Image-to-Image Texture Synthesis with Diffusion Model

January 9, 2024

Federico Barreca

Abstract

This project aims to develop a diffusion model that is data-driven and capable of learning diverse texture patterns, such as brick walls, tatami mats, bark, etc. The proposed model leverages a small input image containing a portion of a texture, allowing it to extrapolate and complete the entire texture. Addressing the challenge of applying textures to large surfaces, the project draws inspiration from the limitations of conventional methods like stretching or tiling, which often result in low-resolution or visually unappealing artifacts. Referencing the need for detailed textures on planar surfaces, the project aligns with the observation that standard texture tiling or stretching can lead to undesirable outcomes. The project seeks innovative solutions to generate larger textures from smaller samples, avoiding visual repetitions or stretches. While existing methods provide some solutions, they often suffer from slow processing speeds and high error rates. The primary objective of this project is to employ deep learning techniques for efficient texture enlargement without pattern repetition. To evaluate the effectiveness of the proposed model, both qualitative and quantitative assessments will be conducted using metrics such as FID, Inception Score, and VIF. Check out `https://github.com/federicoBarreca/2024_Image-to-Image_Texture_Synthesis_with_Diffusion_Model` for an overview of the results and code.

1. Introduction

Textures play a fundamental role in computer graphics and its associated applications, including video games, movies, and digital design. These fields are inherently artistic, and the creation of new textures is a resource-intensive process, demanding both time and financial investment. Video games, in particular, necessitate rapid development cycles, and the historical trend of the industry reveals a prevalent pattern of using repeated textures for 2D models. In the context of expansive virtual environments, the recurrence of identical patterns can lead to monotony, diminishing the immersive quality and, in turn, undermining the intended purpose of video games: to provide entertainment and immersion. To address this issue, deep generative models serve as valuable tools for developers, providing powerful solutions. An inherent strategy for image-to-image translation involves learning the conditional distribution of output images given the input, employing deep generative models capable of capturing multi-modal distributions within the high-dimensional space of images. Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have emerged as the preferred model family for numerous image-to-image tasks (Isola et al., 2017). They demonstrate the ability to generate high-fidelity outputs, possess broad applicability, and support efficient sampling. However, the training of GANs can be challenging (Gulrajani et al., 2017), often leading to the omission of modes in the output distribution (Metz et al., 2016). Diffusion models prove advantageous in achieving superior results than GANs (Dhariwal & Nichol, 2021). This project explores techniques such as image conditioning, with both interclass and intra-class variations, and Exponential Moving Average (EMA) to enhance the efficacy of these models.

2. Related work

My work was inspired by Palette (Saharia et al., 2021), which explored image-to-image translations such as colorization, inpainting, uncropping, and JPEG restoration. Furthermore, my project exploits diffusion models (Ho et al., 2020) while considering the most recent improvements (Nichol & Dhariwal, 2021).

3. Method

The model itself relies on a implementation characterized by typical noise scheduling for the forward pass. Instead

Email: Federico Barreca <barreca.1736423@studenti.uniroma1.it>.

Deep Learning and Applied AI 2023, Sapienza University of Rome, 2nd semester a.y. 2022/2023.

of 1000 steps of noising, I opted for only 300 steps to obtain better-defined images in less time. The network architecture is built upon the 256×256 class-conditional U-Net model proposed in (Nichol & Dhariwal, 2021). There primary distinctions between my architecture and theirs lie in the absence of class-conditioning, the output size of 64x64 and the additional conditioning of the 32x32 source image through concatenation, as outlined in (Saharia et al., 2021). This image-to-timage diffusion model is a conditional diffusion model in the form $p(\mathbf{y}|\mathbf{x})$, where both $\mathbf{x}$ and $\mathbf{y}$ are images. The loss function is build so that given a training output image $\mathbf{y}$ to generate a noisy version $\mathbf{y_t}$. The network ϵ_θ is trained to denoise $\mathbf{y_t}$ given $\mathbf{x}$ and a noise level indicator $\bar{\alpha}_t$.

$$\mathcal{L}(\theta) = \|\epsilon - \epsilon_\theta(\mathbf{x}, \underbrace{\sqrt{\bar{\alpha}_t}\mathbf{y}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon}_{\mathbf{y_t}}, t)\|_2^2 \quad (1)$$

where:

- $\mathcal{L}(\theta)$ is the overall loss;
- $\epsilon \sim \mathcal{N}(0, I)$;
- $t \sim Uniform(\{1, ..., 300\})$;
- $\alpha_t = 1 - \beta_t$
- $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$

After training, a sampling process follows that generates the image by iteratively subtracting predicted noise from initial random noise while conditioning with $\mathbf{x}$. The denoising algorithm is repeated 300 times as follows:

$$\mathbf{y}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{y}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}, \mathbf{y}_t, t)) + \sigma_t \mathbf{z} \quad (2)$$

where:

- $\mathbf{z} \sim \mathcal{N}(0, I)$ if $t > 1$, $\mathbf{z} = 0$ otherwise;
- $\sigma_t = \sqrt{\beta_t}$

Finally, I implemented EMA in order to enforcing a smoother training and lead to a more robust outcome since it's not so susceptible to outliers in terms of updates of the main model. After copying the main model, the EMA model is updated as an interpolation between the old weights w_{old} and the new parameters w_{new} weighted by β:

$$w = \beta * w_{old} + (1 - \beta) * w_{new} \quad (3)$$

4. Evaluation and results

Qualitative evaluation. I provide some generated textures, from both main and EMA model, which can be compared with the real ones:

Figure 1. Real texture

Figure 2. Conditioning partial texture

Figure 3. Generated texture

Figure 4. Generated texture by EMA model

Quantitative evaluation. I use three quantitative metrics of sample quality for image-to-image translation: Inception Score (IS); Fréchet Inception Distance (FID) (Benny et al., 2020); Visual Information Fidelity (VIF) (Sheikh & Bovik, 2006).

Table 1. Performance evaluation.

	FID	IS	VIF
Model	114.62	1.69	0.61
EMA model	106.59	1.30	0.55

5. Conclusions and future work

The project and experiments I conducted demonstrate positive outcomes in terms of shape and pattern learning. There is a significant contrast in colors, with no notable distinctions between the main and EMA models. Both of these issues may be attributed to insufficient training time. I suggest visiting my GitHub repository to see more results.

References

Benny, Y., Galanti, T., Benaim, S., and Wolf, L. Evaluation metrics for conditional image generation. *CoRR*, abs/2004.12361, 2020. URL https://arxiv.org/abs/2004.12361.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *CoRR*, abs/2105.05233, 2021. URL https://arxiv.org/abs/2105.05233.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017. URL http://arxiv.org/abs/1704.00028.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020. URL https://arxiv.org/abs/2006.11239.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. Image-to-image translation with conditional adversarial networks. pp. 5967–5976, 07 2017. doi: 10.1109/CVPR.2017.632.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. Unrolled generative adversarial networks. *CoRR*, abs/1611.02163, 2016. URL http://arxiv.org/abs/1611.02163.

Nichol, A. and Dhariwal, P. Improved denoising diffusion probabilistic models. *CoRR*, abs/2102.09672, 2021. URL https://arxiv.org/abs/2102.09672.

Saharia, C., Chan, W., Chang, H., Lee, C. A., Ho, J., Salimans, T., Fleet, D. J., and Norouzi, M. Palette: Image-to-image diffusion models. *CoRR*, abs/2111.05826, 2021. URL https://arxiv.org/abs/2111.05826.

Sheikh, H. and Bovik, A. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2): 430–444, 2006. doi: 10.1109/TIP.2005.859378.