Proceedings of the 8th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2026.36074

Research and Analysis of Cross-Modal Attention Based on Pareto Optimal Fusion of Sketch Image Generation

Yuhan Wang

School of Software Technology, Dalian University of Technology, Dalian, China
uqdxyokru@outlook.com

Abstract. Conditional Sketch Generation aims to convert abstract sketches into realistic images. This task is challenging due to the modality gap between sketches and photos, weak pixel-level correspondence, and the one-to-many mapping problem. This paper introduce two paradigm shifts within a single framework, replacing ControlNet's zero-convolution with deep cross-modal attention and noise prediction with L1 reconstruction loss. On this basis, a Dual-Path Non-Shared Encoder is presented. During the experiments on SketchCOCO, this paper uncovers three empirical phenomena with universal significance. Firstly, all evaluation metrics exhibit a synchronous improvement phase when $w < 0.7$. Secondly, the guidance weight exhibits a stable Pareto interval, with model performance sustaining high stability for $w \in [0.6,0.9]$. Finally, through dense sampling with a step size of 0.1, $w = 0.7$ achieves the best trade-off across all metrics (FID = 177.90, Corr = 0.7160, SSIM = 0.6419). Meanwhile $w = 0.9$ yields the lowest LPIPS (0.1949) and $w = 0.8$ achieves the highest SSIM (0.6471). This work not only provides a lightweight yet powerful model for sketch generation, but also offers actionable insights for designing fusion mechanisms in cross-modal tasks.

Keywords: cross-modal fusion, diffusion model, pareto optimality

1. Introduction

Conditional sketch generation enables rapid idea visualization in artistic design, entertainment, and conceptual prototyping. Different from natural image translation tasks that possess dense pixel correspondences, the sketch generation task faces three challenges as weak correspondence between sparse strokes and realistic images, a one-to-many mapping problem, and a substantial modality gap.

Early approaches relied on GANs [1]. Pix2pix [2] first applied conditional GANs to image translation but assumes strong structural correspondence, which is invalid for sketches. CycleGAN [3] removes the need for paired data via cycle consistency, yet is parameter-heavy and offers limited controllability. Diffusion models [4] introduced a paradigm shift via denoising and LDM [5] enabled latent-space diffusion. ControlNet [6] injects conditional signals via zero-convolution, but struggles with sketches where undrawn areas are treated as empty space rather than unspecified regions [7]. Recent works distill diffusion models into single-step generators for faster inference [8] or explore dual-path controllability [9].

DOI: 10.54254/2755-2721/2026.36074

Despite these advances, existing models are computationally heavy and lack interpretability. This work proposes a lightweight, interpretable framework that prioritizes quantitative insights over SOTA performance, with two paradigm shifts. Dense sampling of the fusion weight w on SketchCOCO reveals three phenomena. When w is below 0.7, all metrics improve together, showing that the sketch and noise paths reinforce each other before reaching the optimal balance point; a stable Pareto region exists for w between 0.6 and 0.9， and $w = 0.7$ achieves the best overall trade-off (FID=177.90, Corr=0.7160, SSIM=0.6419), $w = 0.8$ maximizes SSIM (0.6471), and $w = 0.9$ yields the best perceptual quality (PSNR=18.15 dB).

2. Method

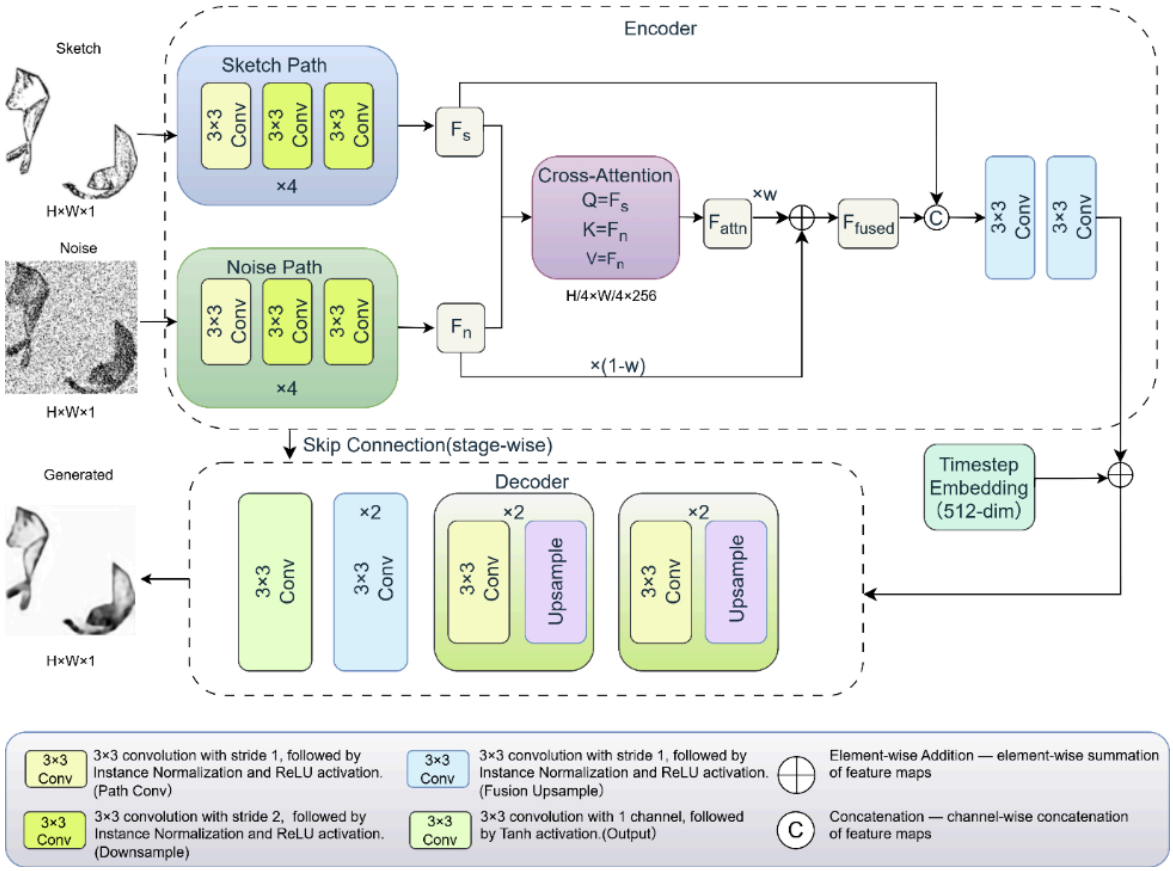

Figure 1. This caption has one line so it is centered

2.1. Motivation and dual-path architecture

This paper reveals two implicit assumptions of ControlNet, including distribution similarity and pixel alignment. The former assumes that sketch features follow the main network feature distribution. However, unlike noised images with random pixel-wise perturbations, sketches contain deliberate human-constructed contours. Regarding the pixel alignment assumption between input and output, sketched lines deviate from true object contours. Through a systematic experiment, ControlNet achieves a correlation score of -0.003, PSNR of 8.36 and FID of 293.58, ranking last among all models the paper tested.

To address this, this work propose a dual-path non-shared encoder. Since sketches and noisy images belong to different modalities with inherent spatial misalignment, cross-modal interference

DOI: 10.54254/2755-2721/2026.36074

needs to be avoided before attention operation. As illustrated in Figure 1, unlike baselines, this model performs cross-modality attention only at the deepest layer, isolating the sketch and noise paths. No weight sharing is employed in the network. The weighted output is applied to fuse with the raw noise feature.

2.2. Paradigm shift：reconstruction over noise prediction and attention over zero-convolution

Sketch-related tasks need stronger structural alignment. However, mainstream models indirectly predict noise ε and denoise to obtain x_0 , causing error accumulation. Therefore, the model directly predicts x_0, which better accommodates conditional image generation. L1 loss better preserves edges. The training method still follows the diffusion forward process. In contrast, the reverse process is replaced with direct regression to the target.

ControlNet injects conditional features via zero-convolution, which is linear static addition. For sketch-based generation, linear addition cannot achieve content-adaptive alignment under spatial misalignment. Therefore, the sketch serves as Q, and the noisy image as K and V. By ta)ing the dot product similarity between Q and K, followed by Softmax normalization, this paper computes the attention matrix A. V is then weighted by A to produce the fused features [8].

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{256}}\right) \tag{1}$$

$$f_{\text{attn}} = AV + Q \tag{2}$$

Preliminary experiments show that f_{attn} aligns sketch structure well but relies heavily on abs2)act lines, degrading realism. The noise feature f_n retains more texture but lacks structural guidance. Therefore, this paper introduces a tunable guidance weight $w \in [0,1]$ to control the mixing ratio between f_{attn} and f_n.

$$f_{\text{fused}} = w \cdot f_{\text{attn}} + (1 - w) \cdot f_n \quad w \in [0,1] \tag{3}$$

3. Result

3.1. Experimental setup

The experiments are conducted on the object-level subset of the SketchyCOCO dataset, providing sketches and their corresponding real photos. Due to inconsistent native resolutions, all images are converted into a uniform resolution preserving essential sketch structures. Since sketches are inherently sparse and line-based, extremely high resolutions offer diminishing returns. A 128×128 resolution is sufficient for conditioning without unnecessary computational overhead. Each configuration is trained three times with fixed random seeds and early stopping, and results are reported as mean±standard deviation.

The experiment includes nine metrics including three dimensions. Condition fidelity is measured by Corr. Pixel-level quality is assessed by PSNR, MSE, and MAE. SSIM evaluates structural

DOI: 10.54254/2755-2721/2026.36074

similarity. For perceptual quality, LPIPS, FID, and KID are used. Semantic alignment is measured by the CLIP score.

3.2. Comparative experiments and ablation experiments

Three baselines are included in the experiments for comparison, namely ControlNet, Pix2pix and CycleGAN [7]. In addition, Table 1 includes an ablation study on learnable adaptive weight gating, though it is found to negatively impact performance. Based on the results presented below, w =0.7 performs as the Pareto optimal trade-off point, balancing all metrics. w =0.8 and w =0.9 are reconstruction optimal points, yielding the best pixel-level quality.

Table 1. Comparative experiments and ablation experiments

Model	Corr↑	PSNR↑	SSIM↑	MSE↓	MAE↓	LPIPS ↓	CLIP ↑	FID↓	KID ↓ ×1000
ControlNet	-0.003	8.36	0.2010	0.1840	0.3580	0.7600	0.512	293.58	2430.9
pix2pix	0.1184	6.59	0.1416	0.9246	0.7161	0.6138	0.7905	358.22	24.5
CycleGAN	0.7339	20.02	0.6510	0.0468	0.1240	0.1010	0.9278	84.46	3.2
Ablation_Single	0.1240	6.69	0.1890	0.2310	0.4010	0.6640	0.548	185.19	639.22
Ablation_NoAttn	0.4460	7.81	0.3240	0.1720	0.3390	0.4770	0.589	151.58	373.15
Adaptive_gating	0.3530	7.03	0.2980	0.1950	0.3710	0.4900	0.581	154.15	441.94
w=0.7	0.7160	17.87	0.6419	0.0837	0.1518	0.2116	0.8829	177.9	8.82
w=0.8	0.7076	18.11	0.6471	0.0794	0.1470	0.2018	0.8835	179.09	8.71

Table 1 compares the proposed model against baselines. ControlNet and Pix2pix perform poorly, while CycleGAN achieves SOTA results. The proposed model achieves a Corr of 0.716, only 0.0179 below CycleGAN, with an order-of-magnitude smaller size and no adversarial training. Its FID is worse, but the framework is designed for lightweight use. Parameter counts: ~6M (ours), 56.6M (Pix2pix), 14.14M (CycleGAN), >1.4B (ControlNet).

Ablation experiments in Table 1 verify the effectiveness of adaptive gating, attention, and the dual-path encoder. A learnable adaptive weight was inserted between the attention output and the noise features. Removing the adaptive gating leads to significant performance improvement. This phenomenon stems from unnecessary competition within the control loop caused by the additional learnable weight confusing the optimization objective. What is more critical, the gating parameter is observed to converge rapidly to sigmoid(3.0) ≈ 0.95 during training and subsequently stop updating. Gradients vanish in the sigmoid saturation region.

3.3. Pareto frontier

Since the gating weight proved unlearnable, the investigation shifted to a manually dense sampling strategy, sweeping the guidance weight w from 0.1 to 1.0 with a step size of 0.1. Table 2 (a) and (b) reports the mean ± standard deviation over three fixed-seed runs. The experiment reveals three clear Pareto fronts, each corresponding to a different dimension.

Table 2. Pareto frontier(a)

w	Corr ↑	PSNR ↑	SSIM ↑	MSE ↓	MAE ↓
0.1	0.6950±0.0063	16.06±0.15	0.5832±0.0039	0.1207±0.0049	0.1906±0.0017

DOI: 10.54254/2755-2721/2026.36074

Table 2. (continued)

0.2	0.6973±0.0088	16.11±0.28	0.5854±0.0104	0.1204±0.0063	0.1892±0.0052
0.3	0.7016±0.0048	16.50±0.08	0.6007±0.0049	0.1107±0.0034	0.1795±0.0023
0.4	0.7032±0.0090	16.83±0.40	0.6102±0.0152	0.1043±0.0080	0.1732±0.0090
0.5	0.7071±0.0025	17.32±0.32	0.6285±0.0110	0.0943±0.0041	0.1624±0.0066
0.6	0.7121±0.0056	17.72±0.19	0.6385±0.0065	0.0873±0.0016	0.1548±0.0036
0.7	0.7160±0.0093	17.87±0.11	0.6419±0.0042	0.0837±0.0022	0.1518±0.0028
0.8	0.7076±0.0037	18.11±0.27	0.6471±0.0058	0.0794±0.0030	0.1470±0.0048
0.9	0.7076±0.0055	18.15±0.22	0.6438±0.0079	0.0784±0.0034	0.1461±0.0047
1.0	0.7078±0.0015	17.84±0.13	0.6243±0.0029	0.0811±0.0027	0.1524±0.0028

Table 2. Pareto frontier(b)

w	LPIPS ↓	CLIP ↑	FID ↓	KID×1000 ↓
0.1	0.2686±0.0030	0.8753±0.0022	190.68±2.87	10.98±0.65
0.2	0.2661±0.0077	0.8761±0.0025	189.95±3.04	10.86±0.52
0.3	0.2552±0.0056	0.8778±0.0010	186.52±2.73	10.15±0.34
0.4	0.2453±0.0090	0.8793±0.0022	185.52±1.88	10.05±0.54
0.5	0.2314±0.0073	0.8810±0.0030	181.49±0.82	9.45±0.37
0.6	0.2199±0.0037	0.8825±0.0018	179.76±1.54	9.15±0.15
0.7	0.2116±0.0025	0.8829±0.0019	177.90±1.95	8.82±0.24
0.8	0.2018±0.0017	0.8835±0.0017	179.09±1.65	8.71±0.22
0.9	0.1949±0.0009	0.8839±0.0019	179.71±1.97	8.74±0.27
1.0	0.2012±0.0007	0.8831±0.0012	187.17±1.20	9.22±0.16

As shown in Figure 2(a) and Figure 3(a), the FID-Corr Pareto front has three stages. Before $w = 0.7$, FID decreases while Corr increases, reaching the Pareto optimum at $w = 0.7$ with minimum FID and maximum Corr. Beyond this point, FID rises while Corr stabilizes, indicating deviation from the true distribution. In Figure 2(b)/3(b), $w = 0.7$ and $w = 0.8$ are non-dominated; $w = 0.7$ favors FID while $w = 0.8$ maximizes SSIM. $w = 0.9$ is dominated due to worse FID. In Figure 2(c)/3(c), $w = 0.9$ is the unique Pareto optimum, achieving the highest PSNR and lowest LPIPS simultaneously. Both metrics improve from $w = 0.1$ to 0.9, but drop sharply at $w = 1.0$.

In summary, $w = 0.7$ is the only point shared by both Pareto fronts, offering the best FID-Corr balance while remaining competitive elsewhere. $w = 0.8$ maximizes SSIM with a slight FID trade-off. At $w = 1.0$ (pure attention), all metrics drop sharply.

DOI: 10.54254/2755-2721/2026.36074

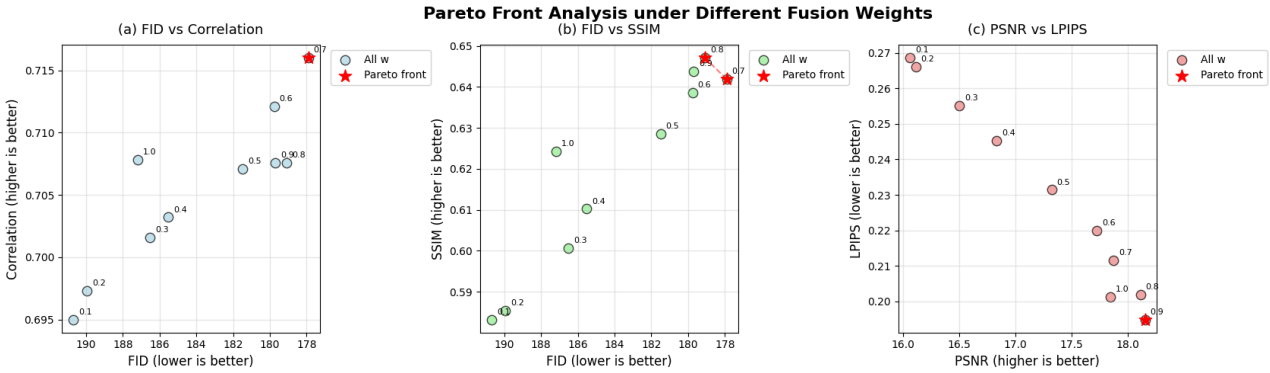

Figure 2. Pareto fronts under different fusion weights. (a) FID vs. Corr, (b) FID vs. SSIM, and (c) PSNR vs. LPIPS (picture credit: original)

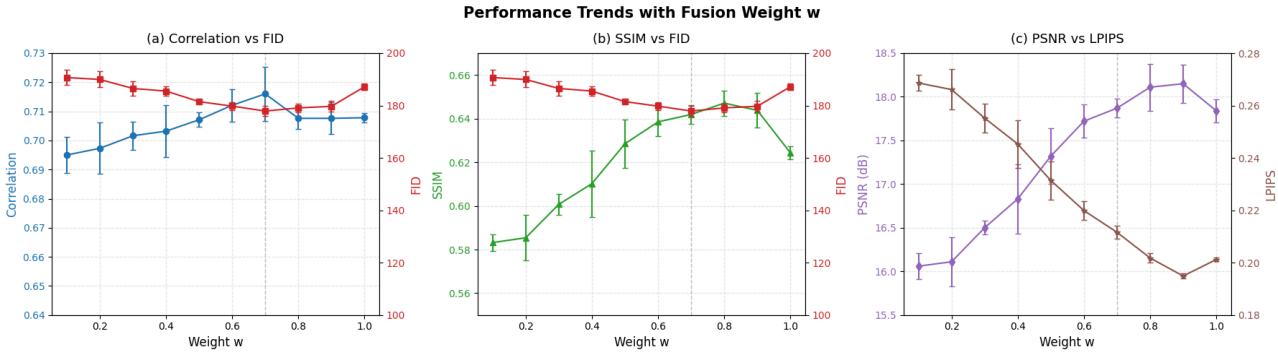

Figure 3. Pareto front in three dimensions under different fusion weights (line plot). Performance trends under different fusion weights. (a) FID vs. Corr, (b) FID vs. SSIM, and (c) PSNR vs. LPIPS (picture credit: original)

3.4. Visualization results

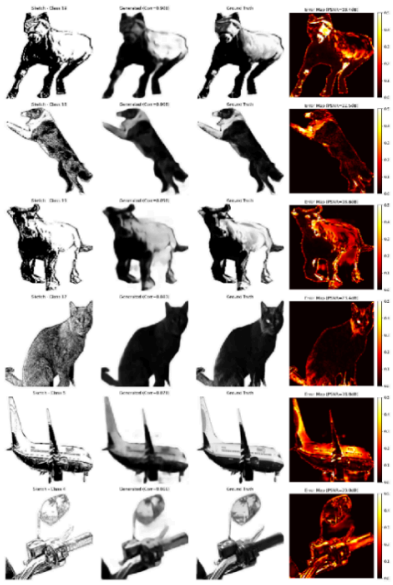

Figure 4. Visual comparison (picture credit: original)

DOI: 10.54254/2755-2721/2026.36074

Figure 4 shows, from left to right, the input sketch, generated image, ground truth, and error map, where brighter regions indicate larger pixel-level errors. The results demonstrate that the proposed model effectively converts abstract sketches into realistic images.

4. Discusssion

4.1. From synergy to optimum and collapse

For $w < 0.7$, FID decreases while other metrics improve, aligning with the observation by Grnarova et al. [10] that multiple evaluation metrics can exhibit correlated improvements within a certain training regime. The diversity from the noise path and structural information from the sketch path complement each other during early fusion. The range $w < 0.7$ corresponds to a free lunch regime where dual-path fusion outperforms single-path designs. Noise features provide stochastic, high-variance components for diversity, while attention features are more deterministic and lock onto the sketch. FID hits its lowest value (177.90) at the point $w = 0.7$. Kynkäänniemi et al. noted that generative models must strike a balance between fidelity and diversity [11]. The ratio $w = 0.7$ mixes attention and noise to match the real distribution, landing precisely at the optimal balance. At $w = 1.0$ (FID = 187.17, SSIM = 0.6243, PSNR = 17.84, LPIPS = 0.2012), performance collapses abruptly across all metrics except Corr. Karras et al. and Kynkäänniemi et al. established that stochastic input is essential for high-frequency texture and diverse outputs. Here, the fusion weight comes entirely from attention, with the noise discarded, stripping diversity and eliminating the high-frequency details essential to the real distribution.

4.2. Pareto analysis in multiple objective spaces

Across three metric pairs (FID vs. Corr, FID vs. SSIM, PSNR vs. LPIPS), a tension exists between distribution matching and structural fidelity. Both $w = 0.7$ and $w = 0.8$ lie on the Pareto front in the FID-SSIM space and neither dominates the other. Faithfulness to input and fidelity to the real images operate at different levels. FID targets global distribution and SSIM focuses on local structure. This is consistent with Sajjadi et al.'s Precision-Recall framework [12]. $w = 0.7$ leans toward realism and diversity, and $w = 0.8$ prioritizes structure and fidelity. The point $w = 0.9$ (PSNR = 18.15, LPIPS = 0.1949) is the single optimal solution in the PSNR-LPIPS space. As Karras et al. [13] observed, both metrics improve continuously from $w = 0.1$ to 0.9. This suggests the generator captures both structure and texture. The contribution of the sketch path increases steadily, and the noise path retains enough texture to drive LPIPS downward although progressively attenuated. Luzi et al. [14] noted that perceptual quality depends more on higher-order moments than pixel-wise fidelity. Hence, even a small amount of noise sustains optimal perception, while zero noise leads to instant degradation.

5. Conclusion

This paper presents a lightweight dual-path fusion framework for sketch generation. Dense sampling of the fusion weight w at 0.1 step size reveals three phenomena. When $w < 0.7$, all metrics improve simultaneously, verifying dual-path complementarity, consistent with Heusel et al. and Kynkäänniemi et al. that fidelity and diversity can be jointly improved before model saturation. At w=0.7, the model achieves an optimal balance between structural fidelity and diversity. At w=1.0, performance collapses, confirming that a single information source degrades quality and diversity.

DOI: 10.54254/2755-2721/2026.36074

Following Saxena et al. [15], this work decompose the objective space into 2D subspaces and identify the Pareto optimal set $\{w = 0.7,0.8,0.9\}$, where $w = 0.7$ optimizes distribution matching, $w = 0.8$ maximizes structural similarity (consistent with Sajjadi et al.), and $w = 0.9$ achieves the perceptual quality Pareto optimum (aligning with Karras et al.), with PSNR and LPIPS improving jointly alongside both low- and high-frequency structure. Users can choose based on application scenario, highlighting the need for multi-dimensional evaluation. The framework generalizes to cross-modal tasks, and incorporating Pareto optimality may enable multi-objective optimization.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*, 139. https://doi.org/10.1145/3422622

[2] Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. arXiv:1611.07004. https://arxiv.org/abs/1611.07004

[3] Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv:1703.10593. https://arxiv.org/abs/1703.10593

[4] Ahsan, M. M., & Raman, S. (2025). A comprehensive survey on diffusion models and their applications. *Applied Soft Computing, 181*, 113470. https://doi.org/10.1016/j.asoc.2025.113470

[5] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. arXiv:2112.10752. https://arxiv.org/abs/2112.10752

[6] Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. arXiv:2302.05543. https://arxiv.org/abs/2302.05543

[7] Sarukkai, V., Yuan, L., Tang, M., Agrawala, M., & Fatahalian, K. (2024). Block and detail: Scaffolding sketch-to-image generation. arXiv:2402.18116. https://arxiv.org/abs/2402.18116

[8] Parmar, G., Park, T., Narasimhan, S., & Zhu, J.-Y. (2024). One-step image translation with text-to-image models. arXiv:2403.12036. https://arxiv.org/abs/2403.12036

[9] Navard, P., Monsefi, A. K., Zhou, M., Chao, W.-L., Yilmaz, A., & Ramnath, R. (2024). KnobGen: Controlling the sophistication of artwork in sketch-based diffusion models. arXiv:2410.01595. https://arxiv.org/abs/2410.01595

[10] Grnarova, P., Levy, K. Y., Lucchi, A., Perraudin, N., Goodfellow, I., Hofmann, T., & Krause, A. (2018). A domain agnostic measure for monitoring and evaluating GANs. arXiv:1805.09528. https://arxiv.org/abs/1805.09528

[11] Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., & Aila, T. (2019). Improved precision and recall metric for assessing generative models. arXiv:1904.06991. https://arxiv.org/abs/1904.06991

[12] Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O., & Gelly, S. (2018). Assessing generative models via precision and recall. arXiv:1806.00035. https://arxiv.org/abs/1806.00035

[13] Karras, T., Laine, S., & Aila, T. (2018). A style-based generator architecture for generative adversarial networks. arXiv:1812.04948. https://arxiv.org/abs/1812.04948

[14] Luzi, L., Marques, C. R. A. O., & Baraniuk, R. G. (2021). Evaluating generative networks using Gaussian mixtures of image features. arXiv:2110.05240. https://arxiv.org/abs/2110.05240

[15] Saxena, D. K., Mittal, S., Deb, K., & Goodman, E. D. (2024). *Machine learning assisted evolutionary multi- and many-objective optimization*. Springer, Singapore. https://doi.org/10.1007/978-981-99-2096-9