From Texture to Semantics: Uncertainty-Gated Data Synthesis for Long-Tailed Recognition

1st Xuantong Liu, 2nd Jiaxin Yang, 3rd Shuohao Li*, 4th Jun Zhang, and 5th Jun Lei

Laboratory for Big Data and Decision
National University of Defense Technology, Changsha City, China
{liuxuantong, yangjiaxin21, lishuohao, zhangjun1975, leijun1987}@nudt.edu.cn

***Abstract*—Real-world visual recognition is frequently challenged by long-tailed distributions, where the scarcity of tail-class samples severely degrades model performance. While classical approaches like re-balancing and Logit Adjustment (LA) mitigate prediction bias, they fail to fundamentally address the information deficiency inherent in tail classes. Data synthesis offers a solution but often suffers from blind ropping (e.g., standard CutMix) or reliance on external pre-trained segmentation models. To bridge this gap, we propose Dual-Prior Mix (DP-Mix), a novel data augmentation strategy that synthesizes diverse tail samples by transplanting precise tail foregrounds onto rich head-class backgrounds. Specifically, we leverage two intrinsic image priors: a Texture Prior for structural integrity and a Semantic Prior for class-discriminative localization. Central to our method is the Uncertainty-Gated Selection mechanism, which dynamically switches between these priors based on mask entropy. This ensures high-quality object extraction throughout the training process—from the "cold start" phase to convergence—without requiring any pre-trained models. Extensive experiments on CIFAR-10/100-LT and ImageNet-LT demonstrate that DP-Mix not only achieves competitive performance but also consistently enhances the efficacy of existing mixing-based strategies.**

***Index Terms*—Long-Tailed Recognition, Data Augmentation, Sample Synthesis, Few-Shot Learning**

I. Introduction

The unprecedented success of deep learning in visual recognition is largely attributed to the availability of massive, manually annotated datasets. However, while academic benchmarks typically exhibit a balanced class distribution, real-world data inevitably follows a long-tailed distribution. In such scenarios, a small portion of head classes dominates the sample count, while the vast majority of tail classes are severely under-represented. This extreme imbalance poses a fundamental challenge: standard training paradigms tend to be biased toward head classes, resulting in poor generalization on the tail.

To mitigate this bias, extensive research has been devoted to class re-balancing. Classical strategies, such as Loss Re-weighting and Logit Adjustment [1], [2], attempt to calibrate the model by adjusting the contribution of tail samples. However, these methods often incur a significant performance trade-off, sacrificing head-class accuracy to improve the tail. More recently, ensemble-based methods [3], [4] have achieved state-of-the-art results by training multiple expert networks for different segments of the distribution. Despite their effectiveness, these approaches typically require complex multi-branch architectures or multi-stage training pipelines, which significantly increases computational overhead and limits their flexibility in standard deployment scenarios.

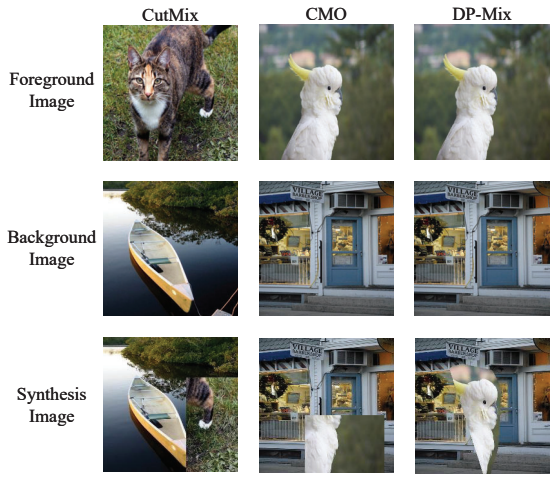

Fig. 1. Visual comparison of different data synthesis strategies. As observed in the first two columns, existing methods like CutMix and CMO rely on blind, box-level cropping. In contrast, our proposed DP-Mix performs precise, content-aware extraction.

To enrich tail-class samples, traditional re-sampling methods typically oversample minority classes. However, this straightforward approach risks overfitting to scarce tail samples by introducing massive redundancy. Consequently, data augmentation has emerged as a promising alternative to enrich the tail-class feature space. Notably, CutMix [5] synthesizes training samples by cutting a random image patch and pasting it onto another image. Building on this, Context-Rich Minority Oversampling (CMO) [6] tailors CutMix for long-tailed scenarios by sampling foreground sources from the tail (based on inverse class frequency) and background sources from the head (following the original distribution) to transfer contextual diversity. However, both methods suffer from a critical limitation: the lack of object localization. Relying on blind, box-level cropping, they frequently fail to capture the discriminative foreground of tail objects or, conversely, introduce irrelevant background noise, as shown in the first two columns in Fig. 1. This issue is particularly detrimental

in long-tailed settings, where tail samples are already scarce; obscuring valuable instances with misleading background patches severely hinders the model's ability to learn robust class boundaries. To overcome this, we propose Dual-Prior Mix (DP-Mix), a precision-driven synthesis strategy that leverages intrinsic texture and semantic priors to precisely extract object foregrounds, as illustrated in the third column of Fig. 1. Crucially, we introduce an Uncertainty-Gated Selection mechanism to dynamically adapt to the model's evolving capability. By evaluating the entropy of the model's activation maps, this mechanism dynamically arbitrates between structural texture cues and high-level semantics, ensuring high-fidelity sample synthesis across the entire training process without relying on external segmentation models.

To summarize, our main contributions are as follows:

- **Dual-Prior Collaboration for Data Synthesis**: We propose DP-Mix, an augmentation framework that integrates intrinsic Texture Priors (low-level structure) and Semantic Priors (high-level class activation). Unlike previous methods that rely on blind cropping or external pre-trained segmentation models, DP-Mix effectively solves the "cold-start" localization problem. By utilizing gradient-based texture cues as a proxy for localization during early training stages, we ensure precise foreground extraction even when the model lacks semantic discrimination.
- **Uncertainty-Gated Selection Mechanism**: We devise a dynamic, self-regulating Uncertainty-Gated Selection strategy. By quantifying the spatial entropy of the model's activation maps, this mechanism acts as an adaptive arbiter that monitors the reliability of the semantic branch. It automatically switches from the texture prior to the refined semantic prior only when the model becomes "confident," ensuring high-fidelity mask generation throughout the entire training process.
- **Extensive Empirical Validation**: We conduct extensive experiments on standard long-tailed benchmarks, including CIFAR-10-LT, CIFAR-100-LT, and ImageNet-LT. The results demonstrate that DP-Mix serves as a generic, cost-effective plugin that effectively enhances the performance and robustness of existing methods utilizing mixing-based strategies.

II. Related Work

A. Long-Tailed Recognition

Current solutions for long-tailed recognition primarily focus on mitigating the prediction bias towards head classes. A prevalent strategy is class re-balancing, which aims to adjust the training distribution or the objective function. While resampling methods balance the dataset by modifying sample counts, they frequently induce overfitting to redundant tail samples or risk discarding valuable head-class information. To avoid discarding data, reweighting methods assign higher importance to minority classes in the loss calculation. Early works adopted inverse class frequency weighting, while subsequent studies introduced more sophisticated metrics. For instance, the Class-Balanced Loss [7] leverages the effective number of samples, and LDAM [8] encourages larger margins for tail classes to improve generalization. At the instance level, Focal Loss [9] and Influence-Balanced Loss [10] dynamically re-weight samples based on their training difficulty.

Beyond simple re-balancing, advanced paradigms like decoupling representation from classifier training [11], [12] demonstrate that high-quality features are learnable from imbalanced data. Building on this, ensemble methods like RIDE [3] and MDCS [4] use multi-branch networks to capture diverse features. Although these strategies effectively calibrate the decision boundary, they still rely on fixed, information-scarce tail samples. Addressing this inherent information deficiency necessitates data augmentation to synthesize new, diverse training examples.

B. Data Augmentation and Mixing Strategies

Data augmentation serves as a fundamental tool to enrich the training set and improve generalization. While conventional geometric transformations like flipping and cropping are standard, modern approaches have shifted towards mixed sample synthesis. Techniques like Mixup [13] and CutMix [5] generate synthetic samples by interpolating images and labels. CutMix, which replaces a random patch of an image with one from another, has proven particularly effective for convolutional networks.

In the context of long-tailed learning, standard CutMix may inadvertently suppress tail classes if not carefully managed. To address this, Context-Rich Minority Oversampling (CMO) [6] adapts CutMix by sampling foregrounds from tail classes and backgrounds from head classes, thereby transferring rich contexts to the minority. However, a critical limitation remains in these mixing strategies. Methods like CMO rely on stochastic region selection, which risks cutting out the tail object entirely or introducing irrelevant background information. Conversely, advanced saliency-based methods [14], [15] often necessitate the use of external pre-trained feature extractors or even heavy pre-trained segmentation models (e.g., Segment Anything Model (SAM)) [16] to obtain high-quality masks. This dependency introduces significant computational overhead and limits their flexibility in standard training-from-scratch settings. This dilemma motivates our proposed method, which utilizes intrinsic texture and semantic priors to ensure high-fidelity synthesis without the need for any auxiliary pre-trained weights.

III. Method

In this section, we propose DP-Mix, a framework enabling precision localization for long-tailed synthesis. As illustrated in Fig. 2, the core lies in the collaboration of two components: 1) Dual-Prior Mask Generation, which derives foreground masks from both intrinsic texture cues and learned semantic activations; and 2) Uncertainty-Gated Selection, a dynamic mechanism that utilizes mask entropy to automatically arbitrate between these priors, thereby selecting the most precise foreground mask for synthesis.

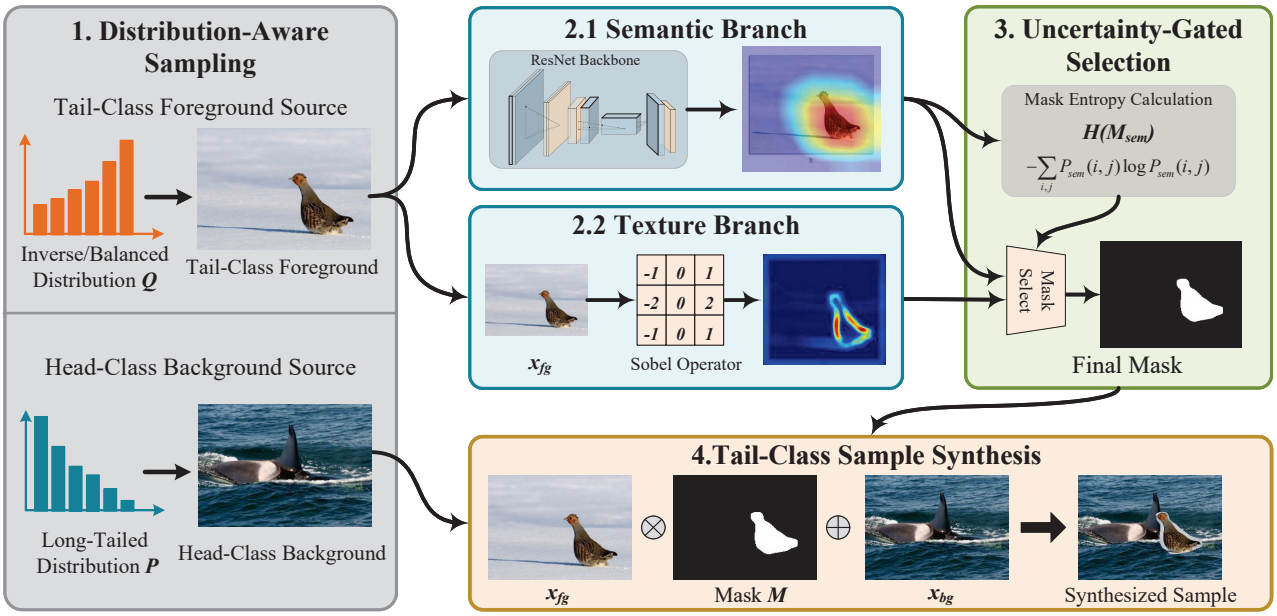

Fig. 2. Overview of the proposed DP-Mix framework. The pipeline consists of four integrated stages: (1) Tail foregrounds and head backgrounds are sampled from different distributions. (2) Candidate masks are generated via a Texture Branch and a Semantic Branch. (3) An Uncertainty-Gated Selection module monitors the spatial entropy of the semantic heatmap, and dynamically arbitrates between the two priors. (4) The final sample is synthesized by compositing the extracted foreground onto a diverse background.

A. Preliminaries

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a long-tailed training dataset, where $x_i \in \mathbb{R}^{H\times W\times C}$ represents an input image and $y_i \in \{1, \ldots, K\}$ is the corresponding class label. The dataset follows a long-tailed distribution where the class frequencies exhibit an exponential decay, satisfying $N_1 \geq N_2 \geq \cdots \geq N_K$, where N_k denotes the number of samples in class k. The degree of imbalance is typically measured by the imbalance ratio $\rho = N_1/N_K \gg 1$.

To synthesize diverse tail-class samples, our DP-Mix framework employs a dual-sampling strategy based on two distinct probability distributions over the classes:

1) Original Distribution P: This distribution reflects the original data frequency, defined as $P(k) = N_k/N$. Sampling from P predominantly yields head-class samples, providing rich contextual variations (backgrounds).

2) Inverse Distribution Q: To prioritize tail classes, we define a class-balanced inverse distribution $Q(k) \propto 1/N_k$. Sampling from Q ensures sufficient coverage of tail categories (foregrounds).

During each training iteration, we sample a **background image** (x_{bg}, y_{bg}) from the original distribution P and a **foreground source** (x_{fg}, y_{fg}) from the inverse distribution Q. Our objective is to generate a synthesized sample $\tilde{x}$ by grafting the semantic foreground of x_{fg} onto x_{bg} via a binary mask $\mathbf{M} \in \{0, 1\}^{H\times W}$:

$$\tilde{x} = \mathbf{M} \odot x_{fg} + (\mathbf{1} - \mathbf{M}) \odot x_{bg}, \tag{1}$$

where $\odot$ denotes element-wise multiplication.

B. Dual-Prior Mask Generation

To overcome the semantic ambiguity caused by blind cropping, we introduce a dual-branch mechanism to extract precise foreground masks $\mathbf{M}$ from the tail-class source image x_{fg}. This mechanism leverages two complementary intrinsic priors: low-level texture structure and high-level semantic activation.

Structural Texture Prior Branch. In the early stages of training, the deep network has not yet learned robust discriminative features, rendering semantic-based localization unreliable. To address this cold-start problem, we utilize an intrinsic texture prior that relies solely on pixel-level gradients, independent of the model's training state.

Specifically, we employ the Sobel operator to calculate the image gradient magnitude, which serves as a proxy for object boundaries. Let G_x and G_y denote the horizontal and vertical gradients of the input image x_{fg}. The gradient magnitude map $\mathcal{G} \in \mathbb{R}^{H\times W}$ is computed as:

$$\mathcal{G}(i, j) = \sqrt{G_x(i, j)^2 + G_y(i, j)^2}. \tag{2}$$

To obtain a binary mask, we apply a thresholding operation. While dynamic thresholding (e.g., Otsu's method) is possible, we find that a mean-based binarization is computationally efficient and effective. The texture mask $\mathbf{M}_{tex}$ is defined as:

$$\mathbf{M}_{tex} = \mathbb{I}(\mathcal{G} > \mu_{\mathcal{G}}), \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $\mu_{\mathcal{G}}$ is the mean value of $\mathcal{G}$. This branch captures high-frequency structural details such as edges and contours, ensuring that the object's shape is preserved even when the model is under-trained.

Learnable Semantic Prior Branch. While the texture prior captures structural boundaries, it is class-agnostic and may include background clutter (e.g., grass texture around a dog). To capture class-specific discriminative regions, we leverage the Class Activation Map (CAM) derived from the on-the-fly model encoder f_θ.

Let $F \in \mathbb{R}^{C' \times H' \times W'}$ represent the feature maps from the final convolutional layer of the encoder, and w_c denote the classifier weights corresponding to the ground-truth class c of x_{fg}. The raw class activation map A^c is obtained by computing the weighted sum of feature maps:

$$A^c = \sum_{k=1}^{C'} w_c^k F^k, \tag{4}$$

where F^k is the k-th feature map. Since A^c typically has a lower resolution than the input image, we bilinearly upsample it to $H \times W$. To generate the semantic mask, we apply ReLU activation to remove negative contributions and normalize the map to the range $[0, 1]$:

$$\mathbf{M}_{sem} = \frac{\text{ReLU}(A^c) - \min(A^c)}{\max(A^c) - \min(A^c) + \epsilon}, \tag{5}$$

where ϵ is a small constant for numerical stability. Unlike the fixed texture mask, $\mathbf{M}_{sem}$ evolves dynamically. As training progresses, the model focuses more accurately on the discriminative parts of the tail object, providing a semantic-aware localization that is superior to the low-level texture cues.

C. Uncertainty-Gated Selection

A core challenge in utilizing the semantic prior is the reliability of the model's activation maps. In the early stages of training, the model cannot distinguish foreground objects from the background, resulting in diffuse and noisy class activation maps (CAMs). Blindly relying on these maps would lead to incorrect cropping. Conversely, the texture prior is robust but lacks semantic specificity. To resolve this dilemma, we introduce an **Uncertainty-Gated Selection** mechanism that dynamically arbitrates between the two priors based on the spatial entropy of the model's attention.

Entropy as Uncertainty Metric. We interpret the normalized semantic mask as a probability distribution over the spatial locations, reflecting where the model focuses on. Let $\mathbf{M}_{sem} \in [0, 1]^{H \times W}$ be the semantic mask generated in Eq. 5. We first normalize it to form a valid spatial probability map $P_{sem} \in \mathbb{R}^{H \times W}$:

$$P_{sem}(i, j) = \frac{\mathbf{M}_{sem}(i, j)}{\sum_{h,w} \mathbf{M}_{sem}(h, w) + \epsilon}. \tag{6}$$

We then calculate the Shannon Entropy $\mathbf{H}(M_{sem})$ to quantify the "peakedness" or concentration of the attention:

$$\mathbf{H}(M_{sem}) = -\sum_{i=1}^{H} \sum_{j=1}^{W} P_{sem}(i, j) \log P_{sem}(i, j). \tag{7}$$

A high entropy value $\mathbf{H}(M_{sem})$ implies a uniform distribution, indicating that the model is "confused" and attending to the entire image diffusely (typical in early epochs). Conversely, a low entropy value indicates that the model's attention is highly concentrated on specific regions, signaling high confidence in localization.

Dynamic Selection and Synthesis. Based on the calculated entropy, we employ a gating mechanism to select the optimal mask $\mathbf{M}^*$. We define a threshold τ to delineate the boundary between uncertainty and confidence:

$$\mathbf{M}^* = \begin{cases} \mathbf{M}_{sem} & \text{if } \mathbf{H}(M_{sem}) < \tau \text{ (Confident: Use Semantic)} \\ \mathbf{M}_{tex} & \text{if } \mathbf{H}(M_{sem}) \geq \tau \text{ (Uncertain: Use Texture)} \end{cases} \tag{8}$$

where τ acts as a hyperparameter controlling the sensitivity of the transition. We empirically set $\tau = 0.6$ based on our experimental results.

This selection mechanism creates a natural curriculum: in the cold start phase, the high entropy forces the model to fall back to the texture prior, guaranteeing the structural integrity of the cropped object. As the model learns to localize discriminative features, the entropy decreases, automatically switching to the semantic prior for more precise, class-aware extraction.

Finally, the synthesized tail-class sample $\tilde{x}$ is generated by compositing the tail foreground x_{fg} and head background x_{bg} using the selected mask $\mathbf{M}^*$:

$$\tilde{x} = \mathbf{M}^* \odot x_{fg} + (\mathbf{1} - \mathbf{M}^*) \odot x_{bg}. \tag{9}$$

This uncertainty-gated synthesis ensures that the augmented samples remain semantically valid throughout the entire training process, effectively preventing the "blind cropping" issues prevalent in prior art.

IV. Experiments

In this section, we perform extensive experiments on three standard long-tailed recognition benchmarks: **CIFAR-10-LT**, **CIFAR-100-LT**, and **ImageNet-LT**, to evaluate the effectiveness of our proposed DP-Mix framework. We systematically compare DP-Mix with a wide range of state-of-the-art approaches, including classical re-balancing strategies, decoupling methods, and advanced data synthesis techniques. Furthermore, we provide comprehensive ablation studies and qualitative visualizations to verify the specific contribution of the Dual-Prior mechanism and the Uncertainty-Gated Selection strategy.

A. Experimental Settings

Datasets. We perform experiments on the long-tailed versions of CIFAR-10 and CIFAR-100. Following standard protocols [27], the long-tailed datasets are constructed by downsampling training samples according to an exponential decay function $n_k = n_1 \mu^k$, where the imbalance factor $\beta = N_{max}/N_{min}$ is set to 100, 50, or 10.

Architectures. For fair comparison with prior art, we adopt standard backbones trained from scratch. Specifically, we use ResNet-32 for all CIFAR-LT experiments and ResNet-50 for ImageNet-LT. Note that our framework relies solely on

TABLE I
PERFORMANCE COMPARISON ON LONG-TAILED DATASETS

Method	Year	Venue	CIFAR-10-LT			CIFAR-100-LT			ImageNet-LT			
			IR=10	IR=50	IR=100	IR=10	IR=50	IR=100	Many	Medium	Few	Overall
Backbone			ResNet-32			ResNet-32			ResNet-50			
cRT [11]	2020	*ICLR*	-	-	-	-	43.8	-	58.8	44.0	26.1	47.3
BBN [17]	2020	*CVPR*	88.32	82.18	79.82	59.12	47.02	42.36	-	-	-	-
LADE [8]	2021	*CVPR*	-	-	-	61.7	50.5	45.4	65.1	48.9	33.4	53.0
PaCo [18]	2021	*ICCV*	-	-	-	64.2	56.0	52.0	65.0	55.7	38.2	57.0
CMO [6]	2022	*CVPR*	-	-	-	60.2	53.0	50.0	66.4	53.9	35.6	56.2
FRS [19]	2022	*ICME*	87.7	82.2	80.0	59.4	48.4	44.2	66.7	45.4	21.6	50.3
GLMC [20]	2023	*CVPR*	94.87	91.04	88.50	73.40	63.78	57.97	**70.1**	52.4	30.4	56.3
AREA [21]	2023	*ICCV*	88.71	82.68	78.88	60.77	51.77	48.83	-	-	-	49.5
MDCS [4]	2023	*ICCV*	-	89.4	85.8	-	60.1	56.1	-	-	-	60.7
FBC [22]	2023	*ICME*	-	-	-	-	58.17	53.88	-	-	-	52.99
DODA [23]	2024	*ICLR*	-	-	-	62.7	53.6	51.0	66.9	54.1	37.4	56.9
LOS [24]	2025	*ICLR*	-	-	-	69.7	58.8	54.9	63.2	50.7	42.3	54.4
FeatRecon [25]	2025	*ICML*	91.6	87.8	85.2	65.3	57.0	52.5	-	-	-	56.8
CALA [26]	2025	*AAAI*	92.47	-	84.79	65.51	-	52.34	-	-	-	-
DP-Mix	-	**Our**	91.15	85.18	83.78	64.60	56.47	54.12	64.72	55.38	38.97	57.1
MDCS + DP-Mix	-	**Our**	94.35	91.27	87.62	72.33	62.59	58.46	68.70	**59.12**	**39.83**	**62.54**
GLMC + DP-Mix	-	**Our**	**95.82**	**92.07**	**89.69**	**74.15**	**64.71**	**58.88**	68.46	53.94	34.12	58.26

these backbones to generate semantic priors (CAM) on-the-fly, without utilizing any external pre-trained feature extractors or segmentation models, ensuring a fair and cost-effective training setup.

Evaluation Metrics. We evaluate the models on the corresponding balanced test/validation sets and report the Top-1 accuracy (%). For detailed analysis on ImageNet-LT, we follow [28] and report accuracy on three subsets based on class frequency: Many-shot (>100 samples), Medium-shot (20∼100 samples), and Few-shot (<20 samples).

B. Main Results

Table I summarizes the performance comparison on CIFAR-10-LT, CIFAR-100-LT, and ImageNet-LT. We compare our method against a wide range of baselines, including classical re-balancing strategies and recent approaches. Crucially, to validate the effectiveness of our proposed precise synthesis strategy, we integrate DP-Mix as a plug-and-play module into two strong mixing-based baselines: GLMC [20] and MDCS [4].

Results on CIFAR-10-LT and CIFAR-100-LT. DP-Mix consistently enhances baseline performance across various imbalance ratios (IR). On CIFAR-100-LT (IR=100), integrating DP-Mix with GLMC yields a solid gain of 0.91%, outperforming standard augmentation strategies. Similarly, on CIFAR-10-LT, our method improves robustness under extreme imbalance. These results empirically verify that replacing the blind cropping of traditional CutMix/CMO with our uncertainty-gated precise localization effectively mitigates feature collapse in tail classes.

Results on ImageNet-LT. On the large-scale ImageNet-LT dataset, the efficacy of DP-Mix is further amplified. When applied to MDCS, our method boosts the overall accuracy by 1.84%. A similar improvement is observed with GLMC. The improvement is primarily driven by the Few-shot subset, where MDCS + DP-Mix achieves 39.83% accuracy. This validates our core motivation: synthesizing high-quality, semantically intact samples for underrepresented classes.

Fig. 3. Qualitative visualization of synthesized samples. DP-Mix effectively leverages dual priors to perform content-aware cropping, preserving the structural integrity of tail objects.

C. Ablation Study and Analysis

To verify the superiority of our strategy, we decouple the method and evaluate the contribution of each component, as shown in Table II. Furthermore, we provide visualizations to intuitively demonstrate how DP-Mix achieves high-quality synthesis.

The vanilla CE baseline struggles on tail classes, achieving only 38.60% accuracy. Our method significantly alleviates

this issue. Specifically, the introduction of dual priors with a random switch strategy alone boosts accuracy by 13.6%, proving the necessity of semantic and texture augmentation. Our full model, DP-Mix, which incorporates the Uncertainty-Gated Selection mechanism, further elevates the gain to +15.5%. This result confirms that our uncertainty-aware gating is not merely a passive switch, but an active component that ensures the optimal prior is utilized for each training instance.

TABLE II
COMPARISON OF DIFFERENT METHODS

Methods	Vanilla	+Dual-prior	DP-Mix
CE	38.6	51.2 (+13.6)	**54.1** (**+15.5**)
LDAM	41.7	52.8 (+11.1)	**56.0** (**+14.3**)

Visualization. Fig. 3 provides a comprehensive visualization of the intermediate artifacts and final outputs generated by DP-Mix. We display the raw foreground and background images, the Class Activation Maps (CAMs) from the semantic branch, the edge contours from the texture branch, the final mask arbitrated by our Uncertainty-Gated Selection, and the resulting synthesized samples. The first three columns illustrate the model's behavior during the early training stages. Here, the semantic heatmaps are diffuse and scattered, indicating high uncertainty. The gating mechanism select texture branch as final mask. In contrast, the last four columns reflect the converged state, where the model focuses intently on discriminative regions. With low entropy, the system automatically switches to the semantic branch for precise extraction. Ultimately, as shown in the final row, the synthesized images successfully capture the holistic foreground objects regardless of the training phase, validating the robustness of our dynamic selection strategy.

V. CONCLUSION

In this work, we present DP-Mix, a robust data synthesis framework designed to address the scarcity of tail-class samples in long-tailed scenarios. By unifying semantic specificity and texture invariance into an adaptive pipeline, our approach effectively overcomes the fragility of single-source priors. Our empirical results demonstrate that the Uncertainty-Gated Selection mechanism plays an important role in this collaboration: it utilizes the texture prior as a structural safety net during the high-entropy cold start phase, while progressively shifting to the semantic prior for precise localization as the model learns discriminative features. This dynamic arbitration ensures the generation of high-quality, semantically valid samples throughout the training process. DP-Mix offers a new perspective on data augmentation and sets a solid foundation for future research in data-centric long-tailed learning.

REFERENCES

[1] Mengke Li, Yiu-ming Cheung, and Yang Lu, "Long-tailed visual recognition via gaussian clouded logit adjustment," in *CVPR*, 2022.

[2] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar, "Long-tail learning via logit adjustment," in *ICLR*, 2021.

[3] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu, "Long-tailed recognition by routing diverse distribution-aware experts," in *ICLR*, 2020.

[4] Qihao Zhao, Chen Jiang, Wei Hu, Fan Zhang, and Jun Liu, "Mdcs: More diverse experts with consistency self-distillation for long-tailed recognition," in *ICCV*, 2023.

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

[6] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi, "The majority can help the minority: Context-rich minority oversampling for long-tailed classification," in *CVPR*, 2022.

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

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

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

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

[11] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *ICLR*, 2019.

[12] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang, "Long-tailed classification by keeping the good and removing the bad momentum causal effect," in *NeurIPS*, 2020.

[13] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[14] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song, "Puzzle mix: Exploiting saliency and local statistics for optimal mixup," in *ICML*, 2020.

[15] Jang-Hyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song, "Co-mixup: Saliency guided joint mixup with supermodular diversity," in *ICLR*, 2021.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al., "Segment anything," in *ICCV*, 2023.

[17] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *CVPR*, 2020.

[18] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia, "Parametric contrastive learning," in *ICCV*, 2021.

[19] Jinmian Cai, Zheng Wang, Huazhu Fu, Jingjing Chen, and Yu-Gang Jiang, "Data-free network debiasing for long-tailed visual recognition," in *ICME*, 2022.

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

[21] Xiaohua Chen, Yucan Zhou, Dayan Wu, Chule Yang, Bo Li, Qinghua Hu, and Weiping Wang, "Area: adaptive reweighting via effective area for long-tailed classification," in *ICCV*, 2023.

[22] Jiaxin Yang, Xiaofei Li, Jun Zhang, and Shuohao Li, "Feature bias correction: A feature augmentation method for long-tailed recognition," in *ICME*, 2023.

[23] Binwu Wang, Pengkun Wang, Wei Xu, Xu Wang, Yudong Zhang, Kun Wang, and Yang Wang, "Kill two birds with one stone: Rethinking data augmentation for deep long-tailed learning," in *ICLR*, 2024.

[24] Siyu Sun, Han Lu, Jiangtong Li, Yichen Xie, Tianjiao Li, Xiaokang Yang, Liqing Zhang, and Junchi Yan, "Rethinking classifier re-training in long-tailed recognition: Label over-smooth can balance," in *ICLR*, 2025.

[25] Lingjie Yi, Jiachen Yao, Weimin Lyu, Haibin Ling, Raphael Douady, and Chao Chen, "Geometry of long-tailed representation learning: Rebalancing features for skewed distributions," in *ICLR*, 2025.

[26] Xiaoling Zhou, Ou Wu, and Nan Yang, "Class and attribute-aware logit adjustment for generalized long-tail learning," in *AAAI*, 2025.

[27] Cao Yue, M Long, J Wang, Zhu Han, and Q Wen, "Deep quantization network for efficient image retrieval," in *AAAI*, 2016.

[28] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu, "Large-scale long-tailed recognition in an open world," in *CVPR*, 2019.