors, 22(17), 6437.

Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).

Hao, Y., Xu, M., Ye, C., Qin, J., Lu, S., Qin, Y., & Han, X. (2026). LoFA: Learning to Predict Personalized Prior for Fast Adaptation of Visual Generative Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21986-21996).

Bai, X., Li, Z., Xu, W., Lam, E. Y., & Peng, Y. (2026). ERF-GS: Reconstructing Fast Motion from Disjoint Event-RGB Viewpoints. arXiv preprint arXiv:2608.08531.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2014). OverFeat: Integrated recognition, localization and detection using convolutional networks. In Proceedings of ICLR.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems.

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Proceedings of ICLR.

Li, Y., Tan, G., & Gou, C. (2024). Cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose. International Journal of Computer Vision, 132(4), 1242-1257.

Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., ... & Wang, J. (2023). Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8834-8845).

Wang, Y., Li, F. W., & Wang, J. (2026, May). Motion-Guided Semantic Alignment with Negative Prompts for Zero-Shot Video Action Recognition. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10712-10716). IEEE.

Chen, S., Zhu, T., Zhou, R., Zhang, J., Gao, S., Niebles, J. C., ... & Li, M. (2025). Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. arXiv preprint arXiv:2503.01773.

Lu, L., Su, J., Wang, J., & Deng, X. (2026). AMSPINN: anisotropic and multiscale cross-attention physics-inspired neural networks for turbulence prediction. Engineering Applications of Computational Fluid Mechanics, 20(1), 2628951.

Zhu, M., He, B., Liu, J., & Zhang, L. (2019). Dark channel: The devil is in the details. ieee signal processing letters, 26(7), 981-985.

Liu, R., Zhang, R., & Wang, S. (2025, December). Transformer-based modeling of user interaction sequences for dwell time prediction in human-computer interfaces. In 2025 7th International Academic Exchange Conference on Science and Technology Innovation (IAECST) (pp. 352-356). IEEE.

Jing, G., Gao, P., Lee, Y., Hu, Y., & Zhang, H. (2025). 3D-aided pedestrian representation learning for video-based person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 35 (12), 12830-12845.

He, H., Sun, Z., Fan, J., & Mok, P. Y. (2025). TiPGAN: High-quality tileable textures synthesis with intrinsic priors for cloth digitization applications. Computer-Aided Design, 183, 103866.

Zhao, K., Li, G., Li, C., Jiang, T., Wang, W., & Yang, Z. (2026). Identification Markers for Salvia miltiorrhiza and Its Close Relatives Based on Cell Wall Component Characteristics. Engineered Science, 40, 2122.

Lyu, S., Mao, S., & Ma, Y. (2025). Unified Long Video Inpainting and Outpainting via Overlapping High-Order Co-Denoising. arXiv preprint arXiv:2511.03272.

Zhang, Y., Pleus, A., & Althoff, M. (2026, July). Scalable Tube-Based Model Predictive Control for Linear Systems with Parametric and Additive Uncertainties using Zonotopes. In 2026 European Control Conference (ECC) (pp. 509-516). IEEE.

Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).

Jielong Tang, Xujie Yuan, Jiayang Liu, Jianxing Yu, Xiao Dong, Lin Chen, Yunlai Teng, Shimin Di, and Jian Yin. 2026. Sake: Self-aware knowl- edge exploitation-exploration for grounded multi- modal named entity recognition. arXiv preprint arXiv:2604.20146.

Z. Wang, R. Yuan, Z. Geng, H. Li, X. Qu, X. Li, S. Chen, H. Fu, R. B. Dannenberg, and K. Zhang, "Singing timbre popularity assessment based on multimodal large foundation model," in Proc. 33rd ACM Int. Conf. Multimedia, Oct. 2025, pp. 12227–12236, doi: 10.1145/3746027.3758148.

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.

Yang, Y., Liao, Y. H., Bortins, I., Baldwin, D. P., & Zhang, S. (2024). Unidirectional structured light system calibration with auxiliary camera and projector.Optics and Lasers in Engineering,175, 107984.

Fan, J., Liang, Y., Liu, L., Huang, S., & Zhang, L. (2023). RCA-NOC: Relative contrastive alignment for novel object captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 15510–15520). IEEE. https://doi.org/10.1109/ICCV51070.2023.01422

Zhao, Y., Li, Z., Guo, X., & Lu, Y. (2022). Alignment-guided temporal attention for video action recognition. Advances in Neural Information Processing Systems, 35, 13627-13639.

Chen, L., Wang, J., Mortlock, T., Khargonekar, P., & Al Faruque, M. A. (2025, June). Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 22306-22316). IEEE.

Liu, Y., Cai, Q., Wang, C., Yang, J., Fan, H., Dong, J., & Chen, S. (2024). Geometry-enhanced attentive multi-view stereo for challenging matching scenarios. IEEE Transactions on Circuits and Systems for Video Technology, 34(8), 7401-7416.

Meng, X., & Lin, J. (2026). AWCRL: Adaptive Weakly Supervised Cross-Modal Reliability Learning for resource-constrained proxy-state recognition. Pattern Recognition, 114998.