k-SALSA: k-anonymous synthetic averaging of retinal images via local style alignment (Supplementary Materials)

1 Benchmark datasets

The APTOS dataset includes 3,662 labeled fundus images. We split the data into a training set of 3,000 images and a test set of 662 images. The EyePACS dataset includes 35,126 labeled fundus images. We split the data into a training set of 28,100 images and a test set of 7,026 images. We rescaled the images in both datasets to 512-by-512 RGB pixels. The training set is used to obtain the GAN and GAN-Inversion models.

2 Implementation details

GAN. We trained our GAN models using the official PyTorch implementation[1] of StyleGAN2-ADA [6]. We set the number of mapping networks to two as recommended based on our image resolution and GPU count. Since the desired size of the generated images is 512×512, we used 16 progressive layers, resulting in the latent space W with dimensions 16×512. We train the models on 5,000 kimgs in each dataset with batch size 64, using 8 3090-RTX GPUs, Pytorch 1.7.1, CUDA 11.1, and CuDNN 8.1.1.

GAN Inversion. To obtain the GAN inversion encoder we built upon the official PyTorch implementation[2] of ReStyle [1]. We incorporated the MOCO-based [3] similarity loss on pSp [12] architecture with ResNetBackboneEncoder [4]. We trained each model for 100,000 iterations with a batch size of 8 and 5 refinement iterations per batch. Similar to the GAN setting, the output image size is set to 512×512. We performed the training using 1 3090-RTX GPU with the same environment as that of GAN training.

***k*-SALSA.** We split the intermediate-level features into a grid of 4×4 patches (16 in total) to construct the local style features in all settings. For the relative ratio of $\mathcal{L}_{\text{content}}$ and $\mathcal{L}_{\text{style}}$, we set the parameter λ to 0.1, 0.05 and 0.03 for $k = 2, 5, 10$, respectively, in APTOS. For EyePACS, we set λ to 0.01, 0.02 and 0.01, respectively for each k. We optimize our model for synthetic averaging using standard stochastic gradient descent with Adam [7], with learning rate 0.1 and $\beta_1 = 0.9$, $\beta_2 = 0.99$. We use the same computational environment as above with a single GPU.

[1] https://github.com/NVlabs/stylegan2-ada-pytorch
[2] https://github.com/yuval-alaluf/restyle-encoder

Downstream classification. For all synthetic datasets, we trained a DR classifier using the ResNet50 model [5] with batch size 32, 60 epochs, stochastic gradient descent (SGD) with Nesterov momentum 0.9 [10], weight decay 0.0005, and cosine annealing in the learning rate schedule [9].

3 Computational costs

One-time pre-training of GAN and inversion models took 10 hrs and 3 days, respectively, for APTOS, and 30 hrs and 3 days for EyePACS. The main runtime of k-SALSA depends on the inference speed of GAN and inversion, only around 0.85 secs/image. Synthetic averaging takes 19 secs/cluster (k=5). Both steps can be parallelized. Same-size clustering takes <1 min. Cosine similarity is computed for 16 patches/image (4x4) for 2 ms/image. Overall, we expect k-SALSA to be practical in realistic settings.

4 Choice of similarity metric for local style alignment

Here we provide additional empirical results supporting our choices of similarity metric in the local style alignment. Recall that once the local style features are constructed for each batch, we find the optimal matches between the target and source images using cosine similarity (COS). Given these optimal matches, we then use mean squared error (MSE) to optimize the synthetic average (see the definition of $\mathcal{L}_{\text{style}}$ in Eq. 7), effectively transferring the local styles from the source image to the target average. We note that these choices are inspired by prior works; dense contrastive learning [13] uses the COS metric to perform the alignment, whereas style transfer [2] is typically done using the MSE—our work combines these two approaches and keeps the respective similarity metrics. As shown in Supplementary Table 1, using either of COS and MSE metrics for both components of the model (COS-COS or MSE-MSE) results in worse averaging performance as measured by the downstream classification task evaluated in our work, which intuitively captures how well the clinically relevant features are preserved in the averaged images. This result supports our hybrid use of both metrics.

Supplementary Table 1: Comparison with alternative similarity metrics in local style alignment with respect to classification performance (APTOS, $k = 5$)

Method	**Accuracy**	**Cohen's κ**
MSE-MSE	0.599	0.717
COS-COS	0.708	0.727
k-SALSA (COS-MSE)	**0.712**	**0.769**

5 Addressing the reduced size of synthetic dataset

Dataset size is an important issue in medical imaging problems. The reduced number of images in the synthetic average dataset constructed by k-SALSA is a potential concern. To mitigate the cost of dataset reduction, we investigated an extension of k-SALSA based on data augmentation, whereby small random noise is added to k-SALSA's average embeddings to generate multiple "views" of each cluster. As shown in Supplementary Table 2, with 5 augmented images per cluster, we observed an improved performance of 0.829 (originally 0.769) for k-SALSA, and 0.809 (0.745) for k-Centroid in APTOS, k=5. This demonstrates the potential of our extension in countering the size reduction of the synthetic dataset with data augmentation. Importantly, the augmented images are independent of private images conditioned on k-SALSA's representative embedding of each cluster, thus there is no additional privacy leakage.

Supplementary Table 2: Addressing the reduced size of synthetic dataset

Method	**Metric**	**APTOS**, $k = 5$	
		Augmented (5×)	Non-Augmented
Centroid	Cohen's κ	0.809	0.745
k-SALSA	Cohen's κ	0.829	0.769

6 Additional comparisons with Original/GAN-Inverted

In our main experiments, we subsampled the original and GAN-inverted images to match the number of training images for the classifiers for comparison with different synthetic average datasets. For completeness, here we include the "best-case scenario" classification performance results for these baselines by using the full training dataset in APTOS. The results are shown in Supplementary

Supplementary Table 3: Classification performance with and without subsampling and using data augmentation with k-SALSA

Method	**Metric**	**APTOS**		
		Full	Subsampled ($k = 5$)	Augmented
Original	Cohen's κ	0.914	0.888	-
GAN-Inverted	Cohen's κ	0.857	0.828	-
k-SALSA	Cohen's κ	-	0.769	0.829

Table 3. We obtained a Cohen's κ of 0.914 and 0.857 for Original and GAN-Inverted, respectively, without subsampling, compared to 0.888 and 0.828 with subsampling with $k = 5$ (i.e., a 20% sampling rate), respectively. These results suggest that, while subsampling does reduce the performance, the impact is relatively small and that k-SALSA performance is still competitive with the best-case scenario, especially when used with our data augmentation strategy described in the previous section.

7 Performance dependence on the cluster size k

To further investigate the effect of cluster size k on downstream classification performance, we compared the classifiers trained on k-SALSA synthetic datasets for different values of k, but subsampled to the same number of clusters. Note that larger k leads to fewer clusters and thus smaller training data for classification. At the same time, larger k increases the potential to retain more salient features from source images. As shown in Supplementary Table 4, we observed a performance improvement for larger k in APTOS, suggesting that summarizing key features across multiple images has a beneficial impact on the classifier training.

Supplementary Table 4: Performance dependence on the cluster size k

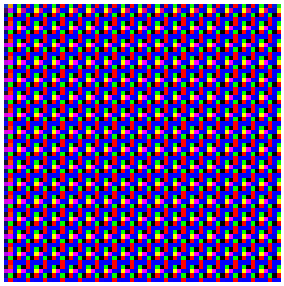
Method	**Metric**	**APTOS**		
		$k = 2$	$k = 5$	$k = 10$
k-SALSA	Cohen's κ	0.688	0.740	0.761

8 Challenges of differentially private GANs

To evaluate the performance of differentially private GAN approach to synthetic generation of retinal images, we trained one of the state-of-the-art models, G-PATE [8], on our fundus image datasets (APTOS) using the official TensorFlow implementation[3]. We follow the setting in the provided code except for the number of teacher networks and batch size, which we changed to 600 and 32 (from 2000 and 64), respectively, to reflect the smaller sizes of our datasets. In order to use the provided code, we downscaled the retinal images to 64×64. Otherwise, we used the following default parameters: number of epochs 1000, σ threshold 600, σ 100, step size 10^{-4}, max ϵ 100 and z-dimension 100. As shown in Supplementary Figure 1, the generated images from the differentially private GAN, even with a lenient privacy parameter of $\epsilon = 100$, are far from resembling retinal

[3] https://github.com/AI-secure/G-PATE

images. We attribute this failure in training to the relatively small size of our datasets (e.g. 3000) and the high resolution of the images, compared to handwritten digit images considered in the original work. Note that the GAN architecture used by G-PATE is DC-GAN [11], which is expected to have difficulties in the limited-data, the high-resolution setting given its low representation power, instability, and vanishing gradients compared with more recent techniques such as StyleGAN2-ADA [6]. Consistent with the visual assessment, the Fréchet Inception Distance (FID) of the images generated by G-PATE is 441.23, which is vastly higher than that of our approach (20.09), indicating the challenges of differentially private training of GANs in our setting.



Supplementary Figure 1: 64×64 generated images from the G-PATE model [8] trained on retinal images (APTOS).

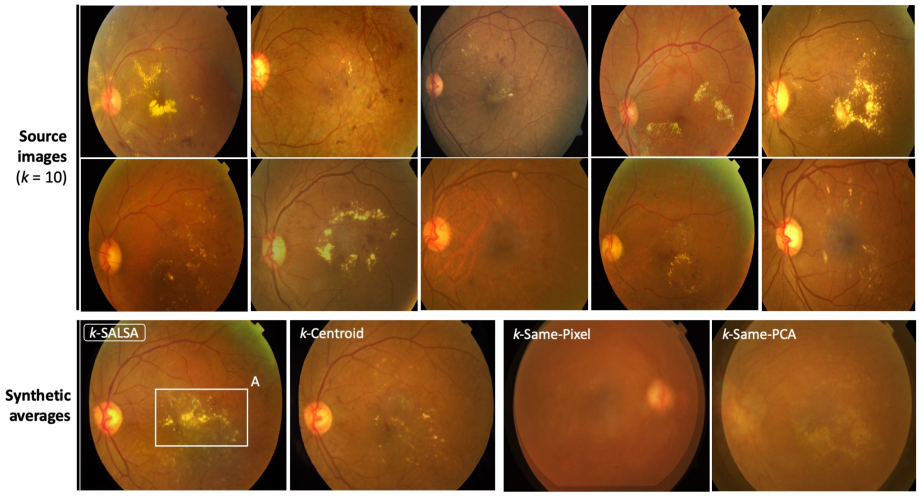
9 Choice of visual fidelity metric

To evaluate the fidelity of synthetic images, we use the Fréchet inception distance (FID), a standard metric for images generated using GANs. Other common metrics include PSNR and SSIM; however, these metrics quantify the degradation of quality when a source image is transformed, whereas FID measures a distributional similarity to a set of reference images based on high-level activations. FID uniquely assesses whether k-SALSA images are realistic compared to real retinal images. Moreover, k-SALSA introduces spatial flexibility of image features, which is not captured by pixel-level metrics like PSNR and SSIM.

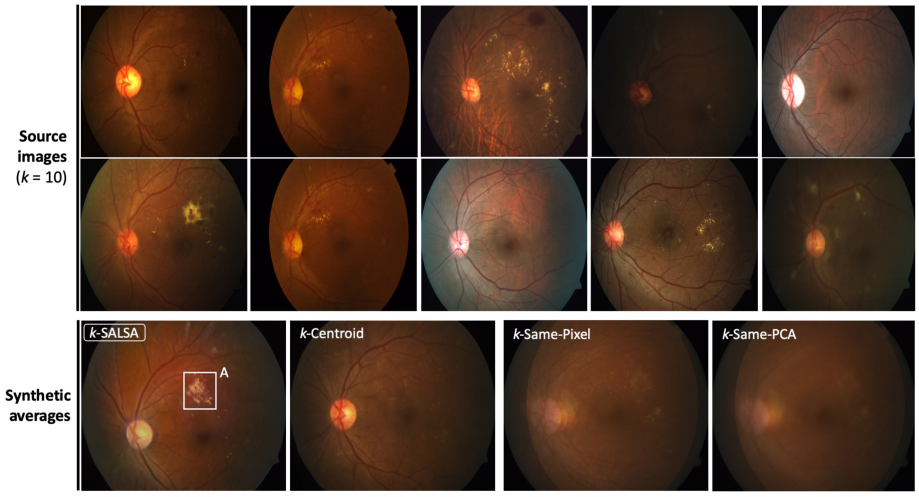
10 Additional examples of synthetic averages

To complement the main results, here we include additional synthetic averages of APTOS images along with their real source images for each $k \in \{2, 5, 10\}$ (4 examples each): Supplementary Figs. 2–5 for $k = 10$; Supplementary Figs. 6 and 7 for $k = 5$; and Supplementary Figs. 8 and 9 for $k = 2$. The results from the EyePACS dataset are analogous. Note that the figures for both $k = 2$ and $k = 5$ include two examples per figure. Source images (*top*) represent the k real original images in the identified cluster, and the synthetic averages (*bottom*) are

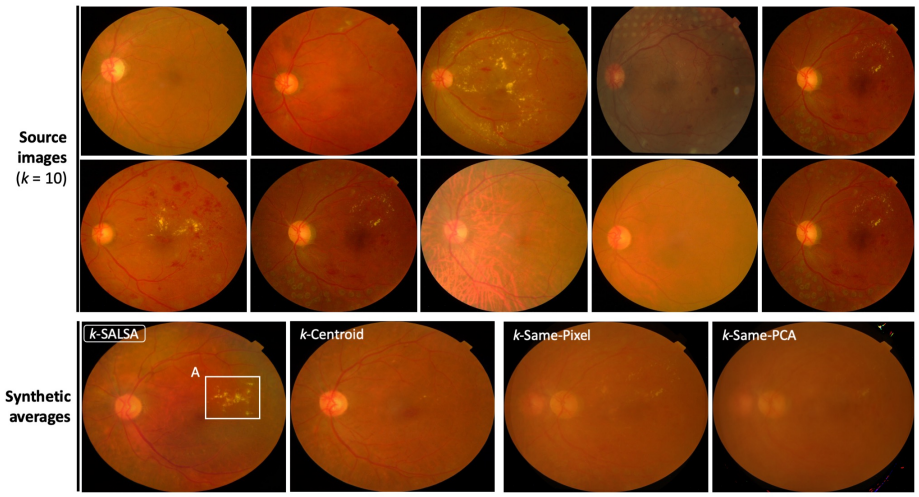
generated using k-SALSA, k-Centroid, k-Same-Pixel, k-Same-PCA, respectively. Overall, k-SALSA can better detect clinically relevant features in all cases.



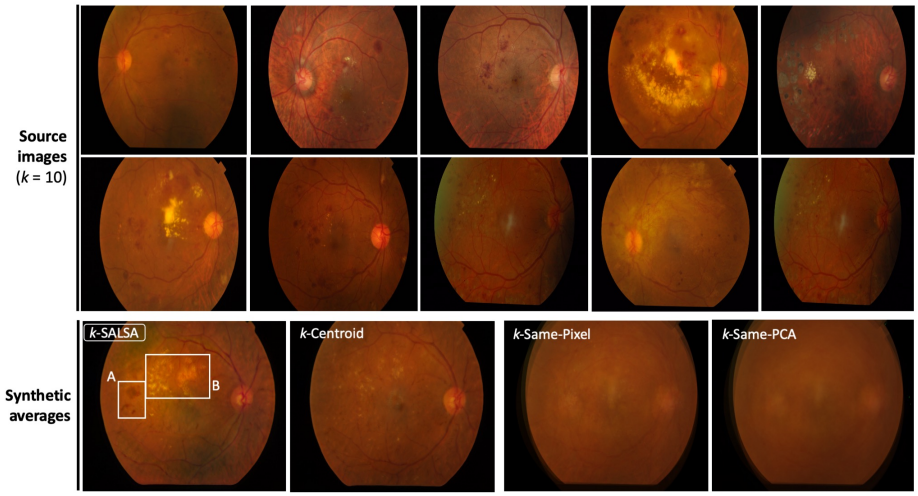
Supplementary Figure 2: **Examples of synthetic average of retinal images ($k = 10$).** One example of $k = 10$ real images (*top*) along with synthetic averages generated by different methods (*bottom*). k-SALSA better captures a disease-related feature (A).



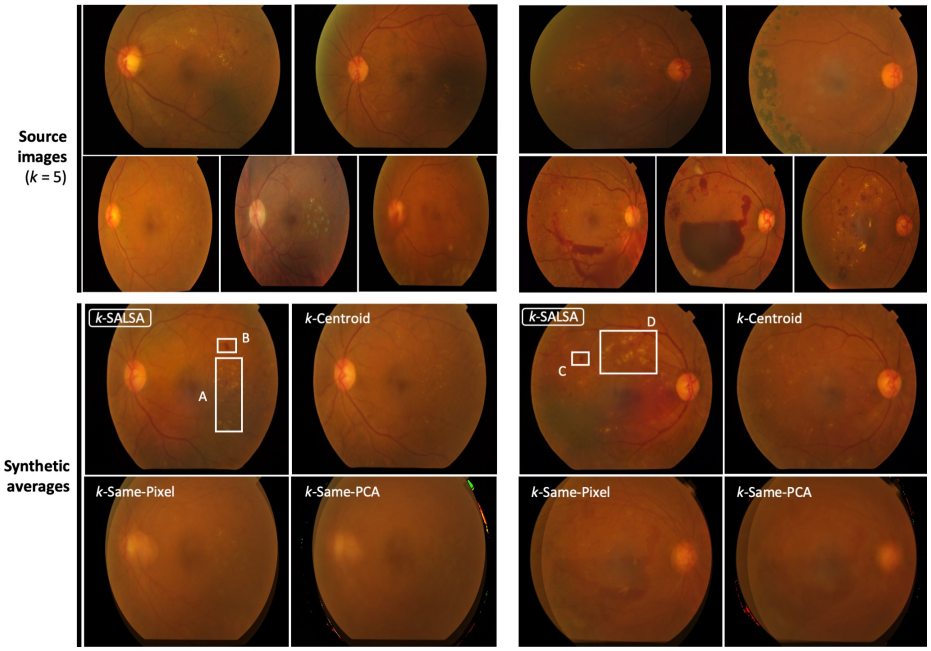
Supplementary Figure 3: **Examples of synthetic average of retinal images ($k = 10$).** One example of $k = 10$ real images (*top*) along with synthetic averages generated by different methods (*bottom*). k-SALSA better captures a disease-related feature (A).



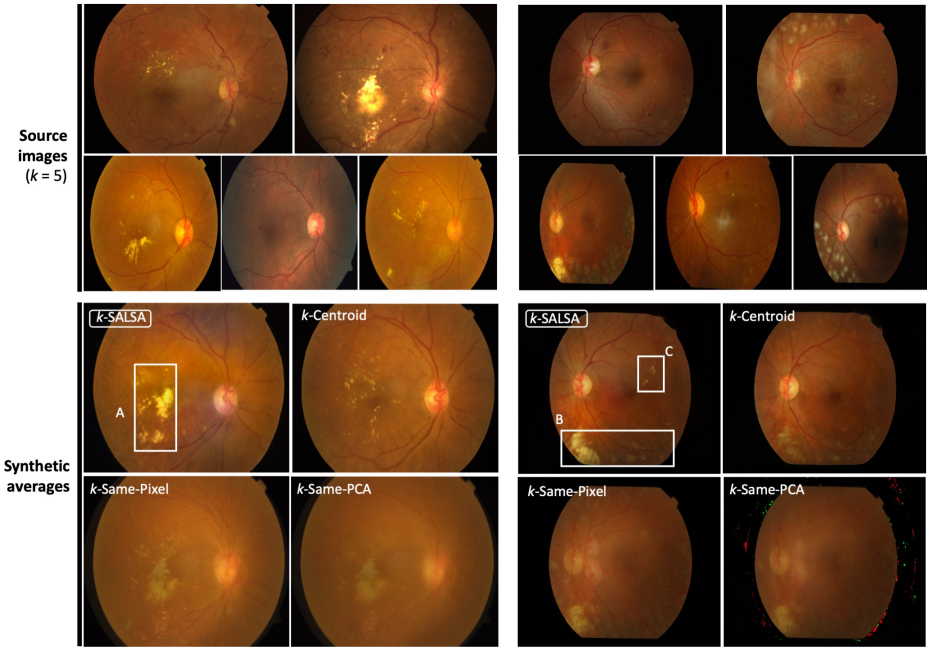
Supplementary Figure 4: **Examples of synthetic average of retinal images ($k = 10$).** One example of $k = 10$ real images (*top*) along with synthetic averages generated by different methods (*bottom*). k-SALSA better captures a disease-related feature (A).



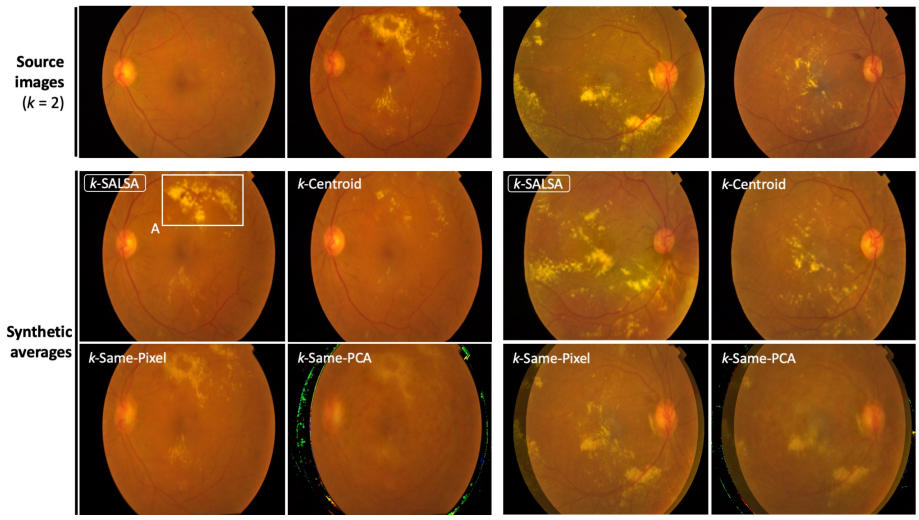
Supplementary Figure 5: **Examples of synthetic average of retinal images ($k = 10$).** One example of $k = 10$ real images (*top*) along with synthetic averages generated by different methods (*bottom*). k-SALSA better captures disease-related features (A, B).



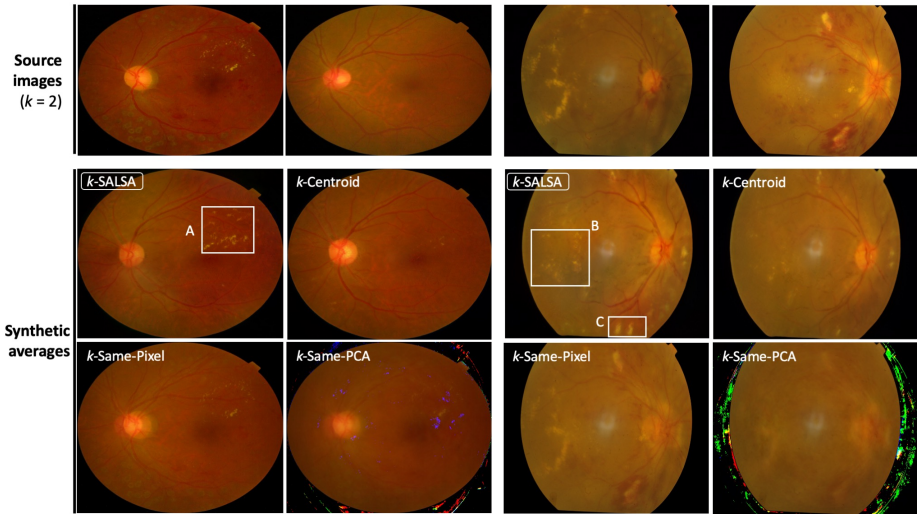
Supplementary Figure 6: **Examples of synthetic average of retinal images ($k = 5$).** Two examples of $k = 5$ real images (*top*) along with synthetic averages generated by different methods (*bottom*). k-SALSA better captures disease-related features (A, B, C, D).



Supplementary Figure 7: **Examples of synthetic average of retinal images ($k = 5$).** Two examples of $k = 5$ real images (*top*) along with synthetic averages generated by different methods (*bottom*). k-SALSA better captures disease-related features (A, B, C).



Supplementary Figure 8: **Examples of synthetic average of retinal images ($k = 2$).** Two examples of $k = 2$ real images (*top*) along with synthetic averages generated by different methods (*bottom*). k-SALSA better captures a disease-related feature (A).



Supplementary Figure 9: **Examples of synthetic average of retinal images ($k = 2$).** Two examples of $k = 2$ real images (*top*) along with synthetic averages generated by different methods (*bottom*). k-SALSA better captures disease-related features (A, B, C).

References

1. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
2. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
3. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
6. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Proc. NeurIPS (2020)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Long, Y., Wang, B., Yang, Z., Kailkhura, B., Zhang, A., Gunter, C.A., Li, B.: G-pate: Scalable differentially private data generator via private aggregation of teacher discriminators. NeurIPS 2021 (2021)
9. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
10. Nesterov, Y.: Introductory lectures on convex optimization: A basic course. Springer Science & Business Media (2003)
11. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
12. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2021)
13. Wang, X., Zhang, R., Shen, C., Kong, T., Li, L.: Dense contrastive learning for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)