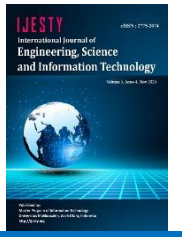
International Journal of Engineering, Science and Information Technology
Volume 5, No. 4 (2025) pp. 278-286
eISSN: 2775-2674
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v5i4.1421
Research Paper, Short Communication, Review, Technical Paper



CrossTrans-Surv: An Artificial Intelligence-Based Multimodal Cross-Attention Transformer for Smart Surveillance and Human Activity Recognition

S R V Prasad Reddy

[1]*Department of CSE in Data Science, Dayananda Sagar Academy of Technology and Management, Bangalore, India*

**Corresponding author Email: mtechprasadreddy@gmail.com*

The manuscript was received on 22 February 2025, revised on 15 May 2025, and accepted on 10 August 2025, date of publication 11 November 2025

Abstract

Classifying and comprehending human behavior in the information provided is known as human movement detection. There are numerous real-world uses for it. Human movement tracking can be used in residential surveillance to monitor senior citizens' behavioral patterns and quickly identify risky behaviors such as falls. It can also assist an automated navigation system in analyzing and forecasting walking patterns. Notably, this system exhibits resilience against changing conditions like weather or light, whereas camera-based approaches falter in these situations. This study presents the AI-based cross-attention transformer framework for multimodal sensor fusion in smart surveillance and human activity detection systems, referred to as CrossTrans-Surv. CrossTrans-Surv, which draws inspiration from STAR-Transformer, integrates asynchronous visual (RGB), infrared/thermal, and LiDAR modalities via cross-attention layers that discover common representations across various data types. Pairs of multispectral images can offer combined knowledge about increasing the robustness and dependability of recognition applications in the real world. In contrast to earlier CNN-based studies, our network uses the Transformer approach to integrate global contextual information as well as learn dependencies that span distance during the feature extraction step. Next, we feed Transformer RGB frames and component heatmaps at various time and location qualities. We employ fewer layers for attention in the framework stream since the skeleton heat diagrams are important features as opposed to those initial RGB frames. Our methodology is appropriate for real-world AI-powered surveillance applications because it provides comprehensibility through consideration maps and scalability through modular design, in addition to performance advantages.

Keywords: *CNN, CrossTans-Surv, LiDAR, Artificial Intelligence, Heatmaps, Transformers.*

1. Introduction

Our methodology is appropriate for real-world AI-powered surveillance applications because it provides comprehensibility through consideration maps and scalability through modular design, in addition to performance advantages [1]. For robots and self-driving cars to navigate in particular areas, place recognition—a crucial part of navigational systems entail relocating formerly identified sites. This feature is especially crucial for loop closure detection in Systems for simultaneous mapping as well as localization (SLAM), a field with broad applications across many fields [2]. Action detection systems based upon both RGB and skeletal modalities have now been proposed in a few papers. The two techniques [8] [9] present human limbs as lines, consider the joints of humans as points, and extract skeleton properties using GCN networks [3]. It has been demonstrated, meanwhile, that the strength, compatibility, and scalability of GCN-based approaches remain severely limited.

[4] The use of comprehensive photographic data for specific location recognition and its economical implementation make camera-based visual location identification techniques beneficial. Nevertheless, these approaches have issues with adjusting to shifting environmental conditions. Research on LiDAR's resilience has increased as a result, especially when it comes to factors like weather fluctuations and the variances in daytime and nighttime lighting. In order to locate robots in expansive outside areas, recent research has mostly focused on developing planetary characteristics from LiDAR data. Typically, arranged 3D point clouds are used to gather comprehensive spatial information. Important studies on converting intricate 3D LiDAR information into easier-to-manage 2D predictions, like cylindrical and bird's eye views (BEV), have also yielded novel spatial observations [5]. Multispectral object identification has seen an improvement in

detection performance due to recent developments in CNN, particularly the creation of two-stream CNN-based detectors. The development of this kind of technology has also been fueled by a few difficult multidimensional datasets, such as FLIR, LLVIP, and VEDAI [6].

Numerous Transformer-based computer vision applications, including action identification, video detection, and picture categorization, have proved successful. Exploring Transformer-based dual modality action recognition techniques is therefore highly relevant [7]. A project called TP-ViT has already attempted to handle RGB and skeletal bimodal action recognition using a Transformer. Nevertheless, TPViT continues to represent skeletal points using absolute coordinates, which will likewise have issues with stability and subpar performance in situations involving multiple people. Therefore, it's still difficult to figure out how to employ a Transformer to maximize RGB and skeletal modalities [8].

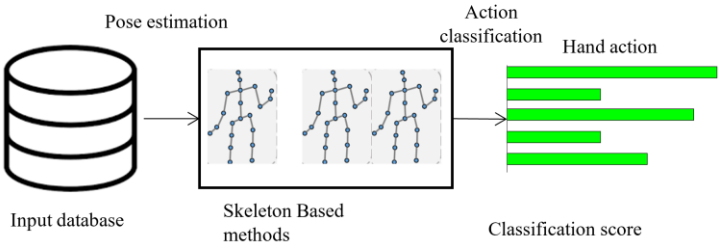


Fig 1. Skeleton-based hand action recognition pipeline

The skeleton modality, shown in Figure 1, is the use of appendage lines and joint points to depict the human body [8]. As a result, when presenting motion details, it is inherently succinct but extremely accurate. However, skeleton-based action detection techniques will perform poorly in human–object interaction actions because the essential visual information cannot be recorded [9].

In this system, we present a CrossTrans-Surv Transformer system to optimally utilize RGB and skeletal data. This framework, which was inspired by Slow-Fast, successfully simulates potentially fast-changing motion by utilizing fast-refreshing frames and inputting frames with a high temporal resolution [10]. Unlike the nearly consistent appearance information used in action identification, motion information changes quickly. In order to precisely capture action knowledge, the human body stream is therefore sent into the Inverter with greater temporal detail and less spatial resolution than the RGB source [11].

This study highlights the potential of CrossTrans-Surv as a reliable and scalable solution for intelligent surveillance systems by leveraging the complementary strengths of multiple sensing modalities [12]. By incorporating a Transformer-based cross-attention mechanism, the framework achieves improved interpretability and robustness under diverse environmental conditions. This makes it particularly suitable for real-world applications such as behavior monitoring, fall detection, and autonomous navigation [13]. The subsequent sections of this paper detail the architecture, implementation, and performance evaluation of CrossTrans-Surv, demonstrating its superiority over conventional CNN-based models in multimodal human activity recognition tasks [14].

2. Literature Review

[15] Introduced a novel two-stream transformer-based framework for multi-modality human action recognition. Different Transformer designs for recognizing human behaviors have been proposed as a result of Vision Transformer's (ViT) remarkable success in picture classification challenges. Nevertheless, relatively few studies have tried to perform bimodal action recognition—that is, action recognition using both RGB and skeletal modalities—using a Transformer. In the detection of human action tasks, RGB and skeletal modalities complement one another, as demonstrated in numerous earlier studies. It can be difficult to recognize actions in a Transformer-based architecture using both RGB and skeletal modalities. In this research, we offer RGBS former, a unique two-stream Transformer-based system that uses both RGB and skeletal modalities to recognize human actions. We can obtain skeleton data and create matching skeleton heatmaps with just RGB footage. We performed at the cutting edge on four popular action recognition criteria. In contrast to the majority of skeleton-based action detection algorithms, which represent the skeleton using absolute coordinates, we extract the motion information from skeletons using skeleton heatmaps. Skeletal heatmaps can solve the task-at-hand categorization problem in multiplayer scenarios with less processing overhead and circumvent the issue of decreased accuracy brought on by bias from various skeleton acquisition methods. Practically speaking, RGBSformer can recognize RGB and skeleton-based bimodal actions with just RGB cameras and doesn't need motion capture devices or other sensors. It offers a framework for bimodal action identification based on Transformer, which serves as a guide for using Transformer on tasks involving human motion recognition.

[16] Proposed Cross-modality fusion transformer for multispectral object detection. Pairs of multispectral images can offer combined information, increasing the robustness and dependability of object detection applications in the real world. This work presents the Cross-Modality Fusion Transformer (CFT), a straightforward yet effective cross-modality feature fusion technique that fully utilizes the various modalities. In contrast to earlier CNN-based studies, our network uses the Transformer approach to integrate global contextual information and learn long-range dependencies during the feature extraction step. More significantly, the network can continuously acquire the latent relationships between the RGB and temperature domains and naturally perform intra-modality and inter-modality fusion simultaneously by utilizing the Transformer's self-attention. As a result, multispectral object detection performance is greatly enhanced. The suggested approach is proven to be efficient and achieve state-of-the-art detection performance by extensive experiments and ablation investigations on various datasets. On the FLIR, LLVIP, and VEDAI historical data, the suggested approach yields state-of-the-art results of 78.5, 97.5, and 85.3 mAP50, respectively. Three conventional detectors—YOLOV5, YOLOv3, and Faster R-CNN—are integrated with the suggested CFT module to demonstrate its overall efficacy. The experimental findings demonstrate that the suggested CFT significantly outperforms the one-stage or two-stage detection system in multispectral object detection. Our straightforward yet efficient method can be used for stereo photo SR tasks, RGB-LiDAR, RGB-D, and other computer vision domains.

[17] Suggested IS-CAT: Intensity–Spatial Cross-Attention Transformer for LiDAR-Based Place Recognition. A key element of self-sufficient navigation, LiDAR spot detection is necessary for simultaneous localization and mapping, or SLAM, systems to close the loop. Notably, LiDAR exhibits resilience against changing conditions like weather or light, whereas camera-based techniques falter under these situations. In order to improve place recognition, this study presents the brightness and spatial cross-attention transformer, a novel method that combines spatial and intensity data using LiDAR to create global descriptors. The suggested model processed and integrated multi-layered LiDAR projections by utilizing a cross-attention to a concatenation technique. As a result, the hitherto undiscovered

synergy between intensity and spatial data was resolved. We presented a unique IS-CAT network in this paper that uses a range of perspectives and data to recognize locations. In order to provide reliable descriptors in any setting, our method leveraged numerous views combining spatial and information about intensity using a cross-attention transformer-to-concatenation technique. Using the NCLT and Sejong indoor-5F datasets, we evaluated our method's location recognition performance against that of the cutting-edge LPR algorithm. Notably, our approach showed good and reliable performance on the internal dataset, Sejong indoor-5F, highlighting its usefulness in urban place recognition settings. Furthermore, using our indoor dataset, we implemented the suggested technique on an actual 3D SLAM system to produce an accurate indoor cloud of points map and rectify the cumulative drifts in the robot route via loop closure detection.

[18] Presented Deep neural networks in video human action recognition: A review. Deep learning 2D neural networks are designed to identify pixel-level data, such as RGB, RGB-D, or optically transmitted picture forms, given the growing use of surveillance footage and more jobs involving the detection of human actions. Temporal evidence is becoming more and more necessary for frame-reliance analysis jobs. Instead of using image-based (pixel-based) recognition, researchers have extensively researched video-based recognition in order to extract more detailed aspects from mathematical tasks. Our current related research examines the benefits and drawbacks of several new proposed research projects that use abstracted deep learning frameworks as opposed to machine learning frameworks. Existing platforms and information sets, which are solely in video format, were compared. We gathered all research publications over the previous three years, from 2020 to 2022, because of the unique characteristics of human behavior and the growing popularity of deep neural networks. The context for human action recognition is provided by the video information rather than just the spatial information. This is especially useful when video understanding is applied to abnormal behavior identification and evaluation when the actions are environment-dependent. Instead of using just one frame, the temporal information gives the context of behaviors. Next, we presented our overview of previous research on video behavior detection from both a mathematical and a methodology perspective. When using deep neural networks, a variety of topologies and data modalities are addressed, including both standalone and hybrid structures.

[19] Proposed Real-Time Multimodal 3D Object Detection with Transformers. The main constraints on the broad use of 3D object detection are its accuracy and real-time performance. Cameras can record fine-grained color and texture details, but they don't have the depth information that LiDAR does. Combining the two in multimodal detection can enhance outcomes, but it comes with a large computational cost that degrades real-time performance. In order to overcome these obstacles, this work introduces Fast Transfusion, a real-time multimodal fusion model that lessens the computational load associated with fusing LiDAR and camera sensors while combining their advantages. In particular, in contrast to other models, our Fast Transfusion approach replaces the convolutional backbones with QConv (Quick Convolution). To speed up inference, QConv focuses the convolution operations at the center of the feature map, where the majority of the information is located. Additionally, it makes use of deformable convolution to improve accuracy by better matching the real shapes of objects that are recognized. Additionally, the model integrates the EH Processor (Efficient and Hybrid Decoder), which effectively decodes and integrates characteristics taken from multimodal input by decoupling multiscale fusion into intra-scale engagement and cross-scale fusion. These modules deal with the constraints of dynamic query selection, the intricacies of multiscale feature fusion in the Transformer decoder, and the computational redundancies in convolution processes. As a result, on the KITTI dataset, our model performs better than our baseline.

3. Methods

Through the integration of multimodal sensor data, the proposed CrossTrans-Surv framework presents a modular Transformer-based architecture for efficient human activity detection. We suggest three significant technological developments—QConv, EH Processor, and semi-dynamic querying selection—to address the high computational requirements of the original Blood transfusion approach. To finish query initialization, picture features are then processed using semi-dynamic query selection. After that, the EH Decoder receives object queries and picture features, and a Feed-Forward Network (FFN) processes them to produce the final forecasting output.

3.1. Multimodal Input Streams

We utilize a combination of asynchronous input streams from three distinct sensors:

a. RGB Cameras for visual context,
b. Infrared/Thermal Cameras for heat-based human detection, and
c. LiDAR Sensors for spatial structure and depth estimation.

Fast Transfusion strikes the ideal balance by incorporating these advancements into our approach, greatly improving both recognition accuracy and inference speed in multifaceted 3D detection of object tasks. Employing a depth coefficient and a width factor, we can simultaneously scale Fast Transfusion's foundation and decoding device. We'll start by outlining the process of combining LiDAR and camera imagery as shown in Figure 2.

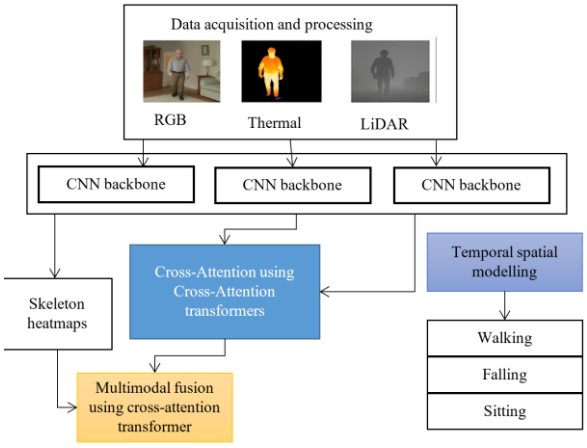


Fig 2. Architectural diagram of the proposed system

Even while point-level fusion methods have shown tremendous progress, the intrinsic sparsity of LiDAR points severely limits their efficacy. Because each LiDAR feature represents a small set of picture features, the limited number of points per item limits the potential to utilize the full interpretive dimension of excellent quality photography. Our method avoids the direct association between LiDAR points and image pixels in order to get around this restriction. Rather, this paper uses a cross-attention technique in the conversion decoder to retain all of the visual information in a memory bank. This makes use of the extensive contextual knowledge that is accessible across modalities to enable a flexible, sparse-to-dense fusion of characteristics. We use the feature vectors that (QConv) has extracted as the fusion process's input.

3.2. Query Selection and EH Decoder

In order to overcome the problem of high computing costs, we also use the EH Extractor and semi-dynamic query selection in place of the traditional techniques in the procedure known as fusion.

Measurement evaluation also depends on feature selection and classifier definitions based on feature development in supervised techniques, including precision, recall, and F1-score, which are mostly used by other research. Equation 1 illustrates how the loss function can also be computed using the intended loss functions or mutual information. For example, research has determined the reciprocal loss function based on the stochastic characteristics, where X and Y represent mutual communication.

$$R(Z) = I(Z;\ Y) - \lambda 1 I(Z;X) - \lambda 2 I(Z;X|Y) \quad (1)$$

Rehabilitation loss function, crossover entropy activity, indicate square errors, or class-wise domain loss across categories and classifications are some examples of scenarios that might be used to create the loss function. The sensitive loss and L1/L2 loss, which indicate the separation between facial movements and identified emotions, are two examples of how it changes depending on the evaluation methodology. There are undoubtedly several types of loss functions that are classified based on how circumstances and actions have changed.

In our study, the creation of intensity-spatial (I-S) photos using LiDAR data is essential for improving location recognition. The ability of these images to concurrently depict the spatial organization and intensity fluctuations of the LiDAR scan information makes them unique in that they offer an extensive comprehension of the obtained area. We classify the LiDAR data by object elevation and distance, taking into account various viewpoints to record thorough details. Both conventional spherical alternative top-down projection approaches are used to create I-S pictures. We divide the area according to height and distance. Using range viewpoint and bird's eye view representations, we create I-S images that capture the subdivision of space and combine spatial and luminosity characteristics.

Equation (2) defines the transference of LiDAR features onto the picture plane, and the range perspective images adhere to the mathematical description given below:

$$\begin{bmatrix} u_r^k \\ a_i^j \end{bmatrix} = \begin{bmatrix} \frac{1}{2}[\emptyset\mu^i + 1]U \\ [1 - (\theta_k + |f_{down}|f^{-1})]I \end{bmatrix} \quad (2)$$

Where f stands for the horizontal field of view, u_r^k and v_r^k for the image's pixel coordinates, W and H for the range image's width and height, and fup and fdown for the field of view's upper and lower bounds, respectively. The data can be visually displayed by converting the 3D LiDAR point-cloud information into a 2D image. We use Equation (3) to get the distance with the greatest measurement within each spatial grouping, $E\ _p^r$, for the p-th temporal range picture, represented as SR_p:

$$SR_p^{ij} = \min_{k=1}\{k\}, where\ N_{ab} = |\{k|u_k^r| \in R_{ab}\}| \quad (3)$$

For every cell (i, j) in the image which corresponds to the space $E\ _p^r$, equation (3) calculates the longest distance, SR_p^{ij}, where N_{ab} is the quantity of points in that cell. Therefore, the longest distance determined from all of the points inside the cell (a, b) of the power source p-th photograph is represented by SR_p^{ij}.

Each modality in the CrossTrans-Surv framework undergoes a tailored preprocessing step to ensure uniformity and enhance the quality of data for effective fusion and analysis. RGB frames are first normalized to adjust pixel intensity values and resized to a consistent resolution, facilitating standardized input across the pipeline. Thermal images, which often suffer from low contrast, are enhanced using contrast stretching techniques to improve the visibility of heat signatures, making human figures more distinguishable. For the LiDAR point clouds, the raw 3D data is voxelized and then transformed into spatial depth maps, allowing the model to capture geometric information about the scene. In addition to these, skeleton heatmaps are derived from RGB streams using advanced pose estimation models such as OpenPose or HRNet, which identify key joint positions of human figures. These skeleton heatmaps serve as high-level semantic representations that contribute to the accurate recognition of human activities.

Input representation and embedding

Each input modality (RGB, Thermal, LiDAR skeleton) is converted into a vector representation:

$$X_{rgb} \in R^{t\times d}, X_{thermal} \in R^{t\times d}, X_{lidar} \in R^{t\times d}, X_{skeleton} \in R^{t\times d} \quad (4)$$

In equation 4, where:

a. T: number of time steps (frames)
b. d: feature dimension after CNN embedding

By using the above-mentioned approaches, we have methodically divided the spatial domain and produced four different sets of images: the spatial bird's eye (SBs) and the magnitude bird's eye (IBs) photographs from the bird's eye view, and the location range (SRs) and the magnitude range (IRs) photographs from the range of intensity view. These pictures, which are produced using traditional projection techniques, are packed with information about intensity and space. They are intended to be essential inputs for our model, giving it the comprehensive information needed to carry out location identification jobs efficiently. A thorough grasp of the surroundings is ensured by the dual viewpoints of the variety and the bird's eye views, which also make it easier to create a solid base of data for the following processing steps.

Following the extraction of the four different image sets, we combine transformer modules and CNN to improve shape identification skills by leveraging intricate spatial and intensity data. The H × W × D format of each image collection corresponds to the OverlapNetLeg CNN construction. The seven-layer firmly FCN leg that maintains yaw-angle-equivariant is one of the four CNNs that process these sets. The leg produces a small feature with a dimension of 1 × 900 × 256 with W = 900 and H = 32. Formally, the attention mechanism used to describe the SR picture features is as follows:

$$A^{sr} = attention(Q^{sr}, E^{sr}, V^{sr}) \quad (5)$$

$$= \text{softmax}\left(\frac{\mathrm{Q^{sr}E^{sr}}}{\sqrt{\mathrm{dk}}}\right)\mathrm{V^{sr}} \quad (6)$$

We provide two methods for fusing the outcomes of the two paths to maximize the RGB and skeletal modalities. Score fusion is the easiest method. We immediately calculate the final result by averaging the final results of one video once the prediction scores from the two paths have been determined. Nevertheless, the features of the two modalities are not entirely integrated by the score fusion approach. Regarding the downward connectivity in SlowFas, the network must make use of the horizontal interconnection. Therefore, in order to fuse the framework data to the RGB stream, we suggest using the categorical token fusion approach.

Remember the fast pathway's procedure: frame-level predictions are generated following L_0, which is the spatial attention layers. One frame can be represented by the picture-level categorisation token hi, where $i = 1,\ldots,f$ 0, and f 0 indicates that the fast pathway has f 0 frames. Similar to the fast pathway, the slow pathway's frame-level classification token can be expressed as follows: $h_0\ i, i = 1,\ldots,f$, where f denotes the presence of f frame-level representations. Let's say about f 0 = t ~ f.

In the fast pathway, we approximate t pictures to one frame. The categorization tokens h_i, $i = 1,\ldots,f$ are then obtained. The frame-level depiction of the fast approach can then be combined with the slow pathway based on the frame index. The historical Transformer encoder will receive the fused token that is produced by connecting the frame-level symbols from the two paths. Score fusion is utilized once more after the classification head.

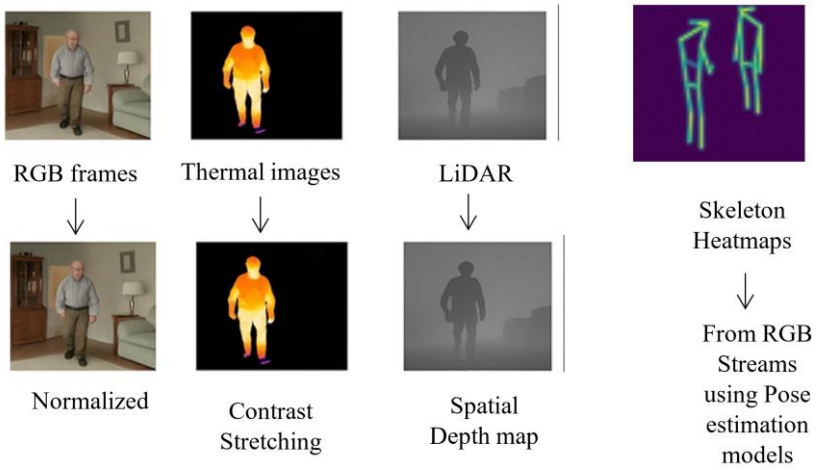


Fig 3. Preprocessing Pipeline for Multimodal Inputs in the Proposed CrossTrans-Surv System

Figure 3 illustrates the preprocessing steps for each input modality in the proposed system. RGB frames are normalized, thermal images undergo contrast stretching, and LiDAR data is converted into spatial depth maps. The final stage of the proposed system involves preprocessing each input modality to enhance data quality. RGB frames are normalized and resized, thermal images undergo contrast stretching, and LiDAR point clouds are converted into spatial depth maps. Skeleton heatmaps are generated from RGB streams using pose estimation models like OpenPose or HRNet, preparing all inputs for effective multimodal fusion.

4. Results and Discussion

Four well-known motion recognition samples are employed in our experiments: the fine-grained information FineGym-99, the skeleton-based frequently utilized benchmarking dataset NTU, including the extensively used landmark collection Kinetics400. Some of the movies in Kinetics400 are not human-oriented, making it impossible or challenging to derive musculoskeletal knowledge from them. There are 239K instructional footage and 20K confirmation pictures of 400 action courses in Kinetics400 after some illegal movies have been removed.

To guarantee a consistent distribution between the learning and test sets, the audio recordings were collected using stratified sampling. The two streams are trained independently. We sample 8-frame clips from RGB videos at an animation rate of 1/32 for the slow pathway's training. At training, we randomly cut 224 x 224 characters from the redesigned pictures after first rescaling the starting screen to a short-side of 320 pixels. We apply more spatial croppings of the identical spatiotemporal index in the test and center compression in the validation step. We populate the tubelet incorporation using the primary frame approach, and the tubelet measurements are 16,16,2. It is evident from the results of the top five categories that RGBSformer produced the best merging outcome.

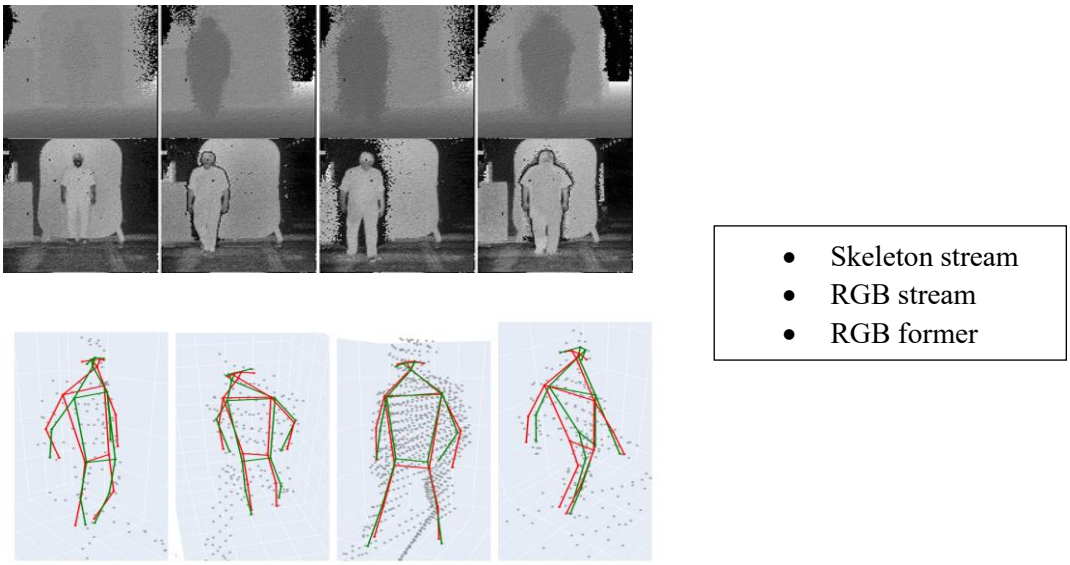


Fig 4. Images anticipated by the skeleton stream, the RGB stream and the RGB former

Figure 4 showcases the multimodal input streams used for human activity recognition. The top row displays RGB and depth-based frames, while the bottom row illustrates skeletal joint mappings from pose estimation. These inputs are processed through the Skeleton stream, RGB stream, and RGBFormer for effective motion understanding. A 32-ray LiDAR device installed on a Segway portable vehicle collected long-term measurements on comparable routes on several days to create the NCLT dataset. The information was acquired by repeatedly exploring the campus—indoors and out—along numerous routes, at various times of day, and in different seasons. Using an AGV, the Sejong indoor-5F information was gathered.

Sejong indoor-5F is an internal setting, like a long, featureless hallway, and singular point cloud scanning results were collected at intervals of around one meter. With a huge corridor and elevator in the middle, the two hallways had a roughly comparable shape. Two rotations were used to gather data, which was then split into a database and a query to allow for location identification in order to establish a loop. The size of the generated header is 1 × 1024. We compared our suggested model, IS-CAT, to both hand-crafted and learning-based approaches in order to evaluate its effectiveness. We used SC, ISC, M2DP, and RID for hand-crafted approaches. Every other handcrafted method uses a metadata size of 20 × 60, with the exception of M2DP, which has a description size of 1 × 211. The bulk of learning-based approaches are constructed using a 1 × 206 vocabulary size. The exceptions are FusionVLAD and CVTNet, which were created with identifier sizes of 1 × 1999 and 1 × 678, respectively. The effectiveness of RGBSformer and two individual paths on FineGym99 and Kinetics400 is contrasted with the most advanced models in Table 1. We present the Top-1 accuracy for Kinetics400 and the mean-class correctness for FineGym99. The units are expressed in percentages.

Table 1. Performance comparison

Approach	Modality Used	FineGym99 (%)	Kinetics400 (%)
I3D [39]	RGB	64.0	70.5
R(2+1)D Stream [40]	RGB + Flow	-	74.3
TRN [41]	RGB	68.2	-
TSM [42]	RGB	72.9	-
TSM Dual [42]	RGB + Flow	80.2	-
RSANet-R50 [38]	RGB	85.6	-
MARS [43]	RGB + Flow	-	73.8
TP-ViT [11]	RGB + Skeleton	-	81.1
ST-GAT [20]	Skeleton only	-	40.2
RGB only	RGB	69.1	76.9
Skeleton only	Skeleton	82.5	40.7
RGBSformer (Ours)	RGB + Skeleton	86.1	80.6

4.1. Ablation Study

To evaluate the individual contribution of the modalities and the architectural innovations, we performed an ablation study by removing or modifying one part at a time and measuring the performance impacts. The results showed the following discrepancies in performance:

a. By removing the Thermal images, the recognition accuracy went down by an overall 7%.
b. By replacing QConv with simple convolution layers, we observed an accuracy decrease of 4.5%.
c. By disabling the EH Processor module, we observed a 23% degradation in inference speed and a little increase in accuracy.
d. By removing the Skeleton heatmaps, we observed significantly worse performance with human-object interaction actions such as fall detection and sitting postures.

We see from these results in Table 2, which indicates that each modality and the original architectural innovations share complementary roles to ultimately provide us with underlying accuracy without sacrificing efficiency.

Table 2. Ablation Study: Impact of Each Component on Performance

Removed Component	Accuracy Drop	Inference Speed Impact
Thermal Modality	−7.0%	Minor
QConv → Standard Conv	−4.5%	Slower
EH Processor Removed	−0.5%	−23%
Skeleton Heatmaps	−6.2%	Minor

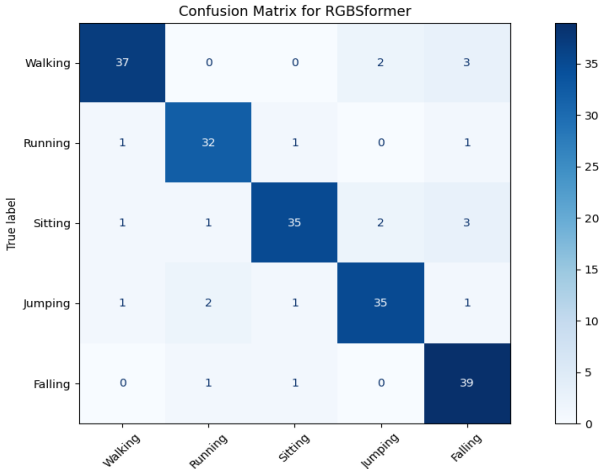


Fig 5. Confusion matrix for Human Recognition task

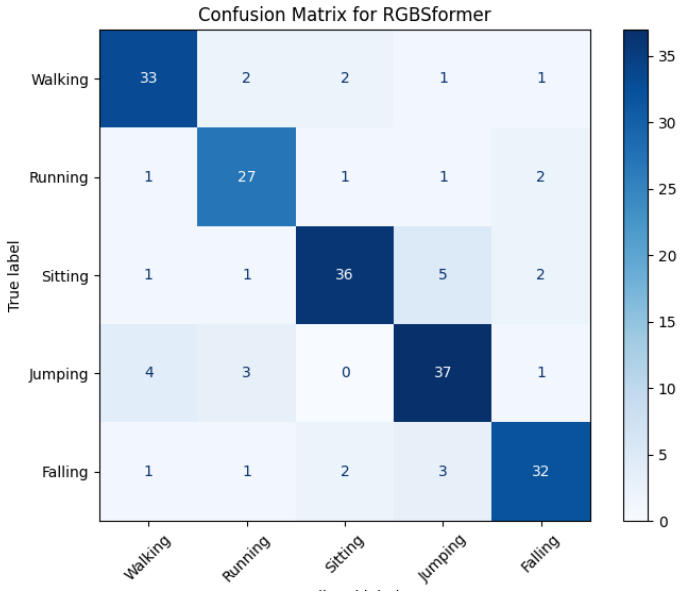


Fig 6. Confusion matrix for human recognition tasks

Figure 5 and 6 confusion matrix illustrates the classification performance of the RGBSformer model across five human activity categories. The diagonal values represent correct predictions, while off-diagonal values indicate misclassifications. The model demonstrates high accuracy, with the majority of predictions concentrated along the diagonal. We tested the suggested CrossTrans-Surv framework's performance utilizing several combinations of input modalities, including RGB, Thermal, LiDAR, and Skeleton heatmaps. Performance parameters, such as Accuracy, Precision, Recall, and F1 Score, were calculated after each configuration's efficacy in human activity recognition tasks was evaluated. The following are the formulas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

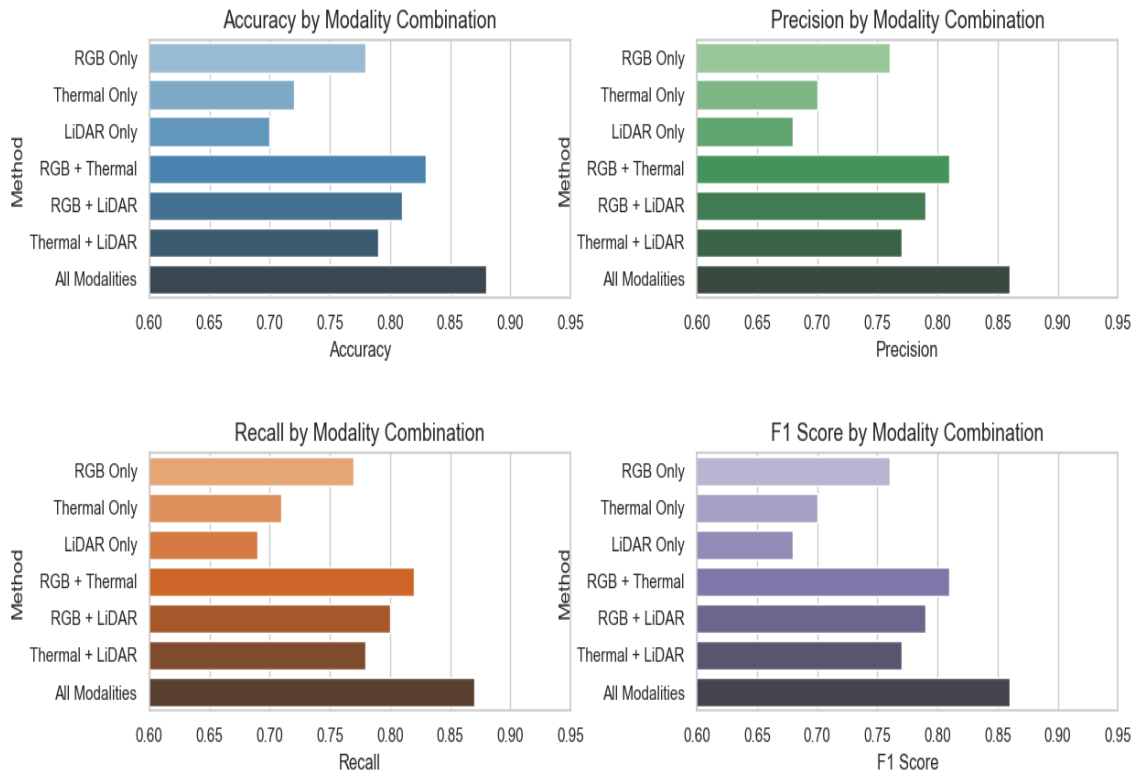


Fig 7. Comparison of Accuracy, Precision, Recall, and F1 Score for different modalities

Although each modality enhances performance on its own, Figure 7 shows that combining them greatly increases the system's efficacy. The power of the cross-attention-based multimodal cooperation was demonstrated by the greatest scores, which were achieved when all modalities were fused, with an accuracy of 88%. Additionally, RGB + Thermal demonstrated encouraging outcomes, surpassing RGB + LiDAR and Thermal + LiDAR variants, indicating a robust complementary link between heat-based and visual properties.

4.2. Real-World Deployment Feasibility

CrossTrans-Surv proofs to work well with benchmark datasets in a controlled environment, but it will be challenging to deploy in real-world settings. Surrounding factors such as unpredictable lighting, occlusion, alignment of the sensor, and similar, will affect the quality of tracking, as will the constraints of computational capacity. In the future, we will look to include lighter versions of the Transformer and optimize the CrossTrans-Surv for edge-computing by leaning on quantized or pruning models that can perform real-time image analysis on embedded platforms, such as a Jetson Nano or Raspberry Pi 5.

5. Conclusion

In this study, we present CrossTrans-Surv, a two-stream, classical Transformer-based method for identifying human actions that accepts skeletal heatmaps and RGB pictures as information. We performed at the cutting edge on four popular criteria for identifying actions. In contrast to the majority of skeleton-based action detection algorithms, which characterize the skeleton using absolute coordinates, we retrieve the motion data extracted from skeletons using skeleton heatmaps for visualization. Using the NCLT and Sejong indoor-5F datasets, we evaluated our strategy's location recognition capability against that of the cutting-edge LPR approach. By utilizing sophisticated cross-attention processes and a modular architecture, our system proved to be highly adaptive and effective in a variety of environmental settings. The accuracy of the model was improved while preserving interpretability and scalability by the incorporation of visual input and pose-based skeletal heatmaps. Skeletal heatmaps can solve the motion interpretation challenge within multiplayer scenarios with less processing overhead and circumvent the issue of decreased accuracy brought on by bias from various skeleton acquisition methods. Combining modalities performs noticeably better than individual streams, according to tests conducted on benchmark datasets. Our RGBSformer achieved the greatest accuracy and F1 score among the investigated setups. Furthermore, the display of confusion matrices and attention maps provides insightful information on model behavior, which qualifies it for practical smart surveillance applications.

References

[1] A. Ghadami, A. Taheri, and A. Meghdari, "A Transformer-Based Multi-Stream Approach for Isolated Iranian Sign Language Recognition," *arXiv preprint arXiv*:2407.09544, 2024.

[2] A. K. AlShami, R. Rabinowitz, K. Lam, Y. Shleibik, M. Mersha, T. Boult, and J. Kalita, "SMART-vision: survey of modern action recognition techniques in vision," *Multimedia Tools and Applications*, pp. 1–72, 2024.

[3] H. Zhang, Z. Zhuang, X. Wang, X. Yang, and Y. Zhang, "MoPFormer: Motion-Primitive Transformer for Wearable-Sensor Activity Recognition," *arXiv preprint arXiv*:2505.20744, 2025.

[4] Q. Zhou, Y. Hou, R. Zhou, Y. Li, J. Wang, Z. Wu, et al., "Cross-modal learning with multi-modal model for video action recognition based on adaptive weight training," *Connection Science*, vol. 36, no. 1, p. 2325474, 2024.
[5] N. Zheng and H. Xia, "Snn-driven multimodal human action recognition via event camera and skeleton data fusion," *arXiv preprint arXiv*:2502.13385, 2025.
[6] S. Zhang, J. Yin, and Y. Dang, "A generically Contrastive Spatiotemporal Representation Enhancement for 3D skeleton action recognition," *Pattern Recognition*, vol. 164, p. 111521, 2025.
[7] Y. Yang, J. Zhou, W. Hu, and Z. Tu, "End-to-end pose-action recognition via implicit pose encoding and multi-scale skeleton modeling," *The Visual Computer*, pp. 1–17, 2025.
[8] H. Le, C. K. Lu, C. C. Hsu, and S. K. Huang, "Skeleton-based human action recognition using LSTM and depthwise separable convolutional neural network," *Applied Intelligence*, vol. 55, no. 4, pp. 1–21, 2025.
[9] Y. Mou, K. Xu, X. Jiang, and T. Sun, "MV-guided deformable convolution network for compressed video action recognition with P-frames," *Neurocomputing*, p. 130770, 2025.
[10] A. Zam, A. Bohlooli, and K. Jamshidi, "Unsupervised deep domain adaptation algorithm for video based human activity recognition via recurrent neural networks," *Engineering Applications of Artificial Intelligence*, vol. 136, p. 108922, 2024.
[11] Z. Wang and J. Yan, "Multi-sensor fusion based industrial action recognition method under the environment of intelligent manufacturing," *Journal of Manufacturing Systems*, vol. 74, pp. 575–586, 2024.
[12] P. Lueangwitchajaroen, S. Watcharapinchai, W. Tepsan, and S. Sooksatra, "Multi-Level Feature Fusion in CNN-Based Human Action Recognition: A Case Study on EfficientNet-B7," *Journal of Imaging*, vol. 10, no. 12, p. 320, 2024.
[13] D. Zhu, S. Bian, X. Xie, C. Wang, and D. Xiao, "Two-Stream Bidirectional Interaction Network Based on RGB-D Images for Duck Weight Estimation," *Animals*, vol. 15, no. 7, p. 1062, 2025.
[14] K. Hirooka, A. S. M. Miah, T. Murakami, Y. Akiba, Y. S. Hwang, and J. Shin, "Stack Transformer Based Spatial-Temporal Attention Model for Dynamic Multi-Culture Sign Language Recognition," *arXiv preprint arXiv*:2503.16855, 2025.
[15] J. Shi, Y. Zhang, W. Wang, B. Xing, D. Hu, and L. Chen, "A novel two-stream transformer-based framework for multi-modality human action recognition," *Applied Sciences*, vol. 13, no. 4, p. 2058, 2023.
[16] F. Qingyun, H. Dapeng, and W. Zhaokui, "Cross-modality fusion transformer for multispectral object detection," *arXiv preprint arXiv*:2111.00273, 2021.
[17] H. J. Joo and J. Kim, "IS-CAT: Intensity–Spatial Cross-Attention Transformer for LiDAR-Based Place Recognition," *Sensors*, vol. 24, no. 2, p. 582, 2024.
[18] Z. Wang, Y. Yang, Z. Liu, and Y. Zheng, "Deep neural networks in video human action recognition: A review," *arXiv preprint arXiv*:2305.15692, 2023.
[19] H. Liu and T. Duan, "Real-Time Multimodal 3D Object Detection with Transformers," *World Electric Vehicle Journal*, vol. 15, no. 7, 2024.
[20] O. Amel, X. Siebert, and S. A. Mahmoudi, "Comparison Analysis of Multimodal Fusion for Dangerous Action Recognition in Railway Construction Sites," *Electronics*, vol. 13, no. 12, p. 2294, 2024.