Learning Dynamics as Feedback: An Adaptive Entropy Flow Dynamics Framework for Long-tailed Human Action Recognition

Yuan Dong[1]*, Zhe Zhao[1,3], Liheng Yu[1], Di Wu[1]†, Pengkun Wang[1,2]†

[1]University of Science and Technology of China, Hefei 230026, China
[2]Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou 215123, China
[3]City University of Hong Kong
dongyuan01221@gmail.com, {zz4543,yuliheng,wdcxy}@mail.ustc.edu.cn, pengkun@ustc.edu.cn

Abstract

Deep human action recognition models trained on real-world data are often challenged by long-tailed distributions, where performance on rare classes is severely degraded. Current solutions typically apply static or heuristic interventions that are disconnected from the model's evolving internal state. To overcome this limitation, we reconceptualize long-tailed human action recognition as a closed-loop, self-regulating system, inspired by ecological theory. We further introduce an Adaptive Ecological Entropy Dynamics (AEED) framework, which is built upon three synergistic components. First, AEED perceives the learning state through entropy flow, providing a robust and directional signal of learning progress. Second, this signal drives an adaptation mechanism, which dynamically adjusts class-specific loss weights to allocate more learning resources to underperforming classes. Finally, AEED facilitates intelligent knowledge transfer via Confidence-Guided Symbiosis (CS-Mix). Extensive experiments demonstrate that AEED achieves state-of-the-art performance on challenging skeleton-based action recognition benchmarks, including NTU-60-LT and Kinetics-400-LT.

Code — https://github.com/ddddddy1221/AEED

Introduction

The inherent imbalance of natural ecosystems, where a few dominant species thrive while many rare ones struggle for survival (Magurran 2021; Wang et al. 2024b; Zhao et al. 2024b), provides a powerful analogy for a critical challenge in human action recognition (Zhang et al. 2023). Real-world action datasets are similarly long-tailed: common activities like 'walking' create a high-frequency head, while a vast number of critical yet infrequent events, such as 'saving a person' or a 'medical emergency', form a data-scarce tail. Models trained with standard empirical risk minimization are prone to developing a strong bias towards these data-rich head classes (Wang et al. 2024a; Zhao et al. 2024c). Consequently, they exhibit impressive performance on common actions but falter on rare ones, leading to poor generalization and diminished reliability in practical applications (Van Horn and Perona 2017).

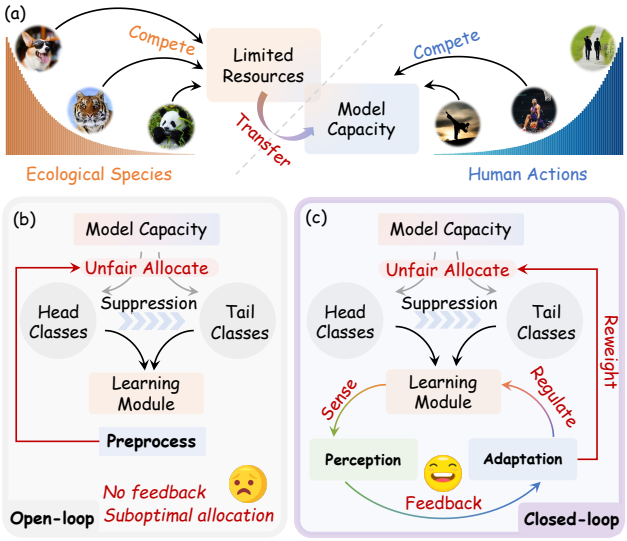

Figure 1: Conceptual overview of our AEED framework, a paradigm shift from open-loop to closed-loop learning. **(a)** We frame long-tailed learning as an ecological competition for limited model capacity. **(b)** Prior open-loop methods use static rules, leading to unfair resource allocation that suppresses tail classes. **(c)** In contrast, our closed-loop approach uses a `Perception-Adaptation` feedback loop to sense the learning state and dynamically regulate resources, enabling autonomous self-regulation.

To address this challenge, a variety of methods have been proposed, often focusing on re-balancing the data distribution (Cui et al. 2019), facilitating knowledge transfer (Zhang et al. 2024), or decoupling training stages (Kang et al. 2020). However, despite their progress, these approaches are fundamentally constrained by an ***open-loop*** design philosophy. Their core interventions are typically static, pre-programmed based on initial dataset statistics or fixed heuristics like epoch counts. Consequently, they treat the model as a passive system, applying rules that are blind to its real-time, evolving cognitive state. ***This inherent lack of feedback prevents the model from intelligently self-regulating its focus, leading to suboptimal resource allo-***

*Work done during a research internship at USTC-DILab.
†Corresponding author.

cation and inefficient learning.

Here, we challenge this passive view by reconceptualizing long-tailed human action recognition as a ***closed-loop***, self-regulating system, as illustrated in Figure 1. Whereas prior methods operate in an open loop, our approach is designed around a continuous feedback mechanism that allows the model to intelligently guide its own learning process. To achieve this, we introduce the **Adaptive Ecological Entropy Dynamics (AEED)** framework, which is built upon three synergistic components that together establish this dynamic feedback loop.

The first component is responsible for ***perception***. For a system to self-regulate, it must first sense its internal state. AEED achieves this by monitoring not just static uncertainty, but its rate of change. We introduce `Entropy Flow`, the temporal derivative of class information entropy, as a robust and directional signal for learning dynamics. A negative flow provides an unambiguous indication of progress, while a positive flow signals growing confusion and triggers the need for intervention. This serves as the critical feedback signal for our system.

This perceptual signal then drives our core feedback controller, which we term Entropy-driven ***Adaptation*** (**E-Adapt**). Inspired by the Lotka-Volterra logistic growth model from mathematical ecology, AEED dynamically adjusts class-specific loss weights, which we conceptualize as `niche widths`. When a class exhibits a positive entropy flow, indicating a learning struggle, its weight is automatically increased to allocate more learning resources. The direct coupling of Entropy Flow perception to niche width adaptation creates a closed-loop feedback system, enabling the model to autonomously correct its learning trajectory.

Finally, AEED facilitates intelligent ***interaction*** to improve knowledge sharing. Standard mixing augmentation often fails because physical interpolation does not guarantee semantic coherence. Our Confidence-Guided Symbiosis (**CS-Mix**) strategy addresses this flaw. It acts as a semantic arbiter, probing the model's own belief state via KL-divergence to determine the semantic kinship of a mixed sample to its parents. This process yields a more reliable supervisory signal, ensuring that knowledge is transferred in a semantically coherent manner. We summarize our contributions as follows:

- ***Brand-new Paradigm:*** For the first time, we propose to reconceptualize long-tailed human action recognition as a *closed-loop, self-regulating* system, breaking the *static and open-loop* limitations of existing methods.
- ***Adaptive Framework:*** We introduce the AEED framework to operationalize this paradigm with two key innovations: E-Adapt for adaptive perception and CS-Mix for intelligent interaction.
- ***Theoretical Analysis:*** We provide a rigorous theoretical proof of the global asymptotic stability of our system's dynamics using LaSalle's Invariance Principle, ensuring robust and predictable convergence.
- ***Generalization Performance:*** We achieve SOTA performance on major long-tailed action recognition benchmarks and demonstrate the framework's strong cross-modal generality on vision and text tasks.

Related Work

Prevailing long-tailed recognition methods are effectively open-loop systems using static or heuristically timed interventions. Strategies like re-weighting the loss (Cui et al. 2019; Cao et al. 2019; Lin et al. 2017) or logits (Ren et al. 2020; Menon et al. 2021; Zhao et al. 2025) apply interventions predetermined by initial dataset statistics. For instance, a rare class's weight remains static throughout training, regardless of whether the model has mastered or struggles with it, making the system oblivious to real-time learning dynamics. Similarly, decoupled learning (Kang et al. 2020; Li et al. 2021) imposes a rigid, multi-stage schedule governed by a preset epoch count. This acts as a coarse proxy for feature maturity, blind to the actual quality of learned representations. Fundamentally, these methods treat the model as a passive system, lacking the feedback mechanism for genuine self-regulation.

A second line of work focuses on knowledge transfer via data augmentation, primarily through mixing samples from different classes (Zhang et al. 2018; Chou et al. 2020). The core limitation here is the reliance on physical interpolation as a proxy for semantic coherence, a strong and often flawed assumption that can lead to noisy or contradictory supervisory signals (Guo, Mao, and Zhang 2019; Yun et al. 2019). Even sophisticated strategies that guide the mixing process based on feature importance or saliency (Zhang et al. 2024; Uddin et al. 2021) often operate with a static mixing policy, failing to dynamically adjust based on the model's real-time mastery of the classes involved. Our work departs fundamentally from both paradigms. While some methods have explored adaptive re-weighting, they typically rely on an external, balanced meta-set to guide the process (Ren et al. 2020). In contrast, AEED internalizes the control loop, empowering the model to perceive its own learning state.

Methodology

AEED consists of three interconnected modules. First, a `Perception` module perceives the model's cognitive state by tracking class-wise Entropy Flow. This signal then drives an `Adaptation` module, which dynamically allocates learning resources by evolving class-specific weights. Finally, an `Interaction` module, CS-Mix, intelligently enriches the data environment to improve knowledge transfer.

Notation and Preliminaries

We consider a training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ over C classes. Our framework augments a backbone model, parameterized by θ, by operating on its output posterior probabilities $p_\theta(y|x_i)$, which are typically obtained via a softmax function. Our method introduces a set of class-wise dynamic variables that evolve over training epochs t: the mean class entropy $H_c^{(t)}$, the smoothed **Entropy Flow** $\Delta\bar{H}_c^{(t)}$, and the **niche width** $w_c^{(t)}$, which serves as a dynamic loss weight. The interplay between these variables forms our self-regulating system, detailed in the following sections.

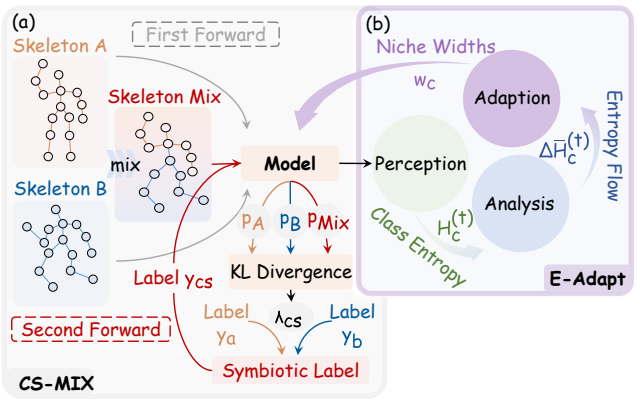

Figure 2: The synergistic workflow of the AEED framework's core modules. **(a) CS-Mix** replaces the arbitrary physical mixing ratio with a model-aware semantic ratio (λ_{CS}), computed via KL-divergence on the model's own posteriors (P_A, P_B, P_{Mix}) to generate a valid symbiotic label. **(b) E-Adapt** forms a feedback loop where the `Perception` stage calculates class entropy ($H_c^{(t)}$). Its temporal trend, the `Entropy Flow` ($\Delta\bar{H}_c^{(t)}$), then guides the `Adaptation` stage to dynamically adjust Niche Widths (w_c) for learning regulation.

`Perception` Module: Sensing Learning Dynamics via Entropy Flow

A core limitation of conventional long-tailed methods is their open-loop nature, as they lack a mechanism to sense the model's evolving internal state. To enable adaptive learning, a system must first perceive its own condition. The AEED perception module provides this crucial capability by generating a real-time feedback signal that reflects the model's cognitive state for each class.

The AEED perception module achieves this by monitoring the model's cognitive uncertainty. We quantify this uncertainty using the per-sample conditional entropy. For a given sample x, the per-sample conditional entropy is:

$$H(Y|x) = -\sum_{k=1}^{C} p_\theta(y=k|x) \log p_\theta(y=k|x). \quad (1)$$

Since our goal is to assess the learning status for an entire class, not just individual samples, we aggregate this information. At each epoch t, we compute the mean conditional entropy for each class c, denoted as $H_c^{(t)}$, by averaging $H(Y|x)$ over all training samples belonging to that class seen during that epoch:

$$H_c^{(t)} = \frac{1}{|S_c^{(t)}|} \sum_{x_i \in S_c^{(t)}} H(Y|x_i), \quad (2)$$

where $S_c^{(t)}$ is the set of samples from class c processed in epoch t. This value, $H_c^{(t)}$, represents the model's average confusion about class c at a specific point in training.

However, using the time-averaged absolute entropy $H_c^{(t)}$ as a direct signal is problematic due to its inherent ambiguity. A high $H_c^{(t)}$ could mean either that a class is intrinsically difficult (e.g., 'running' vs. 'jogging'), or that the model is simply in an early stage of learning it.

AEED resolves this ambiguity by focusing not on the *state* of uncertainty, but on its *trend*. We introduce **Entropy Flow**, the temporal derivative of entropy, as our core perceptual signal. A negative flow ($\Delta H_c < 0$) unambiguously indicates that the model's confusion is decreasing and learning is progressing. A positive flow ($\Delta H_c > 0$) is a clear alarm that learning is deteriorating, signaling a need for intervention. To ensure robustness against the stochasticity of mini-batch training, we use an Exponential Moving Average (EMA) to compute a smoothed, low-variance Entropy Flow:

$$\Delta\bar{H}_c^{(t)} = (1-\beta)(H_c^{(t)} - H_c^{(t-1)}) + \beta\Delta\bar{H}_c^{(t-1)}, \quad (3)$$

where $H_c^{(t)}$ is the mean entropy for class c in epoch t, and β is a smoothing factor. This smoothed signal provides a stable and directional measure of learning dynamics. Further justification for this choice and a formal guarantee of the reliability of the entropy estimate are provided in **Appendix A**.

`Adaptation` Module: Niche Width Evolution via Entropy-driven Adaptation

The Entropy Flow signal generated by the perception module serves as the primary driver for the adaptation module. The objective of this module is to dynamically allocate learning resources to classes where they are most needed.

We conceptualize the learning resource allocated to each class as its **niche width**, a term from ecology representing a species' resource utilization. In our framework, this is implemented as a dynamic, class-specific loss weight, w_c. A principled update rule for these weights should satisfy two key properties: (i) *responsiveness* to the learning signal and (ii) *self-limitation* to prevent instability. The update rule for the niche widths w_c is designed to be responsive to the learning signal while ensuring stability through self-limitation. To achieve this, we adapt the logistic growth model, a component of the *Lotka-Volterra* family of equations from mathematical ecology. The governing equation is defined as:

$$w_c^{(t+1)} = w_c^{(t)} + \eta \cdot w_c^{(t)} \left(\Delta\bar{H}_c^{(t)} - \gamma \cdot w_c^{(t)} \right), \quad (4)$$

where $\eta > 0$ is the adaptation rate and $\gamma > 0$ is a coefficient for the self-limiting term.

Two terms govern the dynamics: (i) $\eta w_c^{(t)} \Delta\bar{H}_c^{(t)}$ adjusts the weight proportionally to the Entropy Flow signal, providing responsiveness. (ii) $-\eta\gamma(w_c^{(t)})^2$ is a quadratic self-limitation factor. This factor, analogous to intra-species competition in ecological models, prevents the unbounded growth of w_c and is a necessary condition for the stability of the system. The perception and adaptation modules collectively form a synergistic feedback controller, which we term Entropy-driven Adaptation **(E-Adapt)**. The workflow of this feedback loop is visualized in Figure 2(b). The principled design of E-Adapt, particularly its self-limitation mech-

anism, ensures that the niche width dynamics are provably stable, as formalized in Theorem 1.

Theorem 1 (Boundedness of niche widths)**.** *The sequence of niche widths $\{w_c^{(t)}\}$ generated by Eq.* (4) *remains within a bounded positive interval for all $t \geq 0$, provided a sufficiently small η is chosen and the weights are initialized appropriately.*

Sketch. A detailed proof via mathematical induction is provided in **Appendix B**. The proof establishes a formal condition on the hyperparameter η that guarantees the weights remain positive, while the self-regulating quadratic term and a practical clipping mechanism ensure an upper bound. □

`Interaction` Module: Confidence-Guided Symbiosis

Beyond adaptive resource allocation, effective long-tailed learning requires robust mechanisms for knowledge transfer from data-rich head classes to data-scarce tail classes. While data mixing strategies like Mixup are popular, they often operate on a flawed premise that linear interpolation in the input space corresponds to a meaningful semantic mix. This can lead to noisy or even contradictory supervisory signals. Our CS-Mix strategy addresses this fundamental issue by letting the model itself act as a semantic arbiter. The process is detailed in Figure 2(a).

The process begins by forming a physically mixed sample, $x_{\text{mix}} = \lambda_{\text{phys}} x_a + (1-\lambda_{\text{phys}}) x_b$, from two parent samples, (x_a, y_a) and (x_b, y_b), where λ_{phys} is drawn from a beta distribution. Crucially, we discard this arbitrary physical ratio for label generation. Instead, we probe the model's current belief system by performing a forward pass to obtain the posterior probability distributions P_a, P_b, and P_{mix}. We then compute a *semantic ratio*, λ_{CS}, based on the informational distance between the mixed sample's posterior and those of its parents, quantified using *Kullback-Leibler* (KL) divergence:

$$\lambda_{\text{CS}} = \frac{\text{KL}(P_{\text{mix}} \| P_a)}{\text{KL}(P_{\text{mix}} \| P_a) + \text{KL}(P_{\text{mix}} \| P_b) + \varepsilon}. \tag{5}$$

Here, ε is a small constant for numerical stability. This formulation has a clear information-theoretic interpretation: as the mixed sample's posterior P_{mix} becomes informationally closer to parent P_a, the numerator $\text{KL}(P_{\text{mix}} \| P_a)$ approaches zero, causing λ_{CS} to also approach zero. This implies that λ_{CS} effectively represents the semantic contribution of parent b. Thus, the weight $(1 - \lambda_{\text{CS}})$ for parent a increases as their posteriors align. By using this semantic ratio to interpolate the labels, we generate a supervisory signal grounded in the model's own understanding, yielding better-calibrated soft labels for more robust knowledge transfer.

The AEED Learning Objective and Algorithm

The perception, adaptation, and interaction modules are unified into a single, cohesive learning objective. For each original sample (x_i, y_i) and its paired parent $(x_{i'}, y_{i'})$, the total loss contribution is a weighted sum of two components. The first component is the standard cross-entropy loss for the original sample, weighted by its class's current niche width $w_{y_i}^{(t)}$. The second component is the cross-entropy loss for the symbiotic sample $x_{\text{mix}}^{(i)}$, which is supervised by the soft label $\mathbf{y}_{\text{mix}}^{(i)} = (1 - \lambda_{\text{CS}}^{(i)})\mathbf{y}_{i,\text{onehot}} + \lambda_{\text{CS}}^{(i)}\mathbf{y}_{i',\text{onehot}}$. This symbiotic component is in turn weighted by an intelligently interpolated niche width, $w_{\text{mix}}^{(i)} = (1 - \lambda_{\text{CS}}^{(i)}) w_{y_i}^{(t)} + \lambda_{\text{CS}}^{(i)} w_{y_{i'}}^{(t)}$. The final loss for the i-th pair in the batch is thus:

$$\mathcal{L}_i = w_{y_i}^{(t)} \cdot \mathcal{L}_{\text{CE}}(f_\theta(x_i), y_i) + w_{\text{mix}}^{(i)} \cdot \mathcal{L}_{\text{CE}}(f_\theta(x_{\text{mix}}^{(i)}), \mathbf{y}_{\text{mix}}^{(i)}). \tag{6}$$

This objective ensures that each optimization step addresses the specific needs of individual classes via dynamic weighting, while integrating shared knowledge from semantically valid interactions. The complete end-to-end training procedure is detailed in Algorithm 1 in **Appendix C**.

Theoretical Analysis of System Stability

Our central claim is that AEED constitutes a well-behaved, stable dynamical system, rather than a mere collection of heuristics. We provide a formal analysis to substantiate it, ensuring our framework is predictable and converges reliably without diverging or oscillating uncontrollably.

To formalize this analysis, we define the complete state of our system at epoch t by the state vector $\mathbf{z}^{(t)} \in \mathbb{R}^{2C}$. This vector is the concatenation of two component vectors: $\mathbf{z}^{(t)} = [\mathbf{w}^{(t)T}, \Delta\bar{\mathbf{H}}^{(t)T}]^T$, where $\mathbf{w}^{(t)}$ is the vector of all class-wise niche widths and $\Delta\bar{\mathbf{H}}^{(t)}$ is the vector of all corresponding smoothed entropy flows. Our analysis begins by establishing the boundedness of this state vector's trajectory.

Proposition 1 (Boundedness of Trajectories)**.** *The state vector of the AEED system, $\mathbf{z}^{(t)} = [\mathbf{w}^{(t)T}, \Delta\bar{\mathbf{H}}^{(t)T}]^T$, is confined to a compact set $\Omega \subset \mathbb{R}^{2C}$ for all epochs t.*

Sketch. The boundedness of the niche widths $\mathbf{w}^{(t)}$ is guaranteed by Theorem 1. The boundedness of the entropy flow $\Delta\bar{\mathbf{H}}^{(t)}$ follows directly from the definition of information entropy on a finite probability space. A more detailed argument is provided in **Appendix D (Section D.1)**. □

With state boundedness established as a prerequisite, we can prove the system's convergence to its ideal equilibrium using principles from nonlinear systems theory.

Theorem 2 (Asymptotic Stability)**.** *The AEED system, as governed by the coupled dynamics in Eq.* (3) *and Eq.* (4)*, is globally asymptotically stable with respect to the equilibrium point $\mathbf{z}^* = (\mathbf{w}^* = \mathbf{0}, \Delta\bar{\mathbf{H}}^* = \mathbf{0})$.*

Sketch. The proof relies on *LaSalle's Invariance Principle* and is detailed in **Appendix D** (Section D.2). The core logic proceeds in three steps. First, we establish the existence of a suitable Lyapunov function $V(\mathbf{z})$ whose change ΔV is non-positive within the compact set Ω, a property guaranteed by the system's inherent dissipative mechanisms. Second, according to LaSalle's principle, the system trajectory must converge to the largest invariant set where this energy dissipation ceases. Third, we analytically determine this set and show that it consists of a single point: the ideal equilibrium $\mathbf{z}^*$. Therefore, the system is guaranteed to converge, ensuring stable and predictable training behavior. □

Method	NTU-60-LT				NTU-120-LT			
	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. 2018)	75.9	86.3	71.7	66.5	66.9	**85.6**	65.6	55.5
ROS (Van Hulse et al. 2007)	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
Remix-DRW (Chou et al. 2020)	78.7	86.6	74.8	72.7	69.3	83.5	67.2	62.6
IB Loss (Park et al. 2021)	76.0	84.0	74.4	66.5	67.8	81.2	67.6	62.4
PaCo (Cui et al. 2021)	76.6	82.0	76.4	69.1	67.9	82.2	66.2	63.8
RIDE (Wang et al. 2021)	76.6	86.7	71.9	68.4	65.2	85.4	66.7	53.8
BS+Max Norm (Alshammari et al. 2022)	77.5	81.2	75.9	74.1	70.3	80.0	68.2	68.8
BCL (Zhu et al. 2022)	77.3	84.4	74.1	71.5	66.9	82.3	64.5	62.8
FSA (Chu et al. 2020)	76.8	85.4	73.3	68.8	66.7	82.4	66.8	59.4
BRL (Liu et al. 2023)	76.9	85.2	73.1	70.0	66.3	84.2	65.0	59.9
GLMC (Du et al. 2023)	78.8	78.6	**81.3**	76.0	71.5	79.5	70.5	69.2
COCL (Miao et al. 2024)	80.7	85.8	78.2	75.4	72.8	84.6	71.1	69.5
LTRL (Zhao et al. 2024a)	80.6	86.5	78.1	75.9	72.8	84.5	71.0	69.5
LDMLR (Han et al. 2024)	80.7	86.6	78.3	75.5	72.9	84.7	70.9	69.4
DSCL (Xuan and Zhang 2024)	80.6	86.7	78.1	75.2	72.7	84.5	71.0	69.3
Shap-Mix (Zhang et al. 2024)	80.8	**86.8**	78.4	75.6	73.0	84.8	71.3	69.7
AEED (Ours)	**81.5**	85.5	79.8	**78.5**	**73.6**	85.1	**72.7**	**71.3**

Table 1: State-of-the-art comparison on long-tailed NTU RGB+D benchmarks (X-Sub, IF=100). We report Top-1 accuracy (%). **Bold** and underline denote the best and second-best results, respectively.

Complexity and Overhead Analysis

► **Time Overhead.** Let ρ denote the probability of invoking CS-Mix (default $\rho = 0.5$). For the mini-batches where CS-Mix is active, AEED performs an additional forward/backward pass to obtain the posterior of the mixed sample. The overall complexity per mini-batch is therefore

$$\text{FLOPs}_{\text{AEED}} = \text{FLOPs}_{\text{baseline}} + \rho \times 2\,\text{FLOPs}_{\text{backbone}}. \quad (7)$$

On NTU-60-LT (100 epochs, 1×RTX 3090), the vanilla CTR-GCN finishes in 8.5 h; AEED requires 10.1 h, an acceptable +18.8% wall-clock increase. Inference speed is unaffected because AEED is activated only during training.

► **Memory Overhead.** AEED maintains four class-wise vectors: niche widths w, current/previous entropies $H_{\text{curr}}, H_{\text{prev}}$, and the smoothed entropy flow $\Delta\bar{H}$. For NTU-120-LT, it's $4 \times 120 \times 4$ bytes $= 1.92$ KB, negligible relative to model parameters; peak GPU memory (11.6 GB) is identical to baselines (11.6 GB).

► **Convergence.** Despite the per-epoch nature of the entropy flow calculation, AEED does not slow down the model's convergence. In our experiments, we observed that the AEED-enhanced model typically reaches the baseline's final performance within a comparable number of epochs. In some cases, the more effective supervisory signals from AEED appear to facilitate a slightly faster convergence in the later stages of training. A comparative convergence curve is provided in **Appendix E** for visualization.

Experiments

Datasets

► **NTU RGB+D 60 & 120** (Shahroudy et al. 2016; Liu et al. 2020a). NTU-60 is a large-scale skeleton-based action recognition dataset containing 56k action clips across 60 classes. We evaluate on its two standard protocols: Cross-Subject (X-Sub), which splits subjects for training and testing, and Cross-View (X-View), which splits by camera views. The larger NTU-120 dataset extends this with 114k clips over 120 classes. For NTU-120, we follow its two protocols: Cross-Subject (X-Sub), with a new subject split, and Cross-Set (X-Set), which splits by setup and replication IDs.

► **Kinetics-Skeleton** (Kay et al. 2017). To evaluate on a large-scale, in-the-wild video dataset, we use the pre-processed Kinetics-Skeleton, containing 2D joint coordinates from 240k training clips across 400 classes.

► **Long-tailed Versions.** To evaluate performance under class imbalance, following (Zhang et al. 2024), we use long-tailed versions denoted as NTU-60-LT, NTU-120-LT, and Kinetics-400-LT for fair evaluation.

Evaluation and Implementation Details

We adopt CTR-GCN (Chen et al. 2021a) as our backbone for all experiments to ensure a fair comparison. Models are trained for 100 epochs using an SGD optimizer with Nesterov momentum (0.9) and a weight decay of 0.0004. The learning rate is initialized to 0.1, undergoes a 5-epoch warm-up, and is decayed by a factor of 10 at epochs 60 and 80. For

Method	Acc. (%)	Method	Acc. (%)
PoseConv3D	46.0	LTRL	48.2
BRL	45.6	DSCL	48.3
GLMC	46.6	COCL	48.9
ShapMix	48.4	LDMLR	<u>49.1</u>
AEED (Ours)			**49.3**

Table 2: Performance comparison (%) on Kinetics-400-LT.

our AEED framework, we use a consistent set of hyperparameters: EMA factor $\beta = 0.9$, eco-adaptation rate $\eta = 0.5$, self-competition coefficient $\gamma = 0.05$, and niche widths w_c are clipped to the range $[0.1, 4.0]$. For evaluation on long-tailed datasets, we report the overall Top-1 accuracy, as well as the accuracy on Many-shot, Medium-shot, and Few-shot subsets. For standard balanced datasets, we report the overall Top-1 accuracy. All experiments are conducted on a server with 1 NVIDIA RTX 3090 GPU.

Comparison with State-of-the-Art Methods

To establish a fair and rigorous benchmark, we conducted extensive experiments to reproduce baselines, from classic LDAM-DRW to recent SOTA models such as Shap-Mix and LTRL. In all reproductions, we strictly ensure a fair comparison by using the same CTR-GCN backbone, the same training schedule, and adhering to the official implementations and hyperparameter settings wherever possible.

► **Analysis on NTU-LT Benchmarks.** Table 1 shows that AEED establishes a new SOTA on both NTU-60-LT and NTU-120-LT. On NTU-60-LT, AEED achieves an overall accuracy of 81.5%. It delivers a boost on the challenging Few-shot set, reaching 78.5% accuracy, a significant +2.5% improvement over the strongest competitor on this metric, GLMC. This trend is amplified on the larger NTU-120-LT dataset, where AEED achieves an overall accuracy of 73.6% and a Few-shot accuracy of 71.3%, a +1.6% improvement over the previous best result from Shap-Mix. This consistent and significant improvement on the tail of the distribution validates our central hypothesis: enabling a model to perceive its own learning dynamics via Entropy Flow allows for a more robust and equitable knowledge distribution.

► **Analysis on Kinetics-400-LT.** As shown in Table 2, AEED's effectiveness extends to the highly challenging Kinetics-400-LT dataset, achieving a Top-1 accuracy of 49.3% and outperforming all competitors. This indicates that the principles of AEED are general and provide a robust solution for large-scale long-tailed action recognition.

► **Performance on Balanced Datasets.** To verify that AEED does not impair performance in standard scenarios, we evaluate it on the original balanced NTU datasets. Table 3 shows that AEED achieves performance on par with or slightly better than SOTA methods. This confirms that AEED's mechanisms are benign in balanced settings, acting as a lightweight and effective regularizer.

Method	NTU-60		NTU-120	
	xsub	xview	xsub	xset
MS-G3D (Liu et al. 2020b)	91.5	96.2	86.9	88.4
EfficientGCN (Song et al. 2021)	91.7	95.7	88.3	89.1
MST-GCN (Chen et al. 2021b)	91.5	96.6	87.5	88.8
CTR-GCN (Chen et al. 2021a)	92.4	96.8	88.9	90.6
InfoGCN (Chi et al. 2022)	93.0	97.1	89.8	91.2
MaskCLR (Abdelfattah et al. 2022)	**93.9**	97.3	87.4	89.5
STC-Net (Lee et al. 2023)	93.0	97.1	89.9	91.3
Stream-GCN (Cheng et al. 2023a)	92.9	96.9	89.7	91.0
MSSTNet (Cheng et al. 2023b)	92.6	97.8	87.4	88.3
JMDA (Xiang and Wang 2024)	93.7	97.2	90.9	91.9
JPFormer (Cui and Hayama 2024)	93.2	96.9	89.4	91.4
Shap-Mix (Zhang et al. 2024)	93.7	97.1	90.4	91.7
LA-GCN (Xu et al. 2025)	93.5	97.2	90.7	91.8
DSTSA-GCN (Cui et al. 2025)	92.8	97.0	89.1	91.0
AEED (Ours)	93.8	**97.3**	**90.9**	**91.9**

Table 3: Performance comparison (%) with SOTA methods on standard balanced NTU RGB+D benchmarks.

Method	Overall	Many	Medium	Few
Baseline (CE Loss)	74.4	**86.4**	69.5	63.8
+ E-Adapt (Ours)	77.9	85.1	75.3	71.2
+ CS-Mix (Ours)	78.6	85.9	76.1	72.0
AEED (Full)	**81.5**	85.5	**79.8**	**78.5**

Table 4: Ablation study of our components on NTU-60-LT.

Ablation Studies and Analysis

► **Effectiveness of Core Components.** We first analyze the individual and combined effects of our two main contributions: E-Adapt and CS-Mix. As shown in the quantitative results in Table 4 and Figure 3 (bottom row), adding E-Adapt or CS-Mix alone provides a significant boost, particularly on Few-shot classes. The full AEED model achieves the best results, demonstrating a powerful synergy. This quantitative improvement is explained by the qualitative results in Figure 3 (top row), where the t-SNE visualizations show that the full framework learns a significantly more compact and separable feature space for the challenging tail classes.

► **Analysis of Driving Signal and Interaction Strategy.** We further investigate the design choices within our two components. In Table 5, we compare different signals for driving the adaptation and different strategies for interaction. The results show that using Entropy Flow ($\Delta \bar{H}_c$) is significantly better than using Absolute Entropy (H_c), confirming our core hypothesis. Furthermore, within the E-Adapt framework, our CS-Mix outperforms standard Mixup and CutMix, demonstrating the superiority of a model-aware semantic mixing approach over purely physical interpolation.

► **Hyperparameter Sensitivity.** We study the sensitivity of hyperparameters, η and γ, on NTU-60-LT. In Table 6, AEED demonstrates robust performance across a reasonable range of values. The overall accuracy varies by less than 1.5%, indicating that our method is not overly sensitive to

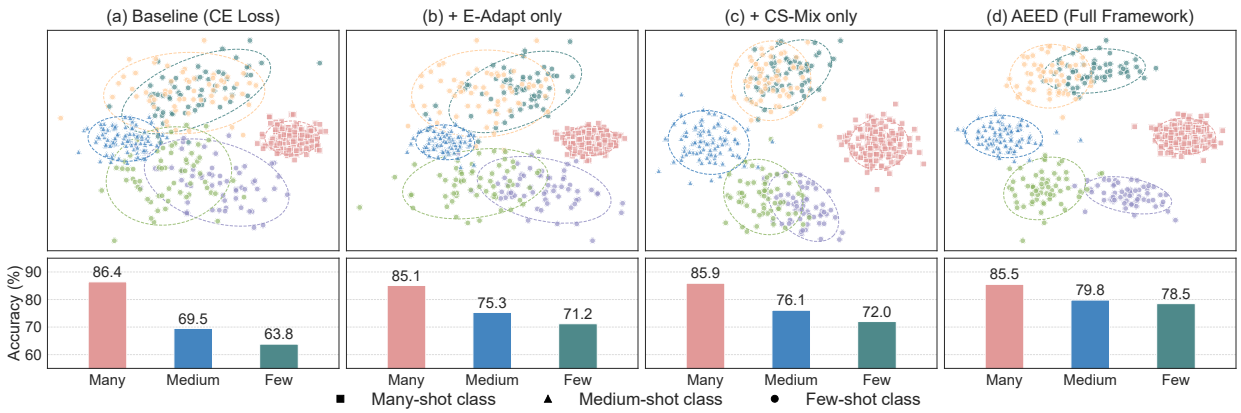

Figure 3: Qualitative (top row) and quantitative (bottom row) analysis of AEED components on NTU-60-LT. The t-SNE visualizations (top) show the feature space evolution, where shape indicates data amount (■ Many, ▲ Medium, ● Few). The full AEED framework (d) learns a significantly more compact and separable feature space for Few-shot classes compared to the baseline (a). This directly translates to the superior Top-1 accuracy shown in the corresponding bar charts (bottom).

Driving Signal	Interaction Strategy	Overall Acc. (%)
H_c	Mixup (Zhang et al. 2018)	76.1
$\Delta\bar{H}_c$	Mixup (Zhang et al. 2018)	77.9
$\Delta\bar{H}_c$	CutMix (Yun et al. 2019)	78.2
$\Delta\bar{H}_c$	ShapMix (Zhang et al. 2024)	81.0
$\Delta\bar{H}_c$	**CS-Mix (Ours)**	**81.5**

Table 5: Analysis of driving signals and interaction strategies on NTU-60-LT. All variants use E-Adapt.

	Self-Competition (γ)		
Eco-Adapt. Rate (η)	0.01	0.05	0.10
0.1	80.4	80.7	80.5
0.5	81.2	**81.5**	81.3
1.0	80.9	81.1	80.8

Table 6: Sensitivity analysis of hyperparameters η and γ on NTU-60-LT. Results are Overall Accuracy (%).

hyperparameter tuning and is practical to deploy. The best performance is achieved at our default setting of $\eta = 0.5$ and $\gamma = 0.05$. A more detailed analysis, including the impact of other hyperparameters like the EMA factor β and the niche width clipping range, is provided in **Appendix F**.

► **Generality Analysis.** To demonstrate the general applicability of AEED beyond skeleton-based action recognition, we evaluated AEED on a diverse set of standard long-tailed benchmarks for vision and text classification. We integrated AEED's adaptive controller directly onto strong, publicly established baselines without any task-specific modifications. The baseline results reported are from recent, widely-used training recipes, ensuring a fair and challenging comparison. For vision tasks, we used ResNet backbones on CIFAR-10/100-LT (Cui et al. 2019) and ImageNet-LT (Liu et al. 2019). For the multi-label text task, we used BERT-base on Amazon-531 (Galke and Scherp 2022).

Dataset	Backbone	Baseline	+ AEED
CIFAR-10-LT (IF=100)	ResNet-32	71.3	**74.1**
CIFAR-100-LT (IF=100)	ResNet-32	38.8	**42.5**
ImageNet-LT	ResNet-50	41.6	**44.2**
Amazon-531 (Text)	BERT-base	42.1	**44.9**

Table 7: Overall performance on cross-modal long-tailed benchmarks. We report Top-1 Accuracy (%) for vision tasks and Sample-F1 (%) for the text task.

Table 7 summarizes the overall performance. The results clearly indicate that AEED provides consistent and significant improvements across all benchmarks. This confirms that the core principle of harnessing learning dynamics via Entropy Flow for self-regulation is a general one that effectively transfers to both image and text domains.

► **Explainability Analysis of CS-Mix.** To ensure that our CS-Mix module is not a black box, we conducted an in-depth explainability analysis, with full details provided in **Appendix G**. The analysis includes: (i) a quantitative study showing a strong and statistically significant correlation ($r = 0.74$, $p < 0.001$) between our KL-divergence based semantic ratio and the geometry of the backbone's feature space; (ii) qualitative visualizations of both successful and challenging mixing scenarios, corroborated by t-SNE projections; and (iii) an ablation study confirming that the semantic guidance mechanism alone brings a substantial performance gain over standard mixing. These results collectively validate the design and robustness of CS-Mix.

Conclusion

We introduced AEED, a framework that reframes long-tailed learning as a closed-loop system. By using Entropy Flow as a feedback signal for resource allocation and CS-Mix for semantically-valid knowledge transfer, AEED achieves SOTA performance on challenging benchmarks.

Acknowledgements

The authors gratefully acknowledge the support from the National Natural Science Foundation of China (NSFC) under Grant Nos. 62402472, and 12227901. This work was also supported by the Natural Science Foundation of Jiangsu Province (No. BK20240461), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No. YSBR-005), and the Academic Leaders Cultivation Program at USTC. The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Sciences.

References

Abdelfattah, M.; Hassan; Mariam; Alahi; and Alexandre. 2022. MaskCLR: Attention-Guided Contrastive Learning for Robust Action Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18678–18687.

Alshammari, S.; Wang, Y.; Ramanan, D.; and Kong, S. 2022. Long-Tailed Recognition via Weight Balancing. In *CVPR*.

Cao, K.; Wei, C.-W.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; and Hu, W. 2021a. Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13359–13368.

Chen, Z.; Li, S.; Yang, B.; Li, Q.; and Liu, H. 2021b. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1113–1122.

Cheng, K.; Peng, K.; Wang, Z.; Xu, A.; and Wang, Y. 2023a. Streaming GCN: A Motion-Based Streaming Graph Convolutional Network for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Cheng, Q.; Cheng, J.; Ren, Z.; Zhang, Q.; and Liu, J. 2023b. Multi-scale spatial–temporal convolutional neural network for skeleton-based action recognition. *Pattern Analysis and Applications*, 26: 1303–1315.

Chi, H.-G.; Ha, M. H.; Chi, S.; Lee, S. W.; Huang, Q.; and Ramani, K. 2022. InfoGCN: Representation Learning for Human Skeleton-based Action Recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20154–20164.

Chou, H.-P.; Chang, S.-C.; Pan, J.-Y.; Wei, W.; and Juan, D.-C. 2020. Remix: Rebalanced Mixup. In *European Conference on Computer Vision Workshops*, 95–110.

Chu, P.; Bian, X.; Liu, S.; and Ling, H. 2020. Feature Space Augmentation for Long-Tailed Data. In *European Conference on Computer Vision*, 694–710.

Cui, H.; and Hayama, T. 2024. Joint-Partition Group Attention for skeleton-based action recognition. *Signal Processing*, 224: 109592.

Cui, H.; Huang, R.; Zhang, R.; and Hayama, T. 2025. Dstsa-gcn: Advancing skeleton-based gesture recognition with semantic-aware spatio-temporal topology modeling. *Neurocomputing*, 130066.

Cui, J.; Zhong, Z.; Liu, S.; Yu, B.; and Jia, J. 2021. Parametric Contrastive Learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 695–704.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *CVPR*.

Du, F.; Yang, P.; Jia, Q.; Nan, F.; Chen, X.; and Yang, Y. 2023. Global and Local Mixture Consistency Cumulative Learning for Long-tailed Visual Recognitions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Galke, L.; and Scherp, A. 2022. On the Challenges of Open-World Recognition under Long-Tailed Class Distribution. In *Advances in Information Retrieval. 44th European Conference on IR Research (ECIR)*, 26–42. Springer International Publishing.

Guo, H.; Mao, Y.; and Zhang, R. 2019. MixUp as locally linear out-of-manifold regularization. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*. AAAI Press. ISBN 978-1-57735-809-1.

Han, P.; Ye, C.; Zhou, J.; Zhang, J.; Hong, J.; and Li, X. 2024. Latent-based Diffusion Model for Long-tailed Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2639–2648.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020. Decoupling representation and classifier for long-tailed recognition. In *Eighth International Conference on Learning Representations (ICLR)*.

Kay, W.; Carreira, J.; Simonyan, K.; Zisserman, A.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Lee, J.; Lee, M.; Cho, S.; Woo, S.; Jang, S.; and Lee, S. 2023. Leveraging Spatio-Temporal Dependency for Skeleton-Based Action Recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10221–10230.

Li, T.; Cao, P.; Yuan, Y.; Fan, L.; Yang, Y.; Feris, R.; Indyk, P.; and Katabi, D. 2021. Targeted Supervised Contrastive Learning for Long-Tailed Recognition. *arXiv preprint arXiv:2111.13998*.

Lin, T.-Y.; Goyal, P.; Girshick, R. B.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007.

Liu, H.; Wang, Y.; Ren, M.; Hu, J.; Luo, Z.; Hou, G.; and Sun, Z. 2023. Balanced Representation Learning for Long-tailed Skeleton-based Action Recognition. *arXiv preprint arXiv:2308.14024*.

Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.-Y.; and Kot, A. C. 2020a. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42: 2684–2701.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-Scale Long-Tailed Recognition in an Open World. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2528–2537.

Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; and Ouyang, W. 2020b. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 140–149.

Magurran, A. E. 2021. Measuring biological diversity. *Current Biology*, 31: R1174–R1177.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2021. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*.

Miao, W.; Pang, G.; Bai, X.; Li, T.; and Zheng, J. 2024. Out-of-distribution detection in long-tailed recognition with calibrated outlier class learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4216–4224.

Park, S.; Lim, J.; Jeon, Y.; and Choi, J. Y. 2021. Influence-Balanced Loss for Imbalanced Visual Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 735–744.

Ren, J.; Yu, C.; Li, S.; Yuan, Y.; Chen, S.; and Ma, X. 2020. Balanced-MetaSoftmax for Long-Tailed Visual Recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1010–1019.

Song, Y.-F.; Zhang, Z.; Shan, C.; and Wang, L. 2021. EfficientGCN: Constructing Stronger and Faster Baselines for Skeleton-based Action Recognition. *arXiv:2106.15125*.

Uddin, S.; Lee, W.-G.; Chuluunbaatar, E.; Kim, M.; Kim, Y.-Y.; and Kim, G.-B. 2021. Saliencymix: A saliency-guided data augmentation strategy for better generalization. In *International Conference on Learning Representations (ICLR) Workshop on Computer Vision for Agriculture*.

Van Horn, G.; and Perona, P. 2017. The Devil is in the Tails: Fine-grained Classification in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5039–5048.

Van Hulse, J.; Khoshgoftaar; M, T.; Napolitano; and Amri. 2007. A systematic study of the class imbalance problem in biomedical data sets. *Expert Systems with Applications*, 32: 1184–1203.

Wang, B.; Wang, P.; Xu, W.; Wang, X.; Zhang, Y.; Wang, K.; and Wang, Y. 2024a. Kill two birds with one stone: Rethinking data augmentation for deep long-tailed learning. In *The Twelfth International Conference on Learning Representations*.

Wang, P.; Zhao, Z.; Wen, H.; Wang, F.; Wang, B.; Zhang, Q.; and Wang, Y. 2024b. Llm-autoda: Large language model-driven automatic data augmentation for long-tailed problems. *Advances in Neural Information Processing Systems*, 37: 64915–64941.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. X. 2021. Long-tailed Recognition by Routing Diverse Distribution-Aware Experts. *ArXiv*, abs/2010.01809.

Xiang, L.; and Wang, Z. 2024. Joint Mixing Data Augmentation for Skeleton-based Action Recognition. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Xu, H.; Gao, Y.; Hui, Z.; Li, J.; and Gao, X. 2025. Language Knowledge-Assisted Representation Learning for Skeleton-Based Action Recognition. *IEEE Transactions on Multimedia*, 1–16.

Xuan, S.; and Zhang, S. 2024. Decoupled Contrastive Learning for Long-Tailed Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6396–6403.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhang; Jiahang; Lin, L.; and Liu, J. 2024. Shap-Mix: Shapley Value Guided Mixing for Long-Tailed Skeleton Based Action Recognition. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations (ICLR)*.

Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; and Feng, J. 2023. Deep Long-Tailed Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45: 10795–10816.

Zhao, Q.; Dai, Y.; Lin, S.; Hu, W.; Zhang, F.; and Liu, J. 2024a. LTRL: Boosting Long-tail Recognition via Reflective Learning. *arXiv preprint arXiv:2407.12568*.

Zhao, Z.; Wang, P.; Wen, H.; Xu, W.; Lai, S.; Zhang, Q.; and Wang, Y. 2024b. Two fists, one heart: Multi-objective optimization based strategy fusion for long-tailed learning. In *Forty-first International Conference on Machine Learning*.

Zhao, Z.; Wen, H.; Wang, P.; Wang, Z.; Zhang, Q.; Wang, Y.; et al. 2025. Balancing Model Efficiency and Performance: Adaptive Pruner for Long-tailed Data. In *Forty-second International Conference on Machine Learning*.

Zhao, Z.; Wen, H.; Wang, Z.; Wang, P.; Wang, F.; Lai, S.; Zhang, Q.; and Wang, Y. 2024c. Breaking Long-Tailed Learning Bottlenecks: A Controllable Paradigm with Hypernetwork-Generated Diverse Experts. *Advances in Neural Information Processing Systems*, 37: 7493–7520.

Zhu, J.; Wang, Z.; Chen, J.; Chen, Y.-P. P.; and Jiang, Y.-G. 2022. Balanced Contrastive Learning for Long-Tailed Visual Recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6898–6907.