Supplementary Material for Label Augmentation as Inter-class Data Augmentation for Conditional Image Synthesis with Imbalanced Data

Kai Katsumata Duc Minh Vo Hideki Nakayama
The University of Tokyo, Japan
{katsumata,vmduc,nakayama}@nlab.ci.i.u-tokyo.ac.jp

1. Additional experiments

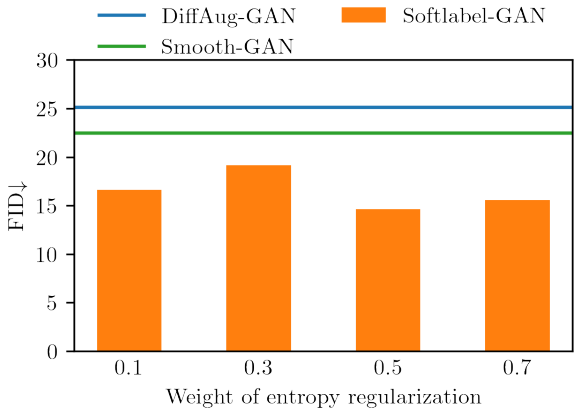


Figure 1. Sensitivity to the hyperparameter of entropy regularization. We test the performance with the weight of the entropy regularization term in the range of 0.1 to 0.7. Our Softlabel-GAN consistently improve the baseline.

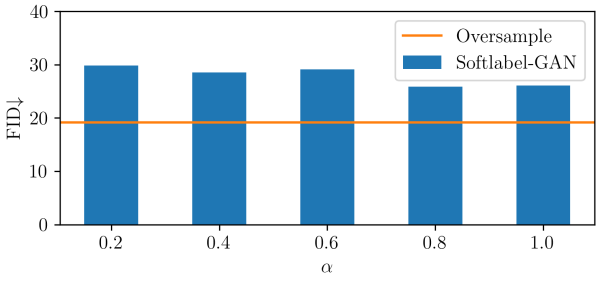


Figure 2. The effects of oversampling parameter. Performance gains by oversampling technique are limited on different hyperparameters.

Effects of hyperparameter. We first test the insensitivity of the hyperparameter of our method: the weight of the entropy regularization term. We conduct experiments on AnimeFace with different weights. Figure 1 demonstrate that Our method outperforms the baselines for any hyperparameter.

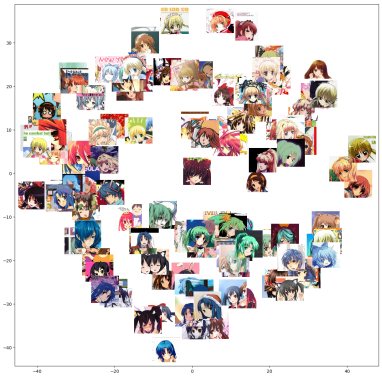
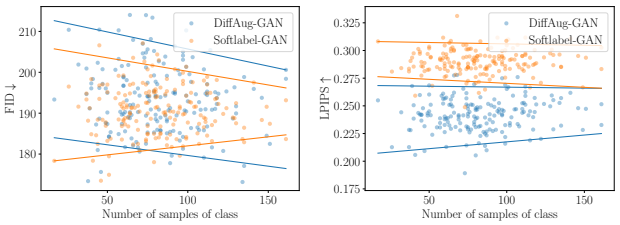


Figure 3. T-SNE visualization. We embed the output of the penultimate layer of the pretrained classifier into 2D using t-SNE. The classifier embeds similar instances, such as those with the same hair color.



(a) Class FID (b) Class LPIPS

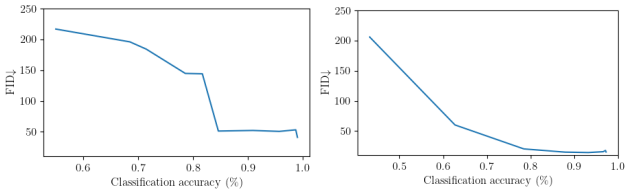
Figure 4. Relationships between per-class scores (i.e., class FID and class LPIPS) and the number of samples per class. The orange and blue points indicate scores of Softlabel-GAN and DiffAug-BigGAN, respectively. The lines indicate 5% quantiles and 95% quantiles. In DiffAug-BigGAN, the number of samples per class is inversely proportional to the class FID and directly proportional to the class LPIPS, meaning fewer samples lead to higher class FID and lower class LPIPS. The results also show that the Softlabel-GAN's performance does not depend on the number of samples of a class compared to DiffAug-BigGAN.

eter.

Effects of sampling strategy in the oversampling baseline. Figure 2 shows the FID scores of the oversampling

Table 1. Comparison of Softlabel-GAN and Online label smoothing (OLS) [4] in terms of intra-FID.

Method	AnimeFace	Cars	Oxford102	CIFAR-10	CIFAR-100	Tiny ImageNet
Softlabel-GAN	**57.43**±16.56	**89.04**±11.64	**126.32**±42.69	**109.79**±24.72	**201.61**±32.50	**238.83**±53.74
OLS [4]	144.34±20.69	132.46±22.19	262.64±28.41	137.98±49.02	263.13±26.08	356.19±19.07



(a) Tiny ImageNet (b) AnimeFace

Figure 5. Dependency of generation performance on classifier performance on Softlabel-GAN with Tiny ImageNet. The method does not require the perfect classifier.

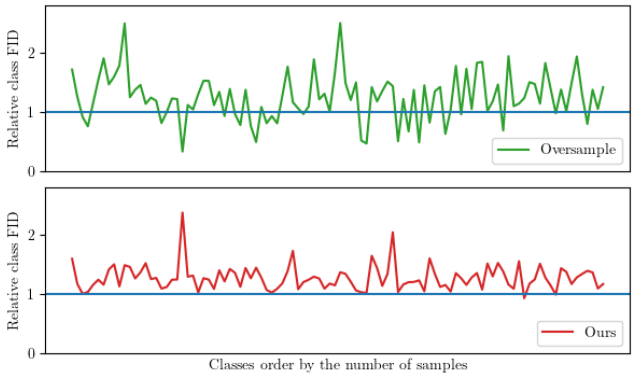


Figure 6. Relative class FID (class FID of DiffAug-GAN/class FID of a compared method) on the Oxford-102 Flower dataset. We compare DiffAug-GAN with oversampling and our Softlabel-GAN. The score above one indicates that the compared method outperforms the DiffAug-GAN baseline. The score below one indicates that the compared method loses to the baseline. We can see that our method almost consistently outperforms the baseline while DiffAug-GAN with oversampling loses the baseline in several classes.

baseline on the experiments with different hyperparameters. The parameter α in $[0, 1]$ interpolates between a uniform distribution ($\alpha = 1$) and data distribution ($\alpha = 0$).

Visualization of Embeddings. We verify the contribution of the probability of vector obtained by a pretrained classifier in enhancing the affinity of augmented data. To this end, we employ t-SNE to visualize the embeddings of the penultimate layer in the classifier. Figure 3 shows that the classifier successfully extracts the shared features between classes.

Dependency of the per-class performance on the number of samples. On the AnimeFace datasets, the per-class

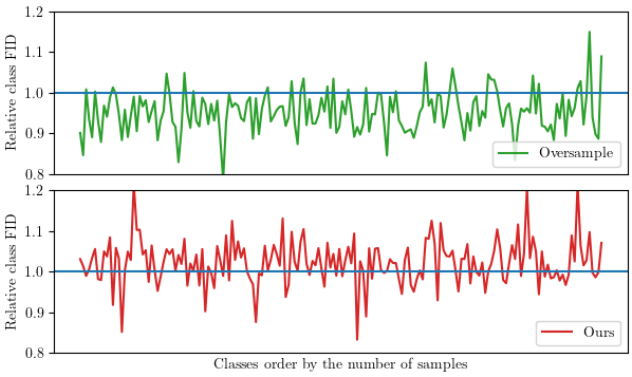


Figure 7. Relative class FID on the AnimeFace dataset.

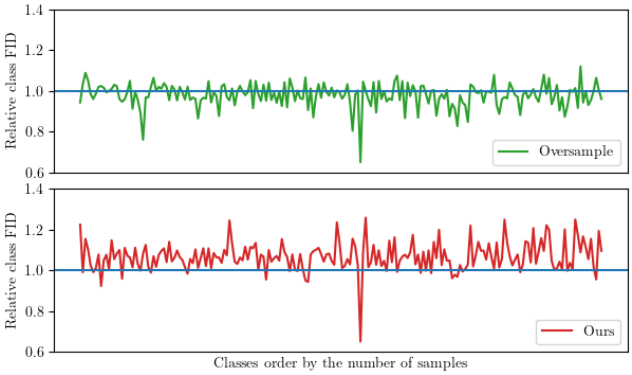


Figure 8. Relative class FID on the Tiny ImageNet dataset.

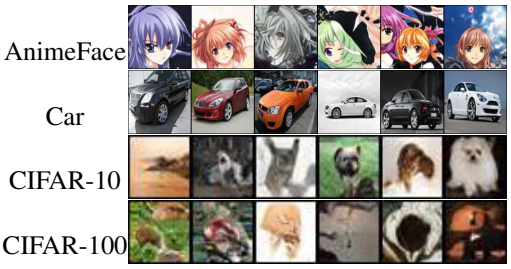


Figure 9. Generated examples by OLS [4]. The images of each row are conditioned by the same class.

FID score of DiffAug-GAN is inversely proportional to the number of samples of class, and the per-class LPIPS score of DiffAug-GAN is directly proportional to the number of samples of class (Fig. 4). The results show that the performance of DiffAug-GAN degrades when the class contains

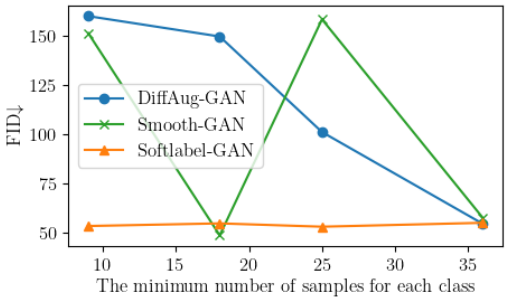


Figure 10. FID scores as a function of the number of samples of the minor class.

a few samples on both metrics. On the contrary, the performance of Softlabel-GAN is quite insensitive to the number of class data. In particular, intra-LPIPS is constant regardless of the number of class samples. The results show that a few samples lead to worse performance, and the performance depends on the number of a class in DiffAug-GAN. In contrast to the baseline, our method levels the per class performance.

Dependency of generation performance on classifier performance. We conduct experiments using labels obtained from classifiers with different performances. Figure 5 shows that a classifier with an accuracy of 85% or above achieves the sufficient performance. We can conclude that we do not require a classifier with sophisticated performance for our purpose. We further tested the performance with ResNet as the classifier. Softlabel-GAN with ResNet on AnimeFace achieves 20.39 of FID, surpassing 22.46 of FID of Smooth-GAN and being comparable to Softlabel-GAN with SpinalNet that achieves 19.14. Thus, our method performs well regardless of the choice of the network architectures.

Comparison with an oversampling method. One of the most commonly used techniques for imbalanced data is an oversampling of minor classes, which balances the number of samples of each class. Figures 6 to 8 show the relative per-class performance to the DiffAug-GAN baseline of the oversampling method and Softlabel-GAN. Our method consistently improves the performance, while oversampling sometimes contributes to the performance of cGANs.

Pretraining method vs. co-training method. We consider a naive variant of Softlabel-GAN, which trains cGANs and a classifier simultaneously, namely the co-training method. We conduct experiments on the AnimeFace dataset. The co-training method achieves an FID of 50.17 and does not reach the performance of DiffAug-GAN, Smooth-GAN, and our Softlabel-GAN.

Comparison with OLS. We show FID of Softlabel-GAN and OLS on six datasets in Tab. 1. Softlabel-GAN is better than OLS on the training cGANs from class-imbalanced

Table 2. The Precision, Recall, intra-Precision (i-P), intra-Recall (i-R), Density, Coverage, intra-Density (i-D), and intra-Coverage (i-C) on the experiments with the Vision-aided GAN baseline.

Method	Precision↑	Recall↑	Density↑	Coverage↑	i-P↑	i-R↑	i-D↑	i-C↑
	AnimeFace							
Vision-aided GAN	0.775	0.278	1.223	0.772	0.769	0.114	0.786	0.975
w/ our label augmentation	**0.812**	**0.297**	**1.522**	**0.819**	**0.778**	**0.183**	**0.803**	**0.978**
	Stanford Cars							
Vision-aided GAN	0.601	0.072	0.497	0.541	0.591	0.020	0.360	0.936
w/ our label augmentation	**0.852**	**0.377**	**1.395**	**0.863**	**0.684**	**0.255**	**0.548**	**0.981**
	Oxford 102 Flowers							
Vision-aided GAN	0.608	0.115	0.417	0.432	0.460	0.003	0.206	0.598
w/ our label augmentation	**0.864**	**0.324**	**1.082**	**0.819**	**0.649**	**0.050**	**0.479**	**0.862**

datasets. While OLS achieves lower FID than Softlabel-GAN on a few dataset, the intra-FID of OLS is considerably worse than that of Softlabel-GAN. Furthermore, OLS generate images that almost do not associated with given class-conditions as shown in Fig. 9.

Performance on balanced dataset. We demonstrate that the Softlabel-GAN work properly even with balanced datasets. We shows the FID scores of DiffAugGAN, Smooth-GAN, and Softlabel-GAN in Fig. 10. Softlabel-GAN outperforms the baselines on imbalanced dataset and achieves the performance comparable with to DiffAug-GAN on balanced dataset.

Comrpaison with P&R [2] and D&C [3] metrics. We show the further metrics of the experiments with a Vision-aided GAN [1] baseline in Tab. 2. Substantial improvements in intra-precision shows that our method synthesizes diverse images and avoids mode collapse.

2. More examples

We further provide qualitative results. Figures 11 and 12 show random examples generated by the methods in minor and major classes, respectively. In comparison with DiffAug-GAN, Softlabel-GAN generates diverse images. Figure 13 shows the generated examples of the imbalanced Tiny ImageNet experiments.

References

[1] Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10651–10662, 2022. 3

[2] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, volume 32, 2019. 3

[3] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *ICML*, pages 7176–7185, 2020. 3

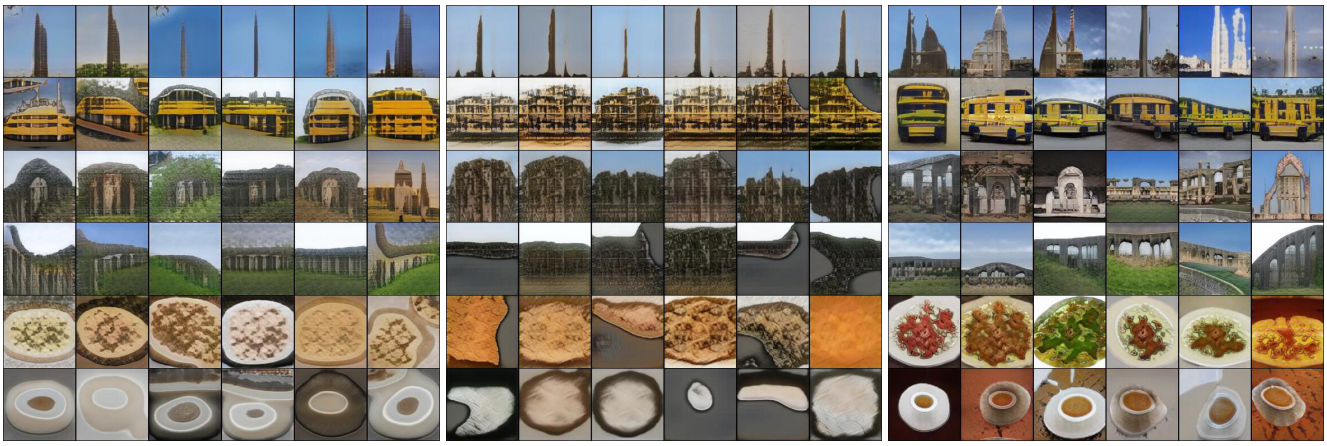
(a) DiffAug-GAN (b) Softlabel-GAN

Figure 11. Random Examples of minor classes obtained by DiffAug-GAN and Softlabel-GAN.

[4] Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. *IEEE TIP*, 30:5984–5996, 2021. 2

(a) DiffAug-GAN (b) Softlabel-GAN

Figure 12. Random Examples of major classes obtained by DiffAug-GAN and Softlabel-GAN.



(a) DiffAug-GAN (b) Smooth-GAN (c) Softlabel-GAN

Figure 13. Visual comparison of conditional image generation on the imbalanced Tiny ImageNet dataset. Our method produces diverse images while achieving the plausibility of images.