HiGFA: Hierarchical Guidance for Fine-grained Data Augmentation with Diffusion Models

Zhiguang Lu[1,2], Qianqian Xu[1,5]*, Peisong Wen[2], Siran Dai[3,4], Qingming Huang[2,1]*

[1]State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences
[2]School of Computer Science and Technology, University of Chinese Academy of Sciences
[3]Institute of Information Engineering, Chinese Academy of Sciences
[4]School of Cyber Security, University of Chinese Academy of Sciences
[5]Peng Cheng Laboratory
{luzhiguang23s, xuqianqian}@ict.ac.cn, wenpeisong@ucas.ac.cn, daisiran@iie.ac.cn, qmhuang@ucas.ac.cn

Abstract

Generative diffusion models show promise for data augmentation. However, applying them to fine-grained tasks presents a significant challenge: ensuring synthetic images accurately capture the subtle, category-defining features critical for high fidelity. Standard approaches, such as text-based Classifier-Free Guidance (CFG), often lack the required specificity, potentially generating misleading examples that degrade fine-grained classifier performance. To address this, we propose Hierarchically Guided Fine-grained Augmentation (HiGFA). HiGFA leverages the temporal dynamics of the diffusion sampling process. It employs strong text and transformed contour guidance with fixed strengths in the early-to-mid sampling stages to establish overall scene, style, and structure. In the final sampling stages, HiGFA activates a specialized fine-grained classifier guidance and dynamically modulates the strength of all guidance signals based on prediction confidence. This hierarchical, confidence-driven orchestration enables HiGFA to generate diverse yet faithful synthetic images by intelligently balancing global structure formation with precise detail refinement. Experiments on several FGVC datasets demonstrate the effectiveness of HiGFA.

Code — https://github.com/ZhiguangLuu/HiGFA

Introduction

Deep learning models have achieved remarkable success across various domains, yet their performance in specialized tasks like fine-grained visual classification (FGVC) often hinges on large, diverse, and accurately labeled datasets. Acquiring such datasets can be prohibitively expensive and time-consuming. Data augmentation techniques are crucial for mitigating data scarcity and improving model generalization. While traditional augmentation methods, such as geometric transforms, color jittering, offer limited diversity, generative models, particularly diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2021; Rombach et al. 2022; Podell et al. 2024), provide a promising alternative to synthesizing novel and realistic data samples. Typically, existing methods utilize class text or image structure as conditions to guide generation and assume the output matches the target category (Trabucco et al. 2024; He et al. 2023; Zhang, Rao, and Agrawala 2023; Michaeli and Fried 2024). Compared to traditional data augmentation, this technical route significantly increases data diversity, thereby improving model generalization in downstream tasks.

Figure 1: Accurately depicting the red shoulder patches of the Red-winged Blackbird is essential for fine-grained classification. More examples are provided in Appendix.

Unfortunately, without careful design, these methods are inferior to traditional augmentation in fine-grained visual classification tasks due to the **insufficient fidelity to fine-grained categories**. Specifically, fine-grained categories often differ only in subtle characteristics, such as the specific shape of a bird's beak, the precise pattern on an insect's wing, or subtle variations in car model designs. However, text descriptions and image contours fail to faithfully capture these subtle and category-defining visual attributes. For instance, as illustrated in Figure 1, generating an image labeled "Red-winged Blackbird" that omits its distinctive red shoulder patches, or depicts them with incorrect coloration or placement, introduces semantic mismatching that will degrade classifier performance.

In light of this, this work explores the following key question:

> ***How to achieve both diversity and fine-grained fidelity in generative augmentation?***

To investigate the solution to the aforementioned question, we propose a ***H**ierarchically **G**uided **F**ine-grained **A**ugmentation method (**HiGFA**)* that progressively incorporates conditions for controlling diversity and fidelity. Specifically, we jointly consider three conditions: **1)** *Text guidance* for diversity, which defines the overall scene, style, and coarse object attributes; **2)** *Diversity-enhanced Contour guidance*

*Corresponding Authors

for structure fidelity, which ensures realistic object structure and integration with scene; **3)** *Fine-grained classifier guidance* for category-specific fidelity, which ensures that the generated objects preserve the discriminative characteristics. Utilizing a conditional diffusion model (Zhang, Rao, and Agrawala 2023) with three guidance, we can generate samples with both high diversity and strong category fidelity.

To avoid conflicts among the three types of guidance, this paper further proposes a dynamic guidance mechanism. Relying on static weighting for these signals is inherently fragile. For example, strict adherence to contour guidance may enforce a subject's pose but conflict with background described in the text prompt, resulting in an unnatural "cutout" effect. Similarly, excessive classifier guidance can make image collapse. Given that the optimal balance between these signals varies on an image-by-image basis, our approach leverages the observation that diffusion models generate images progressively, from coarse to fine(Li and Chen 2024; Raya and Ambrogioni 2023; Benita, Elad, and Keshet 2025; Koulischer et al. 2025).

Consequently, we exploit this inherent temporal dynamic by controlling the guidance strengths with the classification confidences. The generation process begins with a warm-up phase where text and contour guidance predominantly shape low-frequency image components, establishing global structure. Following this initial phase, classifier guidance is activated. The relative strengths of all three guidance signals, text, contour, and classifier, are then dynamically modulated based on the evolving classification confidence. Specifically, when high class fidelity is rapidly attained for a sample, indicated by high classifier confidence, the weight of the classifier guidance is progressively reduced, allowing text and contour cues to exert greater influence and thereby promote diversity. Conversely, for samples where achieving fine-grained fidelity proves more challenging, evidenced by lower confidence, the classifier guidance maintains a pronounced role, ensuring that category-specific details are accurately rendered until satisfactory confidence levels are reached.

In a nutshell, the contributions of this work are as follows:

- We propose `HiGFA`, a unified framework that synergistically integrates hierarchical guidance sources (text, contour, classifier) within the diffusion process.
- We apply geometric transformations on contour maps to enhance sample diversity and implement a dynamic strategy to maintain a balance between global structure and fine-grained details.
- We demonstrate that `HiGFA` generates diverse, high-fidelity augmented images, leading to superior performance on FGVC tasks compared to existing generative augmentation methods.

Related Work

Data Augmentation via Generative Models. Generative models can synthesize novel and diverse examples that potentially lie outside the empirical support of the original training set, enriching the underlying data manifold. Early work employed generative adversarial networks (GANs) (Goodfellow et al. 2014) to synthesize additional samples for minority classes in imbalanced classification and fine-grained classification tasks (Mariani et al. 2018; Zhang et al. 2021). Several studies (Hemmat et al. 2023; Trabucco et al. 2024; He et al. 2023; Shama Sastry, Dumpala, and Oore 2024) have demonstrated that using diffusion models for data augmentation or expansion can improve performance on downstream classification tasks. Some of these works, such as (Zhang et al. 2023; He et al. 2023; Dunlap et al. 2023), primarily rely on CLIP image features and CLIP-encoded text as conditions for data generation. However, due to the limited fine-grained perceptual capabilities of CLIP (Hua et al. 2025, 2024; Jiang et al. 2023; Han et al. 2025), the generated samples often fail to accurately reflect fine-grained visual characteristics. Yang et al. (2024) proposed guiding the generation process using hierarchical prototypes extracted by feature extractors. However, the selection of prototypes in this method depends on hyperparameter choices, and although this approach enforces feature consistency, it does not explicitly encourage diversity in the generated samples. Michaeli and Fried (2024) introduced the use of ControlNet to condition generation based on image structure and subject. However, for fine-grained image classification, preserving only the structure and subject is insufficient; the key performance factors typically lie in subtle visual details, rather than in maintaining a particular pose or object composition. Our proposed method, `HiGFA`, synergistically integrates multiple guidance sources with a hierarchical and adaptive scheduling strategy. We enhance diversity by applying random rotations and thin-plate spline interpolation to manipulate the edge maps input to ControlNet. To ensure consistency in the generated samples, we incorporate guidance from a fine-grained classifier.

Methodology

The primary objective of employing generative models for fine-grained data augmentation is to synthesize diverse images that faithfully represent the subtle characteristics of specific categories, thereby enhancing the performance of fine-grained classifiers. In this section, we first introduce essential preliminaries on diffusion models and guided generation. We then highlight the limitations of existing diffusion-based augmentation methods, particularly their inability to maintain category fidelity. To address this issue, we propose a Hierarchically Guided Fine-grained Augmentation (`HiGFA`) method designed to preserve fidelity while simultaneously improving data diversity.

Preliminaries

Diffusion Models. Diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2021; Lipman et al. 2023) are deep generative models that generate data by reversing a stochastic process with gradually adding noise. Let $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ be a sample from the true data distribution. The forward process is a Markov chain defined over T timesteps, injecting Gaussian noise according to a variance schedule $\beta_t \in (0, 1)$: $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$. As $t \to T$, $\mathbf{x}_T$ approaches an isotropic Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The reverse process aims to denoise $\mathbf{x}_t$ back to $\mathbf{x}_{t-1}$, typically modeled by a neural network $\epsilon_\theta(\mathbf{x}_t, t)$ that predicts the noise

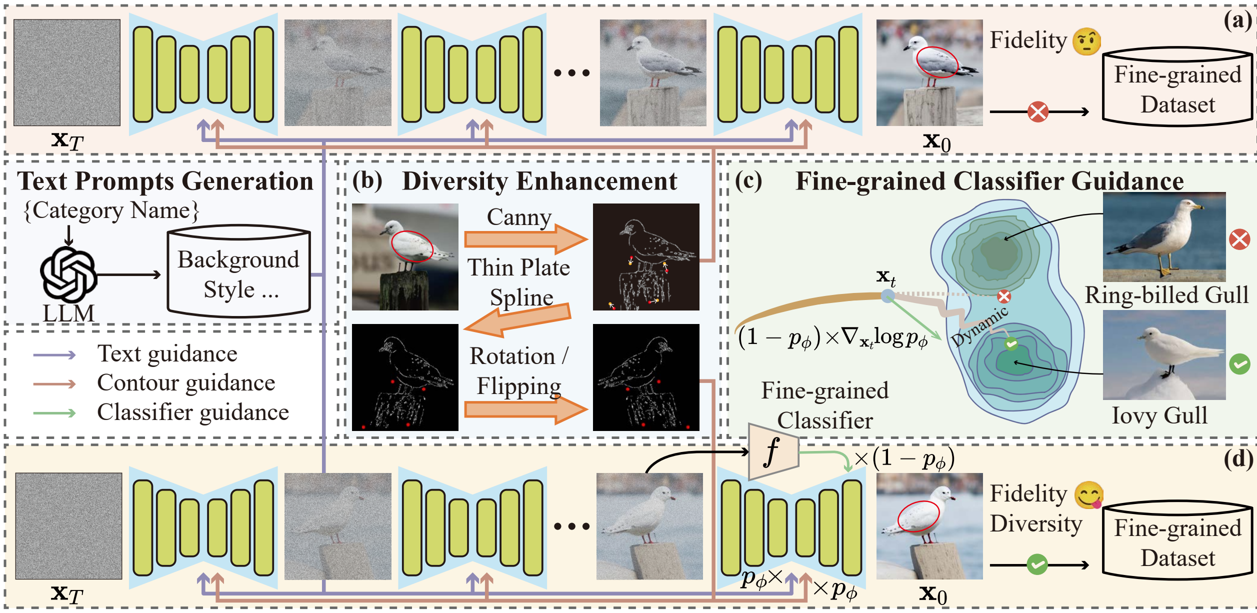


Figure 2: An overview of our `HiGFA`: **a)** Existing diffusion-based augmentation method is insufficient to ensure category fidelity. **b)** Incorporating contour guidance helps preserve scene layout and object structure, but struggles with fine-grained category consistency and limit diversity. To enhance diversity, we apply diversity enhancement such as flipping, rotation, and thin-plate spline interpolation. **c)** To improve fine-grained category consistency, in the later stages, our adaptive strategy incorporates classifier guidance to balance the scale of all three guidance on a sample-wise level. Without classifier guidance, the generation process tends to converge toward the mean of the class distribution. **d)** Under the combination of hierarchical guidance and dynamic adjustment, our method generates augmented images with both high category fidelity and enhanced diversity.

added at step t. The reverse transition $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ can be approximated as:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}_t, t), \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)), \quad (1)$$

where $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ depend on the noise prediction $\epsilon_\theta(\mathbf{x}_t, t)$. At inference, sampling starts from $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively applies the learned reverse process to generate a sample $\mathbf{x}_0$. The basic diffusion model described here is unconditional.

Guided Diffusion. Conditional generation can be achieved using guidance mechanisms. Classifier guidance (Dhariwal and Nichol 2021) uses a pretrained classifier $p_{\boldsymbol{\phi}}(y \mid \mathbf{x})$ to steer the sampling towards a target class y. The mean of the reverse step is adjusted using the classifier's gradient:

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\mathbf{x}_t|y) = \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{x}_t) + s \cdot \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathbf{x}_t)\nabla_{\mathbf{x}_t} \log p_{\boldsymbol{\phi}}(y|\mathbf{x}_t), \quad (2)$$

where s is the guidance scale. Classifier-free guidance (CFG) (Ho and Salimans 2022) avoids external classifiers by jointly training a conditional model $\epsilon_\theta(\mathbf{x}_t, t, y)$ and an unconditional model $\epsilon_\theta(\mathbf{x}_t, t, \emptyset)$ (often by randomly dropping the conditioning y during training). During sampling, the noise prediction is extrapolated from the conditional and unconditional predictions:

$$\hat{\epsilon}_\theta(\mathbf{x}_t|y) = \epsilon_\theta(\mathbf{x}_t|\emptyset) + s \cdot (\epsilon_\theta(\mathbf{x}_t|y) - \epsilon_\theta(\mathbf{x}_t|\emptyset)), \quad (3)$$

where s is the guidance scale, y is the condition (e.g., text prompt), and $\emptyset$ represents the null condition (e.g., empty string). CFG is highly effective and widely used in large-scale text-to-image models (Rombach et al. 2022; Esser et al. 2024; Ramesh et al. 2022; Podell et al. 2024; Han et al. 2024).

Fidelity Challenge in Fine-grained Data Augmentation

A primary challenge when using generative models for fine-grained data augmentation is ensuring the synthetic images accurately reflect the subtle, defining features of the target category. Standard text-to-image models, often guided by text prompts interpreted by models like CLIP (Radford et al. 2021) or LLMs (Raffel et al. 2020), struggle to capture these fine-grained distinctions, potentially leading to generated images with incorrect category attributes.

For instance, consider augmenting images for the "Ivory Gull", a seabird species characterized by entirely white plumage, distinguishing it from most other gulls which typically have gray mantles (Jobling 2010). Existing diffusion-based augmentation methods (Dunlap et al. 2023; He et al. 2023) typically rely on text guidance, often using classifier-free guidance (CFG). Even with the introduction of simple contour guidance (Michaeli and Fried 2024), these methods can only ensure contour similarity but fail to achieve the specificity required to preserve fine-grained key features. Furthermore, maintaining fine-grained category definitions is an image-level challenge, and using fixed guidance weights makes it difficult to balance fidelity and diversity across all images. This can result in generated samples that deviate from the true category characteristics, diminishing the effectiveness of the augmentation.

Hierarchically Guided Fine-grained Augmentation

The core limitation identified above is the lack of precise control at image-level over the generation process to enforce adherence to fine-grained category definitions. To address this, we propose Hierarchically Guided Fine-grained Augmentation (`HiGFA`). Our method is motivated by the observation that diffusion models tend to generate coarse, low-frequency information (e.g., overall structure, scene layout) during the early stages of the reverse sampling process and progressively refine high-frequency details in later stages (Li and Chen 2024; Benita, Elad, and Keshet 2025). `HiGFA` leverages this temporal dynamic by applying multiple forms of guidance with varying strengths across the sampling process, transitioning from coarse-grained control to fine-grained feature enforcement:

- **Text Guidance:** Primarily drives *diversity* by leveraging text prompts via classifier-free guidance to define varied global compositions, scenes, and styles. Its strong initial influence establishes a semantically coherent foundation for image generation.
- **Diversity-enhanced Contour Guidance:** Enforces structural *fidelity* by applying ControlNet (Zhang, Rao, and Agrawala 2023) with transformed edge maps. This ensures realistic structures and seamless integration into scene, while contour transformations further support *diversity* in object pose and layout.
- **Fine-grained Classifier Guidance:** Improves fine-grained *fidelity* by refining category-specific visual details during the final sampling stages. This tailored classifier guidance accurately captures subtle, discriminative features critical to the target class.

Text Guidance. Given a text prompt u, specifying a desired scene, general object category, and optionally an artistic style, we use standard classifier-free guidance (CFG, Eq. 3) to provide coarse-grained control. Following (Michaeli and Fried 2024) prompts are generated using category names and scene descriptions, with optional style modifiers (e.g., "photorealistic", "oil painting") introduced randomly for diversity. The strength of CFG is controlled by a scale parameter $s_{\text{cfg}}(t)$. This scale remains constant during initial sampling steps ($t = T$ down to N_s) to robustly establish the global structure and style, and is dynamically adjusted during later steps.

Diversity-enhanced Contour Guidance. To maintain structural fidelity with real images from the dataset, we incorporate contour guidance using ControlNet (Zhang, Rao, and Agrawala 2023). For a reference image $\mathbf{x}$ from the dataset, we extract its Canny edge map $\mathbf{x}_c$ and condition the diffusion process on it via ControlNet. To enhance diversity beyond the reference image's specific pose and structure, we apply transformations to the Canny edge map $\mathbf{x}_c$ before feeding. For datasets with rigid objects (e.g., cars, airplanes), we apply random horizontal flipping and random rotation (e.g., within $\pm 15°$). For datasets with non-rigid objects (e.g., birds, dogs), we additionally apply thin-plate spline (TPS) warping (Bookstein 1989) to simulate plausible deformations. Specifically, we divide the edge map into patches, randomly select five edge-containing patches, and perturb a point in each to create a target map for TPS. These augmentations encourage pose diversity. ControlNet's influence is governed by a conditioning strength parameter $s_{\text{ctl}}(t)$, which remains constant during the initial phase and is dynamically modulated for $t < N_s$.

Fine-grained Classifier Guidance. To enforce category-specific details, we employ a fine-grained classifier $p_\phi(y \mid \mathbf{x}')$, trained on the original dataset, as a domain-specific expert to ensure the generated images accurately exhibit category-specific details. This guidance is **activated only** during the later sampling stages (for $t < N_s$), when structural elements are largely resolved and noise levels are reduced. Let $p_\phi(y|\mathbf{x}')$ be the fine-grained classifier, where $\mathbf{x}'$ is an image and y is the target fine-grained label. Since the classifier operates on images rather than latent variables, we estimate the corresponding clean image $\hat{\mathbf{x}}_0$ from the noisy latent $\mathbf{x}_t$ using the standard denoising prediction:

$$\hat{\mathbf{x}}_0(\mathbf{x}_t, t) = \frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}}, \tag{4}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$. The predicted latent $\hat{\mathbf{x}}_0$ is then decoded to an image $\mathbf{x}'_t$ using the VAE decoder $\mathfrak{D}$:

$$\mathbf{x}'_t = \mathfrak{D}(\hat{\mathbf{x}}_0(\mathbf{x}_t, t)). \tag{5}$$

The fine-grained guidance adjusts the noise prediction ϵ_θ, which already incorporates CFG and ControlNet, denoted ϵ_θ. For $t < N_s$, the adjusted noise prediction is given by:

$$\tilde{\epsilon}_\theta(\mathbf{x}_t, t) = \epsilon_\theta(\mathbf{x}_t, t) - s_{\text{cls}}(t) \cdot \sigma_t \cdot \nabla_{\mathbf{x}_t} \log p_\phi(y \mid \mathbf{x}'_t), \tag{6}$$

where $s_{\text{cls}}(t)$ is the classifier guidance scale, and $\sigma_t = 1.0$ is the gradient scaling factor, following standard practice (Nichol et al. 2021).

Crucially, accounting for the varying difficulty in extracting fine-grained features from images within the same dataset is vital. For example, images with clean backgrounds require minimal classifier guidance to avoid image collapse due to excessive guidance, whereas images with cluttered backgrounds or multiple objects demand stronger guidance to preserve class fidelity. Therefore, we propose a dynamic guidance adjustment mechanism that adaptively modulates the strength of guidance based on classifier predictions, aligning it with the complexity of individual images. All guidance strengths are dynamically orchestrated in two phases, delineated by a threshold timestep N_s. Let $p_\phi(y \mid \mathbf{x}'_t)$ denote the classifier's confidence in the target class at time t. During the early steps ($t \geq N_s$), the objective is to establish global structure and style. Both s_{cfg} and s_{ctl} are fixed, and fine-grained guidance remains inactive. In the later steps ($t < N_s$), the objective shifts to refining fine-grained details. The guidance strengths are updated as follows:

$$\begin{aligned} s_{\text{cfg}}(t) &= s_{\text{cfg}}(T) \cdot p_\phi(y \mid \mathbf{x}'_{t-1}), \\ s_{\text{ctl}}(t) &= s_{\text{ctl}}(T) \cdot p_\phi(y \mid \mathbf{x}'_{t-1}), \\ s_{\text{cls}}(t) &= s_{\text{cls}}(N_s) \cdot (1 - p_\phi(y \mid \mathbf{x}'_t)), \end{aligned} \tag{7}$$

where $s_{\text{cfg}}(T)$ and $s_{\text{ctl}}(T)$ are the initial CFG and ControlNet scales, respectively, and $s_{\text{cls}}(N_s)$ is the initial classifier guidance scale. When the classifier exhibits uncertain (i.e.,

Type	Methods	Rigid Objects			Non-Rigid Objects		
		Aircraft	Cars	CompCars	CUB	Dogs	DTD
Traditional	NoAug (Rao et al. 2021)	82.6	91.8	67.0	81.5	84.1	68.5
	FRAug (Rao et al. 2021)	83.2	92.1	69.4	82.0	84.2	69.2
	CALAug (Rao et al. 2021)	84.9	92.4	70.5	<u>82.5</u>	84.6	<u>69.7</u>
	RandAug (Cubuk et al. 2020)	83.7	92.6	72.5	81.5	84.2	69.3
	AutoAug (Cubuk et al. 2019)	84.7	92.9	73.2	82.1	84.6	68.7
	RandEarse (Zhong et al. 2020)	82.5	92.3	71.4	81.2	83.8	69.5
	CutMix (Yun et al. 2019)	83.0	91.7	67.3	81.8	84.5	69.2
Generative	Real Guidance (He et al. 2023)	83.2	92.0	72.8	82.2	83.1	68.2
	ALIA (Dunlap et al. 2023)	83.0	91.9	72.2	80.8	83.8	68.9
	DiffuseMix (Islam et al. 2024)	81.8	92.3	72.0	81.6	84.0	68.7
	DistDiff (Yang et al. 2024)	82.4	92.1	71.8	80.9	84.2	68.8
	SaSPA (Michaeli and Fried 2024)	<u>85.2</u>	<u>93.0</u>	<u>74.4</u>	81.7	<u>84.9</u>	<u>69.7</u>
	Ours	**86.1**	**93.6**	**75.2**	**82.7**	**85.8**	**70.8**

Table 1: Performance comparison of data augmentation methods using the CAL (ResNet50) classifier across rigid and non-rigid datasets. Results are reported as top-1 Accuracy (%). Best results are in **bold**, runner-up <u>underlined</u>, respectively.

Dataset	Methods	No Aug	FRAug	CAL-Aug	RandAug	AutoAug	RandEarse	CutMix
Aircraft	Baseline	82.6	83.2	84.9	83.7	84.7	82.5	83.0
	Ours	85.8 **(+3.2)**	86.1 **(+2.9)**	87.5 **(+2.6)**	86.6 **(+2.9)**	86.5 **(+1.8)**	85.4 **(+2.9)**	85.8 **(+2.8)**
Dogs	Baseline	84.1	84.2	84.6	84.2	84.6	83.8	84.5
	Ours	85.1 **(+1.0)**	85.8 **(+1.6)**	86.2 **(+1.6)**	85.7 **(+1.5)**	85.3 **(+0.7)**	84.8 **(+1.0)**	85.2 **(+0.7)**

Table 2: Comparison of traditional augmentation methods on FGVC Aircraft and Stanford Dogs. Our method demonstrates strong compatibility with conventional data augmentation techniques.

$p_\phi(y \mid \mathbf{x}'_t)$ is low), the guidance is amplified to steer generation more strongly toward the target class. Conversely, when classifier confidence is high, the guidance is attenuated to prevent overcorrection, allowing the base generation process to continue with minimal interference. This adaptive mechanism enhances stability and mitigates overfitting to classifier biases.

Experiments

Experiment Setup

Dataset. Six FGVC benchmarks are used to evaluate including FGVC Aircraft (Maji et al. 2013), CUB200-2011 (Reed et al. 2016), Stanford Cars (Krause et al. 2013), Stanford Dogs (Khosla et al. 2011), CompCars (Yang et al. 2015) and DTD (Cimpoi et al. 2014). The FGVC Aircraft, Stanford Cars, and CompCars datasets primarily evaluate the performance of our method on rigid objects, whereas the CUB200-2011, Stanford Dogs, and DTD datasets assess its effectiveness on non-rigid objects.

Comparison Methods. We compare our method against the following competitors:

- **Traditional Augmentation Methods:** FRAug(random horizontal flipping and random rotation), CALAug (Rao et al. 2021)(random flipping, random cropping, and color jittering), RandAug (Cubuk et al. 2020), AutoAug (Cubuk et al. 2019), RandEarse (Zhong et al. 2020), CutMix (Yun et al. 2019).
- **Diffusion Based Augmentation Methods:** Real Guidance (He et al. 2023), a method that uses a low translation strength to preserve fidelity to the original image; ALIA (Dunlap et al. 2023), a method that uses both the original dataset caption and a GPT-generated summary as textual guidance; DiffuseMix (Islam et al. 2024), an image-to-image method that leverages diffusion models to address category ambiguity introduced by mixup; DistDiff (Yang et al. 2024), a method that uses hierarchical prototypes to guide sample generation, ensuring alignment with the distribution of the original dataset; SaSPA (Michaeli and Fried 2024), a method designed for FGVC, utilizing ControlNet to preserve the structure and subject of the original images.

Fine-grained Visual Classification

Implementation Details. For all generative model-based methods, we generate two augmented images per original image. Across all datasets, the training set consists of 40% generated images and 60% original images, except for the CUB dataset, where the augmentation ratio is reduced to 20%. We add the random horizontal flipping and random rotation to all diffusion-based methods.

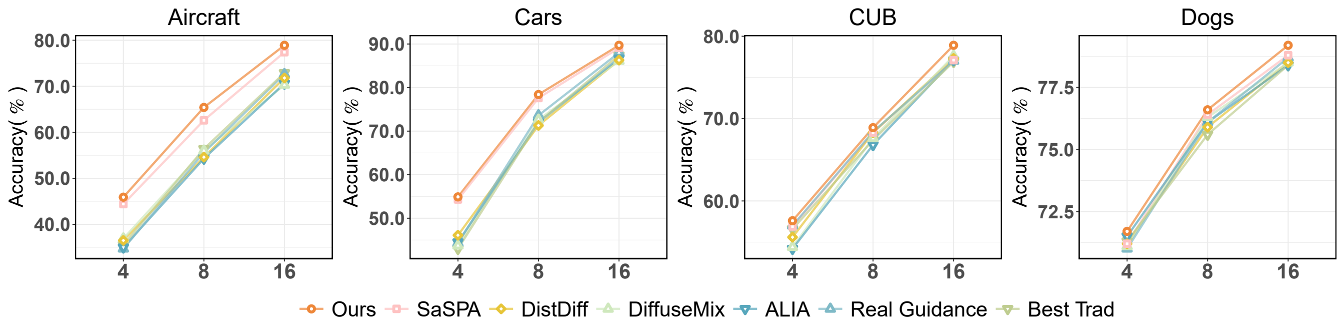


Figure 3: Performance of methods on the FGVC Aircraft, Stanford Cars, CUB, and Stanford Dogs under few-shot settings. The results indicate that our method remains effective even when the classifier is not very reliable.

Results. The results presented in Table 1 demonstrate the effectiveness of our proposed augmentation method across a diverse range of fine-grained visual benchmarks, encompassing both rigid and non-rigid object categories. **1)** First, our method consistently achieves state-of-the-art performance, securing the top accuracy score on all six datasets. This highlights the robustness and general applicability of our approach. **2)** Compared to using no augmentation, our method provides substantial gains, ranging from $+1.2\%$ on CUB to $+8.2\%$ on CompCars, indicating the significant benefit of the generated data. In addition, most data augmentation methods based on generative models are superior to traditional data augmentation methods. These results suggest that the generative approach introduces more beneficial variations than standard geometric and photometric transformations as shown in Figure 4. **3)** Notably, compared to SaSPA, which is specifically designed for FGVC and often the strongest generative competitor, our method shows considerable improvements. **4)** Table 2 presents the results of integrating our method with commonly used traditional data augmentation techniques. The performance improvements observed demonstrate that our approach significantly benefits from these combinations, further highlighting its robustness. This consistent advantage over other state-of-the-art generative methods underscores the effectiveness of our specific augmentation strategy.

Few-shot Learning

Implementation Details. To evaluate the performance of our method under limited data conditions, we conduct experiments in the few-shot learning setting. Specifically, we select four datasets, Aircraft, Cars, CUB, and Dogs, along with several representative comparison methods. And we consider 4-shot, 8-shot, and 16-shot scenarios. In this setting, the guidance classifier is replaced with a ResNet50 model **trained solely** on the available k-shot samples. Additionally, we increased the augmentation ratio to 0.6.

Results. As shown in Figure 3, our method consistently outperforms other baselines in few data setting. This demonstrates that even when the guidance classifier is unreliable, our dynamic adjustment mechanism effectively mitigates over-perturbation, ensuring that the generated images are not overly influenced by the classifier.

Text	Contour	Classifier	Aircraft	Cars	CUB	DTD
✓			83.8	91.9	78.9	63.5
✓	✓		85.3	92.5	80.2	66.6
✓	✓	✓	86.1	93.6	82.7	70.8

Table 3: Ablation study on hierarchical guidance.

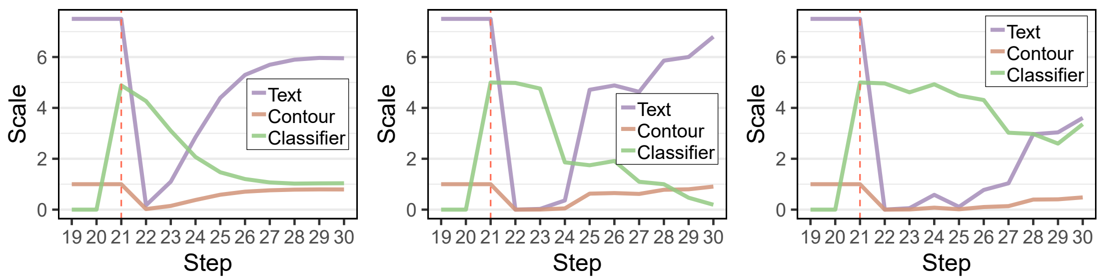
Ablation Study

Hierarchical Guidance. We select FGVC Aircraft, Stanford Cars, CUB, and DTD, to perform ablation experiments on hierarchical guidance. Table 3 shows that each component contributes incrementally to performance improvement. Notably, on non-rigid datasets, omitting classifier guidance results in performance degradation may below the baseline. This indicates that, for non-rigid categories, text and contour guidance alone are insufficient for the generative model to produce discriminative features. Overall, incorporating fine-grained classifier guidance yields substantial performance gains, highlighting its critical role in generating high-fidelity images for fine-grained visual classification tasks.

Classifier Guidance Scale and Step. We conduct ablation studies on the FGVC Aircraft dataset to investigate the effects of the classifier guidance scale and the classifier guidance start step. Figure 6a examines the impact of the classifier guidance scale s_{cls}, with the start step fixed at $N_s = 21$.

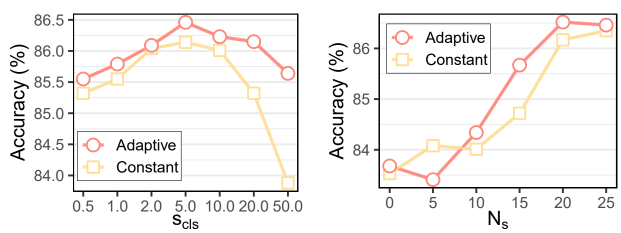
When s_{cls} is small, the influence of classifier guidance is limited, causing the diffusion model to rely more heavily on textual and contour cues. Conversely, when s_{cls} is too large, classifier guidance becomes overly dominant, which may result in partial image collapse and semantic distortion, ultimately degrading classifier performance. Figure 6b analyzes the effect of the classifier guidance start step N_s, with the scale fixed at $s_{\text{cls}} = 5$. Results indicate that applying classifier guidance during the later stages of the diffusion process yields superior outcomes. Introducing guidance too early, when background structures and object contours are still forming, can be harmful, as classifier predictions at this stage are often meaningless. Early intervention may disrupt both textual and contour guidance.

Figure 4: Qualitative comparisons between images generated by `HiGFA` and some baseline generative augmentation methods for fine-grained categories. These comparisons highlight `HiGFA`'s ability to generate diverse images while also better preserving subtle category-defining details. The complete results are provided in Appendix.



(a) Average evolution. (b) Easy sample. (c) Hard sample.

Figure 5: Evolution of the guidance scale during generation. The red dashed line indicates the activation of classifier guidance and the dynamic mechanism. Easy samples with clean backgrounds need only brief classifier signals to guide generation, while hard samples with complex backgrounds and conflicting prompts require extended guidance for generation, highlighting the sample-wise adaptability of our dynamic guidance.



(a) Classifier guidance scale. (b) Classifier guidance start step.

Figure 6: Ablation study on guidance scale, step and adaptive strategy. Enabling classifier guidance in the later stages of diffusion improves efficiency and enhances the robustness.

Adaptive Strategy. Figure 6 present an ablation study evaluating the dynamic strategy. Without this strategy, increasing the guidance scale significantly degrades classification performance. Incorporating it reduces sensitivity to the hyperparameter s_{cls} and improves the performance and stability of `HiGFA`. Figure 5a shows the average evolution of the three guidance during inference on the FGVC Aircraft. At the onset of classifier guidance, both text and contour guidance sharply decline. As inference process continues, all three guidance gradually stabilize. Figure 5b and Figure 5c illustrate the evolution of the guidance scale for an easy and a hard sample, respectively. For the easy sample, images with clean backgrounds and the main subject minimally affected by the surroundings, a brief but strong classifier signal suffices to guide generation toward the target class, after which text and contour guidance dominate. In contrast, the hard sample, which feature complex backgrounds and conflicts between text prompt and the main subject, requires prolonged classifier guidance to progressively align the output with the target distribution.

Conclusion

This paper introduces Hierarchically Guided Fine-grained Augmentation (`HiGFA`) for generating images for fine-grained data augmentation with diffusion models. Existing guidance methods often fail to capture subtle, category-defining features, limiting their augmentation effectiveness. `HiGFA` solves this problem by combining multiple guidance sources: text, diversity-enhanced contours, and a classifier. It adjusts the strength of each source over time to match the diffusion model's coarse-to-fine image generation process. Early stages focus on global structure and style using text and contour guidance, while later stages employ an adaptive, confidence-modulated fine-grained classifier guidance to precisely enforce category-specific features on a cleaner image estimate. Experiments on several FGVC datasets demonstrate the effectiveness of `HiGFA`.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China: 62525212, 62236008, 62441232, U21B2038, and U23B2051, in part by Youth Innovation Promotion Association CAS, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDB0680201, in part by the China National Postdoctoral Program for Innovative Talents, Grant No. BX20250377.

References

Benita, R.; Elad, M.; and Keshet, J. 2025. Designing Scheduling for Diffusion Models via Spectral Analysis. *arXiv preprint arXiv:2502.00180*.

Bookstein, F. L. 1989. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6): 567–585.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3606–3613.

Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. AutoAugment: Learning Augmentation Policies from Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 702–703.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. In *Annual Conference on Neural Information Processing Systems*, volume 34, 8780–8794.

Dunlap, L.; Umino, A.; Zhang, H.; Yang, J.; Gonzalez, J.; and Darrell, T. 2023. Diversify Your Vision Datasets with Automatic Diffusion-Based Augmentation. In *Annual Conference on Neural Information Processing Systems*.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Annual Conference on Neural Information Processing Systems*, volume 27.

Han, B.; Xu, Q.; Bao, S.; Yang, Z.; Zi, K.; and Huang, Q. 2025. LightFair: Towards an Efficient Alternative for Fair T2I Diffusion via Debiasing Pre-trained Text Encoders.

Han, B.; Xu, Q.; Yang, Z.; Bao, S.; Wen, P.; Jiang, Y.; and Huang, Q. 2024. Aucseg: Auc-oriented Pixel-level Long-tail Semantic Segmentation. *Advances in Neural Information Processing Systems*, 37: 126863–126907.

He, R.; Sun, S.; Yu, X.; Xue, C.; Zhang, W.; Torr, P.; Bai, S.; and Qi, X. 2023. Is Synthetic Data from Generative Models Ready for Image Recognition? In *International Conference on Learning Representations*.

Hemmat, R. A.; Pezeshki, M.; Bordes, F.; Drozdzal, M.; and Romero-Soriano, A. 2023. Feedback-guided data synthesis for imbalanced classification. *arXiv preprint arXiv:2310.00158*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Annual Conference on Neural Information Processing Systems*, volume 33, 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598*.

Hua, C.; Xu, Q.; Bao, S.; Yang, Z.; and Huang, Q. 2024. ReconBoost: Boosting Can Achieve Modality Reconcilement. In *International Conference on Machine Learning*, 19573–19597.

Hua, C.; Xu, Q.; Yang, Z.; Wang, Z.; Bao, S.; and Huang, Q. 2025. OpenworldAUC: Towards Unified Evaluation and Optimization for Open-world Prompt Tuning. In *International Conference on Machine Learning*.

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.

Jiang, Y.; Hua, C.; Feng, Y.; and Gao, Y. 2023. Hierarchical set-to-set representation for 3-d cross-modal retrieval. *IEEE Transactions on Neural Networks and Learning Systems*.

Jobling, J. A. 2010. *The Helm Dictionary of Scientific Bird Names: From Aalge to Zusii*. London: Christopher Helm.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Li, F.-F. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, volume 2.

Koulischer, F.; Handke, F.; Deleu, J.; Demeester, T.; and Ambrogioni, L. 2025. Feedback guidance of diffusion models. *arXiv preprint arXiv:2506.06085*.

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.

Li, M.; and Chen, S. 2024. Critical windows: non-asymptotic theory for feature emergence in diffusion models. In *International Conference on Machine Learning*, 27474–27498. PMLR.

Lipman, Y.; Chen, R. T. Q.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2023. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*.

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

Mariani, G.; Scheidegger, F.; Istrate, R.; Bekas, C.; and Malossi, C. 2018. BAGAN: Data Augmentation with Balancing GAN. In *International Conference on Machine Learning*.

Michaeli, E.; and Fried, O. 2024. Advancing Fine-Grained Classification by Structure and Subject Preserving Augmentation. In *Annual Conference on Neural Information Processing Systems*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

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.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Rao, Y.; Chen, G.; Lu, J.; and Zhou, J. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *International Conference on Computer Vision*, 1025–1034.

Raya, G.; and Ambrogioni, L. 2023. Spontaneous symmetry breaking in generative diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 66377–66389.

Reed, S.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 49–58.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Shama Sastry, C.; Dumpala, S. H.; and Oore, S. 2024. DiffAug: A Diffuse-and-Denoise Augmentation for Training Robust Classifiers. In *Annual Conference on Neural Information Processing Systems*, volume 37, 20745–20785.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Trabucco, B.; Doherty, K.; Gurinas, M.; and Salakhutdinov, R. 2024. Effective Data Augmentation with Diffusion Models. In *International Conference on Learning Representations*.

Yang, L.; Luo, P.; Change Loy, C.; and Tang, X. 2015. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3973–3981.

Yang, L.; Yong, J.-H.; Yin, H.; Jiang, J.; Xiao, M.; Zhang, W.; Wang, B.; et al. 2024. Distribution-Aware Data Expansion with Diffusion Models. In *Annual Conference on Neural Information Processing Systems*.

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 *International Conference on Computer Vision*, 6023–6032.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision*, 3836–3847.

Zhang, Y.; Ling, H.; Gao, J.; Yin, K.; Lafleche, J.-F.; Barriuso, A.; Torralba, A.; and Fidler, S. 2021. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10145–10155.

Zhang, Y.; Zhou, D.; Hooi, B.; Wang, K.; and Feng, J. 2023. Expanding small-scale datasets with guided imagination. In *Annual Conference on Neural Information Processing Systems*, volume 36, 76558–76618.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13001–13008.