Cluster-guided Image Synthesis with Unconditional Models - Supplementary material

Clusters in the feature space

We showcase the structure of the feature space by visualizing the clusters that are formed for male and female faces using PCA. In particular, we sample images and manually annotate their apparent gender. We show the formed cluster in Fig. 1a below.

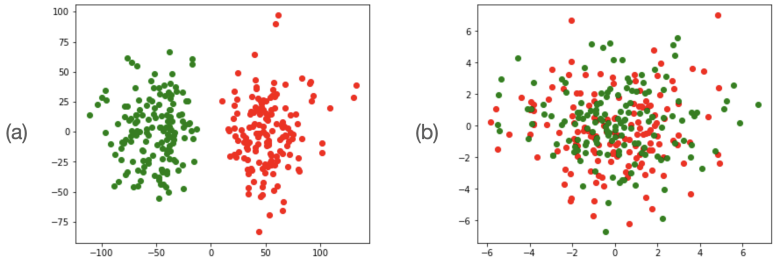

Figure 1: (a) Female (green) and male (red) images in the feature space. (b) The corresponding points in the latent space.

Motivation behind IMLE and alternative models

As shown in [4], maximizing the likelihood of the data (equivalent to minimizing $D_{KL}(p_{data}|p_{model})$) penalizes assigning low probability to data samples, while the alternative $D_{KL}(p_{model}|p_{data})$ penalizes implausible samples, which can lead to mode-dropping in GANs [1]. This is detrimental when modeling the conditional distribution of latent codes (Fig.1b). On the other hand, [4] show that under mild assumptions IMLE maximizes the likelihood of the data. To test our model choice, we replace IMLE with GAN and get a reduced facial pose accuracy of 87%. We also test VAE as an alternative likelihood-based model, which also drops accuracy to 90%. Contrary to IMLE, VAE and GAN focus on a subset of codes z and fail to maintain attribute consistency.

Image quality

Standard metrics (e.g., FID) compare the generated and GT image distributions. However, we cannot directly compare the conditional image distribution generated by our model to the unconditional real image distribution. Therefore, in order to measure the domain gap introduced by IMLE, we train it in an unconditional manner (i.e., using

only 1 cluster). We measure FID for CELEBA-HQ using 10k images: vanilla PGAN[1]: 8.79, Ours: 9.85.

Additional results on LSUN

In this section, we include additional results on LSUN. In particular, we include results for the horse and bird classes. We see on Figure 2 that for the horse class, one cluster contains images of the whole horse, with or without a rider, while the second cluster contains only the head of the horse in various poses. On the other hand, for the bird class, the first cluster contains diverse close-up images of birds, while the second cluster contains flying birds with sky background.

Figure 2: Results for the the horse and bird classes of LSUN.

Additional quantitative results

In addition to the experimental evaluation in the main paper, we provide accuracy for the synthesis of specific hair tone using PGAN [1]: [2]: 91%, [5]: 97%, Ours: 99%. Additionally, we utilise a classifier [3] to evaluate car orientation accuracy for our model: facing left: 97.8% and facing right: 96%.

Means of the clusters

We utilize k-means to perform clustering in the representation space. In Figure 3 we include a visualization of the cluster means for the attributes of pose, gender and hair length for PGAN on CelebA-HQ.

Unnatural clusters and failure cases

Our method utilizes the cluster assignments to condition the image generation. As a result, the content of the generated images depends on the quality of the clustering. The chosen clustering method (i.e., k-means) can be sensitive to outliers and can produce

[1] https://github.com/genforce/genforce

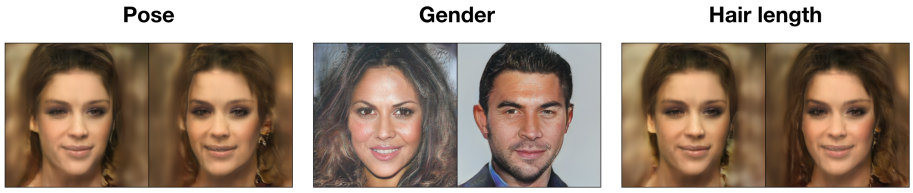

Figure 3: The means of the two cluster for pose, gender and hair length. We see that the means clearly demonstrate the encoded attribute.

clusters with no semantic attribute information. For example, in Figure 4 we show a cluster that only represents images of cats with captions on the top and bottom of the image. Since the images represent outliers, the quality of the images (and the text) is not optimal.

Figure 4: Generated images of cats with captions on the top and bottom of the image. This cluster represents outliers and does not represent any semantically meaningful attributes.

More than 4 clusters

With more clusters, multiple attributes are entangled in each cluster (e.g., pose+gender) which severely hinders the diversity of conditional synthesis due to the smaller support. Although the results are less accurate (per attribute) and less diverse compared to binary clustering per layer, this indicates that attributes appear as minor modes of variation in multiple layers.

Experimental details

In this section, we describe in more detail our experimental setup used to produce the quantitative results in order to further encourage reproducibility. For the classification results, we use 100 synthetic images for all baseline methods.

SeFa To generate the quantitative results for SeFA [5], we use the author's official code and pre-trained weights[2]. We manually identify the interpretable directions for each attribute considered–taking basis vector 1 for the 'gender' attribute, and basis vector 2 for 'pose'. We walk $\alpha := 0.75$ along this basis vector when generating the synthetic images.

GANSpace We train GANSpace [2] on the pre-trained ProgressiveGAN that is used for the other methods. We then take basis vector 1 for the 'gender' attribute, and basis vector 0 for the 'pose' attribute. We set $\alpha := 1.0$ for GANSpace to generate the edited latent codes in the target attribute.

Pretrained models In this section we present links for the pretrained models used in this paper:
PGAN-CelebA-HQ: https://github.com/facebookresearch/pytorch_GAN_zoo
PGAN-LSUN: https://github.com/genforce/genforce/blob/master/models/model_zoo.py
BigGAN-Imagenet: https://github.com/huggingface/pytorch-pretrained-BigGAN
StyleGAN-FFHQ: https://github.com/facebookresearch/pytorch_GAN_zoo

References

[1] Martin Arjovsky and Léon Bottou. "Towards principled methods for training generative adversarial networks". In: *arXiv preprint arXiv:1701.04862* (2017).

[2] Erik Härkönen et al. "Ganspace: Discovering interpretable gan controls". In: *arXiv preprint arXiv:2004.02546* (2020).

[3] Pirazh Khorramshahi et al. "A dual-path model with adaptive attention for vehicle re-identification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 6132–6141.

[4] Ke Li and Jitendra Malik. "Implicit maximum likelihood estimation". In: *arXiv preprint arXiv:1809.09087* (2018).

[5] Yujun Shen and Bolei Zhou. "Closed-form factorization of latent semantics in gans". In: *arXiv preprint arXiv:2007.06600* (2020).

[2]Official SeFA weights: https://github.com/genforce/sefa/blob/master/latent_codes/pggan_celebahq1024_latents.npy