Shap-Mix: Shapley Value Guided Mixing for Long-Tailed Skeleton Based Action Recognition

Jiahang Zhang, **Lilang Lin**, **Jiaying Liu** *

Wangxuan Institute of Computer Technology, Peking University

{zjh2020, linlilang, liujiaying}@pku.edu.cn

Abstract

In real-world scenarios, human actions often fall into a long-tailed distribution. It makes the existing skeleton-based action recognition works, which are mostly designed based on balanced datasets, suffer from a sharp performance degradation. Recently, many efforts have been made to image/video long-tailed learning. However, directly applying them to skeleton data can be suboptimal due to the lack of consideration of the crucial spatial-temporal motion patterns, especially for some modality-specific methodologies such as data augmentation. To this end, considering the crucial role of the body parts in the spatially concentrated human actions, we attend to the mixing augmentations and propose a novel method, Shap-Mix, which improves long-tailed learning by mining representative motion patterns for tail categories. Specifically, we first develop an effective spatial-temporal mixing strategy for the skeleton to boost representation quality. Then, the employed saliency guidance method is presented, consisting of the saliency estimation based on Shapley value and a tail-aware mixing policy. It preserves the salient motion parts of minority classes in mixed data, explicitly establishing the relationships between crucial body structure cues and high-level semantics. Extensive experiments on three large-scale skeleton datasets show our remarkable performance improvement under *both long-tailed and balanced* settings. Our project is publicly available at: https://jhang2020.github.io/Projects/Shap-Mix/Shap-Mix.html.

1 Introduction

Human activity understanding is a crucial problem with wide applications in real life, *e.g.*, healthcare and autonomous driving. 3D skeleton, as a highly efficient representation, describes the human body by 3D coordinates of keypoints. In comparison to other modalities such as RGB videos and depth data, skeletons are lightweight, compact, and relatively privacy-preserving. Owing to these advantages, skeletons have become widely used in human action recognition.

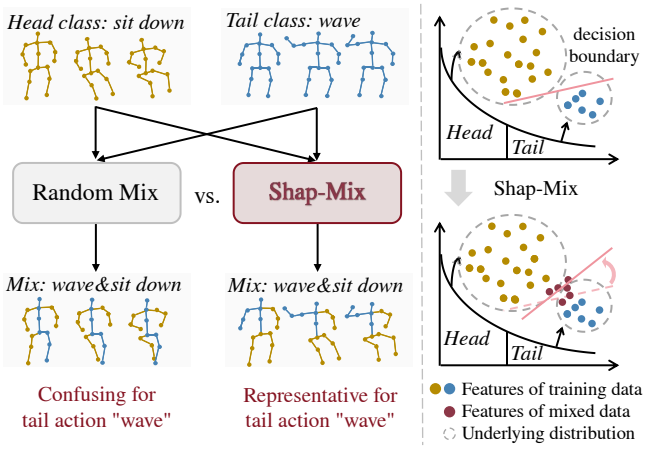


Figure 1: Random mix, *e.g.*, Cut-Mix, treats different classes equally and causes the semantic confusion, degrading the long-tailed performance especially for tail categories. In contrast, our Shap-Mix generates representative samples for tail categories to recover the underlying distribution, obtaining a better decision boundary.

Skeleton-based human action recognition has achieved a remarkable success [Song *et al.*, 2018a; Chen *et al.*, 2021a; Zhang *et al.*, 2023]. However, existing works are mostly targeted at balanced datasets, ignoring the prevalent long-tailed distribution phenomenon in the real world. For example, NTU datasets [Shahroudy *et al.*, 2016; Liu *et al.*, 2019], currently the most popular benchmarks for skeleton-based action recognition, are constructed with a balanced data distribution obtained by human intervention. In contrast, the human actions in real world are under long-tailed distributions [Zhang *et al.*, 2021; Damen *et al.*, 2022]. For example, there are lots of data on actions of "walking" or "standing" while little on some actions such as "abdominal pain" and "falling", which can be more important for some real applications, *e.g.*, health monitoring. This gap results in a huge limitation for the deployment of these works in the real world, *i.e.*, a significant performance degradation on tail categories when transferred to the long-tailed settings directly, due to the triggered biased training [Liu *et al.*, 2023].

To deal with this prevalent challenge in real world, long-

*Corresponding author.

tailed learning has attracted much attention. Generally, the training data under the long-tailed distribution brings two challenges to model [Chu *et al.*, 2020]. The first challenge is the *data imbalance*, which can cause representation bias and negative gradient over-suppression problem on the tail classes [Hsieh *et al.*, 2021; Tan *et al.*, 2020]. To this end, many re-balancing methods have been proposed, including the re-sampling [Van Hulse *et al.*, 2007; Han *et al.*, 2005; Shen *et al.*, 2016] and re-weighting [Lin *et al.*, 2017; Ren *et al.*, 2020; Park *et al.*, 2021], to help the model learn a balanced representation space. They are built upon the assumption that the optimal decision boundary is still well-defined in the given partial data. However, it is often difficult to obtain the optimal decision boundary due to the scarce tail class data. This introduces the second challenge, *i.e.*, *the difficulty of recovering the underlying distributions* from limited tail class samples. As an effective way to provide new training samples, data augmentation [Chu *et al.*, 2020; Chou *et al.*, 2020; Chen *et al.*, 2022] has been widely studied to enhance the distribution learning for the tail categories. However, previous augmentation-based long-tailed methods are designed for images/videos and cannot directly transfer to the skeleton data, which necessitates more careful consideration of its crucial spatial-temporal motion patterns and has been overlooked before. Meanwhile, to get rid of the complicated multiple training stages in previous works [Chu *et al.*, 2020; Chen *et al.*, 2022], we also aim to develop a simpler augmentation-based method with end-to-end training.

In this paper, we focus on the long-tailed augmentation-based method. Considering that human action is often represented as the combination of different motion patterns in skeletons, we attend to the mixing method to integrate different motion parts. Specifically, we propose a novel method, Shap-Mix, as shown in Figure 1. It can generate representative samples for the tail categories guided by the Shapley value [Shapley and others, 1953], a important solution concept for allocation problems in cooperative game theory. To begin with, a simple random mixing technique is developed with effective spatial-temporal design, significantly improving the representation quality of the encoder backbone. Based on this, considering the difference of the head and tail categories, we further develop a saliency-guided mixing strategy, Shap-Mix. Concretely, we first utilize the Shapley value to perform the saliency estimation for different body parts of each action category. Then utilizing the obtained saliency maps, a tail-aware mixing policy is proposed to maintain the representative motion patterns for the tail categories. Shap-Mix promotes decision boundary learning for the minority categories by explicitly establishing the relationship between crucial motion patterns and high-level semantics. We conduct extensive experiments to verify the effectiveness of our method. Remarkably, a significant improvement is achieved with our method under both balanced and long-tailed settings.

Our contributions can be summarized as follows:

- We propose an effective skeleton mixing method, Shap-Mix, for long-tailed action recognition, which consists of a novel skeleton saliency estimation technique based on the Shapley value in cooperative game theory. It jointly considers the relationship between different skeleton joints to obtain the corresponding importance more rationally.
- Based on the obtained joint importance distribution, a tail-aware mixing strategy is proposed, which prefers to produce the mixed samples that are more representative for the tail categories. It alleviates the over-fitting problem of tail categories by explicitly establishing the relationship between crucial motion patterns and high-level semantics.
- A large-scale benchmark for long-tailed skeleton action recognition, covering three popular datasets and different methodological algorithms, is provided to benefit the community. Our method demonstrates the significant performance improvement for the long-tailed learning, and the notable effectiveness is also verified in the balanced setting with full datasets.

2 Related Works

2.1 Skeleton-based Action Recognition

Skeleton-based action recognition aims to classify the action categories using 3D coordinates data of the human body. Previous works are mostly based on the recurrent neural network (RNN), and convolutional neural network (CNN), treating the skeleton in the temporal series [Song *et al.*, 2017; Song *et al.*, 2018a; Song *et al.*, 2018b] or pseudo 2D-image [Ke *et al.*, 2017; Liu *et al.*, 2017]. Recently, inspired by the natural topology structure of the human body, graph convolutional neural network (GCN)-based methods [Yan *et al.*, 2018; Shi *et al.*, 2019; Cheng *et al.*, 2020] have attracted more attention, achieving remarkable performance. Meanwhile, transformer-based models [Shi *et al.*, 2020; Plizzari *et al.*, 2021] also show promising results by learning long-range temporal dependencies, owing to the attention mechanism.

Different from the above works on model architecture, we focus on mixing augmentation design in this paper, which improves the performance as a plug-in design.

2.2 Long-Tailed Visual Recognition

There are two typical methods to tackle long-tailed learning, *i.e.*, re-sampling and re-weighting methods. Re-sampling methods [Van Hulse *et al.*, 2007; Han *et al.*, 2005; Shen *et al.*, 2016] deal with the data imbalance issue by oversampling the tail categories or under-sampling the head categories. Re-weighting methods [Lin *et al.*, 2017; Ren *et al.*, 2020; Park *et al.*, 2021], often assign a higher weight to the tail categories to balance the positive gradients and negative gradients flowing. Recently, valuable efforts have been also made on the data augmentation [Chou *et al.*, 2020; Du *et al.*, 2023], ensemble learning [Wang *et al.*, 2021], decoupled learning [Kang *et al.*, 2019], and contrastive learning [Zhu *et al.*, 2022], providing new perspectives on the image long-tailed learning.

In contrast, the long-tailed issue has not been well explored for skeleton data, especially for the modality-specific data augmentation methods. Previous works targeted at image data can be unsuitable and sub-optimal due to the sparse and compact body structure in the skeleton. Meanwhile, some skeleton augmentation methods [Xu *et al.*, 2022; Zhan *et al.*, 2022] do not well consider the crucial spatial-temporal dynamics in human skeleton, especially under the

long-tailed distribution, leading to the generation of less representative data samples. To this end, we propose a customized mixing augmentation with saliency guidance to generate new data samples to benefit the long-tailed learning.

2.3 Shapley Value

Shapley value [Shapley and others, 1953] is one of the most important concepts in cooperative game theory. It is used to allocate the achieved overall worth in cooperation to each player. Let us consider the set $\mathbb{P}$ and a function $f(\cdot)$ indicating the corresponding worth achieved by some players as a real number. The Shapley value of player i is its average marginal contribution to all possible coalitions $\mathbb{S} \subseteq \mathbb{P}$ that can be formed without i. Concretely, it is formulated as

$$SV(i) = \frac{1}{|\mathbb{P}|} \sum_{\mathbb{S} \subseteq (\mathbb{P}-\{i\})} \frac{f(\mathbb{S} \cup i) - f(\mathbb{S})}{comb(|\mathbb{S}|, |(\mathbb{P} - \{i\})|)}, \quad (1)$$

where the $comb(\cdot, \cdot)$ is the function of combination number. The obtained $SV(i)$ is the final allocated worth to the player i. It well considers the interaction between different players, giving a more rational allocation solution. Meanwhile, it is shown that Shapley value satisfies good properties, *e.g.*, *Efficiency, Symmetry, Linearity* and *Null player* [Hart, 2016].

Due to its solid theory foundation and desirable properties, Shapley value has attracted much attention for machine learning study.However, one drawback is its high computation cost as an NP-complete problem [Deng and Papadimitriou, 1994]. In response, many works adopt the Monte Carlo Sampling to give an estimation.

In this paper, we innovatively apply the Shapley value to the skeleton saliency estimation, which is the first to our best knowledge. Thanks to the sparsity of skeleton data, our approach does not require too much computational overhead.

3 The Proposed Method: Shap-Mix

We introduce our proposed method in this section. First, a simple yet effective skeleton mixing augmentation is presented. Based on this, we further propose a customized rebalanced mixing strategy guided by Shapley value to handle the long-tailed data distribution. Finally, the whole training scheme of the model is provided.

3.1 Preliminaries

Notations. Generally, a skeleton sequence (the i_{th} sample) can be represented as $s_i \in \mathbb{R}^{C \times T \times V \times M}$ with its label as c_i, where C, T, V, M are channel, frame, joint, and performer dimensions, respectively. Note that the number of head categories is often much larger than the tail ones in long-tailed recognition. Our goal is to train a model with good generalization capacity on different categories using long-tailed data.

Mixup and Cut-Mix. Mixup [Zhang *et al.*, 2017] and Cut-Mix [Yun *et al.*, 2019] are proposed as a regularization technique to improve the generalization capacity of deep model. Two skeleton sequences, s_i and s_j, are randomly sampled first. Then for Mixup, the two sequences are mixed in the spatial-temporal dimension $s_{ij}^{Mix} = \lambda * s_i + (1 - \lambda) * s_j$, where λ is the mixing ratio number, sampled from the beta distribution. And for Cut-Mix, a spatial-temporal binary mask m is generated, with a masking ratio of $1 - \lambda$. Then the mixing operation is applied in a copy-paste manner, *i.e.*, $s_{ij}^{Mix} = m \odot s_i + (1 - m) \odot s_j$, where $\odot$ is the element-wise multiplication operator. The corresponding mixed label in the two methods is obtained by $c_{ij}^{Mix} = \lambda * c_i + (1 - \lambda) * c_j$.

3.2 A Simple Skeleton Mixing Strategy

Mixing methods have been proven effective for representation learning and well-explored in image field [Qin *et al.*, 2020; Walawalkar *et al.*, 2020; Liu *et al.*, 2022]. However, for skeleton data, there is still a lack of effective mixing design considering its complex spatial-temporal dependencies. To this end, we introduce a simple yet effective mixing method, ST-Mix, for skeleton data in the following.

Spatial-Temporal Mixing Design. To fully boost the modeling of skeleton spatial-temporal dynamics, we separately consider the mixing strategy on the spatial and temporal dimensions. Note that it is for Cut-Mix to generate a better mask, while is not required in Mixup.

For spatial dimension, many works [Hua *et al.*, 2023; Zhan *et al.*, 2022] have found that meaningful motion patterns are captured at the part level instead of the joint level. Therefore, we first divide the skeleton into five different parts, *i.e.*, *trunk, left arm, right arm, left leg,* and *right leg* following the previous works. Then the mixing operation occurs at the level of body parts to maintain the crucial motion patterns. Specifically, we randomly sampled $N_s \in [N_s^l, N_s^u]$ body parts as the mixing targets of the spatial dimension. N_s^l and N_s^u are the lower- and upper-bound hyper-parameters.

For the temporal dimension, we can simply copy a subclip sampled from s_j into the corresponding position of s_i directly. However, a short clip can not accurately represent the whole action sequence for model to recognize. For example, when only a short clip of a hand in raising is observed, many actions, *e.g.*, *making phone calls* and *drinking*, can be the candidates, leading to the confusion of labels. Therefore, we propose a temporal down-sampling strategy to construct more informative mixed samples. Specifically, supposing N_t frames are to be mixed, a clip is sampled from s_j with its length in $[N_t, T]$ and then down-sampled into N_t frames to paste into s_i to generate the final mixed sample.

Joint Global-Local Mixture Learning. As discussed before, Mixup and Cut-Mix construct the mixed samples in different means, where the former produces the global mixture while the latter performs the copy-paste operation on the local patterns. To introduce more diverse motion patterns and further improve the generalization capacity, we bake these two mixture generation paradigms into our training process. Specifically, for each skeleton, we randomly choose and apply a mixing method, *i.e.*, Mix-up or Cut-Mix, jointly learning the global and local mixture semantics.

Here we highlight these specific points to distinguish our method from other skeleton mixing works [Xu *et al.*, 2022; Zhan *et al.*, 2022], 1) the well-designed mixing policy in spatial-temporal dimensions, 2) the joint learning for the global and local mixed data. These simple yet effective designs significantly improve the model performance compared

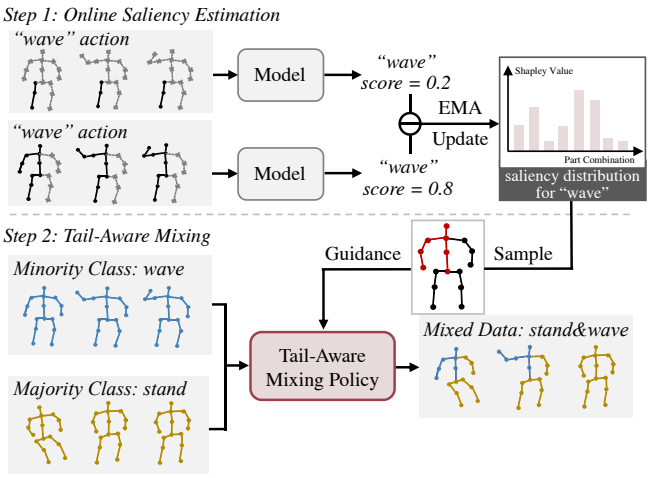


Figure 2: A simplified illustration of Shap-Mix. We first perform the online saliency estimation using Eq. (2). In this example, $r = \{$*right leg*$\}$ and $b = \{$*trunk, right arm*$\}$. For dotted joints, we use the mean of the dataset as the static sequence. The calculated Shapley value is used to update the Shapley value list v^c by EMA. Finally, the mixed data is generated, preserving the representative motion patterns of the minority class (*wave* in this example).

with other mixing methods as shown in Table 6, providing a strong skeleton augmentation technique. Besides, different from other works, we focus on supervised learning, especially under the long-tailed data distribution, which will be discussed in the following sections.

3.3 Shapley-Value-Guided Rebalanced Mixing

Now we have obtained an effective skeleton mixing augmentation. However, it adopts a random mixing manner, *i.e.*, treating all categories equally during training. Considering the data distribution in real world is often long-tailed, it can be unreasonable because tail categories often necessitate more struggle to learn the underlying distribution. As a remedy to this issue, we further propose a customized mixing strategy, Shap-Mix, guided by Shapley value [Shapley and others, 1953] to promote the decision boundary learning for tail categories as shown in Figure 2. Our key idea is to 1) obtain the saliency map of the skeleton joints first and then 2) use it to guide the synthesis of higher quality mixed samples for tail category learning. More details are provided as follows.

Part Saliency Estimation Based on Shapley-Value. Shapley value, as introduced in Section 2.3, is effective on the allocation of the worth to a set of players in cooperation. In addition to its own good properties and solid theoretical foundation, we list three key factors that motivate our choice of the Shapley value for skeleton saliency estimation:

• Shapley value takes into account the interactions between different players, *i.e.*, body parts in our context, to give the saliency. For example, hands (in raising) are important to recognize *saluting*, while it can also be confused with other actions, *e.g.*, *making phone calls* and *drinking* with similar hand motions. Therefore, the interaction between human trunk and hands should be considered jointly when measuring the importance of hands to action *saluting*.

• Shapley value is based on the input instead of the deep feature maps. It can avoid the adverse effect of over-smoothing [Liu *et al.*, 2020a] problem, the exponential convergence of similarity measures on node features, which is widely existing in GCNs. In other words, the saliency estimation can be inaccurate using the methods based on the fusion of deep features, *e.g.*, Grad-CAM [Selvaraju *et al.*, 2017].

• A disadvantage of Shapley value is that its calculation can be computationally costly. Fortunately, skeleton is much sparser compared with image data, which can effectively reduce the computation overhead. Besides, we adopt an online estimation method to further improve the efficiency.

Specifically, we maintain a saliency value list v^c for each category c, which stores v_b^c for every body part combination b, *i.e.*, $v^c = \{v_b^c\}$. Specifically, b denotes a body part combination that may appear in the spatial mixing. It is a subset of the universal set $\mathbb{U}$ containing all pre-defined 5 body parts, *e.g.*, $b = \{$*left arm, right leg*$\}$. The overall worth in our context is defined as the predicted confidence of the model on the correct class c, *i.e.*, $f^c(\cdot)$. Then the Shapley value v_b^c is computed as:

$$v_b^c = \frac{1}{|\mathbb{U}-b|+1} \sum_{r \subseteq (\mathbb{U}-b)} \frac{f^c(r \cup b) - f^c(r)}{comb(|r|, |(\mathbb{U}-b)|)}, \quad (2)$$

where $comb(\cdot,\cdot)$ returns the combination number and $|\cdot|$ is the cardinality of a set. Intuitively, v_b^c represents the average marginal contribution of part combination b to the prediction confidence on the class c it belongs to.

Next, we provide some more implementation details for a better understanding. First, v_b^c is computed and averaged only on the skeletons belonging to class c, and all these skeletons share the same saliency list v^c. When computing the prediction score in Eq. (2), the selected body parts, together with the other part regions which are set to the corresponding mean joint value of the dataset, make up the complete input skeleton for model. For efficiency, we do not consider the Shapley value of temporal dimension, *i.e.*, we always replace the same body parts in all the frames for saliency estimation. Meanwhile, we compute the Shapley value in an online way during training to share the computational overhead. Specifically, we only sample the part combination r and b once to calculate the single marginal contribution at each iteration, and then update the v_b^c using exponential moving average (EMA) to get the average estimation during the whole training process. The obtained saliency results will be utilized to guide the mixing synthesis as introduced in the following.

Tail-Aware Mixing Synthesis. We find that the intra-class saliency estimation is easier than the inter-class decision boundary learning in Figure 3. Therefore, the obtained saliency maps can serve as guidance for the synthesis of mixed data. Due to the scarcity of tail class samples, we suggest explicitly guiding the model to capture the relationships between the crucial motion patterns and the action semantics, revealing the possible underlying distributions of tail categories. Therefore, we introduce a tail-aware mixing policy, to produce more representative mixed data for tail classes.

Specifically, given the source data s_i with label c_i, and the target data s_j with label c_j, the body parts to be mixed are

sampled from a specific importance distribution $d(\cdot)$, which can be formulated as:

$$d(c) = softmax(norm(\{v_b^c/|b|\})/\tau), \quad (3)$$

where $norm(\cdot)$ is the l_1 normalization and $|b|$ is the joint number in the part combination b. τ is the temperature hyper-parameter. A bigger $v_b^c/|b|$ indicates a higher probability that the part combination b will be selected. Specifically, two possible cases are discussed:

• If the sample number of class c_i is more than c_j, a random body part combination would be sampled from $d(c_j)$, which is more likely representative for class c_j. Then the selected body parts of s_j are pasted into s_i to generate the mixed data.

• If the first case is not true, a body part combination would be sampled from $d(c_i)$. Then the actual body parts to be mixed are the complement of the sampled parts in s_j. They are pasted into s_i to preserve the crucial motion patterns in minority categories.

By virtue of this, the mixed sample is expected to be representative for the minority categories. Although the majority categories are not explicitly considered in the mixing synthesis, we assume that the model can learn desirable representations for them using sufficient training samples in datasets. As for the label of mixed data, we simply follow the same definition as discussed in Section 3.1 to make our method more general and compatible with other re-weighting methods.

3.4 Training with Shap-Mix

In *long-tailed* recognition task, we utilize the *Cross-Entropy* loss with *Balanced Softmax* [Ren *et al.*, 2020] to tackle the data imbalance problem as discussed in Section 1. During training, the proposed Shap-Mix technique is integrated, and the model optimizes the loss objective for the mixed data and original data jointly. Note that the saliency value obtained by the model can be inaccurate at the beginning of training. Therefore, we warm up the model in the first few epochs to obtain a stable estimation of Shapley-value, after which we use the estimation to guide the mixing data generation. Meanwhile, we compare our method with previous works in *balanced* setting to verify the effectiveness. Specifically, we utilize the proposed ST-Mix in Section 3.2 for efficiency because of the balanced class distributions in this setting.

4 Experiment Results

4.1 Datasets

NTU RGB+D 60 Dataset (NTU 60) [Shahroudy *et al.*, 2016]. There are 56,578 videos with skeleton data of 25 joints, divided into 60 action categories. Two evaluation protocols are recommended: a) Cross-Subject (xsub): the training data are collected from 20 subjects, while the testing data are from the other 20 subjects. b) Cross-View (xview): the front and two side captured views are used for training, while testing set includes the left and right 45-degree views.

NTU RGB+D 120 Dataset (NTU 120) [Liu *et al.*, 2019]. This is an extension to NTU 60, consisting of 114,480 videos with 120 categories. Two recommended protocols are presented: a) Cross-Subject (xsub): the training data are collected from 53 subjects, while the other 53 subjects are for testing. b) Cross-Setup (xset): the training data use even setup IDs, while testing data use odd ones.

Kinetics Skeleton 400 (Kinetics 400) [Kay *et al.*, 2017]. This dataset contains the 2D skeleton data for action recognition, extracted from the Kinetics 400 video dataset using *OpenPose* toolbox. It is the largest skeleton-based action recognition dataset, containing more than 260k training sequences over 400 classes in a long-tailed distribution. The sample number for each category ranges from 200 to 1000.

For the imbalanced setting, we construct the long-tailed datasets based on NTU 60 and 120 (**LT-NTU 60/120**), and adopt the cross-subject evaluation protocol following [Liu *et al.*, 2023]. The *imbalanced factor* (IF) is defined as the number of training samples in the largest class divided by the smallest [Cui *et al.*, 2019]. Meanwhile, Kinetics 400 is used as a real long-tailed benchmark in the wild.

4.2 Implementation Details

Our method can be applied to any backbone for skeleton-based action recognition. Specifically, CTR-GCN [Chen *et al.*, 2021a] is chosen as our backbone for comparison and we follow its training settings and data processing for fairness. For the implementation of Shap-Mix, we randomly sample 2 or 3 body parts to mix in spatial dimension. The mixed temporal length is from 40% to 70% of the original length. The temperature τ is set as 0.2, and the momentum coefficient in the EMA is 0.9. The warm-up phase is for the first 5 epochs. Due to the less data in the constructed long-tailed dataset, we increase the training epochs to 100.

4.3 Comparison on Long-Tailed Recognition

We compare the popular long-tailed recognition works including different methodological categories. Specifically, decoupling-based methods utilize two-stage training to decouple the learning of the feature extractor and classifier. Contrastive-learning-based methods apply balanced contrastive representation learning. Ensemble-based methods employ multiple experts and perform the knowledge ensemble to obtain the final predictions. Since these methods are not implemented on the skeleton data, we conduct extensive reproductive experiments to provide a benchmark. *We strictly follow the setting fairness including training epochs and utilize the same backbone*, using the official implementation code as possible.

Following [Liu *et al.*, 2023; Du *et al.*, 2023], we first report the accuracy on LT-NTU datasets of three splits of classes, Many-shot classes (training samples$>$100), Medium-shot (training samples 20$\sim$100) and Few-shot (training samples$<$20), to comprehensively evaluate our model. As we can see in Table 1, our method achieves the best overall scores compared with different competitors. Notably, compared with GLMC, which is the SOTA long-tailed augmentation method, our method shows a desirable overall performance improvement, especially for the head categories. Meanwhile, compared with the baseline, our method can largely boost the performance on the medium- and few-shot classes without compromising that on many-shot classes, which is difficult for most long-tailed methods.

Method	Category	LT-NTU 60 xsub (IF = 100)				LT-NTU 120 xsub (IF = 100)			
		Overall	Many	Medium	Few	Overall	Many	Medium	Few
Cross-Entropy Loss	-	74.4	86.4	69.5	63.8	64.2	83.6	64.9	54.7
Mixup [Zhang *et al.*, 2017]	Augmentation	75.9	86.3	71.7	66.5	66.9	**85.6**	65.6	55.5
Remix-DRW [Chou *et al.*, 2020]		78.7	86.6	74.8	72.7	69.3	83.5	67.2	62.6
FSA [Chu *et al.*, 2020]		76.8	85.4	73.3	68.8	66.7	82.4	66.8	59.4
GLMC [Du *et al.*, 2023]		<u>78.8</u>	78.6	**81.3**	**76.0**	<u>71.5</u>	79.5	<u>70.5</u>	<u>69.2</u>
BRL [Liu *et al.*, 2023]		76.9	85.2	73.1	70.0	66.3	84.2	65.0	59.9
ROS [Van Hulse *et al.*, 2007]	Reweighting & Resampling	74.8	86.1	68.1	57.9	61.0	81.3	62.3	50.7
Focal Loss [Lin *et al.*, 2017]		77.6	83.1	75.9	72.0	69.4	81.7	66.7	65.2
CB Loss [Cui *et al.*, 2019]		72.4	78.4	71.2	65.3	63.2	76.0	65.0	56.1
LDAM-DRW [Cao *et al.*, 2019]		76.4	83.7	73.4	69.9	66.8	80.9	65.8	61.2
Balanced Softmax (BS) [Ren *et al.*, 2020]		77.6	83.8	73.9	73.3	69.6	81.4	67.8	66.7
IB Loss [Park *et al.*, 2021]		76.0	84.0	74.4	66.5	67.8	81.2	67.6	62.4
BS+Max Norm [Alshammari *et al.*, 2022]	Decoupling	77.5	81.2	75.9	74.1	70.3	80.0	68.2	68.8
PaCo [Cui *et al.*, 2021]	Contrastive Learning	76.6	82.0	76.4	69.1	67.9	82.2	66.2	63.8
BCL [Zhu *et al.*, 2022]		77.3	84.4	74.1	71.5	66.9	82.3	64.5	62.8
RIDE (3 experts) [Wang *et al.*, 2021]	Ensemble	76.6	<u>86.7</u>	71.9	68.4	65.2	<u>85.4</u>	66.7	53.8
Shap-Mix (Ours)	Augmentation	**80.8**	**86.8**	<u>78.4</u>	<u>75.6</u>	**73.0**	84.8	**71.3**	**69.7**

Table 1: Performance comparison of long-tailed skeleton-based action recognition with single joint stream. IF is the imbalance factor. Top-1 accuracy (%) is reported. The results with **bold** and <u>underline</u> indicate the highest and second-highest value.

Method	Baseline	PoseConv3D	BRL†	GLMC†	Ours†
Acc. (%)	45.0	46.0	45.6	46.6	**48.4**

Table 2: Comparison results on Kinetics 400 of single joint stream. † indicates the long-tailed methods.

Method	IF = 10 (%)	IF =50 (%)	IF = 100 (%)
Baseline	79.6	70.1	64.2
Remix-DRW	81.6	74.4	69.3
GLMC	82.0	75.9	71.5
Balanced Softmax	80.7	74.2	69.6
Shap-Mix (Ours)	**83.0**	**76.8**	**73.0**

Table 3: The results under LT-NTU 120 dataset of different imbalance factors (IF).

We also conduct the experiments on Kinetics 400, which is a long-tailed dataset in the wild, as shown in Table 2. Compared with PoseConv3D [Duan *et al.*, 2022] and GLMC, which are the SOTAs in standard supervised and long-tailed learning methods, Shap-Mix achieves the best scores owing to the utilization of the crucial motion patterns.

Finally, we present the results with different imbalance factors in Table 3. Our method can improve both general representation quality and decision boundary learning for tail categories, achieving the best scores across different imbalance factors compared with other long-tailed methods.

4.4 Comparison on Balanced Recognition

Here we show the effectiveness of our method under balanced setting. The model is trained on the balanced (original) NTU datasets to give a fair comparison with previous skeleton-based action recognition methods. We first compare our method with the state-of-the-art methods, including SGN [Zhang *et al.*, 2020], Shift-GCN [Cheng *et al.*, 2020], MS-G3D [Liu *et al.*, 2020b], MST-GCN [Chen *et al.*, 2021b], CTR-GCN [Chen *et al.*, 2021a], Info-GCN [Chi *et al.*, 2022], EfficientGCN [Song *et al.*, 2022], FR-Head [Zhou *et al.*, 2023], StreamGCN [Yang *et al.*, 2023], HD-GCN [Lee *et al.*, 2023b], STC-Net [Lee *et al.*, 2023a]. Following previous works, we report the results of multi-stream fusion, *i.e.*, joint, bone (2-stream), joint motion and bone motion (4-stream). As shown in Table 4, our method achieves the best performance across different datasets. **Remarkably**, our method with 4 (/2) streams can outperform many latest methods with 6 (/4) streams, verifying the significant effectiveness.

Method	Year	NTU 60 (%)		NTU 120 (%)	
		xsub	xview	xsub	xset
2s-SGN	CVPR'20	89.0	94.5	79.2	81.5
4s-Shift-GCN	CVPR'20	90.7	96.5	85.9	87.6
2s-MS-G3D	CVPR'20	91.5	96.2	86.9	88.4
4s-MST-GCN	AAAI'21	91.5	96.6	87.5	88.8
4s-CTR-GCN	ICCV'21	92.4	96.8	88.9	90.6
4s-Info-GCN	CVPR'22	92.7	96.9	89.4	90.7
6s-Info-GCN	CVPR'22	93.0	**97.1**	89.8	91.2
3s-EfficientGCN	TPAMI'22	91.7	95.7	88.3	89.1
4s-FR-Head	CVPR'23	92.8	96.8	89.5	90.9
6s-StreamGCN	IJCAI'23	92.9	96.9	89.7	91.0
4s-HD-GCN	ICCV'23	93.0	97.0	89.8	91.2
4s-STC-Net	ICCV'23	93.0	**97.1**	89.9	91.3
2s-Ours	-	93.4	96.8	90.2	91.6
4s-**Ours**	-	**93.7**	**97.1**	**90.4**	**91.7**

Table 4: Performance comparison of balanced recognition on NTU datasets in top-1 accuracy. ∗s- means the fusion results of ∗ streams.

Meanwhile, our augmentation method can be equipped with different backbones. Compared with another model-

Method	Params.	xsub (%)	xset (%)
2s-AGCN	3.80M	84.3	85.9
+ FR Head	4.33M	84.6$^{\uparrow 0.3}$	**86.6**$^{\uparrow 0.7}$
+ Ours	3.80M	84.6$^{\uparrow 0.3}$	86.5$^{\uparrow 0.6}$
CTR-GCN	1.46M	84.5	86.6
+ FR Head	1.99M	85.5$^{\uparrow 1.0}$	87.3$^{\uparrow 0.7}$
+ Ours	1.46M	**86.9**$^{\uparrow 2.4}$	**88.3**$^{\uparrow 1.7}$
Info-GCN	1.58M	85.1	86.3
+ FR Head	2.11M	-	-
+ Ours	1.58M	**85.7**$^{\uparrow 0.6}$	**88.0**$^{\uparrow 1.7}$
HD-GCN	1.68M	85.1	87.2
+ FR Head	2.21M	85.4$^{\uparrow 0.3}$	87.7$^{\uparrow 0.5}$
+ Ours	1.68M	**87.0**$^{\uparrow 1.9}$	**88.8**$^{\uparrow 1.6}$

Table 5: Performance of our proposed method using different backbones with single joint stream on NTU 120 dataset. The results of previous work, FR-Head, are given for comparison.

Method	xsub
Baseline	89.9
Mixup	90.8
Cut-Mix	90.7
Mix [Xu *et al.*, 2022]	90.1
Mix [Zhan *et al.*, 2022]	90.7
ST-Mix (Ours)	**91.6**

Table 6: Comparison with other mixing methods. The latter two are also designed for skeleton.

Spa.	Temp.	xsub
✓		91.1
	✓	90.1
✓	✓	**91.6**

Table 7: Ablation study on spatial-temporal mixing design in ST-Mix. "Spatial" is mixing in the part-level or joint-level. "Temporal" denotes to copy directly or after randomly down-sampling.

agnostic method, FR-Head [Zhou *et al.*, 2023], our method can bring consistent performance improvement without introducing additional training parameters as shown in Table 5.

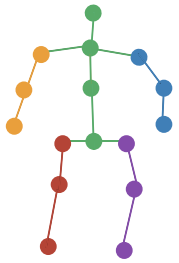
4.5 Ablation Studies

Next we present the ablation results conducted on the (LT) NTU 60 dataset under cross-subject protocol.

Effectiveness of ST-Mix Design. We compare our ST-Mix with different mixing methods for skeleton data in Table 6. The reported results are obtained under the balanced action recognition with the full NTU 60 dataset. Compared with Mixup and Cut-Mix, our method jointly learns these two mixture patterns, yielding a notable performance improvement. Meanwhile, due to the well-designed spatial-temporal mixing policy, our method outperforms other mixing methods, with their respective effects presented in Table 7.

Visualization of Saliency Estimation. We choose the part combination containing 2 or 3 parts and estimate the corresponding importance. As shown in Figure 3, we can obtain a rational skeleton saliency estimation for many-, medium-, and few-shot classes. For example, in *handshaking* action, the most salient part combination is obtained as the two legs with the right arm, which is reasonable because people usually stand with their hands out in this action. Meanwhile, it can be found hands and arms play an important role in most

Part: A: Torso, B / C: Right / Left Arm, D / E: Right / Left Leg



clapping:	B-C, A-B, B-E
hand waving:	B-E, B-C, B-D
kick something:	D-E, B-D, C-D
jumping:	B-D, A-D-E, A-B-D
salute:	B-D, A-B, B-E
kick other person:	D-E, A-D, A-D-E
slap other person:	A-B, B-C, B-D
handshaking:	B-D-E, A-B-D, B-E

Figure 3: The visualization of our saliency estimation, in the form of *action*, the first, second, and third most salient part combination. We choose the actions from many- (first 2), medium- (3-5), and few-shot (last 3) classes.

Mixing Guidance	RS-Acc. (%)	RW-Acc. (%)
Random	76.3	80.0
Grad-CAM	76.4	79.0
Shapley value	**77.9**	**80.8**

Table 8: Ablation study on the mixing with different guidance. The re-sampling (RS-) and re-weighting (RW-) techniques are applied to show the effect, respectively.

human actions, especially the dominant (right) hand of most people. These results are promising for more fine-grained action recognition and spatial localization. We hope that more works will emerge in the future to utilize and explore skeletal saliency maps.

Effectiveness of Shapley Value Guidance. The results of different guidance for mixing augmentation are provided in Table 8. As we can see, the model outperforms the mixing method without guidance either using re-samping or re-weighting method. Meanwhile, we compare with another guidance, Grad-CAM [Selvaraju *et al.*, 2017]. However, due to the over-smoothing problem, it can not achieve desirable performance improvement. In contrast, we use Shapley value as guidance, which is based on input and well considers the relationship between different joints, and brings further performance improvement.

5 Limitation and Conclusion

In this paper, we explore the long-tailed skeleton-based action recognition, which has been largely overlooked before, and propose a novel augmentation-based method, Shap-Mix. Specifically, we develop a saliency estimation method based on Shapley value and a tail-aware mixing policy to preserve more representative motion patterns, improving decision boundary learning of tail classes. One limitation is that due to the inherent computational complexity of Shapley value, our method is primarily for sparse data such as skeleton. This can be alleviated by performing estimation every n iterations, which is a trade-off between estimation accuracy and computation overhead. Extensive experiments on both balanced and long-tailed settings verify the effectiveness.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62172020, and in part by the State Key Laboratory of Media Audio and Video (Communication University of China), Ministry of Education, China.

References

[Alshammari *et al.*, 2022] Shaden Alshammari, Yuxiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *IEEE CVPR*, 2022.

[Cao *et al.*, 2019] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS*, 2019.

[Chen *et al.*, 2021a] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*, 2021.

[Chen *et al.*, 2021b] Zhan Chen, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In *AAAI*, 2021.

[Chen *et al.*, 2022] Xiaohua Chen, Yucan Zhou, Dayan Wu, Wanqian Zhang, Yu Zhou, Bo Li, and Weiping Wang. Imagine by reasoning: A reasoning-based implicit semantic data augmentation for long-tailed classification. In *AAAI*, 2022.

[Cheng *et al.*, 2020] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *IEEE CVPR*, 2020.

[Chi *et al.*, 2022] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *IEEE CVPR*, 2022.

[Chou *et al.*, 2020] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. 2020.

[Chu *et al.*, 2020] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *ECCV*, 2020.

[Cui *et al.*, 2019] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *IEEE CVPR*, 2019.

[Cui *et al.*, 2021] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *ICCV*, 2021.

[Damen *et al.*, 2022] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 130:33–55, 2022.

[Deng and Papadimitriou, 1994] Xiaotie Deng and Christos H Papadimitriou. On the complexity of cooperative solution concepts. *Mathematics of operations research*, 19(2):257–266, 1994.

[Du *et al.*, 2023] Fei Du, Peng Yang, Qi Jia, Fengtao Nan, Xiaoting Chen, and Yun Yang. Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In *IEEE CVPR*, 2023.

[Duan *et al.*, 2022] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *IEEE CVPR*, 2022.

[Han *et al.*, 2005] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*, 2005.

[Hart, 2016] S Hart. A bibliography of cooperative games: Value theory,[e-source], 2016.

[Hsieh *et al.*, 2021] Ting-I Hsieh, Esther Robb, Hwann-Tzong Chen, and Jia-Bin Huang. Droploss for long-tail instance segmentation. In *AAAI*, 2021.

[Hua *et al.*, 2023] Yilei Hua, Wenhan Wu, Ce Zheng, Aidong Lu, Mengyuan Liu, Chen Chen, and Shiqian Wu. Part aware contrastive learning for self-supervised action recognition. *arXiv preprint arXiv:2305.00666*, 2023.

[Kang *et al.*, 2019] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

[Kay *et al.*, 2017] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[Ke *et al.*, 2017] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *IEEE CVPR*, 2017.

[Lee *et al.*, 2023a] Jungho Lee, Minhyeok Lee, Suhwan Cho, Sungmin Woo, Sungjun Jang, and Sangyoun Lee. Leveraging spatio-temporal dependency for skeleton-based action recognition. In *ICCV*, 2023.

[Lee *et al.*, 2023b] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *ICCV*, 2023.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[Liu *et al.*, 2017] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.

[Liu *et al.*, 2019] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE TPAMI*, 42(10):2684–2701, 2019.

[Liu *et al.*, 2020a] Meng Liu, Hongyang Gao, and Shuiwang Ji. Towards deeper graph neural networks. In *ACM SIGKDD*, 2020.

[Liu *et al.*, 2020b] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *IEEE CVPR*, 2020.

[Liu *et al.*, 2022] Zicheng Liu, Siyuan Li, Di Wu, Zihan Liu, Zhiyuan Chen, Lirong Wu, and Stan Z Li. Automix: Unveiling the power of mixup for stronger classifiers. In *ECCV*, 2022.

[Liu *et al.*, 2023] Hongda Liu, Yunlong Wang, Min Ren, Junxing Hu, Zhengquan Luo, Guangqi Hou, and Zhenan Sun. Balanced representation learning for long-tailed skeleton-based action recognition. *arXiv preprint arXiv:2308.14024*, 2023.

[Park *et al.*, 2021] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *ICCV*, 2021.

[Plizzari *et al.*, 2021] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *CVIU*, 208:103219, 2021.

[Qin *et al.*, 2020] Jie Qin, Jiemin Fang, Qian Zhang, Wenyu Liu, Xingang Wang, and Xinggang Wang. Resizemix: Mixing data with preserved object information and true labels. *arXiv preprint arXiv:2012.11101*, 2020.

[Ren *et al.*, 2020] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *NeurIPS*, 2020.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[Shahroudy *et al.*, 2016] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *IEEE CVPR*, 2016.

[Shapley and others, 1953] Lloyd S Shapley et al. A value for n-person games. 1953.

[Shen *et al.*, 2016] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *ECCV*, 2016.

[Shi *et al.*, 2019] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *IEEE CVPR*, 2019.

[Shi *et al.*, 2020] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action recognition, 2020.

[Song *et al.*, 2017] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, 2017.

[Song *et al.*, 2018a] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. Skeleton-indexed deep multi-modal feature learning for high performance human action recognition. In *IEEE ICME*, 2018.

[Song *et al.*, 2018b] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. Spatio-temporal attention-based lstm networks for 3d action recognition and detection. *IEEE TIP*, pages 3459–3471, 2018.

[Song *et al.*, 2022] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE TPAMI*, 45(2):1474–1488, 2022.

[Tan *et al.*, 2020] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *IEEE CVPR*, 2020.

[Van Hulse *et al.*, 2007] Jason Van Hulse, Taghi M Khoshgoftaar, and Amri Napolitano. Experimental perspectives on learning from imbalanced data. In *ICML*, 2007.

[Walawalkar *et al.*, 2020] Devesh Walawalkar, Zhiqiang Shen, Zechun Liu, and Marios Savvides. Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification. *arXiv preprint arXiv:2003.13048*, 2020.

[Wang *et al.*, 2021] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2021.

[Xu *et al.*, 2022] Kailin Xu, Fanfan Ye, Qiaoyong Zhong, and Di Xie. Topology-aware convolutional neural network for efficient skeleton-based action recognition. In *AAAI*, 2022.

[Yan *et al.*, 2018] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, pages 7444–7452, 2018.

[Yang *et al.*, 2023] Yuheng Yang, Haipeng Chen, Zhenguang Liu, Yingda Lyu, Beibei Zhang, Shuang Wu, Zhibo Wang, and Kui Ren. Action recognition with multi-stream motion modeling and mutual information maximization. *arXiv preprint arXiv:2306.07576*, 2023.

[Yun *et al.*, 2019] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

[Zhan *et al.*, 2022] Chen Zhan, Liu Hong, Guo Tianyu, Chen Zhengyan, Song Pinhao, and Tang Hao. Contrastive learning from spatio-temporal mixed skeleton sequences for self-supervised skeleton-based action recognition. In *arXiv: 2207.03065*, 2022.

[Zhang *et al.*, 2017] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2017.

[Zhang *et al.*, 2020] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *IEEE CVPR*, 2020.

[Zhang *et al.*, 2021] Xing Zhang, Zuxuan Wu, Zejia Weng, Huazhu Fu, Jingjing Chen, Yu-Gang Jiang, and Larry S Davis. Videolt: Large-scale long-tailed video recognition. In *ICCV*, 2021.

[Zhang *et al.*, 2023] Jiahang Zhang, Lilang Lin, and Jiaying Liu. Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations. In *AAAI*, 2023.

[Zhou *et al.*, 2023] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Learning discriminative representations for skeleton based action recognition. In *IEEE CVPR*, 2023.

[Zhu *et al.*, 2022] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *IEEE CVPR*, 2022.