Realistic Sketch to Image Synthesis using RePainting-based diffusion model: Final Report

G016 (s2531301, s2506898, s2439263)

Abstract

This research develops a novel conditional diffusion model with RePainting for sketch-to-image synthesis, surpassing existing GAN-based methods in creating high-quality, and high-fidelity images. Apart from building the model from scratch, we created a dataset of paired photos and sketches, enabling advanced model training and evaluation. Our model not only presents a new technique in image synthesis but also expands the current scope of research in this field, with broad applications from creative industries to law enforcement. The study is focused, methodical, and poised to significantly impact the accessibility and quality of digital image creation.

1. Introduction

Sketches represent a fundamental form of visual expression, created without complex tools, focusing primarily on the shape and structure of the observed subject. They serve as a medium for capturing memories and creativity, allowing individuals to recreate and recall scenes over time. In contrast, photographs offer a more detailed and color-rich representation of visuals. The proposed project delves into the transformation of these simple sketches into detailed, realistic images, a process that could revolutionize the way we interact with art and technology.

The field of high-resolution image synthesis, a notable achievement in artificial intelligence, presents an interactive, intuitive, and mind-provoking application across various demographic groups, including children, teenagers, and professionals. This project is particularly intriguing due to its potential to lower technological barriers, enabling those without significant artistic skills to create realistic images. Such technology could significantly boost efficiency in creative industries, assist in graphic design, and even aid in law enforcement, such as in facial reconstruction for crime investigations. The evolution of sketch-based image synthesis has seen a transition from image retrieval methods like Photosketcher (Eitz et al., 2011) and Sketch2photo (Chen et al., 2009), which relied on intricate feature representations and post-processing techniques. Recent developments in AI, particularly in stable diffusion models (Rombach et al., 2021), have offered new pathways for image generation, suggesting an alternative approach to image synthesis from sketches.

The core objective of this project is to develop a novel conditional diffusion model using RePainting techniques for image synthesis from sketches. This approach seeks to improve upon existing methods of generating high-quality, realistic images. Additionally, exploring various network building blocks for model enhancement can be an optional objective.

This project aims to address the following research question:

> How does the proposed RePainting-based diffusion model compare with the existing GAN-based models in generating high-quality, realistic images from sketches?

The project's contribution provides a new approach in the field of sketch-to-image synthesis. First, it introduces a novel approach by employing a diffusion model with RePainting, an alternative to existing GAN-based methods like SketchyGAN and pix2pix, which often struggle to produce high-quality, realistic images. This innovative method holds the potential to achieve superior image synthesis quality. Second, it extends existing work by applying the model to a unique dataset and incorporating a previously unreported RePainting technique, thereby pushing the boundaries of current research in this domain. Finally, the project contributes by creating a new dataset comprising paired photos and sketches, which is crucial for training and evaluating the proposed model, further enriching the resources available for future research in this area.

The methodology involves collecting and preprocessing data by generating sketches for the dataset using a photo-sketching model (Li et al., 2019b) and ensuring data cleanliness for coherence. The training process involves a novel conditional diffusion model tailored for synthesizing images from three distinct classes, namely clothes, car, and dog. The sampling process employs the RePainting technique (Lugmayr et al., 2022), using sketch inputs to guide the denoising process and generate corresponding high-resolution images. The model's performance will be evaluated using metrics like LPIPS, FID, and inception scores for the different classes in the dataset. Due to hardware constraints, the project is limited to working with 64x64 pixel resolution images. The outcomes of this model will be critically analysed to determine its effectiveness in sketch-to-image synthesis, comparing it against current state-of-the-art methods.

2. Related work

The evolution of generative models for image synthesis has seen a significant shift from Generative Adversarial Networks (GANs) to diffusion models, addressing the limitations in image diversity and training stability that GANs present. In the late 2000s, the very first sketch2image model by Chen et al. (2009) proposed that the synthesized picture is generated by seamlessly stitching several photographs in agreement with the sketch and text labels. Afterward, GANs were introduced (goo), and have been conditioned on a variety of inputs, including discrete labels (Mirza & Osindero, 2014), text (Mou et al., 2023), and images (Isola et al., 2017), achieving notable success in image-to-image translation tasks.

Image-to-Image Translation with Conditional Adversarial Networks (Isola et al., 2017), or the pix2pix framework, has been a landmark in this domain, utilizing BEGAN (Berthelot et al., 2017) for image translation through a robust architecture and equilibrium concept that balances the generator and discriminator to control the trade-off between image diversity and visual quality. Additionally, Chen & Hays (2018) proposed a novel approach, SketchyGAN, for synthesizing realistic images from human-drawn sketches. It presents a data augmentation technique for sketches and introduces a new network building block, the Masked Residual Unit (MRU), which significantly improves the quality of the synthesized images by allowing the input image to influence the network at multiple scales. The approach outperforms state-of-the-art image translation methods in realism and diversity, as demonstrated by significantly higher Inception Scores (a metric introduced by Salimans et al. (2016)). However, GANs have been criticized for their lack of diversity and complex training process which collapses without carefully selected hyperparameters and regularisers (Dhariwal & Nichol, 2021).

Diffusion models have emerged as a compelling alternative, offering a solution to the diversity and stability issues associated with GANs. With the advent of multimodality, the text has been embedded into the denoising process along with the noise map to allow more precise controls (Mou et al., 2023; Saharia et al., 2021). This allowed users to generate images based on both spatial conditioning and text prompts, enabling various image synthesis applications that demand image manipulation guided by user constraints (Zhang et al., 2023) and user sketches (Zeng et al., 2021). These likelihood-based models have demonstrated the capability to produce high-quality images, ensuring broad distribution coverage, a stationary training objective, and scalability (Dhariwal & Nichol, 2021). The transition to diffusion models began to gain momentum in the 2020s with the introduction of Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020). DDPMs laid the foundation for this new class of generative models, showcasing their potential to generate high-quality images. This was further evidenced by the work of Dhariwal & Nichol (2021), which highlighted the superiority of diffusion models over GANs in terms of class coverage, image quality, and training stability. The advancements in diffusion models have not only been recognized in academic circles but have also been adopted by cutting-edge generative models such as DALL-E (Ramesh et al., 2021) and Stable Diffusion (Rombach et al., 2022).

Building on the success of DDPMs, classifier-free guidance represents a significant advancement in this domain, eliminating the need for a separate classifier to guide image generation towards a targeted class (Ho & Salimans, 2022). Here, we propose a model that leverages classifier-free techniques, conditional denoising diffusion probabilistic models, and the RePainting sampling algorithm proposed by Lugmayr et al. (2022) with our optimization to predict missing RGB pixels, conditioned on the noised sketch. We use pix2pix as the benchmark. By addressing the trade-off between diversity and fidelity, these improvements seek to establish a new direction of the RePainting technique for the pix2pix model architecture, surpassing the achievements of the previous pix2pix model across various metrics.

3. Dataset

Our research employs a manually-built dataset comprising 10,300 paired images and sketches across three distinct categories, as detailed in Table 1. This dataset was carefully constructed according to the pipeline outlined in Figure 1. We resized the images to 3x64x64 due to the limitation of computational power. We also reduced the dimension of sketches, which are greyscale, to 1x64x64.

Category	train set count	test set count
clothes	3000	100
car	3000	100
dog	4000	100

Table 1. Constructed dataset for the project. In total there are 10,000 sketch-image pairs for training, and 300 pairs for testing.

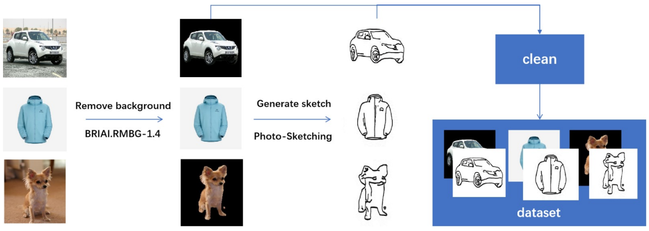

Figure 1. Dataset construction pipeline.

3.1. Why a New Dataset?

In the realm of sketch-based image retrieval (SBIR), foundational datasets such as the Sketchy database (Sangkloy et al., 2016), with its 75,471 sketches across 125 categories, serve as critical baselines. Similarly, other pivotal datasets all focus on the quantity of categories, rather than of the sketch-image pairs for each category (Qi et al., 2015; Yu et al., 2016). These datasets, while comprehensive, do not sufficiently cater to the nuances of diffusion-based modeling due to each class' limited size. Prior initiatives have attempted to bridge this gap by integrating 60,502 natural images from ImageNet into the SBIR datasets (Liu et al., 2017). However, this expanded dataset is not publicly accessible. Consequently, we were motivated to create a handcrafted dataset encompassing a broader spectrum of classes to support our diffusion approach effectively.

3.2. Dataset Construction Pipeline

3.2.1. Selection of Categories

Our selection criteria for the three categories were informed by the quality of the corresponding real-world images, adhering to principles of ablation in the sense that each class represents a different level of difficulty for the model. Each set of images was drawn from a public dataset (Khosla et al., 2011; Li, 2018; Lew, 2023). The dog category, illustrated in Figure 2, represents a highly diverse dataset that not only highlights lighting and shadows but also showcases a variety of poses and angles of different types of dogs. The car category, though also depicting 3D objects, exhibits less diversity. Lastly, the clothing category represents a simpler, 2D dataset with less types of garments. This separation allows us to evaluate our model's performance across varying complexity levels.

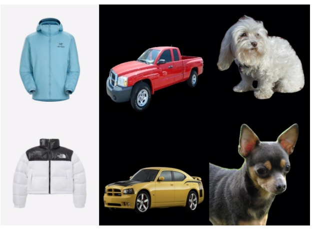

Figure 2. Dataset overview: three categories, each with two examples - clothes, car, and dog (left to right).

3.2.2. Construction of Sketch-Image Pairs

The initial step involved converting selected images into sketches. We employed the photo-sketching algorithm by Li et al. (2019a), which is capable of generating contour drawings akin to human sketches. However, initial trials produced sketches with excessive background noise. To mitigate this, we utilised Bria AI's RMBG model (AI, 2023) for background removal, thereby enhancing the foreground focus of the sketches.

3.2.3. Refinement of Sketch-Image Pairs

Despite the efficiency of the sketch generation process, a subset of the sketches were deemed of insufficient quality and were subsequently excluded. Criteria for exclusion included poor object outline clarity and excessive background noise, as exemplified in Figure 3. After we cleaned the sketches, we removed the corresponding ground truth data as well. This meticulous curation ensured the high quality of our final dataset.

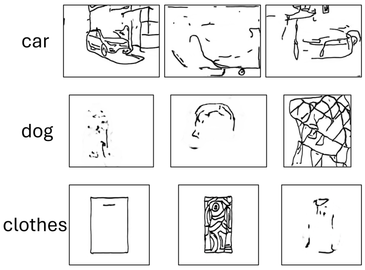

Figure 3. Example sketches for deletion: each is outlined for clear separation, showcasing examples selected for removal.

4. Methodology

We propose a method shown in Figure 4 that combines DDPM (Ho et al., 2020) with the RePaint approach (Lugmayr et al., 2022) for sketch-to-image translation. This enables the generation of images guided by sketches without altering the DDPM model structure. Our approach consists of two steps. Firstly, we converted the sketch images into 1x64x64 grayscale images, which are then stacked with the original 3x64x64 images to form 4x64x64 training data. We trained a conditional DDPM model on the classified stacked data to generate image-sketch pairs. During the sampling process, we iteratively introduce noise to the input sketch to obtain a sequence of T noisy sketch images. These noisy sketches serve as guiding conditions for the sampler, gradually replacing the sketch layer during the sampling process to guide the sampling of the image layer.

4.1. Training

We employed conditional DDPM as our generative model, which exhibits superior image quality and training stability compared to GANs (Goodfellow et al., 2020). However, our task involves generating sketch-image data rather than just images. The model training consists of two parts: a forward process and a backward process. The forward process involves adding Gaussian noise ϵ to the original data x_0 repeatedly until it transforms into a fully noisy image x_t. This process can be described as:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon_1, \tag{1}$$

where α and $\sqrt{1-\alpha}$ represent the weights of the image and noise, respectively. By adjusting α, we can control the intensity of the noise added. Because the whole process is cumulative multiplication, it can be reduced to the formula (2) so that it allows the image x_t at time t to be calculated based on the image x_0 directly:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t. \tag{2}$$

Then, we represent the distribution of noise at time t as

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t)I). \tag{3}$$

To implement the reversion of the noise image, the noise distribution of every step is the key. Assume the noise distribution of step t is known, and the condition is x_0.

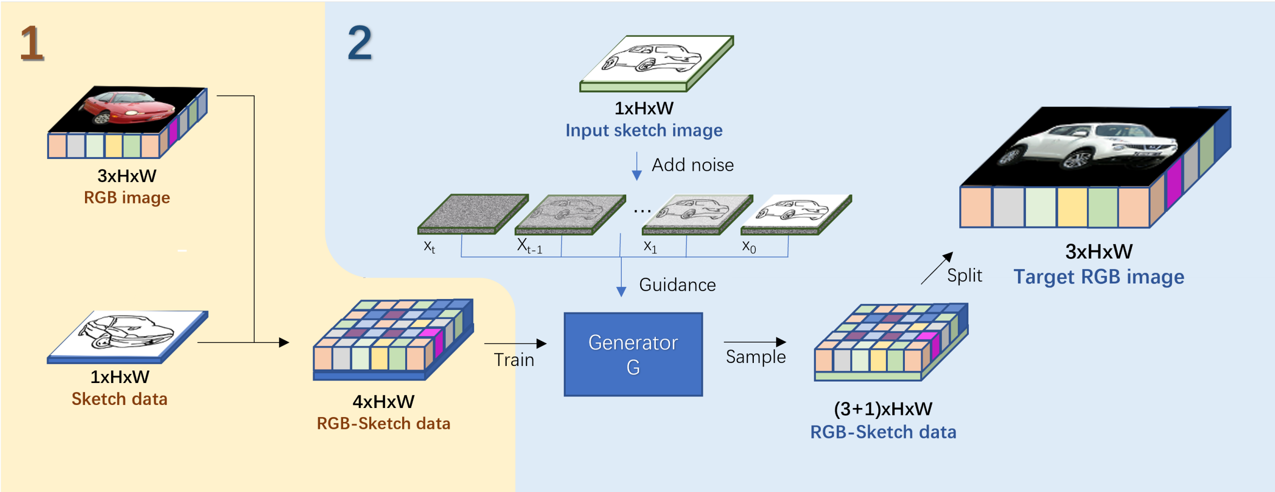

Figure 4. The first part(1) represents the training process, which involves merging corresponding images and sketches to construct 4xHxW training data, and using conditional DDPM for training. The second part(2) involves guided sampling using the repaint algorithm, where different levels of noisy input sketches are used to guide the entire generation (denoising) process, thus achieving the generation of target RGB data

The distribution in x_{t-1} can be expressed as the left part of Formula (4). Then, use Bayes's rule to get the right part of Formula (4),

$$q(x_{t-1}|x_t, x_0) = q(x_t|x_{t-1}, x_0)\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)}. \tag{4}$$

Further derivation leads us to an expression for the mean μ of the noise at time t as follows

$$\mu_t = \frac{1}{\sqrt{a_t}}(x_t - \frac{1 - a_t}{\sqrt{1 - \bar{a}_t}}\epsilon_t). \tag{5}$$

Subsequently, we introduce a U-Net model to learn and predict the mean μ of the noise at the corresponding time t. This allows us to formulate the loss function of the training, which can be expressed as follows,

$$L_t = \mathbb{E}_{t\sim[1,T],x_0,\epsilon_t}[||\epsilon_t - \epsilon_\theta(x_t, t)||^2], \tag{6}$$

where ϵ_t and $\epsilon_\theta(x_t, t)$ are the actual noise and predicted noise.

4.2. Sampling

Once the model is trained, we used a custom sampler edited for guidance sampling. The guidance process is shown in Algorithm 1. First, we employed the forward part of the model to produce a set of T images with the different levels of noise, T being the number of times noise is added. At every step of the sampling process, we replaced the sketch layer in each denoising result with the noisy image x_t. Subsequently, the sampler generates the denoising sketch-image data for the next time step based on the modified samples.

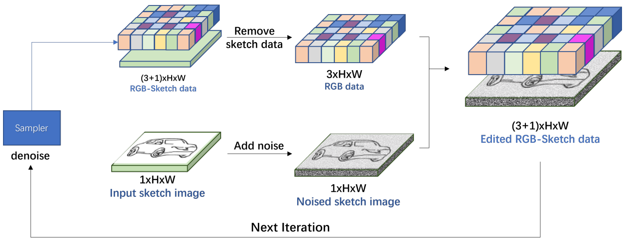

Figure 5. The image depicts the entire process of RePaint sampling. At each step of the sampling, the sketch data is replaced by the corresponding noisily inputted draft sketch, and then proceeds to the next sampling step, repeating until the total number of sampling steps is reached.

In this iterative process shown as Figure 5, the sketch layer in each generated sample is replaced by guiding information, and the sampler proceeds to the next step of sampling based on the modified sketch-image data. This process directs the sampler to synthesize RGB data in the corresponding image layer. This is the key to achieving the translation from sketch to image.

Although this method effectively guides the generation of RGB images, the replacement of the sketch layer during denoising introduces changes in the mean and variance of the original information. This alteration subsequently deteriorates the denoising effectiveness. To mitigate the impact of replacement on data inference, the RePaint algorithm repetitively adds noise to the data after replacing the sketch layer and then denoises it U times. This approach reduces the anomalous fluctuations introduced by the new sketch layer and enhances its guiding effect on the sampled data. A larger value of U often leads to better guidance effects, but it increases the sampling time cost. Thus, selecting an appropriate U is crucial for sampling. However, experimental results indicate that even with an appropriate U, although the RGB images closely match the contour features of the sketch, the representation of image details is inadequate. This issue arises because the influence of the RePaint opera-

Algorithm 1 Sampling process based on RePaint approach

```
$x_T \sim \mathcal{N}(0, I)$
for $t = T, ..., 1$ do
    for $u = 1, ..., U$ do
        $\epsilon \sim \mathcal{N}(0, \mathcal{I})$ if $t > 1$ else $\epsilon = 0$
        $x_{t-1}^{\text{sketch}} = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}} \epsilon$
        $z \sim \mathcal{N}(0, \mathcal{I})$ if $t > 1$ else $z = 0$
        $x_{t-1}^{\text{image}} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(x_t, t)) + \sigma_t z$
        $x_{t-1} = [1, 0, 0, 0]^T \cdot x^{\text{sketch}} + [0, 1, 1, 1]^T \cdot x^{\text{image}}$
        if $u < U$ and $t > 1$ then
            $x_t \sim \mathcal{N}(\sqrt{1 - \beta_{t-1}} x_{t-1}, \beta_{t-1} I))$
        end if
    end for
end for
return $x_0$
```

tion on data variance and mean cannot be simply eliminated through multiple resampling iterations, leading to the loss of image details due to slight perturbations.

4.3. Repaint-mixed

To address the problem of detail loss caused by the RePaint method, we propose an improved sampling method called RePaint-mixed. By observing the denoising process of the generator, we found that the details of the image are often determined in the later stages of denoising, while the early stages of denoising typically determine the contour of the image.

Based on this principle, we innovatively combined the RePaint algorithm with normal denoising algorithms. We divided the entire denoising process into RePaint and normal denoising parts. In the RePaint part, we increased the number of resampling iterations, strengthening the guiding effect of the new sketch layer, ensuring that the entire RGB data obtains an accurate contour that matches the sketch in the early stages of sampling. In the later stages of sampling, the sampler automatically switches to a regular sampling method to avoid introducing fluctuations in the data. Through this method, we can ensure that the RGB data is sufficiently guided by the input sketch, forming an image contour that conforms to the sketch features, while also ensuring the quality of the final image details.

5. Experiments

We implemented our RePaint-DDPM model and compared it with two selected baseline models (Pix2PixGAN and Pix2PixBEGAN) for our experiment. The baseline models are trained using our constructed dataset as well. We applied three different metrics on the test set result for comparison. As we take a novel approach, the aim of this experiment is simply to validate that our model functions as intended and performs comparably with other, existing approaches.

5.1. Evaluation

In the evaluation of the proposed diffusion model's performance, we applied three automatic quantitative metrics, Fréchet Inception Distance (FID) (Heusel et al., 2017), Inception Score (IS) (Barratt & Sharma, 2018), and LPIPS (Learned Perceptual Image Patch Similarity) (Zhang et al., 2018b), to ensure a comprehensive assessment. The FID captures both the fidelity (realism of the target class), quality, and diversity of generated images by comparing statistical distributions of real and generated images' feature representations, calculated as follows:

$$\begin{aligned} FID &= d^2\left((m, C), (m_w, C_w)\right) \\ &= \|m - m_w\|_2^2 + \mathrm{Tr}\left(C + C_w - 2(CC_w)^{\frac{1}{2}}\right), \end{aligned} \tag{7}$$

where m, C and m_w, C_w represent the mean and covariance of the generated and real images' feature representations respectively.

LPIPS quantifies the perceptual similarity between two images, capturing the human visual system's sensitivity to various image distortions (Zhang et al., 2018a), calculated as follows:

$$LPIPS = \frac{1}{N} \sum_{i=1}^{N} \left\| f(\text{Image}_1)_i - f(\text{Image}_2)_i \right\|_2 \tag{8}$$

, where N is the number of image pairs. Image_1 and Image_2 are the first and second images in the i-th pair, $f(\cdot)$ is a function that extracts feature vectors from images, typically obtained from a pre-trained convolutional neural network (CNN), $\|\cdot\|_2$ denotes the L2 norm (Euclidean distance), which measures the distance between the feature vectors of the two images. It is suitable for sketch-to-image synthesis evaluation as it assesses how closely the generated images resemble the target images in a way that aligns with human judgment, making it a robust measure of visual fidelity. IS evaluates the diversity and clarity of generated images through the entropy of the conditional label distribution predicted by an Inception model, which is calculated as

$$IS(G) \approx \exp\left(\frac{1}{N} \sum_{i=1}^{N} D_{KL}(p(y|x^{(i)}) \parallel \hat{p}(y))\right), \tag{9}$$

where D_{KL} is KL divergence. It uses a pre-trained Inception model to measure how distinct the generated images are from one another (diversity) and how confidently the model classifies each image (clarity). This metric is appropriate for our task because it helps ensure that our model not only generates realistic images but also produces a varied set of outputs, reflecting the diverse interpretations possible from a single sketch input. They provide a comprehensive evaluation framework, assessing fidelity, perceptual quality, and diversity of the generated images, which are critical aspects of the sketch-to-image synthesis performance.

5.2. Baseline models

For our comparison, we use two baseline models based on Pix2Pix (Isola et al., 2017), one based on a simple GAN architecture and the other based on BEGAN (Berthelot et al., 2017). We select these as our baseline due to their high quality among GAN-based models and due to their ease of training.

In the base Pix2Pix model, the discriminator maps image-sketch pairs to small "patches" of numbers, and is trained to distinguish generated images (on which it aims to produce a patch of all zeros) from real ones (a patch of all ones) using mean-squared error. The generator maps sketches to images, and is trained firstly to fool the discriminator but also to minimise pixelwise l1 loss between the generated image and the corresponding ground truth image for that sketch. The pixelwise loss is intended to force the generator to balance producing convincing fakes with following the sketch that it is provided. Without this loss, it may act more like a traditional generative model rather than a translator.

The Pix2Pix BEGAN model instead uses a UNet as its discriminator. It learns to approximate the identity function on realistic images within a specific class (say, dogs), but factored through a significantly smaller vector space such that information about that class must be retained in the model weights instead of being simply preserved. The quality of the reconstruction therefore gives an idea of how close the input is to being a realistic image of that class. In this case, the generator learns to produce images on which the discriminator performs well, while the discriminator learns to perform well on real images and poorly on generated images. Again, the generator also learns to minimise the pixelwise L1 loss with the ground truth image for its given sketch.

We implemented both models based on online implementations hosted on GitHub[1]. Both models required some modification, in particular due to breaking changes to the underlying ML frameworks. The discriminator for the BEGAN model, a nine-layer UNet, was implemented from scratch as per Berthelot et al. (2017). Both models are trained for 100 epochs for each class, in total 300 epochs for the training dataset.

5.3. DDPM RePaint models

We trained a conditional DDPM model for 400 epochs using the same dataset. The batch size during training was 28, with the Adam optimiser and a learning rate of 1e-4. Mean squared error was employed as the loss function. Subsequently, utilising the same weights, we devised two guidance sampling strategies as mentioned in the methodology section in our ddpm sampler, resulting in two sketch-to-image models named ddpm-repaint and ddpm-repaint-mixed.

5.4. Result

The partial sampling results of the four models are shown in Figure 6, where the legends from left to right represent input sketch, Ground Truth, Pix2PixGAN, Pix2PixBEGAN, DDPM-RePaint, and DDPM-RePaint-mixed, respectively. Visual inspection reveals that the images generated by RePaint and RePaint-mixed exhibit fewer artifacts, especially in terms of noise. Notably, the samples produced by RePaint-mixed demonstrate lighting and details more in line with the Ground Truth, reflecting higher image quality. Moreover, utilising DDPM as the generator results in more vibrant and rich colors in the images generated by RePaint and RePaint-mixed. Particularly in categories such as car and clothes, the differences among samples from GAN-based models are minimal. Our models not only maintain good sketch fidelity but also ensure sample diversity. However, RePaint-mixed exhibits less stability in the dog category; although it preserves more details, it struggles to accurately represent the shape of dogs. In contrast, Pix2PixGAN depicts dog outlines more accurately with higher sketch fidelity but lacks color richness.This suggests that the task of generating dog images is relatively more complex compared to the other two classes, and the model requires further training. However, it's also possible that the limited quantity of data in the dog dataset is insufficient to provide enough features for the model to learn from. It is worth noting that the background of car samples from RePaint and RePaint-mixed is not pure black. This is attributed to residual background information present in some car images in the training set, which the model learns and incorporates into the generated samples. This disparity highlights the differences in learning capabilities between the two generation models, demonstrating the stronger learning ability of DDPM-based models. As depicted in Table 2, which presents the performance of the four models across three categories and three metrics, our method achieves promising results. In terms of FID, our method's RePaint score is comparable to the baseline, whereas RePaint-mixed signif-

Class	**Model**	**FID**	**IS**	**Lpips**
Car	Pix2PixBeGAN	279.21	1.93	0.35
	Pix2PixGAN	198.54	2.85	**0.31**
	RePaint **(ours)**	170.54	**2.91**	0.44
	RePaint-mixed **(ours)**	**94.09**	2.37	0.43
Clothes	Pix2PixBeGAN	181.96	3.45	0.29
	Pix2PixGAN	152.32	**4.47**	0.29
	RePaint **(ours)**	195.13	3.24	0.30
	RePaint-mixed **(ours)**	**99.21**	4.10	**0.29**
Dog	Pix2PixBeGAN	254.61	3.35	0.33
	Pix2PixGAN	254.21	4.06	**0.32**
	RePaint **(ours)**	258.25	4.08	0.41
	RePaint-mixed **(ours)**	**244.42**	**4.32**	0.38

Table 2. Performance Metrics for Sketch-to-Image Models

[1] https://github.com/Vargha-Kh/Pix2Pix-GAN

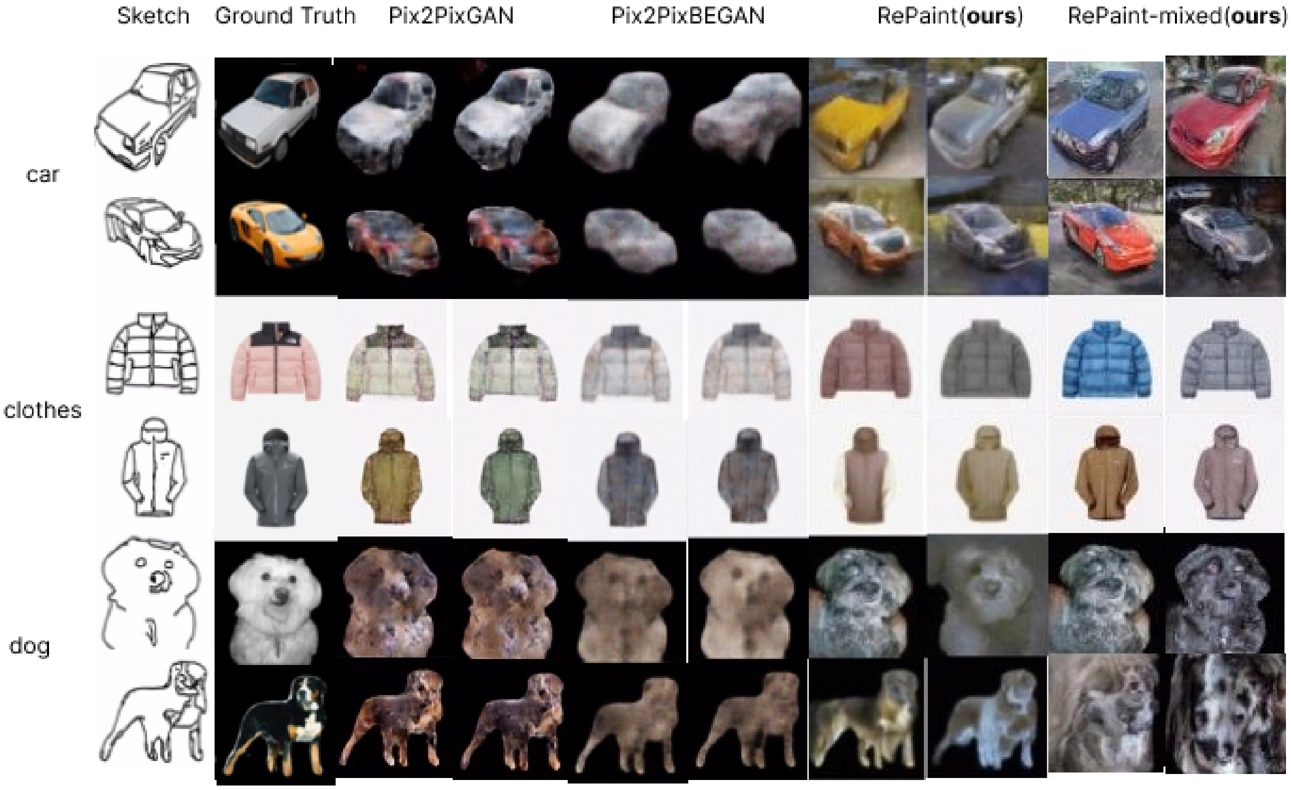

Figure 6. Three classes dataset Qualitative Results, the generated results are of size 64 by 64

icantly outperforms the baseline, particularly with scores of 94.09 and 99.21 for car and clothes categories, respectively, nearly doubling the improvement compared to GAN (198.54) (152.32). This indicates that images generated by RePaint-mixed more closely resemble the feature distribution of real images, thus achieving higher image quality. This is further supported by the clearer details observed in the results of RePaint-mixed compared to other models, as shown in Figure 6. For instance, the reflections on the car body align more closely with the characteristics of real images.

Regarding the Inception score, our method performs consistently well, with the RePaint model (2.91) slightly outperforming other models in the car category, and the RePaint-mixed model (4.32) exhibiting the best performance in the dog category. This suggests that our model not only produces higher-quality images but also demonstrates better diversity. However, our method performs poorly in lpips, especially in the car category, where both RePaint (0.44) and RePaint-mixed (0.43) fall short of the baseline. This indicates a significant perceptual gap between the generated images and the ground truth. We attribute this to the influence of image backgrounds; some car images in the training set have backgrounds that were not completely removed, causing the DDPM model to learn this information, resulting in generated samples often having a blurry background, which differs significantly from the pure black background of the Ground Truth, leading to a decrease in lpips scores.

Overall, RePaint-mixed demonstrates more comprehensive performance compared to the baseline, showcasing superior performance across different categories.

6. Conclusions

In this project, we successfully applied a diffusion model to the task of generating images based on sketches. We constructed and trained a RePaint DDPM model capable of producing high-quality images from input sketches. Additionally, we created a dataset consisting of sketch-image pairs across three different categories for our model. From our results, it is evident that our RePaint DDPM model outperforms baseline models based on GAN in terms of FID and IS metrics.

While our results may not match those of state-of-the-art diffusion models employing meta-learning techniques, our primary contribution lies in introducing a novel approach that combines a basic DDPM diffusion model with the RePaint sampling algorithm for sketch-to-image generation. Our experiments validate the viability of this approach and demonstrate its effectiveness compared to GAN-based models. Furthermore, we refined the RePaint method by introducing RePaint-Mixed, which allocates RePaint sampling and regular sampling to appropriate steps in the sampling process, thereby further improving the quality of generated images.

6.1. Future work

The subsequent phase of our research can concentrate on unifying the noise encoder and decoder into a cohesive end-to-end trainable framework, with the expectation of scaling up image resolution from the present 64x64 to 1028x1028, a transformation necessitated by existing hardware constraints. Our model can also be refined to edit real images (Meng et al., 2021), such as altering the colors of apparel or vehicles, employing noise addition and subsequent denoising aligned with text prompts, and leveraging LoRA for network adaptation to enhance control and reduce overfitting and tuning durations. Additionally, we plan to incorporate Denoising Diffusion Implicit Models (DDIMs) into our existing DDPMs (Song et al., 2020). DDIMs, by introducing a non-Markovian process that permits the skipping of steps in the diffusion sequence, promise to significantly expedite the image generation process without markedly compromising on image quality, offering an efficient post-training integration that maintains the generation of high-quality, diverse images. Additionally, we aim to combine the multimodal space of CLIP to enable text-guided sketch-to-image generation, where sketches are guided by input text prompts.

References

AI, BRIA. Rmbg-1.4: Bria background removal v1.4. https://huggingface.co/briaai/RMBG-1.4, 2023. Accessed: 2024-03-24.

Barratt, Shane and Sharma, Rishi. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

Berthelot, David, Schumm, Tom, and Metz, Luke. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

Chen, T., Cheng, M.-M., Tan, P., Shamir, A., and Hu, S.-M. Sketch2photo: Internet image montage. *ACM Transactions on Graphics (TOG)*, 28(5):124, 2009.

Chen, Wengling and Hays, James. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/papers/Chen_SketchyGAN_Towards_Diverse_CVPR_2018_paper.pdf.

Dhariwal, Prafulla and Nichol, Alex. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

Eitz, M., Richter, R., Hildebrand, K., Boubekeur, T., and Alexa, M. Photosketcher: Interactive sketch-based image synthesis. *IEEE Computer Graphics and Applications*, 31(6):56–66, 2011.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Heusel, Martin, Ramsauer, Hubert, Unterthiner, Thomas, Nessler, Bernhard, and Hochreiter, Sepp. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017. URL https://arxiv.org/abs/1706.08500.

Ho, Jonathan and Salimans, Tim. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, Jonathan, Jain, Ajay, and Abbeel, Pieter. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Isola, Phillip, Zhu, Jun-Yan, Zhou, Tinghui, and Efros, Alexei A. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2017.

Khosla, Aditya, Jayadevaprakash, Nityananda, Yao, Bangpeng, and Fei-Fei, Li. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Lew, Hah Min. Kream product blip captions. https://huggingface.co/datasets/hahminlew/kream-product-blip-captions/, 2023.

Li, Jessica. Stanford cars dataset, Jun 2018. URL https://www.kaggle.com/datasets/jessicali9530/stanford-cars-dataset/data.

Li, Mengtian, Lin, Zhe, Mech, Radomir, Yumer, Ersin, and Ramanan, Deva. Photo-sketching: Inferring contour drawings from images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1403–1412. IEEE, 2019a.

Li, Y., Chen, Q., and Koltun, V. Photo-sketching: Inferring contour drawings from images. *arXiv preprint arXiv:1901.00542*, 2019b.

Liu, Li, Shen, Fumin, Shen, Yuming, Liu, Xianglong, and Shao, Ling. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2862–2871, 2017.

Lugmayr, Andreas, Danelljan, Martin, Romero, Andres, Yu, Fisher, Timofte, Radu, and Van Gool, Luc. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Meng, Chenlin, He, Yutong, Song, Yang, Song, Jiaming, Wu, Jiajun, Zhu, Jun-Yan, and Ermon, Stefano. Sdedit:

Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. URL https://arxiv.org/abs/2108.01073.

Mirza, Mehdi and Osindero, Simon. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. URL https://arxiv.org/abs/1411.1784.

Mou, Chong, Wang, Xintao, Xie, Liangbin, Wu, Yanze, Zhang, Jian, Qi, Zhongang, Shan, Ying, and Qie, Xiaohu. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. URL https://arxiv.org/abs/2302.08453.

Qi, Yonggang, Song, Yi-Zhe, Xiang, Tao, Zhang, Honggang, Hospedales, Timothy, Li, Yi, and Guo, Jun. Making better use of edges via perceptual grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1856–1865, 2015.

Ramesh, Aditya, Pavlov, Mikhail, Goh, Gabriel, Gray, Scott, Voss, Chelsea, Radford, Alec, Chen, Mark, and Sutskever, Ilya. Dall·e: Creating images from text, 2021. URL https://arxiv.org/abs/2102.12092.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2021.

Rombach, Robin, Blattmann, Andreas, Lorenz, Dominik, Esser, Patrick, and Ommer, Björn. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2401.03152.

Saharia, Chitwan, Chan, William, Chang, Huiwen, Lee, Chris A., Ho, Jonathan, Salimans, Tim, Fleet, David J., and Norouzi, Mohammad. Palette: Image-to-image diffusion models. *arXiv preprint arXiv:2111.05826*, 2021. URL https://arxiv.org/abs/2111.05826.

Salimans, Tim, Goodfellow, Ian, Zaremba, Wojciech, Cheung, Vicki, Radford, Alec, and Chen, Xi. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Sangkloy, Patsorn, Burnell, Nathan, Ham, Cusuh, and Hays, James. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016.

Song, Jiaming, Meng, Chenlin, and Ermon, Stefano. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. URL https://arxiv.org/abs/2010.02502.

Yu, Qian, Liu, Feng, Song, Yi-Zhe, Xiang, Tao, Hospedales, Timothy M, and Loy, Chen-Change. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 799–807, 2016.

Zeng, Yu, Lin, Zhe, and Patel, Vishal M. Sketchedit: Mask-free local image manipulation with partial sketches. *arXiv preprint arXiv:2111.15078*, 2021. URL https://arxiv.org/abs/2111.15078.

Zhang, Lvmin, Rao, Anyi, and Agrawala, Maneesh. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. URL https://arxiv.org/abs/2302.05543.

Zhang, Richard, Isola, Phillip, Efros, Alexei A, Shechtman, Eli, and Wang, Oliver. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018a.

Zhang, Richard, Isola, Phillip, Efros, Alexei A, Shechtman, Eli, and Wang, Oliver. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018b.