Confidence-Aware Gated Multimodal Fusion for Robust Temporal Action Localization in Occluded Environments

Masato Takami and Tomohiro Fukuda *

Posted Date: 25 February 2026

doi: 10.20944/preprints202602.1564.v1

Keywords: temporal action localization; multimodal learning; pose estimation; robustness; occlusion handling

Article

Confidence-Aware Gated Multimodal Fusion for Robust Temporal Action Localization in Occluded Environments

Masato Takami and Tomohiro Fukuda *

Division of Sustainable Energy and Environmental Engineering, Graduate School of Engineering, The University of Osaka, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan

* Correspondence: fukuda.tomohiro.see.eng@osaka-u.ac.jp

Abstract

In industrial environments, robust Temporal Action Localization (TAL) is essential; however, frequent occlusions often compromise the reliability of skeletal data, leading to negative transfer in multimodal fusion. To address this challenge, we propose a Gated Skeleton Refinement Module (Gated SRM) that explicitly incorporates OpenPose confidence scores into the network architecture. By applying these scores as a logarithmic bias within a self-attention mechanism, our method achieves soft suppression—dynamically attenuating the attention weights assigned to unreliable joints—before adaptively fusing the refined skeletal features with RGB representations through a learnable gating network. Extensive experiments on the heavily occluded IKEA ASM dataset demonstrate that our approach effectively prevents the catastrophic accuracy degradation typical of naive fusion strategies, improving the mean Average Precision (mAP) to 21.77% and outperforming the RGB-only baseline. Furthermore, the system maintains practical real-time inference speeds of approximately 16 frames per second (FPS). By prioritizing confidence-based data selection over data restoration, this sensor-metadata-driven architecture offers a highly robust and principled solution for real-world action recognition under occlusion.

Keywords: temporal action localization; multimodal learning; pose estimation; robustness; occlusion handling

1. Introduction

In recent years, driven by the advancements in smart cities and Industry 5.0, the significance of video-based Human Action Recognition (HAR) technology has increased rapidly [1]. Within construction and manufacturing sites, there is a surging demand for systems that automatically analyze work activities from surveillance footage to ensure worker safety and optimize operational efficiency [2]. As noted by Luo et al. [3], such industrial environments necessitate wide-area monitoring, which in turn requires robust recognition capabilities from a distance. Furthermore, beyond the mere detection of whether an activity is occurring, Temporal Action Localization (TAL)—which precisely identifies the "when" (start and end times) and "what" (action class) of specific tasks—is indispensable for comprehensive and detailed workflow analysis.

The rapid advancement of deep learning techniques has led to a significant leap in the accuracy of HAR [4,5]. Since the introduction of Transformer architecture by Vaswani et al. [6], models based on Attention mechanisms have become predominant in the field of video recognition. For instance, ActionFormer, developed by Zhang et al. [5], successfully achieves high-precision localization of action segments from untrimmed videos by leveraging local self-attention. Moreover, to further improve the robustness of these systems, there is a growing interest in multimodal approaches that combine RGB video data with skeletal features, which represent the geometric structure of the human

 doi:10.20944/preprints202602.1564.v1

body [7]. Rehman et al. [1] demonstrated that the integration of RGB and skeletal information allows for recognition accuracy exceeding 98% under clean environmental conditions.

However, applying existing multimodal methods to real-world unconstrained environments, such as construction sites, still poses significant challenges. The core issue lies in the "reliability of sensor data." As exemplified by the IKEA ASM dataset [8], industrial work sites are characterized by frequent occlusions that compromise skeletal data quality. While skeleton estimation techniques, such as OpenPose [9], offer effective motion descriptions with advantages like privacy protection and viewpoint invariance [10], they are liable to produce low-confidence or erroneous coordinate values for missing joints under occlusive conditions [10,11]. Conventional fusion strategies [12], including the work by Rehman et al. [1], typically perform feature integration under the assumption that all input streams are consistently valid, rendering them vulnerable to such noise contamination. Consequently, there is a substantial risk of "negative transfer," where degraded skeletal features negate the advantages of the RGB modality, thereby diminishing overall recognition accuracy [13,14].

The objective of this study is to develop a confidence-aware temporal action localization (TAL) system that functions robustly even in real-world environments where occlusions frequently occur. Specifically, we propose a mechanism that directly integrates the confidence scores generated by OpenPose into the attention mechanism of the skeleton feature extraction network. This method functions by adding the logarithm of the confidence score as a bias term to the self-attention layer during the input stage of ActionFormer [5]. This enables the model to adaptively decay the weights of joint information whose reliability is compromised by occlusion. This mechanism effectively prevents negative transfer (adverse effects on RGB features) during multimodal fusion. The result is flexible and seamless integration based on data quality. Specifically, it maximizes the usefulness of geometric features when they are accurate and minimizes their impact when they are inaccurate. We evaluated the effectiveness of this method using the IKEA ASM dataset [8], characterized by severe occlusions, and demonstrated its superiority over conventional methods.

The primary contributions of this work are summarized as follows:

- We propose a confidence-biased self-attention mechanism that incorporates log-transformed pose estimation confidence scores as bias terms in the attention weight computation, achieving continuous soft suppression of unreliable skeletal features.
- We develop a Gated Skeleton Refinement Module (Gated SRM) that purifies skeletal information prior to feature fusion via a learnable gating network, designed to prevent the fused representation from degrading below the RGB-only baseline even under heavy occlusion.
- We validate the proposed method on both the standard THUMOS14 benchmark and the heavily occluded IKEA ASM dataset, demonstrating consistent improvements over RGB-only baselines and conventional fusion approaches.

The remainder of this paper is organized as follows. Section 2 provides an overview of related research on temporal action localization (TAL) and sensor fusion. Section 3 details the architecture of the proposed system. Section 4 describes the verification process using a prototype system, and Section 5 presents the verification test. Finally, Section 6 offers insights derived from the findings, and Section 7 concludes the paper.

2. Related Work

2.1. Video-Based Temporal Action Localization

2.1.1. Evolution from Video Classification to Temporal Action Localization

Research in HAR within the field of computer vision has undergone significant advancements over the past several decades. In early studies, the predominant task was Video Classification, which involves assigning a single action label—such as "walking," "running," or "waving"—to short, trimmed video clips [15,16]. While these methods achieved certain successes under constrained computational resources, they often lacked robustness against complex backgrounds and

 doi:10.20944/preprints202602.1564.v1

illumination variations. Furthermore, they faced inherent limitations in capturing high-level semantic features [4,15].

The emergence of deep learning fundamentally transformed this landscape. The Two-Stream Convolutional Network, proposed by Simonyan et al. [17], introduced a seminal architecture that processes spatial appearance information (RGB images) and temporal motion information (optical flow) through separate convolutional neural networks (CNNs) before eventually fusing them. This approach mimics the human cognitive process of integrating visual and motion information in the brain, leading to a dramatic improvement in the accuracy of Action Recognition. Furthermore, I3D, developed by Carreira et al. [18], enabled the transfer of knowledge from large-scale image datasets to the video domain by "inflating" 2D kernels pre-trained on ImageNet along the temporal dimension. Consequently, this model has become the de facto standard for large-scale video datasets, such as Kinetics-400 [18,19].

However, in real-world deployment scenarios, videos are typically not pre-trimmed; instead, they are provided as long, untrimmed streams. Given this context, the primary research focus has shifted from simple classification to TAL [20]. This task involves identifying both the start and end timestamps of action instances within untrimmed videos while simultaneously identifying their respective action classes.

2.1.2. Introduction of the Transformer Architecture and the Development of ActionFormer

The most significant breakthrough in recent research on TAL is the introduction of Transformer architecture, which has achieved overwhelming success in the field of Natural Language Processing (NLP) [6]. The self-attention mechanism, which serves as the core of the Transformer, allows for the direct modeling of relationships between elements at arbitrary positions within a sequence. Consequently, this mechanism enables the effective utilization of contextual information surrounding action segments [4,5].

ActionFormer, presented by Zhang et al. [5], is a prominent model that optimizes the Transformer architecture specifically for TAL tasks. It has achieved state-of-the-art (SOTA) performance on major benchmarks, such as THUMOS14 and ActivityNet-1.3, significantly outperforming conventional methods [21,22].

Following the success of ActionFormer, several variants offering higher accuracy and efficiency have been proposed in succession. TriDet [23] introduced the concept of "trigger detection" to address the ambiguity of action boundaries, leading to a significant improvement in boundary estimation accuracy. In addition, LGAFormer [24] attains more robust feature representations by integrating local and global attention mechanisms.

However, these SOTA methods primarily focus on structural improvements to the detection head, the refinement of boundary regression, and the optimization of computational efficiency. These approaches inherently assume that the RGB and skeletal features are "reliable" (i.e., clean and noise-free). Consequently, there remains a significant research gap regarding effective countermeasures for scenarios where the reliability of the input data itself is compromised by occlusion or fluctuations in lighting—conditions that directly challenge the robustness of the localization process.

2.1.3. Next-Generation Foundational Technology: State Space Models (SSM) and Mamba

To address the computational cost issues of Transformer—specifically the quadratic increase in computational complexity relative to sequence length—State Space Models (SSM), particularly the Mamba architecture, have recently garnered significant attention [25]. In the field of video recognition, pioneering efforts such as ActionMamba [26] have begun to emerge. ActionMamba leverages the "Selective Scan" capability of Mamba blocks to incorporate a mechanism that retains essential action information within video sequences while discarding irrelevant background noise [26]. This approach suggests the potential to resolve both the limitations of GCN-based methods in global context understanding and the computational overhead inherent in Transformer [19,26]. However, SSM-based TAL methods are still in their infancy; their integration with established

 doi:10.20944/preprints202602.1564.v1

detection heads and post-processing pipelines, such as those employed by ActionFormer, has not yet been sufficiently validated. Furthermore, since SSMs are specialized for one-dimensional sequence modeling, the efficient embedding of spatial structural information, such as image or skeleton data, remains an ongoing challenge [10].

2.2. Action Recognition Using Skeleton Data

2.2.1. Characteristics and Advantages of Skeleton Data

The utilization of skeleton information (Skeleton Data) has established an indispensable status in the field of HAR as a modality that complements the inherent limitations of RGB video [1,10,27]. Skeleton data is represented as a temporal sequence of 2D or 3D coordinates of primary human joint points (keypoints). This representation is of paramount importance from the perspectives of privacy preservation, data efficiency, and background invariance [10,27,28].

2.2.2. Evolution of Graph Convolutional Networks (GCN) and the "Reliability Gap" in Pose Estimation

In skeleton-based action recognition, modeling the human body structure as a graph has become a standard approach [29]. Spatial-Temporal Graph Convolutional Networks (ST-GCN), introduced by Yan et al. [29] in 2018, represents a seminal work in this field. ST-GCN constructs a spatial-temporal graph by integrating a spatial graph structure—where joints are defined as nodes (vertices) and bones as edges—with temporal edges that connect identical joints across successive frames. By applying graph convolution operations to this structure, the model effectively extracts discriminative features of human actions.

Since the inception of ST-GCN, numerous refined methods have been proposed to enhance the representational capacity of graph structures [10,26,30,31]. While skeleton-based approaches are theoretically powerful, their performance is heavily reliant on the quality of the input skeleton data. In real-world scenarios, skeleton features are typically extracted using image-based pose estimation algorithms such as OpenPose [9]; however, these estimators are not infallible. Specifically, the accuracy of pose estimation degrades significantly in the presence of occlusion (where objects hide body parts) or self-occlusion (where body parts overlap), leading to unreliable joint coordinates [10,13].

State-of-the-art pose estimators, such as OpenPose [9], output a confidence score representing the certainty of the estimation alongside the spatial coordinates (x, y) for each joint. These confidence scores can serve as a critical indicator for determining whether a joint is visible or occluded [9,10,14]. However, many existing skeleton-based action recognition methods [29] typically discard these confidence scores at the input stage, relying solely on coordinate information (x, y) as geometric features [32]. This convention treats unreliable, noisy data as equivalent to reliable, accurate data. Consequently, this causes the model to learn erroneous motion patterns when occlusion occurs, resulting in a degradation of recognition accuracy.

2.2.3. Existing Approaches and Their Limitations

To address the noise inherent in skeleton data, several prior studies have focused on "data restoration." Song et al. [13] proposed the "Richly Activated GCN," designed to extract robust features even from incomplete skeleton data, attempting to compensate for missing joint information through multi-stream fusion. Similarly, Yoon et al. [11] developed a method to predict and impute currently missing joints based on information from preceding frames.

These approaches are fundamentally aimed at approximating the Ground Truth (GT) of skeleton data. However, in scenarios characterized by severe occlusion, such as construction sites or complex working environments, the loss of information is often too extensive to allow for accurate restoration or repair [3,10]. Feeding inaccurately restored skeleton data—essentially hallucinations—into the model poses a significant risk of inducing overconfidence in the predictions [33]. We define the term

 doi:10.20944/preprints202602.1564.v1

"Reliability Gap" as the discrepancy between the apparent format of the data output by sensors (or estimators) and the actual reliability of its content. Existing restoration-based methods do not effectively bridge this gap; instead, they risk introducing additional noise in their attempt to fill it. Consequently, they remain problematic for real-world applications where safety and reliability are paramount [3,33].

2.3. Multimodal Sensor Fusion and Reliability

2.3.1. Fusion Strategies and the Manifestation of the Reliability Gap

To overcome the inherent limitations of single modalities (such as RGB-only or skeleton-only approaches), multimodal sensor fusion—which integrates information from multiple sensors—is viewed as a promising solution. Rehman et al. [1] reported that integrating RGB video with skeleton tracking data can significantly improve the accuracy of HAR.

While multimodal integration generally improves average performance [1,7], simple feature concatenation relies on the implicit assumption that all modalities maintain constant reliability [33,34]. In real-world environments where sensor quality fluctuates dynamically, this assumption leads to the negative transfer problem defined in Section 1— a direct manifestation of the Reliability Gap (defined in Section 2.2.3) at the system level. This poses an unacceptable risk in safety-critical applications [3,12,31].

2.3.2. Reliability-Aware Learning

To address this challenge, some studies have employed approaches that utilize attention mechanisms for feature weighting [7,35,36]. Mechanisms such as the cross-attention found in Transformers are capable of learning cross-modal correlations and prioritizing salient information. However, since attention mechanisms are fundamentally trained to identify "discriminative features" relevant to the task [6], they run the risk of assigning high attention weights even to noise if it appears distinctive. In other words, relying solely on internal attention mechanisms within neural networks may be insufficient to accurately assess the physical signal integrity of the sensors.

In contrast, the confidence scores inherently output by the pose estimator serves as objective metadata representing the intrinsic quality of the sensor data itself. However, to the best of our knowledge, research that utilizes this confidence score as an explicit gating signal to dynamically modulate feature fusion when integrating skeleton recognition with state-of-the-art TAL models, such as ActionFormer [5], remains extremely limited [10,19]. Most existing studies relegate confidence scores to simple thresholding operations (e.g., discarding low-confidence joints) and fail to incorporate them directly into the neural network's training process as a fully differentiable gating mechanism [10,37].

2.4. Research Positioning and Contributions

Based on the survey of related work presented above, three critical unresolved issues (Research Gaps) have emerged within the current landscape of TAL and multimodal recognition:

- **Limitations of RGB Dependence:** While SOTA models such as ActionFormer [5] differ in their powerful feature extraction capabilities, they remain susceptible to visual noise and occlusion, often incurring high computational costs.
- **Vulnerability of Skeleton Features:** Skeleton-based models like ST-GCN [29] offer a lightweight alternative; however, they are extremely fragile against data missingness, and noise caused by occlusion, lacking reliability when used in isolation.
- **Neglect of Dynamic Reliability:** Existing multimodal fusion approaches do not explicitly account for dynamic quality fluctuations (i.e., confidence gaps) across sensors, failing to fully eliminate the adverse effects of low-quality data.

 doi:10.20944/preprints202602.1564.v1

To address these challenges, we present a novel sensor fusion framework leveraging Confidence-Weighted Skeleton Features. The uniqueness and contributions of this research are clearly defined in the following aspects:

- **Confidence-Driven Data Selection vs. Data Restoration:** Restoration approaches [11,13] attempt to reconstruct missing joint coordinates, risking hallucinated predictions under severe occlusion. In contrast, our method bypasses reconstruction entirely by utilizing confidence scores as continuous gating signals, representing a fundamental paradigm shift from "data repair" to "data selection."
- **Explicit Confidence Integration:** While conventional attention-based fusion learns "what to look at," our approach explicitly guides the model on "what to trust" based on sensor metadata. This prevents the neural network from erroneously learning from noise.
- **Validation under Realistic Occlusion Conditions:** Unlike prior TAL studies that primarily evaluate on cleanly captured benchmarks [5,23,24], we validate on the IKEA ASM dataset [8] whose heavy occlusion characteristics (detailed in Section 4.2) closely approximate industrial deployment conditions.

3. Proposed Methods

3.1. System Overview

In this study, we present the Gated Skeleton Refinement Module (Gated SRM), designed to effectively integrate skeletal features with RGB features for TAL. The overall architecture of the system is illustrated in Figure 1.

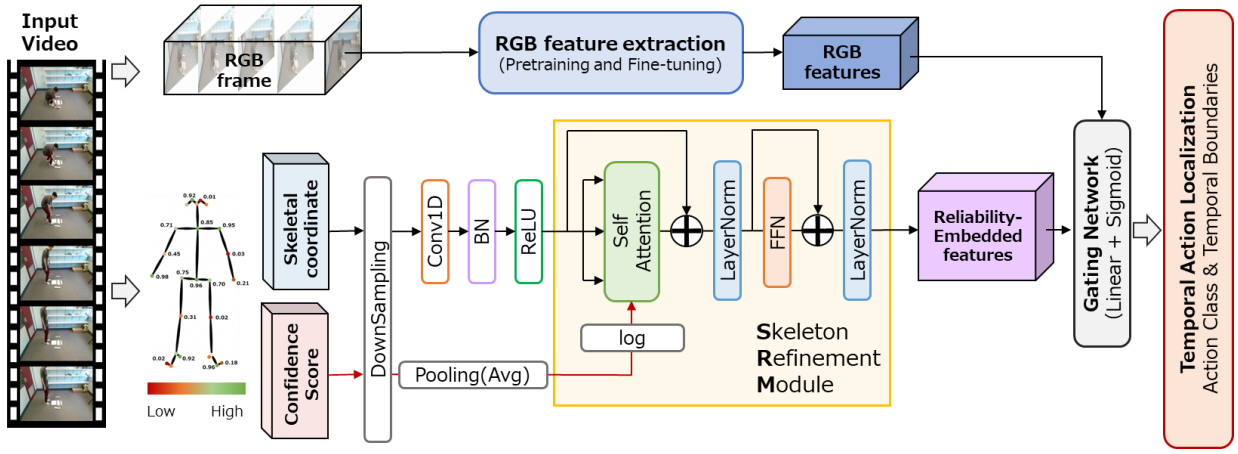

Figure 1. Overall overview of the temporal action localization system incorporating the proposed Gated Skeleton Refinement Module (Gated SRM). The system adaptively fuses RGB features extracted from the input video with skeletal features (Reliability-Embedded features), which are refined through a self-attention mechanism using confidence scores. This fusion is performed via a learnable gating mechanism (Gating Network). The resulting fused features are dimensionally compatible with the base TAL model, requiring no architectural modifications downstream.

The proposed method is a front-end module integrated into an existing Transformer-based TAL model. It takes pre-extracted RGB features and skeletal coordinates as input and generates fused features by adaptively controlling the contribution of skeletal information through a learnable gating mechanism. Because the generated fused features maintain the same dimensionality as the original RGB features, the design allows the system to leverage the benefits of skeletal information without requiring any structural modifications to the subsequent TAL model architecture.

The overall architecture of the proposed method consists of three stages: (1) the Skeleton Refinement Module (SRM), which refines skeletal features using a confidence-biased self-attention

 doi:10.20944/preprints202602.1564.v1

mechanism; (2) dimensional transformation via a projection layer; and (3) adaptive fusion with RGB features through a gating mechanism.

3.2. Skeleton Refinement Module (SRM)

The SRM is a module designed to refine raw skeletal coordinate sequences along with the temporal dimension and extract high-level skeletal features. Since coordinate values obtained from existing pose estimation methods contain variances in reliability caused by occlusion and motion blur, this module introduces Confidence-Biased Self-Attention, which incorporates a confidence bias into the attention scores.

3.2.1. Temporal CNN Projection

First, we apply a projection layer—consisting of a 1D convolution with a kernel size of 3, batch normalization, and ReLU activation—to the input skeletal coordinates $\boldsymbol{S} \in \mathbb{R}^{B\times 50\times T}$ to obtain the embedded features $\boldsymbol{X} \in \mathbb{R}^{B\times d_s\times T}$ $(d_s = 512)$. This preprocessing step captures local temporal context while simultaneously unifying the feature dimensionality with the RGB embedding layer of the base model.

3.2.2. Confidence-Biased Multi-Head Self-Attention

We apply Multi-Head Self-Attention to the obtained embedded features $\boldsymbol{X}$. Following the standard Transformer formulation, the attention score (logit) e_{ij} between a query q_i and a key k_j is typically computed as scaled dot-product attention (Equation (1)):

$$e_{ij} = \frac{q_i {k_j}^T}{\sqrt{d_k}} \tag{1}$$

where d_k is the dimension of the key heads. To explicitly integrate the reliability of skeletal estimates into this mechanism, we inject structural information by adding a learnable bias term to the attention logits. In our framework, we incorporate a bias term derived from the skeletal confidence scores $c \in \mathbb{R}^{B\times 1\times T}$ into the standard attention score on the Key side.

Here, the per-frame confidence score c is obtained by computing the arithmetic mean of the confidence scores across all 25 joints at each time step. The modified attention score ${e'}_{ij}$ is defined as Equation (2):

$${e'}_{ij} = e_{ij} + log(c_j + \varepsilon) \tag{2}$$

where $\varepsilon = 10^{-6}$ is a small constant introduced for numerical stability, and c_j represents the confidence score of the key frame at time step j.

Notably, the logarithmic transformation enables the bias to function as multiplicative weight control within the Softmax function. By substituting e_{ij} into the Softmax operation, the attention weight α_{ij} is derived as Equation (3):

$$\alpha_{ij} = \frac{exp({e'}_{ij})}{\sum_n exp({e'}_{in})} = \frac{exp(e_{ij} + log(c_j))}{\sum_n exp(e_{in} + log(c_n))} = \frac{exp(e_{ij})\cdot c_j}{\sum_n exp(e_{in})\cdot c_n} \tag{3}$$

As shown in this derivation, the confidence scores c_j directly scale the exponential of the raw attention score. This mechanism ensures that features with high confidence $(c_j \approx 1)$ retain their original affinity, while those with low confidence $(c_j \to 0)$ are exponentially suppressed toward zero, effectively filtering out unreliable skeletal information before feature aggregation.

Finally, the output of the attention head is computed as Equation (4):

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}} + log(C + \varepsilon))\mathrm{V} \tag{4}$$

where C is the matrix of confidence scores broadcast across the query dimension.

 doi:10.20944/preprints202602.1564.v1

Following the attention output, we apply residual connections and Layer Normalization, followed by a Feed-Forward Network (FFN) with an expansion ratio of 4. After a subsequent round of residual connections and Layer Normalization, we obtain the refined skeletal features $\boldsymbol{X_{refined}} \in \mathbb{R}^{B\times d_s\times T}$.

3.3. Gated Fusion Mechanism

Rather than employing simple concatenation or addition, we introduce a learnable gating mechanism to fuse the refined skeletal features $\boldsymbol{X_{refined}} \in \mathbb{R}^{B\times 512\times T}$ with the RGB features $\boldsymbol{F_{RGB}} \in \mathbb{R}^{B\times 1024\times T}$. This mechanism enables the model to adaptively adjust the contribution of skeletal information at each time step and channel, depending on the specific video content and action types.

First, we project the skeletal features into the same dimensionality as the RGB features (Equation (5)).

$$\boldsymbol{F_{skel}} = ReLU(BN(\boldsymbol{W_p} * \boldsymbol{X_{refined}})) \in \mathbb{R}^{B\times 1024\times T} \tag{5}$$

where $\boldsymbol{W_p}$ denotes the weights of a 1D convolution with a kernel size of 1.

Next, we concatenate the RGB features and the projected skeletal features to compute per-channel gate values (Equation (6)).

$$\boldsymbol{g} = \sigma(\boldsymbol{W_g} * [\boldsymbol{F}_{RGB}; \boldsymbol{F}_{skel}]) \in \mathbb{R}^{B\times 1024\times T} \tag{6}$$

In this equation, $[\cdot;\cdot]$ represents channel-wise concatenation (yielding 2048 dimensions), $\boldsymbol{W_g}$ is a 1D convolution layer that maps the 2048-dimensional vector to 1024 dimensions, and σ denotes the Sigmoid activation function.

The final fused features are calculated as a weighted addition of the skeletal features, modulated by the gate values (Equation (7)).

$$\boldsymbol{F}_{fused} = \boldsymbol{F}_{RGB} + \boldsymbol{g} \odot \boldsymbol{F}_{skel} \tag{7}$$

where $\odot$ denotes the element-wise product.

To ensure stability during the initial stages of gate network training, we initialize the weights of the gating layer to zero and the bias to -2.0. Consequently, the initial gate output is $\sigma(-2.0) \approx 0.12$, which ensures that the fused features are approximately equal to the RGB feature ($\boldsymbol{F}_{fused} \approx \boldsymbol{F}_{RGB}$) at the onset of training. This initialization strategy offers the following advantages.

- **Preservation of Pre-trained Representations**: It prevents the destruction of the pre-trained RGB backbone's representations during the early phases of training.
- **Progressive Learning:** It enables a progressive learning process where the contribution of skeletal information increases gradually over time.
- **Robustness as a Safety Valve**: It serves as a safety valve that automatically "closes" the gate when the quality of skeletal features is compromised, such as in environments with heavy occlusion.

3.4. Integration with Base Model

The proposed Gated SRM functions as a front-end preprocessing module integrated into the base TAL model. Specifically, for each video, the module receives RGB features $\boldsymbol{F_{RGB}} \in \mathbb{R}^{1024\times T}$, skeletal coordinates $\boldsymbol{S} \in \mathbb{R}^{50\times T}$, and confidence scores $\boldsymbol{c} \in \mathbb{R}^{1\times T}$ as inputs. After generating the fused features $\boldsymbol{F_{fused}} \in \mathbb{R}^{1024\times T}$ through the Gated SRM, these features are fed into the base model in place of the original RGB features.

Because the dimensionality of the fused features remains identical to that of the RGB features (1024 dimensions), no structural or parametric modifications are required for the base model components, such as the backbone network, Feature Pyramid Network (FPN), or the classification and regression heads. This architectural design ensures that the pre-trained weights of the base model can be utilized directly, facilitating either the isolated fine-tuning of the proposed module or comprehensive end-to-end training of the entire system.

 doi:10.20944/preprints202602.1564.v1

4. Implementation of Prototype System

4.1. Development Environment

The development and evaluation of the prototype system were conducted on a workstation running Ubuntu 24.04 LTS. The hardware configuration utilized an Intel Core i7-8700 CPU and an NVIDIA GeForce GTX 1080 GPU. Regarding the software environment, we employed Python 3.9 and the PyTorch 2.5.1 framework, using OpenCV for video input/output and preprocessing.

In this implementation, we prioritized a lightweight design that avoids excessive computational resource demands, anticipating future deployment on edge devices. Specifically, the skeletal processing stream is engineered to maintain a low computational load, ensuring high adaptability for real-time processing applications.

4.2. Datasets

For the evaluation, we utilize the THUMOS14 dataset [38], a standard benchmark in the field of action recognition, and the IKEA ASM dataset [8], which simulates real-world environments characterized by frequent occlusions.

- **THUMOS14 dataset [38]:** This is a large-scale dataset comprising untrimmed videos of sports activities collected from YouTube. It contains 20 action classes, including "Long Jump" and "Cricket Bowling." While the dataset exhibits significant camera motion and diverse backgrounds, the subjects are typically large and clearly visible within the frames, with a relatively low frequency of occlusions. To ensure a fair comparison with existing studies [5,38–42], we utilize the standard subset consisting of the validation set (200 videos) and the test set (213 videos).

Although THUMOS14 exhibits relatively low occlusion frequency, we include it as an evaluation benchmark for two reasons: (1) to verify that the proposed method does not degrade performance on well-established benchmarks, and (2) to demonstrate that the gating mechanism appropriately increases skeletal contribution when skeleton quality is high, thereby improving upon the RGB-only baseline.

- **IKEA ASM dataset [8]:** This dataset consists of 371 video recordings capturing the assembly of various furniture items, such as tables and shelves. It contains a total of approximately 35 hours of footage. The average duration per video is about six minutes. The dataset is densely annotated, yielding approximately 31,000 action instances (clips) across the entire dataset.

It comprises 33 officially defined atomic action classes (typically utilized as 32 foreground action classes plus one background class in Temporal Action Localization tasks). These classes are structured as verb-object pairs (e.g., "attaching legs" and "spinning in"). Sample frames from the IKEA ASM dataset are illustrated in Figure 2. The defining characteristic of this dataset is the frequent occurrence of severe self-occlusion and occlusion by objects, resulting from workers bending forward deeply or being obscured by furniture components. These attributes closely mirror the conditions of real-world environments, such as construction sites and factories, making it an ideal testbed for evaluating the robustness of our proposed method.

 doi:10.20944/preprints202602.1564.v1

Figure 2. Sample frames from the IKEA ASM dataset. This dataset captures furniture assembly tasks where self-occlusion and occlusion by objects occur frequently. It comprises 32 categories of atomic actions, such as "attaching legs" and "flipping the tabletop," exhibiting characteristics highly representative of real-world environments.

For the validation of our research, we have adopted the "Environment-based Train/Test Split" protocol established in the official dataset publication. This strategy is designed to prevent overfitting that could occur if identical subjects or rooms (environments) appear in both the training and testing data. Specifically, among the five assembly environments, we strictly isolate Environments 1 and 2 (the family room and the office) to serve as the test set (117 videos). Consequently, we utilize the remaining environments as the training set (254 videos).

4.3. Base Model: ActionFormer

We adopt ActionFormer [5] as the base model for TAL. ActionFormer is a single-stage detection model that applies a Transformer-based backbone equipped with a Feature Pyramid Network (FPN) to pre-extracted video features, simultaneously performing action classification and temporal boundary regression at each time step.

Table 1 presents the key hyperparameters of the base model. The backbone architecture consists of a convolutional Transformer (convTransformer) with a layer configuration of (2, 2, 5) and a scale factor of 2. Both the embedding dimension and the FPN dimension are set to 512, with the number of attention heads set to 4 and the local attention window size to 19. The detection head comprises three layers with a hidden dimension of 512, and the regression ranges are defined as [0, 4], [4, 8], [8, 16], [16, 32], [32, 64], and [64, 10000]. During inference, we apply Soft-NMS with an IoU threshold of 0.4 and retain the top 200 predictions per video, following the standard ActionFormer configuration [5].

Table 1. Configurations of the ActionFormer Base Model.

Parameter	Value
Backbone Type	convTransformer
Backbone Architecture	(2, 2, 5)
Scale Factor	2
Input Dimension	1024 (IKEA ASM) / 2048 (THUMOS14)
Embedding Dimension	512
FPN Dimension	512
FPN Type	identity
Number of Attention Heads	4
Attention Window Size	19
Detection Head Dimension	512

 doi:10.20944/preprints202602.1564.v1

Number of Detection Head Layers	3
Regression Ranges	[0, 4], [4, 8], [8, 16], [16, 32], [32, 64], [64, 10000]
Maximum Sequence Length	2304

4.4. Implementation of the SRM

To improve recognition accuracy in scenarios with frequent occlusion, we design a SRM that leverages human skeletal features alongside RGB features. The SRM comprises the following three components:

- Temporal CNN Projection Layer: We map the 50-dimensional input skeleton coordinates (25 joints × 2 coordinates) into a 512-dimensional feature space using a 1D Convolutional Neural Network (1D-CNN) with a kernel size of 3 and a padding of 1, followed by Batch Normalization and a ReLU activation function.
- Confidence-Biased Multi-Head Self-Attention: We utilize 8 heads (with a dimension of 64 each). A confidence bias term $log(\boldsymbol{c} + \varepsilon)(\varepsilon = 10^{-6})$ is applied to the Key side during the attention score calculation.
- Feed-Forward Network (FFN): This component consists of two 1D convolutional layers (512 →2048→512) accompanied by ReLU and dropout (with a rate of 0.1).

For both datasets, skeletal coordinates and confidence scores were extracted using OpenPose [9] with the BODY_25 model (25 joints). For THUMOS14, OpenPose was applied to the same frames used for I3D feature extraction, and in scenes with multiple people, the person with the highest average confidence score was selected, and frames with no detected persons were assigned zero-filled coordinates with confidence scores of zero. For the publicly available I3D features, the published 2048-dimensional features were used, which were pre-trained on Kinetics-400, following the standard protocol [5]. For IKEA ASM, the I3D model (RGB stream only, without optical flow) was fine-tuned on the training split for 20 epochs with a learning rate of 0.01, and features with 1024 dimensions were extracted by setting a temporal interval of 4 frames.

4.5. Fusion Strategies

To integrate RGB features and skeletal features, we implement two strategies: a baseline "Naive Concatenation" and our developed "Gated Fusion".

- **Naive Concatenation:** We simply concatenate the 512-dimensional output of the SRM and the RGB features (1024 or 2048 dimensions) along the channel dimension, feeding the result directly into the backbone. Consequently, this approach requires modifying the input dimension of the backbone (e.g., from 1024 to 1536 for the IKEA ASM dataset).
- **Gated Fusion (Proposed Method):** This approach maintains the original input dimension of the backbone, enabling the direct utilization of pre-training weights. It operates in the following three steps:

1. We mapped the SRM output to the same dimension as the RGB features via a projection layer (e.g., a 1x1 convolution).
2. We concatenate the RGB features with the projected skeletal features to generate gate values in the range of [0, 1] using a 1x1 convolution and a sigmoid function.
3. We compute the element-wise sum of the RGB features and the gate-modulated skeletal features as the final fused representation.

The gate initialization follows the strategy described in Section 3.3 (weights = 0, bias = −2.0).

Table 2. Configurations of the SRM and Gated Fusion Module.

Parameter	Value
Skeleton Input Dimension	50 (25 joints x 2 coordinates)
SRM Embedding Dimension	512
Number of Self-Attention Heads	8

 doi:10.20944/preprints202602.1564.v1

Dimension per Head	64
FFN Expansion Ratio	4
Dropout Rate	0.1
Projection Output Dimension	1024 (IKEA ASM) / 2048 (THUMOS14)
Gate Bias Initialization	-2.0

4.6. Training Protocol

All models are trained using the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 0.05. The batch size is set to 2, and the training is conducted for a total of 55 epochs, including 5 warmup epochs. For data augmentation, we employ random cropping with a truncation threshold of 0.5 and a crop ratio of [0.9, 1.0], while applying gradient clipping with an L2 norm of 1.0. We utilize an Exponential Moving Average (EMA) for evaluation, and the model achieving the highest mean Average Precision (mAP) during validation, which is performed every 5 epochs, is selected as the final model.

5. Verification Test

In this section, we evaluate the effectiveness of the proposed Confidence Gating Mechanism through both quantitative and qualitative analyses. For the evaluation, we evaluate the two datasets described in Section 4.2: THUMOS14 and IKEA ASM.

5.1. Experimental Setup

5.1.1. Baselines

In this study, we compare the following configurations: (1) RGB-only baseline (#1): ActionFormer using only RGB features; (2) Naive skeleton (#2): skeleton features projected by CNN and combined with RGB features (without refinement); (3) SRM + fusion (#3): skeleton features refined by SRM and combined with RGB features; (4) Gate fusion without SRM (#4): skeleton features projected by CNN and fused through a gating mechanism without attention or confidence bias; and (5) Proposed method (SRM + gate fusion): the full pipeline proposed in this work.

5.1.2. Evaluation Metrics

This section outlines the evaluation metrics utilized to verify the proposed method. To conduct a multifaceted evaluation of performance in the TAL task, we employ two distinct metrics: mean Average Precision (mAP) and the Boundary-F1 score.

- mean Average Precision (mAP)

mAP is a standard evaluation metric widely adopted in TAL [38,43,44,46]. Specifically, we calculated mAP values at various temporal Intersection over Union (t-IoU) thresholds and utilized their average as the final evaluation metric. The t-IoU was computed according to Equation (8).

$$tIoU = \frac{prediction \cap GT}{prediction \cup GT} \quad (8)$$

where prediction indicates the proposed model's detected actions, and GT refers to the pre-annotated ground-truth activity segments.

In this study, the mAP is calculated according to the following procedure:

1. **Set t-IoU Thresholds**: We establish five levels of temporal Intersection over Union (t-IoU) thresholds: 0.3, 0.4, 0.5, 0.6, and 0.7.
2. **Determine True Positives (TP):** For each detected action instance, the prediction is classified as a True Positive (TP) if its t-IoU with the ground truth segment is equal to or greater than the specified threshold and the action class match correctly.
3. **Rank Predictions:** For each action class, all predictions are ranked in descending order of their confidence scores to compute Precision and Recall.

 doi:10.20944/preprints202602.1564.v1

4. **Calculate Average Precision (AP):** The Average Precision (AP) is determined by calculating the area under the Precision-Recall curve for each class.
5. **Compute mAP:** Finally, the mAP is obtained by averaging the AP values across all action classes.

The mAP value was calculated by Equation (9).

$$mAP = \frac{\sum_{j=1}^{C} AP(j)}{C} \tag{9}$$

where C is the number of action categories.

In this study, we conduct a multifaceted performance evaluation by employing five distinct IoU thresholds. As a final comprehensive assessment, we calculate the average of the mAP values across these five thresholds to serve as the overall performance metric for the system. Furthermore, the Average Precision (AP) for each individual action class obtained during the mAP calculation process is utilized for an in-depth performance.

- Boundary-F1 score

The Boundary-F1 score is an evaluation metric specifically designed to assess the precision of identifying action boundaries, namely, the start and end points of an action. While mAP evaluates the overall overlap of action segments, the Boundary-F1 score directly measures the temporal accuracy of the boundary points themselves [21,47]. In the calculation of the Boundary-F1 score, the success of a prediction is determined by whether a predicted boundary point falls within a specific time window (tolerance τ) of the ground truth boundary. In this study, we established three levels of tolerance: τ = ±0.5 seconds, τ = ±1.0 seconds, and τ = ±2.0 seconds. We computed the Boundary-F1 score in this research according to the following procedure.

1. A predicted boundary point is classified as a True Positive (TP) if it falls within ± τ seconds of a ground truth (GT) boundary point and the action class matches.
2. A prediction is classified as a False Positive (FP) if no ground truth exists within the tolerance window or if the prediction is redundant.
3. A ground truth boundary is classified as a False Negative (FN) if no predicted boundary point exists within the specified tolerance.
4. Precision and Recall are calculated using Equation (10) and Equation (11), respectively.

$$Precision = \frac{TP}{(TP + FP)} \tag{10}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{11}$$

5. The F1 score is then calculated using Equation (12).

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{12}$$

5.2. Quantitative Results

5.2.1. Comparison of THUMOS14

Table 3 presents the mAP comparison results on the THUMOS14 dataset. ActionFormer, serving as the baseline with only RGB input, achieved a mAP of 65.87%. In contrast, a degradation in performance was observed when skeletal data were naively integrated. Specifically, simple concatenation following CNN projection resulted in a mAP of 63.22%, and even with the addition of a SRM, the performance remained at 63.44%; both configurations underperformed relative to the RGB-only baseline. This phenomenon represents a typical case of "negative transfer," where noisy skeletal information is integrated without selective processing, thereby hindering the representation of effective RGB features. Conversely, fusion methods incorporating a gating mechanism demonstrate performance improvements. The combination of CNN projection and gated fusion

 doi:10.20944/preprints202602.1564.v1

recovered the mAP to 64.89%. Furthermore, our proposed approach, "SRM + gated fusion," achieved a mAP of 66.31%, effectively outperforming the baseline.

Table 3. mAP (%) at various t-IoU thresholds on the THUMOS14 dataset.

ID	Skeleton	SRM	Gated Fusion	mAP @ tIoU(%) ↑						Best Epoch
				0.3	0.4	0.5	0.6	0.7	Avg.	
#1				81.17	77.36	70.16	57.62	43.04	65.87	35
#2	✓			79.09	74.96	66.91	54.87	40.24	63.22	35
#3	✓	✓		78.81	74.99	67.47	54.82	41.12	63.44	40
#4	✓		✓	80.92	76.47	68.46	56.28	42.32	64.89	35
Ours	✓	✓	✓	**81.52**	**77.74**	**70.32**	**58.82**	**43.16**	**66.31**	**35**

Furthermore, regarding the Boundary F1 score—an indicator of action boundary detection performance—the proposed method-maintained accuracy levels comparable to the RGB-only baseline. Under the condition of a tolerance threshold $\tau = \pm 0.5$s, the proposed method recorded a score of 0.5665 compared to 0.5659 for the RGB-only configuration. These results confirm that our approach effectively leverages useful skeletal features while simultaneously suppressing the influence of noise (Table 4).

Table 4. Boundary-F1 score on the THUMOS14 dataset.

ID	Skeleton	SRM	Gated Fusion	Tolerance τ (s)		
				±0.5	±1.0	±2.0
#1				0.5659	0.7282	0.8158
#2	✓			0.5256	0.6891	0.7906
#3	✓	✓		0.5607	0.7147	0.8001
#4	✓		✓	0.5664	0.7220	0.8105
Ours	✓	✓	✓	**0.5665**	**0.7281**	**0.8153**

5.2.2. Comparison of IKEA ASM

In contrast to THUMOS14, for the IKEA ASM dataset evaluation (Table 5), we used the fine-tuned I3D model (RGB-only, without optical flow) optimized for furniture assembly tasks.

ActionFormer, serving as the baseline using only RGB features, recorded a mAP of 21.49%. In contrast, the "SRM + simple concatenation" model, which naively integrates skeletal information, suffered a significant performance drop to 19.29%. This 2.20-point degradation exemplifies the negative transfer effect (Section 1), confirming that indiscriminate fusion of occlusion-corrupted skeletal data severely compromises RGB representations.

Conversely, our proposed "SRM + gated fusion" method achieved a mAP of 21.77%, outperforming the baseline (21.49%). These results demonstrate that even in highly noisy environments, the confidence gating mechanism adaptively regulates the contribution of skeletal information. By suppressing negative transfer, the proposed method effectively leverages these features only when they provide beneficial cues.

Notably, while 3DInAction [47] achieves approximately 28.75% mAP on IKEA ASM, it requires 3D point cloud data from depth sensors. Our method achieves 21.77% using only standard 2D RGB video, offering a more cost-effective solution for industrial deployment where depth sensors may be impractical or expensive.

 doi:10.20944/preprints202602.1564.v1

Table 5. mAP (%) at various t-IoU thresholds on the IKEA ASM dataset.

ID	Skeleton	SRM	Gated Fusion	mAP @ tIoU(%) ↑						Best Epoch
				0.3	0.4	0.5	0.6	0.7	Avg.	
#1				30.61	27.23	21.77	17.06	10.79	21.49	25
#2	✓			28.16	24.58	19.74	14.59	9.37	19.29	30
#3	✓	✓		22.32	19.27	16.02	11.65	6.89	15.23	35
#4	✓		✓	**31.21**	**27.94**	22.11	16.27	10.30	21.57	25
Ours	✓	✓	✓	31.01	27.85	**22.24**	**17.08**	**10.49**	**21.77**	**25**

Table 6 presents the comparison of action boundary detection accuracy using the Boundary-F1 score on IKEA ASM. While the simple concatenation method fell below the baseline across all tolerance thresholds τ, the proposed method consistently outperformed the baseline. Specifically, our approach recorded 0.3624 (baseline: 0.3593) at τ = ±0.5s confirming consistent performance improvements ranging from strict boundary evaluation to relaxed conditions.

Table 6. Boundary-F1 score on the IKEA ASM dataset.

ID	Skeleton	SRM	Gated Fusion	Tolerance τ (s)		
				±0.5	±1.0	±2.0
#1				0.3593	0.5354	0.6596
#2	✓			0.3193	0.4842	0.6324
#3	✓	✓		0.2907	0.4399	0.5735
#4	✓		✓	0.3643	0.5430	0.6754
Ours	✓	✓	✓	**0.3624**	**0.5464**	**0.6851**

5.2.3. Statistical Significance Analysis

To rigorously assess the significance of the observed performance differences, we conducted the Wilcoxon signed-rank test—a non-parametric paired test appropriate for per-class AP values that do not necessarily follow a normal distribution—at a significance level of α = 0.05. Table 7 summarizes the p-values comparing the proposed method (SRM + gated fusion) against each baseline configuration.

Table 7. Statistical significance analysis (Wilcoxon signed-rank test, α = 0.05, Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001, n.s. = not significant.).

Comparison (Ours vs.)	IKEA ASM p-value	Sig.	THUMOS14 p-value	Sig.
#1	0.325	n.s.	0.231	n.s.
#2	0.003	**	0.001	***
#3	< 0.001	***	0.005	**
#4	0.665	n.s.	0.019	*

On both datasets, the proposed method achieves statistically significant improvements over all concatenation-based fusion approaches (IDs #2 and #3), confirming that the gating mechanism effectively prevents the negative transfer observed in naive fusion. On THUMOS14, the proposed method also significantly outperforms CNN projection with gated fusion (ID #4, p = 0.019), demonstrating the added value of confidence-biased attention refinement when skeleton quality is relatively high.

Compared to the RGB-only baseline (ID #1), the proposed method did not achieve statistically significant differences on either dataset (p = 0.325 and p = 0.231). This result aligns with the primary

 doi:10.20944/preprints202602.1564.v1

design goal of Gated SRM: the system is designed to safely integrate skeletal information without risking performance degradation, rather than guaranteeing substantial absolute improvement over RGB-only processing. The fact that the proposed method maintains statistical equivalence with the RGB-only baseline, while simple fusion methods suffer significant degradation (6.51 points on IKEA ASM), demonstrates the robustness of the reliability gating mechanism as a safeguard against negative transfer.

5.3. Qualitative Analysis

5.3.1. Attention Map Visualization of IKEA ASM

To provide a detailed analysis of the attention mechanism's behavior within the proposed method, we conducted a visualization study using the IKEA ASM dataset. Figure 3 illustrates the input of RGB images and skeletal features, their corresponding confidence scores, and the attention maps generated by the SRM. As observed in the attention maps, the model assigns higher weights to temporal regions that are crucial for action identification. Specifically, when occlusions occur—such as around timesteps 75–95 where skeletal confidence scores drop below the threshold (0.256)—the attention weights are immediately suppressed (indicated in black). This qualitatively confirms that the proposed method dynamically regulates attention levels in response to the quality of the input data.

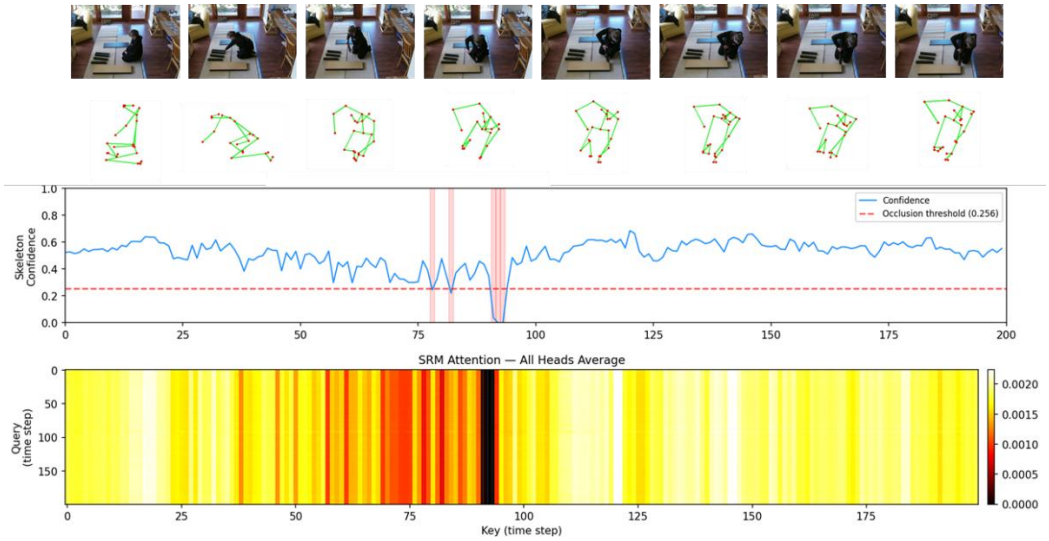

Figure 3. Visualization of the attention mechanism in the IKEA ASM dataset. The top rows show the RGB frames and corresponding skeletal poses. The line plot indicates the skeleton confidence score with the occlusion threshold (0.256). The bottom heatmap displays the SRM attention map, where dark regions represent suppressed attention during low-confidence intervals (occlusions).

Note that this occlusion threshold (e.g., 0.256) is not a hardcoded hyperparameter but is dynamically computed for each video as half of its mean confidence score $Threshold = \mu_{conf} \times 0.5$. For example, with an average score of 0.512, the threshold becomes 0.256. This is an empirical approach based on the assumption that values significantly below the average indicate severe occlusion. However, this dynamic approach allows the model to robustly adapt to baseline confidence distributions that vary across different videos.

5.3.2. Qualitative Evaluation of Robustness to Occlusions

Furthermore, we statistically evaluated the robustness of the proposed method against occlusion. Figure 4 illustrates the cumulative attention weights assigned to each frame as a time series, where "occlusion frames"—defined as frames where the confidence score falls below the threshold—are

 doi:10.20944/preprints202602.1564.v1

highlighted in red. Statistical comparison revealed that while the average attention weight for normal frames (without occlusion) was 0.3536, it decreased to 0.0687 for occlusion frames. This indicates that attention is reduced by an average of 80.6% during the occurrence of occlusions, effectively suppressing unreliable features.

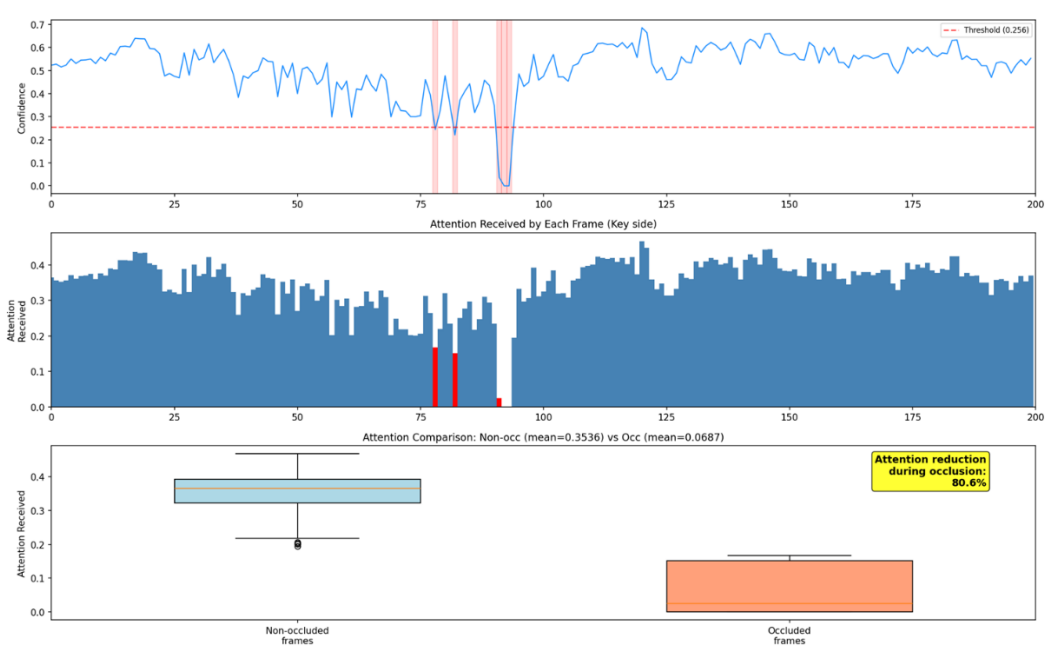

Figure 4. Statistical analysis of attention received by frames during occlusions. The top plot shows the confidence score, and the middle bar chart highlights the attention received by each frame (red bars indicate occluded frames). The bottom box plot compares the attention distribution between non-occluded and occluded frames, demonstrating an 80.6% reduction in attention for unreliable skeletal data.

5.4. Computational Efficiency

To assess the practical deployability of the proposed method, we measured the inference latency and computational resource consumption of each model configuration on an NVIDIA GeForce GTX 1080 GPU. Table 8 presents a comparison of parameter counts, GFLOPs, and inference latency. Note that the reported latency excludes post-processing time for Non-Maximum Suppression (NMS) and refers exclusively to the neural network forward pass.

Table 8. Computational Complexity and Inference Latency on NVIDIA GTX 1080.

Method	Params (M)	GFLOPs	Params Increase	GFLOPs Increase	IKEA ASM Latency (ms)	THUMOS14 Latency (ms)	Inference Speed (FPS)
ActionFormer (RGB-only Baseline)	27.70	83.28	-	-	47.6	46.5	~21
Concat (No SRM Naive Fusion)	28.56	87.25	+3.1%	+4.8%	47.8 (+0.3%)	46.9 (+1.0%)	~21
Gated SRM (Proposed Method)	**33.55**	**121.09**	**+21.1%**	**+45.4%**	**63.6 (+33.6%)**	**63.5 (+36.5%)**	**~16**

The simple concatenation approach introduces minimal overhead (+3.1% parameters, +4.8% GFLOPs), maintaining approximately 21 FPS. The proposed Gated SRM increases parameters to 33.55M (+21.1%) and GFLOPs to 121.09 (+45.4%), resulting in an inference latency of 63.6 ms on IKEA ASM (approximately 16 FPS). The primary source of this overhead is the $\mathcal{O}(T^2)$ self-attention

 doi:10.20944/preprints202602.1564.v1

computation within the SRM over the maximum sequence length of T= 2304. The marginal latency difference between IKEA ASM and THUMOS14 arises from the classification head size (32 vs. 20 classes) and does not significantly affect overall trends.

6. Discussion

This section analyzes the experimental results presented in Section 5 from three perspectives: (1) the mechanism by which the proposed method resolves the reliability gap in multimodal learning (Sections 6.1 and 6.2), (2) the trade-off between computational overhead and inference latency (Section 6.3), and (3) the limitations and prospects for deployment in practical industrial environments (Section 6.4).

6.1. Quantitative Avoidance of Negative Transfer

The IKEA ASM results quantitatively confirm the severity of the reliability gap. Naive fusion degraded mAP by 2.20 points (from 21.49% to 19.29%), demonstrating that increasing input dimensionality without accounting for data quality amplifies negative transfer. Under identical conditions, the proposed Gated SRM fusion achieved a mAP of 21.77%, not only overcoming the limitations of naive fusion (statistically significant; $p < 0.001$ on IKEA ASM) but also marginally surpassing the RGB-only baseline. Although this improvement does not reach statistical significance ($p = 0.325$), the result is consistent with the design philosophy of Gated SRM: the module primarily functions as a safety mechanism that prevents degradation of the RGB due to skeletal noise, rather than as one guaranteeing substantial absolute improvement. The critical distinction lies in the asymmetry of failure modes—whereas naive fusion risks catastrophic degradation, gated fusion ensures a robust performance floor. Additionally, the boundary F1 score improved from 0.3593 to 0.3624 at a tolerance of $\tau = \pm 0.5$ s, indicating the potential for more temporally precise localization. These results collectively demonstrate that the reliability gating mechanism selectively integrates geometric features only from frames with accurately estimated skeletal data, while effectively filtering noisy inputs.

A further notable observation concerns the interaction between the two proposed components. On IKEA ASM, applying the SRM without gated fusion (ID #3: 15.23%) yielded lower performance than naive concatenation without SRM (ID #2: 19.29%). This counterintuitive result suggests that, while the SRM's confidence-biased attention effectively modulates skeletal features, the resulting refined representation exhibits distributional characteristics incompatible with simple concatenation. Without the gating mechanism to regulate the contribution magnitude, the refined features introduce a domain mismatch that exacerbates negative transfer. This finding underscores that the SRM and the gating mechanism are complementary components that must operate jointly to achieve robust fusion.

The evaluations on THUMOS14 further corroborate the generalizability of the proposed method in environments where occlusion is less frequent and skeleton quality is relatively high. While naive integration of skeletal data degraded mAP to 63.22% (ID #2), the proposed SRM with gated fusion achieved 66.31%, outperforming both the RGB-only baseline (65.87%) and CNN projection with gated fusion (ID #4) with statistical significance (p = 0.019). These findings reveal an important complementary property of the Gated SRM: when skeletal features are sufficiently reliable, the module appropriately amplifies their contribution, yielding improvements beyond the RGB-only baseline. Taken together, the results on both datasets confirm that the proposed method operates as an adaptive reliability regulator, suppressing noisy skeletal inputs under heavy occlusion (IKEA ASM) while leveraging high-quality skeletal information to enhance recognition accuracy in cleaner conditions (THUMOS14).

 doi:10.20944/preprints202602.1564.v1

6.2. Confidence Bias in Logarithmic Space and Soft Suppression Behavior

The effectiveness of the proposed method is attributable to the confidence-biased attention mechanism, where log-transformed confidence scores function as multiplicative gates within the Softmax function. Unlike hard thresholding approaches that discard low-confidence joints entirely, this continuous formulation preserves differentiability and enables end-to-end learning of reliability-aware representations. The qualitative evaluations empirically validate this mechanism: during occlusion events on the IKEA ASM dataset, attention weights decreased from 0.3536 to 0.0687, a reduction of 80.6%. This confirms that the soft suppression operates autonomously as a safeguard, preventing unreliable skeletal inputs from adversely affecting the RGB stream.

6.3. Practical Implications of the Computational Overhead

The Gated SRM introduces a moderate increase in latency relative to the RGB-only baseline. We contend that this trade-off is justified from a practical deployment perspective for two reasons.

First, the achieved throughput of approximately 16 FPS on a consumer-grade GTX 1080 GPU (without Tensor Cores) already satisfies the typical requirements for real-time monitoring, where 10–15 FPS provides sufficient temporal resolution [48]. Deployment of newer hardware architectures is expected to reduce this overhead further.

Second, and more critically, this modest computational cost yields a substantial qualitative advantage: complete avoidance of the accuracy degradation observed under heavy occlusion. The naive concatenation approach suffers a 6.51-point mAP drop on IKEA ASM despite incurring only a marginal latency increase (+0.3%), representing an unacceptable failure mode for safety-critical industrial surveillance. In contrast, the proposed method's additional 16 ms overhead ensures a net accuracy gain over the RGB-only baseline, achieving a favorable accuracy–efficiency trade-off for real-world deployment.

6.4. Limitations: Confidence Overreliance Risk and Industrial Deployment Challenges

While the proposed method demonstrates practical inference speed and robustness to occlusion, it possesses a fundamental limitation: its reliance on the confidence scores produced by the upstream pose estimation model. These scores do not always reflect physical reality accurately, owing to three distinct sources of error.

The first source is false detection due to overconfidence; wherein excessively high scores are assigned to incorrect joint locations. This is commonly triggered by adverse lighting, visual variations caused by protective equipment, or misidentification among multiple individuals. The second source is scoring miscalibration: predicted confidence levels may fail to align with empirical correctness probabilities, leading to systematic underestimation of uncertainty. The third source is temporal instability, whereby occlusions and self-occlusions cause confidence scores to fluctuate sharply between consecutive frames, injecting noise into the skeletal feature sequence.

To mitigate these risks, we identify several promising directions for future research. At the algorithmic level, two approaches merit particular attention: post-hoc calibration via temperature scaling to align predicted scores with empirical accuracy, and cross-modal verification to detect inconsistencies between the RGB and skeleton modalities. Additionally, ensemble strategies that aggregate predictions from multiple pose estimation models to quantify reliability through inter-model consistency, combined with temporal smoothing to attenuate sudden score fluctuations, are expected to improve localization robustness.

Beyond algorithmic refinements, practical deployment in industrial settings presents physical and computational challenges. The current framework relies exclusively on visual information and is therefore ineffective in low-visibility conditions such as darkness or heavy dust. Extending the architecture to incorporate non-visual sensing modalities—such as LiDAR or thermal imaging—constitutes a crucial next step toward domain-general applicability. Finally, the quadratic computational complexity of the self-attention mechanism ($\mathcal{O}(T^2)$) limits scalability to long-duration

 doi:10.20944/preprints202602.1564.v1

untrimmed videos. Integrating linear-complexity architectures, such as Mamba, which operates at $O(T)$, represents a critical direction for enabling large-scale, real-time temporal action localization.

7. Conclusion

This study addresses the critical challenge of "negative transfer" in HAR systems deployed in complex, heavily occluded industrial environments (Industry 5.0). To overcome the limitations of conventional multimodal sensor fusion, we propose a Confidence-Aware TAL system driven by a novel Gated SRM.

The main findings are summarized as follows:

- By integrating the confidence scores from the log-transformed pose estimator as a bias term into the multi-head self-attention layer, this system effectively mitigates the impact of highly uncertain skeleton features through a probabilistic and continuous approach. Despite employing this sophisticated attention mechanism, the computational overhead sufficiently meets real-time processing requirements, achieving approximately 16 FPS—a critical temporal resolution requirement for industrial surveillance applications.
- In evaluations using the heavily occluded IKEA ASM dataset, the proposed framework completely avoided the severe performance drop caused by naive fusion methods (where mAP decreased from 21.49% to 19.29%) and instead improved the overall mAP to 21.77%. A Wilcoxon signed-rank test confirmed this improvement is highly significant ($p < 0.001$) and maintains statistical equivalence with the RGB-only baseline, proving its effectiveness as a robust safeguard against negative transfer.
- Furthermore, on the THUMOS14 dataset, the method demonstrated its capability to effectively leverage high-quality skeletal data in environments with less occlusion, improving the mAP to 66.31% and significantly outperforming both the RGB-only baseline and naive fusion approaches.

Ultimately, these findings demonstrate that in sensing environments with inevitable data loss, dynamically weighting information based on its measured reliability is a fundamentally more robust approach than attempting to reconstruct irreversibly lost data.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org and https://doi.org/10.5281/zenodo.18733477. containing the official PyTorch implementation of the Gated SRM, pre-trained model weights, and evaluation scripts to reproduce the results.

Author Contributions: Conceptualization, M.T. and T.F.; methodology, M.T. ; software, M.T. ; validation, M.T. and T.F. ; formal analysis, M.T. ; investigation, M.T. ; resources, T.F. ; data curation, M.T. ; writing—original draft preparation, M.T. ; writing—review and editing, T.F. ; visualization, M.T. ; supervision, T.F. ; project administration, T.F.. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number JP23K11724.

Institutional Review Board Statement: Not applicable. The study utilizes publicly available datasets and does not involve new studies with human or animal subjects.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed during the current study are available in the THUMOS14 repository (https://www.crcv.ucf.edu/THUMOS14/) and the IKEA ASM dataset repository (https://ikeaasm.github.io/).

Acknowledgments: During the preparation of this work, the authors used Gemini (model: Gemini 3 Pro) and Claude (model: Sonnet 4.5) to improve the readability and language of the text. After using these tools/services, the authors reviewed and edited the content as necessary and took full responsibility for the content of the publication.

 doi:10.20944/preprints202602.1564.v1

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TAL	Temporal Action Localization
HAR	Human Action Recognition
RGB	Red, Green Blue
mAP	mean Average Precision
IoU	Intersection over Union
CNN	Convolutional Neural Network
GCN	Graph Convolutional Network
ST-GCN	Spatial-Temporal Graph Convolutional Network
MLP	Multilayer Perceptron
FPS	Frames Per Second
GT	Ground Truth
SOTA	State-of-the-Art
ASM	Assembly (in IKEA ASM dataset)

References

1. Rehman, S.U.; Yasin, A.U.; Ul Haq, E.; Ali, M.; Kim, J.; Mehmood, A. Enhancing Human Activity Recognition through Integrated Multimodal Analysis: A Focus on RGB Imaging, Skeletal Tracking, and Pose Estimation. Sensors **2024**, 24, 4646. https://doi.org/10.3390/s24144646
2. Tian, Y.; Liang, Y.; Yang, H.; Chen, J. Multi-Stream Fusion Network for Skeleton-Based Construction Worker Action Recognition. *Sensors* **2023**, *23*, 9350. https://doi.org/10.3390/s23239350
3. Luo, X.; Li, H.; Yang, X.; Yu, Y.; Cao, D. Capturing and understanding workers' activities in far-field surveillance videos with deep action recognition and Bayesian nonparametric learning. Comput.-Aided Civ. Infrastruct. Eng. **2019**, 34, 333–351.https://doi.org/10.1111/mice.12419
4. Hu, K.; Shen, C.; Wang, T.; Xu, K.; Xia, Q.; Xia, M. Overview of temporal action detection based on deep learning. Artif. Intell. Rev. **2024**, 57, 26. https://doi.org/10.1007/s10462-023-10650-w
5. Zhang, C.L.; Wu, J.; Li, Y. ActionFormer: Localizing Moments of Actions with Transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October **2022**; pp. 492–510. https://doi.org/10.1007/978-3-031-19772-7_29
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Advances in Neural Information Processing Systems; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Long Beach, CA, USA, **2017**; Volume 30.
7. Sun, J.; Zhang, Y.; Xu, W.; Wang, B.; Hu, D. A novel two-stream Transformer-based framework for multi-modality human action recognition. Appl. Sci. **2023**, 13, 2058. https://doi.org/10.3390/app13042058
8. Ben-Shabat, Y.; Yu, X.; Saleh, F.; Campbell, D.; Rodriguez-Opazo, C.; Li, H.; Gould, S. The IKEA ASM dataset: Understanding people assembling furniture through actions, objects and pose. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 5–9 January **2021**; pp. 847–859. https://doi.org/10.1109/WACV48630.2021.00089
9. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. IEEE Trans. Pattern Anal. Mach. Intell. **2019**, 43, 172–186. https://doi.org/10.1109/TPAMI.2019.2929257
10. Liu, T.; Zhu, P.; Zhang, J.; Liu, H.; Yuan, J. A systematic review of skeleton-based action recognition: Methods, challenges, and future directions. IEEE Trans. Neural Netw. Learn. Syst. **2025** (Early Access). 10.1109/TNNLS.2025.3632689
11. Yoon, Y.; Yu, J.; Jeon, M. Predictively Encoded Graph Convolutional Network for Noise-Robust Skeleton-Based Action Recognition. Appl. Intell. **2022**, 52, 2317–2331. https://doi.org/10.1007/s10489-021-02487-z

12. Shafizadegan, F.; Naghsh-Nilchi, A.R.; Shabaninia, E. Multimodal vision-based human action recognition using deep learning: A review. Artif. Intell. Rev. **2024**, 57, 178. https://doi.org/10.1007/s10462-024-10730-5
13. Song, Y.F.; Zhang, Z.; Shan, C.; Wang, L. Richly activated graph convolutional network for robust skeleton-based action recognition. IEEE Trans. Circuits Syst. Video Technol. **2021**, 31, 1915–1925. https://doi.org/10.1109/TCSVT.2020.3015051
14. Wen, H.; Lu, Z.M.; Shen, F.; Lu, Z.; Zheng, Y.; Cui, J. Enhancing skeleton-based action recognition with feature maps from pose estimation networks. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **2025**, E108-A, 1677–1686. https://doi.org/10.1587/transfun.2024EAP1162
15. Wang, H.; Schmid, C. Action Recognition with Improved Trajectories. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December **2013**; pp. 3551–3558. https://doi.org/10.1109/ICCV.2013.441
16. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.L. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. Int. J. Comput. Vis. **2013**, 103, 60–87. https://doi.org/10.1007/s11263-012-0594-8
17. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. Adv. Neural Inf. Process. Syst. **2014**, 27, 568–576.
18. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21-26 July **2017**; pp. 6299–6308. https://doi.org/10.1109/CVPR.2017.502
19. Gong, P.; Luo, X. A survey of video action recognition based on deep learning. Knowl.-Based Syst. **2025**, 309, 113594. https://doi.org/10.1016/j.knosys.2025.113594
20. Chen, H.; Gouin-Vallerand, C.; Bouchard, K.; Gaboury, S.; Couture, M.; Bier, N.; Giroux, S. Enhancing Human Activity Recognition in Smart Homes with Self-Supervised Learning and Self-Attention. Sensors **2024**, 24, 884. https://doi.org/10.3390/s24030884
21. Lin, T.; Liu, X.; Li, X.; Ding, E.; Wen, S. BMN: Boundary-matching network for temporal action proposal generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November **2019**; pp. 3889–3898. https://doi.org/10.1109/ICCV.2019.00399
22. Xu, M.; Chen, C.; Mei, J.; Zhang, Y.; Wu, Y. G-TAD: Sub-graph localization for temporal action detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June **2020**; pp. 10156–10165. https://doi.org/10.1109/CVPR42600.2020.01017
23. Shi, D.; Zhong, Y.; Cao, Q.; Ma, L.; Li, J.; Yan, Y. TriDet: Temporal Action Detection with Trident-head. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June **2023;** pp. 18853–18862.
24. Zhang, H.; Zhou, F.; Wang, D.; Zhan, Q. LGAFormer: Transformer with Local and Global Attention for Action Detection. *J. Supercomput.* **2024**, *80*, 17952–17979. https://doi.org/10.1007/s11227-024-06138-1
25. Gu, A.; Goel, K.; Ré, C. Efficiently Modeling Long Sequences with Structured State Spaces. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April **2022**.
26. Wen, J.; Liu, D.; Zheng, B. ActionMamba: Action Spatial–Temporal Aggregation Network Based on Mamba and GCN for Skeleton-Based Action Recognition. Electronics **2025**, 14, 3610. https://doi.org/10.3390/electronics14183610
27. Huang, Q.; Cui, J.; Li, C. A Review of Skeleton-Based Human Action Recognition. J. Comput.-Aided Des. Comput. Graph. **2024**, 36, 173–194. https://doi.org/10.3724/SP.J.1089.2024.2023-00358
28. Feng, M.; Meunier, J. Skeleton Graph-Neural-Network-Based Human Action Recognition: A Survey. Sensors **2022**, 22, 2091. https://doi.org/10.3390/s22062091
29. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February **2018**; Volume 32.
30. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June **2019**; pp. 12026–12035. https://doi.org/10.1109/CVPR.2019.01230

 doi:10.20944/preprints202602.1564.v1

31. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and unifying graph convolutions for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June **2020**; pp. 143–152. https://doi.org/10.1109/CVPR42600.2020.00022
32. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June **2015**; pp. 1110–1118. https://doi.org/10.1109/CVPR.2015.7298714
33. Sensoy, M.; Kaplan, L.; Kandemir, M. Evidential deep learning to quantify classification uncertainty. In Advances in Neural Information Processing Systems; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; Curran Associates, Inc.: Montreal, QC, Canada**, 2018**, Volume 31.
34. Neverova, N.; Wolf, C.; Taylor, G.; Nebout, F. Moddrop: Adaptive multi-modal gesture recognition. IEEE Trans. Pattern Anal. Mach. Intell. **2015**, 38, 1692–1706. https://doi.org/10.1109/TPAMI.2015.2461544
35. Ghimire, A.; Kakani, V.; Kim, H. SSRT: A sequential skeleton RGB transformer to recognize fine-grained human-object interactions and action recognition. IEEE Access **2023**, 11, 51930–51948. https://doi.org/10.1109/ACCESS.2023.3278974
36. Liu, X.; Yang, X.; Wang, D.; Zhou, J.; Yang, Q. TadTR: End-to-end temporal action detection with transformer. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September **2021**; pp. 541–545. https://doi.org/10.1109/TIP.2022.3195321
37. Xing, Y. Deep learning-based action recognition with 3D skeleton: A comprehensive study. IET Commun. **2021**, 15, 2369–2378. https://doi.org/10.1049/cit2.12014
38. Idrees, H.; Zamir, A.R.; Jiang, Y.G.; Gorban, A.; Laptev, I.; Sukthankar, R.; Shah, M. The THUMOS challenge on action recognition for videos “in the wild”. Comput. Vis. Image Underst. **2017**, 155, 1–23. https://doi.org/10.1016/j.cviu.2016.10.018
39. Shou, Z.; Wang, D.; Chang, S.F. Temporal action localization in untrimmed videos via multi-stage CNNs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June **2016**; pp. 1049–1058.
40. Zeng, R.; Huang, W.; Tan, M.; Rong, Y.; Zhao, P.; Huang, J.; Gan, C. Graph convolutional networks for temporal action localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November **2019**; pp. 7094–7103.
41. Liu, X.; Bai, S.; Bai, X. An empirical study of end-to-end temporal action detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June **2022**; pp. 20010–20019.
42. Yang, L.; Han, H.; Zhao, H.; Tian, Q.; Zhang, M. Background-click supervision for temporal action localization. IEEE Trans. Pattern Anal. Mach. Intell. **2023**, 45, 15159–15175.
43. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. Int. J. Comput. Vis. **2010,** 88, 303–338. https://doi.org/10.1007/s11263-009-0275-4
44. Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Niebles, J.C. ActivityNet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June **2015**; pp. 961–970. https://doi.org/10.1109/CVPR.2015.7298698
45. Lin, T.; Zhao, X.; Su, H.; Wang, C.; Yang, M. BSN: Boundary sensitive network for temporal action proposal generation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September **2018**; pp. 3–19. https://doi.org/10.1007/978-3-030-01225-0_1
46. Tao, K.; Wang, F.; Liu, Z.; Huang, Y. A lightweight spatiotemporal skeleton network for abnormal train driver action detection. Appl. Sci. **2025**, 15, 13152. https://doi.org/10.3390/app152413152

 doi:10.20944/preprints202602.1564.v1

47. Liu, Z.; Zhang, Z.; Cao, Z.; Kan, H.; Zhu, G.; Tan, M. 3DInAction: Understanding Human Actions in 3D Point Clouds. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June **2023**; pp. 10636–10645. https://doi.org/10.1109/CVPR54722.2023.01024
48. Gao, Y.; Lohmann, C.S.; Schiefer, J.; Capanni, F.; Drayss, T.; Schleyer, C.; Selle, S.; Bauerschmidt, S.J.; Dickhaus, H. How Fast Is Your Body Motion? Determining a Sufficient Frame Rate for an Optical Motion Tracking System Using Passive Markers. *PLoS ONE* **2016**, *11*, e0150993. https://doi.org/10.1371/journal.pone.0150993