GENDATAAGENT: ON-THE-FLY DATASET AUGMENTATION WITH SYNTHETIC DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic data is increasingly employed for training dataset augmentation in computer vision. However, prior works typically perform a uniform search across the entire category space, overlooking the interaction between synthetic data generation and downstream task training. Furthermore, balancing the diversity of synthetic data while ensuring it remains within the same distribution as real data (i.e., avoiding outliers) remains a significant challenge. In this work, we propose a generative agent to augment target training datasets with synthetic data for model fine-tuning. Our agent iteratively generates relevant data on-the-fly, aligning with the target training dataset distribution. It prioritizes sampling diverse synthetic data that complements marginal training samples, with a focus on synthetic data that exhibit higher variance in gradient updates. Evaluations across diverse supervised image classification tasks demonstrate the effectiveness of our approach.

1 INTRODUCTION

Generative models (Brock et al., 2018; Razavi et al., 2019; Ho et al., 2020; Saharia et al., 2022; Rombach et al., 2022; Sohl-Dickstein et al., 2015; Ramesh et al., 2022; Nichol et al., 2022) that produce photo-realistic images from text prompts are increasingly used to replace (Besnier et al., 2020; Li et al., 2022; Sarıyıldız et al., 2023; Hammoud et al., 2024; He et al., 2022; Shipard et al., 2023; Tian et al., 2024) or augment (Yuan et al., 2023; Azizi et al., 2023; He et al., 2022; Hemmat et al., 2023; Bansal & Grover, 2023; Dunlap et al., 2024; Astolfi et al., 2023) real data. This shift is largely motivated by the significant time and labor costs associated with collecting and annotating real data (Tian et al., 2024; Yuan et al., 2023; Sarıyıldız et al., 2023; Besnier et al., 2020; Dunlap et al., 2024). However, most approaches provide no feedback to the synthetic data generation process during downstream model training, potentially reducing sample utility (Hemmat et al., 2023). Discrepancies between synthetic and target data distributions (Shin et al., 2023; He et al., 2022; Borji, 2022), alongside limited diversity in generated samples (Hall et al., 2023; Bianchi et al., 2023; Luccioni et al., 2024; Jahanian et al., 2021), can further undermine synthetic data's utility, despite attempts to address these issues through techniques such as prompt engineering (Sarıyıldız et al., 2023; Lei et al., 2023; Azizi et al., 2023), image-conditioned generation (Bordes et al., 2022; Blattmann et al., 2022), diffusion inversion (Zhou et al., 2023; Zhao & Bilen, 2022), and low-density region sampling (Um et al., 2024; Sehwag et al., 2022).

In this work, we propose an approach to training dataset augmentation that addresses these challenges by leveraging *GenDataAgent*, a generative agent. Our method not only samples high-quality synthetic data on-the-fly but also ensures that the generated samples align closely with the distribution of the target training dataset. By prioritizing the sampling of diverse and useful synthetic data that complement marginal real training examples through feedback mechanisms, our approach enhances the generalization performance of downstream models fine-tuned on augmented datasets. Crucially, unlike prior research (Hemmat et al., 2023; Shao et al., 2024; Ye-Bin et al., 2023; Liu et al., 2020; Kozerawski et al., 2020), we do not rely on distributional assumptions, such as long-tailed data or datasets with few to no examples per class. Our key contributions can be summarized as follows:

- We propose GenDataAgent, an on-the-fly framework for augmenting image classification datasets with synthetic data generated via Stable Diffusion (Rombach et al., 2022), ensuring alignment with the target distribution.

- We introduce a sampling algorithm focused on marginal real examples, a variance of gradients (Agarwal et al., 2022) filtering strategy to remove synthetic outliers, and Llama-2 (Touvron et al., 2023) to enhance diversity by perturbing text prompts.
- We demonstrate state-of-the-art generalization performance, increased fairness, and GenDataAgent's ability to complement real-world samples during training.

2 Related Work

2.1 Synthetic Data Generation

Interest in generative models for training data generation has grown, with diffusion-based methods (Bansal & Grover, 2023; He et al., 2022; Shipard et al., 2023; Trabucco et al., 2023; Besnier et al., 2020; Sarıyıldız et al., 2023; Hammoud et al., 2024; Tian et al., 2024; Yuan et al., 2023; Azizi et al., 2023; Hemmat et al., 2023; Dunlap et al., 2024; Astolfi et al., 2023) increasingly supplanting generative adversarial network-based methods (Zhao & Bilen, 2022; Li et al., 2022; Zhang et al., 2021; Kumar et al., 2022; Sharmanska et al., 2020). While text-guided diffusion-generated image data has shown promise, models trained on this data have had varied success due to a persistent distributional gap (Shin et al., 2023; He et al., 2022; Borji, 2022; Hemmat et al., 2023) and lack of diversity (Hall et al., 2023; Bianchi et al., 2023; Luccioni et al., 2024).

To address this, some research has focused on sampling from low-density regions (Um et al., 2024; Sehwag et al., 2022; Samuel et al., 2023), which contain attributes seldom observed in high-density regions. These approaches rely on assumptions about data distributions, such as long-tailed data or datasets with few examples per class, and do not consider the utility of the generated data. Unlike these approaches, we rely on feedback from the downstream classifier, using uncalibrated Marginal scores to generate synthetic data that complements marginal real examples.

2.2 Synthetic Data Augmentation

In parallel, prompt engineering with text-conditioned diffusion models has been proposed for classification (Sarıyıldız et al., 2023; Lei et al., 2023; Azizi et al., 2023). Sarıyıldız et al. (2023) employ *manual*, class-agnostic prompt engineering to reduce semantic issues (e.g., polysemy) and increase diversity, while Lei et al. (2023) leverage *automated* image captioning models. However, neither approach explicitly addresses the inclusion of synthetic data inconsistent with the target distribution or the data's usefulness. As with prior work (Yuan et al., 2023; Sarıyıldız et al., 2023; Lei et al., 2023; He et al., 2022), these methods result in static, bloated datasets containing redundant and uninformative samples, as no information is transmitted from the downstream model into the generation process. Hemmat et al. (2023) propose feedback-guided data synthesis but only consider a single offline feedback cycle. In contrast, we introduce on-the-fly filtering during downstream model training, prioritizing difficult, synthetic data with higher variance in gradient updates (Agarwal et al., 2022) to avoid noisy, unrepresentative samples (He et al., 2022; Hemmat et al., 2023; Shin et al., 2023). Furthermore, our method can be integrated with techniques such as domain adaptation (Tang & Jia, 2023) to further narrow the distribution gap.

Another line of work (Li et al., 2023a; Jiang et al., 2021) utilizes additional real data for non-static training dataset augmentation. For example, Li et al. (2023a) continuously explore the Internet to find relevant data, ranking retrieved images based on their expected *reward*. While internet-sourced data may offer greater diversity, it raises privacy and copyright concerns (Samuelson, 2023; Andrews et al., 2024; Longpre et al., 2023; Besnier et al., 2020; Metcalf & Crawford, 2016; Orekondy et al., 2018; Birhane & Prabhu, 2021; Birhane et al., 2021).

3 GenDataAgent: On-the-fly Dataset Augmentation

We tackle the problem of effectively augmenting vision training datasets with synthetic data by introducing GenDataAgent, a generative agent that augments datasets on-the-fly during model fine-tuning. GenDataAgent generates synthetic data aligned with the target dataset distribution to improve performance in supervised image classification. It prioritizes diverse and useful synthetic data to complement marginal training samples, focusing on those with higher gradient update variance. This approach improves model generalization and fairness while reducing computational and energy costs. An overview of GenDataAgent is detailed in Algorithm 1.

Algorithm 1 `GenDataAgent`

Input: target dataset $\mathcal{T} = \{(x_i, y_i, p_i, c_i)\}_{i=1}^{N}$, pretrained multi-class classfication model f, stable diffusion model `SD`, image feature extractor `CLIP`, image captioning model `BLIP-2`, large language enhance model `Llama-2`
Generate image features Ψ for real data $\mathcal{T}$ by `CLIP`
Adapt `SD` to target distribution with $\mathcal{T}$, $\mathcal{P} = \{p_i\}_{i=1}^{N}$, $\mathcal{C}$ and Ψ (§3.1)
for iter $= 1, 2, \ldots$ **do**
  if iter $<= 3$ **then** // Stage-1
    Fine-tune f **only** on $\mathcal{T}$, save model checkpoints $\{f_i\}_{i=1}^{3}$ (§3.4)
  else // Stage-2
    Sample feedback $\mathcal{M} \subset \mathcal{T}$ of marginal examples (§3.2)
    Perturb image captions for marginal examples $\mathcal{C}'_{\mathcal{M}} \leftarrow$ `Llama-2`$(\mathcal{P}_{\mathcal{M}}, \mathcal{C}_{\mathcal{M}})$ (§3.3)
    Generate diverse synthetic data $\mathcal{S}_c \leftarrow$ `GenData`$^{m}_{x \in \mathcal{M}}(x, \mathcal{C}'_x = \{c'_{x,i}\}_{i=1}^{m})$
    for $j = 1, \ldots, M$ **do** // traverse all categories
      Compute VoG score for each synthetic data $x_i \in \mathcal{S}_c$
      Filter out images of j-th class $\mathcal{S}_{f,j} \leftarrow \underset{x_i \in \mathcal{S}_{f,j} \subset \mathcal{S}_c}{\arg\min} \sum_i \text{VoG}_i$ (§3.4)
    end for
    Combine real and synthetic data as training data $\mathcal{D} \leftarrow \mathcal{T} \cup (\mathcal{S}_c \backslash \bigcup_j \mathcal{S}_{f,j})$
    Fine-tune f on the combined dataset $f_{\theta_{\text{iter}+1}} \leftarrow f_{\theta_{\text{iter}}}(\mathcal{D})$
  end if
end for

3.1 Text-to-Image Generator

Drawing inspiration from text-guided diffusion models, we utilize Stable Diffusion (SD) as a text-to-image generator. To reconcile the distribution gap between SD training data and the target real data, we adapt SD to match the distribution of the target training data, as proposed by Yuan et al. (2023).

Suppose $\mathcal{T} = \{(x_i, y_i, p_i, c_i)\}_{i=1}^{n}$ denotes a target training dataset, with x_i representing a real image, y_i its numerical class label, p_i its semantic class name, c_i its BLIP2-generated (Li et al., 2023b) image caption, and μ_{y_i} the mean vector of image features extracted with CLIP (Radford et al., 2021) for class y_i. Yuan et al. (2023) concatenate semantic class names, p_i, and image captions, c_i, as text prompts, along with ψ_{y_i} to include visual guidance, estimating intra-class feature distributions. The resulting multi-modal condition comprises "`a photo of` p_i`, which is` c_i`,` ψ_{y_i}", which we use to fine-tune SD with LoRA (Hu et al., 2021).

For text-to-image generation, we prioritize sample quality over diversity, employing a prompt guidance value of 7.5, in contrast to prior methods (Sarıyıldız et al., 2023; Yuan et al., 2023) utilizing a value of 2. This choice is driven by our improved approach to introducing sample diversity without compromising quality, as detailed in Appendix A.

3.2 Generator Feedback via Marginal Samples

Previous studies (Yuan et al., 2023; Sarıyıldız et al., 2023) primarily focus on generating synthetic data aligned with the target distribution but overlook its *utility*. The effectiveness of this indiscriminate approach for dataset augmentation is therefore questionable. Similar to sample reweighting techniques (Freund & Schapire, 1995; Johnson & Khoshgoftaar, 2019; Ren et al., 2018), we advocate assigning higher importance to synthetic data that complement real samples near the model's decision boundary. This strategy exposes the model to critical regions of the data space. While prioritizing challenging samples may initially increase losses, it can potentially improve generalization.

To implement this, we adopt a marginal score, which is defined as the target model's predicted probability for class y_i given input x_i. Specifically, the marginal score is calculated as $P(y = y_i \mid x_i) = \exp(z^i_{y_i}) / \sum_j \exp(z^i_j)$, where z_j denotes the logit for class j. We rank each sample in the target dataset by this score and retain the top-k samples with the lowest scores as *feedback* for guiding synthetic data generation. The set of these k marginal samples is denoted as $\mathcal{M}$.

Figure 1 illustrates clear clustering and boundaries in the pretrained (i.e., prior to fine-tuning) target model's feature space of instances with high marginal scores, contrasting with the more challenging

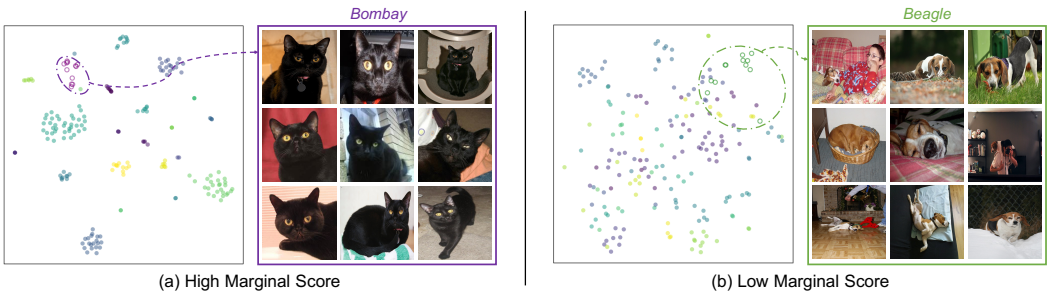

Figure 1: Feature space visualization and qualitative comparison of instances with Top-200 and Bottom-200 marginal scores on Oxford-IIIT Pets dataset. Each color represents a specific class.

separation of samples with low marginal scores. High-scoring images typically feature centrally positioned representative depictions of objects, while low-scoring ones exhibit greater diversity and noise. Moreover, empirical results in Section 4.3 validate the utilization of predicted probabilities over entropy-based selection (Hemmat et al., 2023).

3.3 Enhancing Diversity via Caption Perturbations

Images generated with identical captions but different seeds often exhibit limited diversity, even when using a high guidance value (Figure 2). This redundancy poses a challenge to leveraging synthetic data effectively (Hall et al., 2023; Bianchi et al., 2023). To enhance diversity, we utilize Llama2 (Touvron et al., 2023), a large language model, to modify the image captions. However, since Llama2 tends to produce lengthy sentences, which may result in excessive outliers in synthetic data, we constrain its responses to short sentences by imposing a word count limit. For this purpose, we provide the following tailored prompts to Llama2:

> `"role" : "system", "content": You are an editor tasked with subtly altering the scene described after a comma in a sentence. The goal is to change the scene slightly in no more than 10 words. Respond with` m `versions.`

> `"role" : "user", "content": Given the sentence "a photo of` p_i`, which is` c_i`", slightly alter the scene described after the comma to depict a similar yet different scenario.`

Examples are shown in Figure 2 and the Appendix A. When adapting the stable diffusion to target distribution (Section 3.1), we utilize the same full format (i.e., the combination of classname $\mathcal{P}$ and raw image caption $\mathcal{C}$). By maintaining this consistent prompt format for both adapting stable diffusion and generating synthetic data, we ensure greater alignment between the distribution of the generated data and the target data.

Given the feedback $\mathcal{M}$ that contains marginal examples from the classification model, we first generate m perturbation for the raw caption of each $x \in \mathcal{M}$ by Llama-2:

$$\mathcal{C}'_x = \texttt{CaptionPerturbation}^m(p_x, c_x), \tag{1}$$

where $\mathcal{C}' = \{\mathcal{C}'_x | x \in \mathcal{M}\}$ is the whole caption perturbation set. After obtaining the perturbation, the generating process of GenDataAgent can be formulated as:

$$\mathcal{S}_c = \texttt{GenData}^m_{x \in \mathcal{M}}(x, \mathcal{C}'_x = \{c'_{x,i}\}^m_{i=1}), \tag{2}$$

where the order m denotes m different perturbation versions of the raw image caption, and the size of marginal synthetic dataset $|\mathcal{S}_c| = k \times m$ can be restricted to $1 \times |\mathcal{T}|$, $10 \times |\mathcal{T}|$ or any ratio if needed.

The visual comparison is shown in Figure 2, where we can easily notice that the Llama-2 perturbation here would not totally change the background scene and result in a totally different synthetic image that has nothing to do with the original real image. Yet, the Abyssinian cat's posture and the image's view are slightly different. This ensures that the perturbation is still focusing on marginal examples

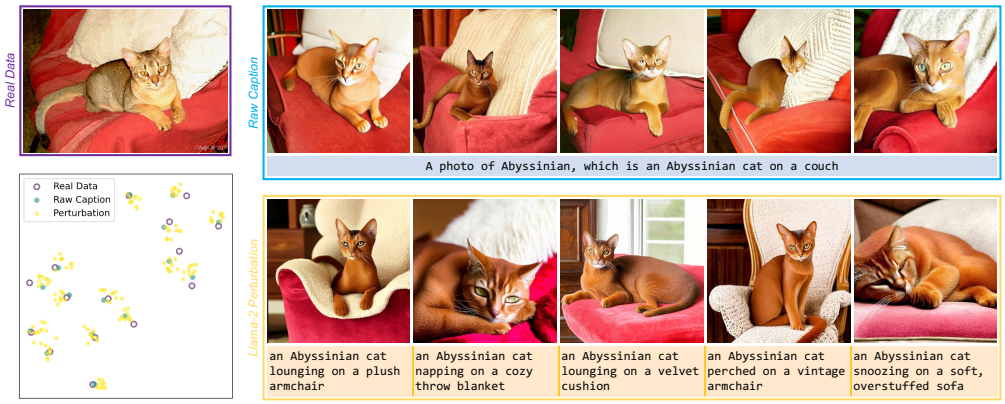

Figure 2: Comparison between real data, synthetic data with identical captions, and Llama-2 perturbation captions on the Oxford-IIIT Pets dataset. t-SNE visualization of feature space is in the lower left corner, where each cluster is related to one class. For simplicity, the perturbed captions omit the prefix "A photo of [*classname*]". All synthetic data is generated with prompt guidance of 7.5.

rather than providing general-purpose augmentations. When turning the perspective to the feature space in Figure 2, it can be easily observed that the synthetic data with raw caption is close to the real data while the synthetic data with Llama-2 perturbation is distributed in a broader region.

3.4 In-Distribution Data Generation

The synthesized data generated by the agent is likely to include out-of-distribution samples (Fig. 3c), which are excessively noisy and disruptive, leading to a performance drop in the target task. Therefore, it is important to control the agent to generate only data that falls within the distribution of the target task. Inspired by (Li et al., 2023a; Robinson et al., 2020; Ge, 2018; Schroff et al., 2015; Agarwal et al., 2022), "harder negatives" samples yield larger gradients. We utilize the variance of the gradient (VoG) for each synthesized data as a criterion to filter out outliers.

Specifically, the classification model fine-tuning operates in two stages. At the first stage, i.e., the first N ($N = 3$) iterations (1 iteration contains 10 epochs), we fine-tune the model with the target dataset $\mathcal{T}$ and save the checkpoint for each iteration. Then during the second stage ($N > 3$), the agent receives the feedback list from the model, as well as the model fine-tune checkpoints from the first N iterations. After generating the synthetic data $\mathcal{S}_c$, the agent computers the gradients of logit z_{y_i} with respect to each pixel of x_i on the synthetic data $\mathcal{S}_c$: $G_i = \frac{\partial z^i_{y_i}}{\partial x_{i,d}}$, where $d = \{1, 2, \dots W\}$, W is the total number of pixels in image x_i, and $(x_i, y_i) \in \mathcal{S}_c$. In Table 5, we conduct experiments using 3, 4, and 5 checkpoints to compute the variance of gradients, which is the so-called VoG score. The experimental results show no significant difference between the different checkpoint numbers. Thus, to save resources, we adopt the minimal requirement, i.e., 3 checkpoints to measure the VoG:

$$\mu_i = \frac{1}{3}(G_i^{10} + G_i^{20} + G_i^{30}), \tag{3}$$

$$\text{VoG}_i = \sqrt{\frac{1}{3}[(G_i^{10} - \mu_i)^2 + (G_i^{20} - \mu_i)^2 + (G_i^{30} - \mu_i)^2]}, \tag{4}$$

where G_i^{10}, G_i^{20}, and G_i^{30} denote the gradients of epochs 10, 20, and 30 respectively.

The model itself can quickly learn the data distribution by first assigning large gradients to in-distribution data and then rapidly decreasing. In other words, in-distribution data tends to own higher variances of gradients. Hence, during the on-the-fly process (fine-tuning stage-2), GenDataAgent rejects the out-of-distribution data with low variances of gradients $\mathcal{S}_{f,j}$ within each class j from the raw synthetic dataset $\mathcal{S}_c$ as:

$$\mathcal{S}_{f,j} = \underset{x_i \in \mathcal{S}_{f,j} \subset \mathcal{S}_c}{\arg\min} \sum_i \text{VoG}_i \quad s.t. \quad |\mathcal{S}_{f,j}| = o_j. \tag{5}$$

As shown in Figure 3 and Appendix B, stable diffusion, even after adapting to the target dataset distribution, might introduce outliers into the synthetic data. Fortunately, the applied VoG filtering

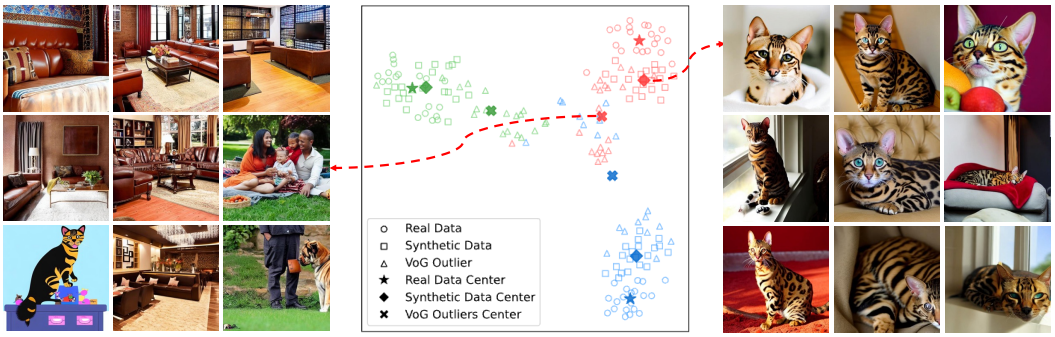

Figure 3: t-SNE visualization of VoG filtering strategy on Oxford-IIIT Pets dataset. For each class, 20 data points are randomly sampled from **Real Data, Synthetic Data w/o outliers, and Bottom 25% VoG Scores** respectively.

strategy can help to remove them by assigning the lowest variances of gradients. The visualization of feature space in Figure 3 provides a more intuitive way to show the distance. Specifically, the center of synthetic data that removes VoG outliers is close to that of the real data. In contrast, the center of VoG outliers is far from that of the real one and gets entangled with other outliers.

Although the caption perturbation helps to increase the diversity of synthetic data, such perturbations might lead to generating more outliers. Yet, our VoG filtering strategy can eliminate most of these outliers, thereby achieving an improved trade-off between diversity and in-distribution.

3.5 On-the-fly Feedback and Fine-tuning

Prior work (Yuan et al., 2023; Sarıyıldız et al., 2023; Lei et al., 2023; He et al., 2022) usually combines the synthetic data with real data in a static way, which ignores the feedback from the classification model to the generation process. LDM-FG (Hemmat et al., 2023) introduces feedback from the classifier but only considers *one* offline feedback cycle. Rather than augmenting with synthetic data statically, our marginal-focused approach is conducted in an on-the-fly manner. As shown in Algorithm 1, during each iteration at stage-2, we resample marginal examples based on the current stage as the feedback. Then send this feedback to GenDataAgent, which will guide the generation of synthetic data. After receiving the generative synthetic data from GenDataAgent, the classification model will be fine-tuned on the combination of real and newly synthetic data as

$$f_{\theta_{\text{iter}+1}} \leftarrow f_{\theta_{\text{iter}}}(\mathcal{T} \cup (\mathcal{S}_c \backslash \bigcup_j \mathcal{S}_{f,j})), \tag{6}$$

where $f_{\theta_{\text{iter}}}$ is the model from the previous iteration. In the next on-the-fly iteration, $f_{\theta_{\text{iter}+1}}$ is used to resample marginal examples and send an updated feedback list to GenDataAgent.

4 Experiments

4.1 Experimental Settings

Previous studies (Yuan et al., 2023; Sarıyıldız et al., 2023; He et al., 2022) that explored the utilization of synthetic data mainly conducted under a supervised learning manner. For a fair comparison, we conduct experiments in this manner with the same multi-class classification model ResNet-50 in two different scenarios: **(i)** Synthetic data alone for image classification, **(ii)** Synthetic data serves as augmentation for real data. For the second scenario, we also conduct experiments on CLIP ResNet-50 and CLIP ViT backbones. More details are shown in Appendix D.

Synthetic Data Only. We evaluate the effectiveness of using synthetic data as a substitute for real training data in the image classification model fine-tuning stage. Note that our GenDataAgent requires real dataset $\mathcal{T}$ for providing feedback. To solve this, we replace the initial real dataset $\mathcal{T}$ with an equivalent number of synthetic training samples and use GenDataAgent to inject augmented generative data iteratively. We compare our method with Real-Fake (Yuan et al., 2023), ImageNet-Clone (Sarıyıldız et al., 2023), CiP (Lei et al., 2023), and Synthetic Data (He et al., 2022).

Table 1: **Top-1 accuracy / worst-case disparity for image classification tasks with only synthetic data**. †The framework of SyntheticData is CLIP but the backbone is the same ResNet-50. ‡We reproduce ImageNet-Clone, CiP, and Real-Fake for all datasets.

Model	Pets	CUB	Flowers	Birdsnap	Food	IN100
*Training with **only synthetic** data*						
ImageNet-Clone‡ (Sarıyıldız et al., 2023)	79.7/0.20	29.4/0.00	27.2/0.00	24.1/0.00	55.6/0.10	64.2/0.20
CiP‡ (Lei et al., 2023)	86.5/0.16	35.1/0.00	25.8/0.00	30.7/0.00	55.5/0.15	64.7/0.20
SyntheticData† (He et al., 2022)	86.8/–	56.9/–	67.1/–	38.1/–	80.4/–	–/–
Real-Fake‡ (Yuan et al., 2023)	89.5/0.40	66.0/0.00	66.9/0.00	54.2/0.00	80.1/**0.42**	82.8/**0.40**
GenDataAgent (Ours)	**90.3**/**0.44**	**71.4**/0.00	**76.9**/**0.11**	**54.7**/0.00	**81.2**/**0.42**	**87.2**/**0.40**

Table 2: **Top-1 accuracy / worst-case disparity for image classification tasks with synthetic data augmentation**. ‡We reproduce the Real-Fake for all datasets. [1]

Model	Pets	CUB	Flowers	Birdsnap	Food	IN100
ResNet-50 backbone						
only real	93.6/0.40	83.1/0.13	87.4/0.40	73.0/0.00	86.8/0.63	87.4/0.20
Real-Fake‡ (Yuan et al., 2023)	94.2/0.48	83.1/0.00	89.0/0.50	73.0/0.00	87.4/0.61	88.6/**0.40**
Internet Explorer (Li et al., 2023a)	94.5/0.52	83.6/**0.25**	90.2/**0.56**	73.9/0.00	87.3/0.63	88.9/**0.40**
GenDataAgent (Ours)	**94.7**/**0.56**	**83.9**/**0.25**	**91.0**/**0.56**	**74.5**/0.00	**87.8**/**0.64**	**90.1**/**0.40**
Δ with only real data	+1.1/0.16	+0.8/0.12	+3.6/0.16	+1.5/0.00	+1.0/0.01	+2.7/0.20
CLIP ResNet-50 backbone						
only real	77.8/0.28	66.3/0.00	69.0/0.24	64.6/0.00	82.2/0.47	87.0/0.20
Real-Fake‡ (Yuan et al., 2023)	80.7/**0.32**	66.9/0.00	71.0/0.24	65.7/0.00	86.1/0.61	88.0/**0.40**
Internet Explorer (Li et al., 2023a)	81.3/**0.32**	67.7/0.00	72.2/0.24	66.2/0.00	86.3/0.61	88.4/**0.40**
GenDataAgent (Ours)	**82.0**/**0.32**	**68.2**/0.00	**72.8**/**0.30**	**66.7**/0.00	**86.5**/**0.63**	**89.1**/**0.40**
Δ with only real data	+4.2/0.04	+1.9/0.00	+3.8/0.06	+2.1/0.00	+4.3/0.16	+2.1/0.20
CLIP ViT backbone						
only real	92.1/0.40	80.5/0.00	86.5/0.33	65.8/0.00	54.4/0.15	63.2/0.20
Real-Fake‡ (Yuan et al., 2023)	92.8/0.40	80.7/0.00	94.9/0.60	67.8/0.00	63.7/0.20	64.9/0.20
Internet Explorer (Li et al., 2023a)	92.9/0.40	81.8/0.00	95.5/0.67	68.4/0.00	65.1/0.24	65.9/0.20
GenDataAgent (Ours)	**93.3**/**0.48**	**82.6**/**0.13**	**96.1**/**0.78**	**69.6**/0.00	**67.0**/**0.26**	**66.3**/0.20
Δ with only real data	+1.2/0.08	+2.1/0.13	+9.6/0.45	+3.8/0.00	+12.6/0.11	+3.1/0.00

Synthetic Data Augmentation. We compare our GenDataAgent with the SOTA method Real-Fake (Yuan et al., 2023), and Internet Explorer (Li et al., 2023a). For a fair comparison with Internet Explorer, we keep all aspects of our method, such as distribution adaptation and Llama-2 Perturbation, unchanged, and only replace Marginal Sampling and VoG Filtering with the 15-NN similarity.

Datasets. GenDataAgent is evaluated on the general ImageNet-100 (IN100) dataset (Tian et al., 2020) and 5 popular fine-grained datasets: Oxford-IIIT Pets (Parkhi et al., 2012), Flowers-102 (Nilsback & Zisserman, 2008), Birdsnap (Berg et al., 2014), CUB-200-2011 (Wah et al., 2011), and Food-101 (Bossard et al., 2014). Following previous work, we adopt backbone models pre-trained on ImageNet for all datasets except the ImageNet-100. Similar to Real-Fake (Yuan et al., 2023), for the ImageNet-100 dataset, we train the classifier models from scratch.

Evaluation Matrics. Following previous work (Yuan et al., 2023; Sarıyıldız et al., 2023), we evaluate our GenDataAgent by Top-1 accuracy across classes. Moreover, we utilize the worst-case disparity (min-max accuracy ratio) (Ghosh et al., 2021), to evaluate the fairness. Since our sampling strategy targets more marginal examples, we anticipate that our GenDataAgent could enhance the fairness of the classifier model, a criterion that holds greater significance in real-world applications.

4.2 Quantitative Result

Synthetic Data Only. Table 1 presents the Top-1 classification accuracy and worst-case disparity of various methods across diverse downstream datasets under ***synthetic-only*** setup. As shown, GenDataAgent substantially outperforms all other methods in classification accuracy, demonstrating that our on-the-fly generation approach effectively operates without requiring a real dataset to initialize the feedback mechanism. Notably, GenDataAgent achieves performance levels comparable to models trained exclusively on real data, particularly on the IN100 dataset.

Synthetic Data Augmentation. Table 2 present the results under ***real + synthetic*** setup with different pretrained backbones. GenDataAgent improves classification accuracy and worst-case

[1]Please refer to Appendix C for the mean and deviation of multiple runs.

Table 3: **Break-down ablation of Marginal score Sampling, Llama-2 Perturbation, and VoG Filtering**. The numbers are Top-1 accuracy / worst-case disparity.

Marginal Sampling Section 3.2	Perturbation Section 3.3	VoG Filtering Section 3.4	Pets	CUB	Flowers	Birdsnap	Food	IN100
-	-	-	93.6 / 0.40	83.1 / 0.13	87.4 / 0.40	73.0 / 0.00	86.8 / 0.63	87.4 / 0.20
			94.2 / 0.48	83.1 / 0.00	89.0 / 0.50	73.0 / 0.00	87.4 / 0.61	88.6 / 0.40
✓			94.4 / 0.56	83.6 / 0.25	90.0 / 0.50	73.6 / 0.00	87.4 / 0.63	89.1 / 0.40
✓	✓		94.6 / 0.56	83.9 / 0.25	90.1 / 0.50	73.6 / 0.00	87.8 / 0.64	89.9 / 0.40
✓	✓	✓	94.7 / 0.56	83.9 / 0.25	91.0 / 0.56	74.5 / 0.00	87.8 / 0.64	90.1 / 0.40

Table 4: Ablation studies on feedback criteria used in Section 3.2.

Feedback	Pets	CUB	Flowers	Birdsnap	Food	IN100
Entropy (Hemmat et al., 2023; Kolossov et al., 2024)	94.6 / 0.52	83.2 / 0.13	**91.0 / 0.56**	73.6 / 0.00	86.9 / 0.58	89.4 / 0.40
Marginal score	**94.7 / 0.56**	**83.9 / 0.25**	**91.0 / 0.56**	**74.5 / 0.00**	**87.8 / 0.64**	**90.1 / 0.40**

disparity over the ***only real*** setup across all benchmarks, with the improvement (Δ) highlighted in green. Furthermore, GenDataAgent outperforms other SOTA methods across all backbones. Comparing our GenDataAgent with Internet Explorer demonstrates the effectiveness of Marginal Score Sampling and VoG Filtering. The worst-case disparity metric in table 2 shows that the synthetic data generated on the fly (e.g., by our method and Internet Explorer) improves the model's fairness. In contrast, Real-Fake surprisingly increases the worst-case disparity between classes in the CUB and Food datasets when using the ResNet-50 backbone. We attribute this to the fact that merely generating synthetic data can amplify existing biases between classes, as it fails to account for biases present in the original real dataset. Our on-the-fly mechanism addresses this issue by interacting with the model during training, thereby mitigating the bias. However, in the ***synthetic-only*** setting, without feedback guidance from the initial real dataset, both our GenDataAgent and Real-Fake experience performance drops in terms of worst-case disparity.

4.3 Ablation Study

Break-down Ablation. We further study the effect of each component in GenDataAgent: Marginal score Sampling, Llama-2 Perturbation, and VoG Filtering separately by a break-down ablation in Table 3. We start with the vanilla static synthetic data augmentation setting and gradually add Marginal score Sampling, Llama-2 Perturbation, and VoG Filtering. It can be observed that each component achieves uniform improvement across these datasets. For further ablation analysis on the hyper-parameters of each module, please refer to Appendix D.

Feedback Criteria. We compare our Marginal score sampling strategy with the Entropy criteria introduced by (Hemmat et al., 2023; Kolossov et al., 2024) in Table 4. The results indicate that the entropy criteria is either worse or on par with our method, while our Marginal score is simpler and more efficient. Thus, we selected the Marginal score as the feedback criterion.

Number of VoG Checkpoints. To figure out the best choice of VoG checkpoint numbers, we conduct experiments with 3, 4, and 5 VoG checkpoints in Table 5, where we find that there is no significant difference between different numbers of checkpoints. Thus we finally adopt 3 checkpoints to save resources.

Table 5: Comparison with different numbers of checkpoints used by VoG filtering.

#ckpts	Pets	CUB	Flowers	Birdsnap
3	94.7/0.56	83.9/0.25	91.0/0.56	74.5/0.00
4	94.5/0.56	84.1/0.25	91.5/0.56	74.4/0.00
5	94.4/0.52	84.0/0.25	91.4/0.56	73.6/0.00

Comparison with Traditional Data Augmentation Method. In Table 6, we compare our GenDataAgent with RandAugment (Cubuk et al., 2020), a traditional transformation-based data augmentation method. The results show that traditional transformation-based augmentation offers negligible performance improvement, whereas our GenDataAgent consistently enhances both Top-1 accuracy and worst-case disparity. This highlights the potential of synthetic data augmentation methods.

Scaling up/down Ablation with Time Analysis. To investigate the impact of different synthetic data ratios, we conduct additional experiments using a real-to-synthetic ratio of 1:10 for Real-Fake, and ratios of 1:0.5 and 1:0.1 for our on-the-fly augmentation framework. Notably, the search space of Real-Fake 1:10 is equivalent in size to that of GenDataAgent 1:0.5, as both generate the same total amount of synthetic data. Furthermore, we break down the time required for each step to analyze the

Table 6: Comparison between data augmentation method RandAugment and our GenDataAgent.

Method	Pets	CUB	Flowers	Birdsnap	Food	IN100
Only Real	93.6/0.40	83.1/0.13	87.4/0.40	73.0/0.00	86.8/0.63	87.4/0.20
RandAugment (Cubuk et al., 2020)	93.7/0.40	83.0/0.13	87.5/0.40	73.5/0.00	87.0/0.63	86.8/0.40
GenDataAgent (Ours)	**94.7/0.56**	**83.9/0.25**	**91.0/0.56**	**74.5**/0.00	**87.8/0.64**	**90.1**/0.40

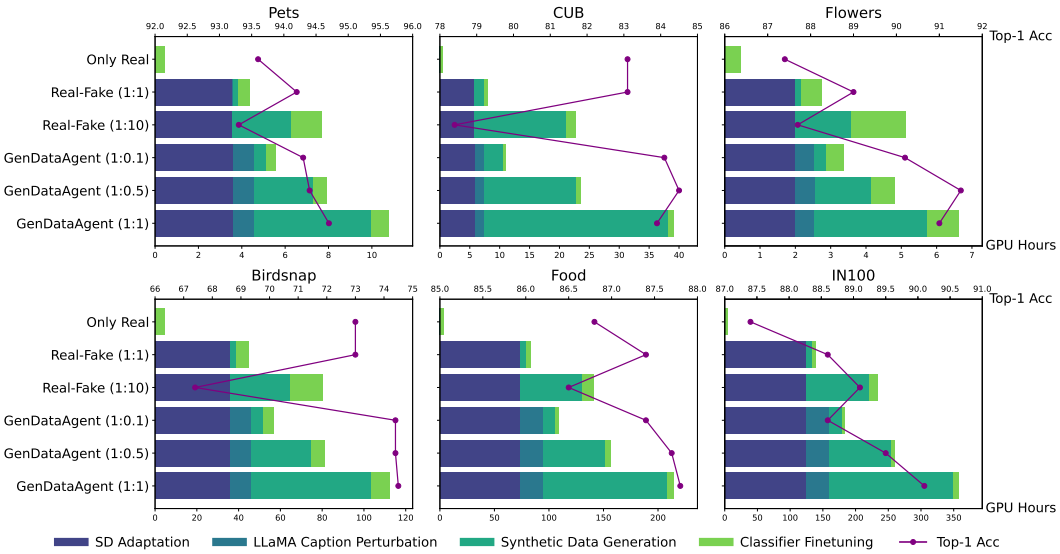

Figure 4: The bars on the left represent the GPU hours required for each step, while the line on the right depicts the Top-1 accuracy for each method. Simply increasing the synthetic data ratio from Real-Fake (1:1) to Real-Fake (1:10) can be detrimental. However, our GenDataAgent demonstrates improvement across both small and large real-synthetic ratios.

efficiency and identify the bottleneck of synthetic data augmentation. As shown in Figure 4, simply expanding the synthetic data search space and incorporating all data into training may be detrimental, as Real-Fake 1:10 underperforms compared to Real-Fake 1:1 on most datasets, particularly on fine-grained, smaller datasets. In contrast, our GenDataAgent 1:0.5 outperforms Real-Fake 1:1, indicating that our on-the-fly framework can effectively handle large volumes of synthetic data. Moreover, GenDataAgent 1:0.1 matches or exceeds Real-Fake's performance, demonstrating its capability with small synthetic data volumes. In terms of time analysis, training with only real data offers the best efficiency, while both Real-Fake and GenDataAgent require significant time to adapt the Stable Diffusion to the target dataset distribution. Although GenDataAgent 1:1 is constrained by the long synthetic data generation process, the lighter GenDataAgent 1:0.1 provides a balanced trade-off between performance and efficiency. In addition, further improvements in time efficiency can be realized by incorporating more efficient generation models.

5 Generating Content Analysis

Is there a relationship between synthetic data volume and category accuracy? In Figure 5, we show that synthetic data sampled by our GenDataAgent can reflect the trend of training accuracy well. From the absolute value perspective, after training the model on real data only (stage-1 in Algorithm 1), GenDataAgent generates more synthetic samples for categories with lower accuracy. The trend in Figure 5 indicates that our GenDataAgent is aware of classification accuracy and can produce synthetic data that complements lower-performing categories. When shifting to the incremental view (right part of Figure 5), the trend of Δ Top-1 accuracy (from the first iteration to convergence) is aligned well with the increment of synthetic data. This suggests that the performance gain is highly correlated to the number of synthetic data.

Does the augmentation by synthetic data relieve the over-fitting problem? We show the training and validation accuracy after convergence in Figure 6 (a). Compared to static augmentation, our on-the-fly method narrows the gap between the training and validation from the yellow region to the red area, suggesting that GenDataAgent can relieve the over-fitting problem.

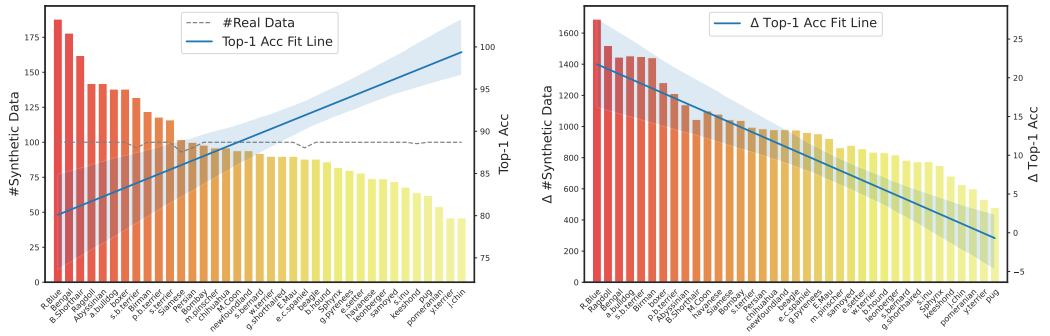

Figure 5: Absolute (left, the first on-the-fly iteration) and increment (right, from the first iteration to the convergence) relationship between the number of synthetic data (bar chart) and training top-1 accuracy on Oxford-IIIT Pets dataset across classes.

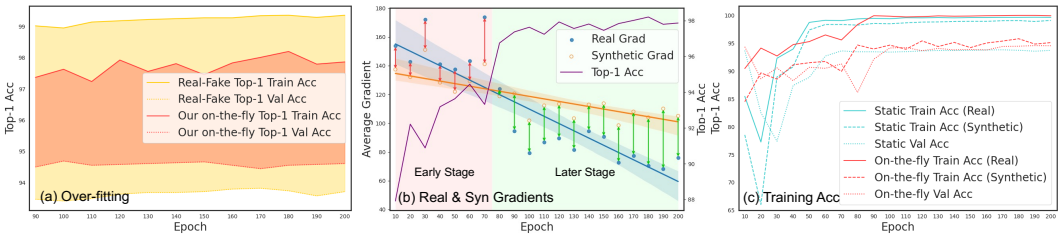

Figure 6: Analysis on Pets dataset. **(a)** The training and validation accuracy of Real-Fake and our method after convergence. The yellow and red regions are the train-val gaps of Real-Fake and our method respectively. **(b)** The average gradient of real and synthetic data, and the training top-1 accuracy during the on-the-fly process. Red arrows indicate the gradient of real data is larger than that of the synthetic data, while green arrows mean the opposite. **(c)** Classification accuracy on real and synthetic data during training on the Pets dataset in both static and on-the-fly settings.

Do real and synthetic data contribute the same to improving the model during on-the-fly stage? The answer is no. As shown in Figure 6 (b), we computer the average magnitude of the gradient of real and synthetic data for each on-the-fly iteration to show their impact on the model. The higher the gradient, the larger the impact (Li et al., 2023a). The initial finding is that the gradient of both real and synthetic data goes down as the model converges, while the gradient of real data decays more significantly. More importantly, there is a distinct boundary evident during the fine-tuning process, which separates the whole on-the-fly process into two stages. The real data is dominant in the early stage, where the expressive ability of the model increases rapidly. When the model starts to slowly converge (the later stage), synthetic samples contribute more to the learning process.

Does the model treat synthetic data differently from real data during fine-tuning? As shown in Figure 6 (c), in Real-Fake's static augmentation setting, the training accuracy of synthetic data is close to that of the real data while far away from the validation set. In comparison, the training accuracy of synthetic data in our on-the-fly augmentation exhibits an obvious gap compared to the real data and is much closer to that of the validation set. In other words, the model treats synthetic data almost the same as real data in the static setting. Yet, in the on-the-fly augmentation, the role of synthetic data differs considerably, and the model is not forced to fit the synthetic data perfectly.

6 Conclusion

In this work, we propose GenDataAgent, an on-the-fly framework for synthetic data augmentation in computer vision. GenDataAgent first aligns synthetic data with target distributions by fine-tuning the Stable Diffusion. Then, it prioritizes diverse samples that complement marginal real examples to narrow the search space. In addition, Llama caption perturbation and VoG filtering are employed to enhance the diversity and keeping the synthetic data within the target distribution respectively. Extensive evaluations across image classification tasks demonstrate its effectiveness, achieving state-of-the-art generalization and increased fairness. Moreover, our content analysis highlights the framework's potential to inspire further advancements in synthetic data techniques.

Ethics Statement

The research conducted in the paper conforms, in every respect, with the ICLR Code of Ethics.

Reproducibility Statement

We have provided implementation details in Section 4. We will also release all the code and models.

References

Chirag Agarwal, Daniel D'souza, and Sara Hooker. Estimating example difficulty using variance of gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10368–10378, 2022.

Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. Ethical considerations for responsible data curation. *Advances in Neural Information Processing Systems*, 36, 2024.

Pietro Astolfi, Arantxa Casanova, Jakob Verbeek, Pascal Vincent, Adriana Romero-Soriano, and Michal Drozdzal. Instance-conditioned gan data augmentation for representation learning. *arXiv preprint arXiv:2303.09677*, 2023.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *Transactions on Machine Learning Research*, 2023.

Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2011–2018, 2014.

Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. This dataset does not exist: training models from generated images. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2020.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493–1504, 2023.

Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1536–1546. IEEE, 2021.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022.

Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about. *Transactions on Machine Learning Research*, 2022.

Ali Borji. How good are deep models in understanding the generated images? *arXiv preprint arXiv:2208.10760*, 2022.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 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*, pp. 446–461. Springer, 2014.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *Advances in Neural Information Processing Systems*, 36, 2024.

Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pp. 23–37. Springer, 1995.

Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 269–285, 2018.

Avijit Ghosh, Lea Genuit, and Mary Reagan. Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, pp. 22–34. PMLR, 2021.

Melissa Hall, Candace Ross, Adina Williams, Nicolas Carion, Michal Drozdzal, and Adriana Romero Soriano. Dig in: Evaluating disparities in image generations with indicators for geographic diversity. *arXiv preprint arXiv:2308.06198*, 2023.

Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022.

Reyhane Askari Hemmat, Mohammad Pezeshki, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Feedback-guided data synthesis for imbalanced classification. *arXiv preprint arXiv:2310.00158*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. In *International Conference on Learning Representations*, 2021.

Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Improving contrastive learning on imbalanced data via open-world sampling. *Advances in Neural Information Processing Systems*, 34:5997–6009, 2021.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Germain Kolossov, Andrea Montanari, and Pulkit Tandon. Towards a statistical theory of data selection under weak supervision. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=HhfcNgQn6p.

Jedrzej Kozerawski, Victor Fragoso, Nikolaos Karianakis, Gaurav Mittal, Matthew Turk, and Mei Chen. Blt: Balancing long-tailed datasets with adversarially-perturbed images. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Ashish Jith Sreejith Kumar, Rachel S Chong, Jonathan G Crowston, Jacqueline Chua, Inna Bujor, Rahat Husain, Eranga N Vithana, Michaël JA Girard, Daniel SW Ting, Ching-Yu Cheng, et al. Evaluation of generative adversarial networks for high-resolution synthetic image generation of circumpapillary optical coherence tomography images for glaucoma. *JAMA ophthalmology*, 140 (10):974–981, 2022.

Shiye Lei, Hao Chen, Sen Zhang, Bo Zhao, and Dacheng Tao. Image captions are natural prompts for text-to-image models. *arXiv preprint arXiv:2307.08526*, 2023.

Alexander Cong Li, Ellis Langham Brown, Alexei A Efros, and Deepak Pathak. Internet explorer: Targeted representation learning on the open web. In *International Conference on Machine Learning*, pp. 19385–19406. PMLR, 2023a.

Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21330–21340, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. 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*, pp. 2970–2979, 2020.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *arXiv preprint arXiv:2310.16787*, 2023.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Jacob Metcalf and Kate Crawford. Where are human subjects in big data research? the emerging ethics divide. *Big Data & Society*, 3(1):2053951716650211, 2016.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pp. 16784–16804. PMLR, 2022.

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

Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8466–8475, 2018.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pp. 4334–4343. PMLR, 2018.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. It is all about where you start: Text-to-image generation with seed selection. *arXiv preprint arXiv:2304.14530*, 2023.

Pamela Samuelson. Generative ai meets copyright. *Science*, 381(6654):158–161, 2023.

Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8011–8021, 2023.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11492–11501, 2022.

Jie Shao, Ke Zhu, Hanxiao Zhang, and Jianxin Wu. Diffult: How to make diffusion model useful for long-tail recognition. *arXiv preprint arXiv:2403.05170*, 2024.

Viktoriia Sharmanska, Lisa Anne Hendricks, Trevor Darrell, and Novi Quadrianto. Contrastive examples for addressing the tyranny of the majority. *arXiv preprint arXiv:2004.06524*, 2020.

Joonghyuk Shin, Minguk Kang, and Jaesik Park. Fill-up: Balancing long-tailed data with generative models. *arXiv preprint arXiv:2306.07200*, 2023.

Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 769–778, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Hui Tang and Kui Jia. A new benchmark: On the utility of synthetic data with blender for bare supervised learning and downstream domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15954–15964, 2023.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Brandon Trabucco, Kyle Doherty, Max A Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

Soobin Um, Suhyeon Lee, and Jong Chul Ye. Don't play favorites: Minority guidance for diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3NmO9lY4Jn.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, Nayeong Kim, Suha Kwak, and Tae-Hyun Oh. Exploiting synthetic data for data imbalance problems: Baselines from a data perspective. *arXiv preprint arXiv:2308.00994*, 2023.

Jianhao Yuan, Jie Zhang, Shuyang Sun, Philip Torr, and Bo Zhao. Real-fake: Effective training data synthesis through distribution matching. *arXiv preprint arXiv:2310.10402*, 2023.

Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10145–10155, 2021.

Bo Zhao and Hakan Bilen. Synthesizing informative training samples with gan. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.

Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on thin air: Improve image classification with generated data. *arXiv preprint arXiv:2305.15316*, 2023.