Diffusion-based Holistic Texture Rectification and Synthesis

Guoqing Hao
University of Tsukuba
Tsukuba, Ibaraki, Japan
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan
hao_guoqing@cvlab.cs.tsukuba.ac.jp

Satoshi Iizuka
University of Tsukuba
Tsukuba, Ibaraki, Japan
iizuka@cs.tsukuba.ac.jp

Kensho Hara
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan
kensho.hara@aist.go.jp

Edgar Simo-Serra
Waseda University
Shinjuku, Tokyo, Japan
ess@waseda.jp

Hirokatsu Kataoka
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan
hirokatsu.kataoka@aist.go.jp

Kazuhiro Fukui
University of Tsukuba
Tsukuba, Ibaraki, Japan
kfukui@cs.tsukuba.ac.jp

Figure 1: Rectified and synthesized texture results from real images. Our framework is able to rectify degraded textures, which include occlusions and geometric deformations, and synthesize holistic textures from selected areas. The first row presents real images, where the masked input to our framework is highlighted, while the second row shows outputs generated by our proposed approach. Our method accomplishes more than merely filling in missing regions; it also rectifies geometric deformations, including perspective variations and distortions to synthesize textures amenable for usage in many different applications such as 3D modelling. Photographs courtesy of Elliott Brown (CC-BY), Scott Meis (CC-BY), denisbin (Public Domain), Jameel Winter (CC-BY), and Bettina Arrigoni (CC-BY).

Authors' addresses: Guoqing Hao, University of Tsukuba, Tsukuba, Ibaraki, 3050085, Japan and National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, 3058560, Japan, hao_guoqing@cvlab.cs.tsukuba.ac.jp; Satoshi Iizuka, University of Tsukuba, Tsukuba, Ibaraki, 3050085, Japan, iizuka@cs.tsukuba.ac.jp; Kensho Hara, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, 3058560, Japan, kensho.hara@aist.go.jp; Edgar Simo-Serra, Waseda University, Shinjuku, Tokyo, 1698050, Japan, ess@waseda.jp; Hirokatsu Kataoka, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, 3058560, Japan, hirokatsu.kataoka@aist.go.jp; Kazuhiro Fukui, University of Tsukuba, Tsukuba, Ibaraki, 3050085, Japan, kfukui@cs.tsukuba.ac.jp.

SA Conference Papers '23, December 12–15, 2023, Sydney, NSW, Australia
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *SIGGRAPH Asia*

ABSTRACT

We present a novel framework for rectifying occlusions and distortions in degraded texture samples from natural images. Traditional texture synthesis approaches focus on generating textures from pristine samples, which necessitate meticulous preparation by humans and are often unattainable in most natural images. These challenges stem from the frequent occlusions and distortions of texture samples in natural images due to obstructions and variations in object

2023 Conference Papers (SA Conference Papers '23), December 12–15, 2023, Sydney, NSW, Australia, https://doi.org/10.1145/3610548.3618233.

surface geometry. To address these issues, we propose a framework that synthesizes holistic textures from degraded samples in natural images, extending the applicability of exemplar-based texture synthesis techniques. Our framework utilizes a conditional Latent Diffusion Model (LDM) with a novel occlusion-aware latent transformer. This latent transformer not only effectively encodes texture features from partially-observed samples necessary for the generation process of the LDM, but also explicitly captures long-range dependencies in samples with large occlusions. To train our model, we introduce a method for generating synthetic data by applying geometric transformations and free-form mask generation to clean textures. Experimental results demonstrate that our framework significantly outperforms existing methods both quantitatively and quantitatively. Furthermore, we conduct comprehensive ablation studies to validate the different components of our proposed framework. Results are corroborated by a perceptual user study which highlights the efficiency of our proposed approach.

CCS CONCEPTS

• **Computing methodologies** → **Image processing**; *Texturing*.

KEYWORDS

Texture rectification, texture synthesis, diffusion models

ACM Reference Format:
Guoqing Hao, Satoshi Iizuka, Kensho Hara, Edgar Simo-Serra, Hirokatsu Kataoka, and Kazuhiro Fukui. 2023. Diffusion-based Holistic Texture Rectification and Synthesis. In *SIGGRAPH Asia 2023 Conference Papers (SA Conference Papers '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3610548.3618233

1 INTRODUCTION

Textures are a crucial visual aspect of real-world scenes, representing surface appearance and consisting of repeating patterns with some inherent randomness. There are numerous applications in computer graphics and vision that use textures, including 3D modelling, image editing [Criminisi et al. 2003], virtual object creation [Chen and Rosenberg 2018], and augmented reality [Isoyama et al. 2021]. Textures can be derived from various sources such as hand-drawn images or natural images. In this work, we concentrate on synthesizing textures from natural images.

Traditional texture synthesis methods [Efros and Freeman 2001; Efros and Leung 1999; Wei et al. 2009; Wei and Levoy 2000] aim to generate arbitrarily large texture images indistinguishable from small input samples. However, these approaches require holistic textures, which are rectangular and free from geometric distortions. Obtaining such holistic textures demands extensive human intervention [Wei et al. 2009] and is often unattainable in most natural images. This limitation stems from frequent occlusions and distortions in real-world objects within natural images, caused by nearby obstructions and variations in the surface geometry of the objects. While recent work [Li et al. 2022b] has automated texture scraping from natural images by grouping texture regions and filling missing regions, it overlooks deformations, resulting in unnatural textures. Consequently, there is a pressing need to both handle occlusions and deformations in texture samples from real images.

We propose a novel framework that addresses these challenges by leveraging Diffusion Models (DM) [Ho et al. 2020] to synthesize holistic textures from degraded samples in real images. Due to pixel misalignment and a lack of correspondence between holistic textures and degraded samples, we empirically find generative adversarial networks (GANs) [Goodfellow et al. 2014] struggle to synthesize holistic textures from degraded samples. We argue that GANs, due to their susceptibility to mode-collapsing and the difficulty in capturing complex data distributions, often produce unnatural results. DM, on the other hand, provides a more efficient training process and produces a superior-quality of image sample, thanks to its stationary training objective and extensive data distribution coverage. Consequently, we adopt DM as the basis of our framework for rectifying and synthesizing holistic textures.

Our framework builds upon Latent Diffusion Models (LDM) [Rombach et al. 2022] and introduces a novel occlusion-aware latent transformer. The LDM operates in the latent space rather than the standard pixel space, drastically reducing computational costs. We build our framework upon LDM to allow further downstream applications such as integration into existing photo editing tools on personal computers. However, operating in the latent space unintentionally entangles valid and difficult usability of the features, which originate from occluded and unobstructed regions in sample textures. Discriminating these features is essential for rectifying degraded textures as the invalid features not only fail to contribute to the rectification process but can also impede it. To address this, we introduce an occlusion-aware latent transformer into the LDM model, delivering effective information to the rectification process. This latent transformer utilizes partial convolutional layers [Liu et al. 2018] to encode the degraded sample into a latent code composed solely of valid features, and incorporates a self-attention block [Zhang et al. 2019] to efficiently model the non-local dependencies of the degraded sample. We empirically demonstrate the effectiveness of the latent transformer and carefully analyze the importance of each component.

Moreover, we introduce a method for generating synthetic training data by applying geometric transformations and free-form mask generation to planar textures, simulating deformations, and occlusions in degraded samples from natural images. Specifically, we employ the homography transformation [Hartley and Zisserman 2003] and the thin plate spline transformation [Bookstein 1989] to simulate perspective variations and geometric distortions, respectively. We also make use of free-form masks [Yu et al. 2019] to mimic occlusions found in natural images. Our approach allows obtaining a vast number of degraded texture images from a finite number of planar texture images by introducing varying scales of the transformations, and enables end-to-end training of our LDM-based framework.

Experimental results attest to the superior performance of our framework compared to existing methods. Additionally, we conduct comprehensive ablation studies to validate the effectiveness of each component within our proposed framework. Finally, we perform a perceptual user study that corroborates the effectiveness of our approach.

Our contributions are summarized as follows:

- The first framework for rectifying occlusions and deformations in degraded sample textures from natural images, expanding the applicability of exemplar-based texture synthesis techniques.
- A novel occlusion-aware latent transformer that provides effective information to the texture rectification process.
- A synthetic data generation method to create training data for rectifying occlusions and deformations in texture samples.
- In-depth evaluation that demonstrates the superior performance of our framework compared to existing methods.

2 RELATED WORK

In this section, we discuss the related work on texture synthesis and generative models for image-to-image translation.

2.1 Texture Synthesis

Here we review the relevant literature on texture synthesis, including exemplar-based texture synthesis, texture exemplar extraction, and shape from texture.

2.1.1 Exemplar-based texture synthesis. Exemplar-based texture synthesis aims to generate arbitrarily large new textures that are perceptually similar to a given input sample texture. Early methods, such as [Efros and Freeman 2001; Efros and Leung 1999; Wei and Levoy 2000], utilized non-parametric techniques, copying pixels or patches sequentially while ensuring neighborhood consistency. These methods, although visually pleasing, were computationally demanding and struggled with complex patterns or large-scale structures. Recently, deep convolutional neural networks (CNNs) were employed by [Gatys et al. 2015] for texture synthesis, iterating between sample textures and random Gaussian noise. Despite revealing CNNs' potential, the optimization process remained slow. Alternative methods [Bergmann et al. 2017; Jetchev et al. 2016; Li et al. 2017] offered texture generation through a single CNN forward process, yet generalizing to unseen textures remained problematic. More recently, [Liu et al. 2020; Mardani et al. 2020] enabled unseen texture synthesis by upsampling textures in the Frourier domain or formulating texture synthesis as transposed convolution operations. Nevertheless, these approaches necessitate pristine samples, which are rectangular and free from geometric distortions.

Recently, [Li et al. 2022b] introduced an automatic texture extraction framework that groups texture regions and synthesizes large textures. Although this method can handle occlusions in degraded texture images, addressing deformations remains a significant challenge. In contrast, our framework efficiently deals with both occlusions and geometric deformations.

2.1.2 Texture exemplar extraction. Texture exemplar extraction is crucial in exemplar-based texture synthesis, as synthesis quality relies heavily on the selection of representative texture samples. Traditionally, this process is labor-intensive, necessitating expert input and significant resources. To mitigate this, [Wu et al. 2018] introduced an automated method for extracting texture exemplars from images, utilizing both global and local textureness measures. Building on this, [Wu et al. 2021] proposed a deep learning-based approach for texture exemplar extraction. However, frequent occlusions and deformations in natural images can hinder the extraction of appropriate exemplars. Therefore, rectifying these occlusions and distortions, as proposed in our framework, is key to improving the texture synthesis process.

2.1.3 Shape from texture. Shape from texture, a subfield of computer vision and image processing, focuses on deriving 3D shape information from 2D images or textures. The goal is to extract depth, orientation, and other geometric properties from the arrangement of texture elements, allowing for the reconstruction of planar textures. Representative work by [Verbin and Zickler 2020] formulates the problem as a three-player game to convert an input image into a 2.5D shape and a planar texture. However, while capable of estimating depth and creating planar textures, shape-from-texture methods struggle with structured textures and require significant computational time per input image.

In summary, despite the advancements in texture synthesis, synthesizing textures from natural images is still challenging due to occlusions and deformations. Our framework addresses these issues by rectifying these elements in sample textures, ultimately enhancing the performance and applicability of texture synthesis from natural images.

2.2 Generative Models for Image-to-image Translation

We tackle the rectification of occlusions and deformations as an image-to-image translation problem, a process that converts an input image from one domain to a corresponding image in another, while preserving crucial structural and contextual details. In our task, we consider converting a degraded texture sample into a planar texture, maintaining the overall texture appearance and structure. Therefore, we delve into several recent generative models for the image-to-image translation problem.

2.2.1 Generative Adversarial Networks (GANs). GANs [Goodfellow et al. 2014] consist of a generator and a discriminator that play an adversarial game to generate realistic samples from a prior distribution. GANs have been extensively employed in image-to-image translation tasks, with notable examples being [Isola et al. 2017], which uses a conditional GAN [Mehdi Mirza 2014] to learn a mapping between input and output images, and [Zhu et al. 2017], which extends this concept to unpaired image translation. Several existing approaches [Liu et al. 2020; Mardani et al. 2020; Zhou et al. 2018] in exemplar-based texture synthesis have also adopted GANs as a basis to generate textures. We find that GANs, due to their susceptibility to the mode-collapse problem and challenges in capturing complex data distributions, often produce unnatural results in difficult tasks like ours.

2.2.2 Diffusion probabilistic models. Recently, diffusion probabilistic models (DM) [Sohl-Dickstein et al. 2015] have taken the lead in the image synthesis field in terms of both sample quality and diversity. [Ho et al. 2020] presented Denoising Diffusion Probabilistic Model (DDPM) for high-quality image synthesis and achieved sample quality comparable to GANs. [Song et al. 2021a,b] exploited advances in score-based generative modeling for accurate score estimation and efficient sample generation. A seminal work [Dhariwal and Nichol 2021] showcased that DM can attain superior image sample quality compared to GANs. With the advent of classifier-free guidance [Ho and Salimans 2021], the necessity of an external

classifier in the generation process of conditional DM was eliminated.

Various applications using DM have since emerged. For instance, [Saharia et al. 2023] utilized DMs for conditional image generation, achieving superior performance in various super-resolution tasks and producing more realistic outputs than GAN-based methods. Notably, [Saharia et al. 2022] introduced a unified framework for image-to-image translation using conditional DM, paving the way for DM in image-to-image translation tasks. However, the use of DM has been limited by its extensive computational resource demands during both training and sampling. This not only impedes progress in the field but also constrains downstream applications. To mitigate this limitation, [Rombach et al. 2022] proposed latent diffusion models (LDM) to reduce computational resources for DM while maintaining their quality and flexibility. By training DM on the latent representation of a pre-trained vector-quantized variational autoencoder (VQ-VAE) [van den Oord et al. 2017], LDM achieves competitive results in various tasks with reduced computational costs. Our framework builds upon the LDM for further downstream applications such as integration into existing photo editing tools on personal computers.

3 APPROACH

In this section, we introduce our proposed framework for rectifying occlusions and deformations in degraded sample textures. Our framework is based on a Latent Diffusion Model [Rombach et al. 2022]. We enable conditional generation by concatenating a latent code of the degraded sample with random noise, while also incorporating features from an occlusion-aware latent transformer using cross-attention layers. An overview of the framework is depicted in Fig. 2.

3.1 Preliminary: Latent Diffusion Models (LDM)

Our framework for rectifying deformations and occlusions is built on the LDM [Rombach et al. 2022]. This allows for integration with existing photo editing tools due to its lower memory consumption compared to pixel-based DMs. The LDM utilizes a VQ-VAE [van den Oord et al. 2017] encoder $\mathcal{E}_{vq}$ to encode a planar texture $\mathbf{P} \in \mathbb{R}^{C\times H\times W}$ into a latent code $z_0 \in \mathbb{R}^{c\times h\times w}$. During training, the forward diffusion process $FwdDiff$ incrementally introduces Gaussian noise to the latent code z_0 at timestep $t \sim \mathcal{U}(1,T)$, and the reverse denoising process subsequently denoises a corrupted latent code z_t at timestep t using a trainable denoising network ϵ_θ. After training, we acquire a trained denoising network ϵ_θ that predicts the noise added at timestep t given z_t. With the trained denoising network, we can synthesize a planar texture from scratch by iteratively denoising on a Gaussian noise $z_T \sim \mathcal{N}(0,\mathbf{I})$.

3.2 Conditional Generation

In addition to the above unconditional generation, we employ conditioning mechanisms to constrain the generated textures on degraded samples. Following the concept of classifier-free diffusion guidance [Ho and Salimans 2021], we train our model by modeling the conditional distributions $p(\mathbf{P} \mid \mathbf{D})$, where $\mathbf{P}$ and $\mathbf{D}$ are planar textures and degraded textures, respectively. More specifically, we incorporate concatenation and cross-attention conditioning mechanisms into our framework to ensure the textures generated correspond to the degraded textures.

In the concatenation mechanism, the diffused latent code $\tilde{z}_t$ is paired with the latent code z_{vq-d} of a degraded sample. Both latent codes, encoded using the same VQ-VAE encoder, are concatenated along the channel dimension. Although this concatenation ensures that the identity of the degraded samples is maintained, the latent code z_{vq-d} unintentionally entangles valid and invalid features coming from occluded and valid regions in the degraded sample. This can mislead the reverse denoising process and produce unnatural results.

We address the entanglement of the latent code z_{vq-d} by introducing an occlusion-aware latent transformer τ_θ. This transformer is trained from scratch to offer valid guidance during the generation process. We integrate this guidance into the generation process using cross-attention layers. Let $z_{lt-d} \in \mathbb{R}^{C_{lt}\times d_{lt}}$ be compensatory feature obtained by the occlusion-aware latent transformer τ_θ, and $\varphi_i(\tilde{z}_t) \in \mathbb{R}^{C^i_{inter}\times d^i_{inter}}$ be flattened intermediate features before the i-th cross-attention layer. Conditional generation with cross-attention layers is implemented as:

$$Q = W_Q^{(i)} \cdot \varphi_i(\tilde{z}_t), \quad K = W_K^{(i)} \cdot z_{lt-d}, \quad V = W_V^{(i)} \cdot z_{lt-d},$$
$$Attention(Q,K,V) = softmax(\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V), \tag{1}$$

where $W_Q^{(i)}$, $W_K^{(i)}$, and $W_V^{(i)}$ are learnable encoding functions.

The final training objective used for training the conditional LDM with pairs of planar textures and degraded samples $\{(\mathbf{P_i}, \mathbf{D_i})\}_{i=1}^K$ can be formally described as follows:

$$z_{vq-d} = \mathcal{E}_{vq}(\mathbf{D}), \quad z_{lt-d} = \tau_\theta(\mathbf{D}), \quad \tilde{z}_t = FwdDiff(\mathcal{E}_{vq}(\mathbf{P})),$$
$$L_{LDM} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t\sim\mathcal{U}(1,T)}\left[\left\|\epsilon - \epsilon_\theta\left(\tilde{z}_t, t, z_{vq-d}, z_{lt-d}\right)\right\|^2\right], \tag{2}$$

where $FwdDiff$ refers to the forward diffusion process that adds noise according to the scheduler. Note that the $FwdDiff$ is only used during training and is replaced with a random Gaussian noise $z_T \sim \mathcal{N}(0,\mathbf{I})$ during inference. With these conditioning mechanisms and conditional training objectives, our framework is able to map the degraded textures to planar textures.

3.3 Occlusion-aware Latent Transformer

Since the entangled features z_{vq-d}, mislead the generation process, we propose to use a novel occlusion-aware latent transformer to compensate for the entangled features. This latent transformer takes as input a degraded texture and outputs valid guidance to the generation process while capturing long-range dependencies. We achieve these capacities with two key components: partial convolutional layers for occlusion elimination and self-attention block for modeling long-range relationships. Table 1 provides full details of our occlusion-aware latent transformer architecture.

3.3.1 Occlusion Elimination. Distilling valid features from degraded samples is important as these samples often carry invalid information stemming from occlusions and geometric deformations. Building on the concept introduced in [Liu et al. 2018], we employ partial convolutional layers for the extraction of valid features from these samples, effectively mitigating the effects of occlusions. In

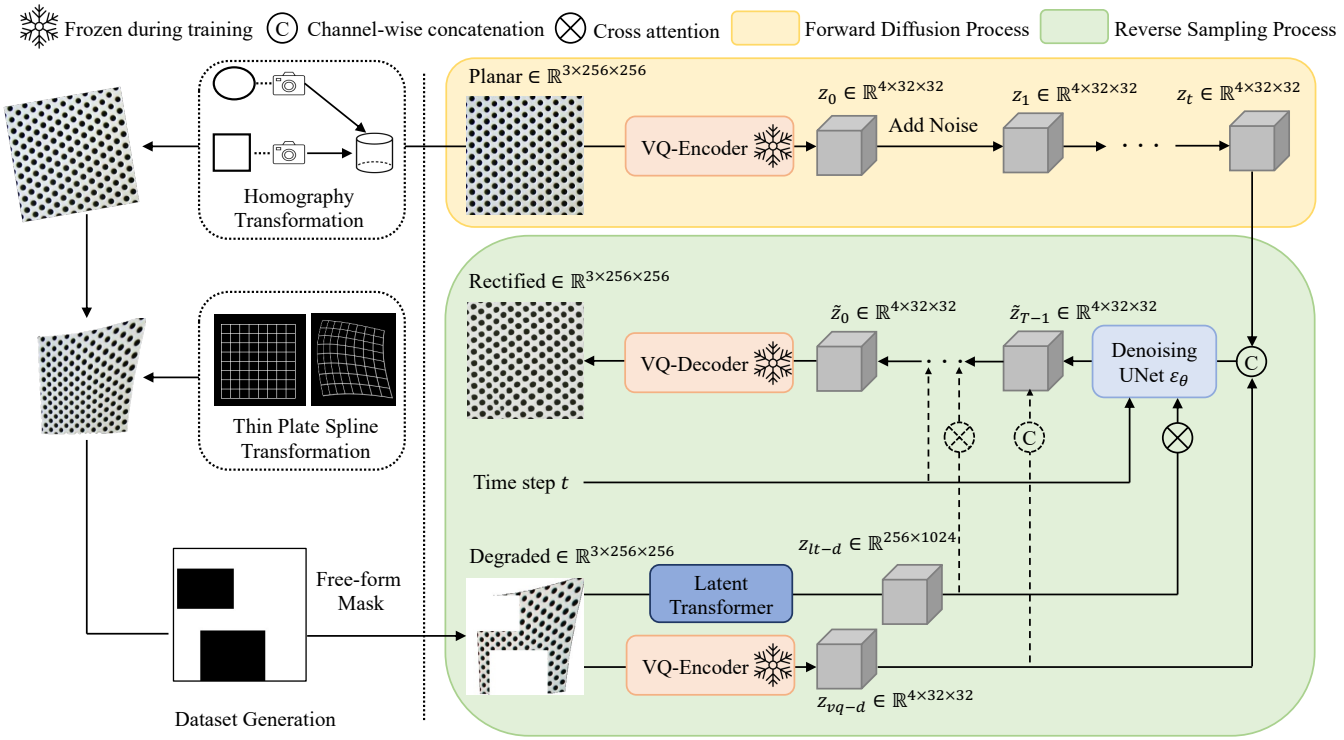

Figure 2: An overview of the proposed framework. Our synthetic training dataset is constructed by applying random geometric transformations and free-form masks on planar textures. During the training phase, our framework takes as input both degraded and planar textures, and performs forward diffusion and reverse sampling processes. Upon completion of the training, our approach takes as input a degraded texture sample and outputs a rectified texture.

Table 1: Architecture of the occlusion-aware latent transformer. The input layer takes as input a concatenation of a sample texture and its corresponding mask. Each *PartialConv* layer consists of a sequence: a partial convolution layer, followed by a Batch Norm layer, and then a *ReLU* layer. At the end of the latent transformer, the output feature is flattened to a size of 256×1024.

Layer Type	Kernel	Strides	Output Resolution
Input&Mask	-	-	$6 \times 256 \times 256$
PartialConv	3×3	2×2	$64 \times 128 \times 128$
PartialConv	3×3	1×1	$128 \times 128 \times 128$
PartialConv	3×3	2×2	$128 \times 64 \times 64$
PartialConv	3×3	1×1	$256 \times 64 \times 64$
PartialConv	3×3	2×2	$256 \times 32 \times 32$
PartialConv	3×3	1×1	$512 \times 32 \times 32$
PartialConv	3×3	1×1	$512 \times 32 \times 32$
Self-attention	3×3	1×1	$512 \times 32 \times 32$
PartialConv	3×3	1×1	$256 \times 32 \times 32$
Flatten layer	3×3	1×1	256×1024

detail, given input features x and a corresponding mask m, in which 0 and 1 indicate invalid and valid regions respectively, the partial convolutional layer is defined as:

$$x' = \begin{cases} \mathbf{W}^T (x \odot m) \frac{\text{sum}(1)}{\text{sum}(m)} + b, & \text{if sum}(m) > 0 \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where $\odot$ denotes Hadamard product, while $\mathbf{W}$ and b represent the convolution filters and the corresponding bias, respectively. The input feature x can either be degraded texture or any intermediate feature. After each partial convolutional layer, the current mask m is updated with the following definition:

$$m' = \begin{cases} 1, & \text{if sum}(m) > 0 \\ 0, & \text{otherwise} \end{cases}. \tag{4}$$

We apply the partial convolutional layer sequentially eight times, where the downsampling operation is performed three times. By repeatedly applying the layer with downsampling operations, we eventually obtain a valid feature in latent representation. This latent representation subsequently offers valid guidance to the texture rectification process.

3.3.2 Modeling Long-Range Dependencies. While the partial convolutional layers are proficient in addressing occlusions, they fall short in modeling long-range dependencies. Modeling long-range dependencies is crucial for rectifying degraded textures, especially since valid information in these textures is often sparse due to occlusions. To address this, we incorporate a self-attention layer [Zhang et al.

2019] at the end of the latent transformer. This allows for the calculation of non-local relationships from sparse information, thereby capturing long-range contextual information. This self-attention layer, which can be construed as a variant of the cross-attention layer (Eq. 1) with a single input feature, can then generate an output feature that guides the texture rectification process through the subsequent cross-attention layers.

Overall, our proposed occlusion-aware latent transformer addresses occlusions using partial convolution layers and captures long-range dependencies through the self-attention layer. The resulting valid guidance is then integrated into the texture rectification process via the cross-attention conditioning mechanism, leading to enhanced performance in the texture rectification and synthesis task.

4 DATASET

We generate synthetic training data by applying homography transformation [Hartley and Zisserman 2003], thin plate spline transformation [Bookstein 1989], and free-form mask [Yu et al. 2019] to planar textures, simulating perspective variations, geometric deformations, and occlusions. We first collect texture images from multiple sources [Abdelmounaime and Dong-Chen 2013; Bell et al. 2013; Burghouts and Geusebroek 2009; Cimpoi et al. 2014; Dai et al. 2014; Kwitt and Meerwald 2008; Mallikarjuna et al. 2006; Picard et al. 2010; Sharan et al. 2014] and manually filter out images that already exhibit degradations. After filtering, we obtain a collection of 22,043 planar texture images. And then, we perform homography transformation and thin plate spline (TPS) transformation on these planar texture images to simulate perspective variations and geometric distortions. Subsequently, free-form masks are applied to mimic occlusions. A visual illustration of the synthetic data generation is shown in Fig. 2.

We incorporate randomness into the generation process by varying the scale of transformations. Specifically, homography transformation is applied with a distortion scale $s_{hmg} \sim \mathcal{U}(0.3, 0.5)$ and a probability of 80%, while the TPS transformation is employed with a distortion scale $s_{tps} \sim \mathcal{U}(0.1, 0.3)$ and a probability of 80%. These transformations are implemented using the Kornia library [Riba et al. 2020]. This random generation process produces a diverse set of degraded texture images from a finite pool of planar textures, which is crucial for learning holistic texture rectification and synthesis in an end-to-end manner. The synthetic dataset is split into a training set with 15,430 images, a validation set with 2,205 images, and a test set with 4,408 images.

5 EXPERIMENTAL RESULTS

In this section, we present a comprehensive evaluation of our texture rectification framework.

5.1 Implementation Details

Our framework is trained for one million iterations on the proposed dataset with a batch size of 32. This takes approximately 4 days on eight A100 GPUs. The sampling process is performed on a single RTX 3090 GPU. The input patch used during the training phase is first resized to 294×294 pixels and then randomly cropped 256×256 pixels from the resized one. We employ the Adam optimizer [Kingma and Ba 2015] with a learning rate of 1e-6. The diffusion process operates with a linear noise schedule, ranging from 0.0015 to 0.0195, which is distributed over 1000 time steps. For sampling, we utilize 200 steps of the Denoising Diffusion Implicit Model (DDIM) strategy [Song et al. 2021a], which requires 4 seconds to generate a 256×256 texture.

5.2 Evaluation Metrics

Following the common metrics used in the field of texture synthesis [Li et al. 2022b; Liu et al. 2020], we employ a set of evaluation metrics that captures various aspects of texture images to assess the occlusion elimination and geometric rectification capabilities of our framework. Specifically, we use the following metrics to assess the content preservation, reconstruction quality, style consistency, and distribution match between generated and real planar textures:

- **Structural Similarity Index Measure (SSIM):** The SSIM [Wang et al. 2004] measures the preservation of structural information in the rectified textures with a larger value indicating higher similarity.
- **Learned Perceptual Image Patch Similarity (LPIPS):** The LPIPS [Zhang et al. 2018] quantifies perceptual differences between images with a lower score indicating higher similarity.
- **Gram Matrix Distance (GMD):** We use the GMD [Johnson et al. 2016] to evaluate the style consistency with a lower score indicating closer matching in texture style.
- **Fréchet Inception Distance (FID):** We employ the FID [Heusel et al. 2017] to measure the statistical similarity between distributions of images with a lower score indicating higher similarity.

5.3 Baselines

Our task inherently relates to the problems of image-to-image translation as we consider converting degraded texture to planar texture. Among these problems, image inpainting is closely related to our task as it also handles occlusions. We compare our approach against several representative methods recognized for their performance in these areas to provide a comprehensive evaluation of our approach. These comparison baselines include pix2pix [Isola et al. 2017] and VQGAN [Esser et al. 2021], well-known for image-to-image translation methods, and a leading method in image inpainting. All approaches are trained on the same dataset as our approach. The implementation details of the comparisons can be found in the supplemental.

- **pix2pix:** A widely-adopted Generative Adversarial Network-based image-to-image translation framework [Isola et al. 2017].
- **VQGAN:** The Vector Quantized Generative Adversarial Network (VQGAN) [Esser et al. 2021] represents the state-of-the-art for diverse image-to-image translation tasks.
- **MAT:** Given the inpainting aspect of our task, we draw a comparison with a leading transformer-based image inpainting method [Li et al. 2022a].

5.4 Quantitative Evaluation

We assess the performance of our method against baselines pix2pix, VQGAN, and MAT using the LPIPS, SSIM, GMD, and FID metrics. As demonstrated in Table 2, our method consistently outperforms

Table 2: Quantitative results. Comparative analysis of our method against other texture generation models, considering different metrics: SSIM, LPIPS, GMD, and FID. The table presents results that highlight the superiority of our method in terms of these metrics. For a fair comparison, all methods were trained on the synthetic training dataset and evaluated on the synthetic test dataset.

Method	SSIM ↑	LPIPS ↓	GMD ↓	FID ↓
pix2pix	0.0141	0.7742	39.29	607.15
MAT	0.2466	0.6751	34.17	187.40
VQGAN	0.4549	0.4407	24.65	45.21
Ours	**0.5096**	**0.3417**	**15.32**	**15.50**

the others. Key aspects contributing to these quantitative results include:

Occlusion and Deformation Handling: Our framework effectively addresses occlusions and deformations, as indicated by the low LPIPS and high SSIM scores. This suggests our method generates textures that are perceptually and structurally more similar to the planar textures compared to the other methods.

Texture Preservation: The lower GMD of our method signifies a higher degree of texture feature preservation, demonstrating the capability of our framework in effectively extracting valid features from degraded textures.

Quality of Generated Images: The FID scores suggest our synthesized textures match the statistical properties of planar texture more closely than other methods, indicating the effectiveness in accurately learning the distributions of planar textures.

In summary, the superior performance of our framework in these quantitative evaluations underscores its effectiveness in holistic texture rectification and synthesis. The qualitative evaluations in the next section provide further visual evidence to support these findings.

5.5 Qualitative Evaluation

We provide visual results of our method alongside the outputs of VQGAN and MAT baselines. As shown in Figure 3, our method consistently generates visually superior results, effectively handling occlusions and distortions, reconstructing detailed texture information, and producing results perceptually closer to the planar textures. We also validate the effectiveness of our method on real-world textures, feeding selected regions from real images to our framework and the baseline methods. As evident in Figure 5, our method successfully preserves the texture and overall structure of the selected regions, delivering visually pleasing results that remain perceptually closer to the original textures.

In contrast to our framework that synthesizes realistic results, other methods often yield unnatural textures. Although MAT can produce texture-like images, it succumbs to the mode-collapse problem, resulting in outputs that disregard the degraded textures. VQGAN, despite effectively capturing data distributions, generates incorrect results relative to the degraded textures. We exclude the results of pix2pix in visual comparison as it persistently produces all-black images.

Table 3: Results of the perceptual user study. The table presents the percentage of times each method was preferred over the others for generating more realistic textures, as determined by human evaluators. A higher percentage indicates a method was often favored due to its superior quality in texture rectification.

	MAT	VQGAN	Ours
vs. MAT	-	85.40%	**90.56%**
vs. VQGAN	14.60%	-	**75.32%**
vs. Ours	9.44%	24.68%	-

5.6 User Study

To further validate the perceptual quality of our synthesized results, we conduct a user study with synthetic data and real images. The goal is to get a human perspective on the effectiveness of our method compared to baselines. We first use synthetic test images for the study, considering the potential complexity of our task for laypersons. We randomly select 50 images from the test set for participants, showing them a degraded texture and a ground truth image. Participants are asked to select the more realistic image from two randomly generated results by different methods. After ensuring participants understand our task through this initial study with synthetic data, we conduct a user study with real images. We prepare 63 real images, manually selected a desired texture region for each, and generate synthesized textures using each of the three methods. We randomly pick 50 images from these 63 images and ask participants to select the more realistic one from two randomly chosen results. 14 laypersons participate in the study, and each of them provides 50 sets of feedback on synthetic test images and 50 sets on real images. As shown in Table 3, our method is preferred 90.56% of the time compared to MAT and 75.32% compared to VQGAN. This user study underscores the perceptual superiority of our framework in holistic texture rectification and synthesis, as it is consistently favored over the baselines.

5.7 Ablation Study

In order to further evaluate the effectiveness of various components of our method, we conduct an ablation study that the occlusion-aware latent transformer and conditioning mechanisms.

5.7.1 Conditioning Mechanism. We explore the conditioning mechanisms of our framework, comparing three different configurations: only concatenation with the latent code z_{vq-d}, only cross-attention with the compensatory feature z_{lt-d}, and our full method combining both. As Table 4 and Figure 4b show, the full method yields the best performance overall. These results validate our hypothesis that while concatenation provides overall guidance, cross-attention offers essential valid features.

5.7.2 Occlusion-Aware Transformer. We also conduct an ablation study on each component of our occlusion-aware latent transformer

Table 4: Ablation study on the occlusion-aware latent transformer and conditioning mechanism. The table presents a performance comparison of different configurations of our framework, including the influence of the self-attention layer and partial convolutional layers in the occlusion-aware latent transformer, and the impact of various conditioning mechanisms. The 'Conditions' and 'Arch.' represent the conditioning mechanism and architecture of the latent transformer. The terms 'Concatenation' and 'Crossattn' refer to the 'only concatenation' and 'only cross-attention' conditioning mechanisms, respectively. The 'PCE' represents removing the self-attention block from the latent transformer, and the 'SAE' indicates replacing the partial convolutions with standard convolutions.

Conditions	Arch.	SSIM ↑	LPIPS ↓	GMD ↓	FID ↓
Full	Full	**0.5096**	**0.3417**	**15.32**	**15.50**
Concatenation	Full	0.4842	0.3621	16.37	17.68
Crossattn	Full	0.4721	0.3857	23.47	21.77
Full	SAE	0.4865	0.3608	15.78	17.85
Full	PCE	0.4879	0.3596	17.61	16.15

to evaluate their contributions to the overall performance. In particular, we evaluate the impact of removing the self-attention layer and replacing partial convolutional layers with standard ones. Table 4 presents the results, and Fig. 4a provides visual outcomes. The results show that both the partial convolutional layer and the self-attention layer proved crucial to holistic texture rectification and synthesis. Their removal or replacement led to significant performance reduction, underscoring the importance of these components in effectively rectifying degraded textures.

6 LIMITATIONS AND DISCUSSION

Despite its effectiveness, our method has limitations, most notably the requirement for fixed-size degraded textures, limiting the flexibility and usage scenarios of our approach. Future work should focus on addressing this limitation. Recent advancements [Bar-Tal et al. 2023] in generating arbitrarily large images with DM present potential solutions for varying input sizes. Such improvements could broaden the applicability of our approach, making it more versatile for tasks such as synthesizing large textures from degraded samples. This advancement is expected to substantially contribute to texture synthesis research. Additionally, our framework occasionally produces imperfect results, particularly in cases with varying lighting conditions and extreme distortions in the sample images. We can address these issues by masking regions with significant lighting changes and incorporating more training data with extreme distortions.

ACKNOWLEDGMENTS

This work was partially supported by JST SPRING (Hao, Grant Number: JPMJSP2124), JST PRESTO (Iizuka, Grant Number: JPMJPR21C1), and JSPS KAKENHI (Hara, Grant Number: JP21H04908).

REFERENCES

Safia Abdelmounaime and He Dong-Chen. 2013. New Brodatz-Based Image Databases for Grayscale Color and Multiband Texture Analysis. *Volume 2013* (2013). https://doi.org/10.1155/2013/876386

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. In *International Conference on Machine Learning*.

Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. 2013. OpenSurfaces: A Richly Annotated Catalog of Surface Appearance. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 32, 4 (2013).

Urs Bergmann, Nikolay Jetchev, and Roland Vollgraf. 2017. Learning Texture Manifolds with the Periodic Spatial GAN. In *International Conference on Machine Learning*.

Fred L Bookstein. 1989. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11, 6 (1989), 567–585.

Gertjan J. Burghouts and Jan-Mark Geusebroek. 2009. Material-specific Adaptation of Color Invariant Features. *Pattern Recognition Letters* 30 (2009), 306–313.

Chin-Fan Chen and Evan Suma Rosenberg. 2018. Virtual Content Creation Using Dynamic Omnidirectional Texture Synthesis. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing Textures in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. 2003. Object Removal by exemplar-based inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Dengxin Dai, Hayko Riemenschneider, and Luc Van Gool. 2014. The Synthesizability of Texture Examples. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*.

Alexei A. Efros and William T. Freeman. 2001. Image Quilting for Texture Synthesis and Transfer. In *SIGGRAPH '01: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. 341–346.

Alexei A. Efros and Thomas K. Leung. 1999. Texture Synthesis by Non-parametric Sampling. In *International Conference on Computer Vision*.

Patrick Esser, Robin Rombach, and Björn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Leon Gatys, Alexander S. Ecker, and Matthias Bethge. 2015. Texture Synthesis Using Convolutional Neural Networks. In *Conference on Neural Information Processing Systems*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Conference on Neural Information Processing Systems*.

Richard Hartley and Andrew Zisserman. 2003. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Conference on Neural Information Processing Systems*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*.

Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Nets. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Naoya Isoyama, Yamato Sakuragi, Tsutomu Terada, and Masahiko Tsukamoto. 2021. Effects of Augmented Reality Object and Texture Presentation on Walking Behavior. *Electronics* 10, 6 (2021).

Nikolay Jetchev, Urs M. Bergmann, and Roland Vollgraf. 2016. Texture Synthesis with Spatial Generative Adversarial Networks. *CoRR* abs/1611.08207 (2016).

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *European Conference on Computer Vision*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Roland Kwitt and Peter Meerwald. 2008. Salzburg Texture Image Database. Online.

Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. 2022a. MAT: Mask-Aware Transformer for Large Hole Image Inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Xueting Li, Xiaolong Wang, Ming-Hsuan Yang, Alexei A. Efros, and Sifei Liu. 2022b. Scraping Textures from Natural Images for Synthesis and Editing. In *European Conference on Computer Vision*.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Diversified Texture Synthesis with Feed-forward Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryzan Catanzaro. 2018. Image Inpainting for Irregular Holes Using Partial Convolutions.

In *European Conference on Computer Vision*.

Guilin Liu, Rohan Taori, Ting-Chun Wang, Zhiding Yu, Shiqiu Liu, Fitsum A. Reda, Karan Sapra, Andrew Tao, and Bryan Catanzaro. 2020. Transposer: Universal Texture Synthesis Using Feature Maps as Transposed Convolution Filter. *CoRR* abs/2007.07243 (2020).

P. B. Mallikarjuna, Alireza Tavakoli Targhi, Mario Fritz, Eric Hayman, Barbara Caputo, and J. O. Eklundh. 2006. THE KTH-TIPS 2 database.

Morteza Mardani, Guilin Liu, Aysegul Dundar, Shiqiu Liu, Andrew Tao, and Bryan Catanzaro. 2020. Neural FFTs for Universal Texture Image Synthesis. In *Advances in Neural Information Processing Systems*.

Simon Osindero Mehdi Mirza. 2014. Conditional Generative Adversarial Nets. *CoRR* abs/1411.1784 (2014).

Rosalind Picard, Chris Graczyk, Steve Mann, Josh Wachman, Len Picard, and Lee Campbell. 2010. Vistex Vision Texture Database. Online.

E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski. 2020. Kornia: an Open Source Differentiable Computer Vision Library for PyTorch. In *WACV*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-Image Diffusion Models. In *SIGGRAPH '22: ACM SIGGRAPH Conference Proceedings*.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. 2023. Image Super-Resolution via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2023).

Lavanya Sharan, Ce Liu, Ruth Rosenholtz, and Edward H. Adelson. 2014. Accuracy and Speed of Material Categorization in Real-World Images. *Journal of Vision* 14, 9 (2014), article 12.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021a. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021b. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. In *Conference on Neural Information Processing Systems*.

Dor Verbin and Todd Zickler. 2020. Toward a Universal Model for Shape from Texture. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

Li-yi Wei, Sylvain Lefebvre, Vivek Kwatra, and Greg Turk. 2009. State of the Art in Example-based Texture Synthesis. In *Eurographics 2009 - State of the Art Reports*.

Li-Yi Wei and Marc Levoy. 2000. Fast Texture Synthesis Using Tree-Structured Vector Quantization. In *SIGGRAPH '00: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. 479–488.

Huisi Wu, Xiaomeng Lyu, and Zhenkun Wen. 2018. Automatic texture exemplar extraction based on global and local textureness measures. *Computational Visual Media* 4 (2018), 173–184.

Huisi Wu, Wei Yan, Ping Li, and Zhenkun Wen. 2021. Deep Texture Exemplar Extraction Based on Trimmed T-CNN. *IEEE Transactions on Multimedia* 23 (2021), 4502–4514.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. 2019. Free-Form Image Inpainting with Gated Convolution. In *International Conference on Computer Vision*.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-Attention Generative Adversarial Networks. In *International Conference on Machine Learning*.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Yang Zhou, Zhen Zhu, Xiang Bai, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2018. Non-Stationary Texture Synthesis by Adversarial Expansion. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 37, 4 (2018), 13 pages.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *International Conference on Computer Vision*.

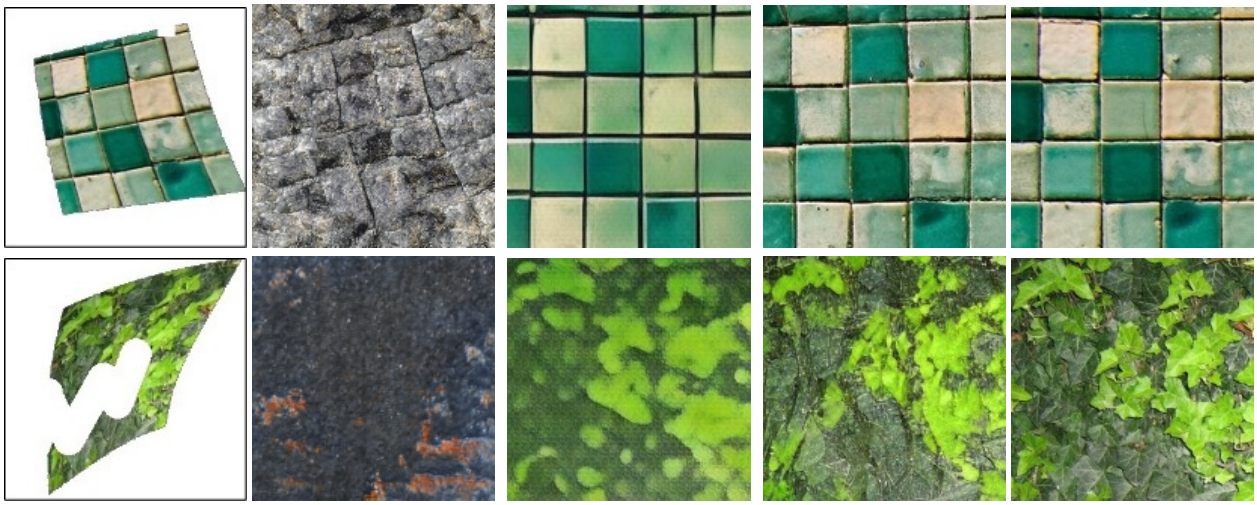

Figure 3: Rectification results on the synthetic test dataset. Our framework can generate texture images that are perceptually closer to the ground truth than other methods. MAT generates textures that are not related to the input as it falls into the mode-collapse problem. Although the VQGAN can rectify degraded textures, it loses details of the input texture. Texture images from [Dai et al. 2014].

(a) Ablation results on the architecture of the occlusion-aware latent transformer.

(b) Ablation results on the architecture of the conditioning mechanisms.

Figure 4: Visual results of the ablation study. Fig. 4b shows ablation results relating to the architecture of the conditioning mechanisms, while Fig. 4a presents ablation results on the architecture of the occlusion-aware latent transformer. It is readily apparent that our full method generates more realistic textures than the other methods. Texture images from [Cimpoi et al. 2014].

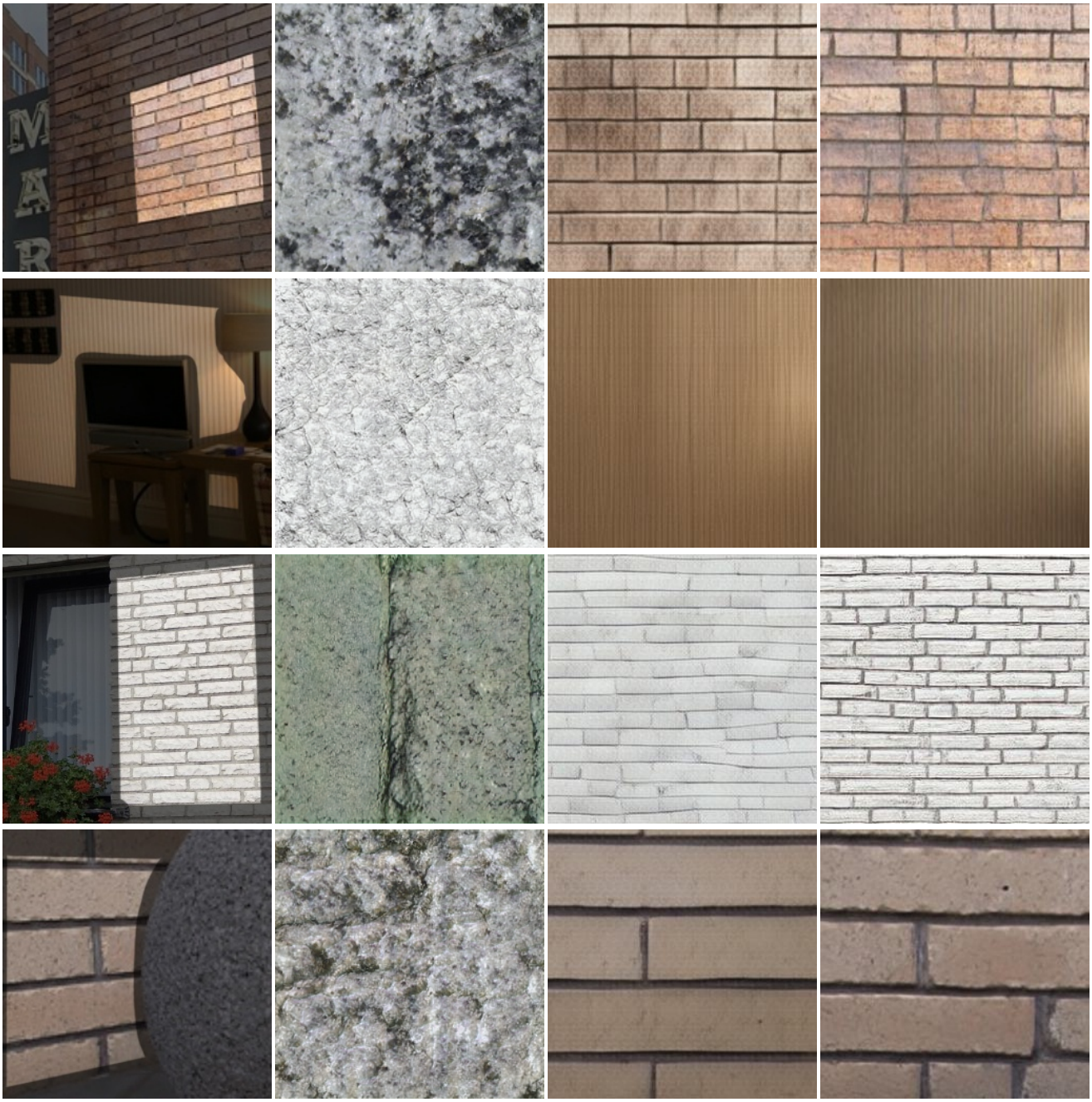

Figure 5: Rectified and synthesized texture results on real images. Our framework can generate texture images from real images. One can easily get a planar texture image by selecting desired regions with brush tools. The input to our framework is highlighted. Compared to other methods, our framework preserves the original appearance and synthesizes holistic texture images. Photographs courtesy of TheTurducken (CC-BY), Alan Light (CC-BY), Andrea Dufrenne (CC-BY), and Toshiyuki IMAI (CC-BY).