CHALMERS
UNIVERSITY OF TECHNOLOGY

AI-driven Single Image Super Resolution for Improved Neuron Segmentation

Bachelor's thesis in computer vision and medical image analysis

Viggo Trobäck
Lukas Karlsson
Arash Shahsavari
Karl Wiklund

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Bachelor's Thesis 2025

AI-driven Single Image Super Resolution for Improved Neuron Segmentation

Viggo Trobäck
Lukas Karlsson
Arash Shahsavari
Karl Wiklund

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden 2025

AI-driven Single Image Super Resolution for Improved Neuron Segmentation
Viggo Trobäck, Lukas Karlsson, Arash Shahsavari, Karl Wiklund

Supervisors: Ida Häggström, Department of Electrical Engineering, Chalmers University of Technology
Valentin Gillet, Department of Biology, Lund University
Examiner: Fredrik Kahl, Department of Electrical Engineering, Chalmers University of Technology

Bachelor's thesis 2025
Department of Electrical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2025

Contents

Abstract

Connectomics research relies heavily on high-resolution imaging of neurons, allowing for the segmentation and tracing of nerve structures. Typically, high-resolution images cover only limited areas, while broader overviews are captured at lower resolutions. As segmentation techniques continue to improve, the analysis of these low-resolution regions has become of increasing interest. AI-driven super resolution models offer the potential to upsample low-resolution neural images, enabling automated segmentation in regions that were previously unusable for analysis and increasing precision in areas with high feature density. As this specific application of super resolution is previously unexplored, and given the growing variety of model architectures available, this work investigates three representative models of different architectures, Real-ESRGAN, SR3, and EDT, and variety of training loss functions. The goal is to compare the strengths and limitations of these architectures when fine-tuned on serial block-face electron microscopy images. This task demands a high degree of structural consistency between low and high resolution outputs. REAL-ESRGAN and SR3 were found to be prone to hallucinations and artifacts, which can hinder downstream applications. In contrast, the fine-tuned EDT models tended to produce overly smooth outputs and in this removed small features. Some improvement was achieved with a task-specific EDT model trained from the ground up and the use of structural similarity–based loss functions.

1 Introduction

Many research fields within biology are becoming increasingly data-driven. To unravel the neuronal processes behind animal behaviour, researchers work on digitally reconstructing parts of insects' central complex, a specific region of the brain that computes navigation among other behaviors. These reconstructions are obtained by photographing select parts of the animals' central complex, layering the images and then tracing individual neurons through this image volume (Heinze and Homberg 2008), as seen in Figure 1. In the emerging field of connectomics, the study of mapping neural connections, the need for automated image assembly, and neuron segmentation within and across layers is rapidly growing (Jurrus et al. 2013). High resolution (HR) images are ideal for reconstruction, but challenging to acquire since photographing large areas with HR is costly, and produces large data volumes that are difficult to store and handle. Due to these factors, only selected regions are photographed in HR, and the larger regions are photographed in lower resolution. In the process of taking serial block-face electron microscopy (SBEM) images the sample gets destroyed as each layer is removed from the sample after being processed (Briggman and Bock 2012). As automation processes are becoming increasingly efficient, larger amounts of data can be processed and regions that previously would not have been considered for segmentation are now of potential interest. It would thus be beneficial if the previously less prioritized low resolution (LR) regions could be upsampled and explored without having to repeat the entire data acquisition and segmentation for a completely new sample.

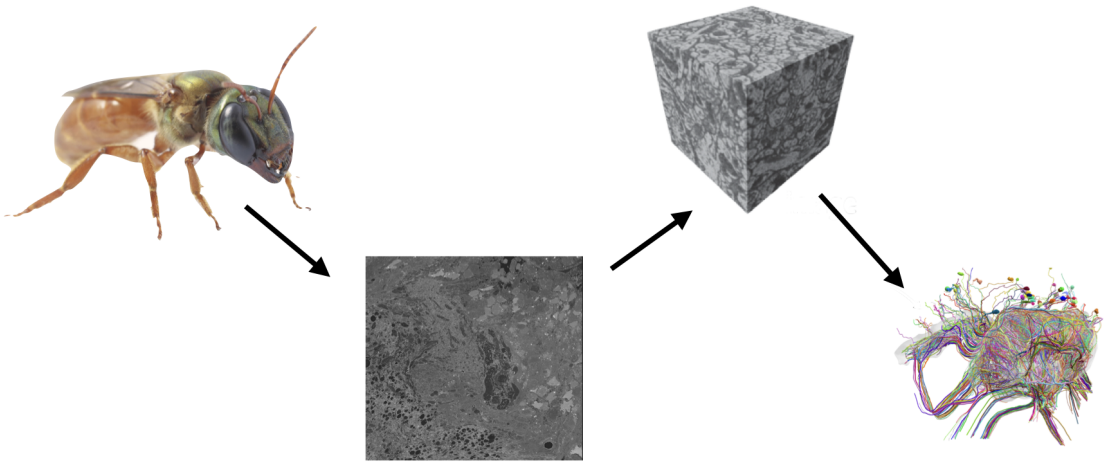

Figure 1: Overview of neuron tracing. Areas of interest are captured as 2D slices and assembled to 3D image volumes from which neuron tracts can be segmented.

Recently, there have been many advances in using Deep Neural Networks (DNNs) for super resolution (SR) tasks, turning LR data into possible HR counterparts (Yang et al. 2021). As this field is still rapidly growing, a variety of possible network architectures are being explored in parallel with different strengths and shortcomings in how well they can be fine-tuned for specific tasks.

2 Theoretical Background

SR is the computer-vision task of transforming LR images, volumes and videos into HR. Specifically, this work considers Single Image Super Resolution (SISR). SISR is an ill-posed problem since there can be many possible HR images for any given LR image, but if done correctly it would enable new possibilities for tasks such as neuron segmentation where feature loss in low quality images is an issue. By enhancing the quality of LR images, automated segmentation and reconstruction may become possible in new areas. It would also make manual reconstruction faster and potentially more accurate. Higher accuracy for both manual and automated neuron reconstruction would lead to larger amounts of data being able to be processed with less manual proofreading. The use of LR images in place of HR images would therefore lower both labor cost, as well as making data acquisition faster and more economical.

SR methods have evolved from simple techniques, such as interpolation, to more sophisticated AI-driven approaches. One of the most common traditional methods is applying an interpolation kernel over the image to approximate the continuous function that underlies the discrete pixel samples. These methods, such as bilinear and bicubic interpolation, are relatively fast and computationally simple. However, these algorithms struggle with preserving fine details, especially edges and lines (van Ouwerkerk 2006). These features are critical in biological data. As a result of the shortcomings of traditional approaches, new algorithms based on deep learning have been developed. These new methods offer improved performance in handling complex image structures by using information learned in training to add details that would have been lost in downsampling.

2.1 Artificial Intelligence in Super Resolution

Recently, there have been many advances in using DNNs for SISR tasks (Yang et al. 2021). Since the field is still emerging, numerous models and approaches are being developed in parallel. This work focused on three categories of DNNs for image processing, namely convolutional neural networks (CNNs), transformer-based neural networks and diffusion models.

Convolutional Neural Networks

CNNs are a foundational deep learning technique, especially for image processing tasks. CNN-based models stack layers such as convolutions, activations and pooling to progressively extract and refine features from input images. CNNs have proven effective in learning spatial hierarchies of features, which makes them very suitable for enhancing the resolution of low quality images. Dong et al. (2016) introduced the SRCNN, which was the first model to successfully apply deep learning to SR. The SRCNN demonstrated that a relatively shallow network could achieve advanced performance by learning a direct mapping from LR to HR images.

A popular class of models that incorporates CNNs is Generative Adversarial Networks (GAN). These use a generator neural network and a discriminator neural network that are trained against each other. The generator tries to produce realistic outputs, while the discriminator attempts to distinguish between real and generated data. This adversarial training process creates a dynamic where both networks improve through competition, but it can also lead to instability if one network overpowers the other. A GAN dedicated to SR, the Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN) shows great potential, but has challenges with training stability due to the adversarial nature of GANs (Wang, Yu, et al. 2018).

ESRGAN was later upgraded to a new version, Real-ESRGAN (Wang, Xie, et al. 2021). The purpose of upgrading the previous model was to improve visual performance and make it more adaptable to real-world applications. To achieve this improvement Wang, Xie, et al. (2021) added a U-Net

discriminator with spectral normalization to increase the capability of the discriminator model. The U-Net discriminator is a symmetric architecture containing an encoder that successively reduces resolution and increases number of features until the center of the network, after which a decoder reverses the process (Ronneberger, Fischer, and Brox 2015). Additionally, U-net uses skip connections, which connect the encoder and decoder at same hierarchical level to allow features and textural details to be preserved.

One prevalent challenge in employing GAN architectures for image upsampling is their susceptibility to checkerboard artifacts, which often arise from the uneven overlap of convolutional kernels during transposed convolution operations (Odena, Dumoulin, and Olah 2016). Checkerboard artifacts appear as grid-like patterns in the resulting image.

Diffusion models

Denoising diffusion probabilistic models, or diffusion models for short, are a class of generative vision models that learn to iteratively go from noise to real images (Ho, Jain, and Abbeel 2020). Inspired by non-equilibrium thermodynamics, there are two processes at work in diffusion models: forward diffusion, where noise is gradually added to the data, and backward diffusion, where the model learns to remove the noise and recreate the original image.

Super resolution via repeated refinement (SR3) belongs to the class of diffusion models, adapted for conditional image generation (Saharia et al. 2023). It produces an HR image from an LR one by starting with an image of pure noise concatenated to the LR image and iteratively denoising, as seen in Figure 2.

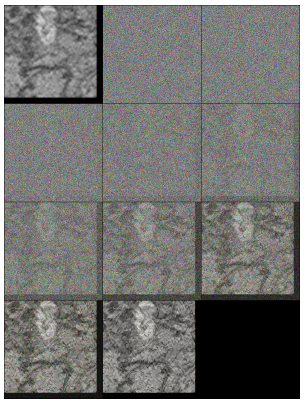

Figure 2: Pre-trained SR3 iteratively producing a super-resolved LR image. LR image in top left, iterations from pure noise to final SR image from left to right, top to bottom.

The backbone of SR3 is a modified U-net architecture (Ronneberger, Fischer, and Brox 2015). While regular diffusion models generate a diversity of images from noise, SR3's concatenation of the LR image helps steer the image generation towards plausible SR alternatives. Furthermore, U-net's skip connections provide structural guidance from the LR image for each resolution level during the refinement.

Transformer-based Neural Networks

Transformer-based neural networks, which have been revolutionary in natural language processing, are also being applied to SR tasks (Chen et al. 2021). Unlike CNNs, which focus on local image features,

transformers use self-attention mechanisms which allow them to capture long-range relationships across the entire image. The transformer models can focus on the most relevant regions and better understand the overall structure of an image, potentially leading to better overall image enhancement. Because attention can be computed across all parts of the input simultaneously, transformer-based models are typically faster to train than recurrent or sequential models (Vaswani et al. 2017. Transformer-based networks are also more efficient in utilizing large pre-training datasets compared to CNNs (Chen et al. 2021). This makes them suitable for fine-tuning with limited data for specific applications.

Chen et al. (2021) propose the Pre-Trained Image Processing Transformer (IPT) model, a general transformer-based network for image processing. The model trained on a very large dataset and then fine-tuned for various low-level computer vision tasks such as SR, denoising and deraining to great success. IPT shows promise but due to its use of full attention transformers it requires massive amount of parameters, 116 million, even for low level computational tasks (Chen et al. 2021). This brings with it a large computational cost and makes exploring different applications through fine-tuning harder. W. Li et al. (2023) proposed as an alternative window based variant in the the form of an encoder-decoder-based transformer (EDT) which is more efficient, only 200 thousand parameters, but still show promising results for low level computer vision tasks (Li et al. 2023). This increased efficiency allows more efficient exploration of applications. W. Li et al. (2023) trained multiple models for varying downstream applications, including SR.

2.2 Fine-tuning deep learning models for SBEM images

Previous model implementations are typically trained on natural image datasets, such as ImageNet (Deng et al. 2009) and CelebA (Liu et al. 2015). The provided model weights are thus poorly suited to the context of SBEM images considered in this work. To investigate the suitability of the chosen models for the SBEM context, each model can be fine-tuned on SBEM images of the central complex. Fine-tuning here refers to training pre-trained models on SBEM data, so they learn the relevant features and textures. As the starting point for fine-tuning is pre-trained models, less training and thus less computational power is required to achieve good performance.

Since most SR models are trained on full-colour datasets with the goal of achieving high perceived image quality for humans (Seif and Androutsos 2018), their loss functions may not be optimized to retain feature consistency and contrasts in HR SBEM images. Common loss functions, such as mean squared error (MSE) and mean absolute error (MAE, also known as L1 loss) tend to produce overly smooth images and do not consider image structures. Other loss functions, such as G-Loss, are edged based and show greater potential retaining salient features, such as edges, that are important for certain computer vision tasks (Ge and Dou 2023). In SR, quality of upsampled images are often measured in Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM). PSNR is calculated by the highest pixel value and the MSE, with 100 being a perfect match and 0 being a complete mismatch or maximum possible error. SSIM considers luminance, contrast, and structure, and ranges from -1 (perfect inverse) to 1 (perfect overlap), with zero being no structural similarity (Ge and Dou 2023).

The models considered use a variety of loss functions. Real-ESRGAN uses a combination of loss functions, with different functions in different steps to combine pixel loss, adversarial loss, and perceptual loss to improve realism and detail (Wang, Xie, et al. 2021). The original EDT implementation uses SSIM and MSE Li et al. 2023. SR3 uses pixel-wise loss, such as L1 (Saharia et al. 2023). The choice of loss function is another factor, beyond model architecture, that affects each model's output.

2.3 Potential issues with AI in the context of SR and practical application

Although AI offers significant advantages in data interpolation, it is not without drawbacks. Hoffman, Slavitt, and Fitzpatrick (2021) mention that we cannot trust every detail that the model outputs as it is theoretically impossible for the model to perfectly conjure more data from less. The DNNs are trained to plausibly interpolate this data, but it is not possible to actually achieve the ground truth. There is a risk of models generating misleading details which do not accurately represent the HR image, such as constructing or removing lines incorrectly.

In the downstream task of neuron segmentation, segmentation models are typically trained on HR images to produce affinity maps, which capture pixel-level relationships (such as correlation across neighboring pixels). Affinity maps are processed with a watershed algorithm to partition the image into fragments, which are successively agglomerated into larger structures (Sheridan et al. 2023). These automated segmentation pipelines allow larger amounts of data to be processed with less human supervision.

However, using SR images generated by DNNs as input to these segmentation pipelines may lead to poorer segmentation results. Artifacts such as different feature contrasts, hallucinations and textures have the potential to degrade the quality of affinity maps and lead to unreliable segmentation. Since the segmentation pipelines are trained on real HR images, their performance on DNN-upsampled SR images remains uncertain.

3 Research Aim and Research Description

This work aims to explore AI-driven SR techniques for enhancing LR SBEM images of insect nervous systems. By improving image quality, the goal is to facilitate neuron segmentation application for higher accuracy and reduce the need for manual proofreading.

After investigating the performance of different DNN architectures for upsampling neuron images, the study will assess resulting image quality as well as downstream impact on neuron segmentation and reconstruction accuracy.

As LR images are faster and cheaper to acquire than HR ones, a successful AI solution would enable connectomics researchers to accelerate and automate parts of their workflow with data for which it previously was not possible.

The primary objective of this study is to identify, fine-tune and train a suitable SR model for upsampling LR SBEM images. This objective can be further divided into the following key research questions:

- Which model architecture is most efficient for this task?
- Are certain loss functions more effective in preserving critical image features?
- What are the most appropriate metrics for evaluating improvements in image quality, particularly in the context of subsequent neuron segmentation?

4 Delimitations

This study will focus exclusively on SISR of SBEM images. It will not cover other types of imaging technologies or tissues other than neurons. Although the assembled data represents 3D volumes, this study will focus solely on processing individual greyscale 2D images to find promising models and methods.

The data will be limited to HR and LR image sets from the tropical bee *Megalopta genalis* (with 4x resolution difference), the South American ant *Eciton burchellii* (with 3.75x resolution difference), and the South African dung beetle *Scarabaeus galenus* (with 3.08x resolution difference). As all data is collected from insects, and no application on human data is within the scope of this study, there is no need for further ethical consideration or data handling protocols.

Only three super-resolution models will be considered: Real-ESRGAN, SR3, and EDT to effectively test state of the art architectures with different approaches to the task of super resolution. Two different types of loss functions will be explored, prioritizing feature retention over smoothness. No pre- or post-processing, such as noise reduction, will be applied to the images in this study. The models will be trained only for SISR, optimizing only for given loss function and not downstream application.

5 Methodology

As the different models differed in training protocols and requirements, each had to have their own fine-tuning pipeline developed. All models used subsets of the same training and validation set, with the specific subset depending on size requirements of input data and computational efficiency of the model and available resources.

5.1 Data preparation

The raw data consisted of large LR overviews of the insect central complex, along with HR images of specific areas. The data was collected in two different passes, giving true LR and HR pairs. These images were generously provided by the Heinze lab, Department of Biology, Lund University along with parameters such as resolution values and offsets in the x, y, and z axes to align the HR image volume with the LR image volume.

As different regions may have slightly different features, image contrast, and feature density, image pairs from two different regions of the central complex were taken from three insects: one set from the central body (CB) and one from the protocerebral bridge (PB). Using Neuroglancer[1], the images were overlaid to visually confirm their alignment and cropped into smaller sections. To minimize distortions caused by the use of different lenses for the different resolutions and ensure precise alignment between the high- and low-resolution images, we compared the cropped sections using Scale Invariant Feature Transform (SIFT) features. These features were matched for as close to optimal overlap as possible (Lowe 2004).

The smaller image sections were resized to match the dimensions required by the respective training pipelines. For those pipelines that needed a precise resolution difference, the high-resolution images were interpolated to the appropriate size using bicubic interpolation. A smaller dataset to test fine-tuning pipelines before final training was also produced with 42 image pairs with LR size 48x48 and HR size of 192x192.

The finished dataset included 4.8 GB of *Eciton* CB, 2.3 GB *Eciton* PB, 5.7 GB *Galenus* CB, and 1.1 GB *Galenus* PB, and the validation set consisted of 5.6 GB of *Megalopta* CB.

To avoid data leakage and introducing bias, the images pairs from *Megalopta* were only used for validation and testing with the images from the two other insects being used exclusively for training. These were labeled by their z-slice along with x, y partition to allow data selection in training if not all slices or areas were desired. This would be applicable in instances where training was very slow and selection of data could be done to make results faster. Because the structural differences between adjacent z-slices are minimal within the image volume, selecting every third slice provided a smaller yet representative training set. This selection reduces redundancy while maintaining a similar variation within the training data.

5.2 GAN

Before training the ESRGAN and Real-ESRGAN models, preliminary evaluations were performed to assess their initial performance and the possibility of using the simpler model based on visual inspection of hallucinations and artifacts.

To construct a training pipeline for fine-tuning, the Real-ESRGANs training pipeline template[2] was

[1] `https://github.com/google/neuroglancer`
[2] `https://github.com/xinntao/Real-ESRGAN`

used with modifications to fit the purposes of this project. These modifications included downloading pre-trained models to use as weight initialisation, removing the degradation process, and adjusting parameters, such as number of epochs and batch size to better fit our computational resources and data size. Although the dataset consists of black-and-white images, the input layer of the Real-ESRGAN model was left unchanged with three input channels. Given the challenges associated with training GAN models, we chose to proceed exclusively with the combination of loss functions that has been proven effective in the pre-training.

Before the final training, training on a small amount of data was performed. The purpose of this was to see if the pipeline worked and that the training gave any results as GAN models are well known to be difficult to train stably. The final training on the larger dataset, as well as the inference, was conducted using NVIDIA L40s GPUs within the Chalmers Minerva computational cluster. The model was trained for 45 epochs with a learning rate of 10^{-4}.

5.3 SR3

The SR3 implementation[3] used as starting point was an architecture built for 8x upsampling, from 16×16 pixels to 128×128. The model is typically trained on an 8x downsampled HR image that can be compared with ground truth, while for real-world inference a regular image would be used for LR.

For this project, the standard pipeline was not immediately applicable since the LR images were not downsampled HR images, and the LR/HR image size ratios varied per organism, as described earlier. To adapt to the dataset of paired LR/HR SBEM images, the procedure was adjusted to slice larger images into smaller patches, and then upscale the LR image with the appropriate ratio. This was achieved with algorithm 1.

Algorithm 1 Batch Super-Resolution using Custom Upsampling and SR3 Refinement

Require: Set of LR/HR image pairs $\{(I_{LR}^{(i)}, I_{HR}^{(i)})\}_{i=1}^{N}$ where $I_{LR}^{(i)}$ has dimensions $H^{(i)} \times W^{(i)}$ and $I_{HR}^{(i)}$ has dimensions $sH^{(i)} \times sW^{(i)}$

Ensure: Set of super-resolved images $\mathcal{O}_{SR} = \{O_{HR}^{(i,j)}\}$, each with resolution 128×128

Part 1: Patch Extraction and Scale-Aligned Upsampling

for each image pair $(I_{LR}^{(i)}, I_{HR}^{(i)})$ **do**
  for each non-overlapping spatial location j **do**
    Extract patch $I_{HR}^{(i,j)} \in \mathbb{R}^{128\times128}$ from $I_{HR}^{(i)}$
    Extract corresponding patch $I_{LR}^{(i,j)} \in \mathbb{R}^{\frac{128}{s}\times\frac{128}{s}}$ from $I_{LR}^{(i)}$
    $I_{up}^{(i,j)} \leftarrow \text{Upsample}(I_{LR}^{(i,j)}, \text{scale} = s, \text{method} = \text{bicubic})$
  end for
end for

Part 2: SR3-based Refinement

for each upsampled patch $I_{up}^{(i,j)}$ **do**
  Initialize $I_{SR}^{(i,j,0)} \leftarrow I_{up}^{(i,j)}$
  for $t = 1$ to T **do**
    $I_{SR}^{(i,j,t)} \leftarrow \text{SR3\_Refinement}(I_{SR}^{(i,j,t-1)})$
  end for
  $O_{HR}^{(i,j)} \leftarrow I_{SR}^{(i,j,T)}$ ▷ Final output after T refinements
end for
return $\mathcal{O}_{SR}$

[3] https://github.com/Janspiry/Image-Super-Resolution-via-Iterative-Refinement

Part 1 of algorithm 1 was used to prepare the data for both training and inference of SR3.

The model was fine-tuned with two passes through the training data, using a batch size of 64 images. The Adam optimizer (Kingma and Ba 2017) was used to minimize L1 loss with a learning rate of 3e-6.

5.4 EDT

Transformed-based architectures are known to demonstrate greater responsiveness to fine-tuning than the alternative model architectures (Li et al. 2023). For this reason, EDT was chosen to explore what potential differences there would be from fine-tuning with the different loss functions MSE, SSIM and G-Loss. MSE and SSIM are classic loss functions while G-Loss is a novel approach (Ge and Dou 2023). To allow for training with G-Loss using the same pipeline as the other loss functions, G-Loss was implemented in the project according to its description.

A custom training pipeline was constructed for fine-tuning EDT, drawing inspiration from the sample test code provided in the EDT source code[4], with several modifications. As the SBEM images are greyscale, each image was converted into a three-channel RGB tensor by replicating the greyscale values across all channels to fit the input size of the pre-trained model. The pipeline was also adapted to support multiple input data folders.

The training on the large dataset was conducted on a NVIDIA L40s GPU for 21 epochs using the ADAM optimizer with a learning rate of 10^{-4}.

5.5 Creating customized EDT model

In addition to the existing models, a new configuration with a larger input window size (16x64) and only one colour channel was added. This was in order to better capture the large features in the images if the self attention didn't generate sufficient results with the original smaller window sizes(6x24). Using one channel instead of three to increase efficiency and taking larger input images (128x128) as training data as the original (48x48) didn't capture larger features in their entirety. Additional layers were added to accommodate this larger input size.

The custom EDT configuration was only trained with the proven MSE-metric to explore if it captured larger structures better than the already implemented configurations. It was trained for 5 epochs until loss decreased and converged on all training data using the ADAM-optimizer and a learning rate of 10^{-4}.

5.6 Evaluation of results

The final model outputs were evaluated using the standard SR metrics, SSIM and PSNR. As texture and feature retention is important for this specific application these were examined visually in sample images to compare how different models perform when faced with tasks such as missing features, textures and performance on areas with high feature density.

Given the downstream application for neuron segmentation, additional metrics were used to assess segmentation performance. Using the segmented labels from HR images as ground-truth a precision/recall score was calculated as the amount of correctly labeled pixels in the segmentation of the upsampled counterpart.

[4] `https://github.com/fenglinglwb/EDT`

6 Results and result analyses

From testing and evaluating the models without task-specific training, it became clear what dependencies were needed or outdated and what documentation that was lacking. All pre-trained models were ultimately successful in generating an image. In comparison with ground-truth and bicubic upsampling it is clear that no model handled the task satisfactorily. All exhibited noticeable artifacts and inaccurate hallucinations.

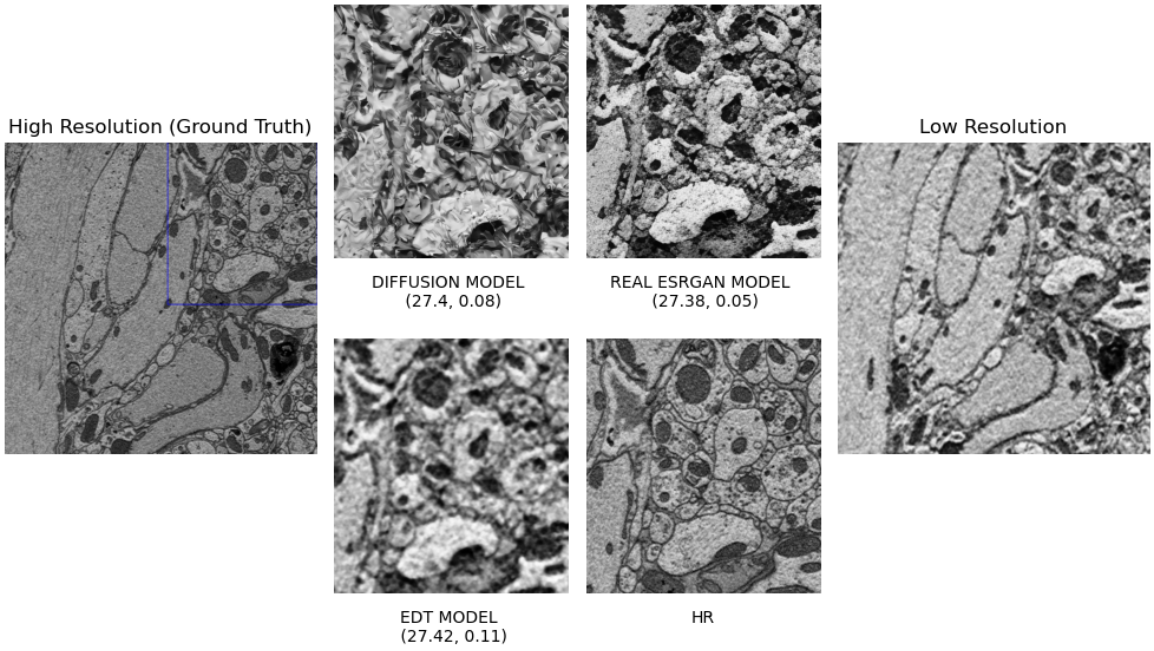

Figure 3: Images created by the models prior to fine-tuning as a baseline along with their evaluation metrics (PSNR, SSIM).

6.1 GAN

Initially, an inference of both the ESRGAN and the Real-ESRGAN models were conducted without fine-tuning the models on the neuron images. The resulting outputs were of low quality and failed to accurately reconstruct the fine structural details, as shown in the comparison in Figure 4. As expected, Real-ESRGAN showed greater suitability for this task as ESRGAN, with its simpler structure generated more artifacts.

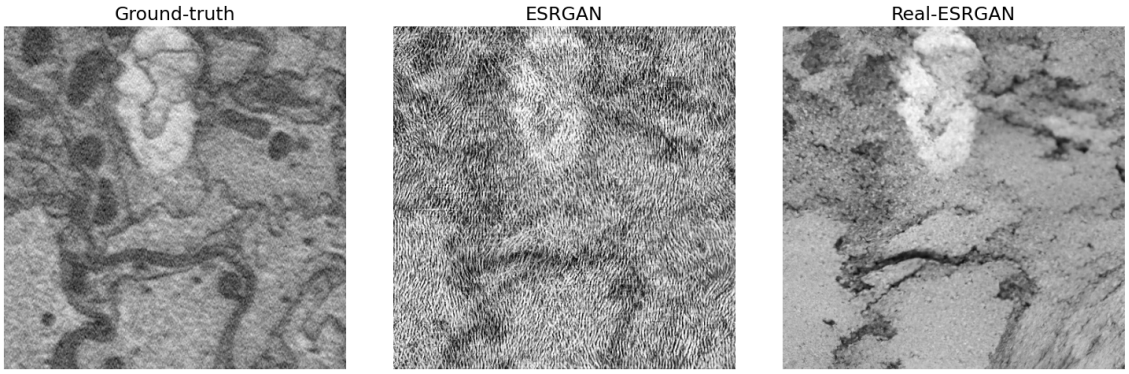

Figure 4: Output from ESRGAN and Real-ESRGAN models without fine-tuning, with pre-trained weights that does not include block-face EM images.

Some features in the HR image are not visible at all in the Real-ESRGAN SR image, especially looking at the areas where there are a lot of details in the structure (top right corner or just below it in Figure 4 for instance). Even though it produces incorrect edges and overly smooth textures, it doesn't share the artifacts of its predecessor.

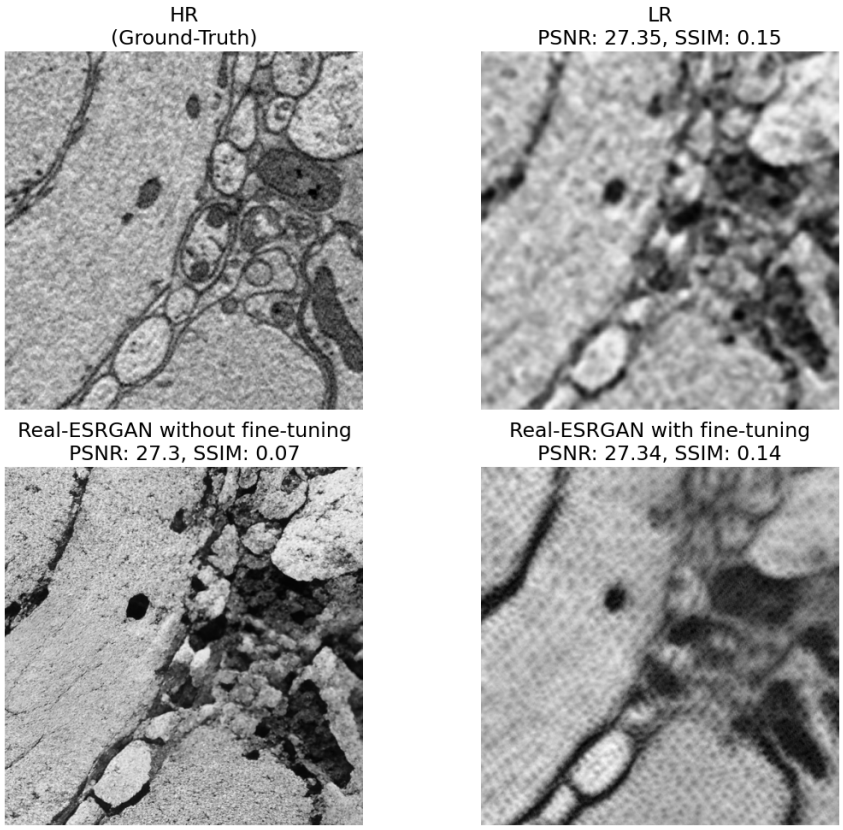

Figure 5: A comparison between the output from the model before the training process and the corresponding LR and HR images of the same area of the image.

A decision was made to use only a third of the larger dataset for training, motivated by time constraints. This was made by taking every third of the z-slices images from each training set. No clear patterns could be seen from the loss outputs observed at each epoch during training. In addition, the results from the validation, as can be seen in Table 1, do not indicate on significant improvements. PSNR and SSIM are implemented in the Real-ESRGAN architecture and it uses a combination of these and others in different training stages making evaluation of individual loss functions difficult.

PSNR	SSIM
14.8980	0.1502

Table 1: Average PSNR and SSIM scores over the entire validation dataset from Real-ESRGAN validation.

The generated SR image (lower right in Figure 5) appears visually sharper compared to a bicubic upsample. Although, it seems like the model have introduced a lot of significant hallucinations, which have generated incorrect features that does not exist in the original image. This is not unexpected, due to the fact that the training process was unstable, with fluctuating loss values. As a result, the perceptual quality might be higher than the actual quality of the images.

Signs of hallucination can be seen in the image to the right of Figure 5. Compared to ground-truth, some features have disappeared, or merged together, like in the area represented in Figure 6.

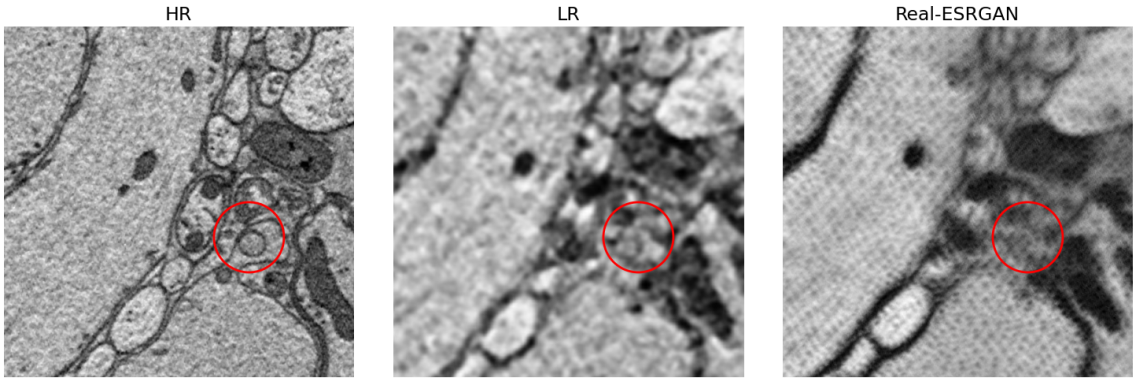

Figure 6: The red circle represents an area of the SR image with few features restored from the ground-truth image.

By looking at the same area in the HR image of Figure 5 several features can be seen which are not visible in Figure 6. These kind of artifacts might not be unexpected with the PSNR and SSIM values obtained for the SR image, which are moderately low. There are also clear signs of artifacts, as numerous spot-like anomalies are spread across the image in a grid.

6.2 SR3

As SR3 upsamples the image one small patch at a time, both the behavior on individual patches and across neighboring patches is of interest for the SBEM SR task considered.

Applying SR3 with the pre-trained model weights, the model generally retains the macroscopic features

of the image while adding a texture that can be described as crystalline or jagged. Figure 7 illustrates this on one patch of Figure 3.

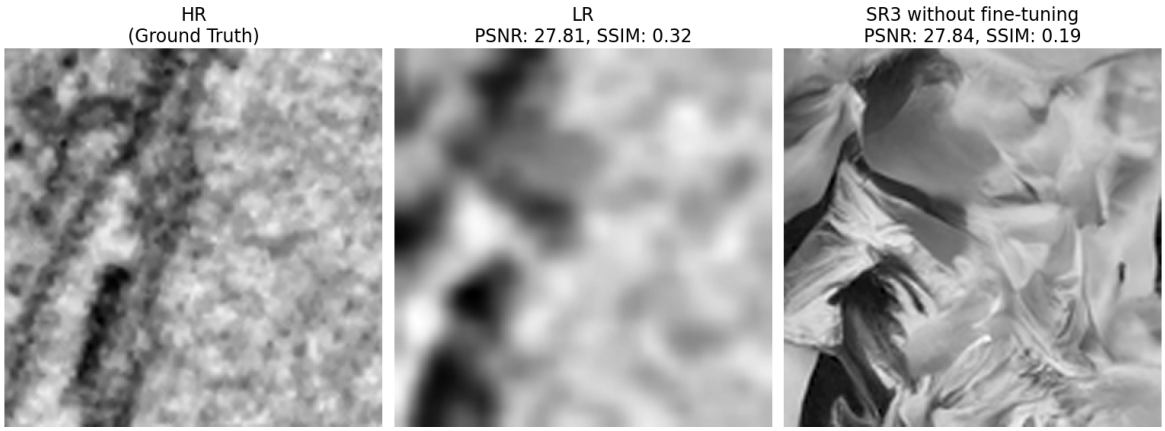

Figure 7: Pre-trained SR3 output on one image patch.

After fine-tuning SR3 for two epochs on the training data, the model improves its ability to recreate the grainy grayscale texture of SBEM images without its previous artifacts, but can be hit-and-miss with the larger image features. In Figure 8, a relatively simple feature of a diagonal line is upsampled well by the model.

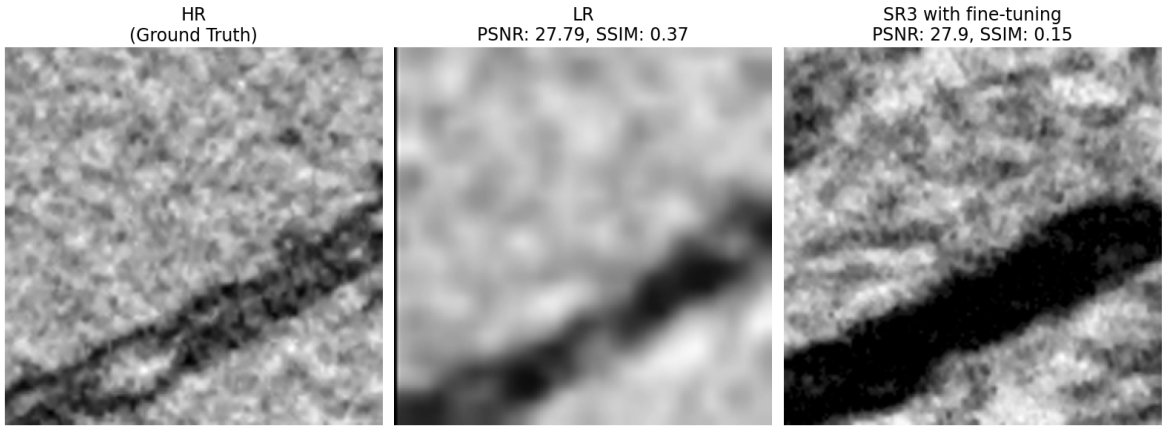

Figure 8: Fine-tuned SR3 manages to recreate a simple image feature.

In instances where there is a displacement or difference in macroscopic features between the LR/HR image pair (see Figure 9 as an example), SR3 struggles to match the HR contents. Rather, SR3 better matches the LR image. Given that SR3 SR3 lacked the architecture of for example EDT that allow it to pay attention to global features when processing chunks, SR3 may have been more sensitive to minor displacements and differences compared to the other models tested. During training, the loss function would punish such deviations between model output and HR image.

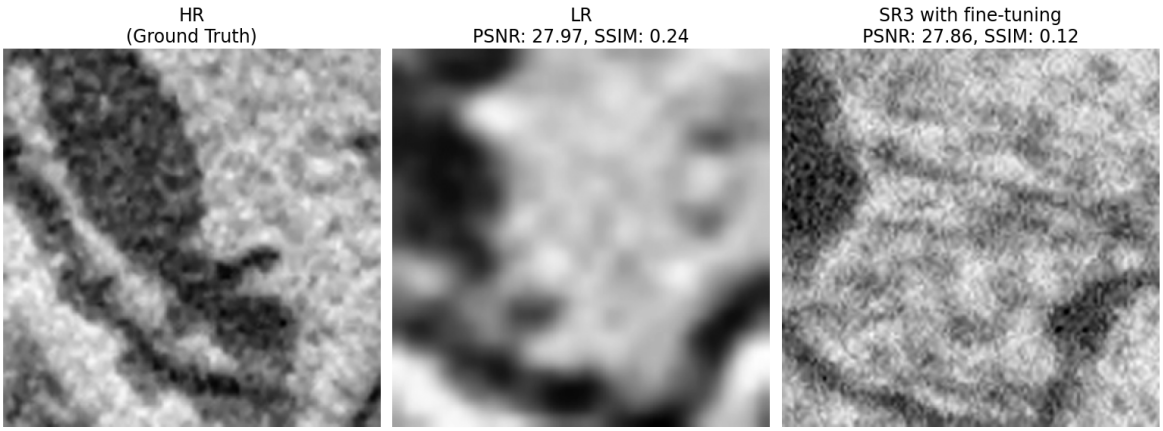

Figure 9: Feature appearing differently in LR than in HR. With the SR3 output better corresponding to the LR image.

Additionally, SR3 occasionally hallucinates images features, as in Figure 10. While the LR/HR pair have some curves in the middle and lower part of the patch, SR3 instead created two large dark areas in the upper part and lower right.

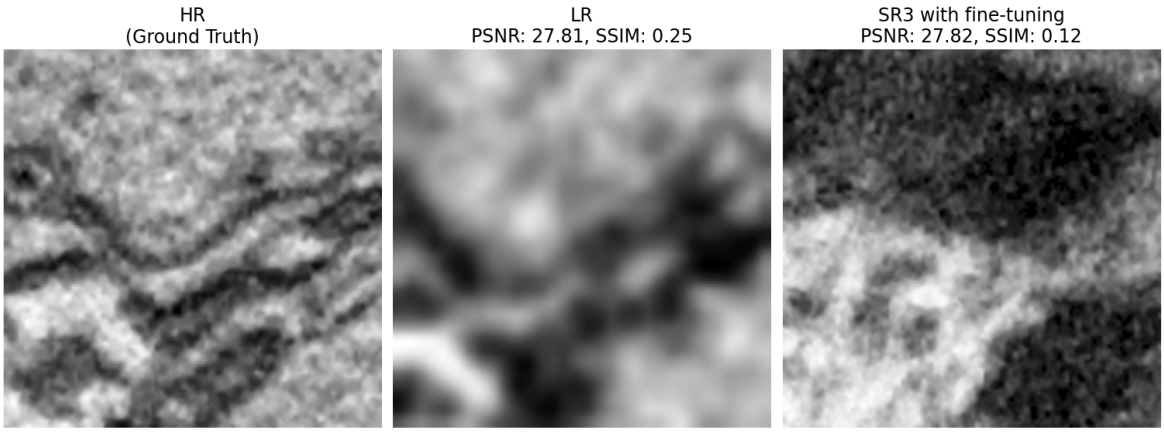

Figure 10: SR3 hallucinating image features on a patch.

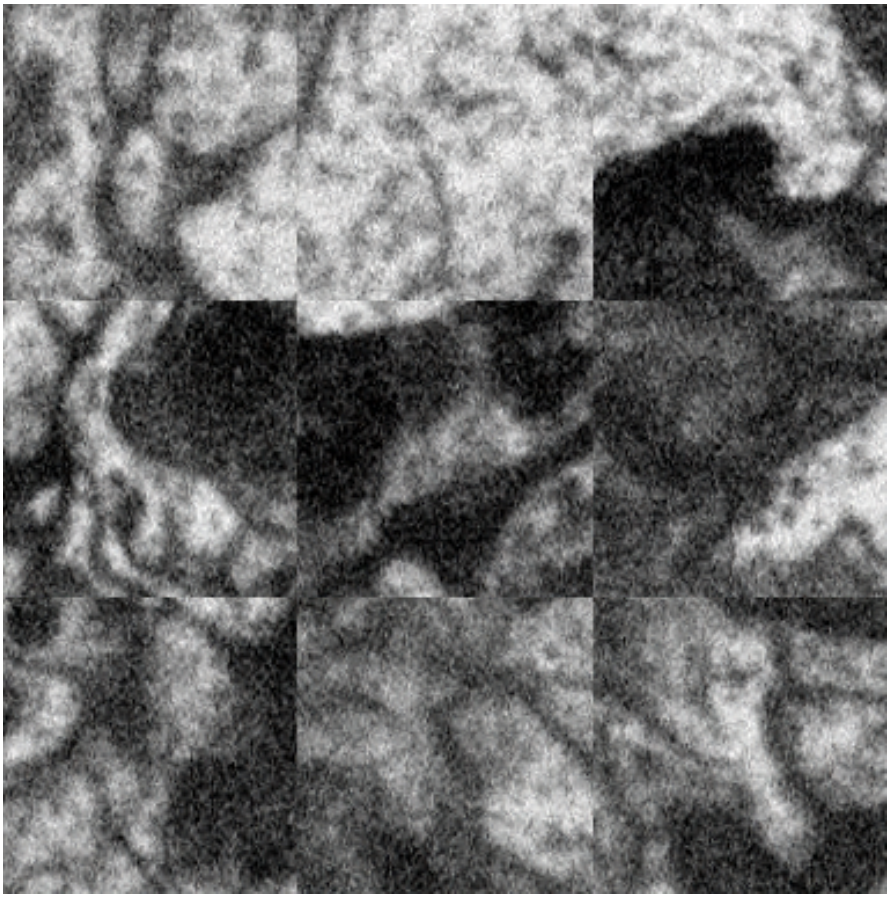

Figure 11: Fine-tuned SR3 fails to maintain consistency between patch borders.

SR3's failure to consistently recreate the contents of smaller patches leads to further problems when putting together a larger image, as seen in Figure 11. Areas that are dark on one side of the border are light on the other and vice versa, and many smaller features are lost.

6.3 EDT

Inference was run on the pre-trained weights without any fine-tuning and the difference to the bicubically upscaled image was minimal, see Figure 13. Some edges from the untrained model have slightly higher constrasts than their bicubic counterparts, but there are many details in the HR image that the EDT model fails to capture.

After initial testing the model was fine-tuned using the training dataset. The model was trained in three instances, using different loss functions (MSE, SSIM and G-Loss). The fine-tuning was run for 21 epochs with a learning rate of 10^{-4}. As figure 12 shows, the training loss decreased very slowly with this learning rate, but fine-tuning with a higher learning rate of 10^{-3} resulted in very unstable training.

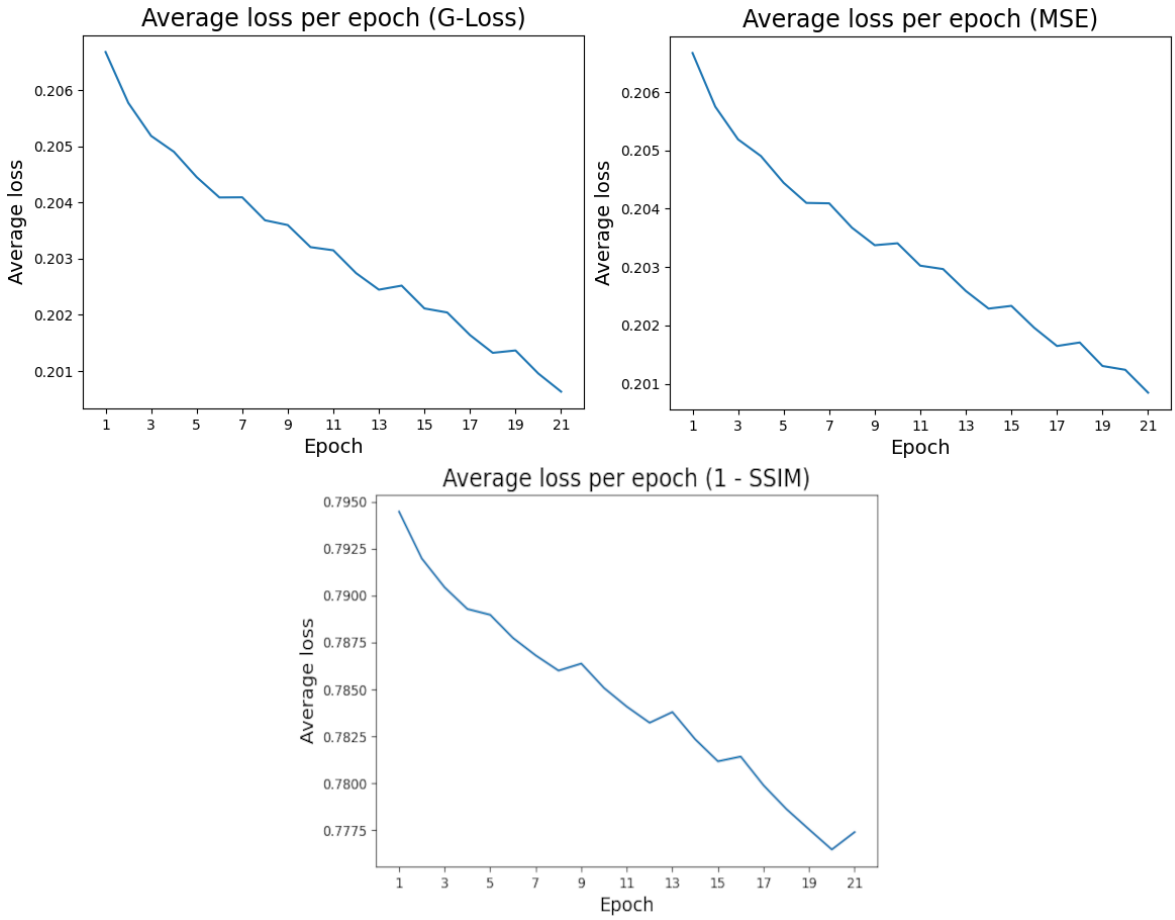

Figure 12: Loss during fine-tuning for each EDT model. Lower final loss values indicate better performance.

The fined-tuned models were validated against the validation set, see Table 2. The validation data show that all models benefited from training, with the exception of the PSNR for the model trained on SSIM. The difference between the model trained on G-Loss compared to trained on MSE was small, with the model trained on G-Loss having a slightly higher PSNR and SSIM than the other. Given that PSNR basically measures the pixel error, it is also logical that the model trained on MSE achieves a high PSNR. What is more unexpected is that the model trained on SSIM achieves the lowest SSIM score of the fine-tuned models.

Loss	**PSNR**	**SSIM**
Pre-trained	14.179	0.191
G-Loss	**15.398**	**0.252**
MSE	15.263	0.248
SSIM	12.83	0.2

Table 2: Average PSNR and SSIM scores over the entire validation dataset from the different EDT models' validation.

There is little perceptual difference between the G-Loss and MSE outputs, as seen in Figure 13. This is also shown in the similar validation results these two models achieved, as shown in Table 2. The SSIM output is much more noisy, which is also reflected in the lower PSNR value in the validation. It does however have higher contrast for the edges of the cells, which is important for the downstream segmentation applications.

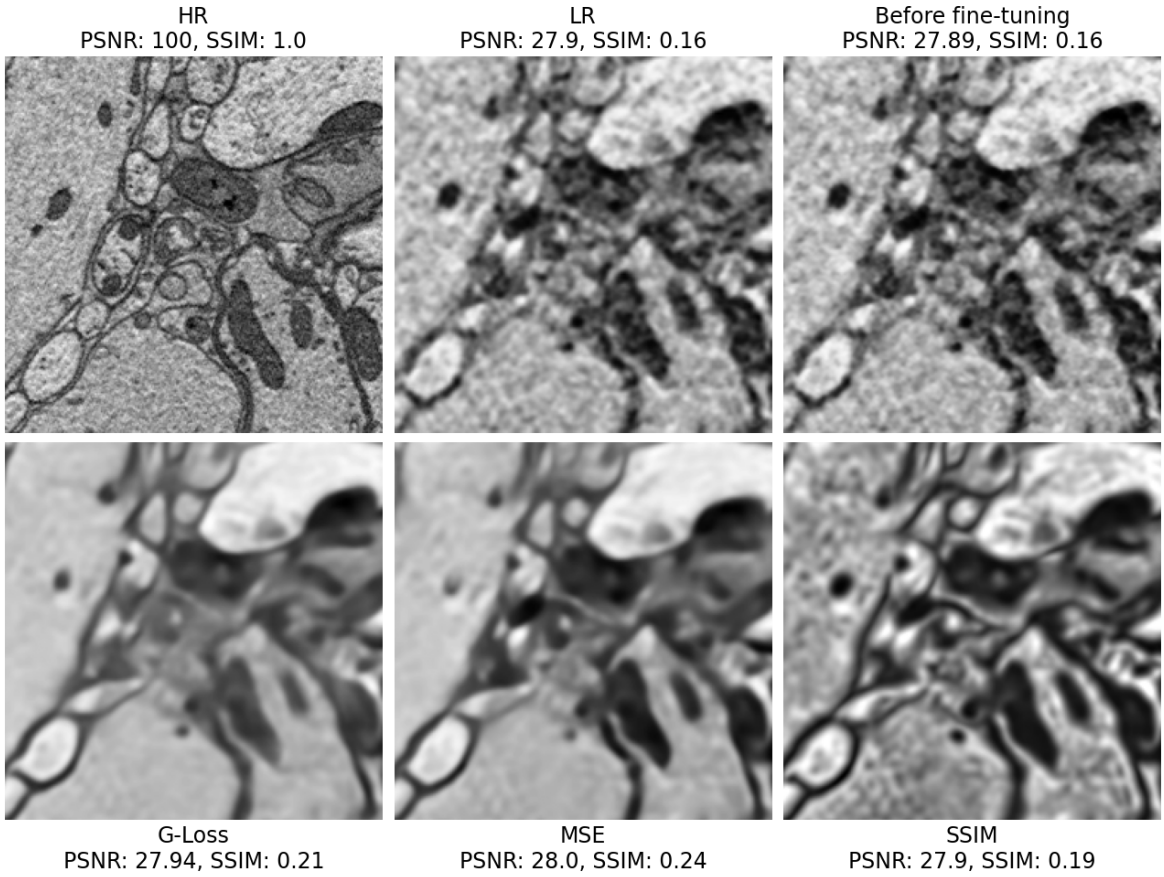

Figure 13: Comparison of EDT models before and after fine-tuning.

Upon inspection of details, the model trained on SSIM better captures sharper gradients in mitochondria and cell membranes as seen in the upper right corner of Figure 14. Even though it is better at capturing some of the desired features, this model also fails to capture details such as a large black feature in the bottom left corner that MSE also struggled with.

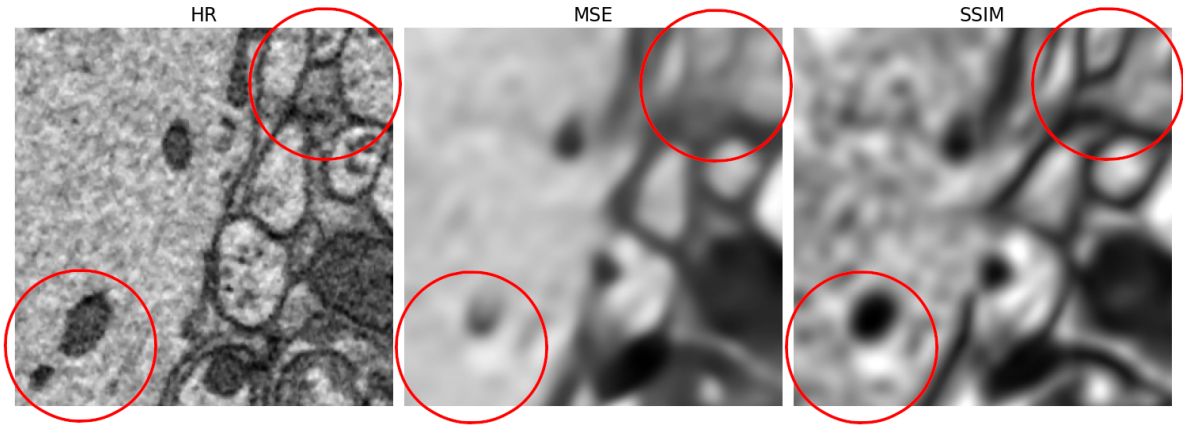

Figure 14: Close comparison of HR and some fine-tuned EDT models.

One point of note is that, despite training loss only decreasing slightly after the first epoch, suggesting early convergence, there were noticeable changes in the models' outputs after 21 epochs of fine-tuning, as seen in Fig 15. The models that were trained for 21 epochs have sharper edges than their less trained counterparts which makes it easier to identify the cells.

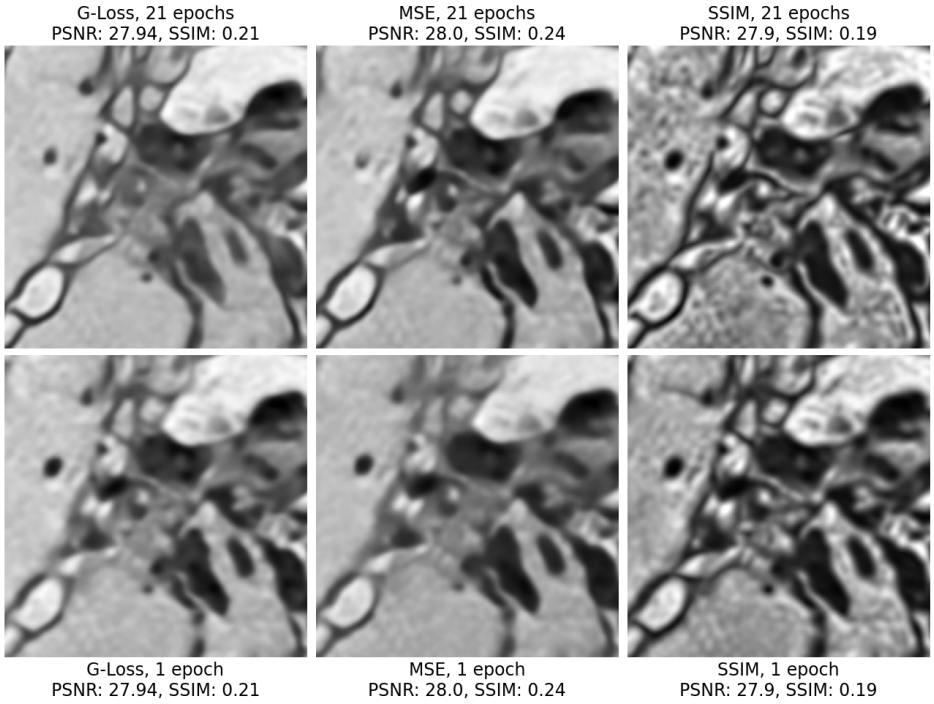

Figure 15: Comparison of fine-tuned EDT models after 1 epoch and 21 epochs of training.

Overall, while all the fine-tuned EDT models seem to struggle with areas with high feature density, they still capture most big features accurately and do not introduce many inaccurate artifacts.

6.4 Custom configured EDT

According to our evaluations metrics shown in Table 3, the quality of the custom model was lower than that of the fine-tuned EDT models. It did however produce improved upsampling compared to the bicubic upsampling, even though it was trained from the ground up on the limited amount of data originally meant for fine-tuning. Upon visual inspection it seems to better capture features that appear overly smooth in the output from the MSE-fine-tuned model (Figure 16).

PSNR	SSIM
27.9659	0.1574

Table 3: Average PSNR and SSIM scores over the entire validation dataset from custom configured EDT validation.

The customised model size was significantly larger than the original EDT configuration, 2.27 GB compared to 90 MB. This was mainly due to the increased input size as the self attention mechanism scales poorly. In the final iteration only one extra layer was added, as more than this didn't show any significance for training and only increased the models size.

6.5 Comparison of trained models

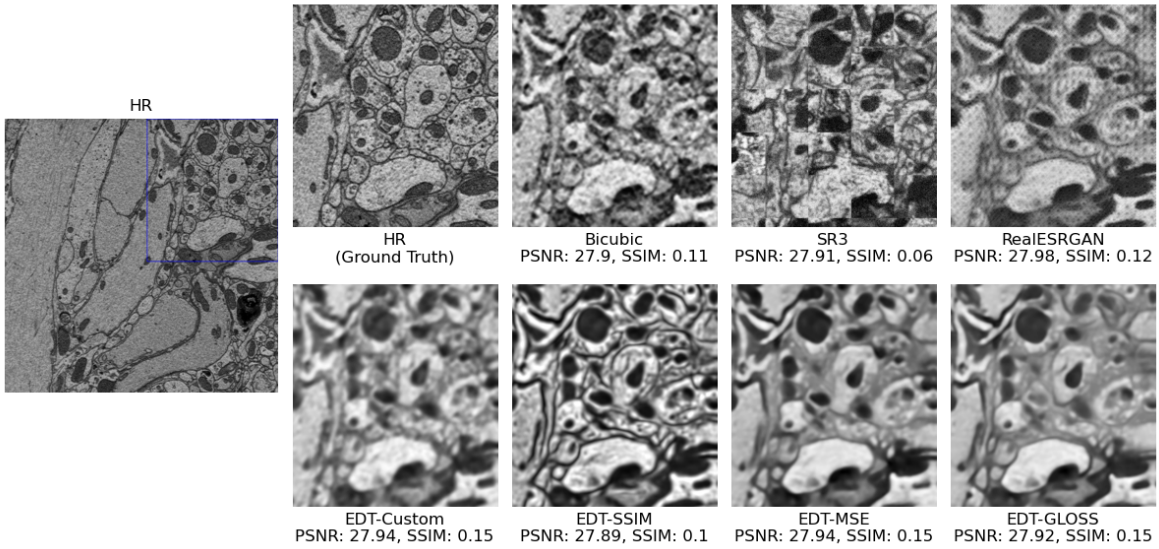

Figure 16: Final comparison

All models trained successfully and produced images closer to the desired HR than before training. It is clear when comparing the outputs side to side that different models and loss functions as expected produced different outputs given the same base information with some producing overly smooth features and others introducing incorrect hallucinations in shapes. Even though the visual impact of training is clear, the models' efficiency of upsampling the complex area visualized in Figure 16 was minimal. The hallucination-prone models introduced incorrect features while other models created overly smooth features.

6.6 Potential segmentation application

An area with small features (Figure 17) from the validation set was chosen to compare if the models improved segmentation over a baseline bicubic upsampling in an area when features are lost due to low resolution which would be the ideal application of SR.

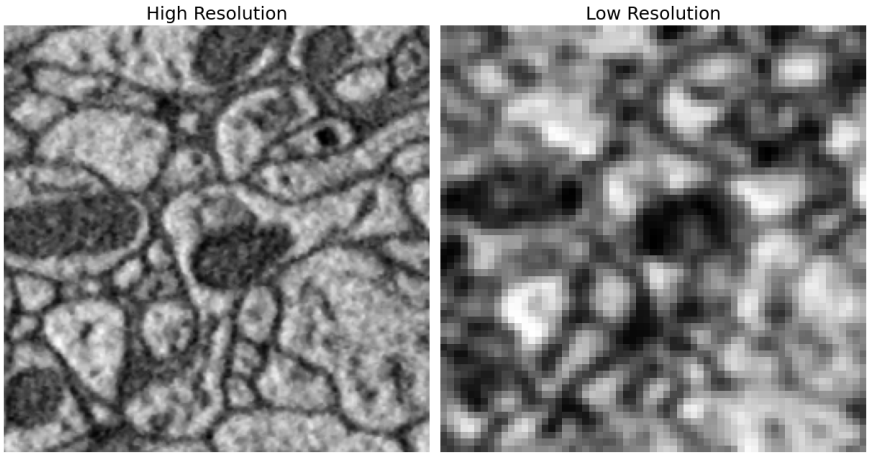

Figure 17: Low and high resolution pair for difficult segmentation.

Running the output of the EDT models with different loss functions through segmentation models and applying watershed with different thresholds show performance on par with, or worse than, standard bicubic upsampling. The noticeable exception being the model trained on SSIM loss that consistently outperforms the segmentation of the bicubic image. With a threshold for watershed set to 0.7 the segmentation of the SSIM output 300 percent better than base line.

Model	**Threshold: 0.1**	**0.3**	**0.5**	**0.7**	**0.9**
Bicubic	**6.37e-05**	3.35e-05	5.63e-05	1.37e-05	1.01e-04
EDT (custom)	5.97e-07	2.18e-06	1.67e-06	1.92e-06	2.03e-06
EDT (G-Loss)	1.10e-06	3.01e-06	5.47e-07	7.78e-07	8.26e-07
EDT (MSE)	1.14e-06	3.26e-06	7.57e-07	9.85e-07	1.05e-06
EDT (SSIM)	1.87e-06	**6.56e-05**	**1.80e-04**	**5.63e-05**	**1.44e-04**

Table 4: Similarity Scores (F1) for Different Models Across Thresholds.

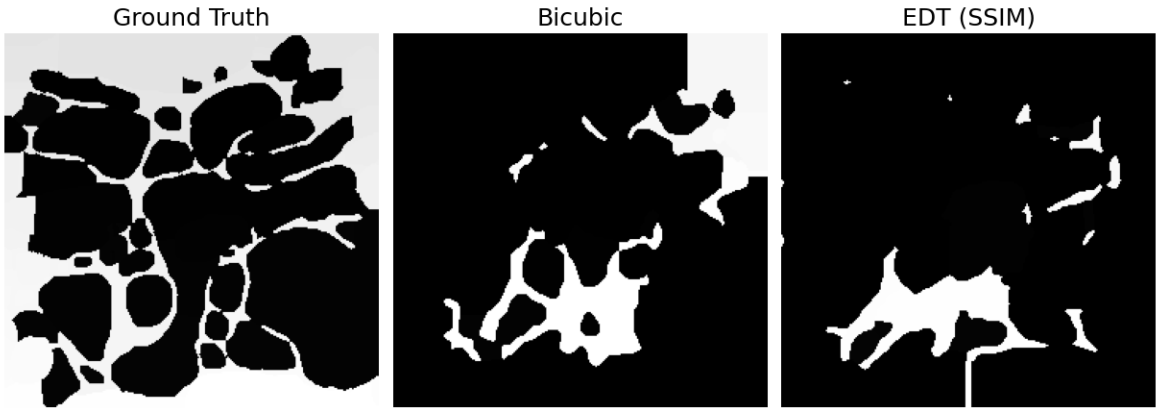

Figure 18: Labels from segmentation showing large merged areas for both the bicubic and SR image.

When comparing the segmented labels, it is not clear that the improvement is sufficient to decrease human proofreading, as seen in Figure 18. The output has a lot of merged labels in both the segmentation of the bicubic and the EDT output. This might be due to the model not adding any details relevant for the segmentation model to adjust the labels more precisely than for the LR image. It might also in part be a result of the model not being trained to segment upsampled images. Training the two models in coalition, might make the segmentation model more efficient in creating labels from upsampled data.

7 Discussion

In this study, we find that all evaluated models struggled with correctly introducing small features to upsampled images. Even if the models can be trained to produce textures and shapes that seem correct it is important to be mindful of this limitation. The models add information through hallucinations, potentially resulting in false features in the images. These features can be misleading or create false results if not handled correctly. Examples of this was very prevalent in the outputs of the GAN and diffusion models. Because of this, we cannot fully trust our models to produce realistic HR representations. If this issue is not properly addressed, the practical implications of applying image-generating AI to the field of biological sciences could be misleading. If it is assumed that all of the HR images output by the model are correct, it could lead to problems when evaluating the validity of reconstructed neurons at a later stage. This is not a consideration in SR:s regular use case as it often is used to only increase the perceived quality of an image where factual correctness is more of an afterthought. As deep learning models are black boxes, there are challenges in understanding where these inaccuracies were introduced given only the outcome.

As shown most prevalently by the GAN model, SR might in some cases also introduce artifacts. These alone might be easier to handle than incorrect hallucinations of for example membranes and mitochondria, but might lower the overall usefulness of the model if it makes other tasks such as segmentation more difficult. These might disappear through further training but would require more data and further fine-tuning.

For SR3 specifically, there was failure to maintain consistency across the borders of image patches. There are SR3 implementations with pre-trained weights available which handle larger image patches[5], which can mitigate the issue. Additionally, border consistency could be introduced as an additional component of the loss function in a future SR3 implementation adapted for this context.

We choose to use true LR and HR pairs in our training. This introduced the additional task of matching and annotating the image pairs. As stated earlier, the convention is to train models on downsampled HR. If we were to use this approach, a larger set of training data could have been used with the downside of it being further from the application. A larger data set could have allowed our more data intensive models, such as the diffusion model, more data to train on, and would maybe yielded better results. One argument against this is the fact that some objects in the HR and LR images are different due to being taken by two separate processes, thus making them difficult to upsample. This effect is clearly showing in the SR3 results with small features not being handled correctly. There are patterns in how these differences occur and affect the image but these would be completely missed if the model was only trained on artificially downsampled images.

Another choice in training that might have impacted the finetuned models performance, both in upsampling accuracy and segmentation potential, was to train on 2D-images. Since the images are collected from volumetric data, treating them as separate images removes information about features that span over several z-slices. This information might have improved the models in correctly upsampling features as they may appear clearer in neighbouring slices. Using 3D-image volumes might also have increased compatibility with the segmentation model that is trained on volumetric data.

If the upsampled images were to be used as an aid instead of a base for further processing, the issue of hallucination might be mitigated. Perceived increase of quality in larger features might be used as an aid for proofreading or studying larger structures where the SR images are not used as a base but rather a more easily interpreted reference. Giving larger basis for decisions than only using one

[5] `https://github.com/Janspiry/Image-Super-Resolution-via-Iterative-Refinement`

alternative, but being harder to combine with automated processes.

One aspect of SR that was not covered in our initial problem description is the matter of inference speed. For SISR tasks, this is often not brought up as an issue due to the limited amount of data it would be applied to. In the potential application of neural imagery, inference speed would matter greatly due to the amount of data that would have to be processed. This is an argument against the use architectures such as SR3 that have inference time much longer than that of the computationally less expensive transformer based models, such as EDT. What model is most suitable to train further for this application will depend on the use-case and amount of data that would need to be processed.

8 Conclusions and future work

Based on the results of this study, we can conclude that the most promising models for SBEM image SR is the transformer based EDT models. This is because the EDT models show limited hallucinations compared to the GAN or Diffusion based models, which is crucial for the final potential for application. As high contrast edges are important for successful segmentation, the use of a feature based loss metric, such as SSIM, show the most potential for producing useful images. In comparison, other tested loss functions produce overly smooth images.

As the end goal for the high-resolution images is neuron segmentation, the success rate of upsampling could be judged as the precision and recall of the segmented labels of the upsampled image compared to the labels of the segmented ground truth. The fine-tuned models in this case didn't show improvement over simpler bicubic upsampling, but this might partly be because of the model used for affinity and fragment detection used in the segmentation is trained solely on high resolution images. Training a segmentation model specifically on upsampled images using ground truth labels as reference, could potentially improve performance. Alternatively, a postprocessing step that adjusts image contrast and intensity ranges to more closely match those of high-resolution images could be explored to help the segmentation-model too better recognize structural features.

As this work aimed to explore the potential application of AI for SR for neuron imagery it did not explore optimization of promising models more than simple feature size adjustment. Further work could be done on balancing promising loss functions, such as SSIM, with other and adjusting the transformer based models core architecture further.

References

Briggman, Kevin L and Davi D Bock (Feb. 1, 2012). "Volume Electron Microscopy for Neuronal Circuit Reconstruction". In: *Current Opinion in Neurobiology*. Neurotechnology 22.1, pp. 154–161. ISSN: 0959-4388. DOI: 10.1016/j.conb.2011.10.022.

Chen, Hanting et al. (2021). "Pre-Trained Image Processing Transformer". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12299–12310. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Pre-Trained_Image_Processing_Transformer_CVPR_2021_paper.html (visited on 02/21/2025).

Deng, Jia et al. (June 2009). "ImageNet: A Large-Scale Hierarchical Image Database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

Ge, Lei and Lei Dou (June 1, 2023). "G-Loss: A Loss Function with Gradient Information for Super-Resolution". In: *Optik* 280, p. 170750. ISSN: 0030-4026. DOI: 10.1016/j.ijleo.2023.170750.

Heinze, Stanley and Uwe Homberg (2008). "Neuroarchitecture of the Central Complex of the Desert Locust: Intrinsic and Columnar Neurons". In: *Journal of Comparative Neurology* 511.4, pp. 454–478. ISSN: 1096-9861. DOI: 10.1002/cne.21842.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 6840–6851. URL: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html (visited on 05/01/2025).

Jurrus, Elizabeth et al. (Jan. 1, 2013). "Semi-Automated Neuron Boundary Detection and Nonbranching Process Segmentation in Electron Microscopy Images". In: *Neuroinformatics* 11.1, pp. 5–29. ISSN: 1559-0089. DOI: 10.1007/s12021-012-9149-y.

Kingma, Diederik P. and Jimmy Ba (Jan. 30, 2017). *Adam: A Method for Stochastic Optimization*. DOI: 10.48550/arXiv.1412.6980. arXiv: 1412.6980 [cs]. Pre-published.

Li, Wenbo et al. (Aug. 19, 2023). "On Efficient Transformer-Based Image Pre-Training for Low-Level Vision". In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. IJCAI '23. Macao, P.R.China, pp. 1089–1097. ISBN: 978-1-956792-03-4. DOI: 10.24963/ijcai.2023/121.

Liu, Ziwei et al. (2015). "Deep Learning Face Attributes in the Wild". In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738. URL: https://openaccess.thecvf.com/content_iccv_2015/html/Liu_Deep_Learning_Face_ICCV_2015_paper.html (visited on 05/12/2025).

Lowe, David G. (Nov. 1, 2004). "Distinctive Image Features from Scale-Invariant Keypoints". In: *International Journal of Computer Vision* 60.2, pp. 91–110. ISSN: 1573-1405. DOI: 10.1023/B:VISI.0000029664.99615.94.

Odena, Augustus, Vincent Dumoulin, and Chris Olah (Oct. 17, 2016). "Deconvolution and Checkerboard Artifacts". In: *Distill* 1.10, e3. ISSN: 2476-0757. DOI: 10.23915/distill.00003.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab et al. Cham: Springer International Publishing, pp. 234–241. ISBN: 978-3-319-24574-4. DOI: 10.1007/978-3-319-24574-4_28.

Saharia, Chitwan et al. (Apr. 2023). "Image Super-Resolution via Iterative Refinement". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4, pp. 4713–4726. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2022.3204461.

Seif, George and Dimitrios Androutsos (Apr. 2018). "Edge-Based Loss Function for Single Image Super-Resolution". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1468–1472. DOI: 10.1109/ICASSP.2018.8461664.

Sheridan, Arlo et al. (Feb. 2023). "Local Shape Descriptors for Neuron Segmentation". In: *Nature Methods* 20.2, pp. 295–303. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01711-z.

Van Ouwerkerk, J. D. (Oct. 1, 2006). "Image Super-Resolution Survey". In: *Image and Vision Computing* 24.10, pp. 1039–1052. ISSN: 0262-8856. DOI: 10.1016/j.imavis.2006.02.026.

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (visited on 05/12/2025).

Wang, Xintao, Liangbin Xie, et al. (2021). "Real-ESRGAN: Training Real-World Blind Super-Resolution With Pure Synthetic Data". In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1905–1914. URL: https://openaccess.thecvf.com/content/ICCV2021W/AIM/html/Wang_Real-ESRGAN_Training_Real-World_Blind_Super-Resolution_With_Pure_Synthetic_Data_ICCVW_2021_paper.html (visited on 05/05/2025).

Wang, Xintao, Ke Yu, et al. (2018). "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks". In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. URL: https://openaccess.thecvf.com/content_eccv_2018_workshops/w25/html/Wang_ESRGAN_Enhanced_Super-Resolution_Generative_Adversarial_Networks_ECCVW_2018_paper.html (visited on 02/21/2025).

Yang, Tianjie et al. (Aug. 31, 2021). "Advancing Biological Super-Resolution Microscopy through Deep Learning: A Brief Review". In: *Biophysics Reports* 7.4, pp. 253–266. ISSN: 2364-3439. DOI: 10.52601/bpr.2021.210019. PMID: 37287757.

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY