SynBalance: Harnessing Synthetic Data in Long-tailed Recognition

Supplementary Material

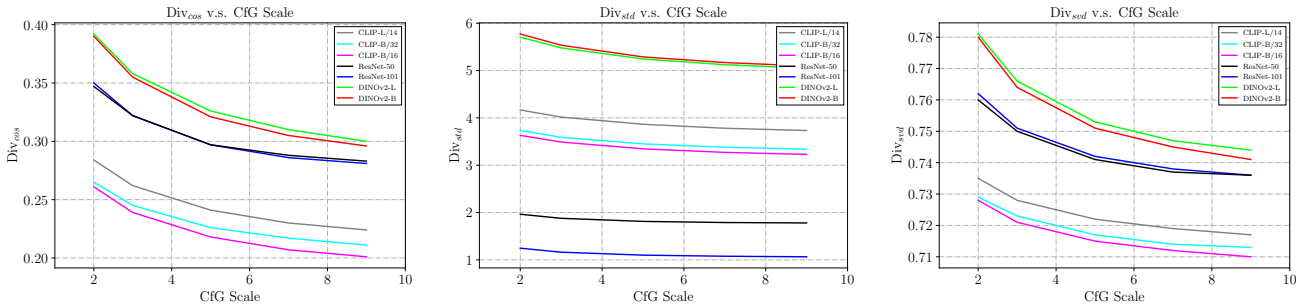

Figure 1. The diversity metrics with different vision foundation models and images generated by different CfG scales.

1. Additional Implementation Details

1.1. Special Template Generation

We design a specialized template in the form: `A photo of {a, several, a school of} {class} in {place}.`. Generating these templates requires identifying potential locations where the object might be found. Inspired by [5], we sample all 365 categories from the Places-365 dataset [6] and prompt LLM (i.e., LLaMA2 [2]) with the query: `Is it possible to find a {class} in the {place}? Answer Yes or No.`. The responses from the LLM are then aggregated to determine the possible backgrounds for each object class and form special templates.

1.2. Optimal Transport for Domain Gap

We adopted Optimal Transport(OT) to calculate the domain gap between real and synthetic images. The cost matrix $\mathcal{M}_c$ of OT between two set of images are defined as:

$$\mathcal{M}_c(i,j) = \frac{1 - \langle \hat{\mathcal{I}}_{i,s,c}, \hat{\mathcal{I}}_{j,r,c} \rangle}{2}, \tag{1}$$

where $\hat{\mathcal{I}}_{i,s,c}$ and $\hat{\mathcal{I}}_{j,r,c}$ are normalized image embeddings from real and synthetic images of class c.

The optimization target of OT is:

$$\mathrm{OT}(\hat{\mathcal{I}}_{s,c}, \hat{\mathcal{I}}_{r,c}) = \underset{\gamma \in \mathbb{R}_+^{m \times n}}{\arg\min} \sum \gamma_{i,j} \mathcal{M}_{c,i,j}. \tag{2}$$

$$\text{s.t.} \gamma \mathbf{1}_n = \frac{1}{n}\mathbf{1}_n, \gamma^\top \mathbf{1}_m = \frac{1}{m}\mathbf{1}_m, \gamma \geq 0$$

The OT between $\hat{\mathcal{I}}_{i,s,c}$ and $\hat{\mathcal{I}}_{j,r,c}$ represents the domain gap between two sets of images.

Method	Backbone	All	Many	Medium	Few
Baseline	ResNet-50	42.0	61.2	35.1	12.6
Only-100	ResNet-50	18.9	19.4	18.7	18.3
Random-100	ResNet-50	41.3	59.5	33.8	13.0
Add-100	ResNet-50	48.1	64.9	42.4	21.4
Only-200	ResNet-50	23.1	23.9	22.8	22.1
Add-200	ResNet-50	50.5	67.0	45.0	23.6
Only-300	ResNet-50	25.5	25.8	25.2	25.1
Add-300	ResNet-50	51.8	68.0	46.1	26.7
Only-400	ResNet-50	27.3	27.9	27.2	26.1
Add-400	ResNet-50	52.3	68.3	46.9	26.3
Only-500	ResNet-50	28.1	29.0	27.6	26.8
Add-500	ResNet-50	52.7	68.6	47.5	26.4

Table 1. The Top-1 accuracy in % on ImageNet-LT dataset. More results of Add-n and Only-n experiments

2. Additional Experiments

2.1. Experimental Results of Random-n, Add-n and Only-n

As we mentioned in Sec. 3.3.1, we conduct Random-n, Add-n and Only-n on ImageNet-LT [4] dataset. The experimental results are shown in Tab. 1. The Random-100 shows worse performance than the Baseline and Add-100 experiment, which indicates the effectiveness of the added synthetic data. Meanwhile, with the n increase, both Add-n and Only-n achieve better performance.

2.2. Diversity Metrics

Following previous works [1, 3], we use the standard deviation of the cosine similarity between image features as the diversity metrics (described in Sec. 3). Additionally, we

Figure 2. The visualization results of synthetic images generated for the ImageNet-LT dataset. Class name: `Sea lion`, `Llamma`, `Otter`, `Container ship`, `School bus`, and `Umbrella`.

Experiment	Prompts	All	Many	Median	Few
Baseline	-	42.0	61.2	35.1	12.6
SynFusion	CLIP template	47.5	64.5	41.4	21.2
SynFusion	LLM generated	46.1	63.7	39.7	19.4
SynFusion	Special template	**48.1**	**64.9**	**42.4**	**21.4**

Table 2. The experimental results on ImageNet-LT with generated images from different prompts, additional data n=100.

explored another two diversity metrics and check the effectiveness of all three metrics with images generated by different Classifier-free Guidance (CfG) scales. The metrics are formulated as follows:

Cosine similarity standard deviation. The diversity metric which we utilized in the paper:

$$\text{Div}_{cos} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1-\langle\hat{\mathcal{I}}_{i,c},\hat{\bar{\mathcal{I}}}_c\rangle}{2}}. \tag{3}$$

Euclidean standard deviation. The standard deviation can also be measured by the Euclidean distance between generated image features:

$$\text{Div}_{std} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\mathcal{I}_{i,c}-\hat{\mathcal{I}}_c\|_2^2}. \tag{4}$$

Singular value. All normalized generated image features $\hat{\mathcal{I}}_c \in \mathbb{R}^d$ can be stacked as a feature matrix $\mathcal{M}_{\mathcal{I}} = [\hat{\mathcal{I}}_{0,c},...,\hat{\mathcal{I}}_{n,c}] \in \mathbb{R}^{n\times d}$. The singular value decomposition (SVD) of $\mathcal{M}_{\mathcal{I}}$ will generate a set of singular value $\sigma = \{\sigma_1,\dots\}$, where $\sigma_1 \in [0,1]$ is the largest singular value. According to Principal component analysis (PCA), the σ_1 represents how the data varies along the first principal component, which can be used as the diversity metric:

$$\text{Div}_{svd} = 1-\sigma_1. \tag{5}$$

As shown in Fig. 1, with the increasing of the CfG scales, all three diversity metrics decrease, which follows the assumption that the lower CfG scales are, generated images are more diverse. However, the scale Div_{std} is highly related to the vision models, and both Div_{cos} and Div_{svd} are

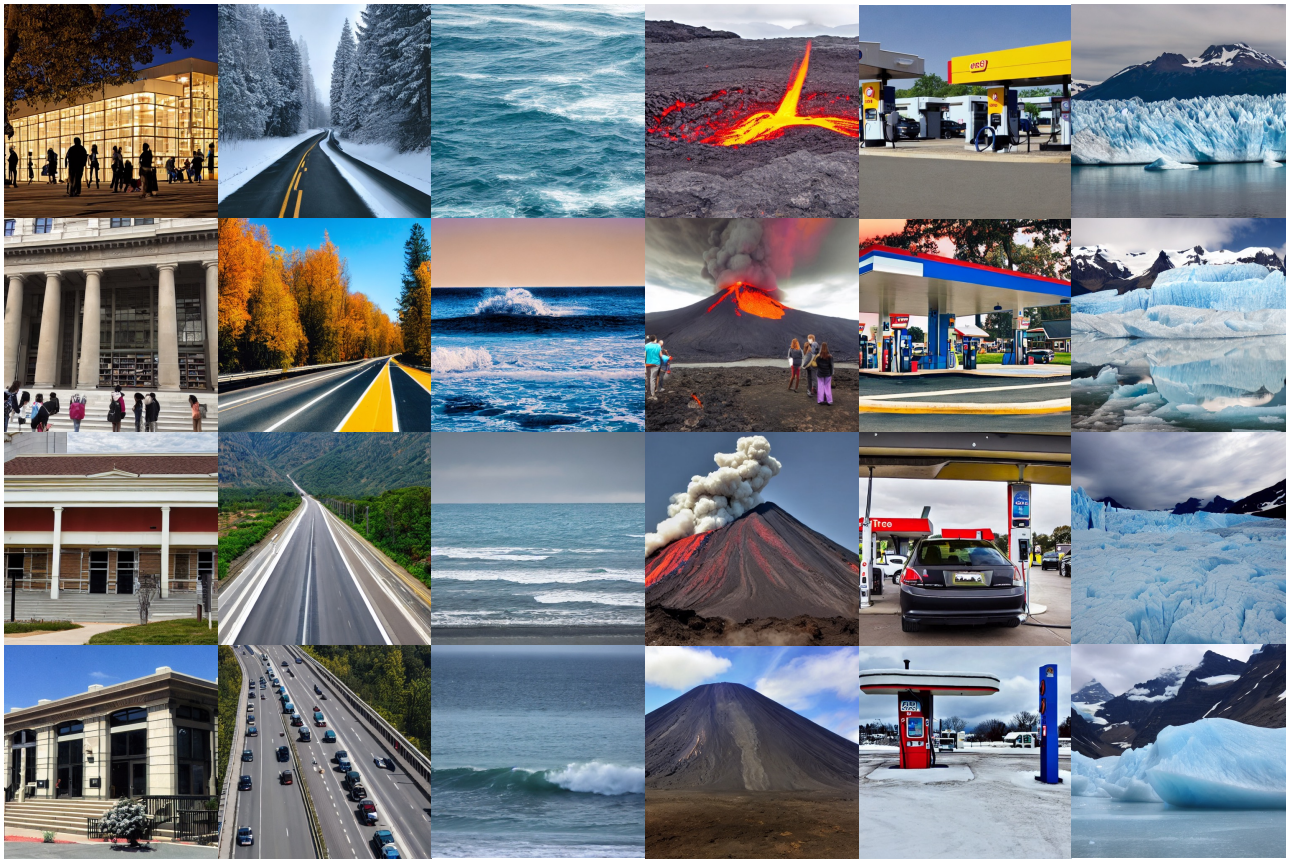

Figure 3. The visualization results of synthetic images generated for the Places-LT dataset. Class name: `Library`, `Highway`, `Ocean`, `Volcano`, `Gas station`, and `Glacier`.

in $[0, 1]$. Therefore, we use Div_{cos} as the diversity metric in the main paper.

3. Visualization of Generated Images

3.1. Synthetic Images for Different Datasets

ImageNet-LT. As shown in Fig. 2, there are some visualization results from synthetic images generated for ImageNet-LT dataset with classes as: `Sea lion`, `Llamma`, `Otter`, `Container ship`, `School bus`, and `Umbrella`.

Places-LT. As shown in Fig. 3, there are some visualization results from synthetic images generated for Places-LT dataset with classes as: `Library`, `Highway`, `Ocean`, `Volcano`, `Gas station`, and `Glacier`.

iNaturalist. As shown in Fig. 4, there are some visualization results from synthetic images generated for iNaturalist 2024 dataset with classes as: `Accipiter cooperii`, `Aepyceros melampus`, `Sylvilagus cunicularius`, `Cygnus buccinator`, `Chaetodon lineolatus`, and `Lacrymaria lacrymabunda`.

3.2. Synthetic Images with Different Recognizability and Diversity

As shown in Fig. 5 and 6, images with different recognizability and diversity are shown. Images with low and high diversity are shown in Fig. 5. Images with low and high recognizability are shown in Fig. 6.

References

[1] Victor Boutin, Thomas Fel, Lakshya Singhal, Rishav Mukherji, Akash Nagaraj, Julien Colin, and Thomas Serre. Diffusion models as artists: are we closing the gap between humans and machines? *arXiv preprint arXiv:2301.11722*, 2023. 1

[2] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[3] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *Proceedings of the IEEE/CVF*

Figure 4. The visualization results of synthetic images generated for the iNaturalist 2018 dataset. Class name: `Accipiter cooperii`, `Aepyceros melampus`, `Sylvilagus cunicularius`, `Cygnus buccinator`, `Chaetodon lineolatus`, and `Lacrymaria lacrymabunda`.

Conference on Computer Vision and Pattern Recognition, pages 7382–7392, 2024. 1

[4] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019. 1

[5] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15887–15898, 2024. 1

[6] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 1

Low diversity↓ High diversity↑

Figure 5. The visualization results of images with different diversity.

Low recognizability↓ High recognizability↑

Figure 6. The visualization results of images with different recognizability.