DECO-Bench: Unified Benchmark for Decoupled Task-Agnostic Synthetic Data Release : Supplementary

Anonymous Author(s)
Affiliation

A Appendix / supplemental material

In this section, we present our supplementary materials that expand on details of our experiments and further discuss our results. In Sec. A.1 we describe our $Tr + V + G$ setup and present the results. Sec. A.2 focuses on our experiments using only front faces of each dataset. In Sec. A.3, we elaborate on our setups and hyperparameters in details. And finally, Fig. 4 presents the high-resolution equivalent of Fig. 3 from the main paper.

A.1 Tr+V+G Experiments

In the main section of the paper, we observed the beneficial effect of adding V split while training the privacy and utility classifiers (i.e., $V + G$ vs G). As we discussed in the main paper, the increased amount of data during classifiers training helps alleviate overfitting and allows for a better observation of the privacy-utility trade-off. To further investigate this, we generated synthetic images from the Tr split and trained the classifiers. Tab. 7 to Tab. 10 extend the results from Tab. 2 to Tab. 5 (main paper) respectively. We observe that only some setups demonstrate improvement. We suggest this may be due to the fact that, despite the significant increase in data, we used the same backbone as before, ResNet18, which might not have sufficient capacity to benefit from the additional Tr set. It is also important to note, that since the base model is trained on Tr, we do not evaluate the privacy using PT for this setup.

Table 7: **Setup A**: Trained Fairface and Tested on Fairface dataset. Classification accuracy percentage for privacy leakage and utility.

	Method	Privacy (PT, Race, ↓)	Privacy (CLS,Race,↓)	Util (CLS,Age,↑)	Util (CLS,Gender, ↑)
Tr+V+G	Baseline (Real)	–	62.38	44.8	83.36
	CE	–	51.03	32.72	80.25
	CE + Mask	–	48.87	38.09	79.81
	Metric Learning	–	54.89	29.56	76.47

Table 8: **Setup B**: Trained FairFace and Tested on UTKFace dataset. Classification accuracy percentage for privacy leakage and utility.

	Method	Privacy (PT, Race, ↓)	Privacy (CLS,Race,↓)	Util (CLS,Age,↑)	Util (CLS,Gender, ↑)
Tr+V+G	Baseline (Real)	–	81.24	56.02	90.55
	CE	–	69.71	46.65	85.97
	CE + Mask	–	61.22	44.35	87.2
	Metric Learning	–	61.33	39.31	84.79

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

Table 9: **Setup C**: Trained (Fairface + UTKFace) and Tested on FairFace dataset. Classification accuracy percentage for privacy leakage and utility.

	Method	Privacy (PT, Race, ↓)	Privacy (CLS,Race,↓)	Util (CLS,Age,↑)	Util (CLS,Gender, ↑)
Tr+V+G	Baseline (Real)	–	65.74	44.53	84.63
	CE	–	46.91	39.42	81.45
	CE + Mask	–	56.35	38.46	79.89
	Metric Learning	–	57.92	41.27	80.9

Table 10: **Setup C**: Trained (Fairface + UTKFace) and Tested on UTKFace dataset. Classification accuracy percentage for privacy leakage and utility.

	Method	Privacy (PT, Race, ↓)	Privacy (CLS,Race,↓)	Util (CLS,Age,↑)	Util (CLS,Gender, ↑)
Tr+V+G	Baseline (Real)	–	81.81	54.84	91
	CE	–	62.01	40.75	88.19
	CE + Mask	–	57.73	42.35	88.92
	Metric Learning	–	66.28	49.49	87.46

A.2 Front Face Experiments

In many studies utilizing datasets of face images, it is common practice to use only front faces. In our main paper, we presented results using both front and profile faces. To accommodate a variety of experimental setups in our benchmark and to study the effect of using only front faces on the privacy-utility trade-off, we repeat our experiments in this section using only front faces.

A.2.1 Front Face Detection:

To detect images with front faces, we processed all splits of the FairFace and UTKFace datasets using dlib library [King, 2009]. We perform front face detection both in original image size and also in 128×128 resolution; given this is the image resolution we train our base model with. Below, we present the statistics of front faces in each data split of the FairFace and UTKFace datasets. As observed, the FairFace dataset contains more profile images compared to the UTKFace dataset.

Table 11: Dataset Statistics for Front Faces of FairFace and UTKFace datasets. The number of classes for the datasets after cross label mapping are demonstrated (see Sec.4.1, main paper).

Dataset	Train (Tr) Samples	Val (V) Samples	Gen (G) Samples	Test Samples	Privacy Attr. (#Classes)	Utility Attr. (#Classes)	Eval Metric
FairFace	11263	1256	4480	2747	Race (C=5)	Age (C=9), Gender (C=2)	Accuracy
UTKFace	12374	1372	5307	3478	Race (C=5)	Age (C=9), Gender (C=2)	Accuracy

A.2.2 Front Face Evaluation:

In our first round of experiments, we maintain the entire pipeline from the main paper, with the exception of using only front faces in the $Test$ split. Specifically, we train the base model on Tr using both front and profile images. We then generate synthetic images from G_{real}, and V, Tr splits using both profile and front images and train the privacy and utility classifiers on them. In the final step, we use only front faces in $Test$ split for evaluation and report the privacy leakage and utility performance. Similarly, for PT evaluation, we use only front faces. The results are presented in Tab. 12 to Tab. 15.

Table 12: **Setup A** : Trained Fairface and Tested on Fairface dataset Front Faces only. Classification accuracy percentage for privacy leakage and utility.

	Method	Privacy (PT, Race, ↓)	Privacy (CLS,Race,↓)	Util (CLS,Age,↑)	Util (CLS,Gender, ↑)
G	Baseline (Real)	–	35.68	39.86	76.01
	CE	18.71	42.34	35.86	78.59
	CE + Mask	15.13	47.11	37.31	72.12
	Metric Learning	–	50.86	32.65	73.79
V+G	Baseline (Real)	–	58.54	41.9	78.05
	CE	18.56	49.91	36.55	75.83
	CE + Mask	15.34	50.24	37.6	80.42
	Metric Learning	–	53.91	37.57	76.37
Tr+V+G	Baseline (Real)	–	67.24	49.33	87.08
	CE	–	45.94	40.84	82.34
	CE + Mask	–	28.76	36.59	82.64
	Metric Learning	–	53.66	40.01	82.71

Table 13: **Setup B** : Trained Fairface and Tested on UTKFace dataset Front Faces only. Classification accuracy percentage for privacy leakage and utility.

	Method	Privacy (PT, Race, ↓)	Privacy (CLS,Race,↓)	Util (CLS,Age,↑)	Util (CLS,Gender, ↑)
G	Baseline (Real)	–	73.23	48.33	86.83
	CE	42.87	66.1	42.9	81.94
	CE + Mask	16.90	52.04	44.31	84.39
	Metric Learning	–	59.8	39.07	83.58
V+G	Baseline (Real)	–	72.94	50.72	85.48
	CE	42.82	69.58	44.77	84.91
	CE + Mask	16.74	57.82	39.59	84.68
	Metric Learning	–	61.18	43.16	83.84
Tr+V+G	Baseline (Real)	–	81.68	56.41	91.14
	CE	–	69.47	46.75	86.06
	CE + Mask	–	61.33	44.51	87.9
	Metric Learning	–	62.13	39.25	84.85

Table 14: **Setup C**: Trained Fairface+UTKFace Tested on FairFace dataset Front Faces only. Classification accuracy percentage for privacy leakage and utility.

	Method	Privacy (PT, Race, ↓)	Privacy (CLS,Race,↓)	Util (CLS,Age,↑)	Util (CLS,Gender, ↑)
G	Baseline (Real)	–	56.28	43.54	81.43
	CE	26.25	48.02	36.33	79.87
	CE + Mask	15.13	46.81	35.57	76.3
	Metric Learning	–	45.61	28.69	73.97
V+G	Baseline (Real)	–	59.23	47.36	83.15
	CE	26.87	50.16	37.02	80.31
	CE + Mask	15.34	42.81	38.26	79.32
	Metric Learning	–	56.13	33.71	78.01
Tr+V+G	Baseline (Real)	–	71.06	50.38	87.19
	CE	–	57.41	42.88	82.71
	CE + Mask	–	55.08	41.35	80.82
	Metric Learning	–	59.56	43.17	83.4

Table 15: **Setup C**: Trained Fairface+UTKFace Tested on UTKFace dataset Front Faces only. Classification accuracy percentage for privacy leakage and utility.

	Method	Privacy (PT, Race, ↓)	Privacy (CLS,Race,↓)	Util (CLS,Age,↑)	Util (CLS,Gender, ↑)
G	Baseline (Real)	–	76.57	50.86	88.18
	CE	13.64	59.49	43.42	85.51
	CE + Mask	16.90	58.77	42.32	87.21
	Metric Learning	–	51.04	41.17	82.78
V+G	Baseline (Real)	–	76.88	52.67	88.04
	CE	13.82	62.1	42.21	86.83
	CE + Mask	16.74	64.17	43.67	87.98
	Metric Learning	–	55.15	42.73	84.88
Tr+V+G	Baseline (Real)	–	80.71	55.78	90.8
	CE	–	62.39	41.09	88.13
	CE + Mask	–	58.05	42.73	89.13
	Metric Learning	–	66.53	49.74	87.61

A.2.3 Front Face Training and Evaluation

In this section, we extend our front face experiments by repeating the entire pipeline using only front faces. Specifically, we train the base model using only front faces of Tr set, followed by generating synthetic images from only front faces of G_{real}, V and Tr splits. We then train the privacy and utility classifiers using these generated images and, finally, evaluate the models on only the front faces of $Test$ split. Tab. 16 to Tab. 19 demonstrate the results.

Table 16: **Setup A (Front Faces only)**: Trained Fairface Front Faces only and Tested on Fairface dataset Front Faces only. Classification accuracy percentage for privacy leakage and utility.

	Method	Privacy (PT, Race, ↓)	Privacy (CLS,Race,↓)	Util (CLS,Age,↑)	Util (CLS,Gender, ↑)
G	Baseline (Real)	–	54.42	34.55	71.17
	CE	18.70	42.52	34.44	72.52
	CE + Mask	15.13	46.01	35.02	72.26
	Metric Learning	–	51.18	36.08	72.84
V+G	Baseline (Real)	–	56.75	38.7	79.43
	CE	18.54	33.49	35.24	72.04
	CE + Mask	15.34	40.88	37.5	76.77
	Metric Learning	–	45.18	36.4	76.88
Tr+V+G	Baseline (Real)	–	67.53	45.9	85.84
	CE	–	56.57	41.32	81.58
	CE + Mask	–	45.69	35.27	82.82
	Metric Learning	–	46.09	40.52	80.85

Table 17: **Setup B (Front Faces only)**: Trained Fairface Front Faces only and Tested on UTKFace dataset Front Faces only. Classification accuracy percentage for privacy leakage and utility.

	Method	Privacy (PT, Race, ↓)	Privacy (CLS,Race,↓)	Util (CLS,Age,↑)	Util (CLS,Gender, ↑)
G	Baseline (Real)	–	72.83	48.88	83.32
	CE	42.87	63.02	43.93	81.08
	CE + Mask	16.90	52.59	38.47	86.26
	Metric Learning	–	60.7	34.76	83.29
V+G	Baseline (Real)	–	74.24	48.91	87.72
	CE	42.82	66.07	41.35	83.84
	CE + Mask	16.74	61.96	44.13	85.48
	Metric Learning	–	62.33	42.73	83.73
Tr+V+G	Baseline (Real)	–	79.61	54.17	88.9
	CE	–	61.87	46.18	84.73
	CE + Mask	–	65.96	47.18	87.26
	Metric Learning	–	63.89	44.48	89.1

Table 18: **Setup C (Front Faces only)**: Trained Fairface + UTKFace Front Faces only and Tested on Fairface dataset Front Faces only. Classification accuracy percentage for privacy leakage and utility.

	Method	Privacy (PT, Race, ↓)	Privacy (CLS,Race,↓)	Util (CLS,Age,↑)	Util (CLS,Gender, ↑)
G	Baseline (Real)	–	59.96	42.77	78.96
	CE	26.25	45.58	35.35	78.41
	CE + Mask	15.13	51.62	38.11	77.18
	Metric Learning	–	42.34	31.67	77.25
V+G	Baseline (Real)	–	59.05	44.05	83.29
	CE	26.87	43.79	39.79	77.9
	CE + Mask	15.34	44.99	39.06	75.06
	Metric Learning	–	45.03	37.68	76.45
Tr+V+G	Baseline (Real)	–	67.09	49.22	86.71
	CE	–	50.16	38.48	80.74
	CE + Mask	–	51.58	41.35	82.64
	Metric Learning	–	57.04	43.17	81.87

A.3 Experiment Details:

In this section, we provide detailed training information and hyperparameters for every component in our pipeline, supplementing Sec.4.2 of the main paper. **Base model:** to train the base model, we resize all images to 128×128. The LoRA adapter layers are initialized with Gaussian distribution, and their rank is set to 4. The base model is trained for 15 epochs using a batch size of 128 and the AdamW [Loshchilov and Hutter, 2019] optimizer with weight decay of 0.01. The learning rate is $1e-5$ for CE and CE+Mask, and $1e-4$ for Metric learning. We stop training before overfitting occurs. We employ a constant learning rate scheduler with 200 warm-up steps and apply gradient clipping with maximum grad norm of 1. The margin for triplet margin loss (metric learning) is 0.3. The masking ratio for CE+Mask setup is 0.6 and the sigmoid temperature in FGN is 1/30. A LoRA adapter is added to the learnable layer of FGN. The configurations for variational autoencoder and CLIP remain unchanged after loading the pretrained models [hf].

Generation: During generation, we load our pretrained base model with trained LoRA adaptors into the SD image-to-image pipeline [Meng et al., 2021]. The strength and guidance scale are set to 0.75 and 7.5, respectively. We fuse the LoRA weights with the scale of 1, meaning the LoRA weights completely replace the weights of the base layer they were added to.

Classification: The classifiers are trained using an image resolution of 128×128 and a batch size of 128. The ResNet18 weights are initialized randomly. We use an SGD optimizer with weight decay of $5e-4$, momentum of 0.9, and learning rate of 0.1. An exponential learning rate scheduler with

Table 19: **Setup C (Front Faces only)**: Trained Fairface + UTKFace Front Faces only and Tested on UTKFace dataset Front Faces only. Classification accuracy percentage for privacy leakage and utility.

	Method	Privacy (PT, Race, ↓)	Privacy (CLS,Race,↓)	Util (CLS,Age,↑)	Util (CLS,Gender, ↑)
G	Baseline (Real)	–	76.77	52.19	86.69
	CE	13.64	63.77	39.59	85.85
	CE + Mask	16.90	58.8	44.88	85.34
	Metric Learning	–	58.05	43.3	84.73
V+G	Baseline (Real)	–	73.75	51.44	88.5
	CE	13.82	67.94	44.45	85.83
	CE + Mask	16.74	53.54	40.91	85.34
	Metric Learning	–	66.5	44.42	84.96
Tr+V+G	Baseline (Real)	–	82.58	56.12	91.17
	CE	–	64.09	45.95	87.92
	CE + Mask	–	64.26	48.39	87.67
	Metric Learning	–	67.4	44.22	88.59

gamma of 0.9 is employed, and the models are trained for 100 epochs, stopping training after 20 epochs of no improvement (delta 0.01) in validation loss.

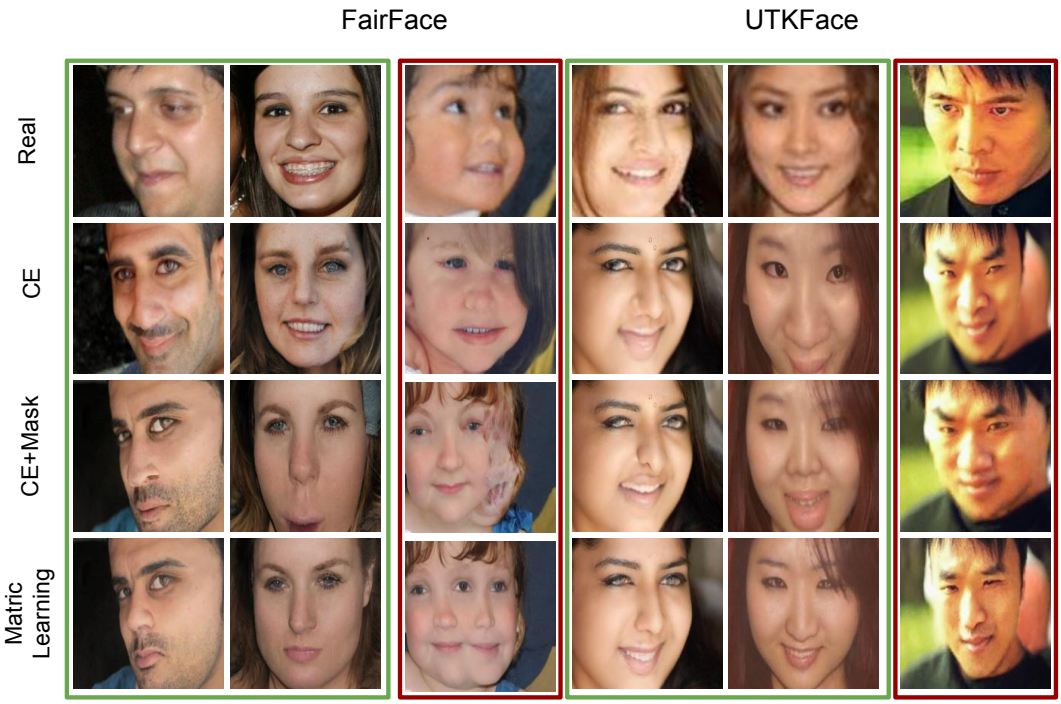

Figure 4: Generated Examples. High resolution image from main paper.

References

URL https://huggingface.co/runwayml/stable-diffusion-v1-5.

Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10: 1755–1758, 2009.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

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.