IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 February 2026, accepted 25 March 2026, date of publication 6 April 2026, date of current version 9 April 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3681228

RESEARCH ARTICLE

Long-Tailed Skeleton-Based Action Recognition via Realistic Mixing and Class-Aware Sampling

JINWOO KIM[1], WONHEE KIM[1], (Student Member, IEEE), MI-SEON KANG[2], AND JONG TAEK LEE[1], (Member, IEEE)

[1]School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, South Korea
[2]Daegu-Gyeongbuk Research Division, Electronics and Telecommunications Research Institute, Daegu 42664, South Korea

Corresponding author: Jong Taek Lee (jongtaeklee@knu.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant RS-2025-25425223; in part by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (26ZD1120, Advancement and Commercialization of Daegu-Gyeongbuk Regional Strategic Industries (Robots, Mobility, AI, Medical, etc.)); and in part by the Regional Innovation System Education (RISE) Program through Daegu RISE Center funded by the Ministry of Education (MOE) and Daegu Metropolitan City, Republic of Korea, under Grant 2025-Glocal30-03-001.

ABSTRACT Skeleton-based action recognition has gained increasing attention due to its robustness to background variations and privacy-preserving properties. However, real-world skeleton datasets often exhibit severe class imbalance, leading to long-tailed data distributions that significantly degrade recognition performance, especially for tail classes. Existing mix-based skeleton augmentation methods partially alleviate this issue by increasing data diversity, but they often generate physically implausible skeletons and fail to explicitly account for class imbalance during augmentation. In this paper, we propose a long-tailed augmentation framework with two key components: 1) realistic skeleton mixing in the relative bone vector space and 2) class-aware sampling that increases effective augmentation frequency of underrepresented classes. Specifically, we perform mixing in the relative bone vector space rather than joint coordinate space to preserve kinematic consistency and further enforce bone-wise length alignment to prevent unrealistic limb deformations. To mitigate class imbalance, we introduce a class-aware weighting mechanism that biases sampling toward tail classes. Extensive experiments on the long-tailed NTU RGB+D and NTU RGB+D 120 benchmarks show that the proposed method consistently outperforms existing approaches across various imbalance factors, delivering notable improvements on the tail classes while remaining robust across different backbone architectures.

INDEX TERMS Augmentation, deep learning, long-tailed distribution learning, skeleton-based action recognition.

I. INTRODUCTION

Human action recognition has long been a fundamental research topic in computer vision [1], [2], [3]. Among various paradigms, skeleton-based action recognition has gained increasing popularity due to its robustness to background variations, computational efficiency, and inherent advantages in privacy-preserving applications [4]. With the availability of balanced large-scale datasets [5], [6], skeleton-based methods have demonstrated strong and competitive performance.

However, in real-world scenarios, constructing class-balanced action datasets is highly challenging [7], [8]. Since each class corresponds to a distinct human action, collecting sufficient and diverse samples for all classes in a balanced manner is inherently difficult [9]. Moreover, even after data collection, extracting and verifying reliable skeleton sequences from raw visual data requires substantial temporal and physical effort [10]. As a result, real-world skeleton action datasets often exhibit severe class imbalance, leading to long-tailed data distributions, which significantly degrade model training and hinder performance improvements in action recognition tasks [11].

To address such data imbalance in practice, extensive research has been conducted in the vision community. Classical approaches include resampling [12] and reweighting [13] strategies to compensate for skewed class distributions.

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

More recent studies aim to enhance tail-class diversity by leveraging majority-class samples for minority-class augmentation [14] or by implicitly modeling intra-class virtual samples at the feature level [15], [16].

Despite these advances, improving performance under data imbalance remains an open challenge. Moreover, most of these long-tailed learning strategies have been primarily developed for image-based recognition tasks [17] and cannot be directly applied to skeleton-based action recognition. Due to the structured and kinematic nature of skeleton data, naive extensions of existing methods often lead to physically implausible or semantically inconsistent skeleton representations [8]. As a result, relatively fewer studies have explored long-tailed learning specifically tailored to skeleton-based action recognition.

One representative approach for the long-tailed skeleton-based action recognition is BRL (Balanced Representation Learning) [7], which augments minority-class samples by combining upper- and lower-body skeleton structures and adjusting labels according to the sample ratios between action classes. In addition, ST-Mix and Shap-Mix [8] have been proposed to address long-tailed skeleton datasets. ST-Mix partitions skeletons into multiple body parts and performs part-level mixing for data augmentation. Shap-Mix further incorporates Shapley value theory to estimate the saliency of each body part, preserving the discriminative information of the target class during mixing.

Despite their effectiveness, existing skeleton-mixing approaches typically construct augmented samples by directly combining joint coordinates across spatial and temporal dimensions. Such coordinate-level mixing often neglects the underlying kinematic constraints of the human body, resulting in physically implausible skeleton sequences that do not correspond to realistic human motions. These unrealistic samples may introduce additional noise during training, which can be particularly problematic for tail classes where training data is scarce. Furthermore, most existing methods select source and target samples uniformly at random for mixing, which fails to explicitly account for class imbalance and is therefore suboptimal for long-tailed skeleton-based action recognition.

To address these limitations, we propose *Realistic Mixing*, a realistic skeleton mixing strategy that generates physically plausible augmented samples. Instead of directly mixing joint coordinates, our method operates in the relative bone vector space, which explicitly preserves the kinematic structure of the human body. We also enforce consistent bone lengths in the mixed skeletons by aligning them with either the source or the target skeleton, thereby preventing unrealistic limb deformations and improving motion realism.

In addition, we introduce a *Class-aware Sampling* strategy that accounts for class-wise data distributions during augmentation. By prioritizing tail classes when selecting target samples, our approach effectively amplifies minority-class representations using information from majority classes, thereby mitigating the long-tailed learning problem.

Extensive experiments under long-tailed settings with various imbalance factors demonstrate that our method consistently outperforms existing skeleton augmentation and imbalance-handling approaches. Notably, the proposed framework remains effective across different backbone architectures, highlighting its robustness and general applicability to long-tailed skeleton-based action recognition.

Our main contributions are summarized as follows:

- We propose *Realistic Mixing*, a realistic strategy that performs skeleton mixing in the bone space. By leveraging relative bone vector representations together with bone length matching, our method generates physically plausible augmented skeletons that preserve kinematic consistency.
- We introduce a *Class-aware Sampling* strategy that explicitly prioritizes tail classes during the augmentation process. By biasing target sample selection according to class frequency, this strategy effectively increases the augmentation frequency of minority classes and alleviates the long-tailed data imbalance.
- We conduct extensive experiments on large-scale long-tailed skeleton action recognition benchmarks. The results demonstrate that the proposed framework consistently outperforms existing skeleton augmentation and imbalance-handling methods across multiple backbone architectures.

II. RELATED WORKS

A. SKELETON-BASED ACTION RECOGNITION

Early skeleton-based action recognition methods primarily adopted convolutional and recurrent neural networks [18], [19], [20], [21], [22], [23], [24], [25], [26] to model spatial and temporal dynamics by transforming skeleton sequences into structured representations or sequential joint trajectories. However, these approaches have limited capability in explicitly modeling the inherent skeleton topology and long-range joint dependencies, which motivates the development of graph-based formulations [4], [27].

A representative work is ST-GCN [4], which introduced spatial–temporal graph convolution to jointly capture spatial joint dependencies and temporal motion patterns. Building upon this paradigm, Two-Stream Adaptive Graph Convolutional Networks (2s-AGCN) [28] further enhanced performance by learning adaptive graph topologies and exploiting complementary joint and bone modalities.

Subsequent studies have sought to improve graph modeling flexibility and representation capacity. CTR-GCN [29] proposed a channel-wise topology refinement mechanism that dynamically adjusts graph connections conditioned on feature channels, achieving strong performance with improved efficiency. Block-GCN [30] further advanced this line of research by decomposing skeleton graphs into semantic blocks, enabling more effective local and global motion modeling. These methods have significantly advanced the state of the art on large-scale benchmark datasets such as NTU RGB+D [5] under balanced training settings.

B. VISUAL RECOGNITION ON LONG-TAILED DISTRIBUTION

Traditional approaches to addressing data imbalance can be broadly categorized into two main strategies. The first is re-sampling [12], [31], [32], which adjusts the sample distribution by modifying the number of training instances per class. Specifically, oversampling is applied to minority classes, while undersampling is used for majority classes to reduce class imbalance. Although effective to some extent, oversampling may lead to overfitting, whereas undersampling risks discarding informative and valuable samples.

The second category consists of re-weighting methods [13], [33], [34], which assign different weights to training samples in the loss function according to class frequency. These methods adjust the importance of samples at either the class level or the instance level, enabling the model to place greater emphasis on minority classes during training. One of the representative works, balanced Softmax [35], corrects the bias of the conventional Softmax under label distribution shift by explicitly incorporating class frequency information, leading to improved performance on minority classes in long-tailed settings.

Beyond these traditional approaches, additional remedies have been proposed to address long-tailed data distributions by leveraging majority-class samples to enhance minority-class learning. The key idea is to increase the diversity of tail-class samples using head-class data. A representative example, GLMC [36], introduces global and local mixture consistency to regularize the model through prediction consistency on mixed samples, while employing a cumulative learning strategy to progressively improve representation learning for tail classes.

C. SKELETON-BASED ACTION RECOGNITION ON LONG-TAILED DISTRIBUTION

Despite growing interest and extensive research on learning from imbalanced data, the long-tailed issue in skeleton-based action recognition [37], [38] has received relatively limited attention. Existing long-tailed data addressing methods developed for other modalities are often non-optimal or sub-optimal when directly applied to skeleton data, due to the unique structural and kinematic properties of human skeletons.

Although several skeleton-specific augmentation techniques have been proposed [9], [39], [40], they are still limited in addressing class imbalance in long-tailed data. This limitation largely stems from their inability to capture the intrinsic spatio-temporal dependencies of skeleton sequences, which leads to the generation of less representative samples for tail classes.

A representative work addressing long-tailed skeleton data is BRL [37], which aims to improve long-tailed learning performance through skeleton mixing. BRL constructs mixed samples by concatenating the upper and lower body parts from skeletons of two different classes, while adjusting label ratios based on the relative dataset sizes of the source and target classes. Additionally, BRL incorporates class-wise weighting and loss adjustment strategies, demonstrating meaningful performance improvements on imbalanced datasets.

In addition, ST-Mix and Shap-Mix [8] have been proposed to better account for the spatio-temporal dependencies inherent in skeleton data. ST-Mix performs mixing at the body-part level in the spatial dimension and replaces down-sampled target sub-clips with corresponding source clips in the temporal dimension. Building upon ST-Mix, Shap-Mix further enhances spatial mixing efficiency by estimating the contribution of each body part to action recognition. During mixing, Shap-Mix preserves highly contributive body parts for tail classes, resulting in improved performance under long-tailed distributions.

However, mixed skeletons generated by ST-Mix and Shap-Mix can still exhibit physically implausible motions, such as inconsistent bone lengths across body parts or anatomically unrealistic poses. Since such augmented actions cannot be observed in real-world scenarios, they may instead hinder model training rather than improve it.

III. PRELIMINARIES

A. PROBLEM FORMULATION

We consider the task of skeleton-based action recognition under a long-tailed data distribution.

Let $\mathcal{C} = \{1, \ldots, C\}$ denote the set of action classes, and let $n_i (1 \leq i \leq C)$ be the number of training samples belonging to the i-th class. In a long-tailed setting, class frequencies are assumed to follow a monotonically decreasing order, $n_1 > n_2 > \cdots > n_C$. Accordingly, the class-imbalance ratio is defined as $\text{IF} := n_1/n_C$.

Each action instance is represented as a skeleton sequence $X \in \mathbb{R}^{V \times T \times D}$, where V denotes the number of joints, T the number of temporal frames, and D the coordinate dimension (e.g., 2D or 3D). Each skeleton sequence X is associated with an action label $y \in \mathcal{C}$.

B. SKELETON MIXING AUGMENTATION

ST-Mix and Shap-Mix [8] are representative augmentation methods for imbalanced skeleton-based action recognition. Our approaches are developed on top of these methods, extending their part-level mixing strategy with more realistic skeleton modeling and class-aware sampling. Notably, our proposed framework is general and can be seamlessly integrated with other skeleton mixing-based augmentation methods.

Before introducing our approaches, we briefly review the part-level skeleton mixing strategy employed by ST-Mix and Shap-Mix. In practice, the source skeleton X_{src} is directly taken from the current mini-batch, whereas the target skeleton X_{tar} is sampled uniformly at random from the same mini-batch. Given two skeleton sequences $X_{\text{src}}, X_{\text{tar}} \in \mathbb{R}^{V \times T \times D}$, these methods generate a new mixed skeleton X_{mix} by exchanging joint coordinates at the body-part level (i.e., left

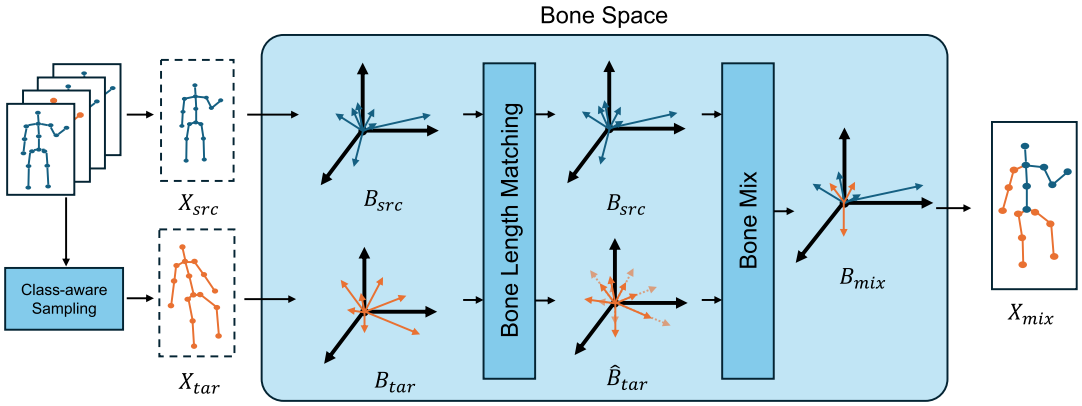

FIGURE 1. Illustration of the proposed *Realistic Mixing*. Source and target skeletons selected by class-aware sampling are transformed into bone representations, aligned via bone length matching, and mixed at the bone level. The mixed skeleton is then reconstructed to the joint space, producing physically plausible augmented samples for long-tailed skeleton-based action recognition.

arm, right arm, trunk, left leg, and right leg). Formally, the mixed skeleton is defined as

$$X_{\text{mix}}[\mathcal{P}] = X_{\text{tar}}[\mathcal{P}], \quad X_{\text{mix}}[\mathcal{U} \setminus \mathcal{P}] = X_{\text{src}}[\mathcal{U} \setminus \mathcal{P}], \tag{1}$$

where $\mathcal{P}$ denotes the selected subset of body parts, $\mathcal{U}$ represents the set of all joints, and $X[\cdot]$ indicates the joint coordinates corresponding to the specified body parts across all temporal frames. ST-Mix selects $\mathcal{P}$ uniformly at random, whereas Shap-Mix determines $\mathcal{P}$ based on body-part saliency estimated via Shapley values.

IV. METHODS

A. REALISTIC MIXING

We propose *Realistic Mixing*, a skeleton augmentation strategy that leverages relative bone vectors instead of absolute joint coordinates for skeleton mixing. Figure 1 illustrates the overall pipeline of the proposed *Realistic Mixing*.

1) BONE REPRESENTATION AND BONE MIX

While bone representations have been widely adopted in prior works [28], [41] as an additional modality for ensemble learning, to the best of our knowledge, they have not been utilized to improve the quality of data augmentation.

To introduce our bone-based augmentation, we first describe the bone representation in detail. We first define a kinematic tree by selecting the central spine joint as the root node and assigning a parent joint to each joint according to the predefined skeleton topology. Let the parent of joint i be denoted as $s(i)$. Then the skeleton representation in bone space, B, is defined as:

$$B[i] = X[i] - X[s(i)], \quad \forall\, i \in \{1, \ldots, V\}. \tag{2}$$

where $X[i]$ represents the coordinate of joint i.

The reconstruction process follows a topological order starting from the root joint to ensure structural consistency. The joint coordinates are recovered as:

$$X[i] = X[s(i)] + B[i], \quad \forall\, i \in \{1, \ldots, V\}. \tag{3}$$

Based on this formulation, we perform data augmentation directly in the bone space. Given two skeleton sequences X_{src} and X_{tar}, we first transform them into the bone space, yielding B_{src} and B_{tar}, respectively. We then apply the part-level mixing operation by replacing the selected set of bone vectors $\mathcal{P}$ as

$$B_{\text{mix}}[\mathcal{P}] = B_{\text{tar}}[\mathcal{P}], \quad B_{\text{mix}}[\mathcal{U} \setminus \mathcal{P}] = B_{\text{src}}[\mathcal{U} \setminus \mathcal{P}]. \tag{4}$$

The mixed bone representation B_{mix} is subsequently mapped back to the joint space using the inverse transformation described above, producing the augmented skeleton sequence X_{mix}. By performing augmentation in the bone space, our method explicitly preserves kinematic constraints of the human skeleton, resulting in physically plausible and anatomically consistent augmented sequences.

2) BONE LENGTH MATCHING

When mixing two different skeleton sequences, the corresponding bones may have inconsistent lengths due to variations in body proportions. Directly replacing bone vectors with mismatched lengths can introduce unrealistic deformations and violate physical plausibility.

To address this issue, we perform bone length matching before the mixing operation. Given the source and target bone representations B_{src} and B_{tar}, we align bone lengths in a *bone-wise* manner. Specifically, for each bone indexed by i, the target bone vector is rescaled to match the length of the corresponding source bone, yielding the length-aligned target bone representation $\hat{B}_{\text{tar}}$:

$$\hat{B}_{\text{tar}}[i] = \frac{\|B_{\text{src}}[i]\|}{\|B_{\text{tar}}[i]\|} \cdot B_{\text{tar}}[i], \quad \forall\, i \in \{1, \ldots, V\}. \tag{5}$$

Furthermore, to avoid biasing the augmentation toward a specific skeleton, we randomly select whether to match the bone lengths of the source or the target skeleton during training.

By aligning each bone individually while preserving its motion direction, the augmented skeleton maintains

realistic local kinematic structures and exhibits smoother joint trajectories. As a result, the proposed bone length matching strategy generates more diverse yet physically plausible samples, leading to improved generalization.

B. CLASS-AWARE SAMPLING

While realistic augmentation improves the physical plausibility of synthesized skeletons, it alone is insufficient to resolve the long-tailed data distribution problem. Existing mix-based augmentation methods typically select source and target samples uniformly at random within a mini-batch, which implicitly favors head classes due to their higher sample frequency. As a result, tail classes remain underrepresented even after augmentation.

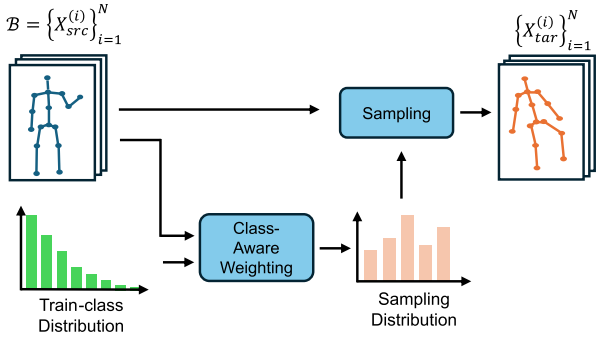

FIGURE 2. Class-aware sampling for skeleton mixing. Class-aware weighting is applied to samples in a mini-batch to construct a sampling distribution that favors tail classes. Target skeletons are then sampled from the mini-batch according to this distribution and paired with source samples for subsequent augmentation.

To explicitly address this issue, we introduce a *Class-aware Sampling* that assigns higher importance to samples from underrepresented classes and guides the subsequent sampling process, as illustrated in Fig. 2. Let N denote the number of samples in the current mini-batch, and let n_{y_i} denote the number of training samples belonging to the class label y_i of sample i. For each sample in a mini-batch, we conduct class-aware weighting by assigning a sampling score inversely proportional to its class frequency:

$$l(i) = \left(\frac{1}{n_{y_i}}\right)^{\alpha}, \quad \forall\, i \in \{1, \ldots, N\}. \tag{6}$$

where α controls the strength of class rebalancing.

The sampling probability is obtained by normalizing the scores over the samples in the current mini-batch using a softmax function:

$$p(i) = \frac{\exp(l(i))}{\sum_{j\in\mathcal{B}} \exp(l(j))}, \quad \forall\, i \in \{1, \ldots, N\}, \tag{7}$$

where $\mathcal{B}$ denotes the set of samples in the mini-batch.

Instead of directly sampling a target class, we sample target indices within the mini-batch to form source–target pairs. For each source skeleton sequence $X_{\text{src}}^{(i)}$ in the mini-batch ($i = 1, \ldots, N$), a target index $r(i)$ is sampled according to the class-aware probability distribution p with replacement. The corresponding target skeleton is then defined as

$$X_{\text{tar}}^{(i)} = X_{\text{src}}^{(r(i))}. \tag{8}$$

The paired skeletons $(X_{\text{src}}^{(i)}, X_{\text{tar}}^{(i)})$ are subsequently used as inputs to the proposed augmentation module.

By sampling targets according to p, tail-class samples within the mini-batch are more likely to be selected as mixing targets, thereby increasing their effective augmentation frequency. This strategy shifts the training emphasis toward underrepresented classes without discarding valuable head-class information, leading to more balanced representation learning under long-tailed distributions.

V. EXPERIMENTS

A. EXPERIMENTAL SETUP

1) DATASETS

We evaluate our method on two large-scale skeleton-based action recognition benchmarks, namely the NTU RGB+D and NTU RGB+D 120 datasets. For both datasets, we follow the Cross-Subject (X-Sub) evaluation protocol.

- NTU RGB+D (NTU-60) [5] is a large-scale dataset for 3D human action recognition. It consists of 56,880 samples from 60 action classes. These actions are performed by 40 subjects and are captured from 80 camera viewpoints. Here, the 40 subjects are divided into two groups: 20 subjects for training and the remaining 20 for testing. The training set contains 40,320 samples, while the testing set contains 16,560 samples.
- NTU RGB+D 120 (NTU-120) [6] is an extended version of the NTU-60, which enlarges the number of action categories and subjects. It contains 114,480 RGB+D video samples from 120 action classes, performed by 106 distinct subjects and captured from 155 camera viewpoints. Following the official split, the 106 subjects are evenly divided into 53 subjects for training and the other 53 subjects for testing.

2) IMPLEMENTATION DETAILS

Our implementation is based on the official CTR-GCN [29] codebase. All models are trained for 100 epochs using the SGD optimizer with a weight decay of 0.0004. The initial learning rate is set to 0.1 and decayed by a factor of 0.1 at the 60th and 80th epochs. A linear warm-up strategy is applied for the first 5 epochs to stabilize training. The batch size is set to 64, and all experiments are conducted on a single NVIDIA A6000 GPU. The class-aware sampling hyperparameter α is set to 0.5 in all experiments. Unless otherwise specified, CTR-GCN is used as the default backbone, and all results are reported in terms of Top-1 accuracy.

3) LONG-TAILED EVALUATION SETTING

Our experiments are designed to simulate real-world scenarios where the class-wise sample distribution is highly imbalanced. Although the original NTU datasets exhibit relatively balanced class distributions, it is necessary to

TABLE 1. Comparisons on Long-Tailed NTU Datasets (IF=100). Top-1 accuracy (%) comparison with representative class-imbalance handling methods on long-tailed NTU datasets (IF=100). All methods are evaluated under the same experimental setting with a single joint stream. Best and second-best results are highlighted in bold and underlined, respectively.

Method	Venue	LT-NTU 60 (IF=100)				LT-NTU 120 (IF=100)			
		Overall	Many	Medium	Few	Overall	Many	Medium	Few
Cross-Entropy Loss	–	74.4	86.4	69.5	63.8	63.2	83.4	64.6	52.2
ROS [42]	ICML'07	74.8	86.1	68.1	57.9	61.0	81.3	62.3	50.7
Mixup [43]	arXiv'17	75.9	86.3	71.7	66.5	66.9	85.6	65.6	55.5
Focal Loss [44]	ICCV'17	77.6	83.1	75.9	72.0	69.4	81.7	66.7	65.2
CB Loss [13]	CVPR'19	72.4	78.4	71.2	65.3	63.2	76.0	65.0	56.1
Balanced Softmax [35]	NeurIPS'20	77.6	83.8	73.9	73.3	69.6	81.4	67.8	66.7
PaCo [45]	ICCV'21	76.6	82.0	76.4	69.1	67.9	82.2	66.2	63.8
BCL [46]	CVPR'22	77.3	84.4	74.1	71.5	66.9	82.3	64.5	62.8
GLMC [36]	CVPR'23	78.8	78.6	**81.3**	76.0	71.5	79.5	70.5	69.2
BRL [46]	MIR'25	76.9	85.2	73.1	70.0	66.3	84.2	65.0	59.9
ST-Mix	IJCAI'24	77.0	**87.5**	73.1	67.1	68.3	**86.8**	68.4	59.6
ST-Mix + Ours	–	77.3	**87.5**	72.5	69.1	68.6	86.0	69.1	60.0
Shap-Mix	IJCAI'24	80.7	86.0	78.3	74.2	73.2	85.1	70.9	70.6
Shap-Mix + Ours	–	**82.1**	86.9	80.2	**76.5**	**74.5**	84.6	**71.7**	**73.4**

reshape them into more realistic long-tailed distributions for evaluation.

We consider three settings, IF $\in \{10, 50, 100\}$, to control the severity of the imbalance. By adjusting IF, we construct long-tailed versions of the training set with increasingly skewed class distributions.

We further divide classes into three groups according to the number of training samples: Many-shot (training samples > 100), Medium-shot (training samples 20$\sim$100), and Few-shot (training samples < 20). We report the performance on these three groups as well as the Overall accuracy.

The values of IF and the split thresholds (20 and 100) are also chosen by following prior works [47], [48]. The long-tailed resampling is applied only to the training set, while the validation and test sets remain unchanged.

TABLE 2. Top-1 accuracy (%) on LT-NTU 120 under different imbalance factors.

Method	IF=10	IF=50	IF=100
Baseline	79.4	69.9	63.2
Balanced Softmax	80.7	74.2	69.6
GLMC	82.0	75.9	71.5
ST-Mix	81.9	72.7	68.3
+ Ours	**82.3**	**74.6**	**68.6**
Shap-Mix	81.9	75.6	73.2
+ Ours	**82.9**	**76.7**	**74.5**

B. COMPARISON WITH OTHER METHODS

1) COMPARISON WITH REPRESENTATIVE CLASS-IMBALANCE METHODS

We compare the proposed method with representative class-imbalance handling approaches on long-tailed NTU datasets under the severe imbalance setting (IF=100). Table 1 summarizes the Top-1 accuracy on LT-NTU 60 and LT-NTU 120, reporting overall accuracy as well as performance on many-, medium-, and few-shot classes.

As shown in the table, conventional imbalance-handling methods provide limited improvements under extreme class imbalance, especially for medium- and few-shot classes. Whereas some methods improve overall accuracy, their gains on tail classes remain relatively small.

By incorporating the proposed framework, ST-Mix improves its overall accuracy on LT-NTU 120 from 68.3% to 68.6%, with consistent gains on medium- and few-shot classes. More notably, when applied to Shap-Mix, the proposed method boosts the overall accuracy from 73.2% to 74.5%, while improving few-shot accuracy from 70.6% to 73.4%. Similar trends are observed on LT-NTU 60, where the proposed method consistently enhances medium- and few-shot performance without sacrificing accuracy on many-shot classes.

These results demonstrate that the proposed method effectively strengthens existing skeleton-based augmentation techniques by improving tail-class performance under severe long-tailed distributions. Rather than serving as a standalone imbalance-handling strategy, the proposed approach acts as a general enhancement that can be seamlessly integrated into existing skeleton mixing frameworks to achieve more balanced and robust recognition performance.

2) PERFORMANCE UNDER DIFFERENT IMBALANCE LEVELS

Table 2 reports the Top-1 accuracy on the LT-NTU 120 dataset under different imbalance factors, which represent progressively more severe long-tailed settings. As the imbalance factor increases, the performance of all methods consistently degrades.

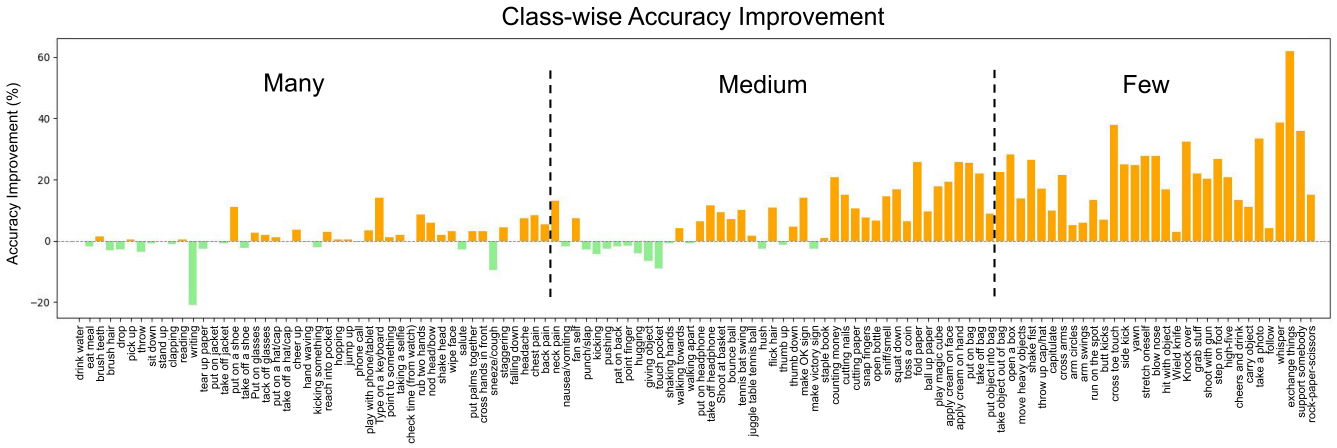

FIGURE 3. Class-wise accuracy improvement of the proposed method over the baseline on the long-tailed NTU-120 dataset (IF=100). Each bar represents the accuracy change of an individual action class. Classes are sorted by the number of training samples and grouped into many-shot, medium-shot, and few-shot categories, separated by dashed lines. While the baseline model shows limited or even negative gains for many-shot classes, the proposed method yields substantial and consistent improvements for medium- and few-shot classes.

TABLE 3. Top-1 accuracy (%) comparison with Different Backbone on NTU RGB+D 120.

Method	CTR-GCN	Block-GCN
Baseline	63.2	64.1
+ ST-Mix	68.3	68.5
+ ST-Mix with Ours	**68.6**	**68.9**
+ Shap-Mix	73.2	73.4
+ Shap-Mix with Ours	**74.5**	**74.5**

TABLE 4. Ablation study on each component.

Realistic Mixing	*Class-aware Sampling*	X-Sub Acc. (%)
		73.2
✓		73.9 (+0.7)
	✓	73.6 (+0.4)
✓	✓	**74.5 (+1.3)**

Notably, integrating the proposed method with both ST-Mix and Shap-Mix consistently improves performance under all imbalance factors. For ST-Mix, our method yields gains of +0.4%, +1.9%, and +0.3% under IF =10, 50, and 100, respectively. Similarly, when combined with Shap-Mix, our method achieves improvements of +1.0%, +1.1%, and +1.3% across the same settings.

These results indicate that the proposed approach is robust to varying degrees of class imbalance and remains effective even under extremely imbalanced conditions.

C. ANALYSIS AND ABLATION STUDY

1) PER-CLASS ACCURACY ANALYSIS

Figure 3 illustrates the class-wise accuracy improvement of the proposed method over the baseline on the LT-NTU 120 dataset under the severe long-tailed setting (IF =100). Each bar corresponds to the accuracy change of an individual action class, with classes sorted by the number of training samples and grouped into many-shot, medium-shot, and few-shot categories.

As shown in Fig. 3, the baseline model exhibits marginal or even negative accuracy changes for many-shot classes, indicating limited benefit from additional augmentation for already well-represented classes. In contrast, the proposed method yields clear and consistent accuracy improvements for medium- and few-shot classes, with particularly large gains observed in the few-shot regime.

This trend demonstrates that the proposed framework effectively redirects the learning focus toward underrepresented classes without overfitting head classes. The substantial improvements in tail classes confirm the effectiveness of the *Class-aware Sampling* in increasing the utilization of minority samples, while the *Realistic Mixing* contributes to generating informative and physically plausible augmented data. Overall, this per-class analysis provides strong evidence that the proposed method specifically benefits tail-class learning under severe class imbalance.

2) GENERALIZATION ACROSS DIFFERENT BACKBONES

Table 3 reports the performance comparison of different backbone architectures on the LT-NTU 120 dataset under the severe long-tailed setting (IF=100). We evaluate the proposed method on two representative graph-based backbones, CTR-GCN and Block-GCN [30], to examine the generalization capability of our approach across different network designs.

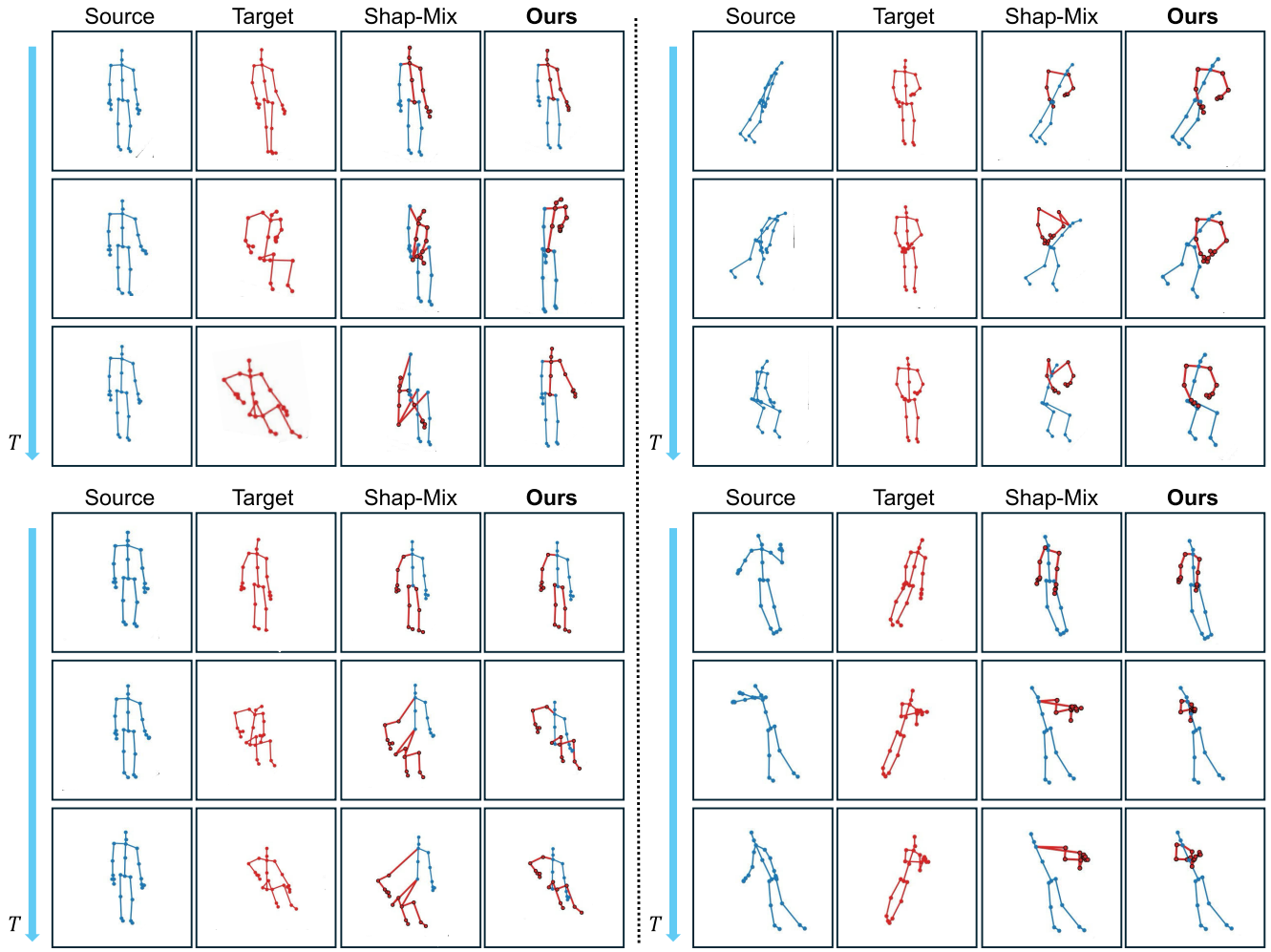

FIGURE 4. Qualitative comparison of skeleton augmentation results. From left to right, we show the source skeleton, target skeleton, baseline joint-coordinate mixing, and the proposed Realistic Mix. While Shap-Mix mixing produces physically implausible poses and distorted limb proportions, our method generates anatomically consistent and smoother skeleton motions by operating in the bone space with bone length matching.

As shown in the table, both ST-Mix and Shap-Mix consistently benefit from the proposed framework on both backbones. When combined with our method, ST-Mix achieves improvements from 68.3% to 68.6% on CTR-GCN and from 68.5% to 68.9% on Block-GCN. Similarly, integrating our method with Shap-Mix yields further gains, improving accuracy from 73.2% to 74.5% on CTR-GCN and from 73.4% to 74.5% on Block-GCN.

These consistent improvements across different backbones indicate that the proposed Realistic Mix and class-aware weighting strategy are not tied to a specific network architecture. Instead, they serve as a general and effective augmentation framework that can be seamlessly integrated into diverse skeleton-based action recognition models under long-tailed distributions.

3) COMPONENT-WISE ABLATION STUDY

To examine the effectiveness of each component in the proposed framework, we conduct a component-wise ablation study on the NTU RGB+D 120 dataset under the severe long-tailed setting (IF=100). Table 4 reports the corresponding results.

Applying *Realistic Mixing* alone improves the baseline performance by +0.7%, demonstrating that realistic skeleton augmentation in the bone space effectively enhances data quality. Using *Class-aware Sampling* alone yields a +0.4% gain, indicating that class-aware sampling contributes to alleviating class imbalance by increasing the effective training frequency of tail classes. When both components are jointly applied, the model achieves a total improvement of +1.3%, which is significantly larger than the gain from either component alone.

These results highlight the complementary nature of *Realistic Mixing* and *Class-aware Sampling*: while *Realistic Mixing* focuses on generating physically plausible augmented samples, *Class-aware Sampling* explicitly addresses the long-tailed data distribution. Their synergy enables the model to better exploit augmented data and learn more

balanced representations, leading to consistent performance improvements under severe class imbalance.

D. QUALITATIVE ANALYSIS OF SKELETON AUGMENTATION

Figure 4 presents qualitative comparisons of skeleton augmentation results generated by Shap-Mix and the proposed method. For each example, we visualize multiple temporal frames, where time progresses from top to bottom. From left to right, the source skeleton, target skeleton, Shap-Mix result, and the augmented skeleton produced by our method are shown.

As illustrated in Fig. 4, Shap-Mix often produces skeletons with distorted limb proportions, abrupt joint transitions, or physically implausible poses, especially when mixing skeletons with significantly different body configurations or motion patterns. These artifacts arise from direct joint-coordinate replacement, which does not explicitly enforce kinematic constraints.

In contrast, the proposed method consistently generates anatomically coherent skeletons with smooth and natural motions. By performing mixing in the bone space and enforcing bone length consistency, our approach preserves relative joint structures and motion continuity across frames. As a result, the augmented skeletons maintain realistic body proportions and temporal smoothness, even under challenging source–target combinations.

These qualitative results provide intuitive evidence that the Realistic Mix produces higher-quality and physically plausible augmented samples, which is critical for effective training under long-tailed data distributions.

VI. CONCLUSION

In this work, we addressed the long-tailed skeleton-based action recognition problem by improving both the quality and distribution of augmented training samples. We proposed *Realistic Mixing*, a realistic skeleton mixing strategy that operates in the bone space with bone length alignment to generate physically plausible and kinematically consistent augmented skeletons. In addition, we introduced a *Class-aware Sampling* strategy that explicitly prioritizes tail classes during augmentation, effectively mitigating class imbalance.

By jointly considering physical realism and class-aware data distribution, our method offers a principled way to improve robustness under severe class imbalance. We believe that this framework can be readily integrated with existing skeleton-based recognition pipelines and further extended to other skeleton-related tasks and imbalance-handling scenarios in future work.

ACKNOWLEDGMENT

(Jinwoo Kim and Wonhee Kim contributed equally to this work.)

REFERENCES

[1] O. Masoud and N. Papanikolopoulos, "A method for human action recognition," *Image Vis. Comput.*, vol. 21, no. 8, pp. 729–743, 2003.

[2] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human action recognition by learning spatio-temporal features with deep neural networks," *IEEE Access*, vol. 6, pp. 17913–17922, 2018.

[3] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *Int. J. Comput. Vis.*, vol. 130, no. 5, pp. 1366–1401, May 2022.

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

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

[6] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[7] H. Liu, Y. Wang, M. Ren, J. Hu, Z. Luo, G. Hou, and Z. Sun, "Balanced representation learning for long-tailed skeleton-based action recognition," *Mach. Intell. Res.*, vol. 22, no. 3, pp. 466–483, Jun. 2025.

[8] J. Zhang, L. Lin, and J. Liu, "Shap-mix: Shapley value guided mixing for long-tailed skeleton based action recognition," 2024, *arXiv:2407.12312*.

[9] S. Lee, A. Fakhry, J. Kim, and J. T. Lee, "Discriminative skeleton-based action recognition via co-learning with motion diffusion model," in *Proc. IEEE Int. Conf. Adv. Vis. Signal-Based Syst. (AVSS)*, Aug. 2025, pp. 1–6.

[10] M. Cormier, Y. Schmid, and J. Beyerer, "Enhancing skeleton-based action recognition in real-world scenarios through realistic data augmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2024, pp. 300–309.

[11] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: Latest research, applications and future directions," *Artif. Intell. Rev.*, vol. 57, no. 6, p. 137, May 2024.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[13] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9260–9269.

[14] S. Park, Y. Hong, B. Heo, S. Yun, and J. Y. Choi, "The majority can help the minority: Context-rich minority oversampling for long-tailed classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6877–6886.

[15] Y. Wang, X. Pan, S. Song, H. Zhang, G. Huang, and C. Wu, "Implicit semantic data augmentation for deep networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019, pp. 1–10.

[16] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, and X. Cheng, "MetaSAug: Meta semantic augmentation for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5208–5217.

[17] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10795–10816, Sep. 2023.

[18] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1110–1118.

[19] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018.

[20] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 816–833.

[21] L. Wang, X. Zhao, and Y. Liu, "Skeleton feature fusion based on multi-stream LSTM for action recognition," *IEEE Access*, vol. 6, pp. 50788–50800, 2018.

[22] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4570–4579.

[23] Y. Xu, J. Cheng, L. Wang, H. Xia, F. Liu, and D. Tao, "Ensemble one-dimensional convolution neural networks for skeleton-based action recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 1044–1048, Jul. 2018.

[24] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 102–106.

[25] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3D action recognition," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2842–2855, Jun. 2018.

[26] C. Cao, C. Lan, Y. Zhang, W. Zeng, H. Lu, and Y. Zhang, "Skeleton-based action recognition with gated convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 11, pp. 3247–3257, Nov. 2019.

[27] H. Chen, Y. Jiang, and H. Ko, "Pose-guided graph convolutional networks for skeleton-based action recognition," *IEEE Access*, vol. 10, pp. 111725–111731, 2022.

[28] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12018–12027.

[29] 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. 13339–13348.

[30] Y. Zhou, X. Yan, Z.-Q. Cheng, Y. Yan, Q. Dai, and X.-S. Hua, "BlockGCN: Redefine topology awareness for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2049–2058.

[31] L. Shen, Z. Lin, and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 467–482.

[32] S. S. Mullick, S. Datta, and S. Das, "Generative adversarial minority oversampling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1695–1704.

[33] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019, pp. 1–12.

[34] S. Park, J. Lim, Y. Jeon, and J. Y. Choi, "Influence-balanced loss for imbalanced visual classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 715–724.

[35] J. Ren, C. Yu, X. Ma, H. Zhao, and S. Yi, "Balanced meta-softmax for long-tailed visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4175–4186.

[36] F. Du, P. Yang, Q. Jia, F. Nan, X. Chen, and Y. Yang, "Global and local mixture consistency cumulative learning for long-tailed visual recognitions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15814–15823.

[37] J. Liu, Y. Sun, C. Han, Z. Dou, and W. Li, "Deep representation learning on long-tailed data: A learnable embedding augmentation perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2967–2976.

[38] S. Zhang, Z. Li, S. Yan, X. He, and J. Sun, "Distribution alignment: A unified framework for long-tail visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2361–2370.

[39] K. Xu, F. Ye, Q. Zhong, and D. Xie, "Topology-aware convolutional neural network for efficient skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 3, pp. 2866–2874.

[40] J. Zhang, L. Lin, and J. Liu, "Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 3, pp. 3427–3435.

[41] 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. 20154–20164.

[42] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proc. 24th Int. Conf. Mach. Learn.*, Jun. 2007, pp. 935–942.

[43] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[45] J. Cui, Z. Zhong, S. Liu, B. Yu, and J. Jia, "Parametric contrastive learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 695–704.

[46] J. Zhu, Z. Wang, J. Chen, Y.-P.-P. Chen, and Y.-G. Jiang, "Balanced contrastive learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6898–6907.

[47] T. Li, L. Wang, and G. Wu, "Self supervision to distillation for long-tailed visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 610–619.

[48] J. Li, Z. Tan, J. Wan, Z. Lei, and G. Guo, "Nested collaborative learning for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6939–6948.

JINWOO KIM received the B.S. degree in mathematics and computer science and engineering from Kyungpook National University, Daegu, South Korea, in 2025, where he is currently pursuing the M.S. degree with the School of Computer Science and Engineering, under the supervision of Prof. Jong Taek Lee. His research interests include deep learning, computer vision, and long-tailed learning.

WONHEE KIM (Student Member, IEEE) is currently pursuing the bachelor's degree in computer science and engineering with Kyungpook National University (KNU), Daegu, South Korea, with a minor in electronics engineering. He was an Undergraduate Research Intern with the Vision and Intelligent Systems Laboratory, KNU, under the guidance of Prof. Jong Taek Lee. His research interests include artificial intelligence and computer vision.

MI-SEON KANG received the B.S. degree in electronic, electrical, and computer engineering from Kyungpook National University, Daegu, South Korea, in 2010, the M.S. degree in electronic, electrical, and computer engineering, in 2012, and the Ph.D. degree in electrical engineering from Kyungpook National University, in 2022.

Since 2012, she has been with the Electronics and Telecommunications Research Institute (ETRI), South Korea, where she is currently a Senior Researcher and a Chief Technical Staff with the AI Infrastructure Research Laboratory, Daegu-Gyeongbuk Research Division. Her research interests include artificial intelligence, smart sensor-based systems, the IoT, computer vision, and digital twin-based big data-driven system development.

JONG TAEK LEE (Member, IEEE) received the B.S. degree in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2005, and the M.S. and Ph.D. degrees in electrical and computer engineering from The University of Texas at Austin, Austin, TX, USA, in 2007 and 2012, respectively.

From 2012 to 2022, he was a Senior Researcher with the Electronics and Telecommunications Research Institute, South Korea. Since 2022, he has been an Assistant Professor with the School of Computer Science and Engineering, Kyungpook National University, Daegu, South Korea. His research interests include computer vision and machine learning, with a focus on 3D vision and human action recognition for mobile video surveillance, robot perception, and rehabilitation healthcare.