Computer Vision Project Report - Group A6

Paul Dobner-Dobenau
k12005600

Nico Filzmoser
k12006215

Markus Frohmann
k12005604

Enes Sovtic
k12005938

Kilian Truong
k12007306

Noah Pichler
k12011672

ABSTRACT

This report documents the details of our computer vision project about identifying a person in a forest from drone images. We implemented a hybrid architecture that uses DINOv2 as an encoder and a DPT decoder, which outperformed our baselines and showed good results both regarding location and sharpness of the person.

1 INTRODUCTION

Image restoration in computer vision describes the process of taking a corrupted image and transforming it into the original, clean image. In recent years, this area has become a focus of research in artificial intelligence.

For this project we analyzed sequences of drone images captured over a forested area. These images may contain a person residing in the forest, who is often obstructed by trees. The task was to restore an image showing an unobstructed view of that person.

The implementation was realized via machine learning models, which will be explained in this report. The input images were available in four focal planes to capture the person correctly in every possible pose: 0m, 0.5m, 1m, 1.5m.

2 RELATED WORK

Building upon the foundational concept of image restoration in computer vision, as outlined in Section 1, a lot of research has been dedicated to advancing techniques and methodologies in this domain.

We explore advancements in neural network architectures for image restoration, particularly relevant to our task of enhancing drone images for person identification in forested areas.

Significant works in this domain include "Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections" [1] highlighting the importance of detail preservation in image restoration via skip-connections.

Similarly, "Uformer: A General U-Shaped Transformer for Image Restoration" [5] and "Restormer: Efficient Transformer for High-Resolution Image Restoration" [6] introduce novel Transformer-based architectures, focusing on high-resolution image processing and efficiency.

Additionally, "DINOv2: Learning Robust Visual Features without Supervision" [2] presents an innovative self-supervised learning approach, which we investigate further to be used as an encoder backbone for our proposed models.

3 IMPLEMENTATION APPROACHES

This section explains the details of our implementation in regards to pre-processing, model and baseline selection, post-processing and other important details.

3.1 Focal stacks

The focal stack consists of multiple images of the same motive taken at different focus distances. In a lot of image processing application, these images are combined into a resulting image with a greater depth of field than the source images. This is especially useful if different areas of an image are of interest and therefore need to be sharp.

In this project, we worked with a focal stack at focal distances of 0m, 0.5m, 1m, 1,5m measured from the ground. The intention behind those numbers was to get a clear picture of every part of the person in the forest.

3.2 Augmentation

We used two different techniques for data augmentation. The first one randomly flips the images horizontally or vertically with a probability of 0.5. The second is called MixUp, which creates an interpolation of two training images as a synthetic sample. Its purpose is generalization to unseen domains, in our case real world images.

3.3 U-Net Baseline

The U-Net model was originally developed for biomedical semantic segmentation[4]. Its architecture consists of a contractive path and an expansive path. The contractive path follows the structure of a typical CNN, consisting of repeated sequences of convolution- and pooling-operations, which project the input to lower dimensions.

The expansive path consists of convolutions and upsampling operations, where features are concatenated with the corresponding feature maps from the contractive path. The output returns a single image with the same width and height as the input.

The structure of the architecture was kept the same as the publicly available version, only the dimensions of the projections were adjusted to our data. The corresponding block diagram can be seen in figure 1. The input focal stack images were standardized and normalized using the mean and standard deviation of the training set.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

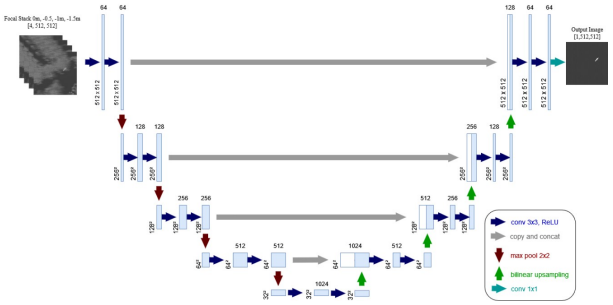


Figure 1: U-Net architecture

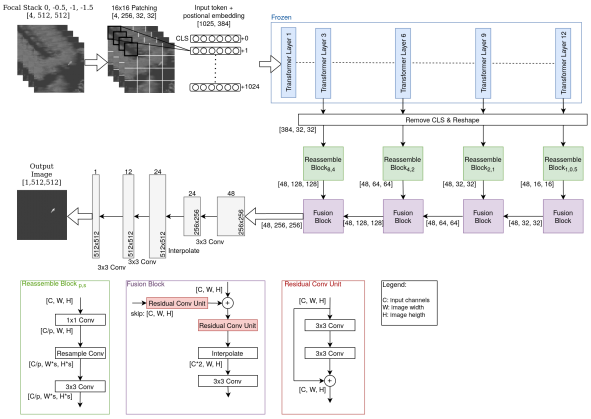


Figure 2: U-Net architecture

3.4 DINOv2 Main Model

Our main model consists of two parts. The first part is a pretrained DINOv2 encoder, which generates the feature embeddings for the input images. DINOv2 is a family of open-source models designed for computer vision tasks. The models come in different sizes in regards to parameter count. Our experiments have indicated no discernible difference in performance between the sizes, so we selected the smallest one with around 21 million parameters. We adapted the published by changing the number of input channels to 4 and the patch size in the second step to 16x16. Because of the resulting change in sequence length we also had to train our own positional embedding.

The second part is made up of a DPT (Dense Prediction Transformer) decoder[3], an architecture designed for dense prediction tasks, particularly depth estimation, which we applied to grayscale image reconstruction. It progressively assembles and fuses the feature representations returned by DINOv2 into the final dense predictions. The full pipeline can be seen in figure 2.

3.5 Loss function

Our loss function has the form $\mathcal{L} = \mathcal{L}_{MSE} + 2\mathcal{L}_{MSGE}$, where $\mathcal{L}MSE$ denotes the mean squared error between prediction and ground truth and $\mathcal{L}_{MSGE}$ denotes the mean squared gradient error, which is defined as follows:

$$\mathcal{L}_{MSGE} = \frac{1}{m} \sum_{i \in M} \left(\nabla_x \left(p_{i,x}^{hv} \right) - \nabla_x \left(\Gamma_{i,x} \right) \right)^2 \quad (1)$$

$$+ \frac{1}{m} \sum_{i \in M} \left(\nabla_y \left(p_{i,y}^{hv} \right) - \nabla_y \left(\Gamma_{i,y} \right) \right)^2 \quad (2)$$

It calculates the mean squared error between the gradients of the pixels of the prediction and the ground truth in horizontal and vertical direction.

Compared to simple MSE, our loss function has yielded better results regarding edges and details around a person.

4 EVALUATION AND RESULTS

The evaluation of our image restoration project involved assessing both the U-Net baseline and the DINOv2-based main model. These models have been primarily trained using the loss function as outlined in Section 3.5.

Model	SSIM	PSNR (dB)
DINOv2	0.9635	36.181
U-Net	0.6822	25.403

Table 1: Comparison of SSIM and PSNR scores evaluated on an independent test set for U-NET and DINOv2-based Main Model

However, to gain a comprehensive understanding of their performance, we also evaluated them using Structural Similarity Index Measure (SSIM) and Peak Signal-To-Noise Ratio (PSNR).

4.1 Training, Validation and Testing

In order to derive meaningful performance comparisons for our models, we opted for a classic 80% training, 10% validation and 10% test split to evaluate our models, for which the training is only done on the training set, the hyperparameter and architecture fine-tuning is only done on the validation set and finally the metrics and the sample images are derived from the test set.

4.2 U-Net Baseline Performance

The U-Net baseline model, despite its simple architecture, showed promising results in certain aspects, however lacked in others:

Person Detection: The U-Net model exhibited a robust capability in detecting the location of a person. It was particularly effective in capturing finer details such as the posture of the person, like if it was sitting or lying on the ground, which seemed quite surprising for a model with such a relatively straightforward architecture.

Model Size: The model, containing approximately 31 million trainable parameters, is quite substantial in size. This large number of parameters, while beneficial for complex feature detection, also makes the training process notably time-consuming and resource-intensive.

False Positives: One notable issue with the U-Net model is its tendency to produce false positives, particularly in scenarios where no person is present in the input. This suggests a weakness in the model's ability to accurately distinguish between relevant and irrelevant features in the focal stack data provided.

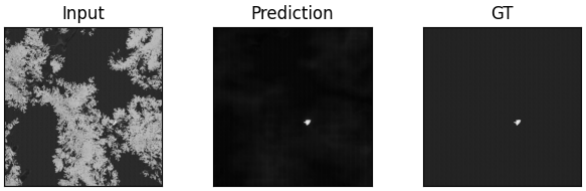


Figure 3: U-Net Baseline Sample Result to compare with DINOv2-based Main Model

Background Artifacts: While the U-Net baseline model produced well detailed images of the person, removing the encompassing forest in the input pictures, it didn't perform as well in terms of background clarity. The images accurately showed the person's position and pose, but they were affected by varying amounts of artifacts and noise in the background. This aspect reveals a limitation in the model's ability to effectively differentiate and cleanly separate the foreground from its background elements, leading to less than ideal outcomes in overall image quality.

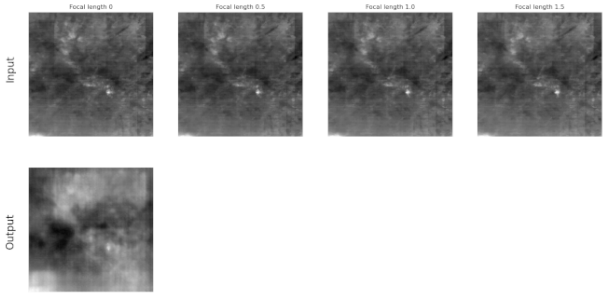


Figure 4: U-Net baseline model result on Real Image Data

4.3 DINOv2 Main Model Performance

Initially, the DINOv2 main model used a simpler convolutional decoder head, which was effective in localization but resulted in softer shapes. Later, the introduction of a Dense Prediction Transformer (DPT) head significantly improved the model, leveraging the full potential of DINOv2 for better localization and detail.

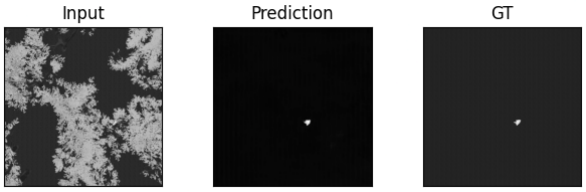


Figure 5: DINOv2-based Main Model Sample Result to compare with U-Net Baseline

Localization Accuracy: The model demonstrated superior performance in pinpointing the exact location of a person within the forest environment, outperforming our U-Net baseline significantly. This capability is particularly notable given input focal stacks with challenging conditions in terms of locating the person under the forest.

Resource Efficiency: Despite the high computational demands associated with generating the embeddings using the pre-trained DINOv2 model, the core image restoration model is notably more compact. This design approach results in a model with significantly fewer trainable parameters (approx. 3M) than found in the U-Net baseline model.

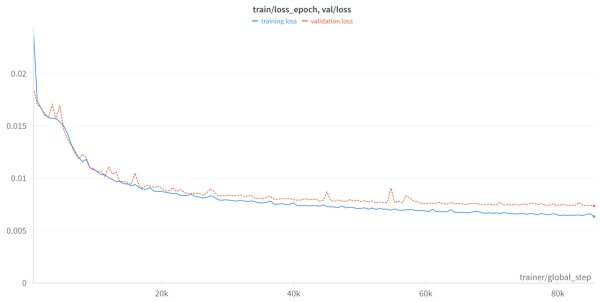


Figure 6: Training and validation losses for main model

High Quality Details: Following a series of initial adjustments and fine-tuning, the model reached its optimal form, producing, compared to the U-Net baseline, very clear and accurate predictions which closely align with the provided ground-truths, showcasing the model's ability to capture and reconstruct high quality details.

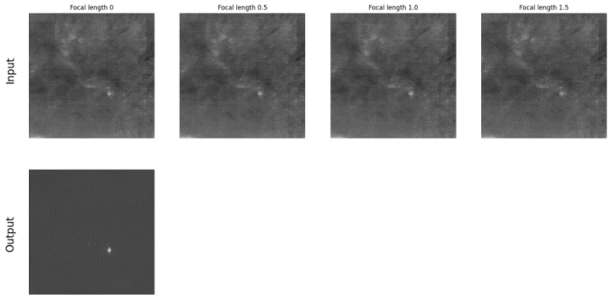


Figure 7: DINOv2-based Main Model Result on Real Image Data

Despite the high resource demand for generating image embeddings from the focal image stack, the DINOv2 model is beneficial due to its lower number of trainable parameters, and it's much better prediction performance, making it efficient once the initial resource-intensive embedding generation is completed.

Additional example outputs for the final model are placed in the appendix.

5 CONCLUSION

In conclusion, our team project demonstrates the potential and the effectiveness of pre-trained self-supervised vision models like DINOv2, as it served as valuable backbone for our image restoration DPT decoder model.

REFERENCES

[1] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. 2016. Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections. (2016). arXiv:cs.CV/1606.08921

[2] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes,

Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. DINOv2: Learning Robust Visual Features without Supervision. (2023). arXiv:cs.CV/2304.07193

[3] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision Transformers for Dense Prediction. (2021). arXiv:cs.CV/2103.13413

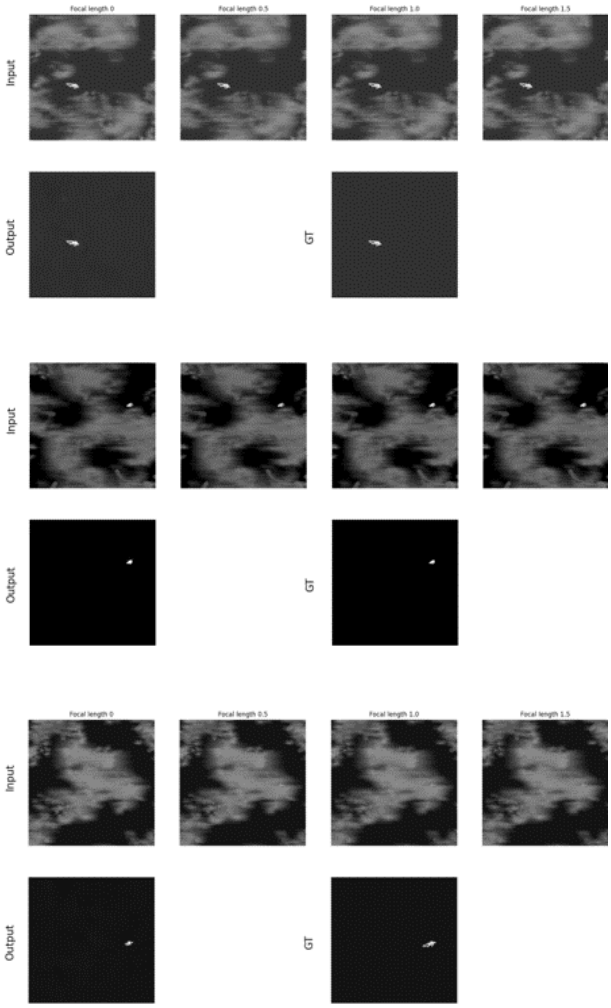
[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. (2015). arXiv:cs.CV/1505.04597

[5] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. 2021. Uformer: A General U-Shaped Transformer for Image Restoration. (2021). arXiv:cs.CV/2106.03106

[6] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient Transformer for High-Resolution Image Restoration. (2022). arXiv:cs.CV/2111.09881

6 APPENDIX

This section features some additional test samples to demonstrate the performance of the main model documented in the report. One can see that that good results are achieved throughout different settings.



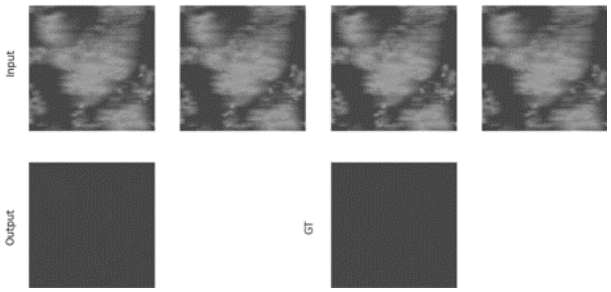


Figure 8: Example outputs