GeNIe: Generative Hard Negative Images Through Diffusion

Soroush Abbasi Koohpayegani[1]   Anuj Singh[2,3]   K L Navaneet[1]
Hadi Jamali-Rad[2,3]   Hamed Pirsiavash[1]

[1]University of California, Davis
[2]Delft University of Technology, The Netherlands
[3]Shell Global Solutions International B.V., Amsterdam, The Netherlands

Abstract

Data augmentation is crucial in training deep models, preventing them from overfitting to limited data. Common data augmentation methods are effective, but recent advancements in generative AI, such as diffusion models for image generation, enable more sophisticated augmentation techniques that produce data resembling natural images. We recognize that augmented samples closer to the ideal decision boundary of a classifier are particularly effective and efficient in guiding the learning process. We introduce GeNIe *which leverages a diffusion model conditioned on a text prompt to merge contrasting data points (an image from the source category and a text prompt from the target category) to generate challenging samples for the target category. Inspired by recent image editing methods, we limit the number of diffusion iterations and the amount of noise. This ensures that the generated image retains low-level and contextual features from the source image, potentially conflicting with the target category. Our extensive experiments, in few-shot and also long-tail distribution settings, demonstrate the effectiveness of our novel augmentation method, especially benefiting categories with a limited number of examples. Our code is available here: https://github.com/UCDvision/GeNIe*

1. Introduction

Augmentation has become an integral part of training deep learning models, particularly when faced with limited training data. For instance, when it comes to image classification with limited number of samples per class, model generalization ability can be significantly hindered. Simple transformations like rotation, cropping, and adjustments in brightness and contrast artificially diversify the training set, offering the model a more comprehensive grasp of potential data variations. Exposure to a broader range of augmented samples enhances model robustness, adaptability, and accuracy in predicting novel instances. Hence, augmentation can serve as a practical strategy to boost the model's learning capacity, minimizing the risk of overfitting and facilitating effective knowledge transfer from limited labeled data to real-world scenarios.

Various image augmentation methods, encompassing standard transformations and learning-based approaches, have been proposed. Some augmentation approaches combine two images possibly from two different categories to generate a new sample image. The simplest ones in this category are MixUp [96] and CutMix [95] where two images are combined in the pixel space. However, the resulting image is usually not in the manifold of natural images, so the model will never see such images at test time.

Recently, generative models have excelled in generating natural images, prompting recent studies to leverage them for data augmentation. Inspired by diffusion-based image editing methods, we are interested in employing text-prompted latent diffusion models [67] to generate hard negative images that may lead to improved model training. Our core idea is to sample source images from different categories and prompt the latent diffusion model with text corresponding to a different target category. By controlling the amount of noise at the latent space and the number of diffusion iterations correspondingly, we can generate images that are from the target category while they resemble the source image in terms of the low-level features. This flexibility allows us to generate challenging examples ("hard negatives") for the classifier, contributing to enhanced training performance. Such hard negatives can improve model performance in handling long-tail distribution of data, as well as avoiding reasoning based on spurious correlations.

Long-tail distribution: In real-world applications, we often find that data follows a long-tail distribution. For example, when collecting data for driving scenarios, most of it will be from regular highway circumstances; situations like jaywalking might not be well-represented. Therefore,
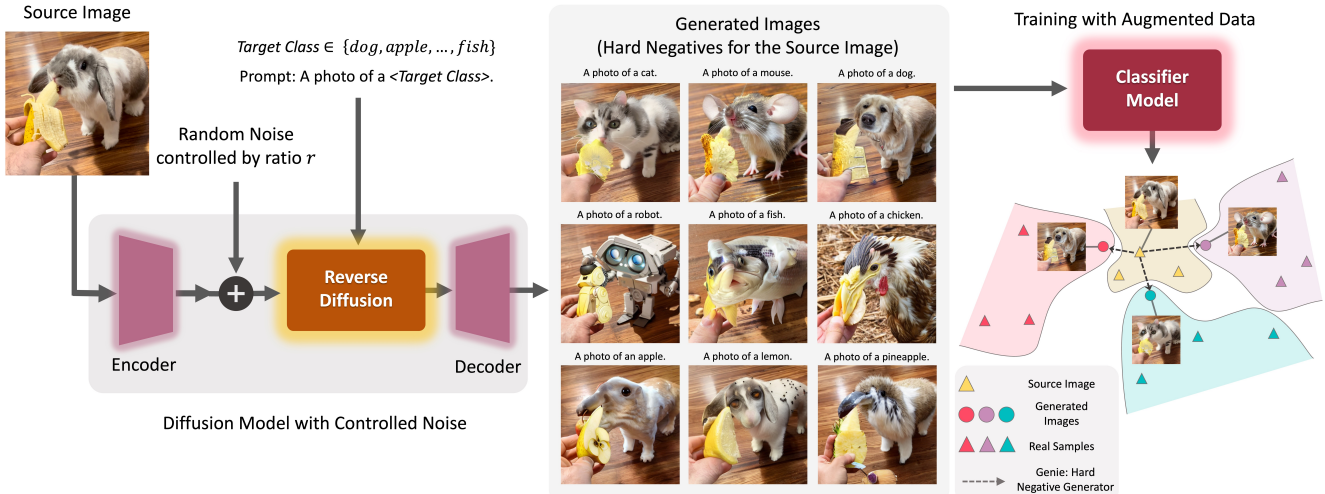
Figure 1. **Generative Hard Negative Images as Data Augmentation (`GeNIe`):** We generate hard negative images that belong to the target category but are similar to the source image in the low level features. We do this by adding noise to a source image and using a diffusion model to remove the noise conditioned on a text prompt that suggests this image is from the target category. By controlling the amount of noise, the diffusion model generates images that are hard negatives for the source category. We expect the generated images to lie close to the boundary between source and target categories (see Fig. 4), making them effective augmentations for training a classifier. To further enhance this process, one can sample the source and target categories using the classifier's confusion matrix.

even though samples of jaywalking are crucial for training self-driving or surveillance models to operate robustly (for instance, to ensure the car brake on time), we might not have enough data for it. This is where data augmentation using generative models can step in to help create more meaningful data for these less common scenarios.

Spurious correlations. Studies show that accidental appearance of certain visual features on images belonging to some categories can drastically impact the final decision of a classifier to be biased to co-occurrence of those spurious features [25]. For example, it is rare to see a giraffe in the middle of a highway. So, the model would pick up on this contextual bias and fail to recognize a giraffe in this uncommon circumstances. To address this, it is crucial to battle against this bias when training the model. One solution is to generate *harder* examples that violate such biases. In this case, our approach takes a noisy image of a street, and prompts the diffusion model with "A photo of a giraffe", resulting in generating out of distribution hard negatives for the street category. This in turn helps the model improve at recognizing giraffes in unconventional situations.

Our extensive experiments, in few-shot and also long-tail distribution settings, demonstrate the effectiveness of the proposed novel augmentation method (`GeNIe`), especially benefiting categories with a limited number of examples. For instance, as shown in Figure 1, `GeNIe` can take a source image of a bunny and generate images augmenting 9 different target categories in which the low-level features or background is similar to the source image.

2. Related Work

Data Augmentations: Data augmentation is a simple, yet effective way to improve generalization of deep learning models. Recent contrastive self-supervised learning methods also heavily depend on data augmentation. Simple flipping, cropping, color jittering, and blurring are some forms of weak and strong augmentations. The strength of augmentation is the extent of image distortion following the application of each augmentation [75]. However, using data augmentation is not trivial on some domains (e.g., medical). For example, using blurring might remove important low level information from images. More advanced approaches, such as MixUp [96] and CutMix [95], mix images and their labels accordingly [14, 29, 37, 48], which seem to offer better model generalization and robustness as well. However, the augmented (mixed) images are now not natural anymore and thus training proceeds on out of distribution images. Unlike the above methods, we propose to utilize pretrained latent diffusion models to augment the training data. Note that generative vision models are indeed trained to generate natural images; however, the typical challenge here is that the generated images might not necessarily belong to same data distribution as the training dataset, which could entail further finetuning them to the specific domain.

Training Data Augmentation with Generative Models: Generative models could be used to generate images within the manifold of natural images. Using synthesized images from generative models to augment training data have been studied before in many domains [24, 70], in-

cluding domain adaptation [32], visual alignment [58], and mitigation of dataset bias [72]. While previous methods predominantly relied on GANs [41, 84, 99] as the generative model, more recent studies promote using diffusion models [68] to augment the data. More specifically, [4, 28, 50, 73, 83] study the effectiveness of text-to-image diffusion models in data augmentation by diversification of each class with synthetic images. [33] uses an SVM in the latent space of the CLIP model to find spurious correlations and generates challenging examples using a diffusion model. [22] use CLIP model to filter generated images. Along the same lines, some studies utilize text-based diffusion for image editing [20]. From this angle, Boomerang [51] and SDEdit [55] are the closest approaches to our method. They they edit each data samples (source images) by adding small amount of noise to each image and applying reverse diffusion to generate an augmented image close to the source image. Unlike Boomerang, we set the negatives classes as source and employ the generative model to transform it to have the semantics of the positive class (Target class) while preserving the visual futures of the negative samples, which is the key reason why we argue these newly generated samples are now *hard negatives*. In a nutshell, the aforementioned studies focus on improving diversity of each class with effective prompts and diffusion models, however, we focus on generating effective hard negative samples for each class by combining two sources of contradicting information (images from the source category and text prompt from the target category.)

Language Guided Recognition Models: Language-vision foundation models have recently received an upsurge of attention [2, 62–64, 68, 69]. These models use human language to guide generating images, or extracting features from images which are aligned with human languages. Due to alignment with human language, these models can be used in downstream recognition tasks. For example, CLIP [62] shows decent zero-shot performance on many downstream tasks by matching image to their text descriptions. Some recent works improve utilization of human language in the prompt [19, 59], and others use a diffusion model directly as classifier [39]. Similar to the above, we use a foundation model (Stable Diffusion 1.5 [68]) to improve the downstream task. Concretely, we utilize category names of the downstream tasks to augment their associate training data with hard negative samples.

Few-Shot Learning: revolves around learning very fast from only a handful of samples. FSL is typically conducted in two stages: pretraining on an abundance of data followed by fast adaptation to unseen few-shot episodes. Each episode consists of a support set allowing the model to adapt itself quickly to the unseen classes, and a query set on which the model is evaluated. In supervised FSL [1, 10, 21, 43, 60, 76, 79, 93, 101], pretraining is done on

a labeled dataset, whereas in unsupervised FSL [3, 31, 34, 36, 49, 54, 61, 74, 87] the pretraining has to be conducted on an unlabeled dataset posing an extra challenge in the learning paradigm and neighboring these methods closer to the realm of self-supervised learning. Even though FSL is not of primal interest in this work, we assess the impact of `GeNIe` on a number of few-shot scenarios and state-of-the-art baselines by accentuating on its impact on the few-shot inference stage.

3. Method: `GeNIe`

Given an image from a source category, we are interested in generating an image for a target category, while preserving low-level visual features or background context of the source image. Assume we have a conditional latent diffusion model that can denoise an input image guided by a text prompt, during the reverse process. With this in place, we pass a source image through the encoder and introduce noise to its latent embedding followed by denoising it while being conditioned on a text prompt. In such a construct, the proximity of the final decoded image to the source image or the text prompt depends on the amount of added noise and the number of reverse diffusion iterations. Hence, by controlling the amount of noise, we can generate images that blend characteristics of both the text prompt and the input image. If we do not provide much of visual details in the text prompt (e.g., desired background, etc.), we expect the final image to follow the details of the source image while reflecting the semantics of the text prompt. This capability offered by conditional latent diffusion models allows us to create novel challenging augmentations, and this sits at the core of the proposed approach.

Our method is shown in Fig. 1. To augment a target category, we select an image from the source category and apply noise equivalent to n diffusion iterations to its latent embedding. Assuming that T is the total number of diffusion iterations, we simply choose $n = \lfloor r \times T \rfloor$, where r is the ratio of the desired amount of noise. The noisy latent embedding of the image, along with a prompt like "A photo of a $<$target category$>$" is then passed through the reverse diffusion model followed by the decoder to generate an image. This denoising process runs for n iterations resulting in an image closer to the manifold of natural images. We argue that the newly generated image is a *hard negative* example for the source category since it shares low-level features with the source image while it represents the target category. We can now use this generated sample to augment the target category when training an image classifier. Note that this approach to image editing with reduced number of iterations $n < T$ is inspired by [51, 55]. What differentiates our proposed approach is extending it to generate hard negatives by combining two contradicting sources of information: i.e., source image and the text prompt referring to
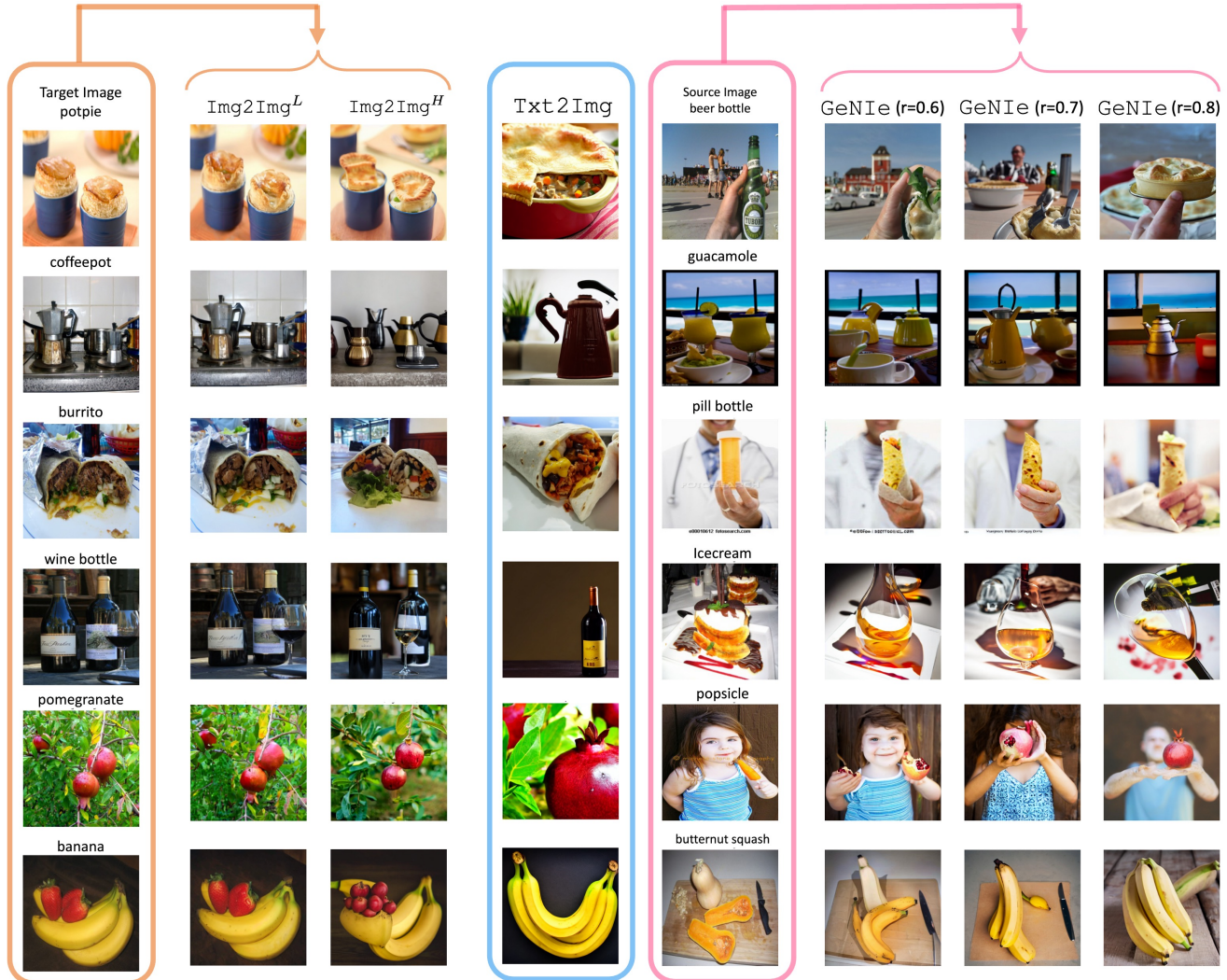
Figure 2. **Visualization of Generative Samples:** We compare `GeNIe` with two baselines: **Img2Img**^{L} **augmentation:** both image and text prompt are from the same category. Adding noise does not change the image much, so they are not hard examples. **Txt2Img augmentation:** We simply use the text prompt only to generate an image for the desired category (e.g., using a text2image method). Such images may be far from the domain of our task since the generation is not informed by any visual data from our task. **GeNIe augmentation:** We use target category name in the text prompt only along with the source image. At appropriate amount of noise (equivalent to 80% of all diffusion iteration), we generate the desired images. Low amount of noise is still far from target category since the source image has to much of effect compared to the text prompt. Note that we never define low-level features concretely and never evaluate if the augmented images preserves those features. That is our hypothesis only. We only evaluate if the augmented images help the accuracy.

the target category. Notably, the source category can be randomly sampled or be chosen from the confusion matrix of a recently trained image classifier based on real training data. The latter will result in *harder negative* samples that can benefit the training the most.

4. Experiments

We evaluate the impact of `GeNIe` on Few-Shot classification in Sec 4.1, Long-Tailed classification in Sec 4.2, and fine-grained classification in Sec 4.3.

Baselines. We use Stable Diffusion 1.5 [68] as our base diffusion model. In all settings, we use the same prompt format to generate images for the target class: i.e., "A photo of a <target category>", where we replace the `target category` with the target category label. We generate 512×512 images for all methods. For fairness in comparison, we generate the same number of new images for each class. We use a single NVIDIA RTX 3090 for image generation. We consider 2 diffusion-based baselines and a suite of traditional data augmentation baselines, as follows.

Img2Img: This baseline follows the data augmentation strategy based on Stable Diffusion proposed in [51, 55].

ResNet-18				
Method	**Augmentation**	**Pre-training**	**1-shot**	**5-shot**
Δ-Encoder [71]	-	sup.	59.9	69.7
SNCA [90]	-	sup.	57.8±0.8	72.8±0.7
iDeMe-Net [12]	-	sup.	59.14±0.86	74.63±0.74
Robust + dist [21]	-	sup.	63.73±0.62	81.19±0.43
AFHN [43]	-	sup.	62.38±0.72	78.16±0.56
ProtoNet+SSL [78]	Weak	sup.+ssl	-	76.6
Neg-Cosine [46]	Weak	sup.	62.33±0.82	80.94±0.59
Centroid Align[1]	-	sup.	59.88±0.67	80.35±0.73
Baseline [10]	-	sup.	59.55±0.75	77.31±0.59
Baseline++ [10]	-	sup.	58.95±0.77	76.65±0.61
PSST [13]	Weak	sup.+ssl	59.52±0.46	77.43±0.46
UMTRA [36]	Weak	unsup.	43.09±0.35	53.42±0.31
ProtoCLR [54]	Weak	unsup.	50.90±0.36	71.59±0.29
SimCLR [9]	Weak	unsup.	62.58±0.37	79.66±0.27
SimSiam [11]	Weak	unsup.	62.80±0.37	79.85±0.27
UniSiam+dist [49]	Weak	unsup.	**64.10±0.36**	**82.26±0.25**
UniSiam [49]	Weak	unsup.	63.14±0.77	81.40±0.53
UniSiam [49]	Heavy	unsup.	62.80±0.77	81.15±0.55
UniSiam [49]	CutMix	unsup.	62.67±0.79	80.63±0.55
UniSiam [49]	MixUp	unsup.	62.14±0.78	80.74±0.55
UniSiam [49]	Img2Img^L	unsup.	63.86±0.77	82.05±0.53
UniSiam [49]	Img2Img^H	unsup.	69.06±0.73	83.89±0.48
UniSiam [49]	Txt2Img	unsup.	74.14±0.63	84.62±0.47
UniSiam [49]	GeNIe (Ours)	unsup.	**75.45±0.62**	**85.38±0.44**

ResNet-34				
Method	**Augmentation**	**Pre-training**	**1-shot**	**5-shot**
MatchingNet [85]	-	sup.	53.20±0.78	68.32±0.66
ProtoNet [77]	-	sup.	53.90±0.83	74.65±0.64
MAML [23]	-	sup.	51.46±0.90	65.90±0.79
RelationNet [79]	-	sup.	51.74±0.83	69.61±0.67
Baseline [10]	-	sup.	49.82±0.73	73.45±0.65
Baseline++ [10]	-	sup.	52.65±0.83	76.16±0.63
SimCLR [9]	Weak	unsup.	63.98±0.37	79.80±0.28
SimSiam [11]	Weak	unsup.	63.77±0.38	80.44±0.28
UniSiam+dist [49]	Weak	unsup.	65.55±0.36	83.40±0.24
UniSiam [49]	Weak	unsup.	64.26±0.79	82.33±0.53
UniSiam [49]	Strong	unsup.	64.48±0.81	82.13±0.55
UniSiam [49]	CutMix	unsup.	63.99±0.80	81.69±0.56
UniSiam [49]	MixUp	unsup.	63.65±0.78	80.12±0.78
UniSiam [49]	Img2Img^L	unsup.	65.52±0.79	82.92±0.51
UniSiam [49]	Img2Img^H	unsup.	70.46±0.75	84.79±0.47
UniSiam [49]	Txt2Img	unsup.	75.36±0.61	85.45±0.45
UniSiam [49]	GeNIe (Ours)	unsup.	**77.08±0.60**	**86.28±0.42**

ResNet-50				
Method	**Augmentation**	**Pre-training**	**1-shot**	**5-shot**
UniSiam [49]	Weak	unsup.	64.55±0.79	83.39±0.51
UniSiam [49]	Strong	unsup.	64.76±0.79	83.24±0.51
UniSiam [49]	CutMix	unsup.	64.25±0.80	83.24±0.45
UniSiam [49]	MixUp	unsup.	63.75±0.80	84.55±0.51
UniSiam [49]	Img2Img^L	unsup.	66.00±0.78	84.01±0.49
UniSiam [49]	Img2Img^H	unsup.	71.14±0.73	85.66±0.46
UniSiam [49]	Txt2Img	unsup.	76.44±0.61	86.50±0.42
UniSiam [49]	GeNIe (Ours)	unsup.	**77.28±0.60**	**87.22±0.40**

Table 1. ***mini*-ImageNet:** We use our augmentations on (5-way, 1-shot) and (5-way, 5-shot) few-shot settings of mini-Imagenet dataset with 3 different backbones (ResNet-18,34, and 50). We compare with various baselines and show that our augmentations with UniSiam method outperforms all the baselines including Txt2Img augmentation. The number of generated images per class is 4 for 1-shot and 20 for 5-shot settings. Note that UniSiam has used only weak augmentation, so we add the other methods for fair comparison.

Concretely, we sample an image from a target class, add noise to its latent representation and then pass it through the reverse diffusion. Notice that the focus here is on a target class for which we generate extra positive samples. Adding large amount of noise (corresponding to large number of diffusion iterations) leads to generating an image less similar to the original image. We use two different noise magnitudes for this baseline: $r = 0.3$ and $r = 0.7$ and denote them by Img2Img^L and Img2Img^H, respectively.

Txt2Img: For this baseline, we omit the the forward diffusion process and only use the reverse process starting from a text prompt to the target classes of interest. This is similar to the base text-to-image generation strategy adopted in [4, 28, 50, 68, 73, 83]. Note that an extreme case of adding maximum noise ($r = 0.999$) in Img2Img degenerates it to Txt2Img.

Traditional Data Augmentation: We consider both weak and strong traditional augmentations. More specifically, for weak augmentation we use random resize crop with scaling $\in [0.2, 1.0]$ and horizontal flipping. For strong augmentation, we consider random color jitter, random grayscale, and Gaussian blur. For the sake of completeness, we also compare against more recent advanced data augmentations such as CutMix [95] and MixUp [96].

Fig. 2 and A1 illustrate a set of generated augmentation examples for Txt2Img, Img2Img, and GeNIe. As can be seen, GeNIe effectively generates hard negatives for the

source image class by preserving its low-level features and transforming its main target class according to the prompt.

4.1. Few-shot Classification

We assess the impact of GeNIe compared to other forms of augmentation in a number of few-shot classification scenarios, where the model has to learn only from the samples contained in the (N-way, K-shot) support set and infer on the query set. Note that this corresponds to an inference-only FSL setting where a pretraining stage on an abundant dataset is discarded. The goal is to assess how well the model can benefit from the augmented data samples while keeping the original $N \times K$ samples intact.

Datasets. We conduct our few-shot experiments on two most commonly adopted few-shot classification datasets: *mini*-Imagenet [65] and *tiered*-Imagenet [66]. *mini*-Imagenet is a subset of ImageNet [17] for few-shot classification. It contains 100 classes with 600 samples each. We follow the predominantly adopted settings of [10, 65] where we split the entire dataset into 64 classes for training, 16 for validation and 20 for testing. *tiered*-Imagenet is a larger subset of ImageNet with 608 classes and a total of $779, 165$ images, which are grouped into 34 higher-level nodes in the *ImageNet* human-curated hierarchy. This set of nodes is partitioned into 20, 6, and 8 disjoint sets of training, validation, and testing nodes, and the corresponding classes form the respective meta-sets.

ResNet-18				
Method	Augmentation	Pre-training	1-shot	5-shot
Transd-CNAPS [5]	-	sup.	65.9 ± 1.0	81.8 ± 0.7
FEAT [94]	-	sup.	70.80	84.79
SimCLR[9]	Weak	unsup.	63.38±0.42	79.17±0.34
SimSiam [11]	Weak	unsup.	64.05±0.40	81.40±0.30
UniSiam + dist [49]	Weak	unsup.	67.01±0.39	84.47±0.28
UniSiam [49]	Weak	unsup.	63.08±0.70	81.04±0.51
UniSiam [49]	Heavy	unsup.	62.84±0.71	80.94±0.51
UniSiam [49]	CutMix	unsup.	62.11±0.71	78.90±0.55
UniSiam [49]	MixUp	unsup.	62.10±0.70	78.35±0.55
UniSiam [49]	$\texttt{Img2Img}^L$	unsup.	63.89±0.69	81.76±0.49
UniSiam [49]	$\texttt{Img2Img}^H$	unsup.	68.68±0.66	83.45±0.46
UniSiam [49]	Txt2Img	unsup.	72.91±0.61	84.15±0.45
UniSiam [49]	GeNIe(Ours)	unsup.	**73.62±0.62**	**85.00±0.43**
ResNet-34				
UniSiam + dist [49]	Weak	unsup.	68.65±0.39	85.70±0.27
UniSiam [49]	Weak	unsup.	65.02±0.71	82.51±0.50
UniSiam [49]	Heavy	unsup.	64.81±0.72	82.42±0.51
UniSiam [49]	CutMix	unsup.	63.77±0.72	80.34±0.55
UniSiam [49]	MixUp	unsup.	64.12±0.73	80.03±0.55
UniSiam [49]	$\texttt{Img2Img}^L$	unsup.	66.12±0.70	83.11±0.49
UniSiam [49]	$\texttt{Img2Img}^H$	unsup.	70.38±0.66	84.74±0.45
UniSiam [49]	Txt2Img	unsup.	74.99±0.61	85.35±0.44
UniSiam [49]	GeNIe (Ours)	unsup.	**75.74±0.62**	**86.02±0.43**
ResNet-50				
UniSiam + dist [49]	Weak	unsup.	69.60±0.38	86.51±0.26
UniSiam [49]	Weak	unsup.	66.75±0.70	84.72±0.47
UniSiam [49]	Heavy	unsup.	66.46±0.71	84.52±0.48
UniSiam [49]	CutMix	unsup.	65.99±0.71	83.29±0.50
UniSiam [49]	MixUp	unsup.	66.12±0.45	84.05±0.78
UniSiam [49]	$\texttt{Img2Img}^L$	unsup.	67.76±0.69	85.28±0.46
UniSiam [49]	$\texttt{Img2Img}^H$	unsup.	72.38±0.65	86.65±0.42
UniSiam [49]	Txt2Img	unsup.	77.06±0.58	87.25±0.41
UniSiam [49]	GeNIe (Ours)	unsup.	**78.01±0.58**	**88.00±0.39**

Table 2. *tiered*-ImageNet: Accuracies in (% ± std) for 5-way, 1-shot and 5-way, 5-shot classification settings on the test-set. We compare against various state-of-the-art supervised and unsupervised few-shot classification baselines as well as other augmentation methods, with UniSiam [49] pre-trained ResNet-18, 34 and 50 backbones. UniSiam+dist indicates that the backbone was pre-trained with a ResNet-50 teacher network.

ResNet-50				
Method	Many	Med.	Few	Overall Acc
CE [16]	64.0	33.8	5.8	41.6
LDAM [8]	60.4	46.9	30.7	49.8
c-RT [35]	61.8	46.2	27.3	49.6
τ-Norm [35]	59.1	46.9	30.7	49.4
Causal [80]	62.7	48.8	31.6	51.8
Logit Adj. [56]	61.1	47.5	27.6	50.1
RIDE(4E)† [89]	68.3	53.5	35.9	56.8
MiSLAS [100]	62.9	50.7	34.3	52.7
DisAlign [97]	61.3	52.2	31.4	52.9
ACE† [7]	71.7	54.6	23.5	56.6
PaCo† [15]	68.0	56.4	37.2	58.2
TADE† [98]	66.5	**57.0**	43.5	58.8
TSC [45]	63.5	49.7	30.4	52.4
GCL [44]	63.0	52.7	37.1	54.5
TLC [40]	68.9	55.7	40.8	55.1
BCL† [102]	67.6	54.6	36.6	57.2
NCL [42]	67.3	55.4	39.0	57.7
SAFA [30]	63.8	49.9	33.4	53.1
DOC [88]	65.1	52.8	34.2	55.0
DLSA [91]	67.8	54.5	38.8	57.5
ViT-B				
LiVT* [92]	76.4	59.7	42.7	63.8
ViT [18]	50.5	23.5	6.9	31.6
MAE [27]	74.7	48.2	19.4	54.5
DeiT [82]	70.4	40.9	12.8	48.4
LiVT [92]	73.6	56.4	41.0	60.9
LiVT + $\texttt{Img2Img}^L$	74.26	56.37	34.31	60.46
LiVT + $\texttt{Img2Img}^H$	73.83	56.35	45.32	61.56
LiVT + Txt2Img	**74.86**	55.59	48.32	62.18
LiVT + GeNIe	74.46	56.71	**50.85**	**62.80**

Table 3. **Long-Tailed ImageNet-LT:** We compare different augmentation methods on ImageNet-LT and report Top-1 accuracy for "Few", "Medium", and "Many" sets. † indicates results with ResNeXt50. ∗: indicates training with 384 resolution so is not directly comparable with other methods with 224 resolution. On the "Few" set and LiVT [92] method, our augmentations improve the accuracy by 9.8 points compared to traditional data augmentation and 2.5 points compared to Txt2Img augmentation.

Evaluation. To quantify the impact of different augmentation methods, we evaluate the test-set accuracies of a state-of-the-art unsupervised few-shot learning method with GeNIe and compare them against the accuracies obtained using other augmentation methods. Specifically, we use UniSiam [49] pre-trained with ResNet-18, ResNet-34 and ResNet-50 backbones and follow its evaluation strategy of fine-tuning a logistic regressor to perform (N-way, K-shot) classification on the test sets of *mini-* and *tiered*-Imagenet. Following [65], an episode consists of a labelled support-set and an unlabelled query-set. The support-set contains randomly sampled N classes where each class contains K samples, whereas the query-set contains Q randomly sampled unlabelled images per class. We conduct our experiments on the two most commonly adopted settings: (5-way, 1-shot) and (5-way, 5-shot) classification settings. Following the literature, we sample 16-shots per class for the query

set in both settings. We report the test accuracies along with the 95% confidence interval over 600 and 1000 episodes for *mini*-ImageNet and *tiered*-ImageNet, respectively.

Implementation Details: GeNIe generates images for each class by using all images from every other class as the source image along with their class labels as the corresponding text prompt. Specifically, we generate 4 samples per class as augmentations in the 5-way, 1-shot setting and 20 samples per class as augmentations in the 5-way, 5-shot setting. For the sake of a fair comparison, we ensure that the total number of labelled samples in the support set after augmentation remains the same across all different traditional and generative augmentation methodologies.

Results: The results on *mini*-Imagenet and *tiered*-Imagenet for both (5-way, 1 and 5-shot) settings are summarized in Table 1 and Table 2, respectively. Regardless of

Method	Birds	Cars	Foods	Aircraft
	CUB200 [86]	Cars196 [38]	Food101 [6]	Aircraft [52]
Baseline	90.34	49.77	82.92	29.19
Img2Img^L	90.73	50.38	87.4	30.99
Img2Img^H	91.31	56.39	91.7	34.73
Txt2Img	92.01	81.29	92.95	41.73
GeNIe (r=0.8)	92.49	87.74	**93.06**	46.49
GeNIe (r=0.7)	**92.52**	**87.92**	92.88	**47.01**

Table 4. **Few-shot Learning on Finegrained dataset:** We train SVM classifier on top of DinoV2 ViT-G pretrained backbone and report Top-1 accuracy on the test set of each dataset. Our baseline here is the SVM trained with traditional augmentations. GeNIe outperforms Baseline across all dataset. Compared to data augmentation with Txt2Img only, GeNIe improves accuracy by 6.6 point and 5.2 point in Cars and Aircraft dataset respectively.

the choice of backbone, we observe that GeNIe helps consistently improve UniSiam's performance and outperform other supervised and unsupervised few-shot classification methods as well as other data augmentation techniques on both datasets, across both (5-way, 1 and 5-shot) settings.

4.2. Long-Tailed Classification

We evaluate our method in Long-Tailed data, where number of instances per class is not balanced and most categories have limited samples (tail). Our goal is to mitigate this bias by augmenting the tail of the distribution with generated samples. Following LViT [92], first, we train an MAE [26] on the unbalanced dataset without any augmentation. We then train the Balanced Fine Tuning stage by incorporating the augmentation data generated using GeNIe or the other baselines. We following the proposed settings in [92] for the Balanced Fine Tuning stage, which includes traditional augmentation including CutMix, MixUp with a Balanced Binary Cross-Entropy (Bal-BCE) loss.

Dataset: We perform experiments on ImageNet-LT [47]. It contains 115.8K images from $1,000$ categories. Maximum and minimum number of images per class is 1280 and 5 respectively. Imagenet-LT classes can be divided into 3 groups: "Few" with less than 20 images, "Med" with $20 - 100$ images, and "Many" with more than 100 images. Imagenet-LT uses the same validation set as ImageNet.

We generate new data for the "Few" categories only. We limit the number of generated images to 50 samples per class. For GeNIe, instead of randomly sampling the source images from other classes, we use confusion matrix on the training data to find top-4 most confused classes and only consider those classes for random sampling. The source category may be from "Many", "Med", or "Few sets".

Implementation Details: We download the pretrained ViT-B of [92] and finetune it with Bal-BCE loss proposed therein on the augmented dataset. We use four NVIDIA RTX 3090 GPUs with the same hyperparameters as discussed in [92] for finetuning: 100 epochs, $lr = 0.008$, batch size of 1024, CutMix and MixUp for the data augmentation.
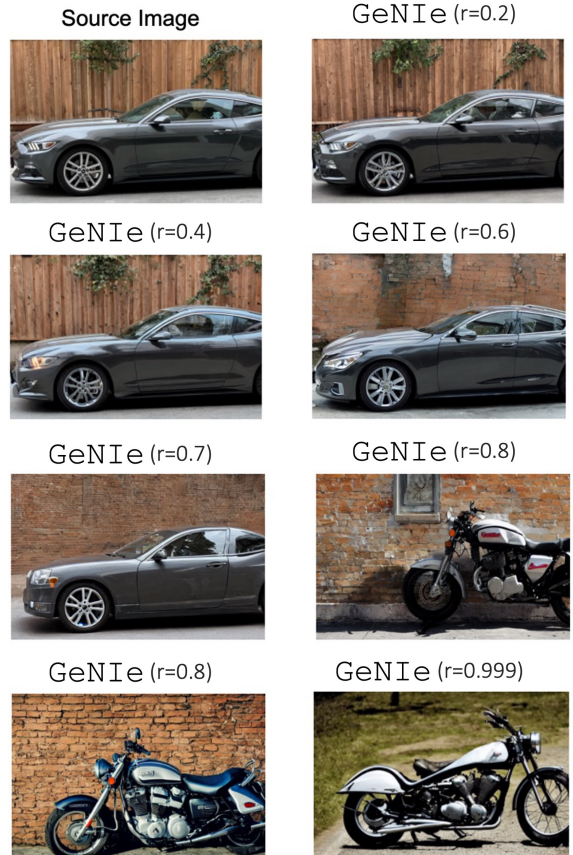

Figure 3. **Effect of noise in GeNIe:** We employ GeNIe to generate augmentations for the target class "motorcycle." As shown in the example above, a lower ratio of noise results in images that closely resemble the semantics of the source images, presenting an inconsistency with the intended target label.

Results: The results are summarized in Table 3. Augmenting training data with GeNIe improves accuracy on the "Few" set by 9.8% and 2.5% compared against LViT and LViT with Txt2Img augmentation, respectively. Moreover, GeNIe improves the overall accuracy by 1.9% compared to the baselines with only traditional augmentation.

4.3. Fine-grained Few-shot Classification

To further investigate the impact of the proposed method, we compare GeNIe with other text-based data augmentation techniques across four distinct fine-grained datasets within a 20-way, 1-shot classification setting. We employ the pretrained DinoV2 ViT-G [57] backbone as a feature extractor to derive features from training images. Subsequently, an SVM classifier is trained on these features, and we report the Top-1 accuracy of the model on the test set.

Datasets: We assess our method on several datasets: Food101 [6] with 101 classes of various foods, CUB200 [86] with 200 bird species classes, Cars196 [38] with 196
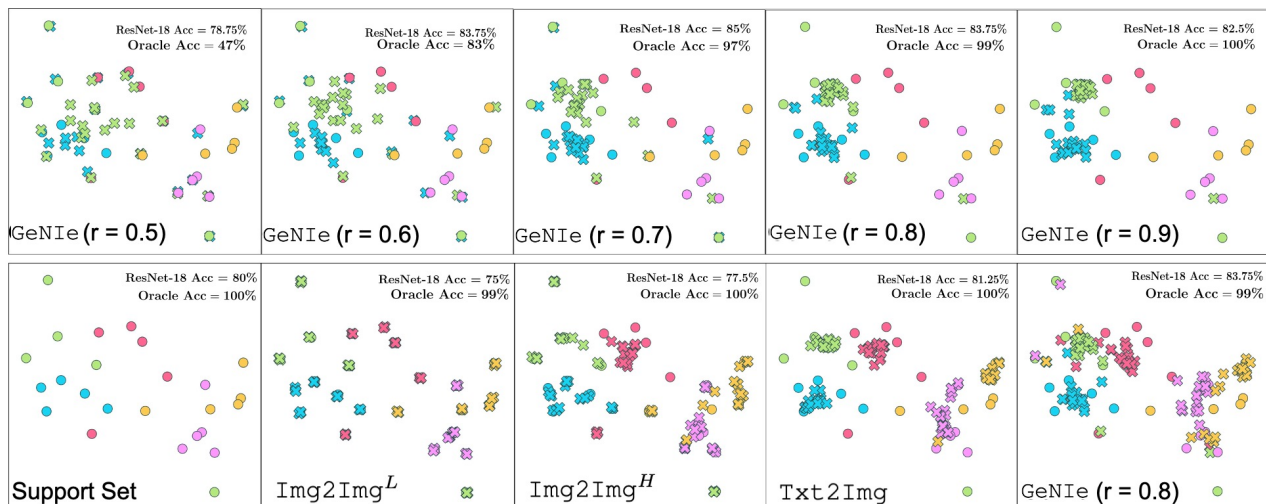
Figure 4. **Embedding visualizations of generative augmentations:** We show UMAP embedding for 5 classes of mini-Imagenet dataset. Colors indicate classes and crosses indicate augmentatedn samples. In the first row, we only display augmentations for the green and blue classes to avoid visual clutter while in the second row, we show augmentations from ours and baseline for all 5 categories. GeNIe-generated samples occupy the boundaries between classes, displaying increased diversity compared to other augmentation methods. Notably, decreasing the noise level within GeNIe enhances sample diversity, albeit with a trade-off in label consistency, as detailed in Table 5.

Method	GeNIe				
	(r=0.5)	(r=0.6)	(r=0.7)	(r=0.8)	(r=0.9)
Oracle Acc	73.43±0.51	85.77±0.43	94.52±0.23	98.21±0.10	99.26±0.06
5-shot Acc	84.66±0.5	85.48±0.4	86.30±0.4	86.28±0.4	86.17±0.4

Table 5. **Consistency of the label for generated samples:** We present both Oracle accuracy and ResNet-34 5-shot accuracy (numbers are copied from Table 1 and Table 6). Notably, reducing the noise level below 0.7 is associated with a decline in Oracle accuracy and a subsequent degradation in the performance of the final few-shot model.

Noise	ResNet-18		ResNet-34		ResNet-50	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
r=0.5	68.47±0.7	83.66±0.5	70.21±0.7	84.66±0.5	70.55±0.7	85.71±0.4
r=0.6	71.59±0.6	84.69±0.5	73.17±0.7	85.48±0.4	73.65±0.7	86.62±0.4
r=0.7	73.96±0.6	85.45±0.4	75.79±0.6	86.30±0.4	76.04±0.6	87.18±0.4
r=0.8	75.45±0.6	85.38±0.4	77.08±0.6	86.28±0.4	77.28±0.6	87.22±0.4
r=0.9	75.96±0.6	85.29±0.4	77.63±0.6	86.17±0.4	78.03±0.6	87.10±0.4

Table 6. **Effect of Noise in GeNIe:** We use the same setting as in Table 1 to study the effect of the amount of noise. As expected, small noise results in worse accuracy since some generated images may be from the source category rather than the target one. As shown in Table 5, for $r = 0.5$ only 73% of the generated data is from the target category. This behavior is also shown in Figure 3.

car model classes, and FGVC-Aircraft [52] with 41 aircraft manufacturer classes. Further details on each dataset can be found in Appendix B. The reported metric is the average Top-1 accuracy over 100 episodes. Each episode involves sampling 20 classes and 1-shot from the training set, with the final model evaluated on the respective test set.

Implementation Details: We enhance the basic prompt by incorporating the superclass name for the fine-grained dataset: "A photo of a <target class>, a type of <superclass>". For instance, in the *food* dataset and the *burger* class, our prompt reads: "A photo of a *burger*, a type of *food*." No additional augmentation is used for generative methods in this context. We generate 19 samples for both cases of our method and baseline with weak augmentation.

Results: Table 4 summarizes the results. Here again GeNIe outperform all other baselines, including Txt2Img augmentation. Notably, GeNIe exhibits great effectiveness in more challenging datasets, outperforming the baseline with traditional augmentation by about 38% for the Cars dataset and by roughly 17% for the Aircraft dataset.

4.4. Ablation and Analysis

Label consistency of the generated samples. Remember that a very high noise level ($r = 0.99$) practically diminishes the impact of the source image, degenerating the the process into text-to-image generation. Thus, to be able to preserve the contextual low-level properties of the source image, we need to start the reverse process from a lower noise level. On the flip side, we observe in Fig. 3 that a moderately low noise level also does not seem to allow the model to appropriately incorporate the prompt in the reverse process, practically leading to a generated images that might not manifest the semantics of the target category. We refer to this matching between the prompt label and the semantics of the generated samples as label consistency. To quantify label consistency, we utilize an ImageNet-pretrained DeiT-Base [81] backbone as an Oracle model to infer the actual label of the generated image as opposed to the text prompt generating it - assuming that the Oracle

8

model provides accurate labels. In Table 5, we present the 5-shot accuracy of the few-shot model (ResNet-34, presented earlier in Table 1) alongside the Oracle Accuracy. We observe a decline in the label consistency of generated data (quantified by the performance of the Oracle model) when decreasing the noise level. As can be seen, this also results in a degradation in the performance of the final few-shot model ($86.28\% \rightarrow 84.66\%$) corroborating that an appropriate choice of r plays a crucial role in our design strategy. We investigate this further in the following.

Effect of Noise in `GeNIe`. To investigate the existence of a sweet-spot in the amount of noise (r), we examine the impact of noise on the performance of the few-shot model, summarizing the findings in Table 6. Our observations indicate that noise levels $r \in [0.7, 0.8]$ yield the best performance. Conversely, utilizing noise levels below 0.7 diminishes performance due label inconsistency, as is demonstrated in Table 5 and Fig 3. As such, determining the appropriate noise level is pivotal for the performance of `GeNIe` to be able to generate challenging hard negatives while maintaining label consistency. To address this, one possible approach could be to leverage additional information from generated images during training as soft labels (e.g., soft cross-entropy loss or ranking loss). We leave the exploration of this avenue to future research.

Why are the generated samples Hard Negatives? Given that the generated images share visual features with the source image, we posit that they should be situated in close proximity to the source image in the embedding space while embodying the semantics of the target class. To provide further insight into our hypothesis, we generate UMAP [53] plots for augmentation samples from various generative methods, as illustrated in Fig 4. Initially, we note that samples generated by `GeNIe` exhibit a more dispersed distribution, spanning the boundaries of classes compared to other augmentation methods. We characterize these samples as hard negatives. Additionally, a reduction in the noise level within `GeNIe` results in more diverse samples, even though this diversity comes with a trade-off in label consistency, as discussed earlier.

5. Conclusion

`GeNIe` leverages diffusion models to generate challenging samples for the target category while retaining some low-level and contextual features from the source image. Our experiments, spanning few-shot and long-tail distribution settings, showcase `GeNIe`'s effectiveness, especially in categories with limited examples. We hope our paper facilitates developing better hard-negative augmentation methods with the advancement of generative AI methods.

References

[1] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné. Associative alignment for few-shot image classification. In *ECCV*, 2019. 3, 5

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. 3

[3] Antreas Antoniou and Amos Storkey. Assume, augment and learn: Unsupervised few-shot meta-learning via random labels and data augmentation. *arxiv:1902.09884*, 2019. 3

[4] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification, 2023. 3, 5

[5] Peyman Bateni, Jarred Barber, Jan-Willem van de Meent, and Frank Wood. Enhancing few-shot image classification with unlabelled examples. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2796–2805, 2022. 6

[6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 7, 13

[7] Jiarui Cai, Yizhou Wang, Jenq-Neng Hwang, et al. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *ICCV*, pages 112–121, 2021. 6

[8] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS*, 32, 2019. 6

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 5, 6

[10] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *ICLR*, 2019. 3, 5

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 5, 6

[12] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. Image deformation meta-networks for one-shot learning. In *CVPR*, 2019. 5

[13] Zhengyu Chen, Jixie Ge, Heshen Zhan, Siteng Huang, and Donglin Wang. Pareto self-supervised training for few-shot learning. In *CVPR*, 2021. 5

[14] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with

a reduced search space. In *Advances in Neural Information Processing Systems*, pages 18613–18624. Curran Associates, Inc., 2020. 2

[15] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *ICCV*, pages 715–724, 2021. 6

[16] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277, 2019. 6

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6

[19] Lisa Dunlap, Clara Mohri, Han Zhang, Devin Guillory, Trevor Darrell, Joseph E. Gonzalez, Anna Rohrbach, and Aditi Raghunathan. Using language to extend to unseen domains. *International Conference on Learning Representations (ICLR)*, 2023. 3

[20] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E. Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation, 2023. 3

[21] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Diversity with cooperation: Ensemble methods for few-shot classification. In *ICCV*, 2019. 3, 5

[22] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning, 2023. 3

[23] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 5

[24] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 2018. 2

[25] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 2

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 7

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 15979–15988. IEEE, 2022. 6

[28] 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. 3, 5

[29] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 2

[30] Yan Hong, Jianfu Zhang, Zhongyi Sun, and Ke Yan. Safa: Sample-adaptive feature augmentation for long-tailed image classification. In *ECCV*, 2022. 6

[31] Kyle Hsu, Sergey Levine, and Chelsea Finn. Unsupervised learning via meta-learning. In *ICLR*, 2018. 3

[32] Sheng-Wei Huang, Che-Tsung Lin, Shu-Ping Chen, Yen-Yi Wu an Po-Hao Hsu, and Shang-Hong Lai. Auggan: Cross domain adaptation with gan-based data augmentation. *European Conference on Computer Vision*, 2018. 3

[33] Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. In *ArXiv preprint arXiv:2206.14754*, 2022. 3

[34] Huiwon Jang, Hankook Lee, and Jinwoo Shin. Unsupervised meta-learning via few-shot pseudo-supervised contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[35] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020. 6

[36] Siavash Khodadadeh, Ladislau Boloni, and Mubarak Shah. Unsupervised meta-learning for few-shot image classification. In *NeurIPS*, 2019. 3, 5

[37] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International Conference on Machine Learning*, pages 5275–5285. PMLR, 2020. 2

[38] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Workshop on 3D Representation and Recognition*, Sydney, Australia, 2013. 7, 13

[39] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier, 2023. 3

[40] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy long-tailed classification. In *CVPR*, pages 6970–6979, 2022. 6

[41] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Adela Barriuso, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations, 2022. 3

[42] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *CVPR*, pages 6949–6958, 2022. 6

[43] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. Adversarial feature hallucination networks for few-shot learning. In *CVPR*, 2020. 3, 5

[44] Mengke Li, Yiu-ming Cheung, Yang Lu, et al. Long-tailed visual recognition via gaussian clouded logit adjustment. In *CVPR*, pages 6929–6938, 2022. 6

[45] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted

supervised contrastive learning for long-tailed recognition. In *CVPR*, pages 6918–6928, 2022. 6

[46] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *ECCV*, 2020. 5

[47] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019. 7

[48] Zicheng Liu, Siyuan Li, Di Wu, Zihan Liu, Zhiyuan Chen, Lirong Wu, and Stan Z Li. Automix: Unveiling the power of mixup for stronger classifiers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 441–458. Springer, 2022. 2

[49] Yuning Lu, Liangjian Wen, Jianzhuang Liu, Yajing Liu, and Xinmei Tian. Self-supervision can be a good few-shot learner. In *European Conference on Computer Vision*, pages 740–758. Springer, 2022. 3, 5, 6

[50] Xue-Jing Luo, Shuo Wang, Zongwei Wu, Christos Sakaridis, Yun Cheng, Deng-Ping Fan, and Luc Van Gool. Camdiff: Camouflage image augmentation via diffusion model, 2023. 3, 5

[51] Lorenzo Luzi, Ali Siahkoohi, Paul M Mayer, Josue Casco-Rodriguez, and Richard Baraniuk. Boomerang: Local sampling on image manifolds using diffusion models, 2022. 3, 4

[52] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 7, 8, 13

[53] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 9

[54] Carlos Medina, Arnout Devos, and Matthias Grossglauser. Self-supervised prototypical transfer learning for few-shot classification. In *ICMLW*, 2020. 3, 5

[55] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3, 4

[56] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021. 6

[57] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 7

[58] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *CVPR*, 2022. 3

[59] Suzanne Petryk, Lisa Dunlap, Keyan Nasseri, Joseph Gonzalez, Trevor Darrell, and Anna Rohrbach. On guiding visual attention with language specification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[60] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, 2018. 3

[61] Tiexin Qin, Wenbin Li, Yinghuan Shi, and Gao Yang. Unsupervised few-shot learning via distribution shift-based augmentation. *arxiv:2004.05805*, 2020. 3

[62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[63] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021.

[64] 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. 3

[65] Sachin Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 5, 6

[66] Mengye Ren, Sachin Ravi, Eleni Triantafillou, Jake Snell, Kevin Swersky, Josh B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *International Conference on Learning Representations*, 2018. 5

[67] 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 (CVPR)*, pages 10684–10695, 2022. 1

[68] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjär¶rn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 4, 5

[69] 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. 3

[70] Swami Sankaranarayanan, Yogesh Balaji, Carlos Domingo Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[71] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogério Schmidt Feris, Raja Giryes, and Alexander M. Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *NeurIPS*, 2018. 5

[72] Viktoriia Sharmanska, Lisa Anne Hendricks, Trevor Darrell, and Novi Quadrianto. Contrastive examples for addressing the tyranny of the majority. *CoRR*, abs/2004.06524, 2020. 3

[73] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Boosting zero-shot classification with synthetic data diversity via stable diffusion. *arXiv preprint arXiv:2302.03298*, 2023. 3, 5

[74] Ojas Kishorkumar Shirekar, Anuj Singh, and Hadi Jamali-Rad. Self-attention message passing for contrastive few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5426–5436, 2023. 3

[75] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 2

[76] Anuj Rajeeva Singh and Hadi Jamali-Rad. Transductive decoupled variational inference for few-shot classification. *Transactions on Machine Learning Research*, 2023. 3

[77] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 5

[78] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? In *ECCV*, 2020. 5

[79] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018. 3, 5

[80] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *NeurIPS*, 33:1513–1524, 2020. 6

[81] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention, 2021. 8

[82] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *ECCV*, 2022. 6

[83] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models, 2023. 3, 5

[84] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[85] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, 2016. 5

[86] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset, 2011. 7, 13

[87] Haoqing Wang and Zhi-Hong Deng. Contrastive prototypical network with wasserstein confidence penalty. In *European Conference on Computer Vision*, pages 665–682. Springer, 2022. 3

[88] Hualiang Wang, Siming Fu, Xiaoxuan He, Hangxiang Fang, Zuozhu Liu, and Haoji Hu. Towards calibrated hypersphere representation via distribution overlap coefficient for long-tailed learning. In *ECCV*, 2022. 6

[89] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X. Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*. OpenReview.net, 2021. 6

[90] Zhirong Wu, Alexei A. Efros1, and Stella X. Yu. Improving generalization via scalable neighborhood component analysis. In *ECCV*, 2018. 5

[91] Yue Xu, Yong-Lu Li, Jiefeng Li, and Cewu Lu. Constructing balance from imbalance for long-tailed image recognition. In *ECCV*, pages 38–56. Springer, 2022. 6

[92] Zhengzhuo Xu, Ruikang Liu, Shuo Yang, Zenghao Chai, and Chun Yuan. Learning imbalanced data with vision transformers, 2023. 6, 7

[93] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *CVPR*, 2020. 3

[94] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8808–8817, 2020. 6

[95] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019. 1, 2, 5

[96] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1, 2, 5

[97] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, pages 2361–2370, 2021. 6

[98] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. *arXiv preprint arXiv:2107.09249*, 2021. 6

[99] 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 *CVPR*, 2021. 3

[100] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *CVPR*, pages 16489–16498. Computer Vision Foundation / IEEE, 2021. 6

[101] Ziqi Zhou, Xi Qiu, Jiangtao Xie, Jianan Wu, and Chi Zhang. Binocular mutual learning for improving few-shot classification. In *ICCV*, 2021. 3

[102] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *CVPR*, pages 6908–6917, 2022. 6

Appendices

A. Visualization of Generative Samples

Additional visuals resembling Fig 2 are presented in Fig A1, and more visuals akin to Fig 3 can be found in Fig A2.

B. Details of Fine-grained Dataset

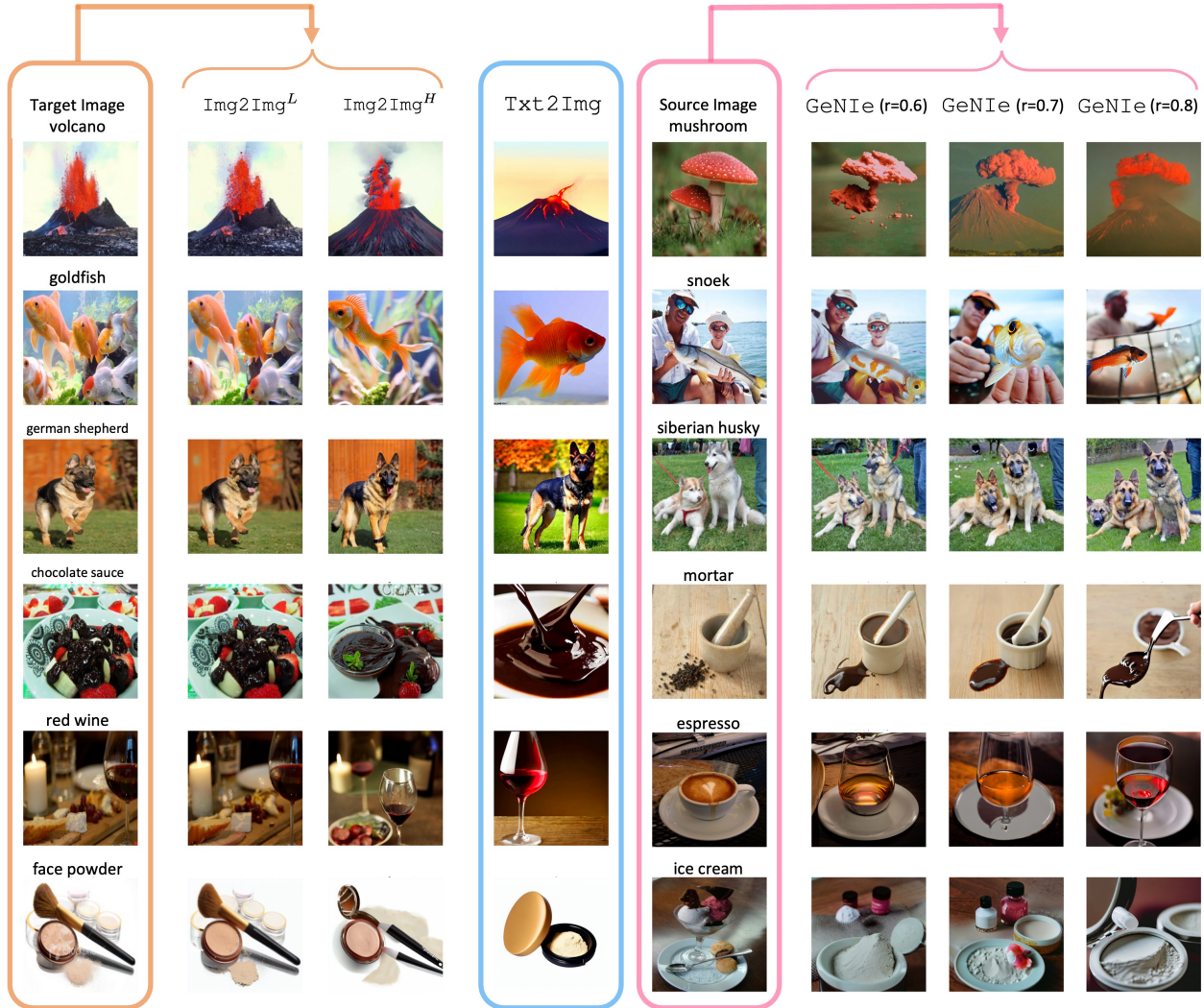We provide detailed information for each Fine-grained dataset in Table A1, as discussed in Section 4.3.

Figure A1. **Visualization of Generative Samples:** We show more visualizations similar to Figure 2. We compare GeNIe with two baselines: **Img2Img^L augmentation** uses both image and text prompt from the same category, resulting in less challenging examples. **Txt2Img augmentation** generates images based solely on a text prompt, potentially deviating from the task's visual domain. **GeNIe augmentation** incorporates the target category name in the text prompt along with the source image, producing desired images with an optimal amount of noise, balancing the impact of the source image and text prompt.

Dataset	Classes	Train samples	Test samples
CUB200 [86]	200	5994	5794
Food101 [6]	101	75750	25250
Cars [38]	196	8144	8041
Aircraft [52]	41	6,667	3333

Table A1. We list the sizes of train and test splits of the Fine-grained dataset. We use the provided train set for few-shot sampling, and the provided test sets for our evaluation. For the Aircraft dataset we use manufacturer hierarchy.
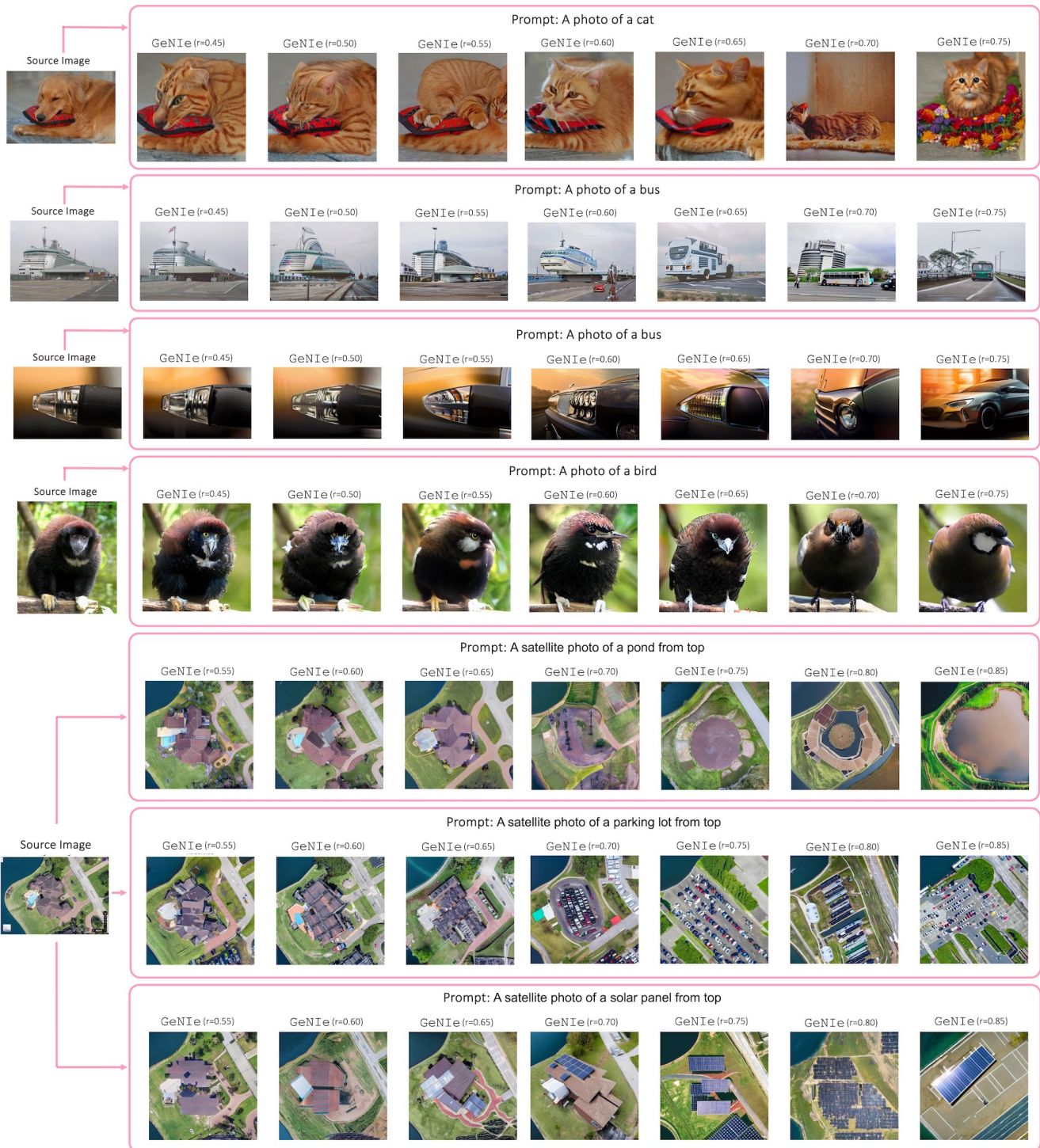
Figure A2. **Effect of noise in GeNIe:** Just like in Fig 3, we use GeNIe to create augmentations with varying noise levels. As illustrated in the example above, a reduced amount of noise leads to images closely mirroring the semantics of the source images, causing a misalignment with the intended target label.