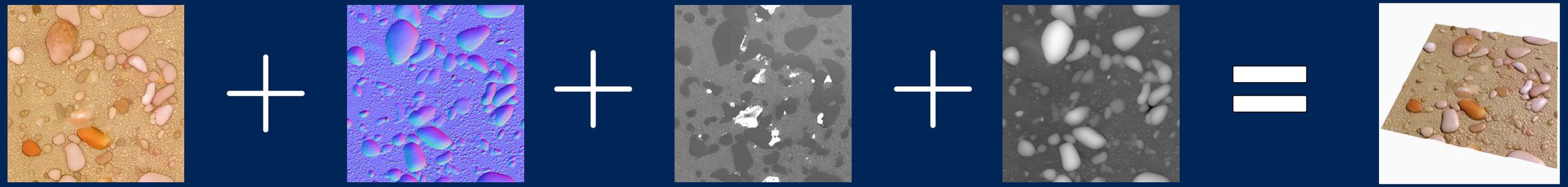
Controlled Texture Map Generation

Siddharth Narsipur
Suchith Hegde

University of Rochester

Background

- Physically Based Rendering (PBR) is a modern approach to shading and rendering images. PBR workflows use different texture maps that encode stylization and detail.



Basecolor (Albedo) | Normal | Roughness | Height | Final Render

All images from ControlMat supplementary material by Vecchio et al. (2023)

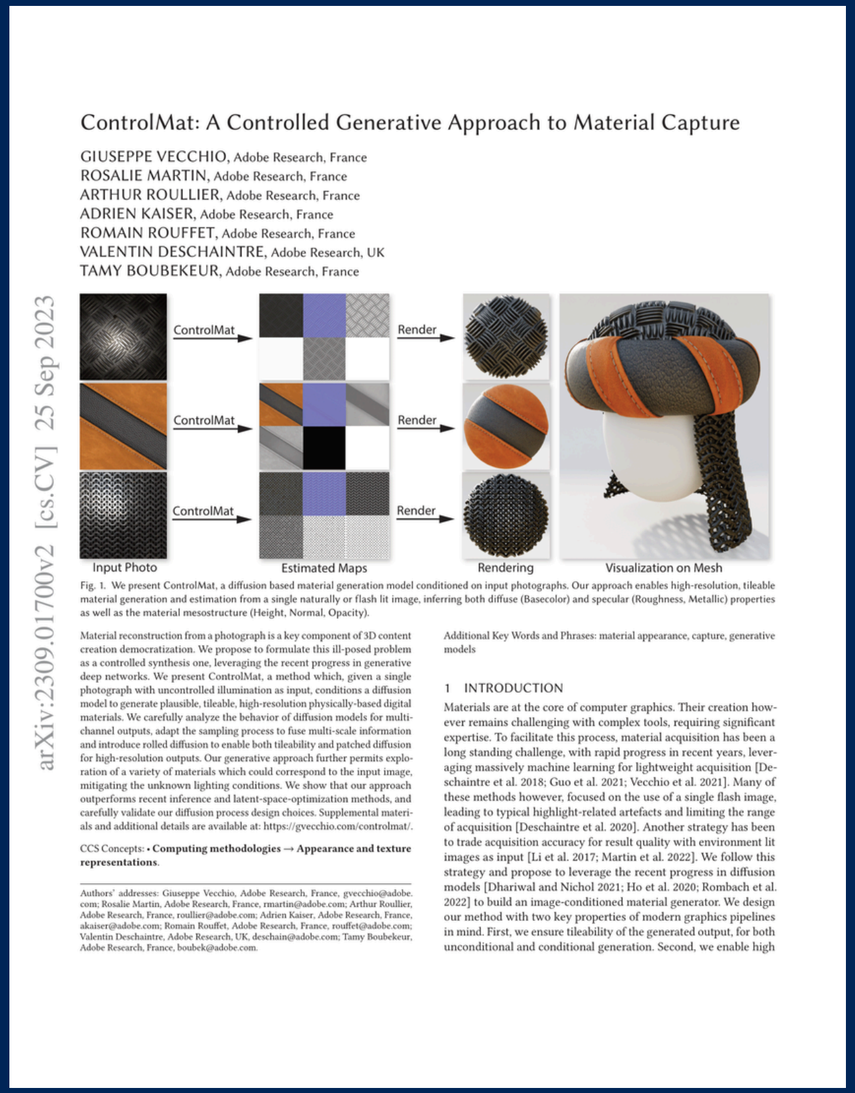
Background

- These materials are at the core of compute graphics and are used in AR/VR, video games, film/tv, 3D design, architecture etc.

- However, their creation remains challenging and requires complex expertise.

ControlMat

- Released in 2023, current state of the art
- Adobe Research
- Custom Diffusion Model + ControlNet
- Closed-source dataset



ControlMat: A Controlled Generative Approach to Material Capture

GIUSEPPE VECCHIO, Adobe Research, France
ROSALIE MARTIN, Adobe Research, France
ARTHUR ROULLIER, Adobe Research, France
ADRIEN KAISER, Adobe Research, France
ROMAIN ROUFFET, Adobe Research, France
VALENTIN DESCHAINTRE, Adobe Research, UK
TAMY BOUBEKEUR, Adobe Research, France

Fig. 1. We present ControlMat, a diffusion based material generation model conditioned on input photographs. Our approach enables high-resolution, tileable material generation and estimation from a single naturally or flash lit image, inferring both diffuse (Basecolor) and specular (Roughness, Metallic) properties as well as the material mesostructure (Height, Normal, Opacity).

Material reconstruction from a photograph is a key component of 3D content creation democratization. We propose to formulate this ill-posed problem as a controlled synthesis one, leveraging the recent progress in generative deep networks. We present ControlMat, a method which, given a single photograph with uncontrolled illumination as input, conditions a diffusion model to generate plausible, tileable, high-resolution physically-based digital materials. We carefully analyze the behavior of diffusion models for multi-channel outputs, adapt the sampling process to fuse multi-scale information and introduce rolled diffusion to enable both tileability and patched diffusion for high-resolution outputs. Our generative approach further permits exploration of a variety of materials which could correspond to the input image, mitigating the unknown lighting conditions. We show that our approach outperforms recent inference and latent-space-optimization methods, and carefully validate our diffusion process design choices. Supplemental materials and additional details are available at: https://gvecchio.com/controlmat/.

CCS Concepts: • **Computing methodologies** → **Appearance and texture representations**.

Authors' addresses: Giuseppe Vecchio, Adobe Research, France, gvecchio@adobe.com; Rosalie Martin, Adobe Research, France, rmartin@adobe.com; Arthur Roullier, Adobe Research, France, roullier@adobe.com; Adrien Kaiser, Adobe Research, France, akaiser@adobe.com; Romain Rouffet, Adobe Research, France, rouffet@adobe.com; Valentin Deschaintre, Adobe Research, UK, deschain@adobe.com; Tamy Boubekeur, Adobe Research, France, boubek@adobe.com.

Additional Key Words and Phrases: material appearance, capture, generative models

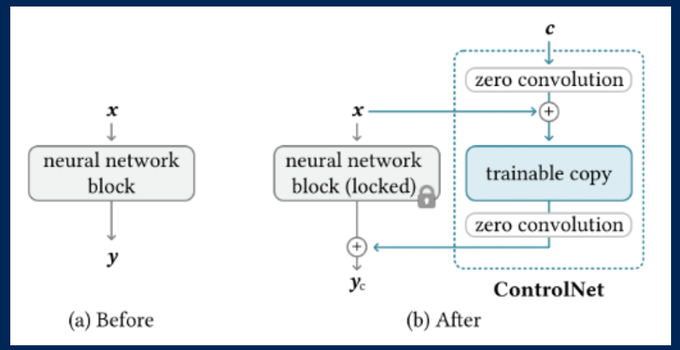
1 INTRODUCTION

Materials are at the core of computer graphics. Their creation however remains challenging with complex tools, requiring significant expertise. To facilitate this process, material acquisition has been a long standing challenge, with rapid progress in recent years, leveraging massively machine learning for lightweight acquisition [Deschaintre et al. 2018; Guo et al. 2021; Vecchio et al. 2021]. Many of these methods however, focused on the use of a single flash image, leading to typical highlight-related artefacts and limiting the range of acquisition [Deschaintre et al. 2020]. Another strategy has been to trade acquisition accuracy for result quality with environment lit images as input [Li et al. 2017; Martin et al. 2022]. We follow this strategy and propose to leverage the recent progress in diffusion models [Dhariwal and Nichol 2021; Ho et al. 2020; Rombach et al. 2022] to build an image-conditioned material generator. We design our method with two key properties of modern graphics pipelines in mind. First, we ensure tileability of the generated output, for both unconditional and conditional generation. Second, we enable high

https://doi.org/10.48550/arXiv.2309.01700

ControlNet

- Add conditional control to diffusion models by “locking” the original NN block and training a copy on new data.



Adding Conditional Control to Text-to-Image Diffusion Models

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala
Stanford University
{lvmin, anyirao, maneesh}@cs.stanford.edu

Figure 1: Controlling Stable Diffusion with learned conditions. ControlNet allows users to add conditions like Canny edges (top), human pose (bottom), *etc.*, to control the image generation of large pretrained diffusion models. The default results use the prompt “a high-quality, detailed, and professional image”. Users can optionally give prompts like the “chef in kitchen”.

Abstract

We present ControlNet, a neural network architecture to add spatial conditioning controls to large, pretrained text-to-image diffusion models. ControlNet locks the production-ready large diffusion models, and reuses their deep and robust encoding layers pretrained with billions of images as a strong backbone to learn a diverse set of conditional controls. The neural architecture is connected with “zero convolutions” (zero-initialized convolution layers) that progressively grow the parameters from zero and ensure that no harmful noise could affect the finetuning. We test various conditioning controls, e.g., edges, depth, segmentation, human pose, etc., with Stable Diffusion, using single or multiple conditions, with or without prompts. We show that the training of ControlNets is robust with small (<50k) and large (>1m) datasets. Extensive results show that ControlNet may facilitate wider applications to control image diffusion models.

1. Introduction

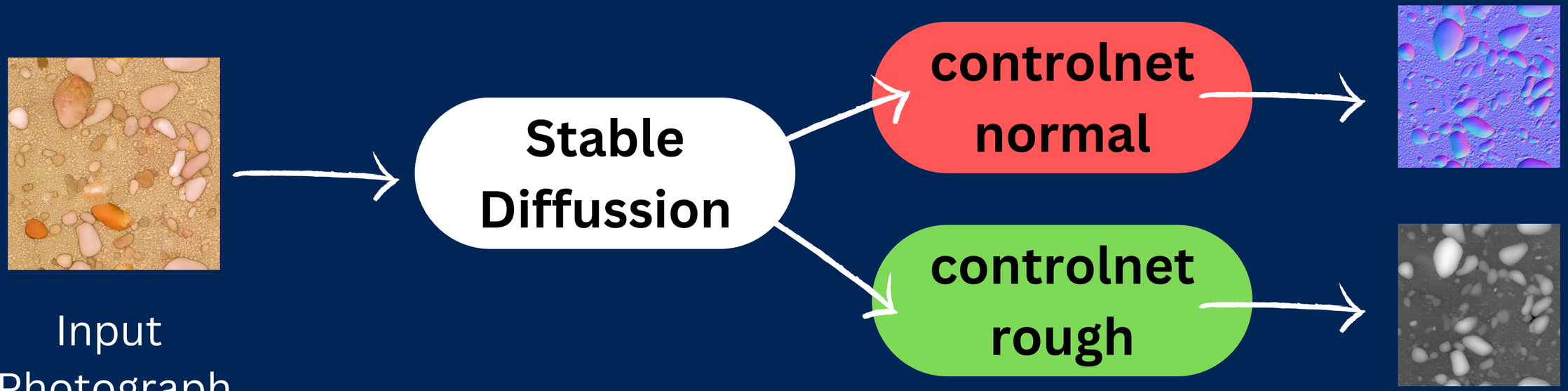
Many of us have experienced flashes of visual inspiration that we wish to capture in a unique image. With the advent of text-to-image diffusion models [54, 62, 72], we can now create visually stunning images by typing in a text prompt. Yet, text-to-image models are limited in the control they provide over the spatial composition of the image; precisely expressing complex layouts, poses, shapes and forms can be difficult via text prompts alone. Generating an image that accurately matches our mental imagery often requires numerous trial-and-error cycles of editing a prompt, inspecting the resulting images and then re-editing the prompt.

Can we enable finer grained spatial control by letting users provide additional images that directly specify their desired image composition? In computer vision and machine learning, these additional images (*e.g.*, edge maps, human pose skeletons, segmentation maps, depth, normals, *etc.*) are often treated as conditioning on the image generation process. Image-to-image translation models [34, 98] learn

https://doi.org/10.48550/arXiv.2302.05543

Motivation & Goal

- Use existing Stable Diffusion model + ControlNet to train a new model to generate roughness and normal maps, given an input photograph.
- First high quality, open-source dataset, MatSynth, released in February 2024.



Dataset

Dataset	Number of materials
MatSynth (Original Dataset)	5300
After Cleaning	3351
Final Dataset (Crop, Flip & Rotate at diferent scales)	52309

Method

Both models trained on one A100 40GB on the school’s Bluehive cluster

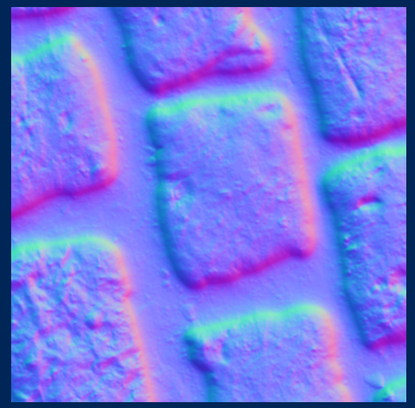
- controlnet_normal
  - learning rate: 1e-4
  - batch size: 15
  - number of epochs: 8
  - training time: 19 hours 30 minutes

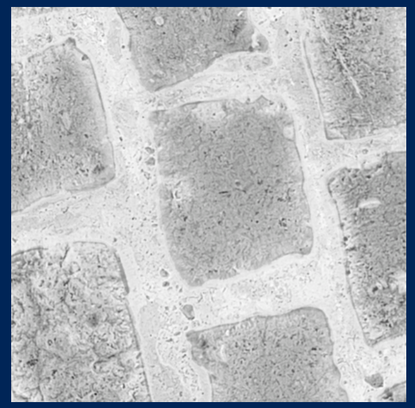
- controlnet_rough
  - learning rate: 1e-4
  - batch size: 12
  - number of epochs: 8
  - training time: 20 hours 11 minutes

Results

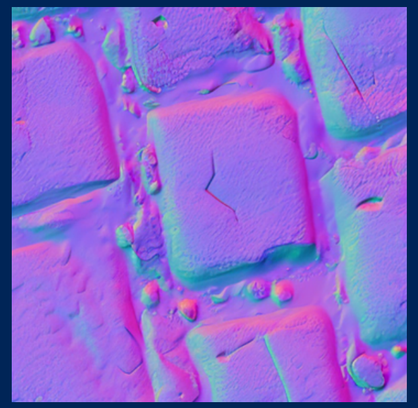
Input Photograph

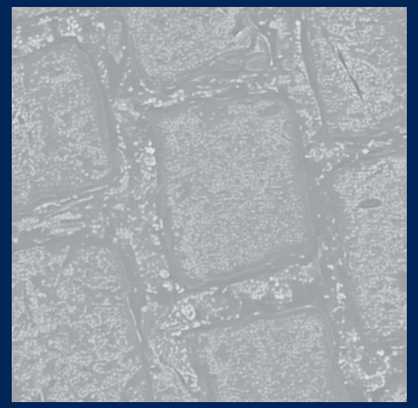
Ground Truth



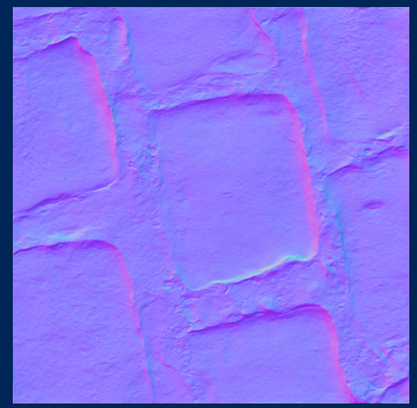


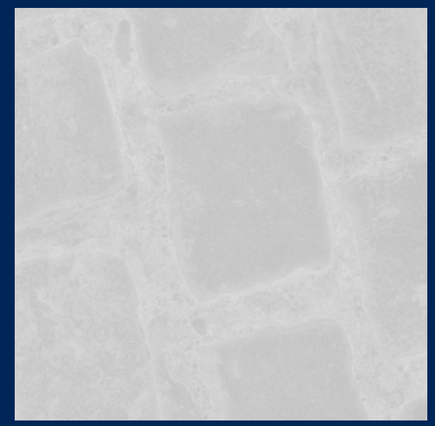
Ours





ControlMat





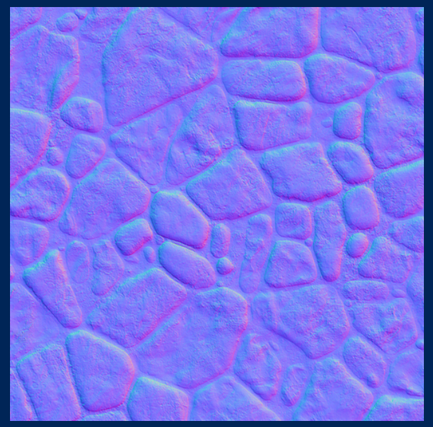
Results

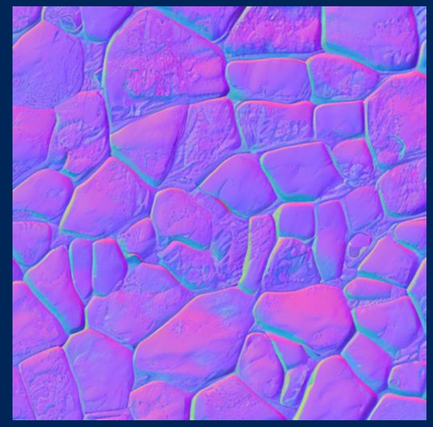
Ground Truth

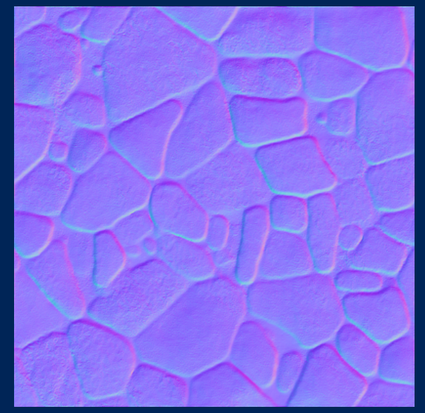
Ours

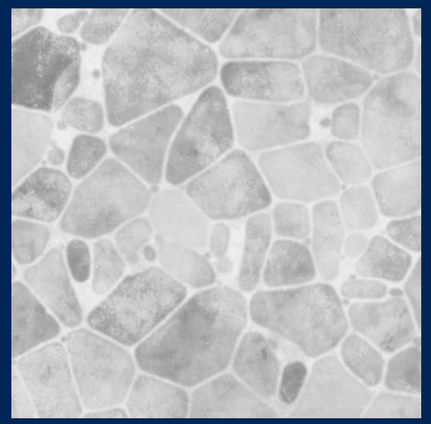
ControlMat

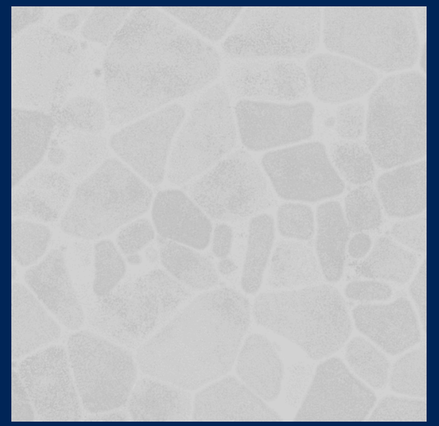
Input Photograph







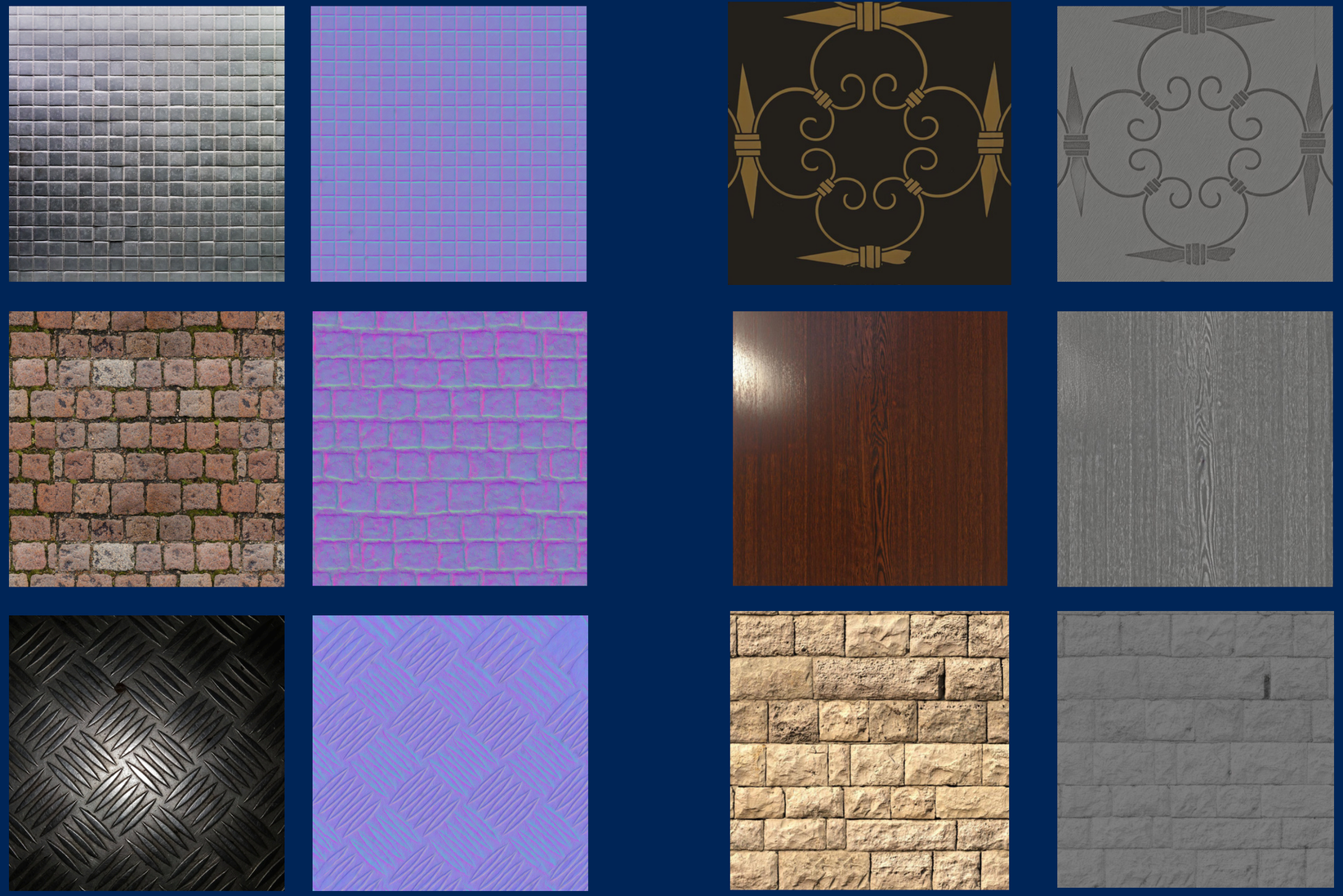




Results

controlnet_normal

controlnet_rough



Results

Evaluated on MatSyth Test Split

(Lower is better)		Error (RMS for Roughness, Cosine for Normal)
ControlNet	Normal Map	0.029
	Roughness Map	0.250
MatForger	Normal Map	0.010
	Roughness Map	0.276

Findings & Limitations

- Captures shape and patterns in textures well. Low error in outdoor categories (ground, stone, concrete).
- Adds too much z-axis “height” for normal maps.
- Struggles with textures showing weak correlation between input photo and normal/roughness maps (leather, metals).
- Longer training and larger dataset will enhance model performance.
- Custom diffusion model offers broader image compatibility and improved performance but requires more compute for training.

Thank You