IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 May 2023, accepted 16 May 2023, date of publication 22 May 2023, date of current version 1 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3278974

RESEARCH ARTICLE

SSRT: A Sequential Skeleton RGB Transformer to Recognize Fine-Grained Human-Object Interactions and Action Recognition

AKASH GHIMIRE[1], **VIJAY KAKANI**[1], **(Member, IEEE), AND HAKIL KIM**[2], **(Member, IEEE)**
[1]Department of Integrated System Engineering, School of Global Convergence Studies, Inha University, Incheon 402-751, South Korea
[2]Department of Information and Communication Engineering, Inha University, Incheon 402-751, South Korea

Corresponding author: Hakil Kim (hikim@inha.ac.kr)

This work was supported by Inha University's Research Grant.

ABSTRACT Combining skeleton and RGB modalities in human action recognition (HAR) has garnered attention due to their ability to complement each other. However, previous studies did not address the challenge of recognizing fine-grained human-object interaction (HOI). To tackle this problem, this study introduces a new transformer-based architecture called Sequential Skeleton RGB Transformer (SSRT), which fuses skeleton and RGB modalities. First, SSRT leverages the strength of Long Short-Term Memory (LSTM) and a multi-head attention mechanism to extract high-level features from both modalities. Subsequently, SSRT employs a two-stage fusion method, including transformer cross-attention fusion and softmax layer late score fusion, to effectively integrate the multimodal features. Aside from evaluating the proposed method on fine-grained HOI, this study also assesses its performance on two other action recognition tasks: general HAR and cross-dataset HAR. Furthermore, this study conducts a performance comparison between a HAR model using single-modality features (RGB and skeleton) alongside multimodality features on all three action recognition tasks. To ensure a fair comparison, comparable state-of-the-art transformer architectures are employed for both the single-modality HAR model and SSRT. In terms of modality, SSRT outperforms the best-performing single-modality HAR model on all three tasks, with accuracy improved by 9.92% on fine-grained HOI recognition, 6.73% on general HAR, and 11.08% on cross-dataset HAR. Additionally, the proposed fusion model surpasses state-of-the-art multimodal fusion techniques like Transformer Early Concatenation, with an accuracy improved by 6.32% on fine-grained HOI recognition, 4.04% on general HAR, and 6.56% on cross-dataset.

INDEX TERMS Multimodality fusion, human action recognition, fine-grained actions, transformer cross-attention fusion.

I. INTRODUCTION

RGB video data encompasses both temporal and spatial information, including details about human limbs and interactions with objects [9], [10]. However, extracting human actions from RGB data can pose a challenge due to the diversity in surroundings, angles of observation, human proportions, and illumination settings. On the other hand, skeleton modality data encodes human body joint movements, capturing motion-related information and making it highly suitable for HAR tasks [21], [43], [44]. This modality is scale-invariant and robust to variations in clothing textures and backgrounds, ensuring reliable action recognition across different subject sizes and situations. Nonetheless, the limitation of skeleton data is its lack of spatial information, which complicates the accurate prediction of fine-grained Human-Object Interaction (HOI). These actions involve similar limb and joint movements but vary in HOI aspects, as illustrated in Figure 1. In this series of images, the action of drinking differs due to the person's interactions with various objects (cup, can, and bottle). To accurately classify fine-grained HOI, it is beneficial to merge the strengths of both RGB and skeleton

The associate editor coordinating the review of this manuscript and approving it for publication was Thomas Canhao Xu.

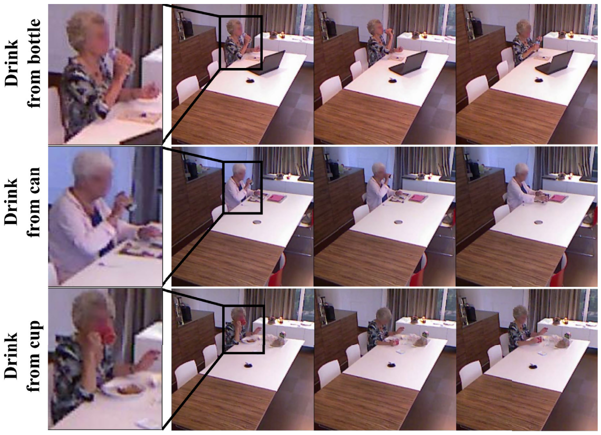


FIGURE 1. Examples of fine-grained human-object interactions.

modalities into a cohesive set of distinguishing features. Consequently, in the field of HAR, combining skeleton and RGB modalities is currently a significant research focus.However, most of the previous studies [12], [13], [16] primarily focused on achieving state-of-the-art performance on well-known datasets, such as NTU-RGB+D [27], and often overlooked the following key concerns:

1) **Recognition of fine-grained HOI** : Most existing HAR studies [12], [13], [36], [37], [42] that fuse skeleton and RGB modalities have mainly concentrated on recognizing broad interaction categories. These studies have assessed datasets like NTU RGB+D [27], which contain only coarse-grained HOI that can be accurately classified using high-quality skeleton modality features alone.
2) **Bias to RGB modality features**: During the training process of a model that utilizes both RGB and skeleton modalities, there is often a prevalent bias towards appearance. This bias can limit the model's generalization capabilities for unseen videos and increase its susceptibility to deception by out-of-context videos [11]. Research on the fusion of these two modalities suggests that the multimodal methods employed in these studies offer only slight enhancements over RGB-based HAR models [12], [14]. Furthermore, none of these studies [12], [14], [15], [16], [37], [42] assess the robustness of their proposed models when facing cross-dataset.

Some limited research [14], [15], [16] has explored the integration of RGB and skeleton modalities using datasets containing fine-grained HOI, such as Toyota Smarthome [14], but their primary focus was not on improving the accuracy of fine-grained HOI. First, these studies were conducted on overall action classes, with only a few actions involving fine-grained HOI. Furthermore, although multimodal strategy in [14] demonstrated enhanced accuracy for certain fine-grained HOI classes, this improvement was not observed in other fine-grained HOI classes possessing the same coarse label. The reason for this might be an uneven distribution of training and testing instances in fine-grained HOI with identical coarse labels.

In the field of HAR, recent advancements are largely driven by the success of transformer-based multimodality architectures [35], [37], [40], [41], [42], demonstrating state-of-the-art results. Building on such success, this paper introduces a new transformer-based architecture called Sequential Skeleton RGB Transformer (SSRT) that fuses skeleton and RGB modalities. First, SSRT harnesses the power of Long Short-Term Memory (LSTM) and a transformer multi-head attention mechanism to obtain abstract features from the skeleton and RGB modalities. Subsequently, SSRT employs a two-stage fusion approach, consisting of transformer cross-attention multimodal fusion [34] and Softmax layer (*SML*) late fusion, to efficiently integrate abstract features from two modalities. The architecture of SSRT is illustrated in Figure 2.

Efficiently complementing heterogeneous modalities such as RGB and skeleton modality presents challenges, as mentioned by Joshi et al. [57]. Unlike conventional fusion techniques like early fusion, late fusion, concatenation, or weighted sum, which do not effectively address the heterogeneity of RGB and skeleton modalities, SSRT utilizes multi-head attention mechanisms. These mechanisms, as demonstrated in [29], can determine information from various representation subspaces at distinct locations. This approach captures the two counterintuitive modalities and provides a more accurate fusion method than traditional techniques shown in [17], [18], [19], and [20].

The primary objective of this study is to solve the problem of recognizing fine-grained HOI. To accomplish this goal, this study utilizes balanced sets of training, validation, and test data for each fine-grained HOI class in the Toyota Smarthome dataset to prevent any bias towards specific actions within the same coarse label. Apart from recognizing fine-grained HOI, the study assesses SSRT on action classes other than fine-grained HOI to confirm that the model generalizes well across various scenarios. Additionally, to ensure that SSRT is robust and not biased toward RGB features, the study evaluates the proposed method on classes from cross-dataset actions in the ETRI-Activity3D dataset [28].

The main contributions of this research include the following:

1) A new method for merging skeleton and RGB data in human activity recognition, SSRT initially extracts high-level features from both the skeleton and RGB modalities using a unique technique that employs LSTM and a transformer encoder to capture improved high-level temporal dependencies of two modalities. Following this, SSRT combines the high-level features from the skeleton and RGB modalities through a two-stage fusion process: transformer cross-attention and softmax layer late score fusion.
2) We evaluate the performance of skeleton and RGB modalities on fine-grained HOI, general, and

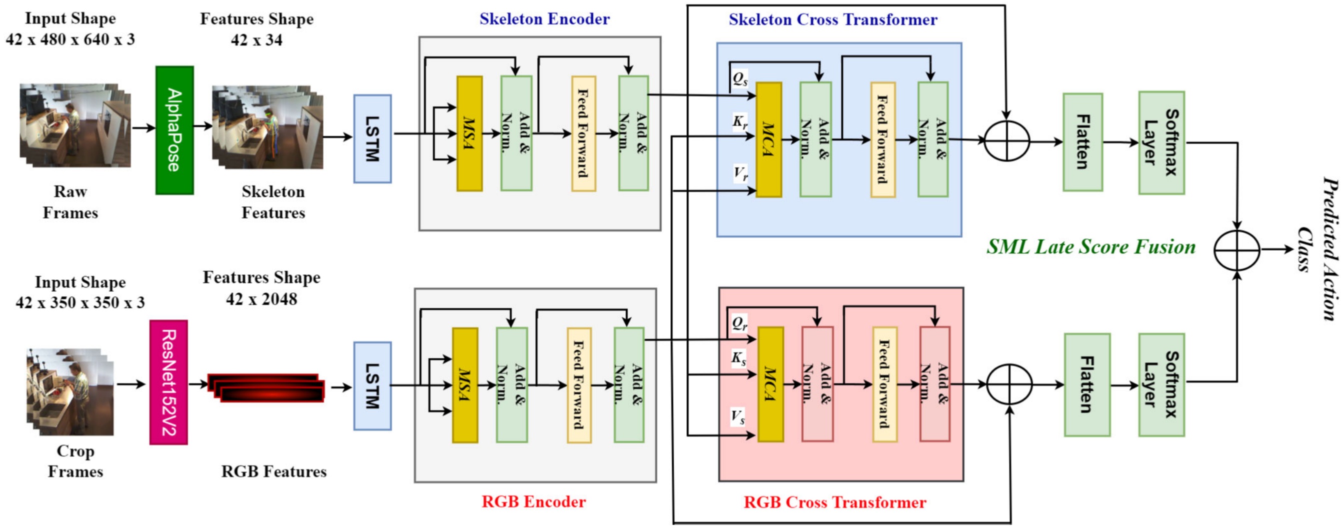


FIGURE 2. Overview of SSRT architecture: A sequential skeleton RGB transformer designed for fine-grained human-object interaction and action recognition.

cross-dataset HAR tasks. Following this, the best-performing single-modality HAR model for each HAR task is compared to SSRT, which incorporates both the skeleton and RGB modalities.

3) To ensure a fair comparison among skeleton, RGB, and multimodal approaches, the study employs a state-of-the-art transformer architecture action recognition model tailored to each modality-based HAR.
4) The novelty of this research is fourfold. First, to the best of our knowledge, SSRT is a pioneering model that employs a combination of LSTM and a transformer architecture as described above. Second, SSRT is the first to fuse skeleton and RGB modalities using a transformer cross-attention multimodal mechanism. Third, this research is the first of its kind to primarily concentrate on recognizing fine-grained HOI and action recognition. Fourth, we are the first to assess the performance of a multimodal features-based HAR model on cross-dataset actions.

II. RELATED WORKS

A. TRANSFORMERS FOR HAR

Research in HAR using transformer-based architectures, such as ViViT [8], TimeSformer [45], MVT [39], AcT [44], and STST [43], has demonstrated state-of-the-art results. Using RGB modality features, ViViT [8] and TimeSformer [45] extract spatiotemporal features separately, whereas MVT [39] uses two transformer encoders to process different views of input frames. Transformer-based models such as AcT [44] and STST [43] have also explored the use of skeleton modality for HAR tasks. AcT utilizes 2D skeleton representations for efficient real-time performance, whereas STST [43] uses spatial and directional temporal transformer blocks for modeling skeleton sequences.

In this study, we select a model resembling the one used in AcT [44] as our primary model for HAR with a single modality. This choice was made because the AcT model not only outperforms state-of-the-art models such as SR-TR [53] and MS-G3D(J+B) [54], but also provides a low latency solution. Furthermore, our proposed multimodal HAR model, SSRT, employs a model resembling the AcT model as a baseline, combined with LSTM, to extract high-level temporal dependencies from the skeleton and RGB modalities. However, our study introduces a distinct transformer architecture that differs from AcT.

B. MULTIMODAL FUSION FOR HAR

1) TRANSFORMER-BASED MULTIMODALITY FUSION FOR HAR

The fusion of multimodal information through transformer architectures has recently garnered considerable interest, particularly in uniting RGB and language modalities to achieve state-of-the-art performance in vision-linguistic tasks [30], [31], [38]. Capitalizing on this achievement, researchers have explored the integration of skeleton and RGB modalities with transformer-based architectures for a range of vision tasks, including HAR. To offer a thorough understanding of the transformer-based multimodality fusion applied in HAR, a summary is given in Table 1, featuring three columns: *Fusion Method*, *Modalities Used*, and *Purpose*. The *Fusion Method* column details three popular fusion techniques using transformer architecture: early summation, early concatenation, and cross-attention. The *Modalities Used* column lists the modalities employed in the studies, while the *Purpose* column highlights the specific tasks for which fusion is executed.

Early summation is a simple yet effective approach to multimodal interaction that involves weighting token embeddings from multiple modalities and subsequently summing them at each token position before processing them through a transformer layer. For example, the authors in [35] utilized early summation to fuse three multimodality features (RGB images, optical flow images, and static skeleton) for group activity recognition. A benefit of using the early summation approach is its low computational difficulty, but the primary drawback lies in the need to manually assign weights for various input multimodal features [34].

As another early fusion method, early concatenation entails concatenating token embedding sequences from multiple modalities and inputting them into transformer layers. Studies in [36], [37], and [42] employed this fusion method with RGB and skeleton sequences for HAR. Early concatenation offers the benefit of relative simplicity compared to other methods. Nevertheless, the drawback of utilizing this fusion technique is the increased computational complexity due to the longer sequence resulting from concatenation [34].

Cross-attention is an efficient fusion method for multiple modalities, enabling each modality to attend to information from the others. This process is achieved by exchanging the key (K) and value (V) vectors of one modality with the query (Q) sequences of another modality within multiple stream transformer layers. Furthermore, this fusion method does not significantly increase computational complexity.

Yan et al. [39] employed cross-attention to fuse RGB images from different views for more robust action recognition. Similarly, Ijaz et al. [40] used cross-attention to integrate skeleton sequence data and acceleration data for action recognition in nursing. Additionally, Zhang et al. [41] applied cross-attention to fuse three modalities-RGB images, text, and audio-for effective facial expression recognition.

Our proposed SSRT represents the first transformer-based multimodal fusion approach to combine skeleton and RGB modalities for HAR using cross-attention. Diverging from prior studies that employed early concatenation for fusing high-level skeleton and RGB features [37], [42], our proposed method focuses on detecting fine-grained HOI and action recognition.

2) MULTISTAGE MULTIMODAL FUSION TECHNIQUES FOR HAR

Cheng et al. [51] employ a two-stage approach to fuse RGB and depth sequences. First, they introduce a novel method called Cross-Modality Interactive Module (CMIM) to enhance the sharing of high-level features between RGB and depth sequences. Subsequently, these features are fused using the late score fusion technique. Yan et al. [39] initially utilize transformer cross-attention to fuse RGB images from different views for more robust action recognition. Later, they use a global transformer encoder to fuse the high-level features derived from the transformer encoder, which processes two distinct views of RGB images. Yuean et al. [52] develop a human monitoring system that integrates PRF and PIR sensor data through sensor fusion. They employ three RNN models for PIR, PRF, and combined PRF-PIR data and implement decision-level fusion with HAP XAI to improve the interpretability of the results. Weiyao et al. [12] and Zhu et al. [13] fuse RGB and skeleton modalities in two stages of fusion for human action recognition. Weiyao et al. first uses the proposed Bilinear Pooling and Attention Network (BPAN) module to fuse the high-level features of RGB and skeleton modalities. The BPAN module is employed to learn potential semantic relationships between RGB and skeleton HAR baseline models. Later, similar to our Softmax layer late score fusion, Weiyao et al. fuse the probability scores from two pathways to obtain a final score prediction. Zhu et al. implement a novel two-stage feature fusion network to combine the knowledge of the RGB and skeleton modalities. First, they fuse skeleton and RGB features in the early stage using element-wise concatenation. Then, Zhu et al. use either GCN or LSTM architecture to fuse these features as late score fusion.

In our proposed method, we leverage the benefits of a two-stage fusion approach to effectively integrate RGB and skeleton modalities for human action recognition. By initially applying transformer cross-attention to capture the complex interactions between the modalities, we are able to learn more meaningful and complementary features. Then, through softmax layer late score fusion, we combine the probability scores of the individual modalities, allowing for a more accurate and robust final prediction. To the best of our knowledge, our method is the first to utilize this combination of transformer cross-attention and softmax layer late score fusion for human action recognition. This unique approach capitalizes on the advantages of both stages of fusion, ultimately leading to improved performance in comparison to previous methods.

III. METHODOLOGY

A. FEATURES PREPROCESSING

This study standardized the experimental datasets by randomly selecting 42 sequences of frames from each video to identify human actions. The frames were then resized to $480 \times 640 \times 3$. To obtain RGB features, the input pixel values of each frame were first rescaled to between -1 and 1 using equation (1) and were then passed through a pre-trained Resnet152 [33] model to generate 42×2048 RGB features. To obtain the skeleton modality, AlphaPose Pose Estimation [32] was utilized to generate 17 2D skeleton features from each input frame. These skeleton features were then normalized along the x and y coordinates utilizing equation (1) and flattened to obtain 42×34 skeleton features. Figure 4 shows the raw images in the first row and their associated skeleton keypoint representations in the second row.

In this study, we employed a specific preprocessing approach to create multimodal features using both RGB and skeleton data. First, we extracted skeleton features following the procedure described earlier. Next, we cropped frames to a size of $350 \times 350 \times 3$ using the x and y coordinates of the skeleton features. To determine the optimal cropping dimensions, we conducted a qualitative analysis, which is illustrated in Figure 3. The figure's columns represent three different cases, with the first row demonstrating the limitations of various cropping dimensions, while the second row

TABLE 1. Related works on transformer-based multimodal fusion for HAR.

Fusion Method		Modalities Used	Purpose
Early Summation	[35]	RGB Images ,Optical Flow Images, Static Skeleton	Group Action Recognition
Early Concatenation	[36]	RGB Images, Skeleton Sequences	HAR
	[37]	RGB Images , Skeleton Sequences	HAR
	[42]	RGB Images, Skeleton Sequences	HAR
Cross-Attention	[39]	RGB Images, RGB Images	Action Recognition
	[40]	Skeleton Sequences, Acceleration Data	Nursing Action Recognition
	[41]	RGB Images, Text, Audio	Facial Expression Recognition
	SSRT(Ours)	**RGB Images, Skeleton Sequences**	**Fine-Grained HOI**

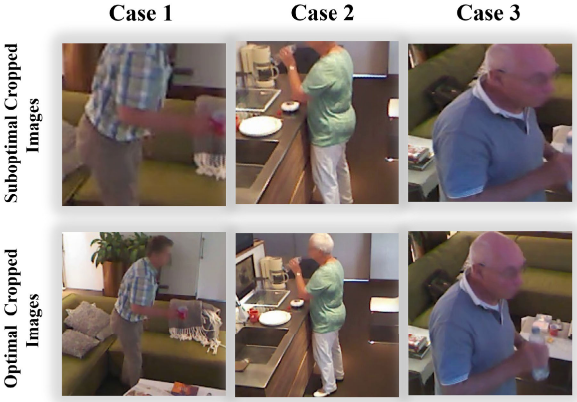


FIGURE 3. Suboptimal crop selection vs. Optimal crop selection.

showcases the optimal results achieved with the proposed cropping dimension in each scenario.

Initially, we attempted cropping with a size of 144 × 144 × 3, as depicted in the *Case 1* of the first row in Figure 3. However, this cropping method proved inadequate in most scenarios, as it failed to fully capture the subject. We then increased the crop dimensions to 224 × 224 × 3 but found this size to be insufficient for cases where the subject is standing, as illustrated in *Case 2* of the first row in Figure 3.

Subsequently, we further increased the crop dimensions to 275 × 275 × 3 and observed that although this size produced optimal results in most scenarios, it struggled to accurately capture human-object interactions, particularly when the person was too close to the camera. In *Case 3* of the first row, it is evident that the bottle was cropped during the process. Ultimately, we settled on a dimension of 350 × 350 × 3, which provided optimal results, as shown in the second row of Figure 3.

To obtain RGB features, we followed the same process as previously outlined, resulting in 42×2048 RGB features from the 42 input frames. By implementing these preprocessing steps, the proposed SSRT can strike a balance between the most relevant and contextual information by cropping the frame to an ideal size.

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}} \times 2 - 1 \tag{1}$$

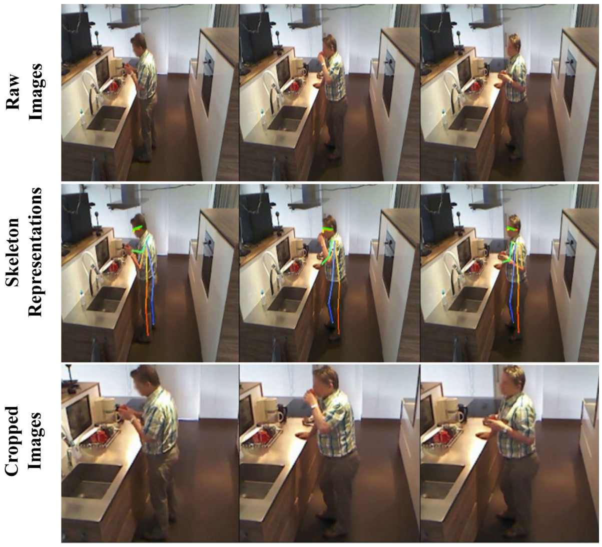


FIGURE 4. Optimizing RGB frame size through skeleton feature-based cropping.

TABLE 2. Different versions of SSRT, skeleton encoder, and RGB encoder.

Model	# H	# d_{model}	# d_{ff}
X1	96	96	192
X2	128	128	256
X3	256	256	512

B. ACTION RECOGNITION MODEL

1) SINGLE MODALITY ACTION RECOGNITION

Section (III-A) outlined the preprocessing procedure employed for classifying HAR using a transformer encoder with either skeleton or RGB modality. Following this, the preprocessed features are channeled through a positional encoding layer. In accordance with the transformer architecture [29], each input feature dimension must be a multiple of the number of heads (H) utilized in the multi-head self-attention (*MSA*) layer. As a result, the positionally encoded input features are projected onto d_{model} dimensions, corresponding to the number of H incorporated in the *MSA* layer, as noted in Table 2. Therefore, the input X for the transformer encoder possesses dimensions of 42 × d_{model}.

The operating mechanism of the transformer encoder is elucidated in [29]. The output from the transformer encoder is flattened into a one-dimensional vector, which is then processed by a Softmax layer, producing a probability output for each human action class. In this study, the terms *RGB encoder* and *skeleton encoder* are used when the RGB modality and skeleton modality, respectively, are leveraged for classifying HAR with the transformer encoder.

2) SSRT: PROPOSED MULTIMODALITY ACTION RECOGNITION

SSRT integrates RGB and skeleton sequences using two parallel pathways: the skeleton modality pathway and the RGB modality pathway, as depicted in Figure 2.

To efficiently extract high-level features from multimodal inputs, feature preprocessing is performed for each modality feature prior to high-level feature extraction. Section III-A offers an in-depth description of feature preprocessing for multimodality fusion. Providing positional information to the preprocessed features from each modality is crucial before inputting them into the transformer encoder. Without this information, the transformer encoder might interpret the features as a bag of features, potentially reducing its effectiveness. Although a Positional Encoding Layer is commonly employed to tackle this issue, it may be unsuitable for time-series tasks like HAR due to its disregard for temporal dependencies between input sequences.

In order to effectively capture the sequence order and timing of input features, this study proposes the use of an LSTM layer instead of a positional encoder. LSTMs process inputs element by element, adeptly capturing the temporal dependencies inherent in feature sequences. This characteristic is crucial for time-series tasks, such as HAR, where recognizing temporal dependencies is vital. By implementing an LSTM layer, the effectiveness of abstract feature extraction can be enhanced compared to traditional positional encoding methods. To maintain the dimensionality of each input sequence instance as a factor of the number of heads H used in the multi-head attention layer, the hidden dimension of the LSTM layer is set to d_{model}. As a result, the LSTM layer generates an output with a shape of $42 \times d_{model}$ for both pathways, which is subsequently directed to the appropriate skeleton or RGB transformer encoder based on the input modality.

Subsequently, the transformer encoder architecture is employed to obtain high-level features from both the skeleton and RGB pathways, as described in Section III-B1. After obtaining the abstract features from the RGB encoder (HLF_{RGB}) and the skeleton encoder (HLF_{SKL}), the fusion process begins through the skeleton cross transformer and the RGB cross transformer. These cross transformers merge features from two modalities using a cross-attention fusion mechanism, which is elaborated upon in Section II-B1. Importantly, the skeleton/RGB cross transformer uses multi-head cross-attention (MCA) layers, distinct from the MSA layers found in the skeleton/RGB encoder.

The skeleton cross transformer incorporates an MCA layer that creates a Q_s vector for each attention head H by multiplying HLF_{SKL} with a trainable weight matrix WQ_s. Simultaneously, the MCA layer generates two contextual vectors, K_r and V_r, by multiplying HLF_{RGB} with trainable weight matrices WK_r and WV_r, respectively. The attention score (A_s) for each attention head H in the skeleton cross transformer is determined using equation (2), where d_k denotes the dimension of each attention head. To acquire the final multi-head attention scores (MCA_s), the attention scores from all attention heads are concatenated and then multiplied by a trainable weight vector W_s, with dimensions $(d_k \times H) \times d_{model}$, as illustrated in equation (3).

$$A_s = Softmax\left(\frac{Q_s K_r^T}{\sqrt{d_k}}\right) V_r \tag{2}$$

$$MCA_s = Softmax\Bigl([A_s1; A_s2; ..; A_sH]\Bigr))W_S \tag{3}$$

In a similar manner, the RGB cross transformer produces three vectors (Q_r, K_s, and V_s). However, the query input vectors are obtained using HLF_{RGB}, while the contextual input vectors are derived from HLF_{SKL}. Then, equations (4) and (5) are used to compute the final multi-head attention score (MCA_r) for the RGB cross transformer, following an approach analogous to that of the skeleton cross transformer.

$$A_r = Softmax\left(\frac{Q_r K_s^T}{\sqrt{d_k}}\right) V_s \tag{4}$$

$$MCA_r = Softmax\Bigl([A_r1; A_r2; ..; A_rH]\Bigr) W_r \tag{5}$$

The following steps in the skeleton/RGB cross transformers conform to the same process as the skeleton/RGB encoder, as detailed in Section III-B1. The outputs from the skeleton cross transformer (O_{SCT}) and the RGB cross transformer (O_{RCT}) are combined with HLF_{SKL} and HLF_{RGB}, respectively, and then normalized. Afterward, the normalized outputs from each modality pathway are flattened and fed into the Softmax layer, resulting in the probability of each action class for the skeleton modality (P_{SKL}) and the RGB modality (P_{RGB}), as shown in equations (6) and (7). Ultimately, the SML late score fusion is carried out by adding P_{SKL} and P_{RGB} to obtain the final probability of each action class (P_{FINAL}) for HAR, as portrayed in equation (8).

$$P_{SKL} = Softmax\Bigl(Flatten(Norm(HLF_{SKL} + O_{SKL}))\Bigr) \tag{6}$$

$$P_{RGB} = Softmax\Bigl(Flatten(Norm(HLF_{RGB} + O_{RGB}))\Bigr) \tag{7}$$

$$P_{FINAL} = ADD\Bigl(P_{SKL}, P_{RGB}\Bigr) \tag{8}$$

3) TRANSFORMER MODEL ARCHITECTURE

This research introduces three distinct transformer architecture versions (*X1, X2, and X3*) for the skeleton encoder, the RGB encoder, and SSRT, as listed in Table 2. In this Table, d_{ff} represents the dimension of the initial layer in the feed-forward network inside the transformer encoder, as described in [29]. Diverging from traditional transformer architectures that use a low number of H and a higher number of d_{model} [22], [29], this study recommends making the number of H in a multi-head attention layer the same as d_{model}. This design enables each model to effectively harness a high number of H without incurring excessive computational demands.

IV. IMPLEMENTATION DETAIL

A. DATASETS

This research employs two datasets for the HAR task. Table 3 compares the commonly used vision-based datasets for this task. Among them, the Toyota Smarthome dataset stands out due to its fine-grained HOI and context-free nature, featuring elderly people performing actions without adhering to a specific script. Furthermore, this dataset exemplifies a real-world situation, presenting a distinct set of challenges, including high intra-class variation, significant class imbalance, and activities exhibiting similar motion patterns and considerable duration disparities [14]. Consequently, the Toyota Smarthome dataset serves as the primary dataset for this study.

As a secondary dataset, ETRI-Activity3D is utilized. This dataset also features elderly subjects performing context-free actions, making it an appropriate choice. Additionally, four general actions are shared between the Toyota Smarthome and ETRI-Activity3D datasets, making ETRI-Activity3D well-suited for the cross-dataset HAR task.

This study enhances dataset quality by filtering out samples in which the interacting object is obscured by the subject or another object.

1) FINE-GRAINED HUMAN-OBJECT INTERACTION

The Toyota Smarthome dataset consists of three sets of fine-grained HOI action classes, which fall under the coarse action labels of *Drink*, *Eat*, and *Pour*. It is crucial to mention that this research omits the fine-grained actions (*Clean dishes*, *Clean up*, *Cut*), (*Pour grains*, *Pour water*), and (*Boil water*, *Insert tea bag*), corresponding to the coarse labels *Cook*, *Make coffee*, *Make tea*, respectively. This exclusion is due to the fact that these fine-grained actions represent composite actions at a fine-grained level, rather than fine-grained HOI.

This study also excluded the fine-grained HOI action class within the *Eat* and *Pour* categories. The *Eat* category is not included because, as depicted in Figure 5, the fine-grained HOI *Eat at the table* involves eating at a table, while *Eat snack* features a few instances of eating while sitting. The posture differences can be detected independently using the skeleton modality. Similarly, as illustrated in Figure 5, the *Pour* category is not considered because the *Pour from kettle* action involves pouring water from a kettle, whereas the corresponding fine-grained HOI action *Pour from bottle* also includes additional actions such as opening and closing the bottle. The skeleton modality can identify these extra actions without requiring RGB features. This study only employs fine-grained HOI from the *Drink* category. As shown in Figure 1, the fine-grained HOI within the *Drink* category share similar motions, with the only variation being the subject's interaction with various drinking objects.

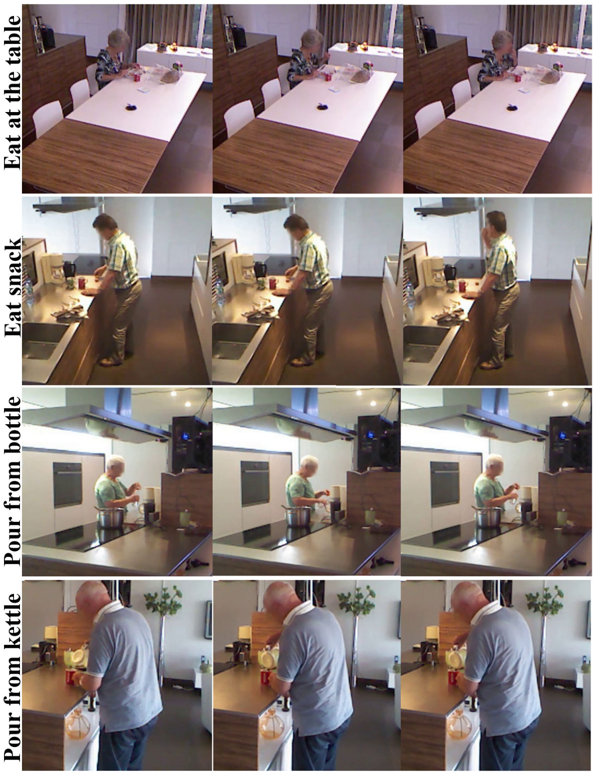


FIGURE 5. Unused action classes in fine-grained human-object interactions from Toyota Smarthome dataset.

This research used 196 video samples from each class within fine-grained HOI and stratified them into training, validation, and testing sets at a 60:20:20 ratio, ensuring unbiased results. This approach provides representative samples of fine-grained HOI in each subset, mitigating the risk of overfitting to any particular class, unlike the methods used in [14], [15], and [16].

2) GENERAL ACTIONS AND CROSS-DATASET ACTIONS

This research extended the evaluation of the proposed SSRT model to two other action recognition tasks. First, the model's ability to generalize beyond the fine-grained HOI class was tested in the General Actions class. Secondly, the model was assessed on the Cross-dataset Action class to evaluate its performance with unseen videos from a different dataset. The objective of this evaluation was to establish if the SSRT model could achieve high accuracy on videos that differ significantly from the training set, demonstrating its ability to generalize to new and diverse contexts.

In this research, an extensive search was conducted of both the Toyota Smarthome and ETRI-Activity3D datasets to identify suitable action classes for general action and cross-dataset HAR tasks. After analyzing the datasets, four classes were identified that exhibited comparable motion and human-object interaction. These classes are *Drink From Cup*, *Readbook*, *Uselaptop*, and *Usetelephone* from Toyota Smarthome, which corresponds to *Drinking water*, *Reading*

TABLE 3. Comparison of vision-based datasets for human action recognition.

Dataset		#Classes	#Samples	#Subjects	Type	Context	Scripted Action	Fine-grained HOI
HMDB	[25]	51	6766	-	Youtube	biased	-	No
Kinetics-600	[23]	600	480000	-	Youtube	biased	-	No
MSRDailyActivity3D	[26]	16	320	10	ADL	free	High	No
CAD-120	[24]	10+10	120	4	ADL	free	High	No
NTU-RGB+D 120	[27]	120	114480	106	ADL	free	High	No
ETRI-Activity3D	[28]	55	112620	100	ADL	free	Low	No
Toyota Smarthome	[14]	31	16129	18	ADL	free	Low	Yes

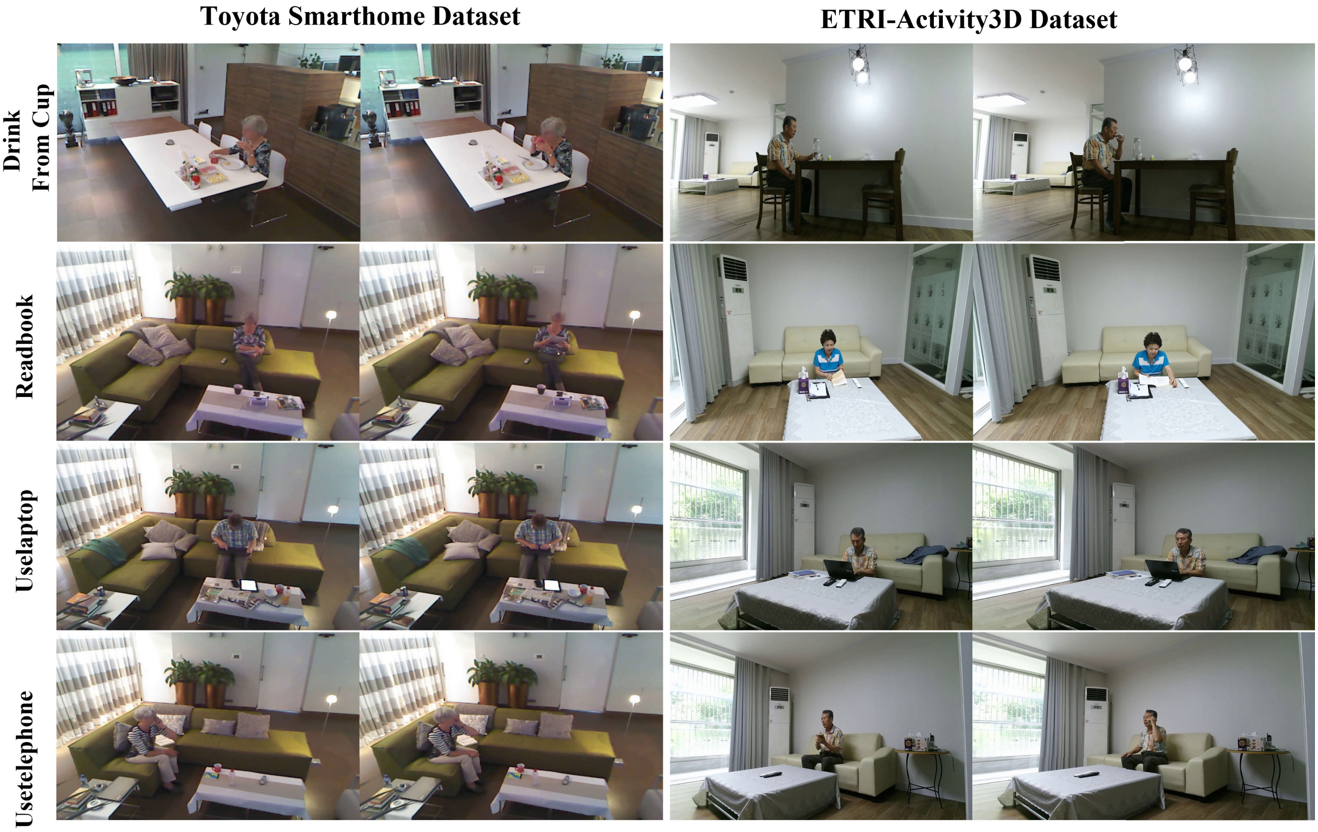


FIGURE 6. Comparison of common actions in Toyota Smarthome and ETRI-Activity3D datasets.

a book, *Using a computer*, and *Talking on the phone*, respectively, from ETRI-Activity3D. Throughout the study, the action class names from the Toyota Smarthome dataset were used for both HAR tasks for ease of analysis. In Figure 6, frames depicting the beginning and middle stages of each action category in the Toyota Smarthome dataset are shown alongside the corresponding class from the ETRI-Activity3D dataset.

This research underscores the significant challenges associated with cross-dataset evaluation, which is more intricate than both cross-view and cross-subject evaluations [14], [27], [28]. The complexity arises due to the divergent perspectives between the Etri-Activity3D dataset, captured from a robotics viewpoint, and the Toyota Smarthome dataset, recorded from a surveillance viewpoint. This variance heightens the difficulty of the cross-view challenge. Furthermore, cross-dataset evaluation proves to be more convoluted than cross-subject testing, as the individuals performing actions differ, and other RGB factors such as illumination, colors, and video quality further compound the evaluation process.

The effectiveness of SSRT on the general HAR task was assessed using the training, validation, and testing datasets

TABLE 4. Traning and regularization hyperparameters.

Traning	
Epochs	[200,225,250,275,300]
Batch Size	512
Optimizer	AdamW
Regularization	
Weight Decay	[0.0001, 0.001]
Optimizer learning rate	0.00001
Dropout	[0.2,0.3,0.5]

from Toyota Smarthome. For selecting the training, validation, and testing datasets, sample videos of selected action classes from the Toyota Smarthome dataset were split into those datasets at a ratio of 60:20:20. Similarly, to evaluate the performance of SSRT on the cross-dataset HAR action recognition task, this research utilized the same training and validation dataset splits as the General HAR task, and tested the model on cross-dataset actions from the ETRI-Activity3D dataset. From ETRI-Activity3D, 99 samples were selected from the action class *Drink From Cup*, 125 samples were selected from *Readbook*, 95 samples from *Uselaptop*, and 123 samples from *Usetelephone*.

This dataset implementation enables benchmarking of the proposed method, SSRT, in a manner akin to that presented by An et al. [55]. An et al. have proposed two experimental protocols; first, they split their dataset into Setting 1 (S1) and Setting 2 (S2). The S1 dataset is obtained from a train-validation-test split, while S2 focuses on cross-subject evaluation. Next, they also divide their actions into two action protocols, P1 and P2. Similar to the S1 setting, our fine-grained and general actions involve a train-test random split for comparison. Likewise, our approach mirrors the S2 setting with a cross-dataset HAR task. However, the cross-dataset setting is considerably more challenging than the S2 protocol, as it encompasses not only cross-subject challenges but also other difficulties such as varying backgrounds, cross-views, and so on.

B. COMPARISION WITH OTHER HAR METHODS

1) COMPARISION WITH SINGLE-MODALITY HAR MODEL

In this study, to compare SSRT with single-modality HAR, LSTM and Transformer encoder HAR models were chosen. Although LSTM is not a state-of-the-art HAR model, it was selected alongside the Transformer encoder due to their critical roles in the SSRT framework. Consequently, this particular selection of single-modality-based HAR models also benefits the ablation study.

As discussed in Section II-A, a Transformer encoder resembling the AcT model [44] was chosen as the primary single-modality HAR model for comparison with SSRT performance. The main reasons for this choice are:

1) **Transformer is well-suited for both RGB and skeleton modality-based HAR models**: As discussed in Section II-A, the transformer architecture demonstrates state-of-the-art performance for both RGB and skeleton modality-based HAR. This makes the transformer architecture one of the few architectures suitable for both the skeleton and RGB-based modalities in HAR tasks. For example, Graph Convolution Network-based models such as STGCN [21] and 2s-AGCN [56] exhibit state-of-the-art performance for skeleton-based modality but fail when RGB modality is used. Similarly, 3D CNN-based models like I3D [4] perform exceptionally well with RGB modality but fall short when used with skeleton modality.
2) **Ensuring a fair comparison of the skeleton, RGB, and multimodal approaches**: This study proposes the same three Transformer architectures for both single-modality-based HAR models and SSRT, as discussed in Section III-B3. In SSRT, a similar Transformer architecture to that of the single-modality HAR model is first used in the RGB/Skeleton encoder for higher feature extraction, and later, the same architecture is employed in the fusion stage using the cross-transformer. This consistent use of comparable Transformer architectures allows for a fair comparison between single-modality-based HAR and SSRT.

2) COMAPRISION WITH OTHER METHODS OF MULTIMODALITY FUSION HAR MODEL

In this study, the SSRT method is evaluated alongside three other prominent fusion methods: LSTM Late Score, Transformer Early Concatenation, and Transformer Late Score. The baseline models for the late score fusion methods are derived from the single-modality HAR models used in this research. The LSTM and Transformer late score fusion models are trained independently on each modality, with the predicted probability scores from each modality combined to predict the action classes using the late score fusion approach. Transformer Early Concatenation employs a Transformer encoder to extract high-level features for each modality. These features are subsequently concatenated and projected onto another Transformer encoder to generate the final prediction.

C. EXPERIMENTAL SETTINGS

The experiments were conducted using Tensorflow platform release 2.11 on a personal computer that has an Intel i7-10700K processor, 48 GB of RAM, and an NVIDIA RTX 3090 GPU that has 24 GB of VRAM. Optuna [46] was employed to optimize the training process because it is a widely recognized hyperparameter optimization tool for machine learning models. The objective function, hyperparameter search space, and optimization algorithm were defined using Optuna. The hyperparameters for training and regularization are detailed in Table 4, where some hyperparameters possess fixed values, but others have a range of values. Optuna performs a random search during training to optimize the hyperparameters based on the listed values and then determines the optimal hyperparameter. Both the

TABLE 5. Overall experimental results.

Model	Fine-grained HOI Actions				General Actions				Cross-Dataset Actions			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
Using Skeleton Modality												
LSTM	63.96	64.04	63.96	63.76	86.53	86.59	86.53	86.53	39.36	40.38	39.36	38.97
Skeleton Encoder	71.17	71.14	71.17	70.45	86.19	86.19	86.07	86.06	43.67	51.54	43.67	37.58
Using RGB Modality												
LSTM	64.86	64.88	64.86	65.85	53.19	43.97	53.19	46.92	33.84	34.16	33.84	29.44
RGB Encoder	74.77	75.38	74.77	74.91	79.46	79.56	79.46	79.42	40.49	52.4	40.49	39.18
Using Skeleton and RGB Modalities												
LSTM (Late Fusion)	45.04	46.83	45.04	45.26	26.59	26.58	26.59	26.57	28.5	28.01	28.5	27.78
Transformer (Early Concat)	78.37	80.12	78.37	78.03	88.88	88.9	88.88	88.83	48.19	42.23	48.19	42.46
Transformer (Late Fusion)	58.59	60.37	58.59	58.76	83.5	83.56	83.5	83.4	36.65	36.79	36.65	34.36
Ours SSRT	**84.69**	**85.11**	**84.69**	**84.77**	**92.92**	**93.01**	**92.92**	**92.95**	**54.75**	**54.98**	**54.75**	**54.65**

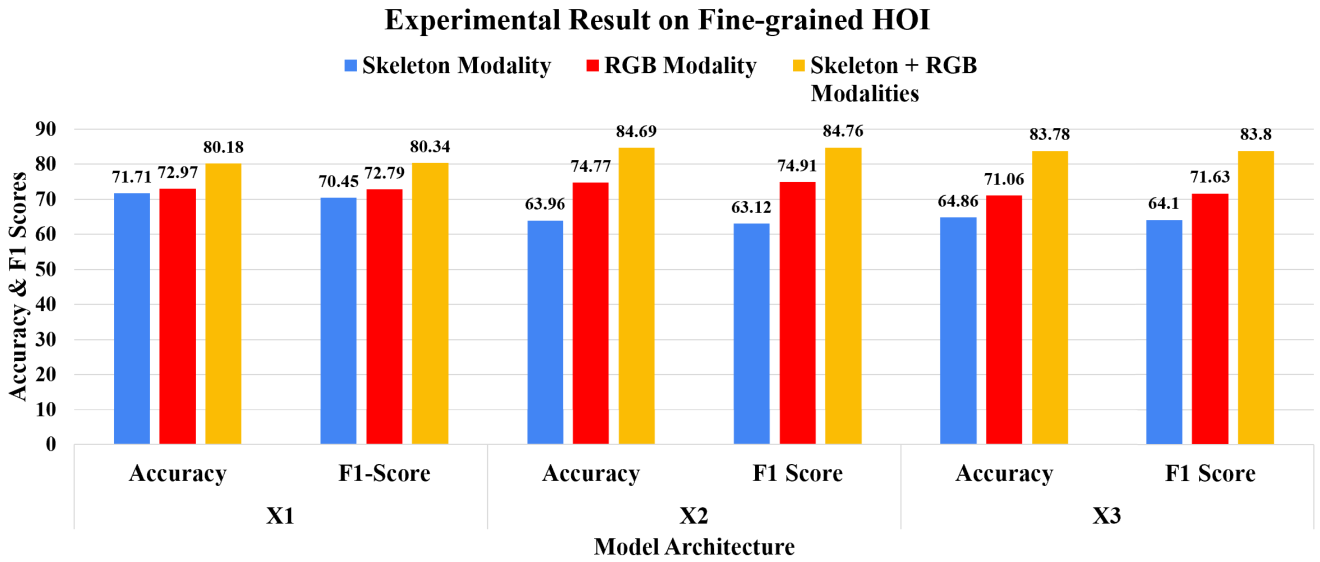


FIGURE 7. Experimental results on fine-grained HOI using different modality features.

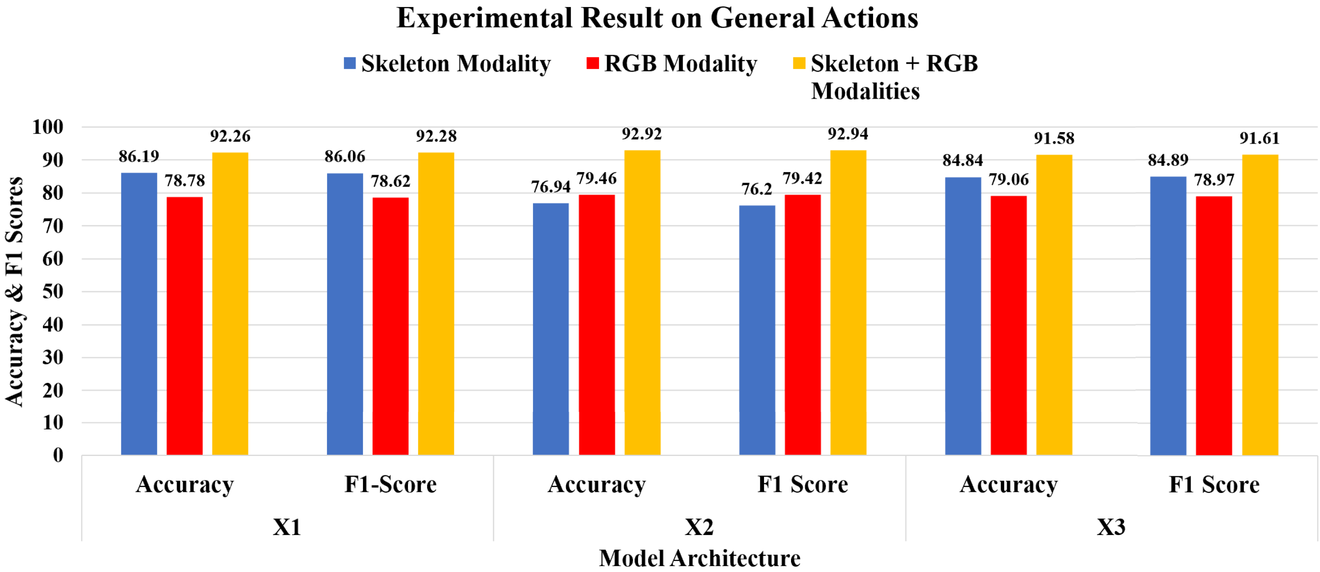


FIGURE 8. Experimental results on general actions using different modality features.

single-modality based transformer encoder and the proposed SSRT were trained using categorical cross-entropy loss. The AdamW [48] optimizer, which had a learning rate of 0.00001, was chosen for network optimization, as shown in Table 4.

V. EXPERIMENTAL RESULT

The comprehensive experimental results obtained from this research are presented in Table 5. This section evaluates SSRT in comparison to both single-modality HAR models and multimodality HAR models. It is organized into two subsections. The first subsection, V-A, examines the extensive experimental results depicted in Table 5. Concurrently, the second subsection, V-B, assesses the three proposed Transformer architectures for the skeleton encoder, RGB encoder, and SSRT within the scope of all three HAR tasks.

A. COMPREHENSIVE EXPERIMENTAL OUTCOMES

Table 5 highlights that when utilizing only the skeleton modality, the transformer encoder achieves superior accuracy in fine-grained HOI action recognition and cross-dataset action recognition tasks. In addition, the transformer encoder yields comparable results to the LSTM model in general HAR

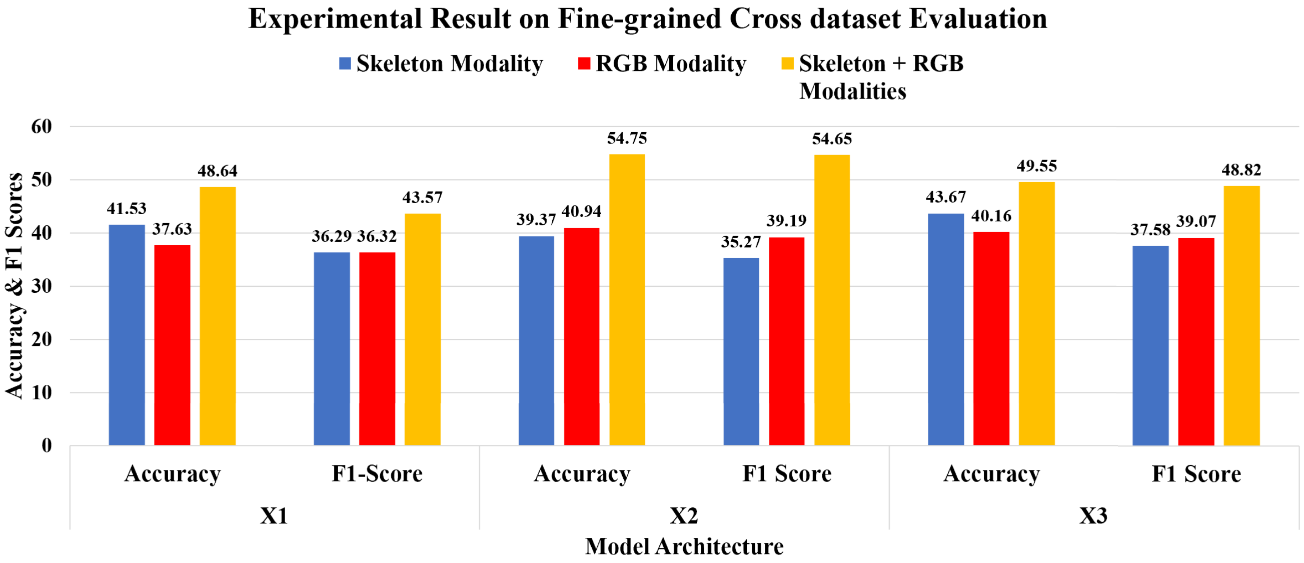


FIGURE 9. Experimental results on cross-dataset actions using different modality features.

tasks. For the fine-grained HOI action recognition task, the transformer encoder displays a substantial improvement with a 7.21% increase in accuracy and a 6.69% enhancement in F1 score compared to the LSTM model. Furthermore, the transformer encoder demonstrates a significant advantage over the LSTM-based model when solely relying on the RGB modality in all action recognition tasks, achieving up to a 32.5% improvement in F1 score for the general HAR task. These observations suggest that the transformer encoder, which employs the multi-head attention mechanism, is more effective at capturing human actions than the traditionally used LSTM model.

Interestingly, when compared to the best-performing single-modality HAR model, the RGB-based modality outperforms the skeleton modality in fine-grained HOI tasks, whereas the skeleton modality achieves better results in general and cross-action HAR tasks. The reasons behind this are:

1) **Lack of spatial information in skeleton modality**: Unlike the RGB modality, the skeleton modality does not encode spatial features, rendering it unable to comprehend human-object interactions within fine-grained HOI tasks. As a result, the RGB modality surpasses the skeleton modality in fine-grained HOI action classes.
2) **Skeleton Modality Captures Human Motions Better than RGB Modality**: The skeleton modality is better at capturing human movement since it encodes various joint and limb movements. Consequently, the skeleton modality outperforms the RGB modality in general HAR tasks.
3) **Skeleton modality is more robust than RGB modality**: The RGB modality faces challenges due to environmental diversity, such as changes in illumination. In contrast, the skeleton modality is robust to variations in clothing textures, and illumination, and is also scale-invariant. Thus, in cross-dataset HAR tasks, the RGB modality is more affected than the skeleton modality.

As illustrated in Table 5, the LSTM late score fusion model exhibits the poorest performance in all HAR tasks. Moreover, it is particularly noteworthy that both late score fusion models (LSTM late score fusion and transformer late score fusion) experience a decrease in all evaluation metrics when compared to their corresponding single-modality HAR methods. In contrast, the state-of-the-art multimodal fusion method, Transformer Early Concatenation, enhanced both accuracy and F1 score when compared to single-modality methods.

Our proposed method, SSRT, achieved the best results in all evaluation metrics for all HAR tasks. SSRT outperformed Transformer Concatenation in accuracy by 6.32% in the fine-grained HOI action recognition task, 4.04% in the general action recognition task, and 6.56% in the cross-dataset action recognition task. Similarly, SSRT surpassed Transformer Concatenation in F1 score by 6.74% in the fine-grained HOI action recognition task, 4.12% in the general action recognition task, and 12.19% in the cross-dataset action recognition task. SSRT outperformed the Transformer Concatenation multimodal fusion primarily because it utilizes both LSTM and transformer encoder to extract higher temporal features from both modalities. Furthermore, SSRT employs two levels of fusion stages compared to the single-stage multimodal fusion in Transformer Early Concatenation. This study examines the multimodal fusion in greater depth in Section VI-B.

From these experiments, we can observe that compared to general HAR tasks, SSRT shows significant improvement in fine-grained HOI and cross-dataset actions. The reasons for this are:

1) **Performance analysis of SSRT on fine-grained HOI task**: Fine-grained HOI actions are complex in nature, as they require an understanding of various human-object interactions that share similar human actions. Capturing fine-grained HOI actions using a single modality is challenging since the skeleton modality lacks spatial understanding, and the RGB modality is not as effective in capturing human actions. Furthermore, the diverse nature of these two modalities makes it difficult to fuse them using simple fusion methods (such as late score fusion). By employing SSRT, we were able to complement the RGB and skeleton modalities, resulting in significant improvement in the fine-grained HAR task.
2) **Performance analysis of SSRT on general HAR task**: General action classes do not contain complex actions. As these actions do not involve fine-grained HOI actions, achieving good accuracy in these action classes is possible using the skeleton modality alone. Although the improvement observed in this HAR task with SSRT is not as substantial as in other tasks, it is still remarkable.
3) **Performance analysis of SSRT on cross-dataset HAR task**: As mentioned in Section IV-A2, both the skeleton and RGB modalities face challenges in this task. The efficient integration of the skeleton and RGB modalities through SSRT demonstrates a significant improvement in overall accuracy compared to alternative HAR methods. Despite SSRT's substantial outperformance of other HAR models in this task, the overall accuracy remains relatively low.

Additionally, the ablation study (Section VI-A) examines the performance of a single modality HAR model with SSRT, providing an in-depth analysis by comparing the accuracy of each action class.

B. TRANSFORMER ARCHITECTURE COMPARISION

Figure 7, Figure 8, and Figure 9 provide a comprehensive comparison of the performance of three distinct transformer architectures employed in the skeleton encoder, RGB encoder, and SSRT. These performances are represented as bar plots for the fine-grained HOI task, general HAR task, and cross-dataset HAR task, respectively. From these bar plots, it is clear that *X2* is the optimal transformer architecture for both the RGB encoder and SSRT in all three HAR tasks. In contrast, for the skeleton encoder, *X1* performs best in fine-grained HOI and general actions, while *X3* excels in cross-dataset actions. These results emphasize that although the *X2* architecture can be recommended for every HAR task involving the SSRT and RGB encoder, the same is not true for the skeleton encoder, as no single transformer architecture consistently outperforms others across all HAR tasks.

In the fine-grained HOI task, when using a single-modality feature, the RGB modality surpassed the skeleton modality, showing a 3.06% increase in accuracy and a 4.46% improvement in the F1 score. The best-performing transformer architecture variant of SSRT significantly outperformed the corresponding version of the RGB encoder, with an impressive 9.92% enhancement in accuracy and a 9.86% boost in the F1 score. In the general action task, when utilizing the skeleton modality, the *X1* version of the transformer encoder achieved better outcomes compared to the *X2* version. The highest-performing transformer architecture variant of SSRT considerably outperformed the corresponding version of the skeleton encoder, with a notable 6.86% enhancement in accuracy and a 9.86% boost in the F1 score. Finally, for the cross-dataset actions task, the optimal transformer architecture version of the skeleton encoder outperformed the optimal version of the RGB encoder with an accuracy difference of 4.3%, although both optimal versions of the skeleton and RGB encoder exhibited comparable F1 scores. The top-performing transformer architecture variant of SSRT exceeded the best-performing skeleton encoder model in accuracy by 11.08% and outperformed the best-performing RGB encoder with a 15.06% improvement in the F1 score. Moreover, from these bar plots, it can be observed that SSRT consistently outperformed both the skeleton and RGB modalities across all transformer architectures.

From these results, a key insight can be drawn: the effectiveness of each modality and transformer architecture variant is highly dependent on the specific HAR task. In the fine-grained HOI task, the RGB modality outperforms the skeleton modality, whereas, in the general action task, certain skeleton encoder variants deliver superior results. Additionally, the optimal transformer architecture for the skeleton encoder varies depending on the task, reinforcing the idea that a one-size-fits-all approach is not ideal. Moreover, the SSRT consistently outperforms both the skeleton and RGB modalities across all transformer architectures, indicating its potential as a robust and versatile solution for various HAR tasks.

VI. ABLATION STUDY

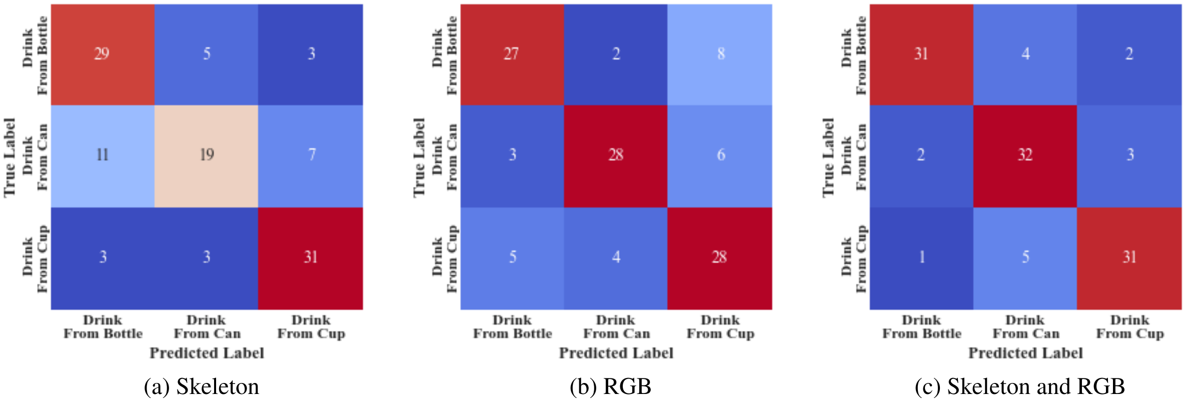
A. SKELETON VS. RGB VS. MULTIMODAL MODALITIES

In this section, a comparison of performance from action recognition models employing skeleton modality, RGB modality, and multimodality is conducted across all three HAR tasks. To achieve this, optimal transformer architecture versions of the skeleton encoder, the RGB encoder, and SSRT were selected for evaluation.

1) FINE-GRAINED HOI

Figure 10 offers a performance comparison of the skeleton modality, the RGB modality, and multimodality in fine-grained HOI recognition. In order to test the performance of the various modalities, we selected 37 samples from each fine-grained HOI class.

The confusion matrix depicted in Figure 10a showcases the results of fine-grained HOI classification performed by



(a) Skeleton

(b) RGB

(c) Skeleton and RGB

FIGURE 10. **Comparative analysis of confusion matrices for fine-grained HOI classification: (a). Skeleton, (b). RGB, and (c). Multimodal (Skeleton + RGB) approaches.**

the skeleton encoder. This figure demonstrates that the *Drink From Cup action* attained the highest classification accuracy (83.78%) by identifying 31 samples. Similarly, the skeleton encoder accurately classified 29 samples of the *Drink From Bottle action* attaining an accuracy of 78.38%. Conversely, the *Drink From Can* class exhibited a diminished classification performance with only 19 samples classified accurately, resulting in an accuracy of 51.35%. From the confusion matrix results, we observe a marked variation in accuracy across different fine-grained HOI classes. This can be mainly attributed to the skeleton modality's inability to encode fine-grained HOI spatial information. Consequently, the high accuracy achieved by the skeleton modality may be unreliable, possibly resulting from coincidental matches rather than a genuine understanding of the underlying patterns.

Next, as observed from the confusion matrix in Figure 10b, the RGB encoder classified 28 samples of *Drink From Can* and *Drink From Cup* with an accuracy of 75.67%, while misclassifying only one additional sample in the *Drink From Bottle* class. This observation highlights that, in contrast to the skeleton modality, the RGB modality did not display a biased accuracy towards any specific action class. The results indicate that RGB modality-based HAR models are more capable of understanding complex HOI patterns compared to skeleton modality. Therefore, if only a single modality must be used, this study suggests employing RGB modality-based HAR models for fine-grained HOI actions.

The performance of SSRT is displayed in the confusion matrix in Figure 10c. SSRT classified 32 samples in the *Drink From Can* action class with an accuracy of 86.48%. It misclassified only one fewer sample in the *Drink from Bottle* and *Drink from Cup* action classes compared to the *Drink from Can* action class. Furthermore, when compared to overall accuracy, Table 5 shows that the RGB modality improved accuracy by 3.6% compared to skeleton modality, and multimodality improved accuracy by 9.92% compared to RGB modality in this human activity recognition task. Similar to the RGB modality, the multimodal setting does not display biased performance towards any specific class.

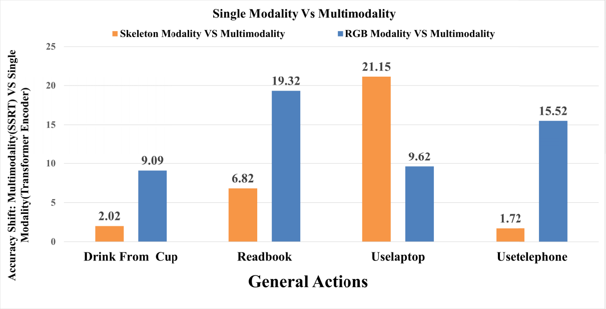


FIGURE 11. **Accuracy shift: Multimodality (SSRT) vs. Single modality (Transformer Encoder) for general actions recognition.**

However, the SSRT modality demonstrates improved accuracy for each action. This is due to the fact that while the RGB modality can comprehend fine-grained HOI, it struggles to encode human movement effectively. Consequently, SSRT, which complements both RGB and skeleton modalities, outperforms the RGB modality in the fine-grained HOI task.

2) GENERAL ACTIONS AND CROSS-DATASET ACTIONS

This section highlights the change in accuracy for each action class when employing SSRT as opposed to single-modality transformer encoders (skeleton and RGB) in Figures 11 and 12 for general actions and cross-dataset actions, respectively. The x-axis lists the action class, while the y-axis displays the change in accuracy between single-modality and multimodalities based HAR models. The orange bar plot illustrates the difference in accuracy when utilizing the skeleton encoder in comparison to SSRT, and the blue bar plot demonstrates the difference in accuracy when using the RGB encoder as opposed to SSRT.

In the general actions recognition task, integrating the skeleton and RGB modalities led to the most significant performance enhancement for the *Uselaptop* class, with a 21.15% increase in accuracy compared to employing only the skeleton modality. Conversely, the *Drink From Cup* class experienced the smallest performance improvement, with an

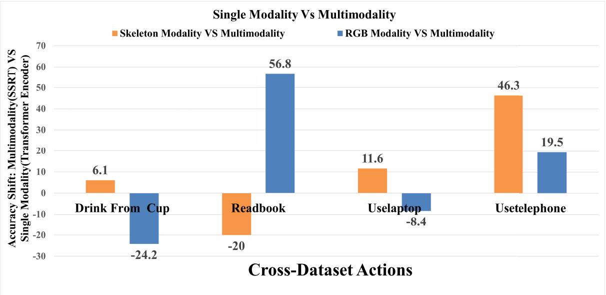


FIGURE 12. Accuracy shift: Multimodality (SSRT) vs. Single modality (Transformer Encoder) in cross-dataset actions recognition.

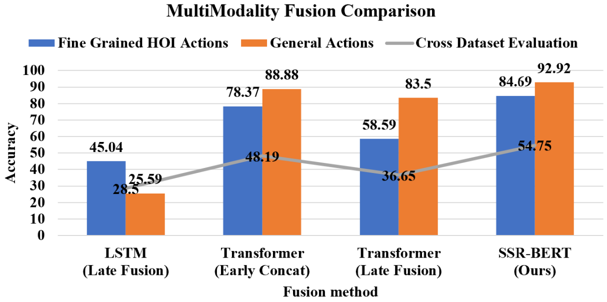


FIGURE 13. Comparison of multimodality fusion methods for HAR.

accuracy increase of just 1.72% when adopting multimodality instead of skeleton modality. When examining the change in accuracy between multimodality and RGB modality, the *Readbook* action showcased the most notable performance improvement, with SSRT boosting accuracy by 19.32% compared to the RGB encoder. In contrast, the *Drink From Cup* class displayed the smallest improvement when using multimodality in comparison to the RGB modality. It can be observed that for general actions tasks, in three out of four actions, SSRT demonstrated greater improvement when compared to using RGB modality alone. This is because these actions do not pose challenges for the skeleton modality, as they can also be classified without utilizing spatial information.

In cross-dataset action recognition, the *Readbook* class experienced the most substantial performance enhancement, with a 56.8% increase in accuracy when employing multimodality as opposed to utilizing only the RGB modality. Nevertheless, a 20% decline in accuracy was noted for the same class when adopting multimodality over skeleton modality. The *Drink From Cup* class exhibited the most substantial negative impact on performance when using multimodality instead of RGB modality, resulting in a considerable decrease in accuracy of 24.24%. In contrast, the multimodality showed better performance in the *Usetelephone* class compared to the skeleton modality, with an improved accuracy of 46.3%.

Although, as illustrated in Table 5, SSRT outperformed every other HAR in cross-dataset HAR tasks, it can be seen from Figure 12 that when compared with the accuracy of each class, SSRT negatively impacts the accuracy of all action classes except for the *Uselaptop* class when compared to the accuracy obtained from the best-performing single-modality (skeleton or RGB) HAR model for each class. This is mainly due to the challenges in cross-dataset HAR tasks, as discussed in Section IV-A2.

B. SSRT VS. OTHER FUSION METHODS

This study compares the accuracy of four multimodal fusion techniques. In Figure 13, the x-axis displays the fusion methods, while the y-axis presents their accuracy for three action recognition tasks. The blue and orange bar plots represent the accuracy of fine-grained HOI and general actions, respectively, and the line plot illustrates the accuracy of cross-dataset actions.

The findings reveal that, among all four multimodal methods, LSTM late score fusion consistently exhibited the lowest accuracy across all HAR tasks.Both late score fusion techniques (LSTM and Transformer) exhibited reduced accuracy compared to early fusion (Transformer Early Concatenation) and SSRT. Moreover, late score fusions were outperformed by their respective single-modality counterparts, as shown in Table 5. This table reveals that Transformer late score fusion modality experiences the most significant drop in accuracy for fine-grained HOI actions, decreasing by 16.18% compared to the RGB encoder. However, it suffers the least in general actions HAR tasks, with a decrease of only 2.69% when compared to the skeleton encoder. The primary reasons for these outcomes can be attributed to:

1) **Distinct nature of RGB and skeleton modality features**: RGB and skeleton modality features exhibit heterogeneity. Although these two features can complement each other in HAR tasks, their distinct nature makes fusion more challenging. To tackle this diversity in feature modalities, a more sophisticated fusion technique must be employed.
2) **Inadequacy of late score fusion**: As outlined in Section IV-B2, both late score fusion methods implemented in this research work merely sum the probability scores from individual single-modality models to obtain the final prediction score. In this multimodal fusion approach, fusion only takes place at the final stage, which is insufficient for effectively capturing the nuanced interplay between the two modalities. As a result, this fusion method struggles significantly in fine-grained HOI actions where understanding both RGB and skeleton features is essential. Consequently, late score fusion methods fail to supplement RGB and skeleton modality features, and instead, they negatively affect the overall performance in comparison to individual single-modality approaches.

Transformer Early Concatenation, a state-of-the-art multimodal fusion method, demonstrated a significant increase in accuracy when compared to the Transformer late score fusion method. The improvement was 19.78% for fine-grained

HOI, 5.35% for general actions, and 11.69% for cross-dataset actions. Furthermore, this fusion approach substantially enhanced both accuracy and F1 score for all HAR tasks when compared to the best single-modality HAR model. One reason for this improvement is that this method employs a multi-head attention mechanism to complement RGB and skeleton modalities.

SSRT outperformed the other three fusion methods, including Transformer Early Concatenation, showcasing 6.32% higher accuracy for fine-grained HOI, 4.04% higher accuracy for general actions, and 6.56% higher accuracy for cross-dataset actions. Two main reasons why SSRT is a better fusion method than Transformer Early Concatenation are:

1) **Early concatenation vs. Cross-attention**: As discussed in Section II-B1, the transformer cross-attention mechanism is much more efficient in understanding features from two different modalities. This is because early concatenation simply employs a single transformer encoder to fuse concatenated multimodal features, whereas SSRT allows skeleton and RGB modalities to attend to each other bidirectionally. This process is achieved by exchanging the key (K) and value (V) vectors of one modality with the query (Q) sequences of another modality within multiple stream transformer layers.
2) **Single-stage fusion vs. Two-stage fusion**: SSRT employs two stages of fusion, which are transformer cross-attention and Softmax layer late score fusion, while Transformer Early Concatenation only employs a single stage of multimodal fusion.

C. THE EFFECTIVENESS OF THE LSTM COMPONENT IN SSRT

This section examines the influence of LSTM on the proposed method. To do this, all experiments in this paper were performed again by substituting the LSTM component with traditional positional encoding. The effects of LSTM on SSRT are depicted using a bar plot in Figure 14. The x-axis in Figure 14 represents the three action recognition tasks, while the y-axis displays the accuracy and F1 score values for each task. The orange bar shows the accuracy and F1 score of the proposed SSRT, whereas the blue bar represents the accuracy and F1 score of SSRT with positional encoding instead of LSTM.

The results indicate that the integration of LSTM has the most significant influence on fine-grained HOI action recognition tasks, contributing to a 6.32% increase in accuracy and a 6.38% improvement in the F1 score. The smallest impact is observed in general action recognition tasks, with increases of 3.36% in accuracy and 3.44% in F1 score. For cross-dataset action recognition tasks, the implementation of LSTM enhances the overall robustness of SSRT, resulting in a 5.37% improvement in accuracy and a 5.26% boost in the F1 score.

These findings imply that the combination of an LSTM with a transformer encoder generates more effective temporal abstract features for each modality. This outcome can be ascribed to the unique sequential processing approach of LSTM. Unlike positional encoding, LSTM processes each input instance in a sequential manner, enabling it to capture temporal information more effectively. On the other hand, the positional encoding layer assigns each element's position based on a predefined sinusoidal function and subsequently processes the entire input sequence in parallel using the transformer encoder. As a result, positional encoding doesn't capture temporal dependencies as efficiently as LSTM.

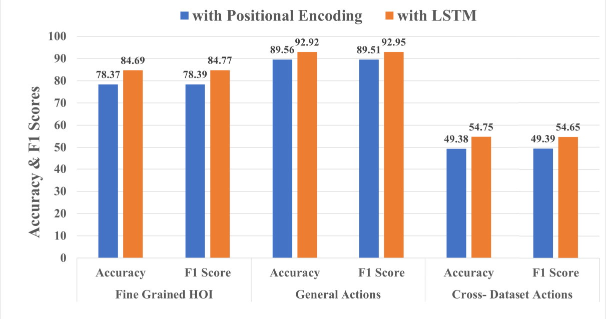


FIGURE 14. Performance comparison of LSTM vs Positional encoding.

Remarkably, SSRT consistently outperforms the Transformer Early Concatenation fusion method across all action recognition tasks, even when only positional encoding layers are incorporated. This superiority is particularly evident in fine-grained HOI actions, where a significant improvement is noted. This result further validates the superiority of SSRT as a fusion method, as detailed in Section VI-B. It is also noteworthy that when LSTM is employed solely as the HAR model, its performance is somewhat lackluster, as illustrated in Table5. The study demonstrates that when a transformer encoder is integrated with LSTM, a notable improvement is observed, corroborating the findings of this research work.

VII. DISCUSSION

This study highlights that the SSRT method surpasses both single-modality and multimodality HAR models in three action recognition tasks, as shown in Table 5. The most significant impact of SSRT is evident in fine-grained HOI action recognition, where it not only achieved a considerable increase in accuracy but also displayed consistent enhancements across all action classes within the same coarse label. The proposed method also generalizes well to action classes beyond fine-grained HOI and proved most robust when evaluated on action classes from other datasets. However, it is important to note that, except for the *Uselaptop* class, SSRT adversely influences the accuracy of all action classes when compared to the top-performing single-modality (skeleton or RGB) HAR model for each class. This can mainly be ascribed to the inherent challenges associated with cross-dataset HAR tasks. In addition to these challenges, the reason for SSRT not performing exceptionally well in cross-dataset actions may be due to its inability to extract superior higher-level features from each modality before performing multimodal fusion.From the overall experiments conducted, the SSRT method demonstrates state-of-the-art results for multimodal fusion techniques; however, SSRT can still be improved by utilizing better higher-level features from both modalities and subsequently integrating them using the proposed two-stage fusion method. This can be achieved in the following ways:

1) **Utilization of 3D skeleton features**: In this study, we implemented Alphapose human pose estimation, a popular human skeleton feature estimator, to extract 2D skeleton features. Although 2D skeleton features are computationally efficient and simpler to implement, they are not as robust as 3D skeleton features. 3D skeleton features provide depth information in addition to 2D skeleton features, resulting in a more accurate representation in 3D space. 3D poses are more robust to the shape and size of humans as well as varying camera angles and heights. For instance, An et al. [55] utilize multimodal approaches, encompassing mmWave, RGB-D, and inertial sensors, to achieve superior 3D human pose estimation representations, which are notably more robust than conventional 2D skeleton features. This information could be vital in addressing the challenges faced by SSRT in cross-dataset actions. In our future work, we plan to use state-of-the-art 3D skeleton feature extraction methods such as those proposed in [49] and [55].
2) **Selection of better baseline for higher features extraction**: In this study, we employed a transformer-based architecture as the baseline for higher-level feature extraction, with the rationale outlined in Section IV-B1. Although, implemented transformer-based architecture shows comparable state-of-the-art results for both RGB and skeleton modality but this architecture may not be the best solution for each modality. For instance, Graph Convolutional Networks like STGCN and 2S-AGCN, specifically tailored for processing skeleton features, might extract more intricate skeleton data compared to the transformer baseline. Likewise, pre-trained 3D Convolutional Networks, such as I3D [4], could potentially derive superior features from the RGB modality. In future work, we plan to investigate the integration of diverse state-of-the-art HAR models to extract sophisticated higher-level features, followed by employing the proposed two-stage fusion for enhanced feature integration. This approach may bolster SSRT's ability to comprehend cross-dataset actions more effectively. Furthermore, we will examine the combination of various HAR baselines to achieve cutting-edge performance on prominent HAR datasets, such as [14], [27], and [28].

The experimental results demonstrate that neither the skeleton nor the RGB modality consistently outperforms the other, as performance varies depending on the action recognition task. In fine-grained HOI action classification, the RGB modality excels, likely because the skeleton modality has difficulty recognizing similar motion dynamics. In contrast, the skeleton modality fares better in general Human Activity Recognition (HAR) and cross-dataset HAR tasks, potentially due to more distinct movement patterns. In cross-dataset HAR tasks, the RGB modality encounters challenges arising from significant differences in the required RGB features.

As highlighted in Section V-A, the experimental study conducted here suggests that multimodal HAR models do not always surpass single-modality models in performance. Late score fusion models, such as Transformer Late Score Fusion and LSTM Late Score Fusion, exhibited decreased overall accuracy and F1 Score compared to their single-modality counterparts. The proposed SSRT outperformed the state-of-the-art Transformer Early Concatenation in all tasks, even with only the traditional positional encoding and without utilizing the LSTM. It is important to note, however, that employing the LSTM in place of positional encoding substantially improved SSRT's performance.

Lastly, this study's primary limitation concerning fine-grained HOI is the scarcity of available datasets. Apart from the fine-grained HOI of drinking from the Toyota smart home dataset, no other related datasets were found for this task. As part of our future work, we aim to collect more fine-grained HOI action classes to enhance dataset diversity.

VIII. CONCLUSION

In this research, SSRT (a novel method specifically designed for fine-grained HOI recognition) is introduced by integrating skeleton and RGB modalities. SSRT first obtains abstract temporal features from each modality using an LSTM and a transformer encoder. Subsequently, SSRT employs two fusion stages: cross-attention multimodality fusion and Softmax late score fusion for effective feature integration.

The study demonstrates that SSRT outperforms state-of-the-art single-modality HAR models (such as transformer encoders) and multimodal based HAR models (such as Transformer Early Concatenation) in fine-grained HOI tasks without any bias towards a particular action class. Moreover, SSRT also excels in general HAR and cross-dataset HAR tasks. The research highlights the significant accuracy improvement achieved by incorporating an LSTM layer instead of a positional encoder layer in SSRT across all three HAR tasks.

Additionally, this study compared the performance of skeleton modality and RGB modality across all three HAR tasks. It revealed that the RGB modality outperformed in a fine-grained HOI task, while the skeleton modality exhibited better results in general and cross-dataset HAR tasks.

Lastly, the adaptability of SSRT enables it to merge various modalities for diverse purposes, including Vision-Language tasks. Thus, in future work, other researchers can explore SSRT's potential for combining different modalities to serve various purposes.

REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 1–43, 2011.

[2] M. R. Sudha, K. Sriraghav, S. S. Abisheck, S. G. Jacob, and S. Manisha, "Approaches and applications of virtual reality and gesture recognition: A review," *Int. J. Ambient Comput. Intell.*, vol. 8, no. 4, pp. 1–18, Oct. 2017.

[3] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vis. Image Understand.*, vol. 115, no. 2, pp. 224–241, Feb. 2011.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[5] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[6] V. Veeriah, N. Zhuang, and G. Qi, "Differential recurrent neural networks for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4041–4049.

[7] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6202–6211.

[8] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6836–6846.

[9] F. Becattini, T. Uricchio, L. Seidenari, L. Ballan, and A. D. Bimbo, "Am I done? Predicting action progress in videos," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 4, pp. 1–24, Nov. 2020.

[10] Y. Zheng, X. Li, and X. Lu, "Unsupervised learning of human action categories in still images with deep representations," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 4, pp. 1–20, Nov. 2019.

[11] G. Moon, H. Kwon, K. M. Lee, and M. Cho, "IntegralAction: Pose-driven feature integration for robust human action recognition in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3339–3348.

[12] X. Weiyao, W. Muqing, Z. Min, and X. Ting, "Fusion of skeleton and RGB features for RGB-D human action recognition," *IEEE Sensors J.*, vol. 21, no. 17, pp. 19157–19164, Sep. 2021.

[13] X. Zhu, Y. Zhu, H. Wang, H. Wen, Y. Yan, and P. Liu, "Skeleton sequence and RGB frame based multi-modality feature fusion network for action recognition," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 3, pp. 1–24, Aug. 2022.

[14] S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond, and G. Francesca, "Toyota smarthome: Real-world activities of daily living," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 833–842.

[15] S. Das, S. Sharma, R. Dai, F. Bremond, and M. Thonnat, "VPN: Learning video-pose embedding for activities of daily living," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 72–90.

[16] S. Das, R. Dai, D. Yang, and F. Bremond, "VPN++: Rethinking video-pose embeddings for understanding activities of daily living," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9703–9717, Dec. 2022.

[17] Z. Wu, Y.-G. Jiang, X. Wang, H. Ye, and X. Xue, "Multi-stream multi-class fusion of deep networks for video classification," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 791–800.

[18] S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie, and Y. Zhuang, "Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2330–2343, Sep. 2018.

[19] R. Zhao, H. Ali, and P. van der Smagt, "Two-stream RNN/CNN for action recognition in 3D videos," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 4260–4267.

[20] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2904–2913.

[21] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–9.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[23] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[24] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *Int. J. Robot. Res.*, vol. 32, no. 8, pp. 951–970, Jul. 2013.

[25] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 2556–2563.

[26] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1290–1297.

[27] A. Shahroudy, J. Liu, T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[28] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, and J. Kim, "ETRI-Activity3D: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 10990–10997.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, vol. 30, no. 1, pp. 5998–6008.

[30] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 32, 2019, pp. 13–23.

[31] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7464–7473.

[32] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2334–2343.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," 2022, *arXiv:2206.06488*.

[35] K. Gavrilyuk, R. Sanford, M. Javan, and C. G. M. Snoek, "Actor-transformers for group activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 839–848.

[36] J. Do and M. Kim, "Multi-modal transformer for indoor human action recognition," in *Proc. 22nd Int. Conf. Control, Autom. Syst. (ICCAS)*, Nov. 2022, pp. 1155–1160.

[37] 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," *Appl. Sci.*, vol. 13, no. 4, p. 2058, Feb. 2023.

[38] T. Rahman, M. Yang, and L. Sigal, "TriBERT: Human-centric audio-visual representation learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 34, 2021, pp. 9774–9787.

[39] S. Yan, X. Xiong, A. Arnab, Z. Lu, M. Zhang, C. Sun, and C. Schmid, "Multiview transformers for video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3333–3343.

[40] M. Ijaz, R. Diaz, and C. Chen, "Multimodal transformer for nursing activity recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2065–2074.

[41] W. Zhang, F. Qiu, S. Wang, H. Zeng, Z. Zhang, R. An, B. Ma, and Y. Ding, "Transformer-based multimodal information fusion for facial expression analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2428–2437.

[42] D. Ahn, S. Kim, H. Hong, and B. Chul Ko, "STAR-transformer: A spatio-temporal cross attention transformer for human action recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3330–3339.

[43] Y. Zhang, B. Wu, W. Li, L. Duan, and C. Gan, "STST: Spatial-temporal specialized transformer for skeleton-based action recognition," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3229–3237.

[44] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge, "Action transformer: A self-attention model for short-time pose-based human action recognition," *Pattern Recognit.*, vol. 124, Apr. 2022, Art. no. 108487.

[45] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. ICML*, 2021, vol. 2, no. 3, p. 4.

[46] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, ''Optuna: A next-generation hyperparameter optimization framework,'' in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2623–2631.

[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ''An image is worth 16 × 16 words: Transformers for image recognition at scale,'' 2020, *arXiv:2010.11929*.

[48] I. Loshchilov and F. Hutter, ''Decoupled weight decay regularization,'' 2017, *arXiv:1711.05101*.

[49] G. Rogez, P. Weinzaepfel, and C. Schmid, ''LCR-Net++: Multi-person 2D and 3D pose detection in natural images,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1146–1161, May 2020.

[50] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, ''Cross-view action modeling, learning, and recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2649–2656.

[51] Q. Cheng, Z. Liu, Z. Ren, J. Cheng, and J. Liu, ''Spatial-temporal information aggregation and cross-modality interactive learning for RGB-D-Based human action recognition,'' *IEEE Access*, vol. 10, pp. 104190–104201, 2022.

[52] L. Yuan, J. Andrews, H. Mu, A. Vakil, R. Ewing, E. Blasch, and J. Li, ''Interpretable passive multi-modal sensor fusion for human identification and activity recognition,'' *Sensors*, vol. 22, no. 15, p. 5787, Aug. 2022.

[53] C. Plizzari, M. Cannici, and M. Matteucci, ''Skeleton-based action recognition via spatial and temporal transformer networks,'' *Comput. Vis. Image Understand.*, vols. 208–209, Jul. 2021, Art. no. 103219.

[54] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, ''Disentangling and unifying graph convolutions for skeleton-based action recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 143–152.

[55] S. An, Y. Li, and U. Ogras, ''MRI: Multi-modal 3D human pose estimation dataset using mmWave, RGB-D, and inertial sensors,'' 2022, *arXiv:2210.08394*.

[56] L. Shi, Y. Zhang, J. Cheng, and H. Lu, ''Two-stream adaptive graph convolutional networks for skeleton-based action recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12026–12035.

[57] G. Joshi, R. Walambe, and K. Kotecha, ''A review on explainability in multimodal deep neural nets,'' *IEEE Access*, vol. 9, pp. 59800–59821, 2021.

AKASH GHIMIRE is currently pursuing the B.E. degree with the Department of Integrated System Engineering, School of Global Convergence Studies, Inha University. His research interests include video understanding, utilizing deep learning, and computer vision techniques, with a focus on identifying human actions through various methods.

VIJAY KAKANI (Member, IEEE) received the B.Tech. degree in electronics and communication engineering from Jawaharlal Nehru Technological University, Kakinada, India, in 2012, the M.S. degree in computer and communication systems from the University of Limerick, Ireland, in 2014, and the Ph.D. degree in information and communication engineering (major) and future vehicle engineering (minor) from Inha University, South Korea, in 2020. He is currently an Assistant Professor with the Department of Integrated System Engineering, School of Global Convergence Studies, Inha University. His research interests include autonomous vehicles, sensor signal processing, applied computer vision, deep learning, systems engineering, and machine vision applications.

HAKIL KIM (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical and computer engineering from Purdue University, in 1985 and 1990, respectively. In 1990, he joined the College of Engineering, Inha University, Incheon, South Korea, where he is currently a Full Professor with the Department of Information and Communication Engineering. In order to retain the balance between academic research and commercial development, he founded Vision Inc., in 2014, where he is also the CEO. His research interests include biometrics, intelligent video surveillance, and embedded vision for autonomous vehicles. Since 2003, he has been actively involved as a Project Editor in the International Standardization of Biometrics at ISO/IEC JTC1/SC37.