IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 April 2024, accepted 21 June 2024, date of publication 3 July 2024, date of current version 15 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3422428

RESEARCH ARTICLE

DDC3N : Doppler-Driven Convolutional 3D Network for Human Action Recognition

MUKHIDDIN TOSHPULATOV[1], (Member, IEEE), WOOKEY LEE[1], (Member, IEEE), SUAN LEE[2], (Member, IEEE), HOYOUNG YOON[3], (Member, IEEE), AND U KANG[3], (Member, IEEE)

[1]Biomedical Science and Engineering, Inha University, Incheon 22212, South Korea
[2]School of Computer Science, Semyung University, Jecheon 27136, South Korea
[3]Department of Computer Science and Engineering, Seoul National University, Seoul 08826, South Korea

Corresponding author: Wookey Lee (trinity@inha.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP), South Korea, funded by Korean Government through the Ministry of Science and ICT (MSIT) (XVoice: Multi-Modal Voice Meta-Learning) under Grant 2022-0-00641 and [Artificial Intelligence Graduate School Program (Seoul National University)] under Grant RS-2021-II211343; and in part by Inha University.

ABSTRACT In deep learning (DL)–based human action recognition (HAR), considerable strides have been undertaken. Nevertheless, the precise classification of sports athletes' actions still needs to be completed. Primarily attributable to the exigency for exhaustive datasets about sports athletes' actions and the enduring quandaries imposed by variable camera perspectives, mercurial lighting conditions, and occlusions. This investigative endeavor thoroughly examines extant HAR datasets, furnishing a yardstick for gauging the efficacy of cutting-edge methodologies. In light of the paucity of accessible datasets delineating athlete actions, we have taken a proactive stance, endeavoring to curate two meticulously datasets tailored explicitly for sports athletes, subsequently scrutinizing their consequential impact on performance enhancement. While the superiority of 3D convolutional neural networks (3DCNN) over graph convolutional networks (GCN) in HAR is evident, it must be acknowledged that they entail a considerable computational overhead, particularly when confronted with voluminous datasets. Our inquiry introduces innovative methodologies and a more resource-efficient remedy for HAR, thereby alleviating the computational strain on the 3DCNN architecture. Consequently, it proffers a multifaceted approach towards augmenting HAR within the purview of surveillance cameras, bridging lacunae, surmounting computational impediments, and effectuating significant strides in the accuracy and efficacy of HAR frameworks. GitHub link: https://github.com/muxiddin19/DDC3N-Doppler-Driven-C3D-Network-for-HAR

INDEX TERMS 3D pose estimation, discriminator, deep neural network, deep learning, generator, mesh estimation, metadata, skeleton, top-down approach, motion embedding, optical flow map, channel-wise, spatiotemporal, doppler, dataset, action recognition.

I. INTRODUCTION

The study of human action recognition (HAR) represents a burgeoning domain within the ambit of artificial intelligence (AI), endeavoring to decipher the intricate tapestry of human activities through the synergistic amalgamation of computer vision (CV) and machine learning (ML) methodologies. Conceptualize a computational system endowed with the capacity to discern and categorize an array of human actions, from pedestrian locomotion to intricate dance maneuvers and even the delineation of accidental falls [1]. The transformative potential encapsulated within HAR's purview extends far beyond academia's confines, permeating diverse realms, including healthcare, sports analysis, security, and entertainment. This burgeoning field, propelled by rapid technological advancements, portends to revolutionize human-computer

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

interaction (HCI) paradigms, heralding an era characterized by seamless integration and intuitive engagement with technological interfaces [2].

Central to the pursuit of deep learning (DL)-based HAR lies the indispensable role assumed by datasets, serving as the quintessential bedrock upon which computational algorithms are honed and refined. Analogous to the ingredients meticulously curated within a culinary compendium, these datasets imbue our computational models with the acumen to discern and classify an expansive spectrum of human actions. However, as we focus on sports athletes, a confluence of complexities emerges, necessitating specialized datasets calibrated to accommodate the nuanced exigencies intrinsic to athletic performance. The athletic milieu, characterized by the dynamic interplay of human kinetics within multifaceted environments such as stadiums, presents formidable challenges, including oblique camera angles, disparate lighting conditions, and occlusions impeding unobstructed visibility [5].

In response to these exigencies, we advocate for curating specialized datasets tailored to the distinctive demands of athletic performance, thereby furnishing our computational models with the requisite granularity to interpret and comprehend the intricacies inherent within athletes' movements. These datasets, far from mere repositories of raw information, emerge as veritable treasure troves brimming with transformative potential, engendering novel insights and refining computational algorithms to elucidate the nuances of athletic performance.

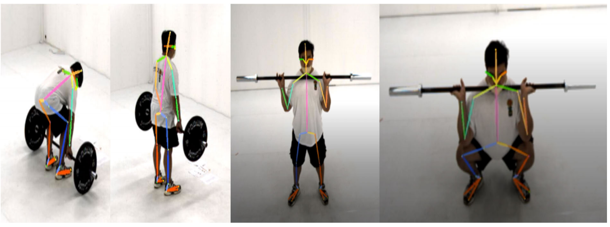


FIGURE 1. Pose estimation examples.

Within the contemporary landscape of DL-based HAR methodologies, 3D convolutional neural networks (3DCNN) and graph convolutional networks (GCN) occupy prominent positions. Our investigations reveal a notable efficacy disparity, with 3DCNN-based techniques outperforming their GCN-based counterparts. However, the computational overhead incurred by 3DCNN networks, particularly when confronted with voluminous datasets, underscores the exigency for innovative solutions to ameliorate computational burden. Our endeavor culminates in developing the Doppler-driven convolutional 3D Network (DDC3N) for HAR, a pioneering architectural framework poised at surmounting existing performance benchmarks. Introducing an efficient solution to alleviate the computational onus on 3DCNN networks, the DDC3N heralds a paradigm shift in HAR methodologies. Remarkably, our study represents the inaugural attempt at incorporating Doppler mechanisms into skeleton-based HAR, yielding a marked enhancement in accuracy vis-à-vis extant state-of-the-art (SOTA) methodologies.

Moreover, cognizant of the indispensable role assumed by meticulously curated datasets, we have embarked on an initiative to craft specialized repositories tailored to the unique requisites of CrossFit and Figure Skating sports activities, as elucidated in Figure 1. These datasets, characterized by their exhaustive coverage of diverse movements and poses, not only serve as invaluable resources for researchers but also furnish a standardized benchmark for evaluating models focused on HAR and pose estimation within the sporting domain. Our scholarly endeavor embodies methodological innovation and empirical rigor, poised to advance the frontiers of human-centric technological endeavors.

II. RELATED WORKS

This section is divided into several related subsections essential to this research field. We will start with human pose estimation (HPE) and HAR, as the current study directly relates to them. Let us begin with the first one.

A. HUMAN POSE ESTIMATION

Pose estimation is a relatively new CV technique that has gained significance in various applications [21]. It involves estimating the 2D posture of individuals in images and has relevance in tasks such as action identification, security, and sports analysis. Initial approaches approximated the stance of a single person in a snap.

With CNNs emerging, HPE accuracy was significantly improved. The Mask Region-Based Convolutional Neural Network (Mask R-CNN) architecture, known for semantic and instance segmentation, extended its capabilities to HPE [22]. It utilized CNNs to extract image features and employed a Region Proposal Network (RPN) for bounding box generation [23]. However, this method still required separate phases for person and part detection.

A gap existed in understanding an individual's orientation through pose estimation, particularly in domains such as sports and fitness. A human pose skeleton, representing an individual's orientation, became a key concept, consisting of data points characterizing their stance. Relationships between coordinates within these pose skeletons were fundamental. OpenPose emerged as an efficient and robust solution for multi-person HPE in crowded scenes [15]. It detected multiple body parts and facial key points, setting high standards for keypoint estimation. Its features included real-time 3D single-person and 2D multi-person keypoint detections and introduced the concept of Part Affinity Fields (PAFs) to represent the affinity between body parts [13]. Furthermore, it leveraged bipartite graphs and PAF values for multi-person pose estimation.

In the context of HPE, challenges persisted, including occlusion, limited training data, and depth ambiguity [8]. Precise 3D pose annotations could have been more problematic. Various approaches, such as those involving inertial measurement units (IMUs), depth sensors, and radiofrequency devices, were proposed to address these challenges, although they often required specialized hardware [29]. Addressing the gap in precise 3D pose estimation from 2D data, a Bayesian formulation using Capsule networks was introduced [29]. This approach aimed to regularize the ill-posed optimization problem and successfully estimate 3D coordinates of human pose joints. Furthermore, a recent approach focused on high-resolution representation learning for HPE [30], maintaining high-resolution representations throughout the network and achieving rich, high-resolution illustrations. Efficiency in multi-person HPE was a significant concern, and single-stage models, such as Single-stage multi-person Pose Machine (SPM), aimed to close this gap [7]. SPM introduced a Structured Pose Representation (SPR) to predict structured poses for multiple individuals in a single stage, and an instance-aware module adapted network parameters for each instance, improving capacity and adaptive-ability [9]. This approach offered a more compact pipeline and enhanced efficiency. HRNet, developed by researchers at the Chinese University of Hong Kong, addressed the gap in high-resolution representations for 2D keypoint estimation [37]. It effectively captured fine-grained and global information about human body parts. Additionally, Facebook AI Research introduced two frameworks, SimpleBaseline and DensePose, to address gaps in HPE [38]. SimpleBaseline offered high accuracy with low computational requirements, while DensePose estimated both 2D keypoints and 3D surface correspondence, accomplishing SOTA result [39]. These HPE methods have different strengths and weaknesses achieving SOTA performance on benchmark datasets. The choice of method depends on specific application requirements, computational resources, and the trade-offs between accuracy and efficiency.

B. HUMAN ACTION RECOGNITION

HAR is another challenging task in CV that involves identifying and classifying human actions in a video or image sequence. Researchers have proposed various methods for this task, including appearance-based and motion-based approaches [31], [32]. One popular and effective approach is skeleton-based HPE, which estimates the 3D coordinates of a person's joints from the given dataset and constructs a skeleton representation of their pose. ML algorithms are also being actively exploited to recognize different actions based on the movement patterns of the joints [31]. Using skeleton-based HPE has several advantages, including computational efficiency, robustness to changes in lighting and background, and a more interpretable representation of the actions being performed. However, it faces challenges such as occlusions, clothing and body shape variations, and recognizing similar movements.

Current advances in this field have focused on using DL models such as CNNs and GCNs for skeleton-based approaches [31], [32], [44]. GCNs are a neural network that can operate on graph-structured data, such as skeleton data, and have been used for skeleton-based action recognition. One example is the Spatial-Temporal Graph Convolutional Networks (ST-GCN) method proposed by Yan et al. [31], which uses a convolutional graph network to model the temporal and spatial relationships between the joints. Another approach is using GCNs to learn the graph structure of the skeleton, capturing more complex relationships between the joints. Zhao et al. proposed a method called ActionGCN [32] that uses a convolutional graph network to learn a global skeleton representation.

Recent research has proposed hybrid approaches that combine skeleton-based and appearance-based methods to improve the accuracy of HAR. These approaches exploit human activities' kinematic and visual information for superior performance. Building on this work, Zhang et al. [16] presented a GCN-based system with a novel graph attention mechanism. It learns discriminative features from skeleton data, achieving comparable results on several benchmark HAR datasets. Chen et al. [79] further improved feature extraction by introducing the CTR-GCN skeleton-based approach. It dynamically learns different topologies, effectively aggregates joint features in other channels, and integrates temporal modeling modules to create a powerful GCN. In contrast, it primarily focuses on encoding skeleton information and embedding it into the latent representations of human action. InfoGCN [78] proposes a novel learning framework that combines an information bottleneck-based learning objective with attention-based graph convolution for HAR using skeleton data to address this limitation.

With further advances in ML algorithms and sensor technology, skeleton-based HPE can become an even more powerful tool for HAR in the future [36], [48], [49]. 3D-CNNs are a type of neural network that can operate on spatiotemporal data, such as skeleton data. They have been used for skeleton-based action recognition because they capture spatial and temporal features simultaneously [33], [45], [46], [47]. Huang et al. proposed a 3D-CNN-based approach that uses a temporal pyramid pooling method to capture multi-scale temporal information from skeleton data, achieving SOTA performance on the NTU RGB+D 60 dataset. Narang et al. [41] proposed a 3D-CNN-based approach that uses a motion attention mechanism to weigh different joints' importance for action recognition, achieving SOTA performance on the NTU RGB+D 120 dataset [41]. The recent method, PoseC3D [1], evaluates the performance of different models on several benchmark datasets and shows that 3D CNN-based methods perform better than previous SOTA methods [33], [34], [35], [43].

III. DATA ENGINEERING

Within AI endeavors, the paucity of requisite data emerges as a salient impediment, precipitating a conundrum that vitally impacts the efficacy and trajectory of AI projects. Indeed, the acquisition and curation of extensive datasets constitute an elemental prerequisite for realizing AI success, with the need for appropriate data to precipitate suboptimal outcomes that impede fulfilling a project's latent potential. Invariably, the need for more suitable data engenders a disconcerting reality wherein the attainment of requisite datasets proves to be a formidable and time-intensive endeavor, thus thwarting the project's advancement. This predicament underscores the imperative for the availability of tailored datasets conducive to the unique demands of AI applications, a shortfall that, when perpetuated, exacts a toll on the enthusiasm and momentum of project stakeholders.

Supervised Learning (SL), hailed as a quintessential paradigm within AI methodologies, exhibits prodigious efficacy in addressing diverse business challenges. However, its operational efficacy is inexorably tethered to the availability of copious and pertinent training datasets [72], [73], [74]. The efficacy of SL hinges upon the meticulous curation of extensive and relevant datasets. This prerequisite underscores the pivotal role assumed by data abundance in augmenting the efficacy and fidelity of AI models. Indeed, the success of SL paradigms is contingent upon the accessibility of expansive datasets endowed with the requisite granularity to facilitate robust model training and validation processes.

Creating large training datasets can be challenging due to limited availability, time consumption, and human and machine expenses. To tackle this issue, we have taken the initiative to build two new datasets, CrossFit and Figure Skating. These datasets help overcome data scarcity, opening up new possibilities for AI research related to HPE and HAR in academia and industry. They are poised to revolutionize action recognition in sports. With a specific emphasis on sports athletes' action recognition, the importance of these datasets extends far beyond the realm of research. They offer a unique opportunity to bridge the gap between scientific datasets and real-world applications, where accuracy and robustness in action recognition are paramount. These datasets can potentially deliver valuable insights to coaches, sports analysts, and healthcare experts, enabling them to create sports analysis, fitness monitoring, and rehabilitation applications. Their availability is set to usher in a new era of excellence in HPE and action recognition, particularly in sports.

In crafting the datasets, meticulous consideration was given to potential biases stemming from various factors, including the criteria for selecting video clips, the breadth of actions represented, and the precision of the HPE process. Acknowledging these biases is imperative to ensure the study's transparency and mitigate their impact on the outcomes. To mitigate the biases above, the authors implemented several strategic measures. Notably, specialists from the respective sports activities were consulted to determine the primary action classes and potential error classes, thus ensuring a comprehensive representation within each dataset. Consequently, each dataset was meticulously curated to encompass normative and comprehensive action classes.

Additionally, installing eight semi-cameras from varied angles addressed issues such as camera occlusion and ensured comprehensive coverage of the action sequences. Moreover, including eight videos from distinct angles for each action further bolstered the dataset's robustness against potential biases. Moreover, a systematic approach was adopted to mitigate biases associated with human pose estimation results, which directly influenced the model's performance. Specific threshold values were meticulously chosen for pose estimation accuracy, and metadata was generated to identify frames where the accuracy surpassed or fell below these thresholds. Frames failing to meet the prescribed accuracy thresholds were systematically excluded from the experiments, thus mitigating the potential biases of inaccurate pose estimations.

Both datasets were meticulously cultivated under controlled indoor conditions, leveraging SOTA deep pose-based Simi high-speed cameras positioned from eight perspectives within the indoor shooting environment. These high-speed cameras, coupled with sophisticated software, hardware, and storage systems, ensured the acquisition of high-quality data essential for robust analysis. Additionally, each dataset class was stratified into beginner, intermediate, and advanced levels, further enhancing the granularity and applicability of the datasets. Notably, the demographic composition of actors in both datasets was meticulously balanced regarding age and sex percentages, fostering representativeness and generalizability. However, variations in the number of activities or classes across datasets were contingent upon the specificities of the respective sports and athletes. This facet will be expounded upon comprehensively in subsequent subsections.

A. CrossFit DATASET

CrossFit, founded by Greg Glassman in 2000, is a high-intensity fitness regimen known for its varied functional movements. Our dataset is crucial for DL-based research in HAR, featuring nine standard actions and twenty-four error movements, serving the research community. With 33 distinct classes, it's a cornerstone for ML in human action classification, offering a nuanced perspective on human actions. This dataset is valuable for recognizing fundamental movements, spanning proficiency levels from beginner to advanced, and serving as a benchmark for assessing understanding (see Table 1).

CrossFit's U.S. Headquarters courses (ON RAMP and LV.1) highlight the practical importance of these actions. Figure 2 provides a clear example of an athlete in action. The dataset consists of two main components, normal and full datasets, aiming to advance DL in HAR, impacting various applications from sports analysis to security. It covers three athlete proficiency levels (32% beginner, 39% intermediate, and 29% advanced) and includes nine key activities. The dataset is evenly distributed, with each action class contributing equally (11.11% each), promoting balanced research (see Table 2).

TABLE 1. Datasets details.

Datasets	Raw videos	Refined videos	Selected videos	Dataset split ratio(8:1:1)		
				Training	Validation	Tesing
CrossFit	192000	189000	186000	148800	18600	18600
Figure Skating	162000	159000	156000	124800	15600	15600

TABLE 2. Classification of CrossFit dataset.

Content			Acquisition ratio		
			Beginner	Intermediate	Advanced
Squatting	Air squat Front squat Overhead	33.3%	32%	39%	29%
Press	Shoulder press Push press Jerk press	33.3%	32%	39%	29%
Deadlift	Dead lift Smod lift	22.2%	32%	39%	29%
Clean	Medicine ball clean	11.1%	32%	39%	29%

TABLE 3. Classification of Figure Skating dataset.

Content			Acquisition ratio		
			Beginner	Intermediate	Advanced
Jump	Waltz 1S (Salchow) 1T (Toe loop) ILo (Loop) 1F (Flip) 1Lz (Lutz)	37.5%	36%	31%	33%
Spin	Two-Foot Spin One-Foot Spin Upright Sit Camel	31.25%	36%	31%	33%
Step	Swizzle Three Turn Stroking Cross Over Spiral	31.25%	36%	31%	33%

B. FIGURE SKATING DATASET

Our new dataset for Figure Skating is a comprehensive resource for DL-based HAR. It comprises fifty-three classes categorized into normal and error behaviors, reflecting the criteria referees use to assess athletes' performances. These actions are centered on the fundamental tasks of jumps, spins, and step-by-step sequences. Within these categories, we've selected sixteen evenly distributed normal movements, including six jumps with detailed phases and five spins. This nuanced categorization provides a strong foundation for sophisticated DL models. The dataset offers a wealth of information for researchers, promising advancements in Figure Skating action recognition and applications such as sports analysis and human-computer interaction. Like the CrossFit dataset, the Figure Skating Dataset is divided into three proficiency levels (36% beginner, 31% intermediate, and 33% advanced), reflecting realistic gender and skill level distributions. It classifies actions into jumps, spins, and steps, with each normal action class having related error action classes. Each normal action class is equally distributed (6.25% each) (see Table 1, and Table 3).

IV. PROPOSED METHODOLOGY

The proposed methodological framework constitutes a paradigmatic synthesis of Doppler-driven and 2D pose extraction blocks, as delineated in Figure 11. This architectural instantiation embodies a nuanced confluence of two distinct modules, CWSTB and the Doppler, synergistically orchestrated to effectuate a holistic fusion of spatiotemporal and Doppler features within a unified framework. The integration of these constituent modules furnishes a robust foundation for action recognition, wherein the amalgamation of spatiotemporal and motion features is seamlessly processed through a Convolutional 3D (C3D) network, thereby epitomizing a harmonious union of computational prowess and methodological ingenuity.

This innovative methodological underpinning stands as a testament to the efficacy of video data processing, leveraging the synergistic amalgamation of Doppler, spatial, and temporal information to engender heightened performance in action recognition endeavors. Central to this architectural instantiation is the CWSTB module, meticulously engineered to optimize the efficacy of 3D CNNs in capturing

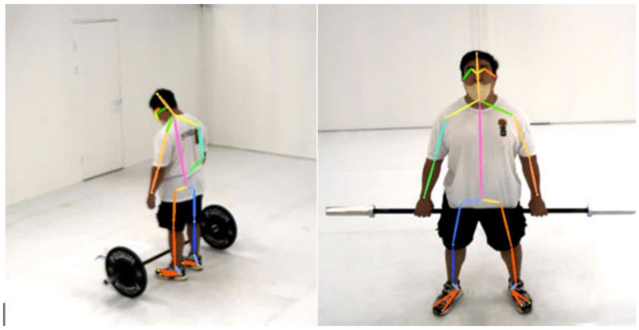


FIGURE 2. Estimated human key point results from the newly created CrossFit video dataset.

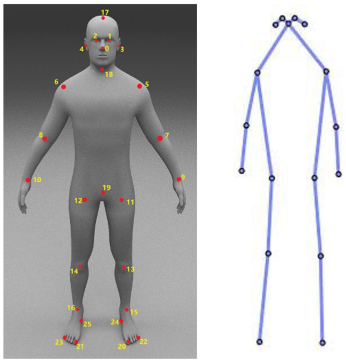


FIGURE 3. Human keypoint structure [19].

spatiotemporal intricacies, thus conferring an augmented discernment of temporal dynamics within video sequences. Concurrently, the Doppler model assumes the mantle of deftly handling motion information between successive frames, leveraging convolutional and pooling layers to distill pertinent motion cues with efficiency and precision.

The confluence of these disparate yet complementary modules is further accentuated by their integration with 2D human pose keypoints, thereby refining the granularity and fidelity of action recognition within the Convolutional C3D network. Specifically, the CWSTB module is architected to capitalize on the synergistic interplay between 1D and 2D convolutional layers, adeptly capturing spatiotemporal relationships with finesse and acumen. In parallel, the Doppler model operates in concert, leveraging convolutional and pooling layers to encapsulate motion dynamics inherent within video sequences efficiently.

This integrative approach manifests in enhanced performance across a spectrum of action recognition tasks, culminating in the generation of a probability distribution encompassing diverse action classes. Ultimately, the overarching objective of this methodological instantiation lies in the efficient optimization of HAR within video data, thus paving the way for transformative advancements within the realm of computational action analysis and understanding.

A. HUMAN 2D KEYPOINT EXTRACTION

In our study, we harnessed cutting-edge deep learning tools for HPE, deploying Faster R-CNN [41] with a ResNet 50 [13] backbone and the updated AlphaPose [16], [18]. To maintain precision, we utilized ground-truth bounding boxes for athletes. Our approach involved extracting two types of keypoint information: 26 keypoints gleaned from the Halpe-FullBody dataset and 17 from the COCO dataset. The Halpe-FullBody dataset endowed us with images adorned with 26 key points, facilitating accurate human pose estimation. On the other hand, the COCO dataset contributed 17 keypoints, a common choice for representing human body poses [14]. The HPE stage mainly involved our two datasets: CrossFit Behavior, comprising 186,000 videos, and Figure Skating, housing 156,000 videos. For each video, we generated skeletons, bolstering HPE accuracy, as illustrated in Figure 2. Using the 26 keypoints from the Halpe-FullBody dataset was particularly beneficial for enhancing HAR, as visually depicted in Figure 3.

We have compared several SOTA models for 2D human keypoint extraction, encompassing LCRNet++, AlphaPose, Mask R-CNN [14], and OpenPose. The data from our newly established CrossFit and Figure Skating datasets revealed a significant advantage for AlphaPose with 26 human joints. This model outperformed its counterparts, including Mask R-CNN, OpenPose, and LCRNet++, despite its higher computational demands [4]. We also provided the assessment of inference times among the libraries mentioned above. Notably, OpenPose exhibited consistent runtime, whereas AlphaPose and Mask R-CNN experienced linear runtime increments in proportion to the number of individuals as depicted in Figure 4. Throughout the provided experiments, we relied on several key metrics for result comparison, including the number of joints, confidence scores, missing frames, and incorrect detection. These metrics were instrumental in evaluating the representative capacity of joints, the accuracy of keypoint estimation, the identification of missing frames, and the detection of mislabeled frames. Table 4 gives the full picture of the process and the performance of the related approaches on the metrics mentioned above.

To mitigate the risk of overfitting during our experiments, we meticulously divided both datasets into training, validation, and test sets, adhering to an 8:1:1 ratio, which we ascertained as optimal through extensive trials. Both datasets featured two categories: normal data and data with error behaviors. The error behavior labels were strategically allocated to data with pose estimation errors, bolstering the model's resilience to previously unseen data. We should not have utilized some instances in the experiments due to human joint keypoint identification inaccuracies. This decision further refined model accuracy and improved the final method's performance, meticulously outlined in Table 5.

The chosen model is modified to adapt to our task and data by employing a top-down approach that prioritizes accurate person detection and focuses on regions containing individuals. By extracting accurate person regions from bounding

TABLE 4. Comparison of the recent SOTA approaches on HPE.

Methods	Joints	Missing Frames		Incorrect detection		Confident Score
		mean	median	mean	median	mean
LCRNet++	26	20.1	6	120.4	71	0.81
OpenPose	25	19.9	5	70.5	62	0.86
AlphaPose	17	20.3	6	95.4	70	0.85
AlphaPose	26	**14.9**	**3**	**65**	**45**	**0.88**

TABLE 5. Detailed explanation of the related datasets.

Dataset	Type	|Labels|	Instances					
			Total	Used	Training	Testing	Val	Unused
CrossFit	Normal data	9	83,246	81,840	65472	8,184	8,184	1,406
	Full data	33	192,000	186,000	148,800	18,600	18,600	6,000
Figure Skating	Normal data	16	70,191	68,640	54,912	6,864	6,864	1,324
	Full data	53	162,000	156,000	124,800	15,600	15,600	6,000

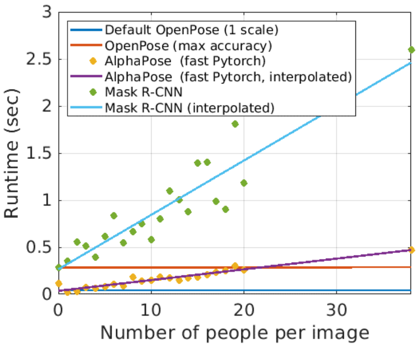


FIGURE 4. Runtime analysis [20].

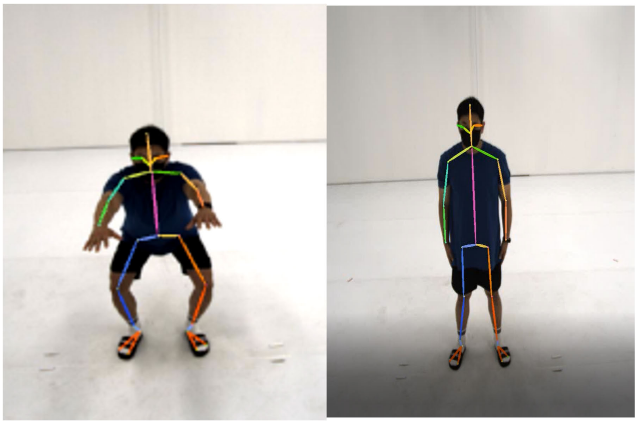


FIGURE 5. Pose estimated input data.

boxes, we enhanced precision. It was accomplished using the Symmetric Spatial Transformer Network (SSTN) [25]. A Single Person Posture Estimator (SPPE) [26] estimated the human pose within these regions. We accomplished the final step of remapping the pose to the original image coordinates by employing a spatial de-transformer network (SDTN) [27] in conjunction with the spatial transformer network (STN) (Figure 6.). The provided experiments show that the STN has demonstrated excellent performance in selecting the

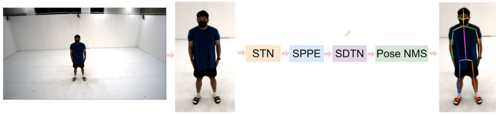


FIGURE 6. 2D human keypoint extraction.

region of interest (ROI) [6]. It extracts high-quality dominant human proposals, and it can be expressed as a 2D affine transformation (Equation 1):

$$\begin{pmatrix} x_j^{b_t} \\ y_j^{b_t} \end{pmatrix} = [\alpha_1, \alpha_2, \alpha_3] \begin{pmatrix} x_j^{a_t} \\ y_j^{a_t} \end{pmatrix} \tag{1}$$

where α_1, α_2, and α_3 are 2D vectors, and $(x_j^{b_t},\ y_j^{b_t})$ and $(x_j^{a_t},\ {a_t}_j^t)$ are related coordinates before and after transformation. After the STN and SPPE networks process the human pose, mapping it back to the original human proposal image requires an SDTN. The SDTN computes σ for de-transformation and generates grids based on σ, as described in Equation 2:

$$\begin{pmatrix} x_j^{a_t} \\ y_j^{a_t} \end{pmatrix} = [\sigma_1, \sigma_2, \sigma_3] \begin{pmatrix} x_j^{b_t} \\ y_j^{b_t} \end{pmatrix} \tag{2}$$

STN and SDTN are inverse procedures to each other, as indicated in Equations 3 and 4:

$$\left[\sigma_1, \sigma_2\right] = \left[\alpha_1, \alpha_2\right]^{-1} \tag{3}$$

$$\sigma_3 = -1 \times \left[\sigma_1, \sigma_2\right] \alpha_3 \tag{4}$$

The process of updating the parameters σ_1 and σ_2 in the SDTN network using backpropagation is described in Equation 5:

$$\frac{\partial K}{\partial(\sigma_1, \sigma_2)} \times \frac{\partial(\sigma_1, \sigma_2)}{\partial(\alpha_1, \alpha_2)} + \frac{\partial K}{\partial(\sigma_3)} \times \frac{\partial \sigma_3}{\partial(\sigma_1, \sigma_2)} \times \frac{\partial(\sigma_1, \sigma_2)}{\partial(\alpha_1, \alpha_2)} \tag{5}$$

while $\frac{\partial K}{\partial \alpha_3}$ with respect to α_3 is derived in Equation 6:

$$\frac{\partial K}{\partial \alpha_3} = \frac{\partial K}{\partial \sigma_3} \times \frac{\partial K}{\partial \alpha_3} \tag{6}$$

Equations 3 and 4 provide the relationships between the partial derivatives $\frac{\partial(\sigma_1,\sigma_2)}{\partial(\alpha_1,\alpha_2)}$ and $\frac{\partial\sigma_3}{\partial\alpha_3}$. After extracting high-quality, dominant human proposal regions can be utilized as off-the-shelf SPPE for accurate proposal estimation. In addition to these transformations, the SSTN was fine-tuned with SPPE during the training process.

The exploited Non-Maximum Suppression (NMS) [28] handled the issue of irrelevant pose deductions. Multiple detections can occur when using detectors to locate people, leading to excessive pose estimations. Pose NMS eliminates redundant results, retaining only essential and unique information.

In the context of pose NMS, the estimated human pose P_i with n joints, denoted as $(x_1^j, y_1^j, c_1^j), \ldots, (x_n^j, y_n^j, c_n^j)$, is considered, where (x_1^j, y_1^j) represents a keypoint, and c_1^j indicates the confidence score. The NMS process begins by selecting the most confident pose as a reference and discarding the pose closely resembling the reference. This procedure iterates on the remaining poses until all redundant ones have been removed, resulting in a collection of exclusively unique poses.

Establishing a quantifiable measure of pose similarity is essential to enabling the removal of closely related poses. It was achieved through a pose distance metric, $D(P_j, P_k)$, which assesses the degree of pose similarity. The threshold value ϵ is determined as a function $D(\cdot)$ parameter. The elimination criterion can be succinctly expressed as Equation 7:

$$g(P_j, P_K|\epsilon) = [D(P_j, P_K|\epsilon) \leq \epsilon] \tag{7}$$

When the value of $D(\cdot)$ is less than ϵ, the result of the function $g(\cdot)$ should be set to 1. This outcome signifies that the pose P_j ought to be removed due to redundancy to the reference poses P_j. Suppose the B_J is a given box for the P_J pose, and the matching function is defined as Equation 8:

$$S_f(P_j, P_k|\beta_1) = \begin{cases} \sum_m tanh(\frac{c_m^j}{\beta_1}) \cdot tanh(\frac{c_m^k}{\beta_1}), & \text{if } B(l_m^j) \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

In this context, $B(l_m^j)$ is a bounding box centered at l_m^j. Each dimension of $B(l_m^j)$ is one-tenth of the size of the original box B_j. The tanh operation filters out low-confidence poses, with a higher confidence score pushing the output towards one. This distance metric gradually incorporates the count of matched joints and accounts for the spatial distance between body parts, as shown in Equation 9:

$$S_d(P_j, P_k|\beta_2) = \sum(exp(-\frac{(l_m^j - l_m^k)^2}{\beta_2})) \tag{9}$$

By combining Equations 8 and 9, we can formulate the ultimate distance function as shown in Equation 10:

$$D(P_j, P_K|(\beta_1, \beta_2, \gamma)) = S_f(P_j, P_k|\beta_1) + \gamma S_d(P_j, P_k|\beta_2) \tag{10}$$

where γ is a weight that balances the two distances, unlike previous techniques that manually set pose distance parameters and thresholds, our approach uses data-driven methods, providing more adaptability and precision. Our approach leverages heatmap-based techniques [19] to effectively address issues arising from asymmetric gradients and size-dependent keypoint scoring, resulting in a novel and highly accurate human keypoint regression method. This is achieved through the utilization of the soft-argmax operation, also known as integral regression, which is a differentiable function that transforms heatmap-based methods into regression-based counterparts, enabling end-to-end training. The soft-argmax operation is expressed in Equation 11:

$$\eta' = \sum(y \cdot P_y) \tag{11}$$

where η' represents the estimated pose, y is the pixel coordinate, and P_y is the pixel likelihood on the heatmap after normalization. We minimize the l_1 norm between the predicted η' and the original η using the loss function in Equation 12:

$$L_{reg} = \|\eta - \eta'\|_1 \tag{12}$$

The gradient of each pixel can be formulated as Equation 13:

$$\frac{\partial L_{reg}}{P_y} = y \cdot sgn(\eta - \eta') \tag{13}$$

However, this results in asymmetric gradients with varying values depending on the keypoint's specific location, which can hinder the CNN network's translation invariance. To overcome this, we introduce the Amplitude Symmetric Gradient (ASG) function in the context of backward propagation, which approximates the gradient. ASG is represented in Equation 14:

$$\rho^{ASG} = Am_{gr} \cdot sgn(y - \eta') \cdot sgn(\eta - \eta') \tag{14}$$

where Am_{gr} is a gradients' amplitude.

In analyzing the stability of the ASG, we employ a Lipschitz analysis to determine the value of the amplitude term, denoted as Am_{gr}. This analysis is crucial to demonstrate that ASG yields more consistent gradients for training. We start with an objective function, denoted as f, which should be minimized, and classify it as L-smooth if it satisfies Equation 15:

$$\|\nabla_\zeta f(\zeta + \Delta\zeta) - \nabla_\zeta(\zeta)\| \leq \|L\Delta\zeta\| \tag{15}$$

where ζ represents the related pose estimation network parameters, and ∇ denotes the gradient. The objective function can be reformulated as Equation 16:

$$\nabla_\zeta f = \nabla_\zeta L(\eta, n_s(l)) = \nabla_l L(\eta, n_s(l))\nabla_\zeta l \tag{16}$$

where l denotes the logits predicted by the network, and $\eta' = n_s(l)$ represents the composition of the normalization and soft-argmax functions. Assume the smoothness of the network's gradient and solely focus on the analysis of the composition function, as Equation 17:

$$\|\nabla_l L(\eta, n_s(l+\Delta l)) - \nabla_l L(\eta, n_s(l))\| \tag{17}$$

In the conventional integral regression framework, we have (Equation 18):

$$\nabla_l L(\eta, n_s(l)) = (y-\eta')\cdot P_y \tag{18}$$

In this scenario, Equation 17 can be equated to Equation 19:

$$\|(y-\eta'-\Delta\eta')(P_y+\Delta P_y) - (y-\eta')\cdot P_y\| \tag{19}$$

where y represents an arbitrary position on the heatmap, considering the heatmap size as S, we have $\|y-\eta'\| \le S$ across the entire dataset. Hence, the Lipschitz constant of integral regression is deduced as Equation 20:

$$\begin{aligned} &\|\nabla_l L(\eta, n_s(l+\Delta l)) - \nabla_l L(\eta, n_s(l))\| \\ &\le \|S(P_y+\Delta P_y) - S(P_y)\| \\ &= S\|\Delta P_y\| = S\cdot L_S\cdot\|\Delta l\| \end{aligned} \tag{20}$$

where L_S is the Lipschitz constant [19], and indicates that the traditional integral regression magnifies the Lipschitz normalization constant by S. Likewise, the Lipschitz constant can be computed, and the gradient of the logits is Equation 21:

$$\begin{aligned} |\nabla_l L(\eta, n_s(l))| &= |Am_{gr}\cdot P_y\cdot(1+\sum_{y_k<\eta'} P_{y_k} - \sum_{y_k>\eta'} P_{y_k})| \\ &\le 2\cdot Am_{gr}\cdot P_y \end{aligned} \tag{21}$$

$Am_{gr} = S/8$ ensures that the average gradient norm matches integral regressions. Specifically as Equation 22:

$$E_y[|(y-\eta')P_y|] = E_y[|y-\eta'|]P_y = \frac{S}{4}\cdot P_y \tag{22}$$

The derived Lipschitz constant is (Equation 23):

$$\begin{aligned} &\|\nabla_l L(\eta, n_s(l+\Delta l)) - \nabla_l L(\eta, n_s(l))\| \\ &\le \|2Am_{gr}(P_y+\Delta P_y) - 2Am_{gr}(P_y)\| \\ &= \frac{S}{4}\cdot L_S\cdot\|\Delta l\| \end{aligned} \tag{23}$$

It reveals that when $Am_{gr} = S/8$, our Lipschitz constant is four times smaller than that of the original integral regression. It implies that the gradient space is more uniform, rendering the model more amenable to optimization.

Building upon the above insights, we have successfully integrated and harnessed a 2D HPE tool into our project. As a result, we have attained exceptional accuracy in detecting and locating human key points within 2D space, which will be exploited in the following stages of the proposed approach.

B. CHANNEL-WISE SPATIO TEMPORAL BLOCK

The innovative CWSTB is a pivotal component within the proposed approach, specifically designed for efficient spatial and temporal modeling. This module excels in extracting dynamic spatiotemporal features, showcasing the potential to substantially enhance the performance of temporal-related action recognition while maintaining a minimal computational cost. As illustrated in Figure 7, given an input video data, and through the 1 × 1 2D CNN feature map $F^* \in R^{N\times T\times C\times H\times W}$ is extracted. We first reshape F^* as Equation 24:

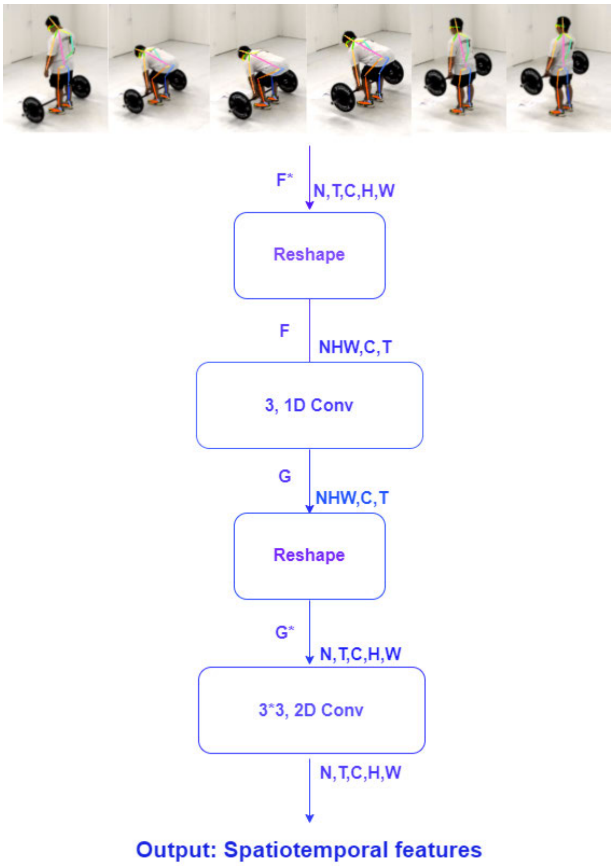


FIGURE 7. Architecture of the proposed CWSTB.

$$F^* \to F \in R^{NHW\times C\times T} \tag{24}$$

Subsequently, we implemented channel-wise 1D convolution along the temporal dimension (T) to integrate temporal information. In the feature map (F), distinct channels often contain varied semantic information, necessitating diverse combinations of temporal information for each channel. To address this, channel-wise convolution is utilized to produce independent kernels. Let G represent a variable denoting the number of groups, equivalent to the number of input channels, distinguishing it from conventional convolution. The channel-wise temporal fusion operation is formulated as Equation 25:

$$G_{c,t} = \sum_i K_i^c F_{c,t+j} \tag{25}$$

where K_i^c represents the temporal combination kernel weights corresponding to channel c, where i denotes the index of the temporal kernel. $F_{c,t+j}$ denotes the input feature sequence, and $G_{c,t}$ represents the resulting updated temporal fusion features. The temporal kernel size is configured to three, implying that i takes values from the set $[-1, 1]$. To restore the shape of G to the original input image shape, we define it as G^* using Equation 26.:

$$G \rightarrow G^* \in R^{N \times T \times C \times H \times W} \tag{26}$$

After that, apply a local-spatial 2D convolution with a kernel size of 3 × 3. Visualize the resulting feature maps of CWSTB to facilitate a better understanding of this module, as illustrated in Figure 7. The experimental findings suggest that CWSTB has effectively learned spatiotemporal features with a heightened focus on the central aspects of actions, such as the hands depicted in the first column. Simultaneously, background features exhibit weaker prominence in the learned representation.

C. DOPPLER MODEL

The Doppler is another essential element of our proposed methodology, and it performs a crucial role by quantizing the motion difference vectors between successive frames into discrete codes. These codes are subsequently embedded within a lower-dimensional space, encapsulating the ongoing action's pose and temporal dynamics. As observed in references [6], [67], [68], [69], [70], and [71], even when separated from spatiotemporal features and directly learned by a 3DCNN from RGB images, the performance can be significantly enhanced through the integration of an optical-flow motion stream. In line with this observation, the Doppler module extracts motion patterns at the feature level from neighboring frames. These extracted features are concatenated with the spatiotemporal and pose-related features and then processed through the C3D network, ultimately enhancing the model's capability. The block derives its name from the Doppler effect, a phenomenon commonly associated with a wave's change in frequency or wavelength as an observer moves relative to the source of the wave. In our context, the proposed module captures the motion dynamics between frames, similar to how the Doppler effect captures the motion of waves in various physical scenarios.

The main objective of the module is to derive an efficient representation of motion for accurate HAR, and it focuses on recognizing actions rather than providing precise motion information between adjacent frames. This strategy allows our approach to solely leverage RGB frames, omitting the need for pre-computed optical flow. This decision reduces computational demands and eliminates the requirement for additional efforts to obtain optical flow data. By relying on RGB frames and building upon prior research detailed in references [72], [73], [74], and [80], we aim to achieve the desired results in HAR.

Like the previously explained CWSTB, the Doppler module also takes input video frame features, passing them through a 1 × 1 2D CNN. This input feature maps $F^{R \times N \times T \times C \times H \times W}$ undergo a 1 × 1 convolution layer to reduce spatial channels by a factor of r, set to 16 in our experiments to mitigate computational load. Subsequently, motion information at the feature level is generated from consecutive feature maps. For instance, considering F_{t-1} and F_t, we apply 2D channel-wise convolution to F_t and subtract it from F_{t-1}, yielding the approximate motion representation $\Delta D_{i,i+1}$ (Equation 27):

$$\Delta D_{i,i+1} = \sum_{i,j} K^c_{i,j} F_{t,c,h+i,w+j} - F_{t-1} \tag{27}$$

where t, h, and w indicate the spatial dimensions of the feature map, respectively, and $K^c_{i,j}$ signifies the c-th motion filter with the spatial index of the kernel. The kernel size is set at 3 × 3, resulting in $i, j \; \in$, the range of -1 to $+1$. This process is applied to every two adjacent feature maps across the temporal dimension (F_{t-1} and F_t, F_t and F_{t+1}, etc.), generating in total $(T - 1)$ motion representations. In the last step, we employ zeros to signify movement absence to ensure temporal compatibility with input feature maps.

These representations are then concatenated along the temporal dimension. Subsequently, another 1×1.2D convolution layer is applied to restore the channel count to C. Despite its apparent simplicity, this approach effectively enhances the overall model's performance, demonstrating that motion features obtained via the Doppler module complement estimated human joint data. Experimental outcomes illustrate that the Doppler module efficiently captures motion features with distinct edges compared to similar approaches. In the final phase of the pipeline, the outcomes from both proposed blocks are combined. These outcomes are then joined with the 2D human key points before entering the action recognition stage. This fusion of diverse information sources is anticipated to heighten the model's accuracy, as experimental results corroborate. The model, thus enriched, capitalizes on spatial-temporal and motion features extracted from video data, coupled with human pose information, to achieve improved HAR performance.

Figure 9 gives a clearer picture of the obtained data from the proposed approach. It can be seen that each model extracts the related features from the provided data, and obtained data are concatenated before joining with pose data and going through the 3DCNN model to be classified.

D. "DOPPLER-DRIVEN DYNAMICS"

To advance the current SOTA, we have elected to employ the Pose Convolutional 3D (PoseC3D) network as our base model, acknowledged as the pinnacle of performance within the pertinent field. Our objective is to amplify computational efficiency and optimize model performance by assimilating the proposed novel components, as elucidated in the preceding subsections IV-A, IV-B, and IV-C. These modules were chosen for their valuable features and are expected to increase the total performance while decreasing the massive computation on 3DCNNs while considering the trade-off between model complexity and computational efficiency. The chosen base model represents a sophisticated variant of the C3D architecture, meticulously designed to integrate human pose information seamlessly. This architectural distinction

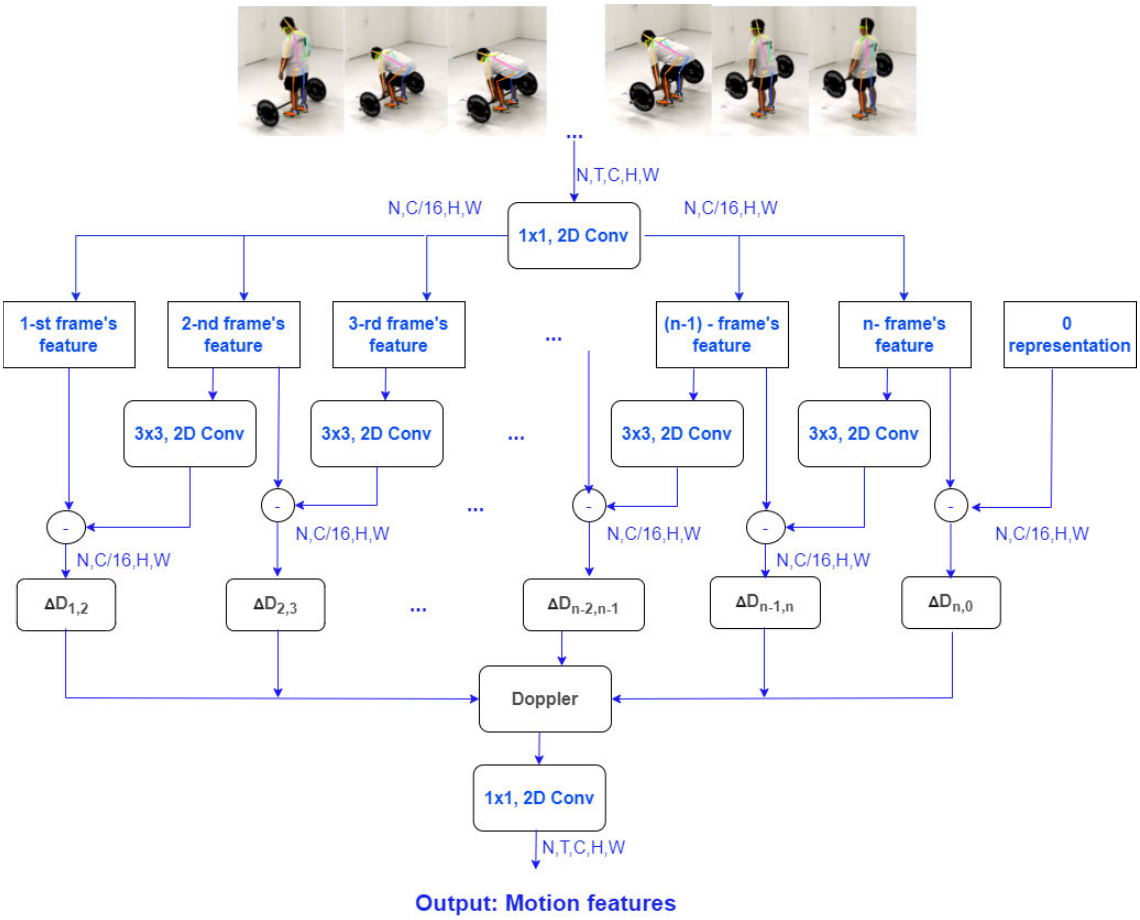


FIGURE 8. The architecture of the proposed Doppler module. The feature maps are shown in the shape of their tensors. ⊖ denotes element-wise subtraction, while $\Delta D_{i,i+1}$ is the difference between motion features of adjacent frames i and i+1, and Doppler is their combination through the whole video.

is visually depicted in Figure 10. The model capitalizes on a combination of 3D convolutional layers and max-pooling operations, adeptly extracting spatiotemporal features inherent within video sequences. These extracted features undergo a flattening process and traverse through a sequence of fully connected layers, culminating in identifying and categorizing human actions. In the context of our study, PoseC3D was earmarked as the baseline model due to its intrinsic compatibility with our data, which amalgamates three distinct feature types. Specific model layers underwent meticulous modification to ensure optimal alignment with our dataset. The objective of these modifications was twofold: firstly, to curtail the overall parameter count, thereby enhancing computational efficiency, and secondly, to maintain, if not augment, the model's performance metrics. A comprehensive exposition of these modifications and their rationales will be expounded upon in the ensuing sections of this manuscript.

1) DATA PREPARATION

The input data under consideration pertains to video data, which necessitates preprocessing techniques to enhance its utility. Two primary strategies are employed for this purpose: centered cropping and uniform sampling. Centered cropping is crucial when the relevant activities occur within a confined spatial region. This technique enables the reduction of spatial volumes while retaining essential features and associated motion data. On the other hand, uniform sampling is adept at preserving the global dynamics of the video. Both methodologies have exhibited significant advantages in skeleton-based action recognition, as validated through empirical experimentation. Notably, the application of uniform sampling has yielded notable improvements in action recognition across various datasets, surpassing traditional fix-stride sampling methods by a substantial margin, as evidenced in Table 6.

This superiority is particularly evident when dealing with datasets characterized by highly variable video lengths. However, uniform sampling's efficacy may diminish under certain conditions, as observed in the context of Kinetics 400 dataset, where slight performance degradation was encountered due to specific input length configurations [1]. Following the implementation of uniform sampling and object-centered cropping, relevant features are extracted

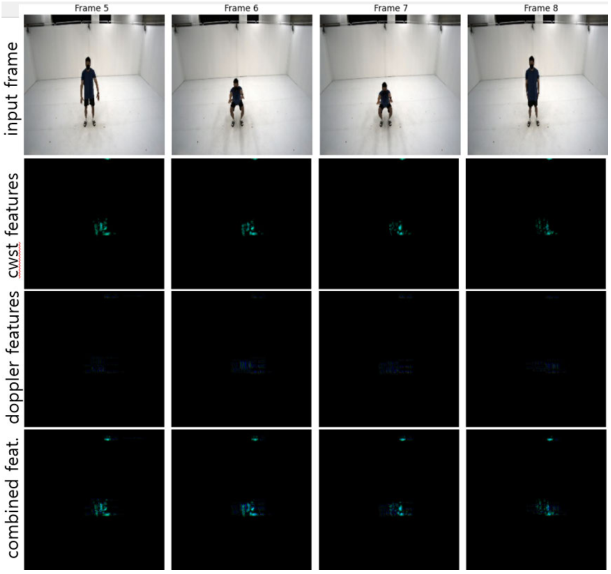


FIGURE 9. Data visualization. The first line is input data, the second line is spatiotemporal features, the third line is Doppler features, and the last one is combined features.

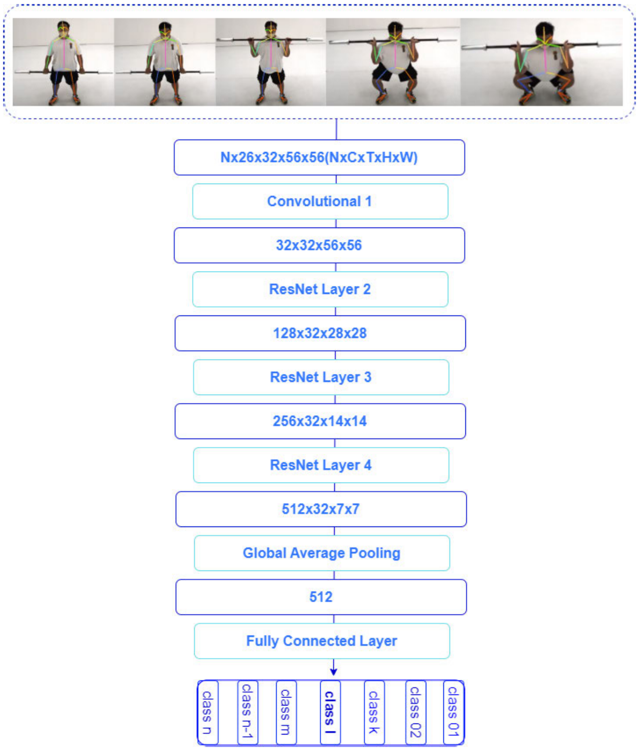


FIGURE 10. Updated architecture of the utilized 3D CNN block.

and structured within a 5D array, representing a single video batch. The standardization of image sizes to 224 × 224 dimensions is recommended, aligning with the requisite format of (N, T, H, W, C) = x (1, 32, 224, 224, 3). Here, N denotes the batch size corresponding to the number of video files, T signifies the number of frames, and H, W, and C denote the dimensions of the frames in terms of height, width, and channels, respectively. Subsequently, data normalization via the MinMaxScaler operation is performed to ensure uniformity of data scales, a prerequisite for DL-based approaches. The data undergoes conversion to the appropriate data type, such as np.uint8, before undergoing BGR conversion, as necessitated by the OpenCV library's BGR color format.

The processed data is then primed for further analysis and forwarded to relevant models, including Doopler, CWSTB, and 2D key extraction blocks. The inputs and outputs of these models are subsequently concatenated before traversing through the 3DCNN network. Experimental findings have indicated that to mitigate computational overhead, the optimal shape for the 3DCNN model is represented by (N, C, T, H, W) = (N, 26, 32, 56, 56). Consequently, data reshaping is undertaken to conform to this requisite format before the data is fed into the 3DCNN-based HAR network. The ultimate output of this process is the classification of the action performed by the subject depicted in the input data.

2) 3D HEATMAP

We have harnessed a 3D heatmap that combines human skeletons, spatiotemporal, and Doppler features from the video frames using our innovative blocks, illustrating the flexibility of our approach. The estimated heatmaps are stored as coordinate triplets (x, y, c), where c marks the maximum score of the heatmap and (x, y) is the corresponding coordinates. Experiments show that coordinate triplets (x, y, c) help save the most storage at the cost of a little performance drop. Here (x, y) are the coordinates of the related joint. At the same time, c is the score, which determines the probability of how correctly the human joint was estimated, and its value is usually between 0 and 1. Technically, we express input data before its preprocessing into 3D heatmap volume as a K × H × W heatmap, where K is the number of joints and H and W are the frame's height and width. We used the top-down posture estimator's heatmap as the target heatmap, which should be zero-padded to match the original frame given the corresponding bounding box. We have created a joint heatmap J_h by assembling K Gaussian maps centered at each joint (Equation 28):

$$J_{h_{nij}} = exp(-\frac{(i-x_n)^2+(j-y_n)^2}{2*\sigma^2})*c_n \tag{28}$$

σ controls the variance of Gaussian maps, and (x_n, y_n) and c_n are the location and confidence score of the n-th joint. We also created a limb heatmap L_h, which exploits the Equation 29:

$$L_{h_{nij}} = exp(-\frac{Dist(i,y),[\alpha_n,\beta_n]^2}{2*\sigma^2})*\min(c_{\alpha_n},c_{\beta_n}) \tag{29}$$

The n-th limb is located between the α_n and β_n joints. Dist computes the distance between the points (i,j) and the segment $[(x\alpha_n, y\alpha_n), (x\beta_n, y\beta_n)]$. Although the preceding technique assumes a single person in each frame, we can easily adapt it to the multi-person scenario by directly accumulating the n-th Gaussian maps of all persons without extending the heatmap. Ultimately, a 3D heatmap volume with the size of $K \times T \times H \times W$ is created by stacking all heatmaps (J_h or L_h) along the temporal dimension.

E. LOSS FUNCTIONS

In this study, we have exploited several loss functions, which will be described in detail in this subsection.

1) TRIPLET LOSS

Triplet loss was used to concatenate the obtained features, such as spatiotemporal, motion features, and 2D human keypoints from the given input data to make proper input data for the model. Triplet loss is a widespread function used in metric learning tasks, particularly in action, facial recognition, and person re-identification tasks. It aims to learn a feature space where samples from the same class are close together while samples from different classes are far apart. Let's denote:

- A_{st}, P_{st}, N_{st} as the spatiotemporal feature embeddings for the anchor, positive, and negative samples, respectively;
- A_{dop}, P_{dop}, N_{dop} as the motion feature embedding obtained from the Doppler module for the anchor, positive, and negative samples, respectively;
- A_{key}, P_{key}, N_{key} are the 2D human key point data for the anchor, positive, and negative samples.

Triplet loss, in our case, was expressed as follows: An anchor sample A, a positive sample P that belongs to the same class as the anchor, and a negative sample N that belongs to a different class from the anchor. The concatenated embeddings for each sample were represented as Equation 30:

$$\begin{aligned} A &= (A_{st}, A_{dop}, A_{key}) \\ P &= (P_{st}, P_{dop}, P_{key}) \\ N &= (N_{st}, N_{dop}, N_{key}) \end{aligned} \tag{30}$$

Now, let's define the distance between embeddings using the L2 (Euclidean) norm (Equations 31, 32):

$$D_{pos} = \|A - P\|_2 \tag{31}$$

$$D_{neg} = \|A - N\|_2 \tag{32}$$

Finally, the triplet loss can be formulated as follows (Equation 33):

$$L_{trip} = max(D_{pos} - D_{neg} + (margin, 0)) : \tag{33}$$

where *margin* is a hyperparameter representing the minimum desired difference between the distances of positive and negative samples from the anchor. If the difference between D_{pos} and D_{neg} exceeds the margin, the loss will be 0, indicating that the triplet is correctly ordered. However, if the difference is smaller than the margin, the loss will be greater than 0, encouraging the network to learn more discriminative features. The overall triplet loss for a batch of triplets is often computed as the average of individual triplet losses (Equation 34):

$$L_{over_{trip}} = mean(L_{trip}) : \tag{34}$$

The margin was set to a positive value to ensure the network learns more robust and separable feature representations. The selection of an appropriate margin value is often based on empirical validation for the specific task at hand; in our case, the optimal margin is equal to 0.5.

2) CATEGORICAL CROSS-ENTROPY LOSS

3DCNN combines all input data as a sequence of human joint positions over time. Its kernels capture essential motion features, generating an output $f_M(M_{1:T}) \in \mathbb{R}^C$, where C is the number of output channels. The temporal stream is adapted to incorporate spatial and temporal features through 2D and 3D convolutions. A 2D CNN followed by a 3D one captures spatial and temporal dynamics, resulting in an output $f_S(S_{1:T}) \in \mathbb{R}^C$. The outputs from the three streams are combined using a fusion method, to create a feature vector f_{comb}. A softmax function predicts the probability distribution over action labels based on this feature vector. The categorical cross-entropy loss is used as the objective function (Equation 35).

$$L_{cat_{cross}} = -1/N \sum_{i=1}^{N} \sum_{y=1}^{C} y_{i,y} log(p_i y | f_{comb}) \tag{35}$$

where N is the number of training samples,

Our architectural intricacy is perceptible in Figure 11, where the Doppler-driven block is essential, which reduces the computational expenses for 3DCNN-based networks while working with massive datasets. Notably, it comprises two constituent modules, namely the CWSTB and the Doppler module, whose outputs are seamlessly integrated within a unified block, culminating in concatenating their respective spatiotemporal and Doppler features. The amalgamation of disparate modalities necessitates a judicious selection of fusion techniques. Herein, we confront the duality of concatenation and summation as viable aggregation mechanisms. Through rigorous experimentation, as corroborated by previous studies [73], [75], summation emerges as the superior fusion strategy for harmonizing the CWSTB and Doppler module outputs. Consequently, an element-wise summation operation is employed to amalgamate the output data before traversing through the C3D network.

Furthermore, to gauge the intrinsic coherence between the CWSTB and Doppler module outputs, we employ a cosine similarity metric, thereby affording insights into the degree of alignment between the encoded features. This meticulous, analytical approach underscores our commitment to both methodological rigor and empirical validation within the framework of our proposed architectural paradigm. Let us use the variables A and B for them, respectively. So the

TABLE 6. Implementing results of uniform sampling technique on the given datasets.

Sampling	NTU-60	FineGYM	CrossFit		Figure Skating	
			normal	full	normal	full
16x2	94.9	87.9	97.2	96.4	96.1	95.4
16x4	95.1	88.7	98.1	97.5	97.8	96.1
uniform-16(1c)	95.7	91.1	99.4	98.4	98.1	97.1
uniform-16	**96.1**	**91.6**	**99.9**	**99.3**	**99.1**	**98.2**

mathematical explanation will be as follows (Equation 36):

$$d = \frac{A \cdot B}{max(\|A\|_2 \cdot \|B\|_2, \epsilon)} \tag{36}$$

where ϵ =1e-08.

3) BINARY CROSS ENTROPY LOSS

Then the Binary Cross Entropy(BCE) loss is used for the traning model (Equation 37):

$$L_{bin_{cross}}(z, d) = -z \cdot log(d) - (1 - z) \cdot log(1 - d) \tag{37}$$

where z is the label, and if the Doppler and spatiotemporal features are paired, the value of z is one. Otherwise, it is equal to zero.

Our approach leverages the advantages of three types of information: pose, Doppler, and spatiotemporal. Pose information helps with body positioning and configuration, while Doppler and spatiotemporal info capture movement patterns and dynamics. This combination improves recognition accuracy compared to a single modality and addresses computation challenges in 3DCNN.

Total loss overall whole pipeline can be described as given Equation 38:

$$L_{total} = L_{reg} + L_{bin_{cross}} + L_{over_{trip}} + L_{cat_{cross}} \tag{38}$$

This composite loss function reflects the combined effects of individual losses during different pipeline stages. By optimizing this total loss, the proposed model is guided to learn effective and discriminative features from different sources, ultimately leading to improved performance in HAR tasks. Certain data points were excluded in the data preprocessing phase because they didn't perform well with the HPE tools. This exclusion was carried out to enhance the accuracy of the proposed HAR model since the model's performance relies on the quality of HPE results. More details on these operations can be found in Section III and Table5. Figure 11 shows our approach, which outperforms prior methods. It excels at learning Doppler and spatiotemporal features, handles pose estimation inaccuracies well and works across different datasets. It's cost-effective for multi-person scenarios and can be combined with other modalities to enhance performance.

V. EXPERIMENT

We conducted exhaustive experiments on well-established benchmark datasets to meticulously evaluate the proposed methodology's efficacy. The datasets encompassed UCF Sports Action (UCF101), Human Motion Database (HMDB51), FineGYM, NTU60, NTU120, and Kinetics 400, renowned for their prominence in action recognition, offering a diverse spectrum of human actions and challenging scenarios for a comprehensive evaluation of our approach. Unless explicitly delineated, the experimental pipeline adhered to the Top-Down approach for pose extraction, wherein the Faster R-CNN architecture, featuring a ResNet50 backbone, served as the detector, complemented by the HRNet pose estimator, pre-trained on the Halpe-FullBody-keypoint dataset. Notably, except for the FineGYM dataset, 2D pose annotations were directly obtained by applying Top-Down pose estimators to RGB inputs. The reported performance metrics encompassed the Mean Top-1 accuracy for the FineGYM dataset and Top-1 accuracy metrics for the remaining datasets, serving as a testament to the meticulousness and comprehensiveness of the experimental evaluation. Central to our experimental framework was using 3DCNN instantiated within the MMAction2 framework, underscoring our commitment to leveraging SOTA approaches in pursuit of methodological rigor and empirical validation.

A. DATASETS

This subsection provides a comprehensive overview of the exploited benchmark datasets from the related research field.

FineGYM is a fine-grained HAR dataset comprising 29,000 videos spanning 99 fine-grained gymnastic action classes. During pose extraction, three types of person bounding boxes are compared: a) bounding boxes predicted by the detector (Detection), b) Ground Truth (GT) bounding boxes for the athlete in the first frame, and tracking boxes for subsequent frames (Tracking), and c) GT bounding boxes for the athlete in all frames (GT). In experiments, human poses extracted with the third type of bounding boxes are used unless otherwise specified.

NTURGB+D, a large-scale human action recognition dataset collected in the lab, has two versions: NTU-60 and NTU-120 (a superset of NTU-60). NTU-60 contains 57,000 videos of 60 human actions, while NTU-120 comprises 114,000 videos of 120 human actions. The datasets are split in three ways: Cross-subject (X-Sub), Cross-view (X-View, for NTU-60), and Cross-setup (X-Set, for NTU-120), where action subjects, camera views, and camera setups differ between training and validation. The dataset provides 3D skeletons collected by sensors. Unless otherwise specified, experiments are conducted on the X-sub splits for NTU-60 and NTU-120.

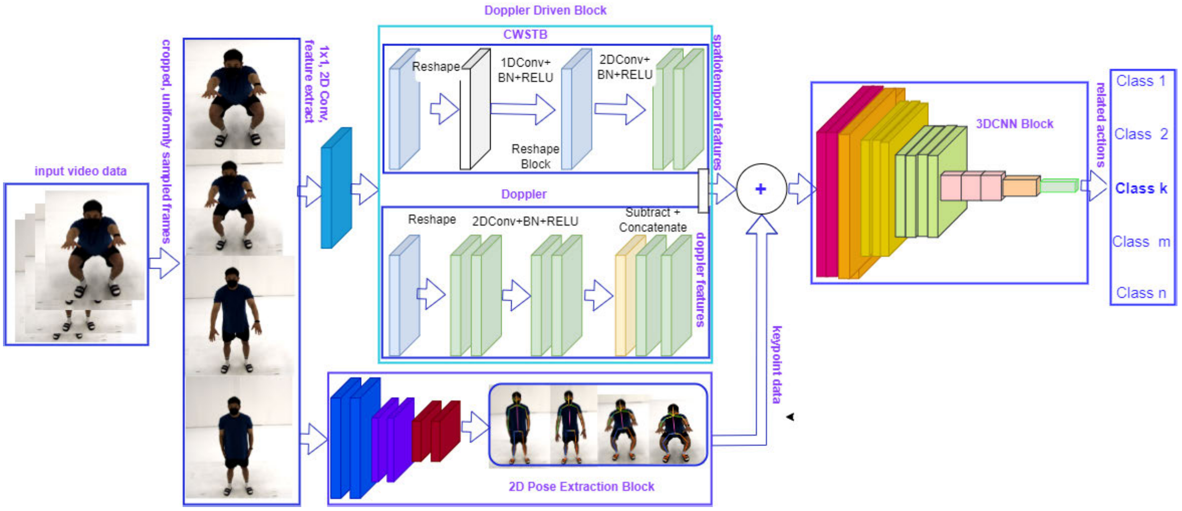


FIGURE 11. Pipeline of the proposed approach.

Kinetics400, UCF101, and HMDB51 are general action recognition datasets sourced from the web. Kinetics400 is a large-scale video dataset featuring 300,000 videos across 400 action classes. UCF101 and HMDB51 are smaller datasets containing 13,000 videos from 101 classes and 6,700 videos from 51 classes, respectively. Experiments are conducted using 2D-pose annotations extracted with the TopDown pipeline.

B. GOOD PROPERTIES OF THE MODEL

We provided several comparisons to elaborate on the good properties of the model over similar approaches. To elaborate on the good properties of 3D convolutional networks over graph networks, we compare the model with PoseC3D, a recent SOTA approach. Two models also input similar data, which is based on coordinate triplets.

1) SUBJECTS-CENTERED CROPPING

Since the sizes and locations of persons can vary a lot in a dataset, focusing on the action subjects is the key to reserving as much information as possible with a relatively small H × W budget. We conduct experiments on FineGYM with input size $32 \times 56 \times 56$, with or without subjects centered cropping to validate this. We find that subjects-centered cropping is helpful in data preprocessing, which improves the 1.1%, from 96.2% to 97.2%.

2) UNIFORM SAMPLING

The input sampled from a small temporal window may not capture the entire dynamic of the human action. To validate this, we conduct experiments on FineGYM and NTU-60. For fixed stride sampling, which samples from a fixed temporal window, we try to sample 32 frames with the temporal stride 2, 3, and 4; for uniform sampling, we sample 32 frames uniformly from the entire clip. In testing, we adopt a fixed random seed when sampling frames from each clip to ensure the test results are reproducible. Figure 6 shows that uniform Sampling consistently outperforms fix-stride Sampling with fixed temporal strides. With uniform Sampling, 1-clip testing can achieve better results than fixed stride sampling with 10-clip testing. Note that the video length can vary a lot in FineGYM and NTU-60. A more detailed analysis found that uniform Sampling mainly improves the recognition performance for longer videos in the dataset.

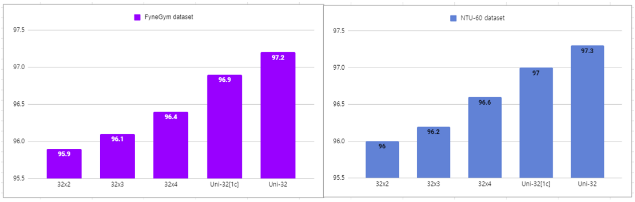


FIGURE 12. Uniform sampling outperforms fix-stride sampling. Left: FineGym dataset, right: NTU-60 dataset results.

3) PERFORMANCE & EFFICIENCY

In the performance comparison, we employ a consistent input shape of $48 \times 56 \times 56$ for both models. Table 7 illustrates that our model is lighter under this configuration than its counterpart, exhibiting fewer parameters and Floating Point Operations (FLOPs). Despite its lightweight nature, the model delivers competitive performance across various datasets. The 1-clip testing result either surpasses or is on par with the SOTA, requiring significantly less computation. During 10-clip testing, it consistently outperforms the SOTA. Moreover, our model uniquely leverages multi-view testing by subsampling all heatmap volumes to construct each input.

4) ROBUSTNESS

To assess the robustness of the model, we conducted experiments involving the random removal of a proportion

TABLE 7. Performance comparision.

Dataset	PoseC3D				DDC3N			
	1-clip	10-clip	Params	FLOPs	1-clip	10-clip	Params	FLOPs
UCF101	97.0	97.2	2.0M	15.9G	**97.1**	**97.4**	**1.98M**	**16.1G**
FyneGYM	92.4	93.2			**92.5**	**93.4**		
NTU-60	93.1	93.7			**96.1**	**95.4**		
NTU-120	85.1	86.0			**92.9**	**93.3**		
Kinematics400	44.8	46.0			**83.9**	**84.3**		

TABLE 8. Robustness test.

Model / P	0	1/8	1/4	1/2	1
PoseC3D	92.4	92.4	92.3	92.1	91.5
+ robust train	92.2	92.1	92.1	92.0	91.4
DDC3N	**92.5**	**92.7**	**92.0**	**91.9**	**92.0**

of keypoints in the input to observe the impact on final accuracy. Given that limb keypoints play a more critical role in gymnastics compared to torso or facial keypoints, we tested both models by randomly dropping one limb keypoint in each frame with a probability denoted as 'p.' Table 8 reveals that DDC3N exhibits high robustness to input perturbations. Specifically, dropping one limb keypoint per frame results in a modest decrease (less than 0.5%) in accuracy. In contrast, for PoseC3D, the accuracy experiences a substantial drop of 0.9%. While training with noisy input akin to the dropout operation, even under this setting, the accuracy of PoseC3D still decreases by 0.5% for the case when 'p' equals 1. Additionally, with robust training, there is an extra 0.2% decrease for the case when 'p' is 0. The experimental results underscore that the proposed approach significantly outperforms its predecessors regarding robustness for HAR.

5) RUNTIME ANALYSIS

When juxtaposed with other techniques, DDC3N has demonstrated state-of-the-art or comparable results across various benchmark datasets. It is a unified CNN framework, eliminating the need for time-intensive optical flow calculations. As illustrated in Table 9, we delineate our model's complexity alongside several SOTA methodologies on the Something-Something v1 dataset. All evaluations were conducted utilizing a single GeForce RTX 3090 GPU. To ensure a consistent comparison, we employed a uniform sampling approach, extracting either 8 or 16 frames from each video followed by center cropping. The notations STM-18, STM-34, and STM-50 denote 18-layer, 34-layer, and 50-layer variants of STM, akin to ResNet18, ResNet-34, and ResNet-50, respectively. TSN 8F and PoseC3D utilize the conventional ResNet-50 as their backbone, while the latter is our baseline model. DDC3N incurs a minimal additional computational overhead (1.3%, 16.1 G FLOPs vs. 15.9 G FLOPs) and parameters (1.0%, 1.98 M vs. 2.0 M) compared to PoseC3D, yet it exhibits enhanced performance. In contrast to I3D 64F and ECO 16F, our 8F model necessitates 19x and 4x fewer FLOPs (16.1 G vs. 306 G, 64 G) and operates at speeds 9.8x and 1.4x faster (106.9 V/s vs. 11.0 V/s, 79.7 V/s). Relative to TSM 16F, our 8F model achieves a 1.7x speed enhancement while halving the FLOPs. Concerning TEA 8F, our 8F model operates at a speed 1.81x faster than TEA (106.9 V/s vs. 59.5 V/s). These results are significant in the field of CV and ML as they demonstrate the superior performance of DDC3N compared to other methodologies. Notably, our model surpasses all STM and PoseC3D variants regarding FLOPs and speed, exhibiting more than a 1.3% improvement in FLOPs while maintaining a 1% reduction in parameters.

TABLE 9. Runtime analysis.

Model	Frame	FLOPs	Params	Speed
I3D [57]	64	306G	28.0M	11.0V/s
ECO [36]	16	64G	47.5M	79.7V/s
TSM [36]	8	32.9G	24.3M	120.3V/s
TSM [36]	16	65.8G	24.3M	63.5V/s
TEA [36]	8	35G	-	59.5V/s
TSN [73]	8	32.9G	23.8M	121.2V/s
STM-18 [36]	8	14.6G	11.0M	161.6V/s
STM-34 [36]	8	29.4G	20.5M	155.3V/s
STM-50 [36]	8	33.3G	24.0M	106.7V/s
STM-50 [36]	16	66.5G	24.0M	52.5V/s
PoseC3D [1]	8	15.9G	2.0M	106.8V/s
PoseC3D [1]	16	15.9G	2.0M	54.8V/s
DDC3N	8	**16.1G**	**1.98M**	**106.9V/s**
DDC3N	16	**16.1G**	**1.98M**	55.9V/s

6) SCALABILITY

The computational efficiency of the GCN-based MS-G3D diminishes proportionally with the growing number of persons in a video, adversely impacting its efficacy in group activity recognition. This assertion is substantiated through experimentation on a pertinent video dataset, where each video comprises 13 persons and spans 20 frames. In the case of MS-G3D, the input shape expands to 13 × 20 × 26 × 3, a size 13 times larger than that for an individual. The GCN incurs a substantial parameter count and FLOPs in this configuration, amounting to 2.8M and 7.2G (×13), respectively. Contrastingly, the proposed model leverages a singular heatmap volume with dimensions 26 × 12 × 56 × 56, effectively representing all 13 persons. Notably, the model maintains a modest base channel width of 16, resulting in a mere 0.52M parameters and 1.6 GFLOPs. Despite the marked reduction in parameters and FLOPs, our approach attains an impressive 95.2% accuracy on the validation set, surpassing the GCN-based MS-G3D by 1.9%.

C. IMPLEMENTATION DETAILS

1) NETWORK DETAILS

Upon receiving an input video, our initial step involves partitioning it into T segments, each of uniform duration, facilitating comprehensive modeling of long-range

TABLE 10. Comparison results with the SOTA models in the HAR field on some familiar benchmark datasets, including our newly created ones.

Models	**Metrics**	**Accuracy %**									
								CrossFit		**Figure Skating**	
		UCF101	**HMDB51**	**FineGYM**	**NTU60**	**NTU120**	**Kinetics400**	**Normal**	**Full**	**Normal**	**Full**
I3D [57]	Top-1	0.9250	0.8270	0.9230	0.7810	0.7900	0.7490	0.9700	0.9689	0.9705	0.9660
	Top-3	0.9400	0.8300	0.9250	0.7820	0.7980	0.7510	0.9716	0.9700	0.9720	0.9686
	Top-5	0.9640	0.8320	0.9260	0.7830	0.8000	0.7520	0.9740	0.9721	0.9735	0.9711
	Mean Class	0.9420	0.8300	0.9240	0.7820	0.7930	0.7510	0.9724	0.9703	0.9720	0.9686
X3D [56]	Top-1	0.9300	0.8520	0.9620	0.9000	0.8080	0.8040	0.9761	0.9742	0.9750	0.9690
	Top-3	0.9350	0.8560	0.9640	0.9090	0.8120	0.8490	0.9796	0.9779	0.9765	0.9706
	Top-5	0.9400	0.8600	0.9660	0.9100	0.8280	0.9460	0.9810	0.9800	0.9790	0.9711
	Mean Class	0.8560	0.8560	0.9640	0.9060	0.8140	0.8690	0.9782	0.9770	0.9775	0.9705
MS-G3D [55]	Top-1	0.9600	0.8020	0.9290	0.9600	0.8690	0.7200	0.9810	0.9800	0.9800	0.9700
	Top-3	0.9650	0.8035	0.9320	0.9615	0.8780	0.7220	0.9830	0.9820	0.9820	0.9710
	Top-5	0.9710	0.8050	0.9350	0.9680	0.8840	0.7290	0.9850	0.9830	0.9850	0.9721
	Mean Class	0.9655	0.8035	0.9320	0.9620	0.8470	0.7230	0.9830	0.9817	0.9830	0.9712
CTR-GCN [79]	Top-1	0.9650	0.8405	0.9245	0.9510	0.9010	0.7245	0.9861	0.9835	0.9815	0.9705
	Top-3	0.9620	0.8455	0.9325	0.9720	0.9120	0.7265	0.9880	0.9855	0.9830	0.9745
	Top-5	0.9650	0.8505	0.9345	0.9790	0.9190	0.7295	0.9945	0.9905	0.9850	0.9775
	Mean Class	0.9625	0.8455	0.9310	0.9680	0.9060	0.7270	0.9895	0.9865	0.9825	0.9740
InfoGCN [78]	Top-1	0.9700	0.8400	0.9240	0.9710	0.9110	0.7240	0.9856	0.9845	0.9810	0.9710
	Top-3	0.9720	0.8450	0.9320	0.9770	0.9130	0.7260	0.9875	0.9855	0.9845	0.9750
	Top-5	0.9750	0.8500	0.9340	0.9780	0.9280	0.7290	0.9950	0.9915	0.9860	0.9785
	Mean Class	0.9716	0.8450	0.9305	0.9760	0.9160	0.7265	0.9895	0.9870	0.9835	0.9750
HD-GCN [77]	Top-1	0.9705	0.8410	0.9245	0.9720	0.9160	0.6350	0.9850	0.9845	0.9825	0.9715
	Top-3	0.9725	0.8445	0.9320	0.9770	0.9200	0.6380	0.9865	0.9860	0.9830	0.9755
	Top-5	0.9755	0.8505	0.9345	0.9780	0.9250	0.6390	0.9900	0.9905	0.9850	0.9785
	Mean Class	0.9720	0.8453	0.9310	0.9760	0.9205	0.6365	0.9870	0.9870	0.9825	0.9750
Baseline [1]	Top-1	0.9700	0.8400	0.9430	0.9710	0.9230	0.7560	0.9861	0.9840	0.9820	0.9711
	Top-3	0.9720	0.8450	0.9480	0.9770	0.9250	0.7710	0.9880	0.9860	0.9835	0.9751
	Top-5	0.9750	0.8500	0.9510	0.9780	0.9310	0.7890	0.9950	0.9909	0.9855	0.9780
	Mean Class	0.9716	0.8450	0.9470	0.9760	0.9260	0.7765	0.9900	0.9870	0.9830	0.9747
Baseline+Doppler	Top-1	0.9700	0.8460	0.9460	0.9710	0.9280	0.8060	0.9910	0.9850	0.9875	0.9770
	Top-3	0.9703	0.8480	0.9485	0.9720	0.9320	0.8270	0.9930	0.9870	0.9895	0.9786
	Top-5	0.9706	0.8409	0.9520	0.9725	0.9350	0.8390	0.9980	0.9970	0.9910	0.9819
	Mean Class	0.9733	0.8404	0.9482	0.9720	0.9320	0.8225	0.9940	0.9915	0.9890	0.9792
Baseline+CWSTB	Top-1	0.9710	0.8450	0.9470	0.9715	0.9270	0.8370	0.9900	0.9855	0.9870	0.9771
	Top-3	0.9720	0.8455	0.9490	0.9720	0.9330	0.8490	0.9960	0.9945	0.9890	0.9790
	Top-5	**0.9740**	0.8460	0.9525	0.9740	0.9350	0.8600	0.9990	0.9958	0.9940	0.9821
	Mean Class	0.9717	0.8455	0.9495	0.9725	0.9315	0.8490	0.9950	0.9920	0.9900	0.9800
DDC3N	Top-1	**0.9720**	**0.8660**	**0.9500**	**0.9730**	**0.9315**	**0.8490**	**0.9961**	**0.9899**	**0.9890**	**0.9811**
	Top-3	**0.9730**	**0.8680**	**0.9520**	**0.9750**	**0.9350**	**0.8560**	**0.9996**	**0.9909**	**0.9910**	**0.9946**
	Top-5	**0.9740**	**0.8690**	**0.9550**	**0.9760**	**0.9380**	**0.8690**	**0.9999**	**0.9978**	**0.9940**	**0.9971**
	Mean Class	0.9730	0.8670	0.9525	0.9750	0.9350	0.8580	0.9985	0.9931	0.9915	0.9822

temporal structures. Subsequently, we employ a uniform sampling technique to select a single frame from each segment, assembling an input sequence comprising T frames and, consequently, object-centered cropping. In the context of our experimentation, we typically configure T to assume values of either 8 or 16. Additionally, within the Doppler framework, the parameter r is consistently set to 16, aligning with established experimental protocols.

2) TRAINING DETAILS

The entire model was trained using two GeForce RTX 3090 GPUs. Each GPU processed a mini-batch consisting of 8 video clips when T = 8 or 4 when T = 16. In the first stage, we initialized training with a learning rate 0.01 for the Kinetics dataset. We reduced this learning rate by a factor of ten at 30, 40, and 60 epochs and concluded training at 70 epochs. We utilized the ImageNet pre-trained model to initialize these large-scale datasets. For UCF101 and HMDB-51, we employed the Kinetics pre-trained model as initialization and began training with a learning rate of 0.001 for 50 epochs. The learning rate decayed by a factor of 10 every 15 epochs. We employed mini-batch SGD as the optimizer with a momentum of 0.9 and a weight decay of 5e-4. The short side of input frames was fixed at 256, with augmentation applied, and the cropped regions were resized to 224 × 224 for network training. Consequently, the input size of the network was N × T × 3 × 224 × 224, where N represents the batch size, and T denotes the number of sampled frames per video. We set β to 0.0039 for the framework and λ to 0.01. The data augmentation (A) strategy included corner cropping, scale-jittering, horizontal flipping, color jittering, and grayscale. A dropout ratio of 0.5 was identified as optimal and consistently used across experiments. In the second stage, where the 3DCNN network performed action recognition, we reshaped the concatenated outputs of the first-stage blocks into an appropriate shape for the given model. Specifically, we deliberately chose a shape of T × 56 × 56 to accommodate the operations' intricacy. The 3DCNN model was trained for 20 to 500 epochs, depending on the complexity of the datasets and the achieved performance.

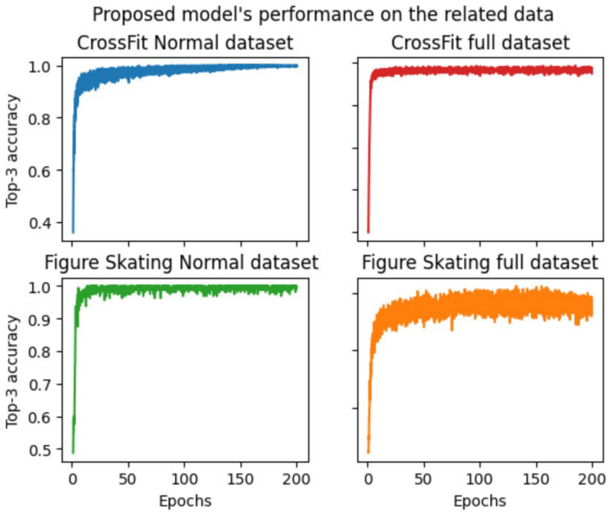


FIGURE 13. Performance of the proposed model on the related datasets.

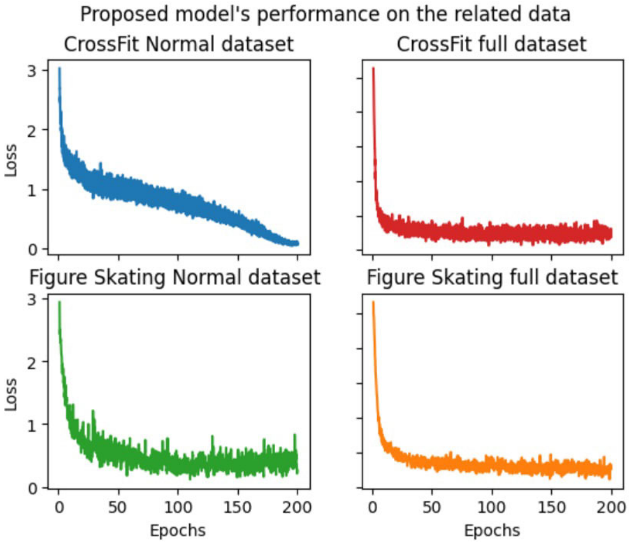


FIGURE 14. Performance of the proposed model on the related datasets in term of loss.

3) COMPREHENSIVE VIEW

Due to time and space constraints in the paper, and also an emphasis on the newly created datasets, a detailed presentation of experimental results is provided exclusively for the newly introduced CrossFit and Figure Skating datasets in Figures 13 and 14. Each experiment is conducted for 200 epochs, a duration determined as optimal through extensive experimentation. Notably, the training performance consistently surpasses the testing performance. Consequently, our focus lies predominantly on presenting and comparing testing performances against SOTA results in the relevant research domain. These accuracy and convergence results are pivotal for evaluating the model's efficacy. Upon closer analysis, the Top accuracy graphs for the CrossFit normal and Figure Skating normal datasets exhibit similar trends. However, discrepancies arise in the full data case. Accuracy consistently registers higher values in normal data scenarios compared to full data scenarios for both the CrossFit and Figure Skating datasets. This divergence can be attributed to including diverse error behavior classes in the full data, presenting additional challenges for the model in effectively recognizing human actions. In the examination of testing loss across the four experiments, it is observed that loss reduction occurs at a slower rate in normal data cases compared to full data cases. Nevertheless, the final loss value is lower in the normal data case. This behavior can be ascribed to the intricacies introduced by various error behavior classes in the full data, posing challenges for the human-action recognition model. To mitigate potential overfitting during the experiments, the model employs cross-entropy loss in its final stage. This strategic choice serves to regularize the model, ensuring effective generalization to unseen data.

D. COMPARISION WITH THE SOTA

To ascertain the efficacy and performance of our proposed methodology, we embarked upon a comprehensive comparative analysis against a cadre of SOTA models that have garnered prominence within the domain of HAR. This comparative roster includes illustrious architectures such as the Inflated 3D ConvNet (I3D), Expanding Architectures for Efficient Video Recognition (X3D), Multi-Scale Gating 3D (MS-G3D), Channel-wise Topology Refinement Graph Convolution (CTR-GC), InfoGCN, Hierarchically Decomposed Graph Convolutional Networks (HD-GCN), and PoseC3D.

Among these, I3D and X3D emerge as preeminent deep neural network (DNN) architectures celebrated for their prowess in feature extraction explicitly tailored for video action recognition tasks. The MS-G3D framework, a dedicated instantiation of 3DCNNs, is characterized by its adeptness in capturing spatiotemporal intricacies inherent within video sequences. Conversely, CTR-GC stands distinguished by its innovative approach of aggregating joint features and integrating temporal modeling modules to tackle the challenges of modeling channel-wise topologies. Meanwhile, InfoGCN introduces a novel fusion of information bottleneck-based learning objectives with attention-based graph convolution, facilitating the effective embedding of skeleton information for action recognition tasks through the strategic utilization of node decomposition and the construction of an attention-guided hierarchy aggregation module. HD-GCN endeavors to extract structurally adjacent and distant edges within graph representations. Finally, PoseC3D adopts an integrative stance by synergistically fusing pose estimation methodologies with 3D CNNs, augmenting action recognition capabilities by incorporating human pose information.

The empirical findings stemming from our extensive experimentation, as delineated in Table 10, unequivocally underscore the superiority and efficacy of our proposed model within the domain of HAR. Leveraging a rigorous experimental methodology, we conducted comprehensive evaluations against a panoply of recent models within the field, diligently replicating implementations where codes were openly available and meticulously analyzing original works in cases where code accessibility was limited. Our proposed methodology consistently surpasses current SOTA models across a spectrum of datasets, demonstrating outstanding Top-1 accuracy metrics. Specifically, we achieve a remarkable accuracy of 97.3% on the NTU60 dataset, closely followed by 97.2% on UCF101. Moreover, our model attains accuracies of 95.0%, 93.2%, 86.0%, and 84.0% on the FineGym, NTU120, HMDB51, and Kinetics400 datasets, respectively, outperforming previous SOTA models which registered lower accuracies. The performance discrepancy observed in the Kinetics400 dataset can be attributed to its non-human-centric nature for action recognition. The variability in person locations, scales, and the number of individuals in Kinetics videos complicates the extraction of human skeletons, thereby affecting performance. In contrast, NTURGB+D and FineGym datasets, which exhibit less variability, yield higher accuracies. Notably, NTURGB+D, comprising NTU60 and NTU120 subsets, achieves higher accuracy in NTU60 due to its smaller action set. HMDB51 also outperforms Kinetics400, despite its complexity, featuring 13K videos across 101 classes and 6.7K videos across 51 classes. Among the benchmarked SOTA models, PoseC3D consistently achieves the highest accuracy, followed by four GCN-based approaches: HD-GCN, InfoGCN, CTR-GCN, and MS-G3D. The 3DCNN-based X3D and I3D models also demonstrate competitive performance. Furthermore, on our custom CrossFit and Figure Skating datasets, our model achieves an accuracy of 98.9%, surpassing the previous SOTA accuracy of 99.0%. When comparing results from the normal and full datasets, our model consistently performs better on the normal dataset. This can be attributed to the increased data complexity in the full dataset, which includes normal and error behavior classes. In summary, our empirical findings substantiate the superior accuracy and efficacy of our proposed approach over competitors across a diverse range of datasets. These results offer invaluable insights into the evolving landscape of SOTA action recognition methodologies.

E. ABLATION STUDY

Conducting an ablation study constitutes a fundamental aspect of our methodological inquiry aimed at discerning the discrete contributions of each constituent component within the ambit of our proposed approach. The rudimentary framework, serving as our baseline model, encompasses a C3D network architecture, which exclusively processes the sequential pose information. Building upon this foundational structure, we endeavored to augment its efficacy by the incremental inclusion of two pivotal components: the Doppler module and the CWSTB. While the former is tasked with encoding motion information, the latter assumes the mantle of elucidating temporal and spatial relationships across all channels. A series of ablation experiments were meticulously conducted to unravel the distinct impacts of these augmentative components:

Baseline: This experiment entailed training the baseline model in its pristine form, thereby eschewing the incorporation of both the Doppler and CWSTB modules.

Baseline + CWSTB: The CWSTB module was introduced into the baseline framework singularly, and the combined model was subsequently trained.

Baseline + Doppler module: In this iteration, the CWSTB module was omitted, and instead, the Doppler module was integrated into the baseline model, with subsequent training.

Complete Model: The zenith of our experimentation, this iteration encapsulated incorporating both the Doppler and CWSTB modules into the baseline model architecture, thereby furnishing a holistic rendition of our proposed approach.

Table 10 furnishes a detailed overview of the outcomes derived from our comprehensive ablation study. Notably, the baseline model exhibits a discernible superiority over alternative methodologies across all evaluated datasets. However, individual integration of the Doppler or CWSTB modules engenders substantial enhancements in performance metrics, outstripping several existing methodologies, including the baseline. This enhancement is further accentuated when these modules are synergistically incorporated into the holistic model framework, culminating in the most optimal results. This underscores the pivotal significance of our proposed approach within the realm of human action recognition tasks.

The Doppler block augmentation elevates performance across the HMDB51, FineGYM, NTU60, NTU120, Kinetics, CrossFit, and Figure Skating datasets. Conversely, the performance on the UCF101 dataset remains invariant. In contrast, when solely incorporating the CWSTB module, excluding Doppler, improved performance across all datasets is observed. However, the performance differential between the baseline and the CWSTB-augmented model is less pronounced than the Doppler-enhanced results. Notably, synergistic integration of Doppler and CWSTB modules into the baseline model yields a marked improvement, substantiating the efficacy of our proposed approach over extant methods.

The ablation study's findings robustly validate the proposed techniques' effectiveness, underscoring their pivotal role in enhancing model accuracy. This empirical substantiation unequivocally establishes the superiority of our approach over previous SOTA methodologies. Such methodological validation accentuates our proposed framework's robustness and practical efficacy, thereby eliciting significant implications for advancing HAR systems in academic and industrial contexts.

VI. DISCUSSION

This investigation propels the field of HAR forward by pioneering a novel approach to extracting motion and spatiotemporal features from video inputs. At the heart of this advancement lie the CWSTB and Doppler modules, seamlessly integrated into the architecture of Conv3D networks. This integration is meticulously crafted to mitigate computational demands while concurrently enhancing accuracy. The prior one is engineered to adeptly capture spatiotemporal information through channel-wise 1D and 2D convolutional layers. Simultaneously, the latter orchestrates the swift extraction of motion information, representing a notable advancement over conventional optical flow methodologies.

A series of meticulously conducted experiments validated the proposed model's efficacy comprehensively. These include a subsection as Robustness (in p. 14), wherein the model's performance without specific pivotal points is juxtaposed against its baseline counterpart. A runtime analysis compares the model against its counterparts with similar parameters and FLOPs, demonstrating its clear superiority. Furthermore, a scalability comparison with the GCN-based MS-G3D model showcases the enhanced performance of our approach.

Moreover, the proposed approach undergoes rigorous training on bespoke datasets curated specifically for this research endeavor and a curated selection of open benchmark datasets. Experiment results indicate that our approach outperforms recent methodologies across exploited benchmark datasets. Incorporating an end-to-end training regimen featuring binary and categorical cross-entropy alongside triplet loss significantly culminates exceptional performance.

However, it is imperative to acknowledge certain limitations inherent to the suggested method. Its reliance on pose information may only capture some facets of an action, including object interactions or scene context. Future research endeavors could address these limitations by integrating additional modalities, such as RGB without skeleton data or depth images, thereby enhancing the recognition of more intricate human actions.

VII. CONCLUSION

The network offers a seminal advancement in HAR by amalgamating the pioneering modules that efficiently capture motion dynamics and spatiotemporal intricacies. It mitigates the computational overhead inherent to traditional 3DCNNs, eclipsing the SOTA benchmarks. The model surpasses its counterparts in FLOPs and speed metrics, notably outclassing the recent PoseC3D by over 1.3 % in FLOPs while boasting a reduction of 1.0 % in parameter count.

Furthermore, developing two novel datasets explicitly tailored for HAR is a testament to the commitment to advancing the field. It furnishes invaluable resources for researchers seeking to explore and evaluate DL-based methodologies. These advancements augur exciting prospects for enhancing the precision and efficiency of HAR systems and heralding a new epoch of innovation. They propel the field toward more sophisticated solutions, thus fortifying its relevance across various disciplines in tandem with the inexorable march of technological evolution.

As a harbinger of future research trajectories, it behooves scholars to address the inherent limitations of the present study and embark upon endeavors aimed at prognosticating impending human actions. These predictions hold immense potential, from identifying rising stars in sports to forecasting game outcomes. This reinforces the transformative power of this network in various human-centered activities.

REFERENCES

[1] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, LA, LA, USA, Jun. 2022, pp. 2969–2978.

[2] R. M. Yusupov, S. A. Tavboyev, M. A. Tashpulatov, and Z. R. Akhmedov, "Linguistic modelling of the theory of indistinct sets as the basis of the estimation of quality of formation," in *Young Scientist USA*, vol. 1. USA: Lulu Press, 2014, ch. 5, pp. 22–27.

[3] T. Mukhiddin, H. R. Arousha, A. Ubaydullo, L. Wookey, and S. Lee, "Privacy-preserving of human identification in CCTV data using a novel deep learning-based method," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Jan. 2022, pp. 211–214, doi: 10.1109/BigComp54360.2022.00048.

[4] 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.

[5] M. Toshpulatov, W. Lee, S. Lee, and A. Haghighian Roudsari, "Human pose, hand and mesh estimation using deep learning: A survey," *J. Supercomput.*, vol. 78, no. 6, pp. 7616–7654, Apr. 2022.

[6] T. Mukhiddin, W. Lee, S. Lee, and T. Rashid, "Research issues on generative adversarial networks and applications," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Feb. 2020, pp. 487–488, doi: 10.1109/BigComp48618.2020.00-19.

[7] M. Toshpulatov, W. Lee, and S. Lee, "Generative adversarial networks and their application to 3D face generation: A survey," *Image Vis. Comput.*, vol. 108, Apr. 2021, Art. no. 104119.

[8] J. Gong, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "Meta agent teaming active learning for pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 11079–11089.

[9] Y. Wang, M. Li, H. Cai, W. Chen, and S. Han, "Lite pose: Efficient architecture design for 2D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13126–13136.

[10] G. Moon, H. Choi, and K. M. Lee, "Accurate 3D hand pose estimation for whole-body 3D human mesh estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2308–2317.

[11] Y. Zhan, F. Li, R. Weng, and W. Choi, "Ray3D: Ray-based 3D human pose estimation for monocular absolute 3D localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13116–13125.

[12] Y. He, Y. Wang, H. Fan, J. Sun, and Q. Chen, "FS6D: Few-shot 6D pose estimation of novel objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6814–6824.

[13] H. Liu, T. Liu, Z. Zhang, A. K. Sangaiah, B. Yang, and Y. Li, "ARHPE: Asymmetric relation-aware representation learning for head pose estimation in industrial human–computer interaction," *IEEE Trans. Ind. Informat.*, vol. 18, no. 10, pp. 7107–7117, Jan. 2022.

[14] Z. Wang, X. Nie, X. Qu, Y. Chen, and S. Liu, "Distribution-aware single-stage models for multi-person 3D pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13096–13105.

[15] N. U. R. Malik, S. A. R. Abu Bakar, and U. U. Sheikh, "Multiview human action recognition system based on openpose and KNN classifier," in *Proc. 11th Int. Conf. Robot., Vis., Signal Process. Power Appl.* Singapore: Springer, 2022, pp. 890–895.

[16] D. Feng, Z. Wu, J. Zhang, and T. Ren, "Multi-scale spatial temporal graph neural network for skeleton-based action recognition," *IEEE Access*, vol. 9, pp. 58256–58265, 2021.

[17] S. Juraev, A. Ghimire, J. Alikhanov, V. Kakani, and H. Kim, "Exploring human pose estimation and the usage of synthetic data for elderly fall detection in real-world surveillance," *IEEE Access*, vol. 10, pp. 94249–94261, 2022.

[18] J. Park, S. Cho, D. Kim, O. Bailo, H. Park, S. Hong, and J. Park, "A body part embedding model with datasets for measuring 2D human motion similarity," *IEEE Access*, vol. 9, pp. 36547–36558, 2021.

[19] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7157–7173, Jun. 2023.

[20] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," 2018, *arXiv:1812.08008*.

[21] A. Rohan, M. Rabah, T. Hosny, and S.-H. Kim, "Human pose estimation-based real-time gait analysis using convolutional neural network," *IEEE Access*, vol. 8, pp. 191542–191550, 2020.

[22] X. Zhang, X. Wang, and R. Zhang, "Dynamic semantics SLAM based on improved mask R-CNN," *IEEE Access*, vol. 10, pp. 126525–126535, 2022.

[23] M. Toshpulatov, W. Lee, and S. Lee, "Talking human face generation: A survey," *Expert Syst. Appl.*, vol. 219, Jun. 2023, Art. no. 119678.

[24] D. Groos, H. Ramampiaro, and E. A. F. Ihlen, "EfficientPose: Scalable single-person pose estimation," *Appl. Intell.*, vol. 51, pp. 2518–2533, Nov. 2021.

[25] Y. Li, D. Yang, Y. Chen, C. Peng, Z. Sun, and L. Jiao, "A lightweight top-down multi-person pose estimation method based on symmetric transformation and global matching," *IEEE Access*, vol. 10, pp. 22112–22122, 2022.

[26] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3D human pose estimation with spatial and temporal transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11656–11665.

[27] J. Janardhanan and S. Umamaheswari, "A comprehensive study on human pose estimation," in *Proc. 8th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, vol. 1, Mar. 2022, pp. 535–541.

[28] S. Niu, W. Ou, S. Feng, J. Gou, F. Long, W. Zhang, and W. Zeng, "Designing compact convolutional filters for lightweight human pose estimation," *Wireless Commun. Mobile Comput.*, vol. 2021, no. 1, 2021, Art. no. 1333250.

[29] Y. Yang, Z. Ren, H. Li, C. Zhou, X. Wang, and G. Hua, "Learning dynamics via graph neural networks for human pose estimation and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8074–8084.

[30] N. Garau, N. Bisagno, P. Bródka, and N. Conci, "DECA: Deep viewpoint-equivariant human pose estimation using capsule autoencoders," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11677–11686.

[31] H. Chen, M. Li, L. Jing, and Z. Cheng, "Lightweight long and short-range spatial–temporal graph convolutional network for skeleton-based action recognition," *IEEE Access*, vol. 9, pp. 161374–161382, 2021.

[32] C. Zhao, M. Chen, J. Zhao, Q. Wang, and Y. Shen, "3D behavior recognition based on multi-modal deep space-time learning," *Appl. Sci.*, vol. 9, no. 4, pp. 716–722, 2019.

[33] C. Ding, S. Wen, W. Ding, K. Liu, and E. Belyaev, "Temporal segment graph convolutional networks for skeleton-based action recognition," *Eng. Appl. Artif. Intell.*, vol. 110, Apr. 2022, Art. no. 104675.

[34] H. Kataoka, K. Hara, R. Hayashi, E. Yamagata, and N. Inoue, "Spatiotemporal initialization for 3D CNNs with generated motion patterns," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1279–1288.

[35] M. Faraji, S. Nadi, O. Ghaffarpasand, S. Homayoni, and K. Downey, "An integrated 3D CNN-GRU deep learning method for short-term prediction of PM2.5 concentration in urban environment," *Sci. Total Environ.*, vol. 834, pp. 155324–155331, Aug. 2022.

[36] M. Wang, J. Xing, J. Su, J. Chen, and Y. Liu, "Learning spatiotemporal and motion features in a unified 2D network for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3347–3362, Mar. 2023.

[37] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, and J. Wang, "Lite-HRNet: A lightweight high-resolution network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10440–10450.

[38] A. Zeng, X. Ju, L. Yang, R. Gao, X. Zhu, B. Dai, and Q. Xu, "DeciWatch: A simple baseline for 10× efficient 2D and 3D pose estimation," in *Proc. 17th Eur. Conf. Comput. Vis.*, 2022, pp. 607–624.

[39] L. Ma, L. Liu, C. Theobalt, and L. V. Gool, "Direct dense pose estimation," in *Proc. Int. Conf. 3D Vis. (3DV)*, Dec. 2021, pp. 721–730.

[40] Z. Yu, B. Ni, J. Xu, J. Wang, C. Zhao, and W. Zhang, "Towards alleviating the modeling ambiguity of unsupervised monocular 3D human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8651–8660.

[41] V. Narang and A. Solanki, "An efficient algorithm for human abnormal behaviour detection using object detection and pose estimation," in *Proc. 5th Int. Conf. Saf. Secur. IoT (SaSeIoT)*, 2022, pp. 47–64.

[42] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3200–3225, Mar. 2023.

[43] J. Cai, H. Liu, R. Ding, W. Li, J. Wu, and M. Ban, "HTNet: Human topology aware network for 3D human pose estimation," 2023, *arXiv:2302.09790*.

[44] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.

[45] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 591–600.

[46] P. Wang, "Research on sports training action recognition based on deep learning," *Sci. Program.*, vol. 2021, no. 1, 2021, Art. no. 3396878.

[47] B. Zhang, J. Yu, C. Fifty, W. Han, A. M. Dai, R. Pang, and F. Sha, "Co-training transformer with videos and images improves action recognition," 2021, *arXiv:2112.07175*.

[48] O. Moutik, H. Sekkat, S. Tigani, A. Chehri, R. Saadane, T. A. Tchakoucht, and A. Paul, "Convolutional neural networks or vision transformers: Who will win the race for action recognitions in visual data?" *Sensors*, vol. 23, no. 2, p. 734, Jan. 2023.

[49] M. Wu, B. Jiang, D. Luo, J. Yan, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and X. Yang, "Learning comprehensive motion representation for action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 2934–2942.

[50] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, pp. 1–74, Mar. 2021.

[51] H. Xia and X. Gao, "Multi-scale mixed dense graph convolution network for skeleton-based action recognition," *IEEE Access*, vol. 9, pp. 36475–36484, 2021.

[52] H. Ou and J. Sun, "Multi-scale spatialtemporal information deep fusion network with temporal pyramid mechanism for video action recognition," *J. Intell. Fuzzy Syst.*, vol. 41, no. 3, pp. 4533–4545, 2021.

[53] W. Yang, T. Zhang, Z. Mao, Y. Zhang, Q. Tian, and F. Wu, "Multi-scale structure-aware network for weakly supervised temporal action detection," *IEEE Trans. Image Process.*, vol. 30, pp. 5848–5861, 2021.

[54] Z. Li, J. Li, Y. Ma, R. Wang, Z. Shi, Y. Ding, and X. Liu, "Spatio-temporal adaptive network with bidirectional temporal difference for action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 5174–5185, Sep. 2023, doi: 10.1109/TCSVT.2023.3250646.

[55] 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.

[56] C. Feichtenhofer, "X3D: Expanding architectures for efficient video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 203–213.

[57] G. Liu, C. Zhang, Q. Xu, R. Cheng, Y. Song, X. Yuan, and J. Sun, "I3D-shufflenet based human action recognition," *Algorithms*, vol. 13, no. 11, p. 301, Nov. 2020.

[58] F. Iodice, E. De Momi, and A. Ajoudani, "HRI30: An action recognition dataset for industrial human–robot interaction," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 4941–4947.

[59] Z. Li, L. He, and H. Xu, "Weakly-supervised temporal action detection for fine-grained videos with hierarchical atomic actions," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 567–584.

[60] M. Chen, F. Wei, C. Li, and D. Cai, "Frame-wise action representations for long videos via sequence contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13801–13810.

[61] A. B. Tanfous, A. Zerroug, D. Linsley, and T. Serre, "How and what to learn: Taxonomizing self-supervised learning for 3D action recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 2888–2897.

[62] U. Asif, D. Mehta, S. Von Cavallar, J. Tang, and S. Harrer, "DeepActsNet: A deep ensemble framework combining features from face, hands, and body for action recognition," *Pattern Recognit.*, vol. 139, Jul. 2023, Art. no. 109484.

[63] Y. Liu, J. Yuan, and Z. Tu, "Motion-driven visual tempo learning for video-based action recognition," *IEEE Trans. Image Process.*, vol. 31, pp. 4104–4116, 2022.

[64] X. Yin, S. Wu, Y. Liu, Z. Qin, L. Bi, and R. Fan, *Underwater Target Tracking Algorithm Based on Optical Flow*. Singapore: Springer, 2022, pp. 25–34.

[65] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7912–7921.

[66] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3316–3333, Jun. 2022.

[67] J. Tu, M. Liu, and H. Liu, ‘‘Skeleton-based human action recognition using spatial temporal 3D convolutional neural networks,’’ in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[68] X. Ji, J. Cheng, W. Feng, and D. Tao, ‘‘Skeleton embedded motion body partition for human action recognition using depth sequences,’’ *Signal Process.*, vol. 143, pp. 56–68, Feb. 2018.

[69] M. Alibayev, D. Paulius, and Y. Sun, ‘‘Estimating motion codes from demonstration videos,’’ in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 4257–4262.

[70] D. Paulius, N. Eales, and Y. Sun, ‘‘A motion taxonomy for manipulation embedding,’’ 2020, *arXiv:2007.06695*.

[71] S. Yang, W. Heng, G. Liu, G. Luo, W. Yang, and G. Yu, ‘‘Capturing the motion of every joint: 3D human pose and shape estimation with independent tokens,’’ 2023, *arXiv:2303.00298*.

[72] M. Dong, Z. Fang, Y. Li, S. Bi, and J. Chen, ‘‘AR3D: Attention residual 3D network for human action recognition,’’ *Sensors*, vol. 21, no. 5, p. 1656, Feb. 2021.

[73] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, ‘‘STM: SpatioTemporal and motion encoding for action recognition,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2000–2009.

[74] H. Yang, Z. Ren, H. Yuan, W. Wei, Q. Zhang, and Z. Zhang, ‘‘Multi-scale and attention enhanced graph convolution network for skeleton-based violence action recognition,’’ *Frontiers Neurorobot.*, vol. 16, Dec. 2022, Art. no. 1091361.

[75] A. Abdelbaky and S. Aly, ‘‘Two-stream spatiotemporal feature fusion for human action recognition,’’ *Vis. Comput.*, vol. 37, p. 1821–1835, Aug. 2021.

[76] J. Huh, J. Chalk, E. Kazakos, D. Damen, and A. Zisserman, ‘‘Epic-sounds: A large-scale dataset of actions that sound,’’ in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[77] J. Lee, M. Lee, D. Lee, and S. Lee, ‘‘Hierarchically decomposed graph convolutional networks for skeleton-based action recognition,’’ 2022, *arXiv:2208.10741*.

[78] H.-G. Chi, M. H. Ha, S. Chi, S. W. Lee, Q. Huang, and K. Ramani, ‘‘InfoGCN: Representation learning for human skeleton-based action recognition,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20186–20196.

[79] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, ‘‘Channel-wise topology refinement graph convolution for skeleton-based action recognition,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13359–13368.

[80] M. Toshpulatov, W. Lee, C. Tursunbaev, and S. Lee, ‘‘Human action recognition utilizing Doppler-enhanced convolutional 3D networks,’’ in *Proc. IEEE Int. Conf. Big Data Smart Comput.*, Feb. 2024, pp. 475–478.

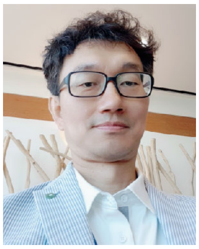
[81] K. Bae, S. Lee, and W. Lee, ‘‘Diffusion-C: Unveiling the generative challenges of diffusion models through corrupted data,’’ 2023, *arXiv:2312.08843*.

[82] J. Kim, W. Lee, J. J. Song, and S.-B. Lee, ‘‘Optimized combinatorial clustering for stochastic processes,’’ *Cluster Comput.*, vol. 20, no. 2, pp. 1135–1148, 2017.

[83] J. Afshar, A. H. Roudsari, and W. Lee, ‘‘Top-*k* team synergy problem: Capturing team synergy based on C3,’’ *Inf. Sci.*, vol. 589, pp. 117–141, 2022.

[84] J. Yoo, J. Kim, H. Yoon, G. Kim, C. Jang, and U. Kang, ‘‘Accurate graph-based PU learning without class prior,’’ in *Proc. Int. Conf. Data Mining (ICDM)*, 2021, pp. 827–836.

MUKHIDDIN TOSHPULATOV (Member, IEEE) received the B.S. and M.S. degrees from Samarkand State University, Uzbekistan, in 2000 and 2002, respectively. He is currently pursuing the Ph.D. degree with the Department of Biomedical Science and Engineering, Inha University, South Korea. His research interests include machine learning and its applications to computer vision, sensor data science, human pose estimation, action recognition, talking human face generation, and natural language processing.



WOOKEY LEE (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Department of Industrial Engineering, Seoul National University, and the M.S.E. degree from Carnegie Mellon University. He is currently a Professor with the Department of BMSE and IE, Inha University, Incheon, South Korea, and the Director of the Voice AI Institute. His research interests include deep learning, meta-learning, patent data, data privacy, and voice AI. He received the best paper awards, such as ACM BigDas, in 2016, 2019, and 2021; IEEE TCSC, in 2012; ACM BigDas; and the Statistics Korea Commissioner Award. He is the Chair of several international conferences, such as ASONAM, BigComp, BigData, CIKM, Concept, Dasfaa, Dexa, ER, ICDE, PAKDD, SAC, SocialCom, and VLDB. He is the EIC of *Big Data Service* journal; a Steering Committee Member of IEEE BigComp; the EC of the IEEE Technical Committee on Data Engineering; and an Associate Editor of *World Wide Web* journal, DKE, SUPE, CLUS, C&IE, and *Sensors*.

SUAN LEE (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from Kangwon National University, South Korea, in 2010 and 2017, respectively. He is currently an Assistant Professor with the School of Computer Science, Semyung University, South Korea. He has developed an in-memory database engine and real-time stream data processing engine as a Researcher with Altibase Company for over three years, since 2012. He was a Senior Researcher with the Information and Communication Research Center, Kangwon National University, in 2018, and a Research Professor/a Visiting Professor with the National Program of Excellence in Software, Kangwon National University, in 2019. He was a Principal Researcher with the Voice AI Institute, Inha University, in 2020. His research interests include machine learning, deep learning, recommender systems, spatio-temporal, time-series, tensors, graphs, language models, and computer vision.

HOYOUNG YOON (Member, IEEE) received the B.S. degree in electronic and electrical engineering (EEE) from the University of Seoul, South Korea, in 2014. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Seoul National University (SNU), South Korea. His research interests include wireless networks, device-to-device (D2D) communication, and vehicle-to-everything (V2X).

U KANG (Member, IEEE) received the B.S. degree in computer science and engineering from Seoul National University and the M.S. and Ph.D. degrees in computer science from Carnegie Mellon University. He is currently an Associate Professor with the Department of Computer Science and Engineering, Seoul National University. He has published over 100 refereed articles in major data mining, database, and machine learning venues. He holds seven U.S. patents. His research interests include data mining, graph mining, and voice meta-learning. He won the 2013 SIGKDD Doctoral Dissertation Award, the 2013 New Faculty Award from Microsoft Research Asia, and the 2016 Korean Young Information Scientist Award. He received the Six Best Paper Awards, including the 2018 ICDM 10-Year Best Paper Award, the 2021 KDD Best Research Paper Award, and the 2022 ICDE Best Research Paper Award.

• • •