Style Quantization for Data-Efficient GAN Training

Supplementary Material

A. Analysis and discussion

A.1. Sinkhorn divergence algorithm

The optimal transport (OT) problem is formally defined as follows:

$$\begin{aligned} D(\boldsymbol{p}, \boldsymbol{q}|\mathbf{C}) &= \min_{\gamma \in \mathbb{R}^{n\times m}} \langle \gamma, \mathbf{C} \rangle \\ &= \min_{\gamma \in \mathbb{R}^{n\times m}} \sum_{i=1}^{n}\sum_{j=1}^{m} \gamma_{ij} c(\mathbf{f}_i^x, \mathbf{f}_j^y), \\ \text{subject to} \quad & \sum_{j=1}^{m} \gamma_{ij} = p_i \quad \forall i \in \{1, \dots, n\}, \\ & \sum_{i=1}^{n} \gamma_{ij} = q_j \quad \forall j \in \{1, \dots, m\}, \end{aligned} \tag{1}$$

where $c(\mathbf{f}_i^x, \mathbf{f}_j^y)$ represents the cost of transporting a unit of mass from feature $\mathbf{f}_i^x$ to $\mathbf{f}_j^y$, and $\langle \gamma, \mathbf{C} \rangle$ denotes the Frobenius inner product of γ and the cost matrix $\mathbf{C}$.

Solving this optimization problem directly can be computationally prohibitive, especially in high-dimensional feature spaces. To mitigate this issue, we incorporate entropic regularization, which leads to the Sinkhorn distance and smooths the optimization landscape. The regularized OT problem is formulated as:

$$\begin{aligned} D_\epsilon(\mathbf{p}, \mathbf{q}|\mathbf{C}) &= \min_{\gamma \in \mathbb{R}^{n\times m}} \langle \gamma, \mathbf{C} \rangle - \epsilon h(\gamma), \\ \text{subject to} \quad & \gamma \mathbf{1}_m = \mathbf{p}, \quad \gamma^\top \mathbf{1}_n = \mathbf{q}, \\ & \gamma \in \mathbb{R}_+^{n\times m}, \end{aligned} \tag{2}$$

where $h(\gamma) = -\sum_{i,j} \gamma_{ij} \log \gamma_{ij}$ is the entropy of γ, and $\mathbf{1_m}$, $\mathbf{1_n}$ are all-ones vectors. The Sinkhorn distance allows for more efficient optimization through the iterative Sinkhorn-Knopp algorithm, where the optimal transport plan γ is updated iteratively as:

$$\gamma^{(t)} = \text{diag}(\mathbf{u}^{(t)}) \mathbf{K} \text{diag}(\mathbf{v}^{(t)}), \tag{3}$$

where $\mathbf{K} = \exp(-\mathbf{C}/\epsilon)$ and $\mathbf{u}$ and $\mathbf{v}$ are updated as:

$$\begin{aligned} \mathbf{u}^{(t+1)} &= \mathbf{p} \oslash (\mathbf{K}\mathbf{v}^{(t)}), \\ \mathbf{v}^{(t+1)} &= \mathbf{q} \oslash (\mathbf{K}^\top \mathbf{u}^{(t+1)}), \end{aligned} \tag{4}$$

with $\oslash$ denoting element-wise division and the initial condition $\mathbf{v}^{(0)} = \mathbf{1}$.

A.2. Detailed architecture design

To ensure compatibility with the Transformer architecture in our approach, the quantized sub-vectors $\{\hat{\boldsymbol{w}}_i^q\}_{i=1}^m \subset \mathbb{R}^{m\times s\times (d_w/s)}$ are processed through a learnable MLP layer, which adjusts them to match the embedding dimension of the input tokens. These transformed variables are then input into a Transformer model, utilizing the CLIP [11] text encoder. In the original CLIP model, absolute positional embeddings are added to the token embeddings to encode positional information within the sequence. As depicted in Fig. 1, to align the pre-trained positional embeddings with the size of our transformed variables, We modify the context length in the CLIP model to correspond to the number of quantized sub-vectors, denoted by s, and interpolate the pre-trained positional embeddings to match this new length. This adjustment also necessitates corresponding changes in the attention mask within the Transformer, ensuring proper functionality.

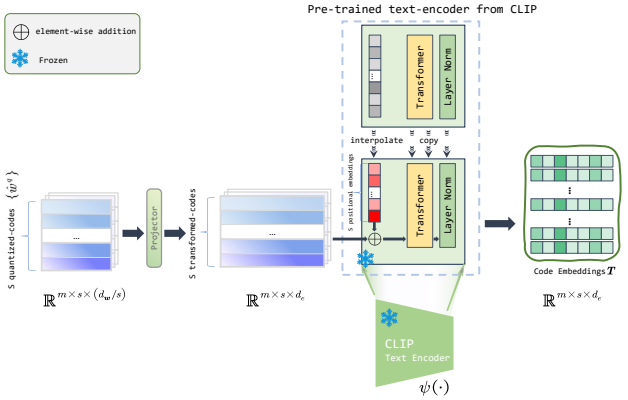


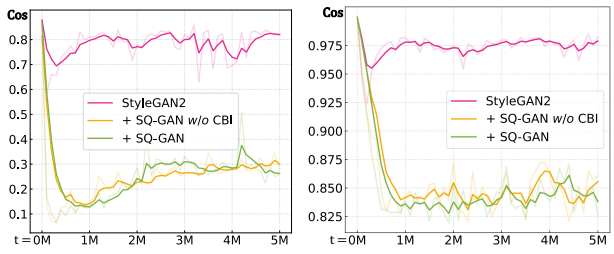
Figure 1. Modifications to the positional embedding of CLIP's text encoder.

A.3. Usage of codebook

Our primary objective in this work is to learn a compact and thoroughly explored latent space. To achieve this, we aim to maximize the activation of codes within the codebook. The utilization rate of the codebook is defined as the proportion of active codes. Our results (Table 4 in primary text) show that smaller code dimensions are more effectively utilized, likely due to the simplicity and decoupling of the information they encode. Furthermore, our codebook initialization method significantly enhances the utilization rate, indicating that embedded prior knowledge facilitates better use of the codebook information.

A.4. Embedding space

We evaluate the diversity within the discriminator's embedding space by analyzing the similarity between extracted features. Feature vectors are drawn from the discriminator's



(a) Discriminator (b) CLIP

Figure 2. Evolutions of the cosine similarity for the discriminator's embedding space and CLIP's embedding space.

penultimate layer, and cosine similarity is computed across all pairs of feature vectors in the dataset. By tracking the average cosine similarity throughout training, we gain insights into the evolution of the embedding space. Fig. 2a illustrates the cosine similarity's progression in the discriminator's embedding space. Compared to the baseline method, our approach demonstrates a lower average cosine similarity, indicating a more diverse and discriminative embedding space.

Additionally, we evaluate the similarity between the CLIP model's features on the generated images, as shown in Fig. 2b. The results reveal that our method achieves a lower average cosine similarity, indicating greater diversity in the generated images, aligning with our approach's objectives.

B. Experimental settings

B.1. Datasets

Our experiments are conducted using four distinct datasets: Oxford-Dog (derived from the Oxford-IIIT Pet Dataset [9]), Flickr-Faces-HQ (FFHQ) [4], MetFaces [5], and BreCaHAD [1]. Detailed descriptions of each dataset are provided below:

B.1.1. OxfordDog

We utilize dog images from the Oxford-IIIT Pet Dataset [9]. A dog face detection model is employed to crop and standardize these images to a uniform resolution of 256×256 pixels, ensuring the dog's face is centered as accurately as possible. A total of 4,492 images are randomly selected for training, while the remaining 498 images constitute the test set. This dataset includes approximately 25 dog breeds, presenting a challenging variety of poses and backgrounds, thereby offering a diverse set of conditions for our experiments.

B.1.2. FFHQ

The Flickr-Faces-HQ (FFHQ) dataset [4] comprises 70,000 high-resolution images of human faces, representing a broad spectrum of ages, ethnicities, and backgrounds. These images, sourced from Flickr, have been meticulously aligned and cropped to ensure high consistency and quality across the dataset. The dataset also includes various accessories such as eyeglasses and hats, further enhancing its diversity.

B.1.3. MetFaces

The MetFaces dataset [5] contains 1,336 high-resolution images of faces from the Metropolitan Museum of Art's collection[1]. These images are used primarily for research and analysis in facial representations across different artistic styles, making this dataset a unique resource for evaluating generative models.

B.1.4. BreCaHAD

The BreCaHAD dataset [1] is specifically designed for breast cancer histopathology research. It contains 162 high-resolution images (1360×1024 pixels) of histopathology slides. For our experiments, these images are restructured into 1,944 partially overlapping crops, each with a resolution of 512×512 pixels.

B.1.5. FFHQ-2.5k

We apply a pre-trained BLIP-base model [7] to the original 70,000 images from the FFHQ dataset to extract features. These features are then aggregated to facilitate the application of the K-means clustering algorithm. To simulate a low-data scenario, we set the number of cluster centers to K=14, resulting in an average of 5,000 images per cluster. We present the distribution of these clusters (Fig. 3) and visualize the features using t-SNE (Fig. 4). From these clusters, we select the smallest, comprising 2,500 images (referred to as FFHQ-2.5K), for further experimentation.

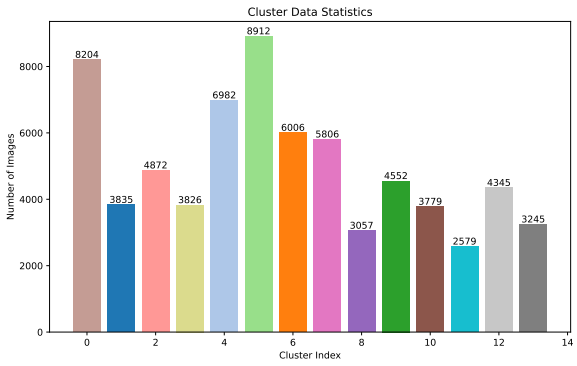


Figure 3. Clustering results for the FFHQ dataset.

B.2. Evaluation metrics

To evaluate the efficacy of our proposed method and benchmark it against existing baselines, we employ three widely recognized metrics: Inception Score (IS) [12], Fréchet Inception Distance (FID) [3], and Kernel Inception Distance

[1] https://metmuseum.github.io/

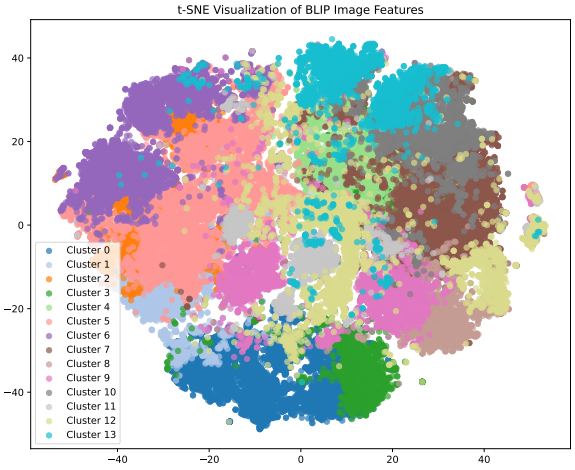


Figure 4. t-SNE visualization of features from the FFHQ dataset.

(KID) [2]. These metrics provide a comprehensive assessment of the quality, diversity, and distribution alignment of the generated images, enabling a thorough comparison of the models' performance. The following sections provide a detailed overview of each metric:

B.2.1. Inception Score (IS)

The IS metric [12] evaluates both the quality and diversity of the generated images. It is a widely recognized measure in the early stages of GAN development, which evaluates the generated images by analyzing the conditional entropy of class labels predicted by an Inception network, with a higher score indicating better performance.

B.2.2. Fréchet Inception Distance (FID)

The FID metric [3] is extensively used to measure the similarity between the distributions of generated and real images. It computes the Fréchet distance—also known as the Wasserstein-2 distance, between Gaussian distributions fitted to the hidden activations of an Inception network for both the generated and ground-truth images. FID is sensitive to both the quality and diversity of the images, with a lower score reflecting superior performance. It is considered a more comprehensive and reliable indicator than IS, particularly in capturing discrepancies in higher-order statistics.

B.2.3. Kernel Inception Distance (KID)

The KID metric [2] is similar to FID but offers distinct advantages. It measures the squared maximum mean discrepancy (MMD) between Inception representations of the generated and real images. Unlike FID, KID does not assume a parametric form for the activation distribution, and it provides a simple, unbiased estimator. This makes KID especially informative when the available ground-truth data is limited in scale. A lower KID score indicates better alignment between the generated and real data distributions, signaling higher image quality and consistency.

B.3. Implementation details

We utilized the implementations from [8] and [6] to train SNGAN and StyleGAN2, respectively. Additionally, we implemented a 64×64 version of DCGAN based on the approach outlined in [10]. All experiments were conducted using consistent hyperparameter settings across models, and performance was evaluated using the evaluation framework provided by [6]. During training, we adjusted the settings for λ_{sq} and λ_{qcr} to 0.01 each, reflecting the additional constraint terms introduced by StyleGAN2. It is worth noting that our results exhibited some discrepancies when compared to the scores reported in the literature, which may be attributed to variations in hardware or differences across experimental runs.

C. Additional Results

The superior performance of our method is further validated by the qualitative results presented in Figs. 5, 6, 7 and 8, which showcase the high-quality images generated by our approach. For consistency, identical hyperparameters and the same random seed were maintained across all experiments. The images presented were randomly selected from the generated outputs, with no specific selection criteria other than a global random seed. "Best FID" refers to images generated at the step with the best FID score. Our results indicate that our method produces more realistic images compared to baseline models.

References

[1] Alper Aksac, Douglas J. Demetrick, Tansel Ozyer, and Reda Alhajj. Brecahad: a dataset for breast cancer histopathological annotation and diagnosis. *BMC Research Notes*, 2019. 2

[2] Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD gans. In *6th International Conference on Learning Representations*, 2018. 3

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 6626–6637, 2017. 2, 3

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2

[5] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural*

Figure 5. Quantitative results on the OxfordDog dataset (best FID).

Figure 6. Quantitative results on the FFHQ-2.5k dataset (best FID).

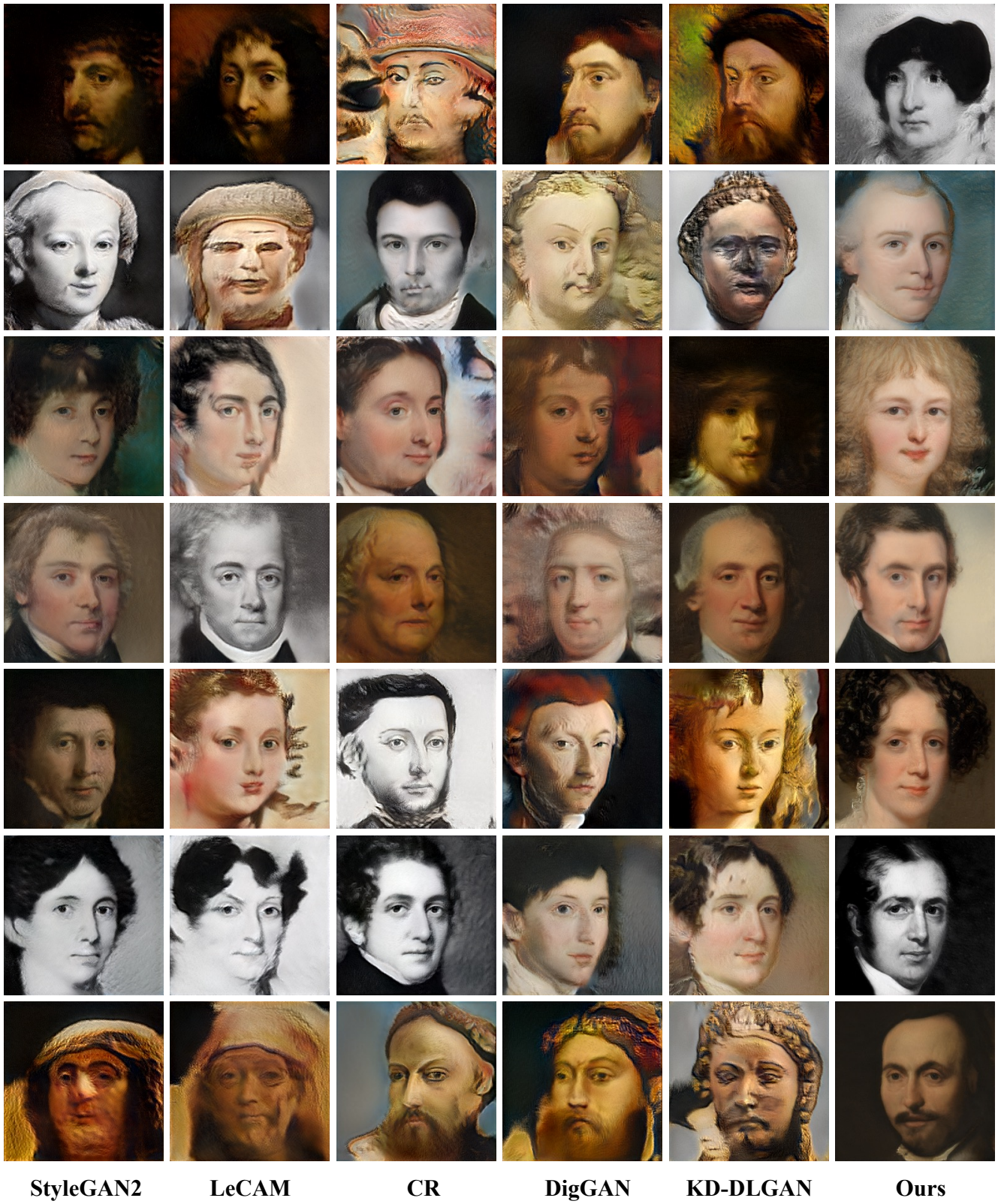


Figure 7. Quantitative results on the MetFaces dataset (best FID).

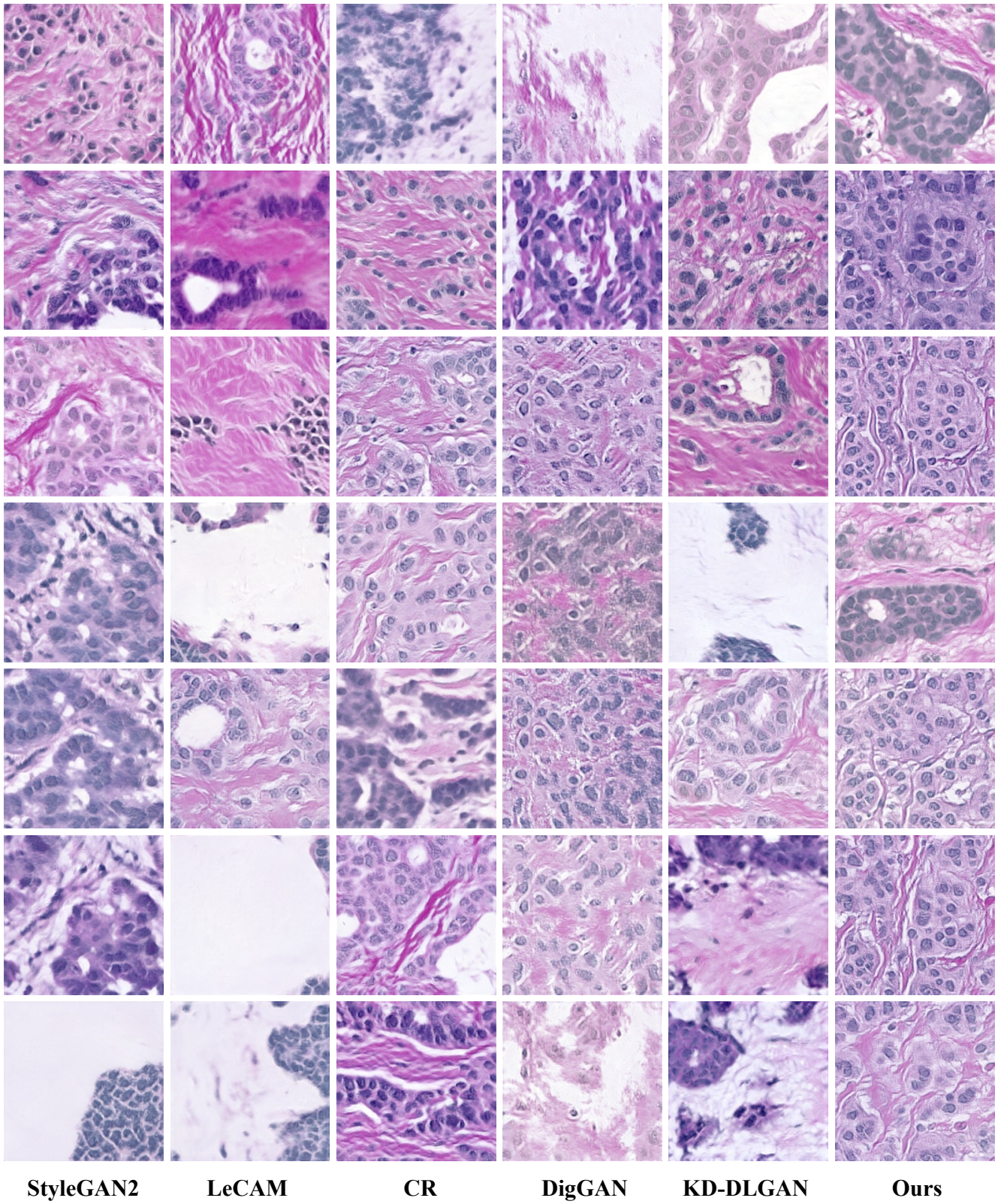


Figure 8. Quantitative results on the BreCaHAD dataset (best FID).

Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, 2020. 2

[6] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8107–8116, 2020. 3

[7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023*, pages 19730–19742, 2023. 2

[8] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *6th International Conference on Learning Representations*, 2018. 3

[9] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012. 2

[10] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations*, 2016. 3

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021. 1

[12] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pages 2226–2234, 2016. 2, 3