Compositional Attribute Imbalance in Vision Datasets

Yanbiao Ma[1,2,3*], Jiayi Chen[4*], Wei Dai[4], Dong Zhao[4], Zeyu Zhang[6], Yuting Yang[4], Bowei Liu,[5] Jiaxuan Zhao[4], Andi Zhang[7†]

[1]Gaoling School of Artificial Intelligence Renmin University of China Beijing, China
[2]Beijing Key Laboratory of Research on Large Models and Intelligent Governance
[3]Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE
[4]Xidian University, [5]Tsinghua University, [6]The Australian National University, [7]The University of Manchester

Abstract

Visual attribute imbalance is a common yet underexplored issue in image classification, significantly impacting model performance and generalization. In this work, we first define the first-level and second-level attributes of images and then introduce a CLIP-based framework to construct a visual attribute dictionary, enabling automatic evaluation of image attributes. By systematically analyzing both single-attribute imbalance and compositional attribute imbalance, we reveal how the rarity of attributes affects model performance. To tackle these challenges, we propose adjusting the sampling probability of samples based on the rarity of their compositional attributes. This strategy is further integrated with various data augmentation techniques (such as CutMix, Fmix, and SaliencyMix) to enhance the model's ability to represent rare attributes. Extensive experiments on benchmark datasets demonstrate that our method effectively mitigates attribute imbalance, thereby improving the robustness and fairness of deep neural networks. Our research highlights the importance of modeling visual attribute distributions and provides a scalable solution for long-tail image classification tasks.

Introduction

Addressing data imbalance in computer vision tasks remains a core challenge for improving model performance (Zhang et al. 2021b; Ma et al. 2025). Imbalances in the number of training samples across classes often lead to biases during the learning process, making it difficult for deep learning models to accurately recognize underrepresented classes. To tackle this issue, researchers have proposed various approaches, such as class-aware sampling strategies, loss reweighting, and balanced data augmentation techniques (Tan et al. 2020; Sinha, Ohashi, and Nakamura 2020; Ren et al. 2020; Ma et al. 2024b,a; Yin et al. 2019; Liu et al. 2020; Huang et al. 2016; Dong, Gong, and Zhu 2017; Kang et al. 2020). However, these methods primarily focus on inter-class imbalances, assuming that achieving balance at the class level suffices to ensure fairness and efficacy in learning. This assumption, however, overlooks intra-class attribute imbalances, particularly the problem of compositional attribute imbalance.

Attribute imbalance refers to the uneven distribution of image attributes (e.g., color, texture, and shape) within a single class. This imbalance can bias the learned representations of a model. While limited studies (Tang et al. 2022; Liu et al. 2021b; Li et al. 2024) have qualitatively discussed the challenges posed by attribute imbalance, no prior research has systematically quantified or analyzed its prevalence, severity, and impact on model performance. To address this gap, our study aims to answer three core questions:

(1) How prevalent is attribute imbalance in commonly used vision datasets?

(2) What is the impact of attribute imbalance on model performance?

(3) How can attribute imbalance be effectively mitigated?

To automatically assess the degree of attribute imbalance in image datasets, two key challenges must be addressed: how to define attributes and how to determine the attributes corresponding to each image. First, based on previous studies (Zhong et al. 2021; Zhang et al. 2024), we define 20 primary attributes (e.g., color) and their corresponding 300+ secondary attributes (e.g., black, white). Second, leveraging the CLIP (Radford et al. 2021), we construct a visual attribute dictionary that aligns low-level visual attributes of images with specific textual descriptions, enabling automated attribute annotation for each image. Using this dictionary, we assign the most suitable secondary attribute from each primary attribute category to each image, resulting in a total of 20 secondary attributes per image. We then compute the frequency of all secondary attributes within each class—the lower the frequency, the higher the scarcity. Based on this, we propose the concept of Compositional Attribute Scarcity (CAS) to comprehensively evaluate the overall attribute scarcity of an individual image. Specifically, for each image, we calculate the scarcity of its contained secondary attributes and sum them to obtain its CAS score.

Through experiments on 12 commonly used vision datasets, we reveal that intra-class attribute imbalance and compositional attribute imbalance are pervasive. Furthermore, we systematically analyze how these imbalances affect model performance. Our experimental results demonstrate a consistent pattern: images with higher CAS tend to have lower recognition accuracy. This finding underscores the potential for improving model generalization by addressing attribute

*These authors contributed equally.
†Corresponding author

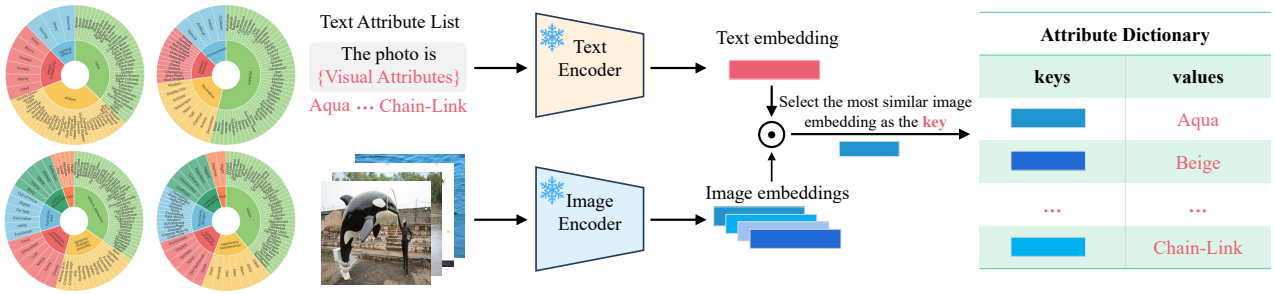

Figure 1: The left side shows all primary attributes we defined and their corresponding secondary attributes. The right side illustrates the process of constructing the visual attribute dictionary based on CLIP.

imbalance, beyond the improvements achievable by resolving inter-class imbalance alone.

To mitigate attribute imbalance, we propose a novel sampling adjustment strategy for data augmentation. Specifically, we adjust the sampling probability of each image based on its compositional attribute scarcity, with rarer images being sampled more frequently. This adjustment increases the representation of rare attributes in augmented datasets, enabling data augmentation methods to generate more samples emphasizing rare attributes (e.g., white dogs). As a result, the proposed method facilitates better learning of diverse intra-class attributes. Notably, our method introduces no additional computational overhead and requires only a simple modification of the sampling strategy, making it easily integrated into existing frameworks. The key contributions of this work are as follows:

(1) A Visual Attribute Framework: We define a comprehensive visual attribute framework encompassing 20 primary attributes and over 300 secondary attributes. We also propose a CLIP-based visual attribute dictionary to automate the evaluation of attribute imbalance, revealing its widespread prevalence in general vision datasets.

(2) Impact Analysis of Attribute Imbalance: We reveal the significant impact of attribute imbalance on model performance. Specifically, images with higher CAS exhibit lower recognition accuracy, highlighting the necessity and importance of addressing intra-class attribute imbalance.

(3) We propose a sampling adjustment method based on CAS. This method, requiring only a custom sampler, integrates seamlessly with existing data augmentation frameworks. Experiments on 12 benchmark datasets demonstrate the effectiveness and generalizability of the proposed approach.

Attribute Imbalance

In this section, we first systematically define the visual attributes of images. Then, we propose using CLIP to construct a visual attribute dictionary, enabling automatic evaluation of image attributes. Finally, we reveal the prevalence of attribute imbalance and compositional attribute imbalance across 12 commonly used visual datasets and analyze their impact on model performance.

Definition of Visual Attributes

Visual attributes refer to the fundamental characteristics that constitute an image, such as color, texture, and shape. These attributes not only define the visual appearance of an image but also play a critical role in the representation learning process of deep learning models. In this study, we define visual attributes based on a comprehensive analysis of prior research (Zhao et al. 2019; Pham et al. 2021) and practical insights. These attributes are categorized into 20 primary attributes (e.g., color, material, shape, size) and over 300 secondary attributes (e.g., "black" and "white" under color). Figure 1 illustrates all primary and secondary attributes. This hierarchical design ensures both the comprehensiveness and granularity of attribute definitions.

Constructing a Visual Attribute Dictionary

To enable the automated evaluation of attribute distributions in image datasets, we leverage the CLIP model to construct a visual attribute dictionary on ImageNet-21k. As shown in Figure 1, we first organize all secondary attributes into a textual attribute list, such as "The photo is Brown," and generate corresponding text embeddings. Next, we calculate the similarity between each text embedding and the image embeddings, matching the most similar image embedding to the respective text attribute. The matched image embeddings serve as the keys in the visual attribute dictionary, while the corresponding text attributes serve as the values. To query the visual attributes of a given image, its embedding is extracted and compared with the dictionary keys using cosine similarity. The value corresponding to the key with the highest similarity score is then returned as the predicted attribute for the image.While CLIP enables both image-text and image-image matching, we deliberately adopt the image-to-image retrieval strategy in our dictionary construction. The main reason is that many secondary attributes, such as colors (e.g., "beige", "brown") or textures (e.g., "striped", "metallic"), are semantically similar in textual form yet visually distinct. Matching directly via textual prompts (e.g., "The object is brown") risks conflating attributes that are linguistically close but visually divergent. Moreover, prompt-based approaches often rely on handcrafted sentence templates and suffer from prompt sensitivity and inconsistency across diverse attributes.

By contrast, constructing a visual dictionary using representative image anchors allows us to build a more stable and interpretable embedding space. These anchors are retrieved from ImageNet-21K and provide consistent visual prototypes for each attribute. This design not only avoids the ambiguity of natural language prompts but also enhances

generalizability across datasets.

Single-Attribute Imbalance

At the single-attribute level, the imbalance manifests as certain attributes (e.g., "black") dominating a large proportion of the dataset, while other attributes (e.g., "purple") are represented by only a few samples. We conducted a systematic analysis of 12 commonly used visual datasets, including ImageNet and CIFAR-100. Using the visual attribute dictionary, we calculated the distribution of different attributes in each dataset and quantified the degree of attribute imbalance. As shown in Figure 2, the distribution of secondary attributes under each primary attribute typically exhibits a long-tailed pattern, with a large number of low-frequency attributes having significantly fewer samples than high-frequency attributes.

To investigate the impact of attribute frequency on model performance, we first trained standard ResNet-18 and ResNet-50 models on each dataset. Within each category, for each primary attribute, we divided the samples into subsets based on their associated secondary attributes and evaluated the recognition accuracy of both models on each subset. The experimental results, shown in Figure 2, reveal that samples with higher-frequency attributes generally achieve higher and more stable recognition accuracy. **Conversely, samples with low-frequency attributes do not consistently exhibit the expected low recognition accuracy**. Merely analyzing single-attribute imbalance is insufficient to explain this phenomenon. We hypothesize that the rarity of one type of secondary attribute in an image does not necessarily imply the rarity of other secondary attributes (e.g., a rare color may coexist with a common shape).

Compositional Attribute Imbalance

In the analysis of single-attribute imbalance, we observed that samples with high-frequency attributes are more likely to be correctly identified by the model. However, merely relying on single-attribute statistics cannot fully explain the model's performance on low-frequency attribute samples. Considering that an image often contains multiple visual attributes, we further introduce the concept of compositional attributes to explore the impact of multi-attribute scarcity on model performance.

Compositional attributes refer to the specific combination of multiple (20 in this study) primary attributes in an image, such as {Blue, Metallic, Round, ... } or {Red, Wooden, Square, ... }. These combinations not only describe the visual characteristics of an image but also capture the interrelationships between attributes. However, the free combinations of attributes are not uniformly distributed in datasets, and many compositional attributes are extremely scarce in the training data. Such scarcity may lead to significantly degraded model performance on images with these rare compositional attributes. We define **Compositional Attribute Scarcity (CAS)** as follows:

(1) For each primary attribute, calculate the frequency of its secondary attributes and rank them in descending order of frequency.

(2) The scarcity of each secondary attribute is indicated by the ranked position, the lower the rank, the rarer it is.

(3) The CAS of an image is calculated as the sum of the scarcity ranks of its 20 secondary attributes.

Figure 3 illustrates the process of calculating the CAS of an image. We further investigate the impact of compositional attribute scarcity on model performance across 12 datasets using ResNet-18 and ResNet-50. Samples were divided into subsets based on their CAS values, and classification accuracy was evaluated for each subset. The experimental results, shown in Figure 4, reveal the following:

(1) Compositional attribute imbalance is pervasive, with many attribute combinations represented by only a few samples in the entire dataset.

(2) As Compositional Attribute Scarcity increases, model performance gradually deteriorates.

It is evident that samples with high compositional attribute scarcity are often under-learned. To address this issue, we propose a simple yet effective solution to mitigate the impact of compositional attribute imbalance.

Leveraging Compositional Attribute Scarcity to Guide Data Augmentation

To mitigate the negative impact of compositional attribute imbalance on model performance, we propose a simple yet effective solution. The core idea is to adjust the sampling probability during data augmentation based on a sample's Compositional Attribute Scarcity (CAS), thereby increasing the exposure of rare attribute combinations in the augmented dataset. This directly addresses the root cause of the performance degradation identified: if high-CAS samples are under-learned due to their scarcity, the most straightforward intervention is to make them more likely to be selected as the "source" image in augmentation techniques like CutMix, thus generating new composite samples that inherit and emphasize these rare attributes . To amplify the differentiation among samples of varying scarcity, we apply a power transformation to the CAS scores. In practice, our method can be implemented by simply customizing the data sampler.

Sampling Strategy Based on CAS

Assume the total number of samples is M, and the compositional attribute scarcity of sample i is r_i. To enhance the differentiation of scarcity, we apply a power transformation to r_i: $r'_i = r_i^b$, where b is the power parameter controlling the degree of nonlinear amplification. When $b > 1$, the differentiation of high-scarcity samples is significantly increased. Our empirical studies recommend setting b to 1.2 (see Section). Based on the transformed scarcity r'_i, the sampling probability for each sample is defined as:

$$p_i = \frac{r'_i}{\sum_{k=1}^{M} r'_k}.$$

Our proposed strategy is deliberately designed for simplicity and practicality. It introduces no additional computational overhead, requires no modification to the model architecture,

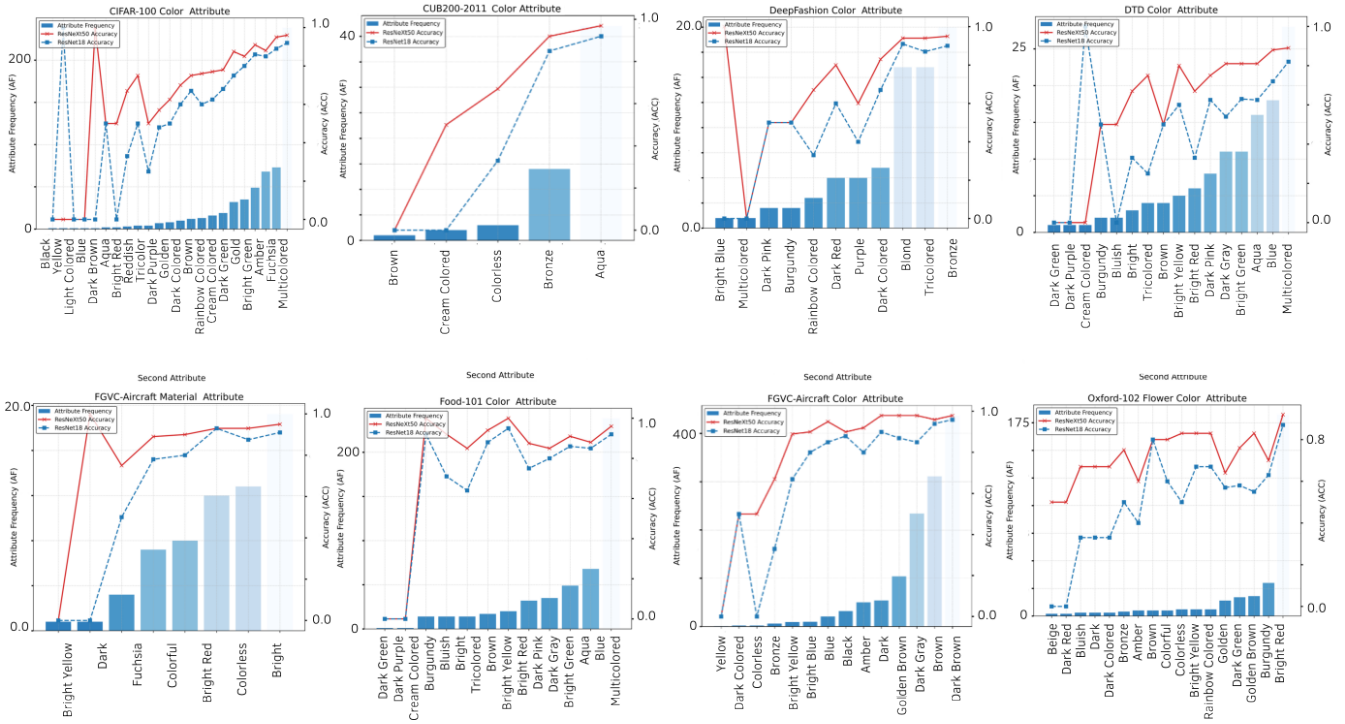

Figure 2: The distribution of secondary attributes under color categories across 12 visual benchmark datasets, along with the performance of ResNet-18 and ResNeXt-50 on each secondary attribute. For further details, please refer to the appendix.

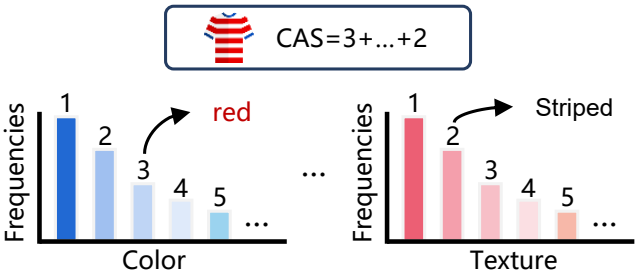

Figure 3: Illustration of the computation process for image omposi-tional attribute scarcity (CAS).

and can be seamlessly integrated as a plug-and-play module with a wide range of existing data augmentation frameworks (e.g., CutMix, FMix, SaliencyMix). This makes our method highly accessible and easy to deploy in real-world scenarios.

Seamless Integration with Data Augmentation

During training, we first compute the compositional attribute scarcity and sampling probability for each sample. These probabilities are then used to customize the sampler. Subsequently, data augmentation techniques (e.g., CutMix, FMix, SaliencyMix) are applied to preferentially generate more samples with rare attributes. Algorithm 1 provides the implementation details using CutMix as an example.

Furthermore, Table 1 compares the level of compositional attribute imbalance in the dataset before and after applying our method. The results show that using only standard data augmentation strategies yields little improvement in reducing attribute imbalance within the dataset.

	ImageNet-1K	+CutMix	+Our Method
Mean	124.6	120.4	129.8
Standard Deviation	37.6	36.8	29.3

Table 1: Comparison of sample CAS statistics on ImageNet-1K. Our method significantly reduces the standard deviation of CAS, indicating a more balanced and less dispersed distribution of compositional attributes.

Empirical Study

Datasets

To comprehensively evaluate the performance of the proposed method, twelve diverse image classification datasets were selected. These datasets encompass tasks ranging from large-scale image classification to fine-grained classification, which effectively validate the model's performance across various scenarios. The `ImageNet-1K` (Deng et al. 2009) dataset contains 1.2 million training images and $50,000$ validation images across $1,000$ categories. The `CIFAR-100` (Krizhevsky, Hinton et al. 2009) dataset includes $50,000$ training images and $10,000$ test images spanning 100 categories. The `Oxford-IIIT Pet` (Parkhi et al. 2012) dataset comprises 37 pet categories with $7,349$ images. The `Stanford Dogs` (Khosla et al. 2011) dataset contains 120 dog breeds with a total of $20,580$ images. The `DTD (Describable Textures Dataset)` (Cimpoi

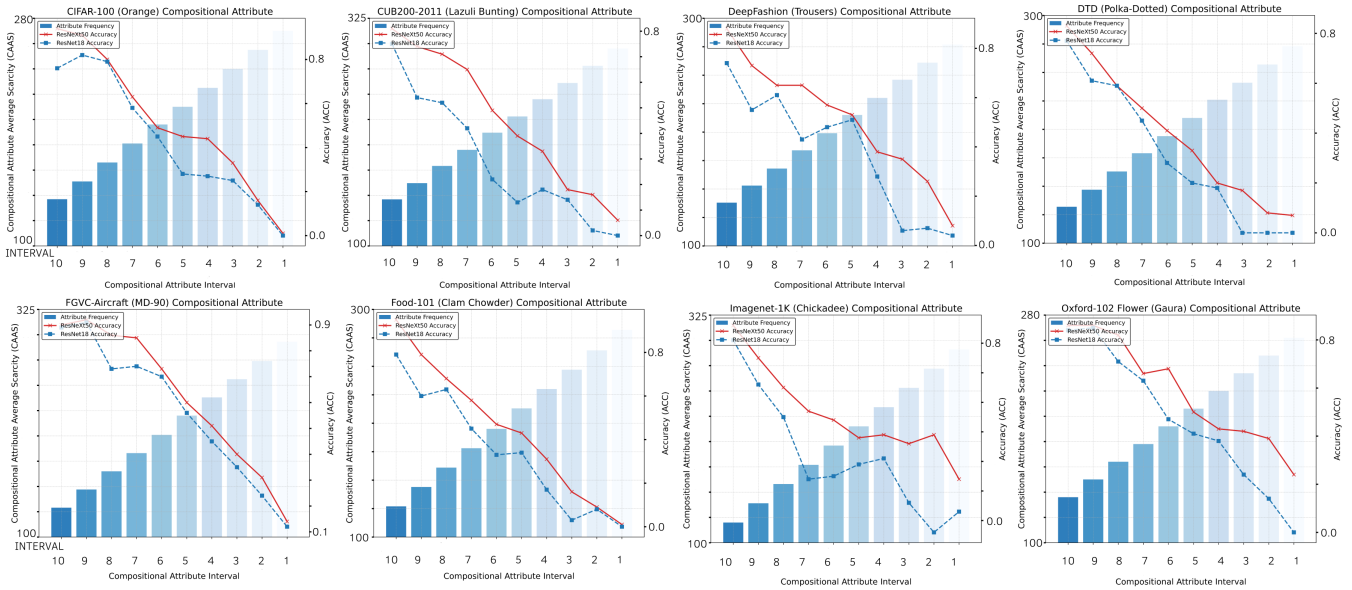

Figure 4: The long-tailed distribution of sample composite attribute sparsity across certain categories in 12 visual benchmark datasets, along with the performance of ResNet-18 and ResNeXt-50 across different compositional attribute scarcity (CAS) intervals. The horizontal axis represents 10 evenly divided intervals based on different CAS values, increasing from left to right. The left vertical axis indicates the average compositional attribute scarcity of all samples within each interval.

et al. 2014) includes 47 texture categories with $5,640$ images. The `Oxford-102 Flower` (Nilsback and Zisserman 2008) dataset contains 102 flower categories with $8,189$ images. The `Food-101` (Bossard, Guillaumin, and Van Gool 2014) dataset covers 101 food categories with a total of $101,000$ images. The `Stanford Cars` (Krause et al. 2013) dataset includes 196 car categories with $16,185$ images. The `FGVC-Aircraft` (Maji et al. 2013) dataset contains 100 aircraft categories with $10,000$ images. The `SUN397` (Xiao et al. 2010) dataset features 397 scene categories with $108,754$ images. The `DeepFashion` (Liu et al. 2016) dataset consists of 50 clothing categories with $50,000$ images. Finally, the `CUB200-2011` (Wah et al. 2011) dataset includes 200 bird species categories with $11,788$ images. Through comprehensive testing across these datasets, we can thoroughly assess the model's performance in a variety of tasks and environments.

Implementation Details

In this experiment, we employed `ResNet-18` and `ResNeXt-50` as the baseline models and configured the hyperparameter α for different data augmentation methods (Qin et al. 2024). Specifically, α was set to 0.2 for CutMix, FMix, and SaliencyMix, and the mixed hyperparameters were generated by sampling from a Beta(α, α) distribution at each training iteration. The batch size for all models was set to 64, and the initial learning rate was 0.1, with `cosine annealing` used as the learning rate scheduling strategy (Qin et al. 2024; Islam et al. 2024).

Evaluation Metrics

To comprehensively evaluate the performance of the proposed method, in addition to testing overall classification accuracy (Qin et al. 2024), we also introduce a sample partitioning strategy based on compositional attribute scarcity (CAS) to further investigate the model's performance at different CAS levels. We first calculate the CAS for all samples and rank them. Samples with higher CAS correspond to rarer visual features, thus posing greater challenges to the model's discriminative ability. Based on the CAS value of each sample, we divide the test set into three subsets:

- **High subset**: Contains the top 40% of samples with the highest CAS values, representing the most challenging samples due to their rare visual features.
- **Middle subset**: Includes the next 30% of samples, with moderate CAS values, representing samples with less rare visual features compared to the first subset.
- **Low subset**: Consists of the remaining 30% of samples with the lowest CAS values, representing the easiest samples with more common visual features.

For each subset, we calculate and test the classification accuracy of the model before and after the improvements. This sparsity-based subset division allows us to more precisely analyze the model's performance under varying information conditions, particularly in terms of classification ability between low sparsity (information-rich) and high sparsity (information-scarce) samples, as well as the differences in model enhancement. Through this method, we can effectively assess the model's robustness and generalization when faced with samples of varying information density.

CIFAR-100								
Method	ResNet18				ResNeXt50			
	low	middle	high	all	low	middle	high	all
CutMix	68.31	49.65	43.05	54.57	71.54	51.73	44.62	58.07
CutMix+weight	68.96	50.32	43.94	55.28	71.85	52.36	45.6	58.52
Δ	+0.65	+0.67	+0.89	+0.71	+0.31	+0.63	+0.98	+0.45
FMix	64.15	52.36	33.58	51.30	69.44	43.25	38.45	52.98
FMix+weight	66.73	56.63	39.38	52.85	73.69	45.59	42.30	55.74
Δ	+2.58	+4.27	+5.80	+1.55	+4.25	+2.34	+3.85	+2.76
SaliencyMix	79.24	64.28	41.73	57.25	72.65	70.28	43.26	59.65
SaliencyMix+weight	79.56	65.86	44.15	57.88	73.41	71.21	45.02	59.94
Δ	+0.32	+1.58	+2.42	+0.63	+0.76	+0.93	+1.76	+0.29

(a) CIFAR-100

Imagenet-1k								
Method	ResNet18				ResNeXt50			
	low	middle	high	all	low	middle	high	all
CutMix	68.24	47.26	21.68	48.36	76.30	71.14	30.25	49.78
CutMix+weight	68.41	48.22	23.26	49.04	77.28	72.21	33.79	50.96
Δ	+0.17	+0.96	+1.58	+0.68	+0.98	+1.07	+3.54	+1.18
FMix	58.72	49.36	38.25	47.26	62.95	56.31	41.36	49.78
FMix+weight	59.07	50.83	40.00	48.31	63.59	58.18	43.97	51.36
Δ	+0.35	+1.47	+1.75	+1.05	+0.64	+1.87	+2.61	+1.58
SaliencyMix	61.54	46.57	40.38	47.10	58.32	49.62	43.28	47.35
SaliencyMix+weight	63.22	48.55	42.36	49.78	61.00	52.68	48.17	50.42
Δ	+1.68	+1.98	+2.92	+2.68	+2.68	+3.06	+4.89	+3.07

(b) Imagenet-1k

DTD								
Method	ResNet18				ResNeXt50			
	low	middle	high	all	low	middle	high	all
CutMix	95.36	91.74	85.55	91.63	95.22	94.36	86.32	99.61
CutMix+weight	95.79	92.33	88.43	93.01	95.98	95.05	87.80	99.98
Δ	+0.43	+0.59	+2.88	+1.38	+0.17	+0.69	+1.51	+0.37
FMix	93.66	86.20	79.65	88.61	94.62	86.31	70.89	99.56
FMix+weight	95.58	87.89	81.02	88.76	95.38	87.99	72.83	99.65
Δ	+1.92	+1.69	+1.37	+0.15	+0.76	+1.68	+1.94	0.09
SaliencyMix	92.99	81.45	69.02	89.33	91.36	79.26	65.35	98.14
SaliencyMix+weight	93.45	82.99	70.67	90.65	91.68	80.04	69.03	99.99
Δ	+0.46	+1.54	+1.65	+1.32	+0.32	+0.78	+3.68	+1.85

(c) DTD

FGVC-Aircraft								
Method	ResNet18				ResNeXt50			
	low	middle	high	all	low	middle	high	all
CutMix	80.77	71.36	60.98	76.25	89.31	79.36	61.28	86.47
CutMix+weight	81.46	73.24	63.60	77.29	89.42	80.59	63.13	87.43
Δ	+0.69	+1.88	+2.62	+1.04	+0.11	+1.23	+1.85	+0.96
FMix	81.33	71.02	69.35	74.25	89.36	78.86	70.25	82.72
FMix+weight	82.55	73.38	72.05	75.89	89.61	80.54	73.19	84.58
Δ	+1.22	+2.36	+2.70	+1.64	+0.25	+1.68	+2.94	+1.86
SaliencyMix	85.56	78.89	68.36	72.32	91.36	81.22	79.69	85.21
SaliencyMix+weight	85.87	79.87	70.33	72.87	92.02	82.56	81.21	85.48
Δ	+0.31	+0.98	+1.97	+0.64	+0.66	+1.34	+1.52	+0.27

(d) FGVC-Aircraft

CUB200-2011								
Method	ResNet18				ResNeXt50			
	low	middle	high	all	low	middle	high	all
CutMix	90.86	91.68	86.35	92.36	98.35	89.63	75.36	94.31
CutMix+weight	91.16	92.66	87.93	93.32	99.69	92.31	78.34	95.83
Δ	+0.33	+0.98	+1.58	+0.96	+1.34	+2.68	+2.98	+1.52
FMix	92.88	89.36	90.58	90.32	91.31	89.56	86.77	96.79
FMix+weight	93.43	89.68	92.26	91.00	91.64	90.92	86.96	96.97
Δ	+0.55	+0.32	+1.68	+0.68	+0.33	+1.36	+1.25	+0.18
SaliencyMix	90.05	91.02	89.25	93.68	92.86	95.44	89.65	96.59
SaliencyMix+weight	90.70	91.68	90.50	94.00	93.08	95.89	90.67	97.05
Δ	+0.65	+0.66	+1.25	+0.32	+0.22	+0.45	+1.02	+0.46

(e) CUB200-2011

Oxford IIIT Pet								
Method	ResNet18				ResNeXt50			
	low	middle	high	all	low	middle	high	all
CutMix	95.02	89.95	83.36	85.58	93.25	89.95	82.63	87.76
CutMix+weight	95.70	91.88	84.98	86.44	93.37	90.27	84.31	88.28
Δ	+0.68	+1.93	+1.62	+0.86	+0.12	+0.32	+1.68	+0.48
FMix	92.25	91.58	81.16	84.45	93.03	91.12	84.50	85.68
FMix+weight	93.93	92.04	83.15	85.43	93.54	92.74	86.38	86.94
Δ	+1.68	+0.54	+1.99	+0.98	+0.51	+1.62	+1.88	+1.26
SaliencyMix	90.04	91.12	86.65	88.32	94.60	89.99	87.75	88.53
SaliencyMix+weight	91.66	93.01	88.68	89.52	95.14	90.65	88.99	89.92
Δ	+1.62	+1.89	+2.03	+1.20	+0.54	+0.66	+1.24	+1.39

(f) Oxford IIIT Pet

Stanford Dogs								
Method	ResNet18				ResNeXt50			
	low	middle	high	all	low	middle	high	all
CutMix	74.43	69.98	62.24	68.00	85.53	71.15	62.25	69.14
CutMix+weight	77.11	74.93	65.89	71.37	89.78	77.48	69.50	74.57
Δ	+2.68	+4.95	+3.65	+3.37	+4.25	+6.33	+7.25	+5.43
FMix	69.94	71.15	60.35	61.54	81.16	72.25	68.87	74.30
FMix+weight	71.56	74.93	64.37	64.2	84.39	78.07	74.13	79.13
Δ	+1.62	+3.78	+4.02	+2.66	+3.23	+5.82	+5.26	+4.83
SaliencyMix	75.56	68.89	60.04	66.24	81.15	74.65	70.05	72.42
SaliencyMix+weight	77.90	72.22	66.68	71.5	82.80	76.47	72.21	74.27
Δ	+2.34	+3.33	+6.64	+5.26	+1.65	+1.82	+2.16	+1.85

(g) Stanford Dogs

Food-101								
Method	ResNet18				ResNeXt50			
	low	middle	high	all	low	middle	high	all
CutMix	86.66	90.05	81.16	84.25	92.25	89.99	91.16	87.76
CutMix+weight	87.61	91.93	82.92	85.57	92.93	90.65	94.57	89.75
Δ	+0.95	+1.88	+1.76	+1.32	+0.68	+0.66	+3.41	+1.99
FMix	90.02	89.95	76.65	81.66	91.17	84.45	80.00	87.85
FMix+weight	91.70	90.20	78.19	82.01	92.13	85.79	81.66	88.12
Δ	+1.68	+0.25	+1.54	+0.35	+0.96	+1.34	+1.66	+0.27
SaliencyMix	92.22	91.14	89.95	85.54	96.65	91.99	88.82	87.75
SaliencyMix+weight	93.07	92.39	91.12	86.17	97.40	93.61	90.58	89.75
Δ	+0.85	+1.25	+1.47	+0.63	+0.75	+1.62	+1.76	+1.02

(h) Food-101

Stanford Cars								
Method	ResNet18				ResNeXt50			
	low	middle	high	all	low	middle	high	all
CutMix	84.89	81.36	80.02	82.49	93.33	91.47	82.24	92.77
CutMix+weight	90.25	89.6	86.68	89.57	94.01	92.72	83.99	93.86
Δ	+5.36	+8.24	+6.66	+7.08	+0.68	+1.25	+1.75	+1.09
FMix	81.14	82.25	70.02	75.72	92.26	91.14	89.92	90.79
FMix+weight	85.99	88.02	76.91	86.51	92.91	92.34	91.57	92.12
Δ	+4.85	+5.77	+6.89	+10.79	+0.65	+1.20	+1.65	+1.33
SaliencyMix	92.13	89.99	84.45	87.39	93.36	91.12	84.45	92.08
SaliencyMix+weight	95.37	97.55	91.33	93.93	94.33	92.81	86.33	92.96
Δ	+3.24	+7.56	+6.88	+6.54	+0.97	+1.69	+1.88	+0.88

(i) Stanford Cars

Deep Fashion								
Method	ResNet18				ResNeXt50			
	low	middle	high	all	low	middle	high	all
CutMix	89.94	91.14	84.46	87.19	93.33	91.25	88.86	89.55
CutMix+weight	92.81	96.09	88.81	90.73	94.79	92.82	91.54	92.29
Δ	+2.87	+4.95	+4.35	+3.54	+1.46	+1.57	+2.68	+1.74
FMix	94.44	93.36	81.23	87.37	92.28	91.14	87.76	88.95
FMix+weight	94.80	94.8	84.10	82.74	93.17	92.48	89.34	89.62
Δ	+0.36	+1.44	+2.87	+1.54	+0.89	+1.34	+1.58	+0.67
SaliencyMix	93.20	90.02	86.63	87.24	90.03	91.15	86.65	90.53
SaliencyMix+weight	96.88	92.06	88.85	89.81	90.67	92.46	87.85	91.85
Δ	+3.68	+2.04	+2.22	+2.57	+0.64	+1.31	+1.20	+1.32

(j) Deep Fashion

SUN397								
Method	ResNet18				ResNeXt50			
	low	middle	high	all	low	middle	high	all
CutMix	78.85	73.36	68.95	70.31	79.98	78.89	71.62	76.35
CutMix+weight	81.53	76.92	71.26	73.31	81.20	77.21	73.58	77.83
Δ	+2.68	+3.56	+2.31	+3.00	+1.22	+0.86	+1.96	+1.48
FMix	63.34	59.98	61.04	60.03	73.36	69.94	65.58	67.76
FMix+weight	65.29	62.34	63.74	62.42	74.02	71.15	67.26	68.63
Δ	+1.95	+2.36	+2.70	+2.39	+0.66	+1.21	+1.68	+0.87
SaliencyMix	63.33	64.45	58.85	60.91	69.99	68.85	60.02	61.99
SaliencyMix+weight	65.37	66.13	62.71	64.35	70.32	70.11	61.13	63.33
Δ	+2.04	+1.68	+3.86	+3.44	+0.33	+1.26	+1.11	+1.34

(k) SUN397

Oxford-102 Flower								
Method	ResNet18				ResNeXt50			
	low	middle	high	all	low	middle	high	all
CutMix	93.22	94.41	91.15	93.89	96.63	95.53	90.24	97.68
CutMix+weight	95.91	99.40	96.51	98.88	97.25	96.77	92.92	99.33
Δ	+2.69	+4.99	+5.36	+4.99	+0.62	+1.24	+2.68	+1.65
FMix	93.33	92.25	86.62	90.00	94.43	95.56	91.12	94.62
FMix+weight	95.49	97.2	90.30	93.21	97.42	99.04	95.78	99.51
Δ	+2.16	+4.95	+3.68	+3.21	+2.99	+3.48	+4.66	+3.44
SaliencyMix	93.34	90.06	89.95	91.14	90.01	93.32	89.95	94.62
SaliencyMix+weight	96.98	90.22	91.63	93.15	93.99	98.98	95.61	99.63
Δ	+3.64	+0.16	+1.68	+2.01	+3.98	+5.66	+5.31	+5.01

(l) Oxford-102 Flower

Table 2: Evaluation results on 12 visual benchmark datasets. The overall performance improvement of CutMix, FMix, and SaliencyMix using our method is reported. Additionally, the performance of our method on three different CAS-based subsets is presented.

Algorithm 1: Enhancing CutMix with CAS

Input: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, rarity scores $\{r_i\}_{i=1}^{N}$, CutMix parameter $\alpha > 0$, scaling factor $\beta > 0$, training epochs T
Output: Trained model $\mathcal{M}$
// Step 1: Compute Sampling Weights
for each r_i in $\{r_1, r_2, \ldots, r_N\}$ **do**
    $w_i \leftarrow r_i^{\beta}$
end for
Define weight vector $\mathbf{w} = \{w_1, w_2, \ldots, w_N\}$
// Step 2: Initialize Weighted Sampler
Initialize sampler $\mathcal{S}$ using $\mathbf{w}$
// Step 3: Training with CutMix
for $t = 1$ to T **do**
    Sample batch $\mathcal{B}$ from $\mathcal{D}$ using sampler $\mathcal{S}$
    for each pair (x_i, y_i) and (x_j, y_j) in $\mathcal{B}$ **do**
        Sample $\lambda \sim \text{Beta}(\alpha, \alpha)$
        Compute CutMix bounding box B
        Generate binary mask M based on B
        $x_{\text{mix}} \leftarrow x_i \cdot M + x_j \cdot (1 - M)$
        $y_{\text{mix}} \leftarrow \lambda y_i + (1 - \lambda) y_j$
    end for
    Perform forward and backward propagation on mixed samples
    Update model $\mathcal{M}$
end for
return Trained model $\mathcal{M}$

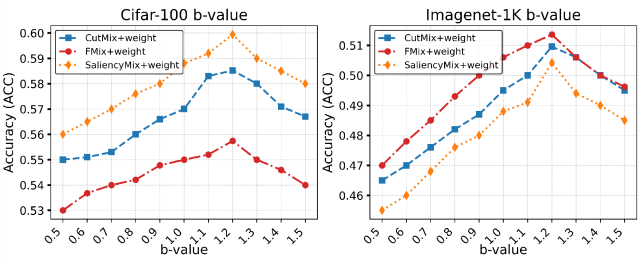

Figure 5: Performance of ResNeXt-50 with our method combined with CutMix, Fmix, and SaliencyMix under different values of b.

Selection of Hyperparameter b

The power parameter b of scarcity augmentation controls the nonlinear amplification of the compositional attribute scarcity. We explored the optimal value of b by setting it within the range of 0.5 to 1.5 on `CIFAR-100` and `ImageNet`. As shown in Figure 5, when $b = 1.2$, our method achieves the highest performance gains for CutMix, FMix, and SaliencyMix. Therefore, we set $b = 1.2$ for main experiments.

Main Results

Table 2 shows the classification results of the model before and after improvements across different datasets. We observed that on all datasets, the performance improved after applying our method, demonstrating the effectiveness of our approach in mitigating attribute imbalance. Impressively, with just our sampling strategy, on `ImageNet-1k` using `ResNeXt-50` as the backbone network, our method improved the overall performance of CutMix, FMix, and SaliencyMix by 1.18%, 1.58%, and 3.07%, respectively. This highlights the necessity of addressing the combined attribute imbalance issue in general-purpose vision datasets.

In fine-grained classification tasks (e.g., `Stanford Dogs`, `Stanford Cars`, and `Oxford-102 Flower`), the performance improvement was most pronounced. Specifically, on `Stanford Dogs`, using ResNeXt-50 as the backbone network, our method improved the overall performance of CutMix, FMix, and SaliencyMix by 5.43%, 4.83%, and 1.85%, respectively. On `Stanford Cars`, using `ResNet-18` as the backbone network, our method achieved performance gains of 7.08%, 10.79%, and 6.54% for CutMix, FMix, and SaliencyMix, respectively. On `Oxford-102 Flower`, using `ResNet-18` as the backbone network, our method improved the overall performance of CutMix, FMix, and SaliencyMix by 4.99%, 3.21%, and 2.01%, respectively.

Impact on Rare Samples

To further analyze the effectiveness of our method across different sparsity levels, we divided the test set into three subsets: high sparsity, medium sparsity, and low sparsity, and evaluated the classification accuracy for each subset. As shown in Table 2, standard data augmentation methods perform poorly on high-sparsity samples, leading to a significant performance gap between low-sparsity and high-sparsity samples. However, after applying our sparsity-based sampling strategy, we observed a notable improvement in classification accuracy for high-sparsity samples, effectively reducing the performance gap between low- and high-sparsity samples.

For instance, on `ImageNet-1k`, using `ResNeXt-50` as the backbone network, our method improved the performance of CutMix, FMix, and SaliencyMix on the high-sparsity subset by 3.54%, 2.61%, and 4.89%, respectively. On the fine-grained image dataset `Stanford Dogs`, with ResNet-18 as the backbone, our method enhanced the performance of CutMix, FMix, and SaliencyMix on the high-sparsity subset by 3.65%, 4.02%, and 6.64%, respectively. Similarly, on `Stanford Cars`, our method boosted the performance of CutMix, FMix, and SaliencyMix on the high-sparsity subset by 6.66%, 6.89%, and 6.88%, respectively. These results demonstrate that sparsity-guided data augmentation effectively improves the model's ability to represent sparse attributes. In summary, our experimental results validate the effectiveness of the sparsity-guided data augmentation approach across multiple datasets and augmentation techniques. This method not only enhances overall classification performance but also significantly improves the model's performance on sparse attribute samples, providing an effective solution to address attribute imbalance issues in real-world applications.

Evaluation under Stronger Vision Backbones

To further examine the impact of *Combinatorial Attribute Scarcity (CAS)* and the generalizability of our sampling strategy, we extend our experiments to a stronger vision backbone — **DINOv2**. This aims to verify whether the CAS problem still exists under improved feature representations,

and whether our sampling method remains effective under high-capacity models. We evaluate on three representative datasets: DTD, Food101, and Oxford-102 Flowers. For each, we apply two widely used augmentation methods: CutMix and SaliencyMix. Table 3 reports performance gains (Overall and High) for CLIP and DINOv2 backbones.

Dataset	Augmentation	Base Model	Δ Overall (%)	Δ High (%)
DTD	CutMix	CLIP	+1.38	+2.88
		DINOv2	+0.97	+1.84
	SaliencyMix	CLIP	+1.32	+1.65
		DINOv2	+1.03	+1.21
Food101	CutMix	CLIP	+1.32	+1.76
		DINOv2	+0.88	+1.22
	SaliencyMix	CLIP	+0.63	+1.47
		DINOv2	+0.59	+1.00
Oxford-102	CutMix	CLIP	+4.99	+5.36
		DINOv2	+3.42	+4.21
	SaliencyMix	CLIP	+2.01	+1.68
		DINOv2	+1.95	+0.87

Table 3: Consistent Gains from CAS-Aware Sampling on Stronger DINOv2 Backbones.

Although DINOv2 improves overall accuracy, the High-scarcity region still lags significantly, validating the persistence of CAS. Our sampling strategy consistently improves performance, showing robustness across architectures. Interestingly, CLIP benefits slightly more, likely due to its superior semantic alignment, especially for fine-grained or attribute-heavy datasets like DTD and Flowers102. These results affirm the generalizability of our CAS-aware sampling method, even under stronger vision backbones.

Conclusion

In this work, we explore the impact of visual attribute imbalance on image classification and propose a sampling strategy based on compositional attribute scarcity (CAS) to improve model performance on rare attributes. By integrating CAS-based sampling with data augmentation techniques, our method effectively enhances the representation of underrepresented attributes. Extensive experiments on twelve benchmark datasets validate its effectiveness in improving both robustness and fairness. Our findings emphasize the importance of attribute-aware learning and provide insights for future research on long-tailed and imbalanced learning.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

The work Supported by Public Computing Cloud, Renmin University of China, Supported by fund for building world-class universities (disciplines) of Renmin University of China.

References

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, 446–461. Springer.

Cai, J.; Wang, Y.; and Hwang, J.-N. 2021. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 112–121.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321–357.

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. Springer.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Cui, J.; Zhong, Z.; Liu, S.; Yu, B.; and Jia, J. 2021. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 715–724.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9268–9277.

Cui, Y.; Song, Y.; Sun, C.; Howard, A.; and Belongie, S. 2018. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4109–4118.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dong, Q.; Gong, S.; and Zhu, X. 2017. Class rectification hard mining for imbalanced deep learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 1851–1860.

Doonan, J. H.; Williams, K.; Corke, F. M.; Zhang, H.; Liu, Y.; et al. 2025. Handling intra-class imbalance in part-segmentation of different wheat cultivars. *Computers and Electronics in Agriculture*, 230: 109826.

Elkan, C. 2001. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, 973–978. Lawrence Erlbaum Associates Ltd.

Estabrooks, A.; Jo, T.; and Japkowicz, N. 2004. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20(1): 18–36.

Hu, X.; Jiang, Y.; Tang, K.; Chen, J.; Miao, C.; and Zhang, H. 2020. Learning to segment the tail. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14045–14054.

Huang, C.; Li, Y.; Loy, C. C.; and Tang, X. 2016. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5375–5384.

Islam, K.; Zaheer, M. Z.; Mahmood, A.; and Nandakumar, K. 2024. DiffuseMix: Label-Preserving Data Augmentation with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27621–27630.

Kang, B.; Li, Y.; Xie, S.; Yuan, Z.; and Feng, J. 2020. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Li, F.-F. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, X.; Zheng, Y.; Ma, H.; Qi, Z.; Meng, X.; and Meng, L. 2024. Cross-modal learning using privileged information for long-tailed image classification. *Computational Visual Media*, 10(5): 981–992.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Liu, B.; Li, H.; Kang, H.; Hua, G.; and Vasconcelos, N. 2021a. Gistnet: a geometric structure transfer network for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8209–8218.

Liu, J.; Sun, Y.; Han, C.; Dou, Z.; and Li, W. 2020. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2970–2979.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1096–1104.

Liu, Z.; Wei, P.; Wei, Z.; Yu, B.; Jiang, J.; Cao, W.; Bian, J.; and Chang, Y. 2021b. Handling inter-class and intra-class imbalance in class-imbalanced learning. *arXiv preprint arXiv:2111.12791*.

Ma, Y.; Jiao, L.; Liu, F.; Li, Y.; Yang, S.; and Liu, X. 2023. Delving into Semantic Scale Imbalance. In *The Eleventh International Conference on Learning Representations*.

Ma, Y.; Jiao, L.; Liu, F.; Wen, M.; Li, L.; Ma, W.; Yang, S.; Liu, X.; and Chen, P. 2025. Predicting and Enhancing the Fairness of DNNs with the Curvature of Perceptual Manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ma, Y.; Jiao, L.; Liu, F.; Yang, S.; Liu, X.; and Chen, P. 2024a. Feature Distribution Representation Learning Based on Knowledge Transfer for Long-Tailed Classification. *IEEE Transactions on Multimedia*, 26: 2772–2784.

Ma, Y.; Jiao, L.; Liu, F.; Yang, S.; Liu, X.; and Chen, P. 2024b. Geometric Prior Guided Feature Representation Learning for Long-Tailed Classification. *International Journal of Computer Vision*, 1–18.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, 722–729. IEEE.

Ouyang, W.; Wang, X.; Zhang, C.; and Yang, X. 2016. Factors in finetuning deep model for object detection with long-tail distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 864–873.

Park, S.; Hong, Y.; Heo, B.; Yun, S.; and Choi, J. Y. 2022. The Majority Can Help The Minority: Context-rich Minority Oversampling for Long-tailed Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6887–6896.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Pham, K.; Kafle, K.; Lin, Z.; Ding, Z.; Cohen, S.; Tran, Q.; and Shrivastava, A. 2021. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13018–13028.

Qin, H.; Jin, X.; Zhu, H.; Liao, H.; El-Yacoubi, M. A.; and Gao, X. 2024. Sumix: Mixup with semantic and uncertain information. In *European Conference on Computer Vision*, 70–88. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, J.; Yu, C.; Ma, X.; Zhao, H.; Yi, S.; et al. 2020. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems*, 33: 4175–4186.

Sinha, S.; Ohashi, H.; and Nakamura, K. 2020. Class-wise difficulty-balanced loss for solving class-imbalance. In *Proceedings of the Asian Conference on Computer Vision*.

Sinha, S.; Ohashi, H.; and Nakamura, K. 2022. Class-Difficulty Based Methods for Long-Tailed Visual Recognition. *International Journal of Computer Vision*, 130(10): 2517–2531.

Tan, J.; Wang, C.; Li, B.; Li, Q.; Ouyang, W.; Yin, C.; and Yan, J. 2020. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11662–11671.

Tang, K.; Tao, M.; Qi, J.; Liu, Z.; and Zhang, H. 2022. Invariant feature learning for generalized long-tailed classification. In *European Conference on Computer Vision*, 709–726. Springer.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, T.; Li, Y.; Kang, B.; Li, J.; Liew, J.; Tang, S.; Hoi, S.; and Feng, J. 2020a. The devil is in classification: A simple framework for long-tail instance segmentation. In *European Conference on Computer Vision*, 728–744. Springer.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. X. 2020b. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Yang, S.; Chen, Z.; Chen, P.; Fang, X.; Liang, Y.; Liu, S.; and Chen, Y. 2024. Defect spectrum: a granular look of large-scale defect datasets with rich semantics. In *European Conference on Computer Vision*, 187–203. Springer.

Yang, Y.; and Xu, Z. 2020. Rethinking the value of labels for improving class-imbalanced learning. *Advances in Neural Information Processing Systems*, 33: 19290–19301.

Ye, H.-J.; Chen, H.-Y.; Zhan, D.-C.; and Chao, W.-L. 2020. Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprint arXiv:2001.01385*.

Yin, X.; Yu, X.; Sohn, K.; Liu, X.; and Chandraker, M. 2019. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5704–5713.

Zang, Y.; Huang, C.; and Loy, C. C. 2021. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3457–3466.

Zhang, Y.; Hooi, B.; Hong, L.; and Feng, J. 2021a. Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. *arXiv preprint arXiv:2107.09249*.

Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; and Feng, J. 2021b. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*.

Zhang, Y.; Zhang, C.; Yu, K.; Tang, Y.; and He, Z. 2024. Concept-Guided Prompt Learning for Generalization in Vision-Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7377–7386.

Zhang, Z.; and Pfister, T. 2021. Learning fast sample re-weighting without reward data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 725–734.

Zhao, B.; Fu, Y.; Liang, R.; Wu, J.; Wang, Y.; and Wang, Y. 2019. A large-scale attribute dataset for zero-shot learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, 0–0.

Zhao, P.; Zhang, Y.; Wu, M.; Hoi, S. C.; Tan, M.; and Huang, J. 2018. Adaptive cost-sensitive online classification. *IEEE Transactions on Knowledge and Data Engineering*, 31(2): 214–228.

Zhong, Z.; Cui, J.; Liu, S.; and Jia, J. 2021. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16489–16498.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9719–9728.

Zhou, Y.; Yang, B.; Lin, X.; Higashita, R.; and Liu, J. 2023. Global-Local Framework for Medical Image Segmentation with Intra-class Imbalance Problem. In *Proceedings of the 2023 2nd Asia Conference on Algorithms, Computing and Machine Learning*, 366–370.

Zhou, Z.-H.; and Liu, X.-Y. 2005. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1): 63–77.

Related Work

Long-tailed image recognition

In practice, the dataset usually tends to follow a long-tailed distribution, which leads to models with very large variances in performance on each class. It should be noted that most researchers default to the main motivation for long-tail visual recognition is that classes with few samples are always weak classes. Therefore, numerous methods have been proposed to improve the performance of the model on tail classes. (Zhang et al. 2021b) divides these methods into three fields, namely class rebalancing (Sinha, Ohashi, and Nakamura 2022; Cui et al. 2019; Lin et al. 2017; Elkan 2001; Zhou and Liu 2005; Zhao et al. 2018; Ye et al. 2020; Chawla et al. 2002; Wang et al. 2020a; Estabrooks, Jo, and Japkowicz 2004; Zhang and Pfister 2021; Zhong et al. 2021), information augmentation (Ma et al. 2024b,a; Chu et al. 2020; Liu et al. 2021a; Park et al. 2022; Cui et al. 2018; Yang and Xu 2020; Hu et al. 2020; Zang, Huang, and Loy 2021), and module improvement (Cui et al. 2021; Ouyang et al. 2016; Zhou et al. 2020; Wang et al. 2020a; Cai, Wang, and Hwang 2021; Wang et al. 2020b; Zhang et al. 2021a). Unlike the above, (Sinha, Ohashi, and Nakamura 2022) and (Ma et al. 2023) observe that the number of samples in the class does not exactly show a positive correlation with the accuracy, and the accuracy of some tail classes is even higher than the accuracy of the head class. Therefore, they propose to use other measures to gauge the learning difficulty of the classes rather than relying on the sample number alone.

Discussion on Intra-Class Imbalance

The fundamental goal of exploring intra-class imbalance is to identify factors that cause differences in recognition performance among samples within the same class, thereby enabling targeted model improvements. (Liu et al. 2021b) attempted to define an imbalanced distribution of learning difficulty within a class, where learning difficulty is determined by the model's prediction confidence. However, prediction confidence varies across different models, leading to inconsistent quantification of learning difficulty, which lacks reproducibility, transparency, and interpretability. (Tang et al. 2022) proposed investigating the long-tail distribution of attributes within a class but only provided qualitative analyses of how attribute imbalance might negatively affect model performance.

In practical applications, (Doonan et al. 2025) addressed the imbalanced distribution of plant traits in wheat recognition by applying weighted point cloud sampling to increase the proportion of rare plant traits. Similarly, (Yang et al. 2024) focused on generating data for specific defect types in industrial defect detection to balance subclass distributions, effectively improving defect detection accuracy. (Zhou et al. 2023) explored the relationship between noise interference and camera angle imbalance with segmentation performance in medical image segmentation tasks. However, these studies are limited to specific domains and lack generalizability.

To date, no research has systematically defined general visual attributes, thoroughly investigated the prevalence of attribute imbalance, or examined whether its negative impact on models warrants widespread attention from researchers.