Application of Image restoration task based on Guided UNet3+ inspired Architecture – Project Report – Group C3

Marcus Pertlwieser *Utsaha Joshi* *Alfi Xavier* *Tamás Scheidl* *Vivekanand Ramakrishnan*

Abstract

This report outlines the architecture and presents the results of our proposed Guided-Unet3+ model, which addresses the given image restoration task. The model is designed to take integrated images and seeks to produce accurate predictions. In summary, the proposed model achieves excellent performance and delivers accurate predictions. However, there are limitations to the model, most notably that it may experience difficulties in predicting targets in images with significant occlusion. Addressing this issue would require generative capabilities for the model and was therefore beyond the original intention of our project.

Introduction

The following report concisely summarizes the proposed algorithm, outlines the results and discusses the limitations of the suggested guided UNet3+ approach applied by Group C3 for the given Image Restoration task using integrated images.

Preprocessing

The first stage of our implementation is aimed at preprocessing the given input data. This step can be broadly distinguished into Filtering, AOS Integration and Train-Test splitting.

Filtering: The Filtering step is handled based on the completeness of the given input samples. If the provided input data is complete, i.e., if the 11 images, the GT (ground truth) and the parameter file are existent, the code will create a new folder for each pair (based on their “row-number”) in the specified destination folder, and consequently copy the data into the corresponding folders.

AOS Integration: Afterwards, the AOS integrator takes the pre-processed folders and generates images for the focal plane lengths of 0, -1.5 and -3m. The resulting images are subsequently merged and saved together, as one image with three channels, to the destination folder along with the associated GT and parameter files.

Splitting: Finally, the code for the for splitting the dataset is then used to load the integrated images and group the files in the source directory based on their row number. Subsequently, these groups will be randomly shuffled and partitioned into a training, evaluation and testing set, which are split into 3 different directories called “Train”, “Eval and “Test”. The splitting is performed in a ratio of 70%-15%-15% (Train, Test, Eval) and is then used to train the model.

Normalization: It should also be noted that we used simple max normalization for both the input focal stack and the ground temperature values.

Architecture

Encoder: The architecture consists of an encoder like VGG16D, however, with the exception that it does only contain the first three 512-channeled stacks at the 4th Encoder stage. For the last encoder stage Strided-Maxpooling with 3x3 Convolutions were used to subsequently generate three 1024 feature stacks. This might be overkill and would be one of the first considerations for better parameter efficiency.

Decoder: On the decoder side, as with UNet3+, each of the encoder feature stacks gets convolved to 64 channels of their corresponding resolution, to build up a skip-stack.

The first stage of the decoder (seen from bottom to up) utilizes all skip-stacks. Entries of skip-stacks from stages above are nearest-sampled down to the appropriate resolution. However, stacks of same stage are not sampled as they already have the appropriate size. Stages from below are up-sampled accordingly.

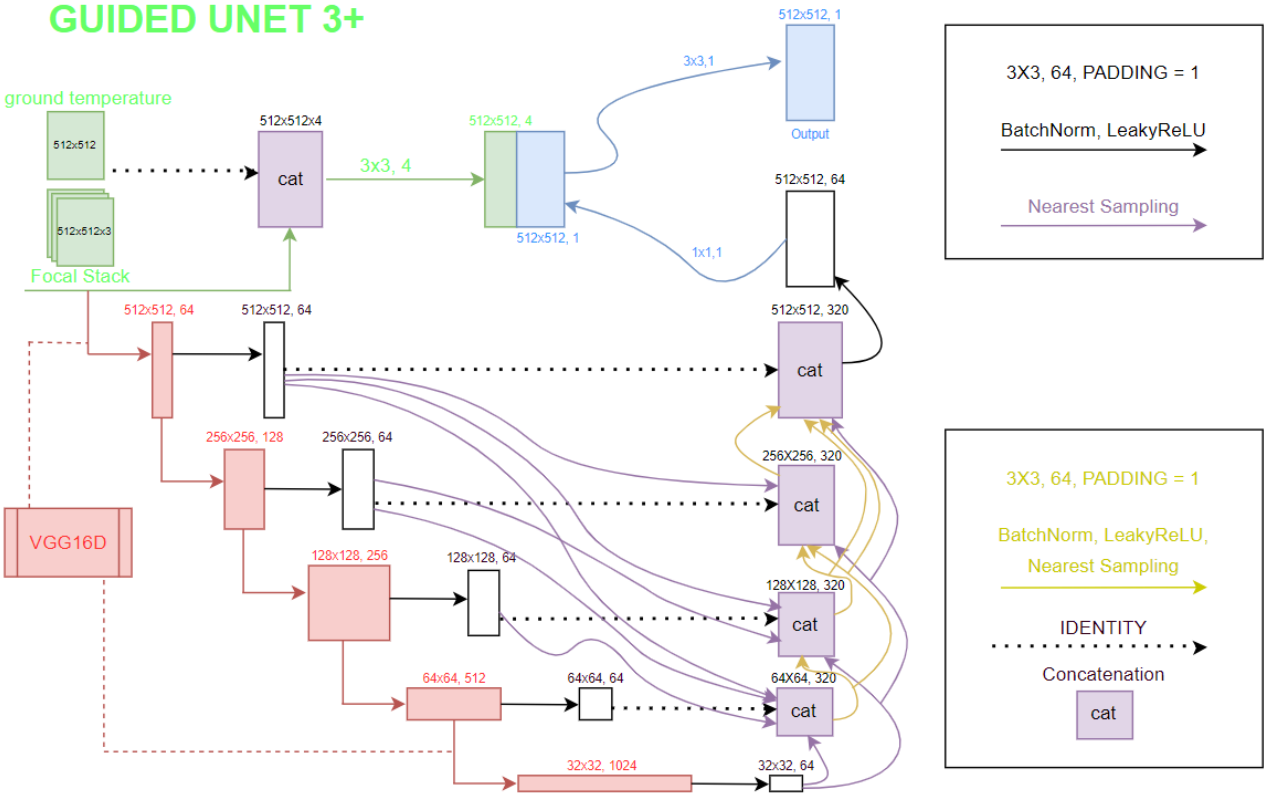
The first and subsequent decoder stages receive 320 channels as inputs. These are then again convolved to 64 channels.

After the first decoder stage, instead of reusing the same 64 channels from the encoder that were already incorporated, the 64 output channels from previous decoder stages are reused. This is intended to provide enriched feature maps as we delve deeper into the decoder.

Following the last decoder, we end up at 64 channels. These are consequently combined linearly with a 1x1 convolution to form 1 channel.

Guidance: Since ground temperature seems to be one of the key values that should pose as a good reference value for removing unnecessary background information, this was included as an additional guidance. We applied max normalisation to the ground temperature values. These values were copied to the pixel values of the same size as the original input images. We concatenated this temperature image with the original focal stack input. After a 3x3 convolution to 4 channels, we arrive at a feature map that should incorporate information that will give the model the ability to correctly assess the appropriate ground temperature values. After concatenation with the output of the UNet3+-like part, we arrive at 5 channels. Now we should have information about the correct values and position of the background and target, thus another 3x3 convolution to 1 channel should do the trick (However, there could also be the argument for a 1x1 convolution). It should be noted that we did not use any activation in the guiding module.

Figure 1: Flow-Chart depicting the Guided-UNet3+ Architecture



Training

Initial Training: Initial Training was performed using a batch size of 8, an Adam optimizer with an initial learning rate of 0.0001, an initialized VGG16 Encoder that was trained on ImageNet-1K, and MSE loss. We chose the batch size of 8 for initial learning because it ensures sufficient exploration of the loss function space. It should also be noted that since this is a regression task with a high dimensional output (512x512 in our case), that the stochastic effect can be significantly dampened by the fact that we are averaging over large spatial dimensions. The idea behind the rather low learning rate of the optimizer is to allow the encoder an adequate amount of time to adapt to the new data, with momentum increasing the effective learning rate after a few iterations. Initial training on an MSE-SSIM combination was also attempted, though this did not yield satisfactory results as the architecture tended to get stuck in local minima that focused on ignoring the target and only removing the background.

Figure 2A: Results for the Initial Training

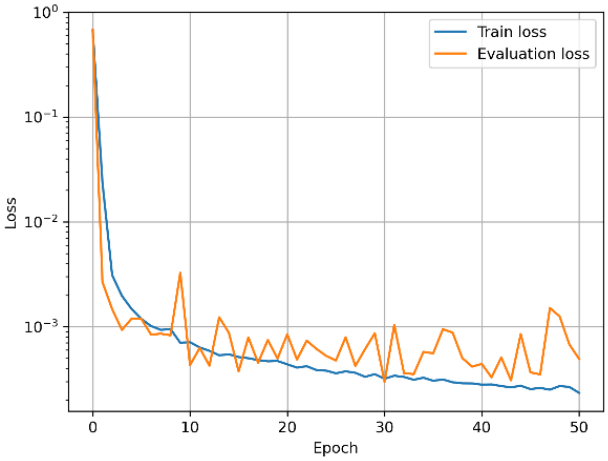
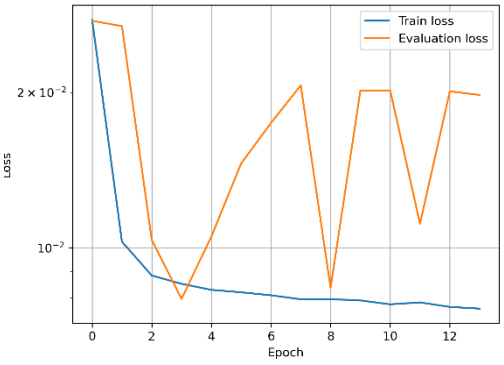


Figure 2B: Results after Fine-tuning



Fine-tuning: Consequently, we have put some effort into fine-tuning the model. The fine-tuning was applied using a batch size of 128, an Adam Optimizer with a learning rate of 0.00001 and a combination of MSE-SSIM with the following form:

$$0{,}5MSE + 0{,}25(1 - SSIM)$$

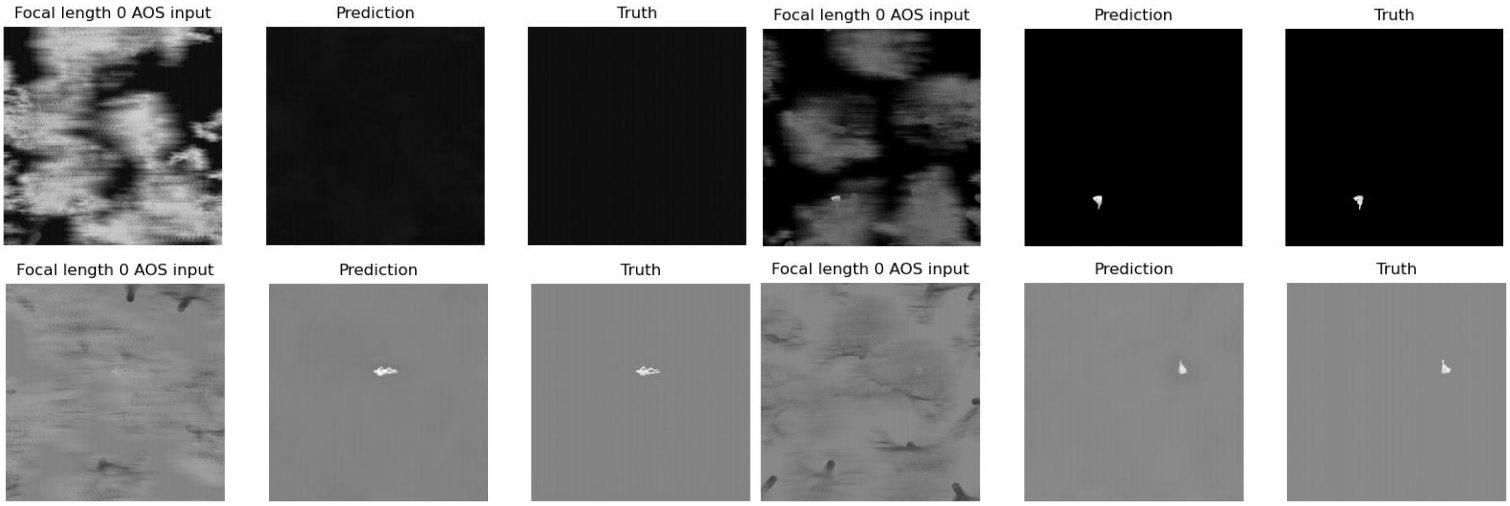
The final losses were:
Train loss MSE SSIM: 0.00715
Eval loss MSE SSIM: 0.00745
Test loss MSE SSIM: 0.00743

The results of the initial training and after conducting the fine-tuning are illustrated in Figures 2A and 2B respectively. The mean PSNR value on our test set was 34.92.

Results & Discussion

In general, it can be concluded that the performance of the proposed guided UNet3+ model demonstrates remarkable accuracy and yields excellent results for the given task of image restoration. In the following, a collection of different images from the test set with diverse backgrounds, temperatures, poses and even images that do not actually contain the main targets, i.e., humans, together with the predictions and the given ground truths are presented (Figure 3).

Figure 3: Predictions for various backgrounds, temperatures and poses



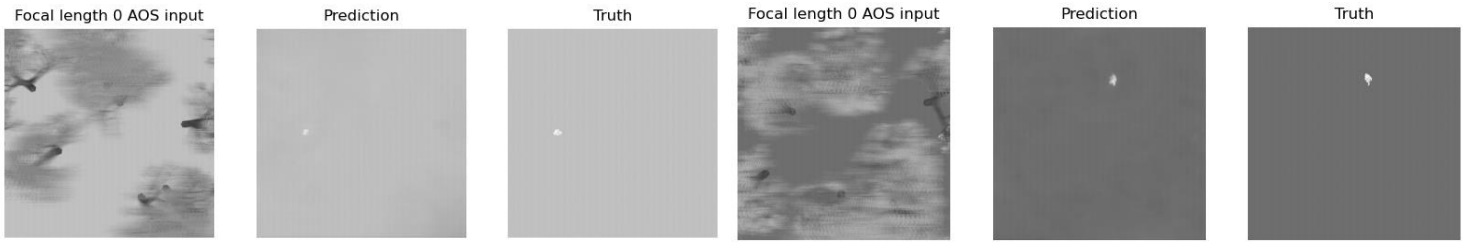
The results presented in Figure 3, underline that the model clearly exhibits strong generalization capabilities. They show that the model succeeds in reconstructing the images a reasonably close manner. It can detect whether humans are present, and in most cases generates an accurate prediction resembling the ground truth, irrespective of the ground temperature, the amount of occlusion, or the poses, as it can be seen in Figure 3. It also seems to cope extremely well with the wide variety of backgrounds in which it generates its predictions, and even with the absence of people in a given image.

In conclusion, it can be clearly stated that the proposed guided-UNet3+ model has been able to fulfill the task in identiying the targets and producing predictions that resemble the ground truths in a very adequate and accurate fashion. However, the model is certainly not without shortcomings. There are some limitations, which will be presented in the subsequent section.

Limitations

Although the model is fairly accurate and manages to do a very good job in general to acclimatising to various scenarios, there can be a small number of occurrences in which the model might yet lack capabilities to accurately reproduce particular ground truths.

Figure 4: Images with highly occluded targets



The illustrated examples in Figure 4 are examples of the behaviour described above. The model still lacks accuracy when it has to extrapolate the target to completely occluded areas. However, this is to be expected as generative capabilities have not been the focus of our approach in architecture design, training and evaluation. Therefore, incorporating generative capabilities to the model could potentially contribute to achieve even greater accuracy and consistency, even for instances in which the targets are highly occluded.