DUNHUANG IMAGE RESTORATION

MODEL

In order to achieve the task of image restoration, a U-like autoencoder with **Partial Convolution** Downsampling, **Skip Connections** and **Nearest-Neighbour-Upsampling** is implemented.

After the forward pass, the predicted image is merged with the masked input image to restore unmasked pixels.

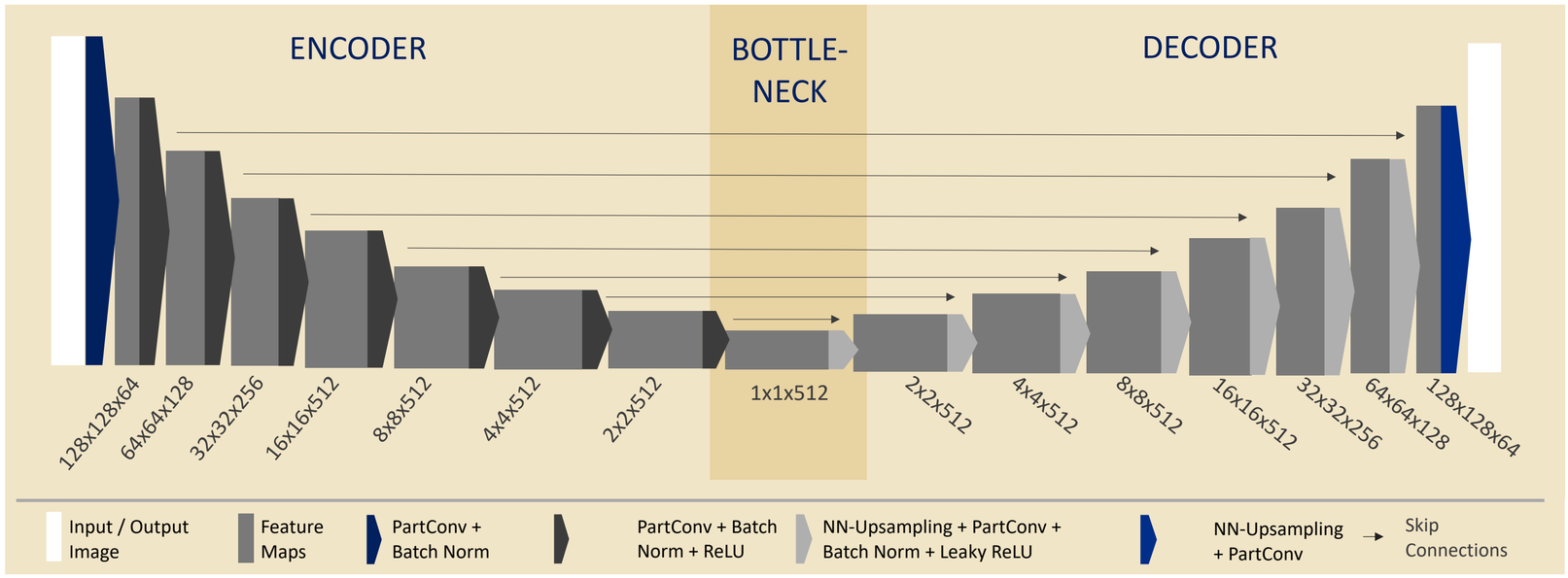

Figure 1: Overview of model architecture

DATA PREPROCESSING

Since only 400 training images of different shapes were provided, preprocessing was needed in order to increase the number of samples and to adjust the size of extracted patches to a standard of 256 x 256 pixels.

First, random patches were selected from the ground truth training images (see Figure 2). To ensure that most of the image is covered by the patches, the distance between patches was maximized.

Figure 2: 25 random patches per ground truth image

According to the patch size, new masks were created using a random walk algorithm. The results of the mask and the respective masked image are plotted in the following Figure:

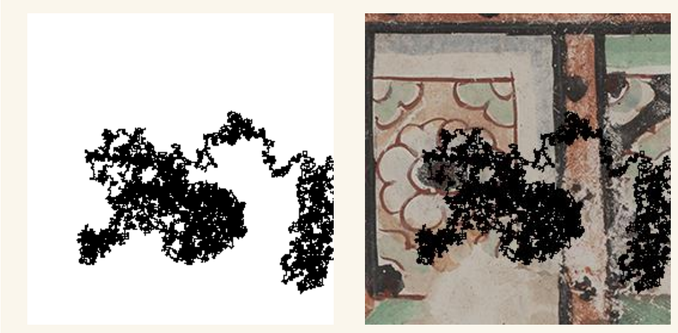

Figure 3: Mask generated with random walk (Left). Random Patch with mask (Right)

Using the method described above, the amount of 400 ground truth training images could be increased to 10.000 training patches, resulting in a data augmentation factor of 25.

Nevertheless, the patches do have notable overlaps. To counteract overfitting, further preprocessing was conducted. Masked input images were transformed by

> **Horizontal Flip**, with probability = 0.5

> **Random Rotation** between -30° and 30°, with p = 0.5

TRAINING & PERFORMANCE

> **MSE: 36.3145**

> **SSIM: 0.8016**

> Pretraining on Place2 dataset with 40.000 images and learning rate = 2e-4

> Fine-tuning on Dunhang dataset with 10.000 images (see preprocessing) and learning rate = 5e-5

METHODOLOGY

PARTIAL CONVOLUTION

In order to enhance inpainting performance, not usual Convolution, but a novel technique called Partial Convolution was used in the autoencoder which was developed by NVIDIA researchers in 2018 [1]. Besides masked images, Partial Convolution also takes corresponding binary masks as input and restricts the learnable patterns on non-deteriorated regions (conditional inverse attention). It includes a convolutional update of the feature maps:

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

where W denotes the convolution filter weights, b the bias, X the feature values, M the binary mask and $\odot$ element-wise multiplication. 1 has same shape as M but with all elements being 1. In addition, Partial Convolution includes an update of the input mask, if the Partial Convolution was able to inpaint holes based on intact regions:

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

After several Partial Convolution layers, masks will only contain ones and holes will be inpainted properly.

LOSS FUNCTIONS

Our total loss per iteration is calculated by a linear combination of 3 loss terms, covering differences in content and style as well as regularization for spatial smoothness:

$$L = \lambda_{content} L_{content} + \lambda_{style} L_{style} + \lambda_{TV} L_{TV}$$

First, ground-truth and predicted images are fed into a pretrained network to retrieve deep feature representations:

> **Content Loss $L_{content}$** measures the content difference between deep features of an image pair.

> **Style loss L_{style}** measures the difference of style features computed by Gram matrix.

> **Total Variation Loss L_{TV}** measures the spatial smoothness of the predicted regions by comparing neighboring pixels.

EXPERIMENTAL RESULTS

To start with, a simplified model with conv. layers was implemented. The model was trained with the MSE loss and reached MSE of 40.67 and SSIM of 0.75. In order to improve the model, partial conv. layers in combination with skip connections were implemented instead. Aiming to boost MSE and SSIM, different losses were tested. One attempt focused on these two evaluation criteria by combining them into a single loss which results in achieving MSE of 36.42 and SSIM of 0.80.

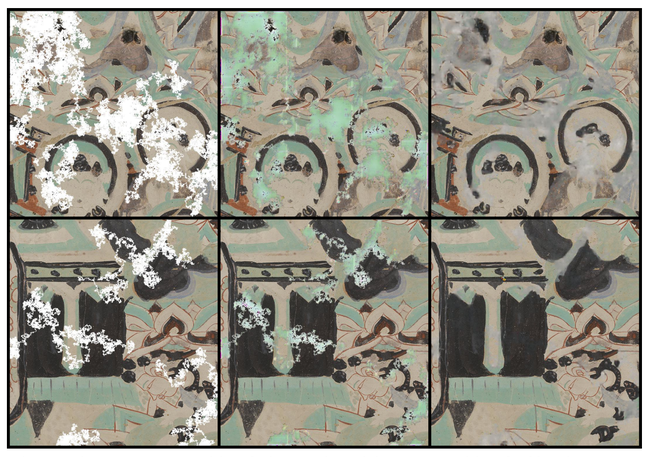

Figure 4: input, conv. net with MSE loss and partial conv. net with MSE-SSIM loss

The final model was pretrained using MSE loss and then switched to the above combination of losses in fine-tuning. Although training with the combined loss did not result in improving the baseline much, it enhances the inpainted texture as can be seen in comparing the two output figures.

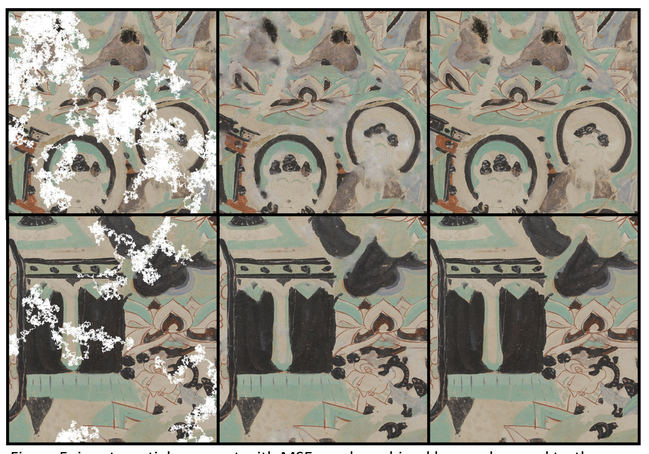

Figure 5: input, partial conv. net with MSE - and combined loss and ground truth

REFERENCES

[1] “Image Inpainting for Irregular Holes Using Partial Convolutions”, Guilin Liu et al., In proceedings of The European Conference on Computer Vision (ECCV), 2018

[2] "Dunhuang Grottoes Painting Dataset and Benchmark", Tianxiu Yu et al., arXiv:1907.04589, 2019

[3] “End-to-End Partial Convolutions Neural Networks for Dunhuang Grottoes Wall-painting Restoration”, Tianxiu Yu et al., In proceedings of The IEEE International Conference on Computer Vision (ICCV), 2019