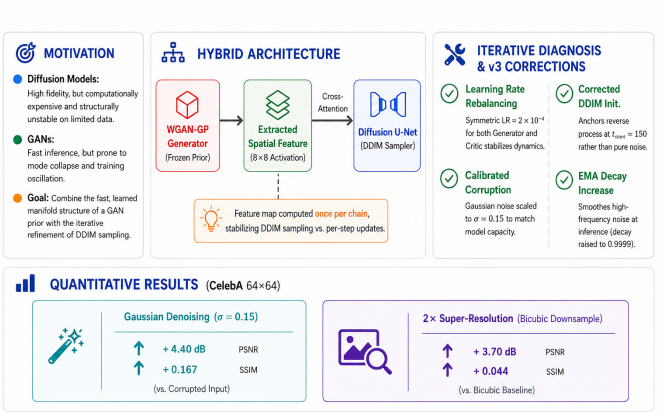
GAN-Diff : Coupling Pretrained WGAN-GP Features with Conditional Diffusion U-Nets

Saif Ahmed, Ashadulla Hil Galib, S.M. Riaz Rahman Antu, Ahmed Faizul Haque Dhrubo*,
Souvik Pramanik, Mohammad Abdul Qayum, Mohsin Sajjad, and Mohammad Ashrafuzzaman Khan
Department of Electrical and Computer Engineering
North South University,Dhaka, Bangladesh
{saif.ahmed03, ashadulla.galib01, riaz.antu, ahmed.dhrubo,
souvik.pramanik, mohammad.qayum, mohsin.sajjad, mohammad.khan02}@northsouth.edu

***Abstract*—Generative adversarial networks (GANs) can provide efficient image generation, while diffusion models offer high-quality image restoration but require iterative sampling. This paper presents a hybrid GAN-guided diffusion framework that uses a pretrained Wasserstein GAN with gradient penalty (WGAN-GP) as a feature prior for conditional diffusion-based image restoration. Intermediate features from the frozen WGAN-GP generator are incorporated into a diffusion U-Net through cross-attention and remain fixed during the DDIM sampling process. The framework is evaluated on two restoration tasks, Gaussian denoising and $2\times$*super-resolution*, using CelebA face images. During development, several sources of instability were identified and addressed, including adversarial learning-rate imbalance, inappropriate diffusion initialization, excessive corruption, and insufficient parameter averaging. The resulting framework consistently improves the quality of both degraded and low-resolution images. In particular, it improves denoising performance by 4.40 dB in PSNR and super-resolution performance by 3.70 dB over their respective input baselines. These results demonstrate the potential of a frozen GAN feature prior to guide diffusion models toward stable and effective image restoration.**

***Index Terms*—generative adversarial networks, diffusion models, image restoration, image denoising, super-resolution, WGAN-GP, cross-attention**



I. Introduction

Deep generative models have emerged as essential tools for image synthesis and recovery, and two types of such models, namely generative adversarial networks (GANs) and diffusion models, complement each other. In particular, GANs [1] generate images using only one forward pass and thus provide efficient inference. Objective functions based on Wasserstein distance with gradient penalty, for example WGAN-GP [2], also provide additional benefits in terms of stable training. On the other hand, GANs may still exhibit mode collapse, oscillations during training, and instability due to the interplay between the learning dynamics of the generator and the discriminator. Diffusion probabilistic models (DDPMs) [3], [4] serve as another method of image generation by training a model to reverse the noising process in a series of denoising steps. While diffusion models show high accuracy and high variety [5], their iterative nature of image generation requires much higher computational efforts than the one-time generation performed by GANs. Diffusion models may experience instability with poor initialization and restricted settings, whereas GANs are highly capable of modeling structural and semantic manifolds. Integrating the two approaches capitalizes on the strengths of both by allowing a pretrained GAN to serve as a supplementary feature-level prior for the diffusion model's optimization iterations without compromising its restoration procedure. With this in mind, this paper suggests a GAN-aided diffusion model for Gaussian denoising and $2\times$ super-resolution using a WGAN-GP generator that has been pretrained on CelebA faces and is subsequently frozen before being plugged into a conditional diffusion U-Net via cross-attention to provide facial structure guidance while maintaining the denoising ability. This final architecture was empirically developed based on multiple iterations—first, deciding which architecture of GAN is to be used out of DCGAN, WGAN-GP, and StyleGAN-lite (which resulted in choosing WGAN-GP for stability and quality reasons) and then resolving issues arising in training and inference due to generator-critic learning rates, DDIM initialization, corruption strength, and exponential moving average smoothing.

The main goal of this work is to examine whether a pretrained WGAN-GP provides an adequate feature-level prior for conditional diffusion-based image restoration through the design of a hybrid approach in which intermediate spatial features of the frozen GAN generator steer the diffusion U-Net via cross-attention. The second goal is to apply this architec-

ture to the tasks of Gaussian image denoising and $2\times$ super-resolution as two distinct low-level vision problems. Moreover, it is necessary to uncover and mitigate pipeline instabilities associated with the proper tuning of the adversarial learning rate ratio, diffusion initialization, corruption normalization, and exponential moving averages.

This paper proposes a novel GAN-prior guided conditional diffusion network, in which a frozen WGAN-GP generator generates the intermediate spatial features before the diffusion U-Net through cross-attention, calculated once for each sampling chain and fixed during the DDIM sampling process. For the stability of the model, this paper gives a diagnosis-driven solution to the following problems: the imbalance between the learning rate of generator and critic, DDIM initialization problem, excessive corruption intensity, and lack of EMA smoothing. We evaluate the effectiveness of our learned GAN-prior for image restoration on Gaussian denoising and $2\times$ super-resolution tasks on the CelebA dataset.

II. Related Work

Modern techniques of image restoration have become more inclined towards the use of GANs together with diffusion models. The use of DCGAN [6] and StyleGAN [7] have led to the development of strong convolutional and style-based generator functions that provide realistic samples. At the same time, WGAN-GP [2], which is based on Wasserstein critic and gradient-norm regularizer, almost eliminates mode collapsing. On the contrary, diffusion models, which were first proposed by Sohl-Dickstein et al. [4] and later described theoretically by Ho et al. [3] as denoising diffusion probabilistic models (DDPM), offer stable training and good iterative performance. Song et al. [8] have developed DDIM or denoising diffusion implicit models for efficient inference, using the cosine noise schedule from Nichol and Dhariwal [9].

In case of image-to-image translation and low-level restoration tasks, SR3 [10] and Palette [11] use conditional diffusion model by channel-wise concatenation of the degraded conditioning input. In addition, SDEdit [12] shows that noisying only part of the corrupted input at an intermediate time step is more effective compared to starting with pure Gaussian noise, which we adopted for our task as well.

Past literature has also looked into hybrids of GANs and diffusion methods like using discriminators as learned denoisers [13] or taking advantage of pre-trained latent space as semantic priors. Our work is different from these works by design as, instead of substituting the diffusion reverse time step with the generated output from the GAN, we make use of an intermediary feature map extracted from a pre-trained WGAN-GP generator network as a key value context which the diffusion U-Net can attend to through cross attention.

III. Dataset

A. Dataset Overview

The framework is trained and evaluated using a 50,000-image subset of the CelebA celebrity face dataset. All images are center-cropped, resized to 64×64 resolution using bicubic interpolation, and normalized to the $[-1,1]$ range per channel. To ensure consistent visual and quantitative tracking across checkpoints, five fixed indices $(0, 2501, 8000, 15000, 31337)$ are designated as tracked evaluation samples. Using these fixed indices allows for direct before-and-after comparisons on the same faces throughout training rather than relying on randomly resampled evaluation batches.

Fig. 1: Vertical overview of the CelebA dataset preparation pipeline. The 50,000-image subset is center-cropped, resized to 64×64, normalized to $[-1,1]$, and evaluated using five fixed benchmark samples for consistent checkpoint-wise comparison.

B. Dataset Statistics

The statistics of the dataset is shown in Figure 2. The principal characteristics of the dataset and its preprocessing configuration are summarized in Table I.

TABLE I: Dataset statistics and preprocessing configuration used in the experimental pipeline.

Statistic	Configuration
Dataset	CelebA celebrity face dataset
Training subset	50,000 images
Image type	RGB face images
Image resolution	64×64 pixels
Cropping	Center crop
Resampling	Bicubic interpolation
Normalization	$[-1,1]$ per channel
Evaluation subset	5 fixed benchmark images
Tracked indices	0, 2501, 8000, 15000, 31337
Evaluation purpose	Consistent quantitative checkpoint tracking

IV. Methodology

A. Proposed Methodology

The proposed framework relies on a sequential multi-stage pipeline designed to extract robust structural priors from a pre-trained generative model and leverage them within conditional diffusion restoration paths, as outlined in Figure 3.

DATASET STATISTICS SUMMARY

Dataset
CelebA-50k

Total image files	: 50,000
Valid images	: 50,000
Corrupted images	: 0
Dominant resolution	: 178x218
Dominant format	: JPEG
Mean width	: 178.0 px
Mean height	: 218.0 px
Median width	: 178.0 px
Median height	: 218.0 px
Mean aspect ratio	: 0.8165
Median aspect ratio	: 0.8165
Mean image size	: 6.79 KB
Median image size	: 6.69 KB
Total image storage	: 331.61 MB
RGB images	: 50,000
Grayscale images	: 0
RGBA images	: 0

Fig. 2: Data Statistics.

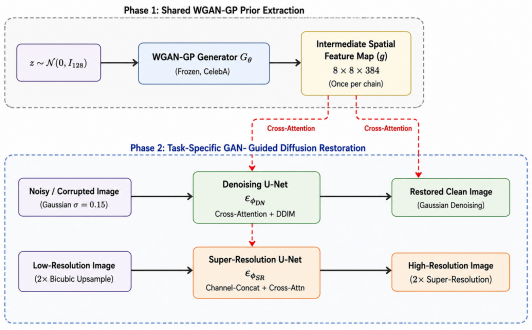


Fig. 3: Proposed Methodology

1) Phase 1 : WGAN-GP Prior Generator: The generator $G_\theta : \mathbb{R}^{128} \to \mathbb{R}^{3\times64\times64}$ maps a latent vector $z \sim \mathcal{N}(0, I_{128})$ through five transposed-convolution/BatchNorm/ReLU blocks with a base channel width of 96, doubling spatial resolution at each stage from 4×4 to 64×64, terminated by a $\tanh$ output. The critic C_θ mirrors this with strided convolutions, spectral normalization [14] on every convolutional layer, instance normalization on intermediate layers, and LeakyReLU activations. The critic is optimized via the WGAN-GP objective,

$$\mathcal{L}_C = \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}[C(\tilde{x})] - \mathbb{E}_{x\sim\mathbb{P}_r}[C(x)] + \lambda_{\text{gp}}\,\mathbb{E}_{\hat{x}}\left[(\|\nabla_{\hat{x}}C(\hat{x})\|_2 - 1)^2\right], \tag{1}$$

with $\lambda_{\text{gp}} = 10$, $\hat{x} = \alpha x + (1-\alpha)\tilde{x}$, and $\alpha \sim \mathcal{U}(0,1)$. The generator is trained using $\mathcal{L}_G = -\mathbb{E}_z[C(G(z))]$ every $n_{\text{critic}} = 5$ critic updates. An exponential moving average (EMA) copy of G_θ is maintained throughout training to reduce parameter variance at inference time [15].

2) Phase 2 : GAN-Guided Conditional U-Net and Diffusion Optimization: Both diffusion U-Nets share a U-Net backbone incorporating FiLM-style timestep conditioning [16], bottleneck self-attention, and cross-attention blocks where U-Net spatial features attend to a fixed WGAN-GP feature map (g) computed once per sampling chain. Optimization follows a cosine noise schedule ($T = 1000$).

Algorithm 1 GAN-Guided Diffusion Training and Sampling

Input: Frozen generator G_θ, clean data x_0, total steps $T = 1000$, schedule parameter $s = 0.008$.
Phase 1: Feature Prior Extraction
Sample latent vector $z \sim \mathcal{N}(0, I_{128})$
Extract fixed feature map $g \leftarrow G_\theta^{(8\times8)}(z) \in \mathbb{R}^{B\times384\times8\times8}$
Phase 2: Training Objective
Sample timestep $t \sim \mathcal{U}\{0,\dots,T-1\}$ and noise $\varepsilon \sim \mathcal{N}(0, I)$
Compute forward diffusion: $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\,\varepsilon$
Compute cross-attention: $Q(x_t)$, $K(g)$, $V(g)$
TextCrossAttn$(x_t, g) = x_t + W_o\left[\text{softmax}\left(\frac{Q(x_t)K(g)^\top}{\sqrt{d}}\right)V(g)\right]$
Minimize objective: $\mathcal{L}_{\text{diff}} = \mathbb{E}\left[\|\varepsilon - \epsilon_\phi(x_t, t, g)\|_2^2\right]$
Phase 3: DDIM Inference (Single-Pass Prior Reuse)
Initialize $x_{t_{\text{start}}}$ (via corrupted input injection for denoising or noise for SR)
for $t = t_{\text{start}}, \dots, 1$ **do**
  Predict clean image: $\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\phi(x_t,t,g)}{\sqrt{\bar{\alpha}_t}}$ (clipped to $[-1,1]$)
  Update deterministically to $x_{t'}$ using fixed g
end for

B. Experimental Setup

The implementation is evaluated across stable configurations and optimized hyperparameters to resolve prior pipeline instabilities:

TABLE II: Training hyperparameters for the restoration pipeline.

Parameter	Value
Image resolution	64×64
Training subset size	50,000 images
Batch size (WGAN-GP / Diffusion)	64 / 32
WGAN-GP Prior	
Latent dimension z	128
Generator / critic width	96
Optimizer (G and D)	Adam (β_1=0.0, β_2=0.9)
Learning rate (LR$_G$, LR$_D$)	2×10^{-4}, 2×10^{-4}
Gradient penalty λ_{gp}	10
Diffusion Backbone	
Timesteps T (Cosine schedule)	1000
DDIM sampling steps	50
Optimizer	AdamW (wd $= 10^{-4}$)
EMA decay	0.9999
Task Specifications	
Denoising epochs / σ / t_{start}	30 / 0.15 / 150
Super-resolution epochs / factor	25 / 2×

C. System Architecture

The framework features a hybrid design where a pre-trained generator acts as a reusable prior. The system architecture is present in Figure 4.

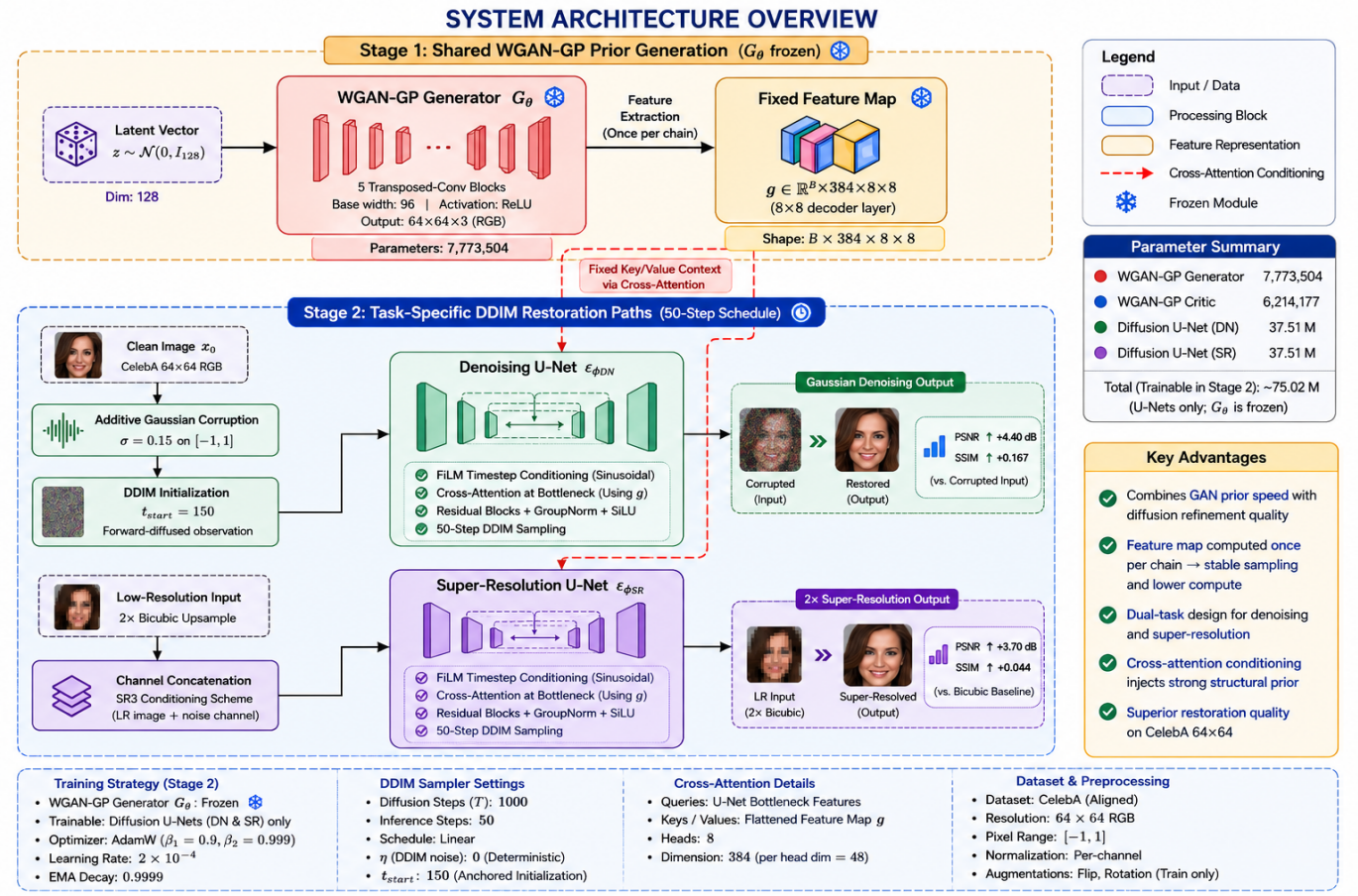


Fig. 4: System Architecture

D. Evaluation Metrics

Restoration performance is measured using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [17]. PSNR evaluates the reconstruction quality by measuring the ratio between the maximum possible power of an image and the corrupting noise power, defined as:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right), \tag{2}$$

where MAX_I is the maximum possible pixel value of the image, and MSE represents the mean squared error between the ground-truth clean image and the evaluated output.

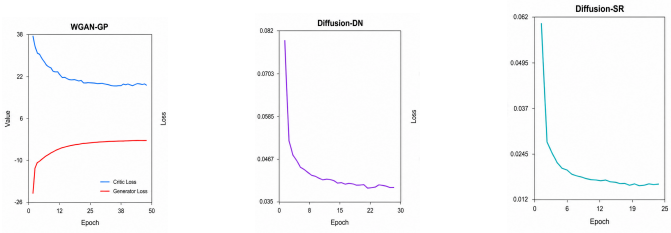
To assess perceived structural degradation, SSIM models image quality changes by combining luminance, contrast, and structural terms:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{3}$$

where μ_x and μ_y are the local means, σ_x^2 and σ_y^2 are the variances, and σ_{xy} is the covariance of images x and y. Constants C_1 and C_2 stabilize the division with weak denominators. These metrics are computed on the five tracked faces between the ground-truth clean image and both the degraded inputs and the final diffusion-restored outputs to quantify the improvement achieved by the GAN-guided refinement process.

V. Results

A. Training Dynamics



(a) WGAN-GP Losses (b) Denoising Loss (c) Super Resolution Loss

Fig. 5: Training loss curves for the pipeline.

The loss curves for WGAN-GP generator and critic as well as diffusion U-Net MSE losses are shown in Figure 5. The critic loss in the case of WGAN-GP steadily increases from -23.99 at epoch 1 to -4.71 at epoch 50, and the generator loss steadily decreases from 36.65 to 18.44 in the same period of time. It is clear that adversarial convergence occurs smoothly

TABLE III: Restoration results on 5 tracked CelebA faces (mean values; see Table IV for per-sample values).

Task	Stage	PSNR (dB)	SSIM
Denoising (σ=0.15)	Corrupted input	22.76	0.6755
	Restored	**27.17**	**0.8429**
Δ (input $\rightarrow$ restored)		+4.40 dB / +0.167	
Super-res. (2×)	Bicubic baseline	27.24	0.9122
	Restored	**30.94**	**0.9566**
Δ (baseline $\rightarrow$ restored)		+3.70 dB / +0.044	

TABLE IV: Per-sample PSNR (dB) on the 5 fixed tracked faces.

Sample	Noisy	Denoised	Bicubic	SR out
S1	23.0	26.2	24.8	27.5
S2	22.8	26.8	28.8	31.2
S3	22.6	27.6	27.9	33.4
S4	22.6	27.2	29.0	30.6
S5	22.7	28.0	25.7	32.0
Mean	22.76	27.17	27.24	30.94

without oscillations in the current implementation, where the problem of learning rate imbalance has been fixed. Moreover, both diffusion U-Nets show steady monotonically-decreasing losses for predicting noise from the image corrupted by it: denoising U-Net converges from 0.079 to 0.037 in 30 epochs, while super-resolution U-Net converges from 0.060 to 0.014 in 25 epochs. It makes sense since super-resolution is a better-posed inverse problem (since bicubic upsampling already contains most of the low-frequency information) than blind Gaussian denoising.

Validation performed on each epoch (finite check, standard deviation of pixel greater than 10^{-3}, dynamic range greater than 10^{-2}) shows that the generator does not fall into mode collapse for all ten WGAN-GP samples recorded at epochs 10, 20, ..., 50. In particular, the sample standard deviation increases monotonically from 0.430 at epoch 10 to 0.476 at epoch 50.

B. Restoration Quality

Table III outlines the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) averaged across the five tracked faces before and after diffusion-based restoration for both tasks .

The figures 6 and 7 depict the absolute value of the metrics and their corresponding gain per sample . All the faces for which the metrics have been computed in both the tasks are exhibiting a PSNR gain, and no degenerate or negative improvement case is observed . This proves the robustness of the corrected pipeline, and how it generalizes restoration over different poses and occlusions (the occluded face wearing sunglasses from the tracked sample set), and does not overfit itself to easy cases . Also, the super-resolution task exhibits higher PSNR and SSIM absolute values both before and after

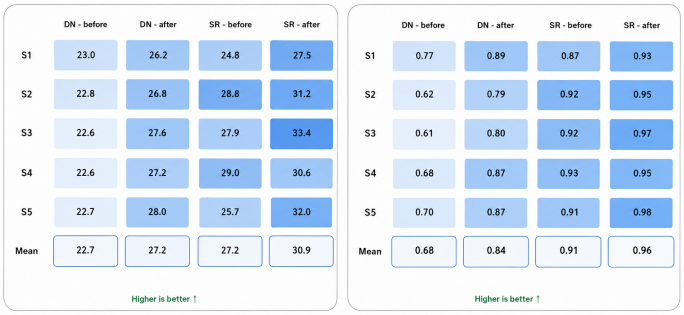


Fig. 6: Per-sample and mean PSNR (left) and SSIM (right) before and after restoration. Lighter cells indicate higher metric values; both tasks improve consistently across the five fixed tracked samples.

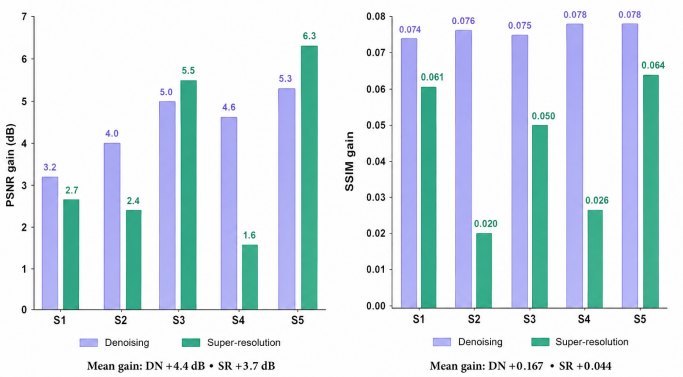


Fig. 7: Per-sample restoration gains in PSNR and SSIM for denoising and super-resolution. Every tracked sample improves under both restoration tasks.

restoration than denoising, since 2× bicubic downsampling is a mild degradation process compared to Gaussian noise of $\sigma = 0.15$.

C. Qualitative Observations

The figures 8, 9 show visual examples of the five tracked faces (organized into three rows: clean reference, degraded, and restored images) . It is evident from visual examination that the GAN-based denoiser has successfully eliminated the visible speckle noise and retained the important identity features, which include the eyes, hair contours, and mouth region . Likewise, the super-resolution approach has been able to synthesize realistic high-frequency details such as the fine hair strands and skin details, which are not present in the bicubic upsampled inputs, without generating any checkerboard artifacts.

VI. Discussion

The most important architectural insight to be taken away from the final update is that the GAN conditioning signal g should not be sampled independently at each of the 50 DDIM steps, but rather considered as a *constant* input throughout the entire DDIM sampling process. This follows from the fact

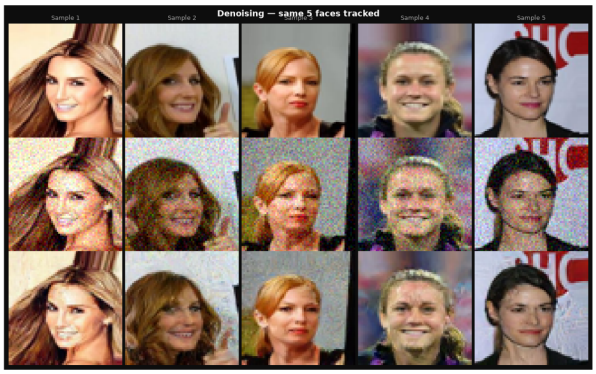


Fig. 8: Denoising results on the 5 fixed tracked faces. Top: clean reference. Middle: corrupted input ($\sigma = 0.15$ Gaussian noise). Bottom: GAN-guided diffusion restoration via the corrected DDIM initialization.

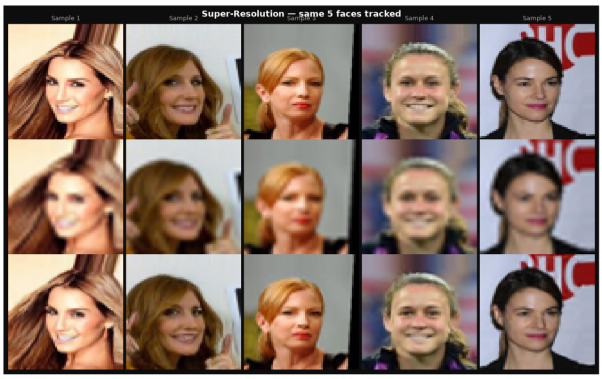


Fig. 9: Super-resolution results on the 5 fixed tracked faces. Top: clean reference. Middle: $2\times$ bicubic downsample/upsample baseline. Bottom: GAN-guided diffusion $2\times$ super-resolution output.

that, being deterministic, the DDIM sampling procedure is a function mapping x_T ($x_{t_{\text{start}}}$) to x_0, and allowing stochastic changes to g at each step essentially makes the trajectory inconsistent, providing the network with contradictory instructions regarding semantics at subsequent steps; the new version of the architecture solves this problem by calculating the g only once, right before running the reverse loop.

The asymmetry between the 4x critic and generator LR imbalance observed in v2 is a likely cause for instability in light of the WGAN-GP loss surface, since having an overly fast critic compared to the generator will push the critic into a state where its gradients on generated samples become meaningless or saturated because the critic "overfits" to discriminate between generated samples produced by the current generator faster than the generator learns. Fixing this issue by going back to the standard symmetric $2 \times 10^{-4}/2 \times 10^{-4}$ ratio advocated by the WGAN-GP paper [2] solves the problem without any modifications to the architecture.

VII. Limitations

There are several limitations associated with this paper. First, the evaluation is limited to five tracked faces; even though it allows to evaluate the effect of the method precisely before and after its application, there is no way to measure the quality of the restoration on a larger population, and the full evaluation could be done using a test set of hundreds or thousands of faces. Second, no perceptual metric (LPIPS [18], included in v1 evaluation harness but absent in v3) or FID [19] scores are provided for the final pipeline, even though these metrics are known to correlate better with human perception of face realism than PSNR/SSIM scores. Third, the corruption model for the task of denoising is limited to one fixed level of Gaussian noise ($\sigma = 0.15$); robustness to different levels of corruption is not tested. Fourth, there is no ablation study that shows the individual effect of the GAN cross-attention conditioning.

VIII. Conclusion

We proposed a conditional diffusion pipeline guided by a pre-trained WGAN-GP to denoise and up-sample face images, where a feature map extracted from a frozen GAN generator acts as a condition for the diffusion U-Net model through the cross-attention mechanism, calculated at one point per sampling and then kept constant during all DDIM reverse steps. Using an iterative approach to development, driven by a thorough analysis of each identified problem, we discovered four particular instabilities, such as GAN/GAN-critic learning-rate mismatch, incorrect DDIM initialization for denoising task, uncalibrated corruption severity and non-optimal EMA scheduling, which caused the instability of v2 pipeline and allowed us to develop a stable v3 pipeline. In particular, on five CelebA face images selected for tracking, the proposed pipeline shows a PSNR gain of 4.40 dB (SSIM +0.167) for denoising task and 3.70 dB (SSIM +0.044) for super-resolution in comparison with naive baselines, where every single tracked image improved in both cases. Further research may involve testing our method on a larger held-out test set with additional perceptual (LPIPS/FID) metrics, comparing with an unconditional diffusion pipeline (ablation study) and corruption severity robustness tests.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[2] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[3] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Machine Learning (ICML)*, 2015.

[5] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[7] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[9] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. Int. Conf. Machine Learning (ICML)*, 2021.

[10] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2022.

[11] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *Proc. ACM SIGGRAPH*, 2022.

[12] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "SDEdit: Guided image synthesis and editing with stochastic differential equations," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2022.

[13] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion GANs," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2022.

[14] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.

[15] Y. Yazıcı, C.-S. Foo, S. Winkler, K.-H. Yap, G. Piliouras, and V. Chandrasekhar, "The unusual effectiveness of averaging in GAN training," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2019.

[16] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artificial Intelligence*, 2018.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[18] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018.

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.