Cap2Aug: Caption guided Image data Augmentation

Aniket Roy[1], Anshul Shah*[1], Ketul Shah*[1], Anirban Roy[2], Rama Chellappa[1]
[1]Johns Hopkins University, [2]SRI International

Abstract

Visual recognition in a low-data regime is challenging and often prone to overfitting. To mitigate this issue, several data augmentation strategies have been proposed. However, standard transformations, e.g., rotation, cropping, and flipping provide limited semantic variations. To this end, we propose Cap2Aug, an image-to-image diffusion model-based data augmentation strategy using image captions to condition the image synthesis step. We generate a caption for an image and use this caption as an additional input for an image-to-image diffusion model. This increases the semantic diversity of the augmented images due to caption conditioning compared to the usual data augmentation techniques. We show that Cap2Aug is particularly effective where only a few samples are available for an object class. However, naively generating the synthetic images is not adequate due to the domain gap between real and synthetic images. Thus, we employ a maximum mean discrepancy loss to align the synthetic images to the real images to minimize the domain gap. We evaluate our method on few-shot classification and image classification with long-tail class distribution tasks. Cap2Aug achieves state-of-the-art performance on both tasks while evaluated on eleven benchmarks. Code: https://github.com/aniket004/Cap_2_Aug.git

1. Introduction

Supervised image classification approaches have achieved near-human performance [19, 26] by leveraging large-scale datasets [8, 12]. However, learning from limited data remains challenging, such as in few-shot setups, where only 1-5 samples could be available for each class. To address this challenge, existing approaches consider various data augmentation approaches to expand the training set. For example, [23] generates pseudo labels for the base class samples and uses these samples to increase the number of novel class samples. Assoalign [2] uses base-class samples in addition to the novel class samples to generate new samples in an adversarial framework. [40] use hard-mixup to combine existing samples to generate additional samples.

*denotes equal contribution.

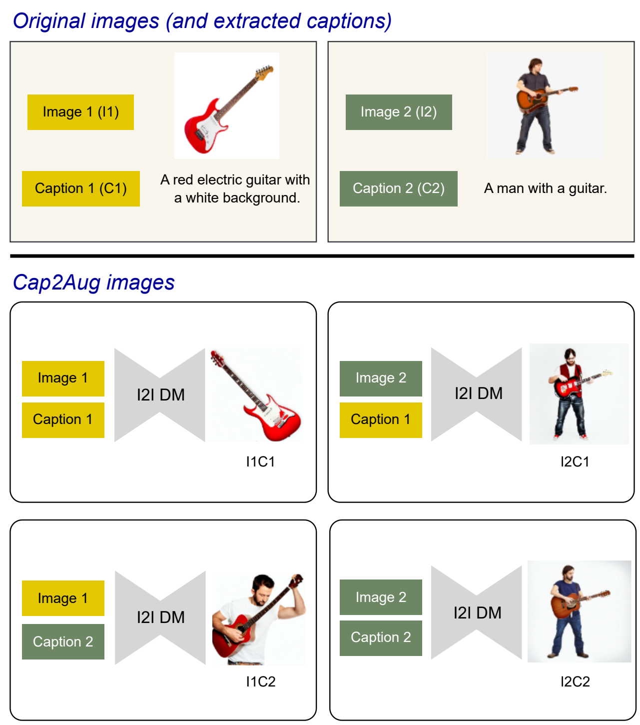

Figure 1. Idea of Cap2Aug: real images I_1 (guitar) and I_2 (person playing guitar) are fed to a captioning model to generate "a red electric guitar with a white background" (C_1) and "a man with a guitar" (C_2) as captions, respectively. Image I_1 and caption C_1 when fed to an image-to-image diffusion model generates a synthetic image I_1C_1 - the image of a guitar similar to I_1 with minor changes (fine changes in guitar head, body). On the other hand, when image I_1 and caption C_2 are fed to the image-to-image diffusion model, it generates a man with a guitar in his hand (I_1C_2). This is replicated using image I_2 for generating synthetic images I_2C_1 and I_2C_2 respectively using captions C_1 and C_2.

Recently, generative models, such as DALL-E [38] and stable diffusion [39], are shown to be successful in generating realistic images. The vision and language models, such as CLIP [37] and BLIP [27], can effectively capture detailed visual cues from images in the form of captions. In this context, we investigate the following question: "Can these large vision language models be leveraged to generate semantically diverse augmented images?" Inspired by the efficacy of these generative models, we develop Cap2Aug - a data augmentation strategy that provides semantic variations in the augmented samples aided by generative models. We first generate a set of captions from an image using a captioning model. Then we use text-conditioned image-to-image diffusion models with these captions as prompts to create additional images. This results in a semantically diverse set of augmented images that can be used for training. The idea is also motivated by "back-translation" [14], an effective data augmentation strategy used in natural language processing, where a sentence is translated to a different language, and then back-translated to the same language providing an augmented version of the sentence itself. Cap2Aug performs back-translation across image-text modalities, which is simple yet effective. Finally, the classifier can be trained with the augmented dataset. We present the Cap2Aug framework in Fig. 1. While the generated images can be directly used to augment the training set, we notice that this is suboptimal due to the domain gap between real and generated images. Thus, we propose a maximum mean discrepancy (MMD) loss [28] to align the features of the synthetic images to real images for better performance.

Cap2Aug leverages generative models that are trained on large-scale datasets. Thus, our approach is not directly comparable to few-shot approaches [3, 4, 23, 40] that do not consider external sources of supervision. Our goal is to develop a data-augmentation framework that leverages existing generative models. Thus, Cap2Aug can be compared to existing approaches [55, 57] that use additional datasets or models to improve classification performance. We primarily consider image classification in a few-shot setup to evaluate our approach. Cap2Aug is also shown to be effective for image classification with long-tail distribution over the classes. Thus, our contributions include:

- We propose Cap2Aug - a simple, training-free, plug-and-play data augmentation strategy leveraging image-to-image generative models with image captions as text prompts. We validate this approach for long-tail and few-shot classification tasks. Cap2Aug is particularly effective for few-shot setups where only a few training images are available.
- We use an MMD-based loss function to align synthetic images to real images to reduce the domain gap between real and synthetic images.
- We validate our approach on standard long-tail classification on ImageNet-LT and eleven few-shot classification benchmarks and achieve improvements over the state-of-the-art.

2. Related Work

Multi-modal few-shot learning. Semantic information is useful for few-shot classification [1]. Padhe et al. [33] use multi-modal prototypical networks for few-shot classification and Yang et al. [50] utilize semantic guided attention to integrate the rich semantics into few-shot classification. [49] generates representative samples for few-shot learning using text-guided variational autoencoder. Wang et al. [47] uses multi-directional knowledge transfer for multi-modal few-shot learning. Text-guided prototype completion [52] also helps few-shot classification.

Vision-language models. Recent advancements in large-scale vision language pretrained models enable significant improvements in multi-modal learning with CLIP [37], GPT-3 [7], DALLE [38], stable diffusion [39] etc. Diffusion models are state-of-the-art text-to-image generative models [22, 31, 38, 39, 41], which are trained on large-scale image and text corpus and produce surprisingly high-quality images just from texts. The vision-language pretraining model CLIP [37] helps to improve zero-shot performance across several datasets. Prompt tuning method CoOp [57] optimizes learnable prompts for better few-shot adaptation. CoCoOp [56] and VT-CLIP [54] used a text-conditioned intermediate network for joint image-text training. The CLIP-adapter [17] uses the powerful CLIP features with a lightweight residual style network adapter for few-shot adaptation. The Tip-Adapter [55] extends this using a training free key-value based cache model and obtained a performance boost. CALIP [18] uses parameter-free attention to elevate CLIP performance in both zero-shot and few-shot settings. SuS-X [46] extends the Tip-adapter using image-text distance and dynamic support set. Recently, several methods [5, 13, 15, 20, 34, 36, 43, 53, 58] have been proposed to use synthetic data to improve visual recognition tasks. In contrast to these works, we use an effective way of using the pretrained vision-language model for effective data augmentation and also investigate the real to synthetic domain adaptation issue.

3. Proposed Approach

In this section, we describe a simple, training-free, plug-and-play data augmentation strategy (Fig. 2). Cap2Aug provides semantic diversity through the collaboration of pretrained captioning and image-to-image diffusion models. The steps of the augmentation scheme are: 1) Generate captions from the images using a pretrained caption model, 2) Generate synthetic augmentations of the images using pretrained text-guided image-to-image diffusion model, where captions from previous step are provided as text prompts.

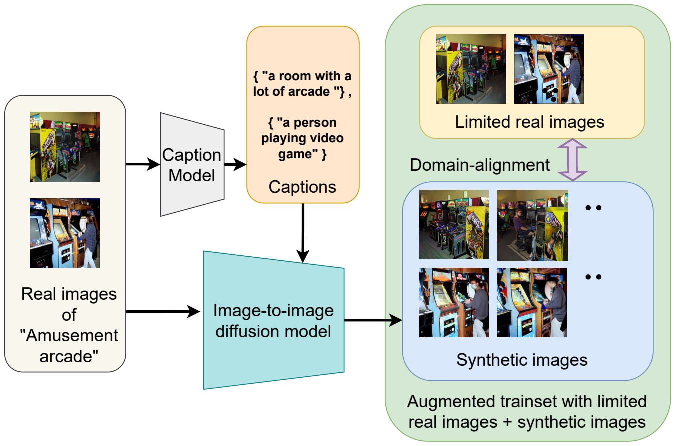

Figure 2. Overview of Cap2Aug. We generate captions from the real images using the BLIP caption model [27]. The generated captions and real images are fed to the image-to-image diffusion model [39] to generate plenty of synthetic images. The combined set of limited real images and abundant synthetic images are used to learn a classifier for the novel class. We also align the synthetic images with the real images to reduce the domain gap using MMD.

3.1. Image-to-text generation using Captioning

Captions capture the semantic information of images with succinct texts. Current large-scale vision and language methods, e.g., CLIP-based [37], BLIP [27] achieve impressive performance in image captioning. We capture the diversity and class-specific information residing in the few training examples by captioning the images using off-the-shelf BLIP caption model [27]. Now, using the BLIP caption model, we generate captions C_i for each image I_i in the training set.

$$C_i = \textit{Caption model}(I_i) \tag{1}$$

Captions provide diverse class-dependent semantic information across samples. For example, in Sun397 dataset [48], there exists a class "youth hostel", which contains images of a group of people sitting on a bed and a couple of bunk beds as shown in Fig. 3. These are typical characteristics of the "youth hostel" class. We see that the generated captions capture the semantic characteristics of the class information as shown in Fig. 3.

3.2. Text-to-image generation with caption guidance

Traditional image augmentation methods rely on fixed transformations e.g., translation, rotation, etc. To the contrary, we generate an augmented version of images using an image-to-image diffusion model by editing these images using captions.

Stable diffusion [39] is a diffusion model conditioned on text embedding of CLIP ViT-L/14 [37] text encoder and trained on LAION-400M dataset [39] of image-text pairs. Prior works have leveraged this model to generate realistic images from textual descriptions. In this work, the stable diffusion model is conditioned on the text prompts that are based on the diffusion-denoising mechanism proposed by SDEdit [30]. The method generates images by iterative denoising through a stochastic differential equation conditioned on the encoded version of the text prompt. Examples of the input image and caption pairs and corresponding generated images are shown in Fig. 1.

Now, the augmented versions of the images are generated by,

$$I_iC_j = \mathit{I2I}(I_i, C_j) \quad for \qquad i, j = 1, .., K \tag{2}$$

where $I2I$ is the pretrained image-to-image diffusion model, I_i is the i-th image and the corresponding caption is denoted by C_i. In an N-way K-shot classification problem, for each class, we have K training images I_1, I_2,... I_K. The corresponding captions generated by the BLIP-caption model are C_1, C_2, ... C_K, respectively. Now, we can generate diverse images pairing (I_i, C_j) denoted by I_iC_j for $i, j = 1, .., K$ as shown in Fig. 1.

3.2.1 Cross image-caption pair generation

The generated images with *self-captions* i.e., using image-caption pairs (I_i, C_i) are denoted by I_iC_i. These images I_iC_i generated using captions from the image itself would still result in some style or content difference in the image. In Fig. 1 (I_1C_1), image-to-image translation of the "guitar image" with its own caption (i.e., "a red electric guitar with a white background"), still generates an image with a different semantic content (i.e., the difference in the guitar head). Hence, this can also be considered as a useful augmentation.

More interesting and diverse images are generated by cross image-caption pairs (I_i, C_j) $(i \neq j)$, where the style of image caption C_j is translated to generated images from I_i through image-to-image stable diffusion model. For instance, an image of a guitar is translated to a person playing a guitar using the caption "a man with a guitar" as shown in Fig. 1 (I_1C_2).

We generate augmented versions of the training images conditioned on the class information captured by captions. Our objective is to provide semantic variations of the existing training images, not generating new samples using the off-the-shelf generative models. Note that, we are not explicitly using the class labels for generating the images. Since the diffusion models are trained on large-scale datasets, therefore generating images using class labels might violate the inherent problem of low-data regime e.g., long-tail or few-shot setting.

3.3. Domain alignment of real and synthetic images using MMD.

Despite the high-quality of synthetic images, there exists a domain gap between real and synthetic images in terms of the background, color, and intensity distribution as shown in Fig. 4. The average color histograms of the real and synthetic images are shown in Fig. 5, which exhibits a distinction between these sets of images. To reduce the domain gap, we use a multi-kernel Maximum Mean Discrepancy (MMD) [28] loss, which minimizes the domain gap by reducing the distance of the mean feature embeddings of the real and synthetic images.

Let's assume, given a source domain ($\mathcal{D}_s$) and the target domain ($\mathcal{D}_t$), samples are drawn from these domains with distributions P and Q, respectively over a set $\mathcal{X}$. The features of the samples from these domains are denoted as $\{z_i^s\}$ and $\{z_i^t\}$, respectively. A multi-kernel MMD ($D_k(P, Q)$) between probability distributions P and Q is defined as [28]: $D_k(P, Q) = \|\mathbb{E}_p[\psi(z^s)] - \mathbb{E}_q[\psi(z^t)]\|^2_{\mathcal{H}_k}$ where k is the kernel function in the functional space, i.e., $k = \sum_{p=1}^{P} \alpha_p k_p$, where k_p is a single kernel. The feature map $\psi : \mathcal{X} \rightarrow \mathcal{H}_k$ maps into a reproducing kernel Hilbert space. $k = \{\mathcal{N}(0, 0.5), \mathcal{N}(0, 1), \mathcal{N}(0, 2)\}$. If the kernel is $k(x, y) =< \psi(x), \psi(y) >_{\mathcal{H}_k}$, then using the kernel trick, MMD can be estimated without directly learning $\psi(\cdot)$ as:

$$\bar{D}_k(P, Q) = \frac{1}{n_s^2} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} k(z_i^s, z_j^s) + \frac{1}{n_t^2} \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} k(z_i^t, z_j^t) - \frac{2}{n_s n_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} k(z_i^s, z_j^t) \quad (3)$$

Therefore, the MMD-loss between the real examples (I_{NK}) and synthetic examples ($I_{NK'}$) will be,

$$\mathcal{L}_{MMD} = \bar{D}_k(I_{NK}, I_{NK'}) \quad (4)$$

Finally, the model will be learned based on the task-specific classification loss (i.e, cross-entropy loss $\mathcal{L}_{CE}$) and the MMD loss ($\mathcal{L}_{MMD}$).

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha * \mathcal{L}_{MMD} \quad (5)$$

The scaling parameter α is set experimentally and ablation on this parameter is shown in the experiments section.

4. Experiments

We evaluate the proposed approach on two tasks: 1) few-shot classification and 2) long-tail classification.

4.1. Few-shot classification

Data augmentation is crucial in a data scarce regime. Hence, we validate our augmentation strategy for the few-shot classification task. We perform few-shot experiments

(a) Caption: a group of people are sitting in a room with bunks.

(b) Caption: A room with bunks and table.

(c) Caption: a room with two beds and a small bed in the middle in bronx, ny

(d) Caption: a woman in a room with a bunk and a bed

Figure 3. Illustration of image captioning using BLIP model: Images from the class "youth hostel". Several images capture the characteristics of the class "youth hostel". For example, images contain a bunk bed and a group of people sitting as shown in the captions generated from the images.

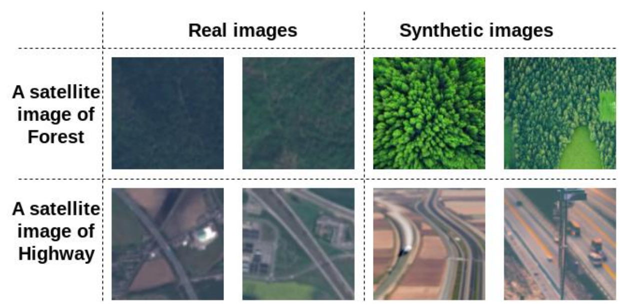

Figure 4. Real images from the original dataset (left) and synthetic images generated using Cap2Aug (right) have different color, intensity and background distributions.

on eleven benchmark datasets - ImageNet-1K [12], StanfordCars [25], UCF101 [44], Flowers102 [32], SUN397 [48], DTD [9], EuroSAT [21], FGVCAircraft [29], OxfordPets [35], Food-101 [6] and Caltech-101 [16]. We follow the protocol of Tip-Adapter [55] to train models with 2, 4, 8, and 16 shots and test on the full test set. Following standard practice, we consider classification accuracy as the metric. For a fair comparison with Tip-Adapter [55], we use CLIP [37] with ResNet-50 as the visual encoder. On top of the feature extractor, an adapter is initialized as a 2-layer MLP with cache keys as learnable parameters. We train the adapter using an AdamW optimizer with an initial learning rate of 0.001 with a cosine scheduler. For generating image captions, we use the open-source implementation of BLIP-caption generator [27] provided in diffusers library from HuggingFace. We also use the same library for generating images from image-to-image stable diffusion model with the "stable-diffusion-v1-5" model. For image-to-imag diffusion model, we use SDEdit [30]. More details are provided in the supplementary material. For the N-way K-shot setup (i.e., N classes and K shots), Tip-Adapter uses NK images. In Cap2Aug, we augment K images of each class with p (p $\leq$ K) different captions, therefore, generating NKp synthetic images. In total, NK(p+1) images are used for training (i.e., NK real, NKp synthetic). To ensure the same number of images seen during training in both approaches, we train Tip-Adapter $(p+1) \times E$ epochs, while training Cap2Aug for E epochs.

We compare our method with state-of-the-art Tip-Adapter [55] and CoOp [57], in Table. 3, Table. 6, Table. 13, Table. 7, and Table. 8 for few-shot classification tasks on eleven different benchmarks. We also compare with naive data augmentation methods e.g., random crop, resize, flip, color saturation, and observe the generative model guided augmentations perform significantly better. We perform experiments with three random seeds and the mean and standard deviation are reported. Our method consistently outperforms state-of-the-art in most cases including the challenging fine-grained classification datasets.

The two recent and relevant baselines are - SuS-X [46], TaskRes [51]. In Table. 1, we have compared with the 16-shot classification performance with respect to SuS-X [46], in various datasets, e.g., ImageNet, Food-101, OxfordPets, Caltech-101, Flowers-102, FGVC. Our approach outperforms the baselines across the datasets. We have also compared TaskRes [51] against our approach on different datasets and different settings in Table. 2. Our approach outperforms TaskRes across dataset and settings as evident from Table. 2.

Table 1. Comparison with SuS-X (16-shot)

Dataset	SuS-X-LC [46]	TIP-X [46]	Ours
ImageNet	61.89	<u>62.16</u>	**66.43**
Food-101	<u>77.62</u>	75.96	**79.32**
OxfordPets	86.59	<u>87.52</u>	**90.11**
Caltech-101	89.65	<u>90.39</u>	**92.93**
Flowers-102	67.97	<u>90.54</u>	**95.21**
FGVC	21.09	<u>29.61</u>	**34.92**

4.1.1 Complexity analysis

We have performed the comparison of the number of parameters and time complexity in Tab. 5. In Tab. 5, we present the results on ImageNet 16-shot experiment and compare them with ZeroShot CLIP [37], Zero-Shot CALIP [18], CoOp [57], CLIP-Adapter [17], Tip-Adapter [55] and observe that our method improves performance with a small increase in model training time.

Table 2. Comparison with TaskRes [51]

Approach	Dataset	2-shot	4-shot	8-shot	16-shot
TaskRes	ImageNet	62.17	62.93	64.03	64.75
Ours	ImageNet	**62.83**	**63.74**	**64.98**	**66.43**
TaskRes	OxfordPets	84.43	86.27	87.07	88.10
Ours	OxfordPets	**87.67**	**88.12**	**88.60**	**90.11**
TaskRes	Food-101	75.30	76.23	76.90	78.23
Ours	Food-101	**77.86**	**78.05**	**78.62**	**79.32**
TaskRes	FGVC	23.07	24.83	29.50	33.73
Ours	FGVC	**23.88**	**25.11**	**29.86**	**34.92**
TaskRes	SUN	64.33	66.67	68.70	70.30
Ours	SUN	**64.72**	**67.39**	**69.10**	**71.20**
TaskRes	EuroSAT	65.77	72.97	77.07	82.57
Ours	EuroSAT	**67.35**	**77.27**	**77.82**	**83.77**

Table 3. Comparison on ImageNet (best results in **bold**, second best in <u>underline</u>, improvement in Δ), DA means data augmentation.

Method	2-shot	4-shot	8-shot	16-shot
Tip [55]	60.96	60.98	61.45	62.03
CoOp [57]	50.88	56.22	59.93	62.95
Tip-F [55]	61.69	62.52	64.00	65.51
Tip-F [55] + Naive DA	61.73	62.56	64.06	<u>65.61</u>
TaskRes [51]	<u>62.17</u>	<u>62.93</u>	<u>64.03</u>	64.75
Ours	**62.83** $\pm$ **0.3**	**63.74** $\pm$ **0.2**	**64.98** $\pm$ **0.2**	**66.43** $\pm$ **0.3**
Δ	+0.66	+0.81	+0.95	+0.82

Table 4. Ablation of Number of Synthetic images (K) on ImageNet

K	4	16	40	80
2-shot	62.7	63.1	63.6	64.2
4-shot	63.0	63.5	63.9	64.5
8-shot	63.4	64.1	64.9	65.6
16-shot	64.1	64.6	65.7	66.3

Table 5. Training parameter and complexity analysis

Method	#Param (million)	Train time	Acc.
ZS CLIP [37]	0	0	60.33
ZS CALIP [18]	0	0	60.57
CoOp [57]	0.02	14h 40 min	62.95
CLIP-Adapter [17]	0.52	50 min	63.59
Tip-Adapter-F [55]	6.2	13 min	64.59
Ours	6.2	15 min	**66.43**

4.1.2 Ablation studies

We conduct an ablation study on the novel components of our method in Table. 14. As expected, adding synthetic images generated by the diffusion model and MMD improves the performance of EuroSAT, and Oxford-flowers datasets in low-data settings. MMD seems to be particularly helpful in extremely low data cases (e.g., 2 shot) as evident from Table. 14. We also provide the ablation of the MMD loss coefficient α in Table. 15. It appears that for low-shot cases, higher α works better. Ablation on different backbones and the number of generated synthetic images for few-shot classification on ImageNet have been provided in Tab. 12 and Tab. 4 respectively. Additionally, we also perform ablation studies on diffusion and caption models (Tab. 9), and diffusion guidance scale (Tab. 10).

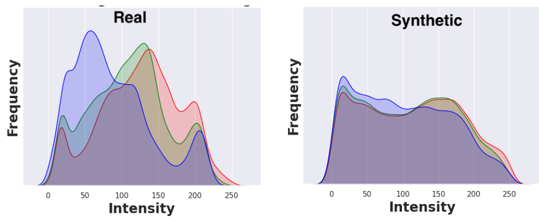

Figure 5. Color histogram of real (left) and synthetic samples (right)

4.1.3 Why MMD?

Diffusion model generated images have prominent domain differences, e.g., color, saturation (as shown in Fig. 4) w.r.t the real images. We improve the classification performance by minimizing the domain discrepancy using MMD loss. MMD is simple, effective and easy to integrate to any training pipeline. In addition to the generated synthetic images, using MMD further improves performance especially in few-shot settings for EuroSAT images as shown in Tab. 14 (EuroSAT). For EuroSAT 4-shot classification, the class-wise accuracy for "forest" class was 72%, adding synthetic samples improves the accuracy to 74%, but there is significant domain gap w.r.t color (Fig. 4). Adding MMD improves the accuracy of that class to 78%.

4.1.4 Is generated caption needed?

Generic text as "class names" can be used, but is not sufficient to capture the diversity in the images. For EuroSAT 4-shot classification, only generating images from class names provide an accuracy of 75.00%, while generating images from captions improves the accuracy to 77.37% (+2.37%).

4.2. Long-tail classification

Long-tail classification has both data-scarce and data-abundant classes, therefore is a good test case for validating our data augmentation strategy. We conduct experiments on large-scale long-tailed ImageNet-LT benchmark and obtain performance improvements over SOTA [45] using Cap2Aug data augmentation as shown in Tab. 11. We provide results for overall accuracy, many-shot (100 samples), medium-shot (20-100 samples), and few-shot (20 samples) cases. In this experiment, 40 images are generated for all the classes and used those as augmented data. We observe the performance gain is higher for few-shot classes in Tab. 11. Approach specific details are provided in the supplementary material.

4.3. Qualitative results

We show examples where image-to-image generation using caption provides diverse training examples and thus helps provide generalization. E.g., in Fig. 6 the real image, showing a person playing guitar, and the caption "person playing guitar" generates images of different people playing guitar, which helps the model to focus more on "playing guitar" (actual class label), than people or background. Similarly, diverse examples for "bee-hive" and "pizza" classes are generated by the image and the corresponding captions in Fig. 6.

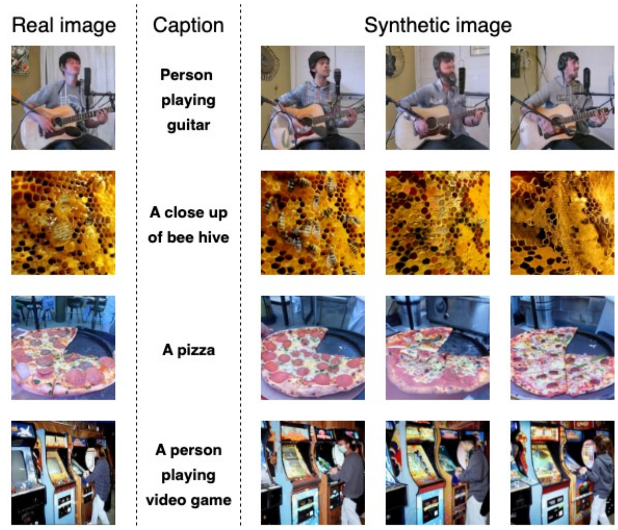

Figure 6. Image to image generation using captions.

4.4. Discussions

Our method attempts to capture the variations within the class through captions and translate that to generate diverse augmented samples from the training samples using image-to-image diffusion model. For instance, in Fig. 8 (first row) the training image is a picture of a man having his haircut and the corresponding classname is "haircut" (from UCF101 dataset). If we provide a caption "a woman is getting her haircut" to this image and fed it to the image-to-image diffusion model, it indeed generates an image of a woman having a haircut (second row, right figure). Therefore, such cross-caption-based image generation provides diversity in the training set and help generalization. Similarly, in the last row (Fig. 8) using caption as "a person in a arcade" to an image of arcade generates image of an arcade with a person in it, providing more diverse and natural augmented instances.

4.5. Analyzing Bias of diffusion model

Our method is more effective where the caption model generates diverse captions across classes, e.g., in EuroSAT dataset, the classes are quite distinctive, e.g., forest, highway, etc., and therefore the caption model is able to generate descriptive captions and diffusion model generates appropriate images, and the overall performance improves (Fig. 4, Table. 6). However, in the case of FGVC dataset, where the classes are different fine-grained airplane classes denoted by their names, with fine difference in details. In that case,

Table 6. Comparison on EuroSAT, SUN397 and UCF101 (best results in **bold**, second best in underline, DA means data augmentation)

Shots	EuroSAT				SUN397				UCF101			
	2	4	8	16	2	4	8	16	2	4	8	16
Tip [55]	61.68	65.32	67.95	70.50	62.70	64.15	65.62	66.85	64.74	66.46	68.68	70.58
CoOp [57]	61.50	70.18	76.73	82.53	59.48	63.47	65.52	69.26	64.09	67.03	71.92	75.71
Tip-F [55]	66.15	74.12	77.30	82.54	63.64	66.21	68.87	70.47	66.43	70.55	74.01	77.03
Tip-F [55] + Naive DA	<u>66.21</u>	<u>74.32</u>	<u>77.53</u>	<u>82.71</u>	<u>63.82</u>	<u>66.29</u>	<u>68.94</u>	<u>70.53</u>	<u>66.55</u>	<u>70.68</u>	<u>74.25</u>	<u>77.22</u>
Ours	**67.35**$\pm$**0.2**	**77.27** $\pm$ **0.3**	**77.82** $\pm$ **0.2**	**83.77** $\pm$ **0.3**	**64.72** $\pm$ **0.2**	**67.39** $\pm$ **0.3**	**69.10** $\pm$ **0.3**	**71.20** $\pm$ **0.2**	**68.77** $\pm$ **0.3**	**71.68** $\pm$ **0.2**	**74.72** $\pm$ **0.2**	**77.63** $\pm$ **0.3**
Δ	+1.14	+2.95	+0.29	+1.06	+0.90	+1.10	+0.16	+0.67	+2.22	+1.00	+0.47	+0.41

Table 7. Comparison on OxfordPets, OxfordFlowers and FGVC (best results in **bold**, second best in underline, DA means data augmentation)

Shots	OxfordPets				OxfordFlowers				FGVC			
	2	4	8	16	2	4	8	16	2	4	8	16
Tip [55]	87.03	86.45	87.03	88.14	79.13	83.80	87.98	89.89	21.21	22.41	25.59	29.76
CoOp [57]	82.64	86.70	85.32	87.01	77.50	85.20	90.18	94.51	18.68	21.87	26.13	31.26
Tip-F [55]	87.03	87.54	88.09	89.70	82.30	85.83	90.51	94.80	23.19	24.80	29.21	34.55
Tip-F [55] + Naive DA	<u>87.22</u>	<u>87.73</u>	<u>88.26</u>	<u>89.88</u>	<u>82.51</u>	<u>85.98</u>	<u>90.69</u>	<u>94.97</u>	<u>23.42</u>	<u>24.95</u>	<u>29.39</u>	<u>34.76</u>
Ours	**87.67** $\pm$ **0.3**	**88.12** $\pm$ **0.2**	**88.60** $\pm$ **0.2**	**90.11** $\pm$ **0.2**	**83.23** $\pm$ **0.3**	**86.83** $\pm$ **0.4**	**91.37** $\pm$ **0.3**	**95.21** $\pm$ **0.2**	**23.88** $\pm$ **0.2**	**25.11** $\pm$ **0.3**	**29.86** $\pm$ **0.2**	**34.92** $\pm$ **0.3**
Δ	+0.45	+0.39	+0.34	+0.33	+0.72	+0.85	+0.68	+0.24	+0.46	+0.16	+0.47	+0.16

Table 8. Comparison on Caltech101

Method	Caltech101			
	2-shot	4-shot	8-shot	16-shot
Tip [55]	88.44	89.39	89.83	90.18
CoOp [57]	87.93	89.55	90.21	91.83
Tip-F [55]	89.74	90.56	91.00	91.86
Tip-F [55] + Naive DA	<u>89.82</u>	<u>90.63</u>	<u>91.12</u>	<u>91.93</u>
Ours	**90.11** $\pm$ **0.2**	**90.97** $\pm$ **0.2**	**91.54** $\pm$ **0.2**	**92.93** $\pm$ **0.3**

Table 9. Ablation on diffusion and caption model

Method	EuroSAT			
	2-shot	4-shot	8-shot	16-shot
SD1.5 + BLIP-2	67.35	72.27	77.82	83.77
SD1.5 + LLaVA	69.18	78.56	79.33	85.22
SDXL + BLIP-2	72.53	80.11	82.45	88.67
SDXL + LLaVA	**75.31**	**83.12**	**84.96**	**90.23**

Table 10. Ablation on guidance scale of diffusion model

Guidance scale	EuroSAT			
	2-shot	4-shot	8-shot	16-shot
5	65.11	75.88	76.15	81.82
7.5	**67.35**	**77.27**	**77.82**	**83.77**
10	66.73	76.32	76.92	82.17

Table 11. Comparison on ImageNet-LT (best results in **bold**, second best in underline, improvement in Δ)

Method	Overall Acc.	Many-shot	Medium-shot	Few-shot
ResLT [10]	55.1	63.3	53.3	40.3
PaCo [11]	60.0	68.2	58.7	41.0
LWS [24]	51.5	62.2	48.6	31.8
DRO-LT [42]	53.5	64.0	49.8	33.1
VL-LTR [45]	<u>70.1</u>	<u>77.8</u>	<u>67.0</u>	<u>50.8</u>
Ours	**70.9**	**78.5**	**67.7**	**51.9**
Δ	+0.8	+0.7	+0.7	+1.1

Table 12. Various backbones for ImageNet 16-shot classification

Method	RN50	RN101	ViT/32	ViT/16
Tip-F [55]	65.51	68.56	68.65	73.69
Ours	**66.32**	**69.20**	**69.70**	**74.70**

the caption model is unable to capture diverse class information, and therefore, diffusion model generates similar images across classes and hence the overall performance doesn't improve much (Fig. 7, Table. 7).

4.6. Limitations

While we see improvements over prior works on classification tasks, our results indicate that this approach might not be suitable for fine-grained classification, e.g., FGVC, Food101 datasets. One potential reason could be that captions are unable to extract the fine-grained details which could be important for fine-grained recognition. E.g., in Fig. 7 top row, the airplane is E-195, which has more fine-grained characteristics (e.g., the shape of the plane and wings), than what the caption captures (i.e., "a white and blue jet"). The synthetic images might confuse with other fine-grained airplane categories and thus degrade performance. Similarly, in the Food101 dataset, the class "samosa" (Fig. 7 (second row) is miscaptioned as "plate of chicken wings", therefore the generated images are not semantically helpful for classifying food items. For the fine-grained pet recognition task, captions are unable to distinguish pet categories, i.e., "a small dog" does not differentiate across pet species and therefore our model fails in these cases. We would like to address these limitations in future work.

5. Conclusion

We have proposed Cap2Aug - a data augmentation approach exploiting the image-to-image generative model using captions. Compared to traditional data augmentation strategies, our proposed augmentation method utilizes semantic information in the images, captured by image captions. Our study has shown that the domain gap between real and synthetic images can pose additional challenges. To mitigate this, we have proposed a multi-kernel MMD-based loss function to align synthetic images to real images. We have validated our approach for long-tail and few-shot classification tasks. For long-tail classification on the standard ImageNet-LT benchmark, Cap2Aug improves over SOTA methods. Our method outperforms the state-of-the-art approaches on few-shot classification on eleven benchmarks. We have performed ablation studies to justify the contribu-

Table 13. Comparison on StanfordCars, Food101 and DTD (best results in **bold**, second best in underline, DA means data augmentation)

	StanfordCars				Food101				DTD			
Shots	2	4	8	16	2	4	8	16	2	4	8	16
Tip [55]	57.93	61.45	62.90	66.77	77.52	77.54	77.76	77.83	49.47	53.96	58.63	60.93
CoOp [57]	58.28	62.62	68.43	73.36	72.49	73.33	71.82	74.67	45.15	53.49	59.97	63.58
Tip-F [55]	61.10	64.50	68.25	74.15	77.60	77.80	78.10	79.00	53.72	57.39	62.70	65.50
Tip-F [55] + Naive DA	61.18	64.60	68.32	74.21	77.68	77.87	78.19	79.06	53.79	57.44	62.75	65.57
Ours	**61.45** $\pm$ **0.2**	**65.00** $\pm$ **0.3**	**69.25** $\pm$ **0.2**	**74.85** $\pm$ **0.2**	**77.86** $\pm$ **0.2**	**78.05** $\pm$ **0.2**	**78.62** $\pm$ **0.2**	**79.32** $\pm$ **0.1**	**54.55** $\pm$ **0.2**	**59.38** $\pm$ **0.2**	**63.49** $\pm$ **0.2**	**66.33** $\pm$ **0.3**
Δ	+0.15	+0.20	+0.90	+0.65	+0.06	+0.09	+0.37	+0.05	+0.78	+1.89	+0.77	+0.63

Table 14. Ablation Study on contributions

	EuroSAT			OxfordFlowers		
Shots	2	4	8	2	4	8
Tip-F [55]	66.15	74.12	77.30	82.30	85.83	90.51
Tip-F + Syn	66.80 (+0.65)	75.93 (+1.81)	77.30 (+0.0)	**82.86** (+0.56)	86.19 (+ 0.36)	90.89 (+ 0.38)
Tip-F + Syn + MMD	**67.03** (+0.23)	**77.37** (+1.44)	**77.50** (+0.20)	83.06 (+0.20)	**86.64** (+0.45)	**91.44** (+0.55)

Table 15. Ablation on MMD coefficient α

	EuroSAT				SUN397				UCF101			
α	2	4	8	16	2	4	8	16	2	4	8	16
0	66.08	75.93	76.45	**83.64**	64.46	**67.45**	68.91	70.88	67.90	**71.76**	73.56	73.77
0.01	65.86	73.50	77.08	83.02	64.30	**67.45**	68.90	**70.90**	68.27	**71.76**	73.51	77.21
0.1	65.29	76.64	77.38	82.75	64.31	**67.45**	68.63	70.61	67.93	**71.76**	**74.12**	**77.24**
1	**67.03**	**77.37**	**77.50**	83.01	**64.60**	67.39	**68.93**	70.61	**68.57**	**71.76**	73.88	77.08

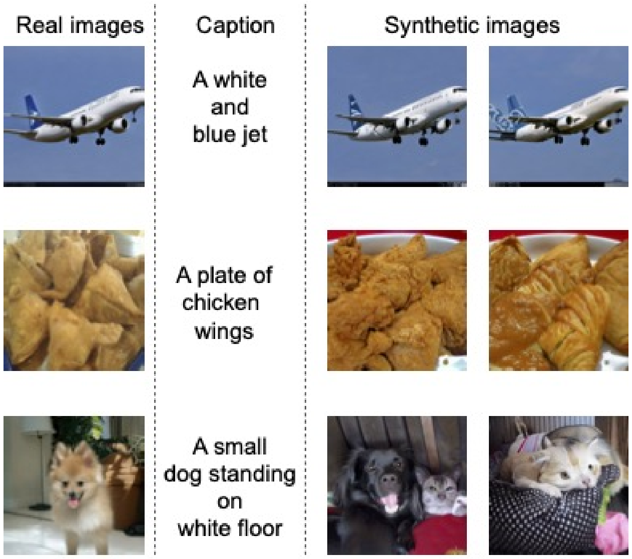

Figure 7. Failure cases: captions for the images are not specific to a particular fine-grained class of images(top row, bottom row) or are not correctly generated (middle row). Hence, synthetic images are not helpful in classification.

tion of various components of our approach. Finally, we investigate the failure cases and discuss the limitations of our approach.

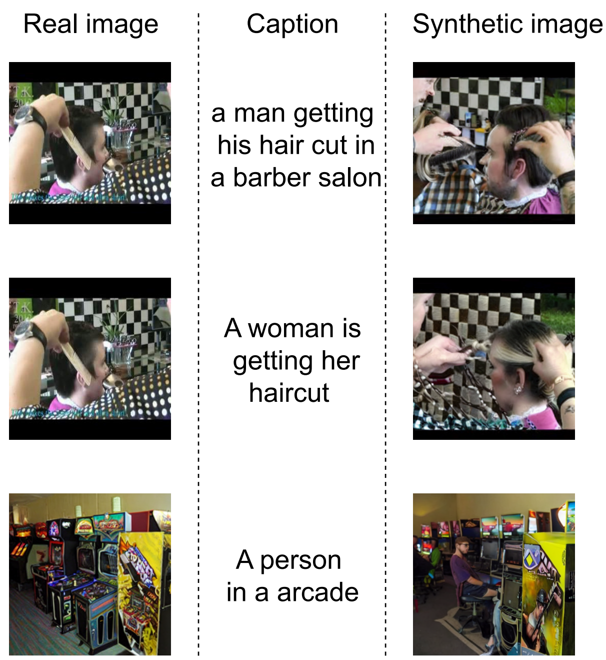

Figure 8. Diverse caption generation. The real image of a man getting a haircut + “a man getting his hair cut in a barber salon” when fed to the image-to-image diffusion model produces another image of a man getting a haircut. A real image of a man getting a haircut + “a woman is getting her haircut” produces an image of a woman getting a haircut. Therefore, we can do image editing using captions and generate diverse images.

6. Acknowledgement

Aniket Roy and RC acknowledge support through a fellowship from JHU + Amazon Initiative for Interactive AI (AI2AI) and ONR MURI grant N00014-20-1-2787.

References

[1] Mohamed Afham, Salman Khan, Muhammad Haris Khan, Muzammal Naseer, and Fahad Shahbaz Khan. Rich semantics improve few-shot learning. *arXiv preprint arXiv:2104.12709*, 2021. 2

[2] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné. Associative alignment for few-shot image classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[3] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné. Mixture-based feature space learning for few-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9041–9051, 2021. 2

[4] Arman Afrasiyabi, Hugo Larochelle, Jean-François Lalonde, and Christian Gagné. Matching feature sets for few-shot image classification. *arXiv preprint arXiv:2204.00949*, 2022. 2

[5] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 2

[6] 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*, pages 446–461. Springer, 2014. 4

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[8] Kaidi Cao, Maria Brbic, and Jure Leskovec. Concept learners for few-shot learning. In *International Conference on Learning Representations*, 2021. 1

[9] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 4

[10] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3695–3706, 2022. 7

[11] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 715–724, 2021. 7

[12] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 1, 4

[13] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *arXiv preprint arXiv:2305.16289*, 2023. 2

[14] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018. 2

[15] Trabucco et al. Effective data augmentation with diffusion models. *arXiv*, 2023. 2

[16] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 4

[17] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 2, 5

[18] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. *arXiv preprint arXiv:2209.14169*, 2022. 2, 5

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[20] 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. 2

[21] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 4

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[23] Yiren Jian and Lorenzo Torresani. Label hallucination for few-shot classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 1, 2

[24] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 7

[25] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 4

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 1

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1, 3, 4

[28] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 2, 4

[29] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 4

[30] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3, 5

[31] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021. 2

[32] 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*, pages 722–729. IEEE, 2008. 4

[33] Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. Multimodal prototypical networks for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2644–2653, 2021. 2

[34] Basudha Pal, Aniket Roy, Ram Prabhakar Kathirvel, Alice J O'Toole, and Rama Chellappa. Diversinet: Mitigating bias in deep classification networks across sensitive attributes through diffusion-generated data. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2024. 2

[35] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 4

[36] Shraman Pramanick, Aniket Roy, and Vishal M Patel. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3930–3940, 2022. 2

[37] 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*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 5

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

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 3

[40] Aniket Roy, Anshul Shah, Ketul Shah, Prithviraj Dhar, Anoop Cherian, and Rama Chellappa. FeLMi : Few shot learning with hard mixup. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1, 2

[41] Aniket Roy, Maiterya Suin, Anshul Shah, Ketul Shah, Jiang Liu, and Rama Chellappa. Diffnat: Improving diffusion image quality using natural image statistics. *arXiv preprint arXiv:2311.09753*, 2023. 2

[42] Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In *ICCV*, 2021. 7

[43] Anshul Shah, Aniket Roy, Ketul Shah, Shlok Mishra, David Jacobs, Anoop Cherian, and Rama Chellappa. Halp: Hallucinating latent positives for skeleton-based self-supervised learning of actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18846–18856, 2023. 2

[44] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4

[45] Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *ECCV 2022*, 2022. 6, 7

[46] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. *ICCV 2023*, 2023. 2, 5

[47] Shuo Wang, Xinyu Zhang, Yanbin Hao, Chengbing Wang, and Xiangnan He. Multi-directional knowledge transfer for few-shot learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3993–4002, 2022. 2

[48] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 3, 4

[49] Jingyi Xu and Hieu Le. Generating representative samples for few-shot classification. In *CVPR*, pages 9003–9013, 2022. 2

[50] Fengyuan Yang, Ruiping Wang, and Xilin Chen. Sega: semantic guided attention on visual prototype for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1056–1066, 2022. 2

[51] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023. 5

[52] Baoquan Zhang, Xutao Li, Yunming Ye, Zhichao Huang, and Lisai Zhang. Prototype completion with primitive knowledge for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3754–3762, 2021. 2

[53] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15211–15222, 2023. 2

[54] Renrui Zhang, Longtian Qiu, Wei Zhang, and Ziyao Zeng. Vt-clip: Enhancing vision-language models with visual-guided texts. *arXiv preprint arXiv:2112.02399*, 2021. 2

[55] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-

adapter: Training-free adaption of clip for few-shot classification. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 493–510. Springer, 2022. 2, 4, 5, 7, 8

[56] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2

[57] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 2, 5, 7, 8

[58] Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on thin air: Improve image classification with generated data. *arXiv preprint arXiv:2305.15316*, 2023. 2