Supplementary Material
Hierarchical Semantic Regularization of Latent Spaces in StyleGANs

Tejan Karmali[1,2*], Rishubh Parihar[1], Susmit Agrawal[1], Harsh Rangwani[1], Varun Jampani[2], Maneesh Singh[3†], and R. Venkatesh Babu[1]

Table of Contents

1 Implementation Details

Our implementation utilized the StyleGAN2-ADA [4] code. We perform all our experiments on 256×256 resolution images. Therefore, we use `paper256` architecture of StyleGAN2-ADA in all our experiments.

Warm-up. In order to leverage the rich feature space of pretrained networks using the generated images, we turn on the HSR regularizer after training the GAN for 500kimgs. By this point in the training, the GAN learns to generate images that start to look like the real images.

HSR. We use the feature extractor of ViT-DINO [2] to extract features. We resize the image to 224×224 before feeding into ViT-DINO. We use the intermediate output (after disarding CLS token) of its 3, 6, 9, and 12th transformer blocks, to supervise the generator's output at the 64, 32, 16, and 8 resolution respectively to align the semantics at various hierarchical levels, since ViT-DINO has been to shown to have a high-to-low level semantics emerging in its stack of transformer blocks [1]. We resize the generator's intermediate outputs to 14×14 before applying the loss function (l_2).

* Work done while at Indian Institute of Science
† Work done while at Verisk Analytics

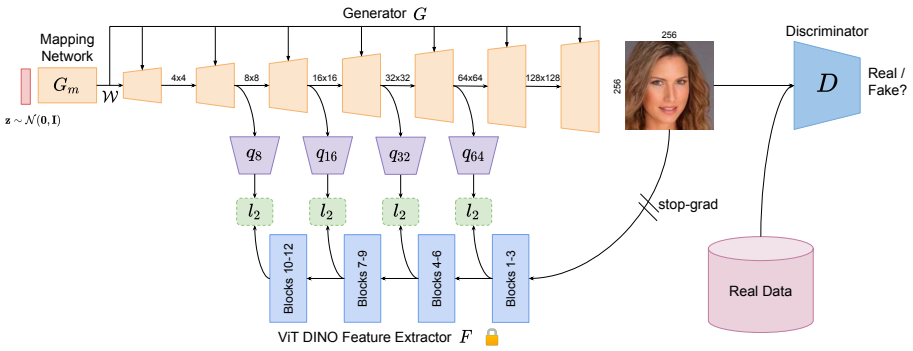

Fig. 1. Architectural Overview

2 Additional Quantitative Results

2.1 LSUN-Church

Table 1. Applying the HSR metric yields significant improvements over StyleGAN2 on quality as well as diversity metrics on the LSUN-Church dataset.

LSUN-Church	FID	Precision	Recall	PPL
StyleGAN2	4.08	**0.60**	0.34	916.15
+ HSR	**3.82**	**0.60**	**0.41**	**678.55**

In addition to the FFHQ-140k results presented in the main paper, we also perform experiment of adding HSR to StyleGAN2 training on LSUN-Church [9] which contains 1.2M images. We present the quantitative results of this experiment in Table 1. We observe a 25.93% improvement in the PPL over the baseline. Additionally, we also observe a significant boost in the recall metric [5], indicating increased diversity in the generated images.

2.2 Additional Metrics for Latent Space Evaluation

Table 2. Improved Disentanglement

	Disentanglement	Completeness	Informativeness
SG2-ADA	0.57	0.61	0.98
+HSR	**0.61**	**0.62**	0.98

To measure the quality of the $\mathcal{W}$ space after applying HSR, we use the DCI metric [3]. DCI stands for disentanglement, completeness, and informativeness. Disentanglement measures the extent to which each dimension in the latent captures at most one attribute. Completeness measures the extent of each attribute is controlled by at most one latent dimension. Informativeness measures the classification accuracy of the attributes using latent representation. We use the procedure same as Wu *et al.* [8] to compute DCI on $\mathcal{W}$ space. We present the results in Table 2, where it is observed that applying HSR improves the disentanglement in $\mathcal{W}$ space while also showing marginal gains in the completeness metric.

2.3 Evaluation on Real Images

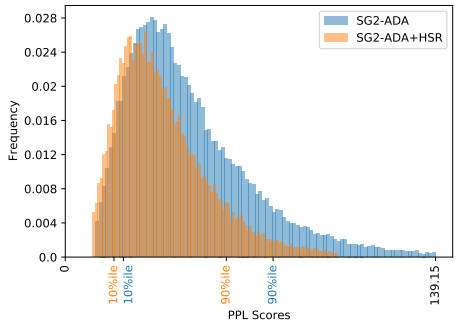

Fig. 2. Distribution of PPL scores over 50k real image pairs from CelebA-HQ [6]. Baseline: 46.29, Baseline+HSR: 36.80

Smoothness of Latent Space. In the main paper, we presented results of PPL over 50k generated image pairs. Generated images can also include qualitatively worse images from the learnt distribution, which elongates the tail in PPL distribution. To show PPL on perceptually better images, we use 50k pairs of real images and find their latents. We project 400 randomly sampled images from CelebA to obtain its latents in $\mathcal{W}+$ space. We then randomly sample pairs of inverted latents to compute PPL. We visualize the distribution of PPL scores in Fig. 2. Since real face images are used, the PPL is low as there are no out-of-distribution images (non-face, unnatural face, artefacts). Yet, it can be seen that PPL is lesser when HSR is applied (36.80), compared to the baseline which leads to larger average distance to traversed (46.29) while interpolating between 2 real images.

Reconstruction of Real Images. Table 3 presents the effectiveness of the StyleGANs to reconstruct/invert real images without and with the application of HSR. Reconstruction, as measured by PSNR and LPIPS metrics, improves when

HSR is applied, thus showing that the latent space obtained after regularizing by the HSR leads to more natural-looking images.

Table 3. Effect of HSR on reconstruction of inverted real images

Method	PSNR (↑)	LPIPS (↓)
SG2-ADA	47.04	0.2296
SG2-ADA+HSR	**47.16**	**0.2281**

2.4 Effect of HSR under Distribution of High Quality Images

Table 4 presents the results under the truncation trick [7], which is used to produce higher quality images. This comparison, using the FID, Precision-Recall, and PPL metrics in rows 3,4, shows that HSR leads to performance gains even on such higher quality generations. Truncation trades off diversity (recall) for increased quality (precision), irrespective of the use of HSR. On the other hand, using HSR on SG2 (or SG2-Trunc) significantly improves diversity (recall) and FID without sacrificing the quality (precision). Truncation on StyleGAN2, trained on FFHQ140k data, is performed using the commonly used truncation value of $\Psi = 0.7$.

Table 4. Comparison of quality and diversity metrics under Truncation

Method	FID (↓)	Precision (↑)	Recall (↑)	PPL (↓)
SG2	3.92	**0.68**	0.45	175.09
SG2+HSR	**3.74**	**0.68**	**0.48**	**144.59**
SG2-Trunc	21.46	0.83	0.22	109.96
SG2+HSR-Trunc	16.94	0.82	0.25	96.46

3 Qualitative results

3.1 Improvement in Worst Images

We have shown quantitatively that applying the HSR regularization improves the quality of worst images that the generator can produce. We also demonstrate this qualitatively in 2 ways. First, we compare the Mahalanobis distance between the generated images and the moments of the real data from a set of 5000 generated images. We present the results of 30 farthest images in Fig. 3. It can be seen that unnatural, non-face images are being generated by the baseline, which are virtually absent when HSR is applied. In the case of the LSUN-Church dataset,

the images lack in structural aspects related to churches and shows presence of unnatural colors in the image. While after applying HSR, the images reproduce the structure faithfully, for e.g., in the edifices.

Secondly, we present the results of images sampled from the bottom 10% according to the PPL score. We present these results in Fig. 4. A similar trend is observed with the presence of artefacts in and around the facial regions. In both cases (faces and churches), the artefacts are greatly reduced after the application of the HSR regularizer, making the images look more natural.

3.2 Linearity of Edits

We present the results of editing along the attributes like "gender", "age", and "smile" in Fig. 5. Note the significant improvement in linearity upon applying the HSR regularizer, as compared to the baseline.

4 Shortcomings of ALS

As noted in the main paper, we propose the ALS score to measure the linearity of change in the attributes. This requires the attributes should not be binary/ categorical but can be continuously varied. Common attributes like age, smile, gender, hair, beard, and bangs fit this description and hence they are a natural choice for evaluation using the ALS metric. Other attributes may be binary or categorical but they still may allow us to evaluate diversity. This is the case for attributes corresponding to "wearables", for *e.g.* eyeglasses, earrings, headgear etc. The quality of corresponding latent space w.r.t. this class of attributes is better measured with diversity-measuring metrics like recall [5].

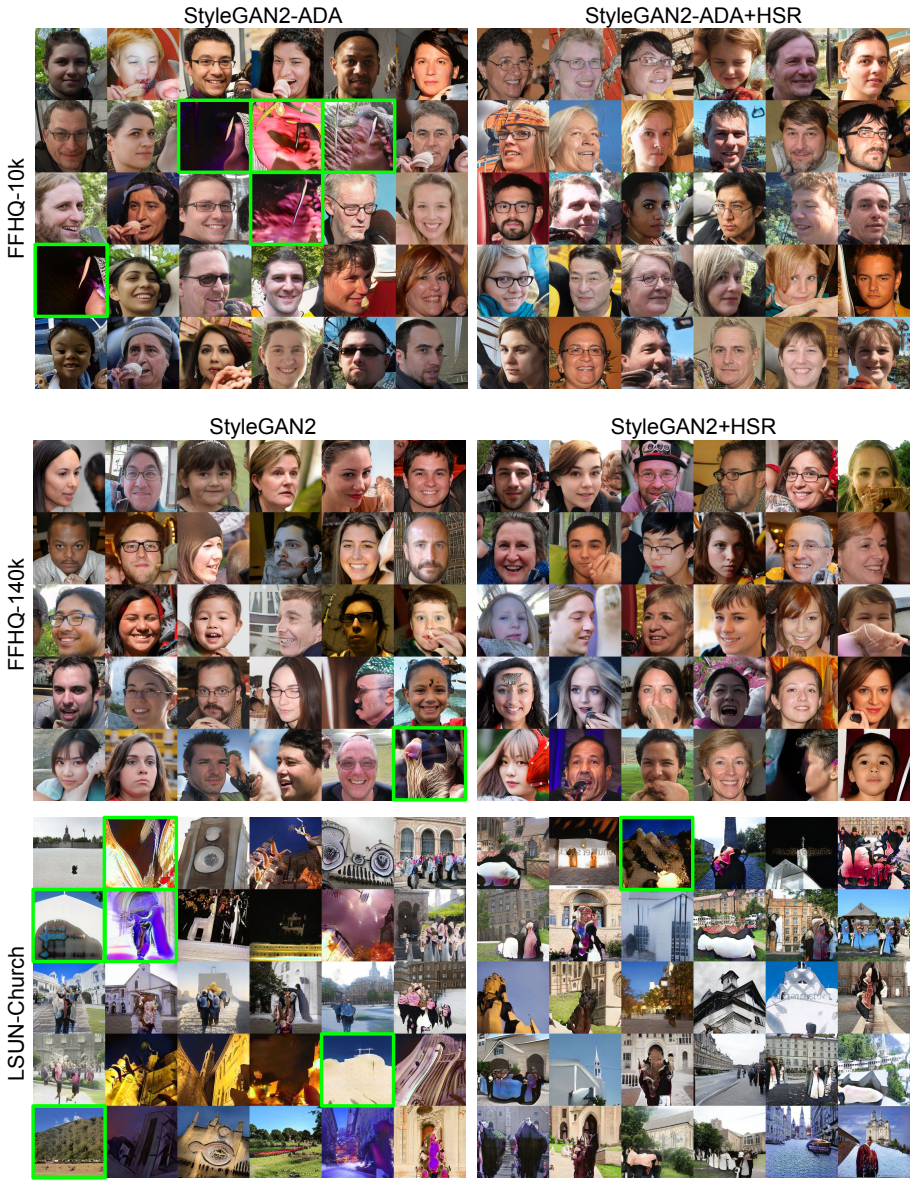

Fig. 3. Worst 30 Images according to the Mahalanobis distance to Inception moments of respective datasets. Highlighted images show structural irregularities in the respective image category (face/church).

Fig. 4. Worst Images according to the PPL scores. Highlighted images have high degree of artefacts.

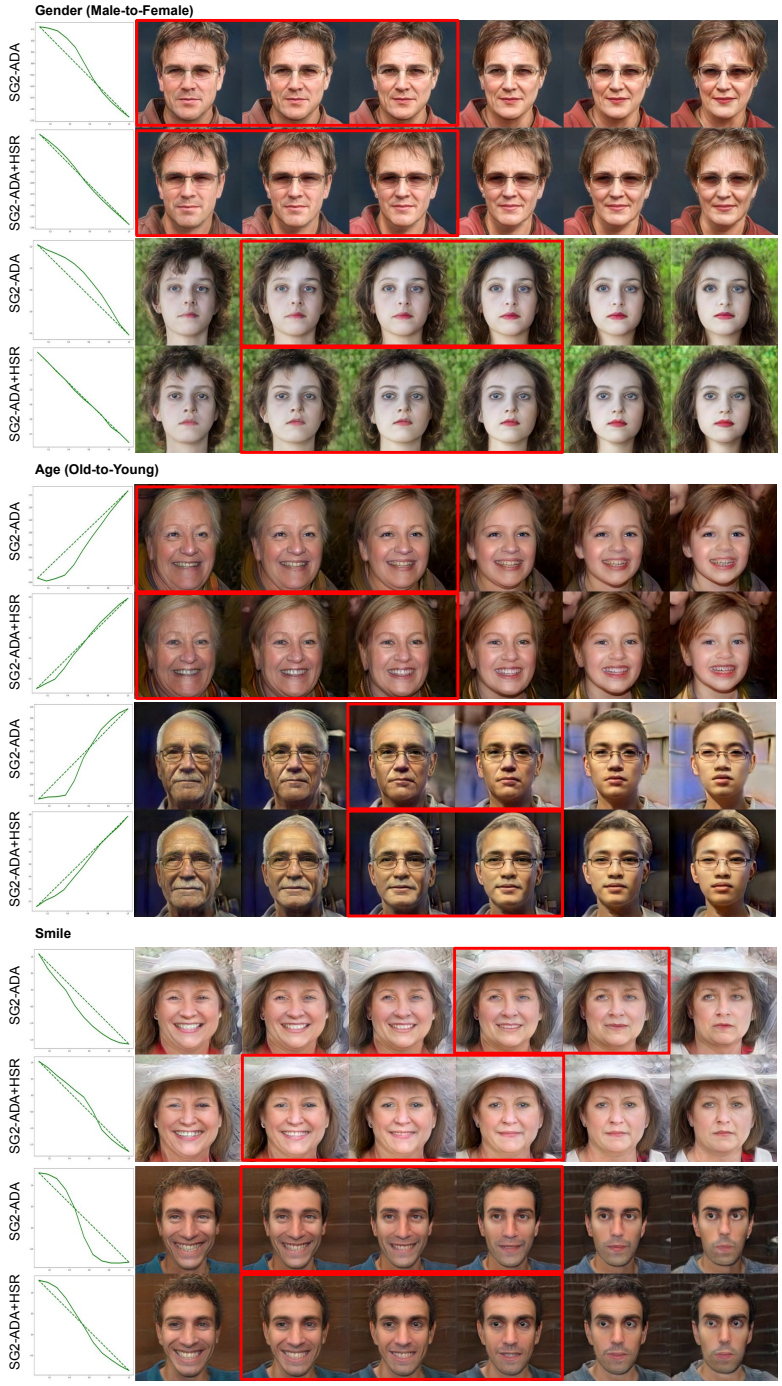

Fig. 5. Linearity of Edits. Plots indicate the intensity of attribute . We observe that the rate of change of a particular editable attribute is more close to the identity after applying HSR. In the red inset, we can observe that HSR has added additional smoothness in the transitions in comparison to the SG2-ADA.

References

1. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep vit features as dense visual descriptors. CoRR **abs/2112.05814** (2021), https://arxiv.org/abs/2112.05814 1
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 1
3. Eastwood, C., Williams, C.K.I.: A framework for the quantitative evaluation of disentangled representations. In: International Conference on Learning Representations (2018) 3
4. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Proc. NeurIPS (2020) 1
5. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. In: NeurIPS (2019) 2, 5
6. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015) 3
7. Marchesi, M.: Megapixel size image creation using generative adversarial networks. ArXiv **abs/1706.00082** (2017) 4
8. Wu, Z., Lischinski, D., Shechtman, E.: Stylespace analysis: Disentangled controls for stylegan image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021) 3
9. Yu, F., Zhang, Y., Song, S., Seff, A., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015) 2