Low bit rate generative face compression using inverse GAN

Leonardo Monchieri and Simone Milani
Dept. of Information Engineering, University of Padova, Padova, Italy
leonardo.monchieri@studenti.unipd.it, simone.milani@dei.unipd.it

***Abstract*—The paper investigates the adoption of generative adversarial networks (GANs) for low-bit-rate learned image compression, aiming to both preserve fine visual details and maintain semantic and biometric consistency. The designed strategy is based on the inversion of a StyleGAN image generation network and the characterisation of noise and visual features through strong quantisation and a side information channel. Experimental results show that the images can be stored with a lower amount of bits and present better visual details with respect to standard state-of-the-art learned coding schemes.**

***Index Terms*—low-bit rate, learned image compression, StyleGAN, Inverse GAN, face compression.**

I. Introduction

Learned image compression has recently experienced significant advancements and is widespread, thanks to the capability of high compression gains (surpassing traditional coders like JPEG or JPEG2000 [1]) and the adaptability that training procedures can achieve. For these reasons, several standardization efforts have been carried out, leading to the recent finalization of coding standards such as *JPEG AI* [1]. Most of the proposed solutions rely on the core autoencoder architecture, where an encoder network maps the input image into a set of latent features that can be reconstructed by a following decoder network. Parameters are optimized by a training process to implement an efficient transformation that both reduces data redundancy and grants a satisfying visual quality [2].

Unfortunately, like traditional coding schemes, such solutions fail in modelling high frequencies whenever the coding bit rate significantly lowers: reconstructed image looks highly blurred leading to the corruption of finer details. To overcome this, researchers have started investigating the possibility of characterizing images at very low bit rates employing generative strategies like Generative Adversarial Networks (GANs) [3] or Diffusion Models (DMs) [4], where the corrupted information can be re-generated by the decoding network itself. Although such solutions are capable of granting a very high perceptual quality, fidelity can not be guaranteed posing new challenges when adopted for semantic-oriented applications (e.g., object recognition) or biometrics (e.g., people identification [5], [6]). These issues have, therefore created a divide between *coding for humans* and *coding for machines* architectures whenever some generative strategy is going to be applied in the compression network.

This work investigates the possibility of overcoming such distinction by designing generative learned image coding systems that provide very good compression performances at low bit rates while preserving the semantic information carried by the original image. To this purpose, our architecture is based on the StyleGAN network [7] and the capability of inverting it [8], i.e., reversing the generation process that maps noise sequences into images by training an encoder network. Since StyleGAN images are modelled by the input pseudo-random noise and the set of style features that control generation at the different resolution layers, the approach proposed in this work reduces the coded bit rate by adopting strong quantization and signalling style features to the decoder with the transmission of a compressed low-resolution version of the input picture (side information).

In a nutshell, we can summarize the main innovations as follows.

1) We designed a scheme based on an Inverse-StyleGAN architecture [9] that first embeds an image into pseudo-random noise space and a set of style features, which are lately coded through strong quantization and side information generation.
2) We conducted a comparative analysis on face datasets showing that, although image fidelity can not be granted, the proposed solution allows a good perceptual quality in the reconstructed images (under different no-reference metrics) reducing the overall bit rate concerning other state-of-the-art learned compression strategies.
3) We verified that the proposed generative coding solutions preserve biometric identification capabilities.

In the following, Section II overviews some state-of-the-art strategies about learned compression and GAN inversion, while Section III presents the proposed coding schemes. Section IV reports some experimental evaluations and conclusions are drawn in Section V.

II. Related works

In recent years, learned image compression has been tackled using variational autoencoders [1], [2], [10], where an hourglass network generates a set of latent vectors whose features are regularized through a statistical prior that is used to estimate symbols probabilities. Despite this, the adopted distortion

This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, with a partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART"). The work was also supported by the SYCURI Project, funded by the University of Padua, in the Program "World Class Research Infrastructure".

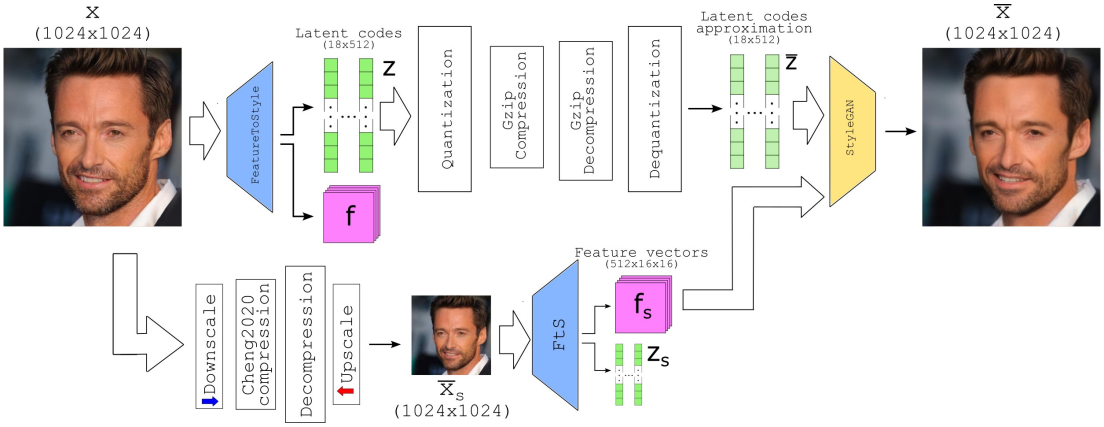

Fig. 1. *FeatureStyleEncoder* and *styleGAN* encoding-decoding approach

metrics and convolutional architectures lead to unsatisfactory perceptual quality at low bitrates as the reconstructed images display evident low-frequency artifacts. For this reason, generative compression algorithms have been developed involving GANs [11] or Diffusion models [4], [12]. These typically exploit generative networks to reconstruct a visually pleasing image at the price of higher reconstruction error.

As for Generative Adversarial Networks, the estimated latent space (i.e., the input noise sequence of the generator) provides an efficient framework for image editing [9] or quality enhancement after standard reconstruction [13]. Other works on GAN inversion show that this is useful in image inpainting [14], outpainting [15] or denoising [16] as well.

The approach in [3] was designed for attribute editing and finds a good trade-off among rate, distortion, and editability with an adaptive distortion alignment and an editing map. In [17], a 3D NeRF-based GAN is inverted to infer both camera pose and latent; the approach in [18] inverts a 3D GAN using symmetric prior.

As for diffusion models, several investigations focused on the inversion of DMs for image editing [19] or reconstruction [20]. Indeed, such approaches rely on the fact that such strategies can be inverted very easily (the inverse of the diffusion process is a diffusion process itself), but the computational requirements imposed by iteration make such algorithms less suitable for low bit-rate compressions (often the devices involved are low power).

Some additional research has been focusing on the adopted noise models. As an example, the work in [21] discusses the possibility of using heavy-tailed distribution in place of Gaussian values within the image generation process. The work in [22] investigates the impact of noise dimension on the final reconstructed image, while the solution in [23] shows how multivariate noise distributions permit handling multiple image categories. Furthermore, recent work such as [24] highlights the benefits of operating in a perceptually aligned latent space. Their Generative Latent Coding framework employs a semantically rich latent representation—learned via VQ-VAE—which better aligns with human perception and enables high-fidelity, high-realism compression at ultra-low bitrates.

Unlike the aforementioned work, this paper aims to develop a lightweight low-bit rate face coder that enables representing a single image employing inverse GAN latent compression.

III. The proposed architecture

A. Inverting a Generative Adversarial Network

An inverse GAN (IGAN) network is a neural architecture that reconstructs a latent noise z from an image x such that the associated generator $G(z)$ provides a realistic reconstruction of the original image, i.e. :

$$z^* = \underset{z}{argmin}\ \ell(G(z), x) \tag{1}$$

where ℓ is a distance metric used to compare the quality of the reconstruction of the image (typically, a fidelity metric).

In the considered framework, the inverse GAN IG maps the input image into a noise vector (encoder), while the original GAN transforms noise back into the image (decoder). To produce an optimal result, the training of IG [9] can be done over a pre-trained GAN (e.g., StyleGAN [7]) or simultaneously with the GAN. Note that, in the case of StyleGAN, inverting the generative architectures implies mapping the input image into a noise z (which can be considered an embedding for the finest details) and a set of style features f (spatial and structural characteristics) that operates at different resolution levels [9]. Such architecture resembles that of an *autoencoder* notably differing in the latent structure, which aims to preserve meaningful structure in its latent space, making it more interpretable or useful for tasks requiring semantic consistency.

Further compression over z and f has been taken into account to reach a lighter representation. Such possibility can be operated with reasonable complexity by approximating the noise z via an invertible pseudo-random sequence created via a Linear Congruential Generator (LCG) and designing an efficient inversion strategy that can map the noise z back to its originating seed value (see the following sections). Alternatively, noise can be approximated via a linear combination of pseudo-random noise sequences or scalarly-quantized into a sequence of integers. As for the specific use of an inverted

StyleGAN, the style features f can be transmitted by inferring such data from some properly transmitted side information.

B. From image to noise: designing and training the inverse GAN

In our implementation, GAN and IGAN were trained separately to avoid model collapse and the extra complexity due to simultaneous training.

Once the GAN has been created, IG parameters can be trained using the *MSE* loss function to preserve the noise structure:

$$\mathcal{L}_{MSE} = ||z - IG(G(\hat{z}))||^2 \; . \tag{2}$$

We considered a first vanilla implementation that used a simple DCGAN to generate small resolution images (64×64), where the generator and inverse network (encoder) have a structure similar to that of the discriminator. In the encoder (inverse generator), we replaced the activation functions of the last layers with a *hyperbolic tangent* (*Tanh*) to produce a latent sequence in the range $[-1, 1]$ (LCG/Uniform compliant).

Then, in order to test the proposed solution on larger images, we focused on a StyleGAN architecture capable of generating faces sized 1024×1024, using the implementation in [9].

C. Coding latent representation

To enhance storage efficiency and reduce computational costs, two parallel compression techniques are applied to the latent noise and style features used in the generation process: one designed for the latent noise sequences z, and another one which is specifically used to signal style features $\mathbf{f}$ to the decoder.

Coding latent noise
Given the noise z resulting from the inversion process, it is necessary to generate entropically efficient formatting to map it into a constrained bitstream. At the beginning of such process, sequences are normalized and centred ($z' = z - \mu_z$). The initial approach explored is based on the idea of representing any noise as a linear combination of LCG sequences, as seed parameters can be easily estimated. Therefore, an image is represented by pairs of seeds and weights

$$\dot{z} = w_1 LCG(s_1) + w_2 LCG(s_2) + ... + w_n LCG(s_n) \tag{3}$$

where $\dot{z}$ approximates $\hat{z}$ using a Gram-Schmidt orthogonalization procedure starting from w_1, s_1 up to w_N, s_N. The computational complexity can be tuned by selecting different seed datasets; in our work, we tested three different seed ranges.

Despite this, further experiments have shown that, at low-rate operating points, simple scalar quantization performs better, followed by a standard *gzip* compression.

Coding style information
To transmit style information with minimum impact of the final bitstream, we use a low-resolution version of the input image x as side information to infer f using the StyleGAN. The overall scheme is reported in Fig. 1, as can be seen, image x (1024×1024) is rescaled into x_s (256×256 or 512×512), and Cheng2020 compression is the applied on it [25]. At the decoder, the reconstructed x_s is upscaled to their original resolution and used to extract feature vectors (f_s) through the IG process. The combination of these extracted feature vectors with the de-quantized latent serves as the foundation for generating high-quality reconstructions.

In the following, we will refer to this compression set-up as proposed method (or proposed), in order to distinguish it from noise composition.

IV. Experimental results

Similarly to other works in literature, performance is measured using different reference and no-reference quality metrics[1]:

- Mean Squared Error (MSE) or the logarithmic equivalent for reconstruction accuracy Peak Signal-to-Noise Ratio (PSNR)
- Structural Similarity Index (SSIM)
- Learned Perceptual Image Patch Similarity (LPIPS)
- Fréchet Inception Distance (FID)
- Natural Image Quality Evaluator (NIQE)
- Deep bilinear convolutional neural network (DBCNN)
- Multi-scale image quality (MUSIQ).

The first four account for the fidelity in reconstruction (referenced), while the last three perform a no-reference evaluation of image quality. The generated bit rate was measured in bit-per-pixel (bpp).

To verify the efficiency of GAN inversion on faces, we first considered the synthetic dataset *Anime Faces* of 64×64 grayscale hand-drawn sketchy images; data were processed using a vanilla GAN since the limited size of images did not allow an easy adaptation to the most recent GAN networks. In this case, no style information is present, and therefore, we need to simply characterize the noise z via linear approximation. To this purpose, we considered two types of noise generators: pseudo-random Uniform and LCG. Moreover, quality results were compared with those of standard autoencoder (with the same later structure of the considered GAN for the sake of fairness), where the number of latent features corresponds to the number of basis noise and weights in the linear composition (thus equalizing the bit rate). This choice was motivated by the fact that small-resolution images are not effectively compressed by state-of-the-art learned codecs.

The results in Tab. (I) show that despite fidelity can not be ensured (as previous works highlighted) since AE embedding ensures minimum MSE and maximum SSIM, the sharpness and visual quality are much better for the GAN-based approach.

The approach has been tested on a face dataset made of pictures from real people (*Celeb HQ* dataset) using the *FeatureStyleEncoder* derived from the inversion of StyleGAN.

Experimental tests were performed on a selection of 20 images (10 male and 10 female) using the Cheng2020 model with quality 2 and the proposed method using style embedding

[1] The references to the different metrics can be found in [4].

TABLE I
MSE, SSIM, AND FID VALUES FOR IGAN CODERS (WITH UNIFORM AND LCG NOISE) COMPARED TO A STANDARD AUTOENCODER.

Composition of uniform noise			
Compression	**Image MSE**	**SSIM**	**FID**
5	0.056232	0.246840	115.374329
10	0.043571	0.300170	114.668144
20	0.035309	0.354097	113.415405
Composition of LCG noises			
Compression	**Image MSE**	**SSIM**	**FID**
5	0.069271	0.195917	165.312332
10	0.124997	0.106405	163.898788
20	0.124159	0.105760	163.612579

Standard autoencoder			
AE(7)	0.033821	0.364243	311.198334
AE(14)	0.028392	0.435497	285.047241
AE(28)	0.025243	0.494204	249.512772

TABLE II
EVALUATION OF COMPRESSION WITH DIFFERENT APPROACHES

Model	↓Bpp	↓lpips	↑PSNR	↑DBCNN	↓Niqe
Cheng2020	0.0663	0.2292	34.16	41.86	6.55
IG(LCG(256))	0.3513	0.4980	14.99	55.41	4.57
Proposed	0.0620	0.3093	24.48	43.92	5.97

from downscale of size 512 × 512 and once again Cheng2020 quality 2.

As observed in Table II, the linear LCG composition appears inefficient both in terms of quality and compression, providing slightly better results only in visual metrics. This suggests that, with larger noise representations, the trade-off between quality and compression becomes less favourable concerning the proposed configuration. Furthermore, the proposed approach demonstrates slightly better performance in almost all no-reference quality metrics compared to *cheng2020* (implemented through the CompressAI library) despite the lower Bpp, made exception for fidelity metrics (PSNR and LPIPS) as the generative reconstruction limits such performance.

This underscores the fact that employing a compression framework in combination with generative embeddings can provide an extremely efficient way to represent good quality images, albeit at the cost of reconstruction fidelity.

To test the algorithm at different rate points, we generated the rate-distortion/quality curves reported in Figure 2. It is possible to notice that proposed method consistently delivers better results for NIQE and DBCNN across all levels, reaffirming its advantage in visual quality. These conclusions are also motivated by a visual inspection of the reconstructed images: details reported in Fig. 3 show that facial parts reconstructed by standard learned codecs look very smooth and lack sharpness.

Since pixel-level fidelity can not be granted by our generative proposed method, it is worth investigating if biometric consistency is preserved, i.e., faces can be accurately identified or recognized even though PSNR values are lower as can be noticed by details in the picture 3. To this extent, we tested our promising framework by comparing the reconstructed images with other pictures of the same person and computing a biometric matching score, which was obtained by means of a face recognition app Recognito [26].

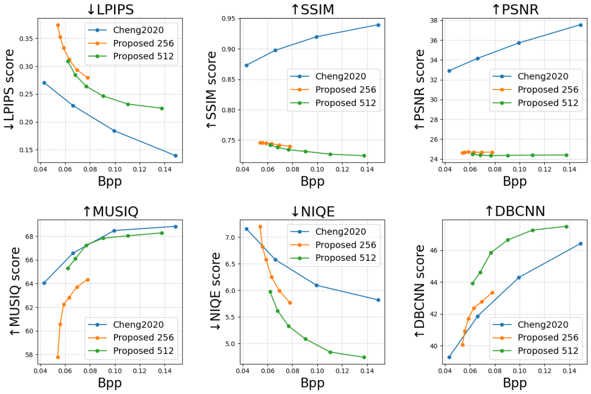

Fig. 2. Fidelity and visual-quality scores across different Bpp for Cheng2020 and the proposed method.

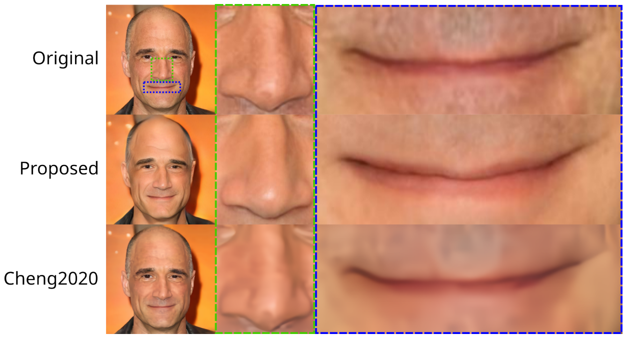

Fig. 3. Reconstruction details using proposed and Cheng2020 at approximately BPP 0.064.

The average score across 15 samples (see the results in Table III) indicates that for the face-matching task, results are very close, showing that a correct authentication is still possible.

A. Ablation

We evaluate the design choices of our proposed method by first demonstrating the effectiveness of each selected component. Then, we analyze the components individually and justify our selections.

As observed in Table IV, the ablation study confirms our initial statement: style features primarily define the spatial and structural details of the image, while latent noise controls style attributes such as texture, colour, and structure. This is evident from the fact that the model, which uses only style features, performs well in fidelity metrics but suffers significant quality degradation. Conversely, using only latent noise produces high-quality images but with lower structural fidelity. This is visually displayed by Fig. 4, where the proposed method image without latent noise looks highly blurred concerning its counterpart where style features were removed. It is also possible to notice that traditional learned compression (Cheng2020) highlights some artifacts (e.g., the top part of the eye), giving a less natural look to the overall picture.

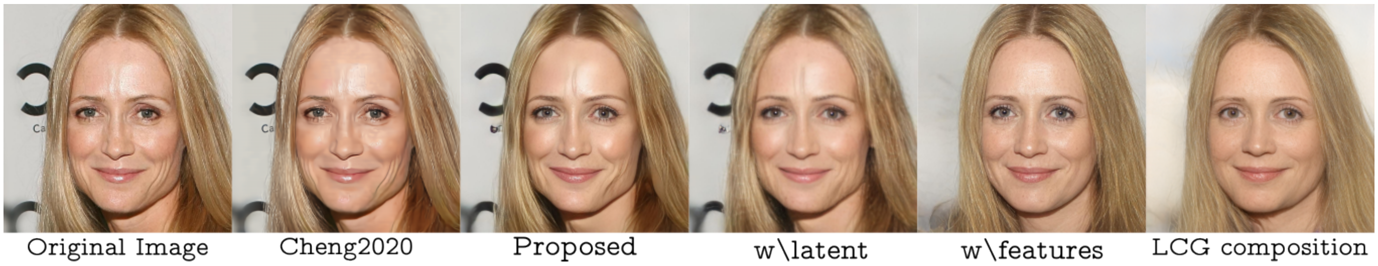

Fig. 4. Comparison between the different approaches analyzed.

TABLE III
RECOGNITO MATCHING SCORE FOR FACE IDENTIFICATION.

Model	**Original**	**Cheng2020**	**Proposed**
Average score	0.9939	0.9941	0.9931

TABLE IV
ABLATION STUDY: THE IMPACT OF DIFFERENT COMPONENTS IN RECONSTRUCTION

Model	**↓Bpp**	**↑PSNR**	**↓LPIPS**	**↑DBCNN**	**↓Niqe**
Full model	0.0620	24.48	0.3093	43.92	5.97
w/latent	0.0131	24.59	0.3433	34.60	6.90
w/feature	0.0489	19.82	0.3188	50.49	4.38

Our proposed approach leverages both elements, ensuring a balance between image quality and structural accuracy, resulting in more perceptually coherent reconstructions.

V. CONCLUSIONS

In this work, we propose novel image embedding approaches for low-bit-rate compression of face images using an Inverse Generative Adversarial Network (IGAN). The proposed scheme uses the generating input noise of GANs as a latent feature representation for the image, thanks to the inversion mechanism and approximates it employing linear composition and quantization. When applied to StyleGAN, the compression framework provides a suitable solution to embed images w.r.t. visual quality, even with complex datasets like Celeb HQ, outperforming learned approaches and making the reconstructed images suitable for tasks such as face recognition. This result suggests that further exploration should investigate generative compression techniques that balance rate-distortion performance while improving generalization across diverse datasets.

REFERENCES

[1] J. Ascenso, E. Alshina, and T. Ebrahimi, "The JPEG AI standard: Providing efficient human and machine visual data consumption," *IEEE MultiMedia*, vol. 30, no. 1, pp. 100–111, 2023.

[2] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in Neural Information Processing Systems*, 2018, vol. 31.

[3] T. Wang, Y. Zhang, Y. Fan, J. Wang, and Q. Chen, "High-fidelity GAN inversion for image attribute editing," in *Proc. of CVPR*, 2022.

[4] D. Mari and S. Milani, "Enhancing the rate-distortion-perception flexibility of learned image codecs with conditional diffusion decoders," *arXiv preprint arXiv:2403.02887*, 2024.

[5] D. Mari, S. Cavasin, S. Milani, and M. Conti, "Effectiveness of learning-based image codecs on fingerprint storage," in *Proc. of IEEE WIFS 2024*. 2024, pp. 1–6, IEEE.

[6] S. Bergmann, D. Moussa, and C. Riess, "Trustworthy compression? impact of ai-based codecs on biometrics for law enforcement," 2024.

[7] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2019.

[8] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang, "GAN inversion: A survey," *IEEE Trans. on PAMI*, vol. 45, no. 3, pp. 3121–3138, 2023.

[9] R. Abdal, Y. Qin, and P. Wonka, "Image2stylegan: How to embed images into the stylegan latent space?," in *Proc. of ICCV*, 2019, pp. 4432–4441.

[10] H. Zhang, F. Mei, J. Liao, L. Li, H. Li, and D. Liu, "Practical learned image compression with online encoder optimization," in *Proc. of PCS*, 2024, pp. 1–5.

[11] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. Van Gool, "Generative adversarial networks for extreme learned image compression," *Proc. of ICCV 2019*, pp. 221–231, 2018.

[12] R. Yang and S. Mandt, "Lossy image compression with conditional diffusion models," in *Proc. of NeurIPS 2023*, 2023.

[13] F. Mentzer, G. D. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[14] Y. Yu, L. Zhang, H. Fan, and T. Luo, "High-fidelity image inpainting with GAN inversion," in *Proc. of ECCV*, Cham, 2022, pp. 242–258, Springer Nature Switzerland.

[15] Y.-C. Cheng, C. H. Lin, H.-Y. Lee, J. Ren, S. Tulyakov, and M.-H. Yang, "InOut: Diverse image outpainting via GAN inversion," in *Proc. of CVPR*, June 2022, pp. 11431–11440.

[16] L. D. Tran, S. M. Nguyen, and M. Arai, "GAN-based noise model for denoising real images," in *Proc. of Computer Vision – ACCV*, 2020, p. 560–572.

[17] J. Ko, K. Cho, D. Choi, K. Ryoo, and S. Kim, "3D GAN inversion with pose optimization," in *Proc. of WACV*, January 2023, pp. 2967–2976.

[18] F. Yin, Y. Zhang, X. Wang, T. Wang, X. Li, Y. Gong, Y. Fan, X. Cun, Y. Shan, C. Oztireli, and Y. Yang, "3D GAN inversion with facial symmetry prior," in *Proc. of CVPR*, June 2023, pp. 342–351.

[19] Z. Huang, T. Wu, Y. Jiang, K. C. K. Chan, and Z. Liu, "ReVersion diffusion-based relation inversion from images," in *Proc. of SIGGRAPH Asia*, 2024.

[20] B. Wallace, A. Gokul, and N. Naik, "EDICT: Exact diffusion inversion via coupled transformations," in *Proc. of CVPR*, June 2023, pp. 22532–22541.

[21] K. Pandey, J. Pathak, Y. Xu, S. Mandt, M. Pritchard, A. Vahdat, and M. Mardani, "Heavy-tailed diffusion models," in *Proc. of ICLR*, 2025.

[22] Z. Zhu, T. Xu, L. Li, and Y. Wang, "Noise dimension of GAN: An image compression perspective," in *Proc. of ICME*, 2024, pp. 1–6.

[23] M. Yang, J. Tang, S. Dang, G. Chen, and J. A. Chambers, "Multi-distribution mixture generative adversarial networks for fitting diverse data sets," *Expert Systems with Applications*, vol. 248, pp. 123450, 2024.

[24] Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu, "Generative latent coding for ultra-low bitrate image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26088–26098.

[25] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proc. of CVPR*, 2020, pp. 7936–7945.

[26] Recognito, "Face recognition sdk," https://recognito.vision/face-recognition-sdk/, 2025, Accessed: January 2025.