university of groningen / faculty of science and engineering / artificial intelligence

Semi-Supervised Contrastive Learning for Organoid Microscopy Image segmentation

Bachelor's Project Thesis

Julian Zwijghuizen, s3799492, j.zwijghuizen@student.rug.nl,
Main supervisor: Prof. Lambert Schomaker & Daily supervisor: Asmaa Haja

Abstract: This project employs a semi-supervised learning approach for segmenting organoid microscopy images, involving two distinct stages: pre-training and fine-tuning. The pre-training stage can be further divided into a global unsupervised contrastive learning stage and a local supervised contrastive learning stage. The objective is to investigate whether the semi-supervised approach outperforms the supervised approach. To evaluate this, the models are trained on varying amounts of data during the pre-training stage to determine the minimum quantity required to develop a model that outperforms the supervised learning approach. Additionally, the study examines whether two different loss functions (SSIM loss and SSIM-MAE loss) positively contribute to the segmentation performance when being used in the fine-tuning stage. Finally, the effect of freezing (vs not freezing) the U-Net encoder of the global stage when training on the local stage is examined in the context of segmentation performance. Results showed that the local stage of the semi-supervised learning approach has a more positive impact on the F1-scores as more data is used compared to the global stage, with F1-scores around 0.9. The SSIM(-MAE) is a better choice in terms of the organoids' coherent structure and the frozen models outperform the frozen models in capturing the larger organoids.

1 Introduction

Organoids are 3D structures of tissue that are created from multiple cells *in vitro* and are designed to simulate the properties and functions of a particular organ *in vivo* (Kretzschmar & Clevers, 2016). Organoids serve as a valuable tool for researchers because they can be used to discover biological mechanisms, model diseases, and to test the efficacy of treatments (de Souza, 2018). For organoids research, it is crucial to periodically observe microscopic images of the organoids to gather data on their physical- and growth patterns (Bian et al., 2021).

In recent years, the potential of deep learning in medical imaging tasks has been recognized (Suzuki, 2017). While convolutional neural networks have shown success in image classification tasks, such as distinguishing between cats and dogs (Deperlioglu, 2018) (Gavali & Banu, 2019), medical imaging tasks require more complex analysis such as image segmentation to locate tumors. Supervised learning approaches have proven to be successful in image segmentation (Oktay et al., 2018) (He et al., 2017). However, supervised learning approaches require all data to be labeled, which can be time-consuming and expensive. It also requires domain-specific knowledge for labeling the data and the expertise of machine learning experts or data scientists (Hu et al., 2021).

Semi-supervised learning offers a promising solution for these challenges by using the unlabeled data to reduce the reliance on labeled data (Bai et al., 2017). In (Hu et al., 2021), a semi-supervised learning approach is used to segment medical images. This model utilizes semi-supervised contrastive learning to efficiently label medical images of CT and MRI scans. According to the paper, prior to training a U-Net on segmenting the medical images, both the labeled and unlabeled data are used in a two-step pre-training stage. First, a self-supervised learning algorithm is utilized

to learn **global features** of the unlabeled data. Next, a supervised learning algorithm is employed to learn **local features** of the labeled data. The global features can be seen as the high-level features such as the organoids' shape, size and overall appearance, while the local features are the fine details and local structures within the images. By implementing the pre-training stage into the learning procedure, the model significantly achieves better performance than self-supervised learning techniques (Chaitanya et al., 2020). When fine-tuning the network, Hu et al. (2021) use a combination of the Dice loss and the BCE-loss, as it seems to give the best results. A Dice score of **0.866** is achieved when only 20% of the train data is labeled.

In this project, the goal is to use the above approach of semi-supervised learning to identify organoids and classify organoid images into "organoid" and "non-organoid" pixels. The model is built upon the existing model of Hu et al. (2021). The **first** contribution to the paper is to evaluate the effectiveness of the semi-supervised learning approach in organoid segmentation by comparing it to the supervised learning approach. This involves analyzing whether the pre-training stage, which is unique to the semi-supervised learning approach, has a positive impact on the segmentation task. The evaluation is conducted by training the models on varying amounts of data during the pre-training phase to determine the minimum amount required to develop a model that outperforms the supervised learning approach. A **second** contribution to the paper is to examine whether two different loss functions (SSIM loss and SSIM-MAE loss) positively affect organoid segmentation when being used in fine-tuning the model. A **third** contribution to the paper is to examine whether freezing (vs not freezing) the encoder of the global stage when training on the local stage has a positive impact on the final segmentation results.

This paper is divided into 6 sections. Section 2 offers a more detailed explanation of semi-supervised learning and the rationale behind using the SSIM(-MAE) loss in the semi-supervised learning approach. Section 3 provides an overview of the data that is being used, the different stages of the experiment and the loss functions that are being used. Section 4 discusses the experimental settings. Section 5 provides all results of the experiments being done and section 6 is associated with the conclusion and further work.

2 Related Work

Semi-supervised learning is a relatively new learning approach that gained popularity around the 00s (Seeger, 2000) (Nigam, 2001). Semi-supervised learning works with both labeled and unlabeled data to improve the performance of the model (Zhu, 2005). There are various methods used in semi-supervised learning, such as combining transfer learning with semi-supervised learning (Cai et al., 2021). Transfer learning is a learning technique where a pre-trained model is utilized as a starting point to solve a new task. The pre-trained model has already extracted useful features from the data, which can be leveraged during the training of the model for the new task (Weiss et al., 2016). Semi-supervised learning combined with transfer learning has proven to perform well (Xie et al., 2020).

Earlier, Wang et al. (2004) proposed the so-called SSIM loss as a new loss function for image quality assessment. The SSIM index takes into account that the human visual system is more sensitive to changes in the patterns and structures of an image rather than changes in the individual pixel intensities. It measures the structural similarity between the ground truth image and the generated image produced by the neural network. The paper showed that the SSIM index correlated better with human perception of image quality than traditional methods such as the mean squared error (MSE) and the peak signal-to-noise ratio (PSNR). Thereby, the SSIM loss takes into account the structural information and texture similarities of the image, making it more robust to noise and distortion (Tao et al., 2017). The SSIM index outperformed other state-of-the-art methods on a set of distorted images, and is since then being used for a wide range of applications in image processing and computer vision (Tao et al., 2017) (Zhao et al., 2015).
However, little research is available where the

SSIM loss has specifically been used for image segmentation. In (Huang et al., 2020), a U-Net architecture was used for segmenting liver and spleen CT images. A hybrid loss combination including the SSIM loss was designed to operate at three different levels; pixel-, patch- and map-level. This hybrid loss combination aimed to promote accurate segmentation of both large-scale structures and fine structures. The results of the paper indicated that the U-Net architecture in combination with the hybrid loss function outperformed all previous state-of-the-art approaches, such as the PSPNet and the DeepLab architecture. Ahamed & Imran (2022) employed different implementations of the U-Net architecture for segmenting cell nucleus and retinal vessel images. The U-Net architectures were used in combination with the proposed Image-to-Patch/Patch-to-Image (IPPI) framework, which consisted of an image segmentation branch and a patch segmentation branch. The image segmentation branch was responsible for segmenting the input image X while the patch segmentation branch was responsible for segmenting all non-overlapping patches x generated from input image X. To achieve global consistency at the local patch-level, the SSIM loss was calculated between Y' (i.e. the reconstructed image of the segmented patches x) and $\hat{Y}$ (i.e. the segmented input image X). Similarly, to ensure local consistency at the image-level, the SSIM loss was calculated between y' (i.e. the generated patches from the segmented input image X) and $\hat{y}$ (i.e. the segmented patches x). The results indicated that using the IPPI framework in combination with the SSIM loss outperformed the regular fully-supervised approach (i.e. without the IPPI framework), which indicated that the SSIM loss positively contributed to the segmentation performance.

As opposed to previous research, the idea in this project is that the SSIM loss can effectively elaborate on the neural network parameters learned from the pre-training stage of the semi-supervised approach.

3 Methods

The semi-supervised learning approach is predominantly based on the paper of Hu et al. (2021). The approach involves a three-stage training process, where all stages utilize (part of) the U-Net. The first two stages are considered as the **pre-training stage**, while the last stage is the segmentation **fine-tuning stage**. Section 3.1 explains the data used in the project. Section 3.2 and 3.3 provide a deeper insight in the U-Net and SIMCLR algorithm as these are important concepts for understanding the full model. Section 3.4 explains the three different stages used in the model. Section 3.5 provides an overview of the different loss functions used in the fine-tuning stage.

3.1 Data

The data used in this project consists of grey organoid images obtained from the University Medical Center Groningen (UMCG). Images were captured at five different time points within a time span of 96 hours, with a 24-hour interval between each time point. From all these images, 10 CZI files were created. A CZI file is a file format used for saving 3D microscopic images which are build up from 2D horizontal slices taken at different depths in the culture (Haja et al., 2023). As most of the organoids were present in the middle stacks, only these stacks were used.
The initial images have a resolution of around 3830 x 2900 pixels. To get a good balance between the information content and computational cost, the images were divided into smaller crops of 320 x 320 pixels using a sliding window technique. To generate more data, the sliding window technique was only moved a fraction of the image size which led to overlapping regions between images. Images that contained less than 3% of organoid pixels were excluded from the data set. To further increase the data size, the resulting images were rotated by 90°, 180° and 270°.

3.2 U-Net

U-Net is a convolutional neural network architecture designed for biomedical image segmentation. It has a U-shape structure, hence its name. The network consists of a contracting path that downsamples the image and a symmetric expanding path that restores the resolution for pixel-wise prediction (Ronneberger et al., 2015). Since the initial deployment of the U-Net, the architecture has undergone

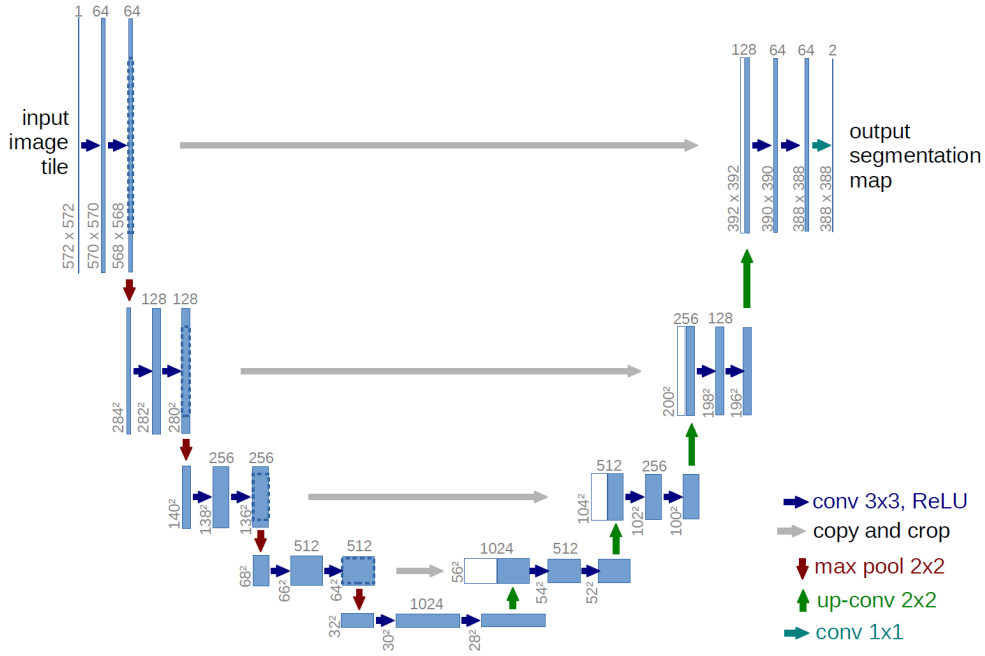

Figure 3.1: The original U-Net structure as used in the paper of Ronneberger et al. (2015)

continuous development, resulting in various forms of the network being used (Wu et al., 2022) (Zhou et al., 2018) (Zhang et al., 2021). Figure 3.1 shows the original U-Net structure as used in the paper of Ronneberger et al. (2015).

3.3 SIMCLR

SIMCLR (i.e. "Simple Contrastive Learning") is a self-supervised framework, which can be utilized for training neural networks on large unlabeled data sets to learn representations that can be used for downstream tasks (Chen et al., 2020). Figure 3.2 shows the procedure of the SIMCLR algorithm. An image x is transformed to two different augmented images $\tilde{x}_i$ and $\tilde{x}_j$. The goal of the SIMCLR algorithm is to maximize agreement using a contrastive loss between the output vectors z_i and z_j by training the base encoder $f(\cdot)$ and projection head $g(\cdot)$. As explained by Chen et al. (2020), the SIMCLR algorithm uses a contrastive loss function that tries to maximize similarity between positive pairs and minimize similarity between negative pairs in the vector space. A positive pair refers to two different versions of the same input image that are obtained through different data augmentation transformations, while a negative pair refers to two independent augmented versions of different input images.

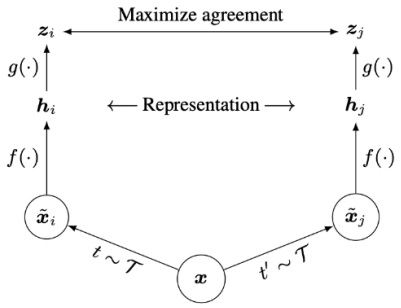

Figure 3.2: A simple framework illustrating the SIMCLR algorithm as described in the paper of Chen et al. (2020). A series of data augmentations from τ are applied to a single image x twice to generate two separate versions $\tilde{x}_i$ and $\tilde{x}_j$ of the image x. These two images $\tilde{x}_i$ and $\tilde{x}_j$ are then passed through a base encoder network f(·) and a projection head g(·), which produces two corresponding vectors z_i and z_j. The goal of the training process is to maximize agreement using a contrastive loss between the output vectors z_i and z_j by training the base encoder and projection head.

3.4 Stages of the model

The first two stages are largely similar to the paper of Hu et al. (2021), but they will still be discussed for clarification.

3.4.1 First stage: Global unsupervised contrastive learning

The first stage focuses on global unsupervised contrastive learning. The process largely follows the structure of the SIMCLR algorithm. During this stage of the experiment, the encoding path of the U-Net network is utilized as a feature extractor and the Multi-layer perception (MLP) head is used to convert the U-Net output into vector representations of the images. Figure 3.4 visualizes the structure of the this stage. The four green blocks resemble the encoding layers of the U-Net and the *head* is the MLP. Only the procedure of the positive pair is shown in this image but in reality the batch consists of both positive- and negative pairs. To train our model using the SIMCLR algorithm, both positive and negative examples are required. To obtain the positive and negative examples, a sequence of augmentations is applied twice for each image in a given batch B. The augmentations for this stage consist of:

1. Random translation or random zooming
2. Random brightness adjustment
3. Random Gaussian blurring
4. Random Gaussian noise or random salt and pepper noise

The augmentations are applied in a random order but translation/zooming is always done at the first step. Figure 3.3 illustrates the different augmentations that are utilized in this project. Figure 3.5 shows an example of a positive pair and

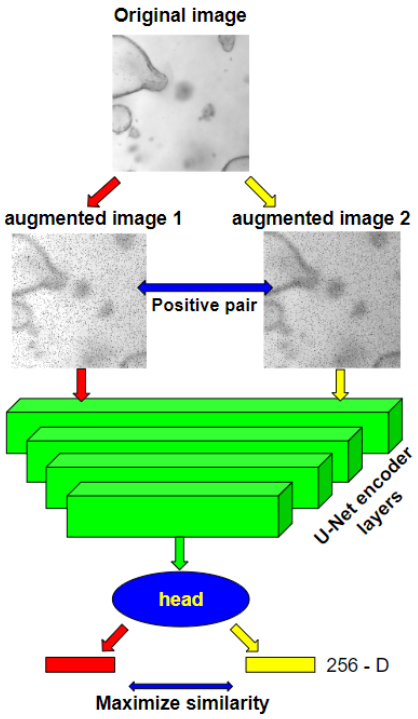

Figure 3.4: The structure of the global unsupervised contrastive learning stage. The four green blocks resemble the global U-Net encoding layers. The head is the MLP.

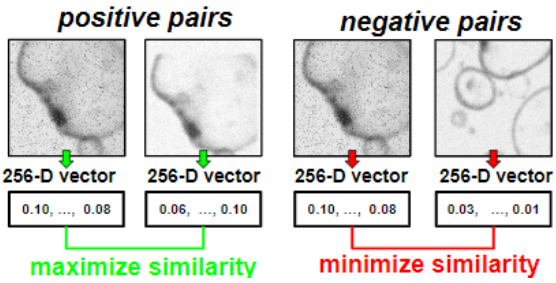

Figure 3.5: Two examples of a positive pair and two examples of a negative pair in the global unsupervised contrastive stage.

a negative pair of the data being used in the global supervised contrastive learning part. A positive pair is a pair of augmented images which are both derived from the same ground truth. A negative pair is a pair of augmented images which are both derived from different ground truths. The first stage is given the name "global" because the comparison is done between the two 256-D output

Figure 3.3: The different data augmentations that are used in this project.

vectors, which represent the whole images.

To put it formally, let $B = \{x_1, x_2, ...x_b\}$ represent an input batch, in which x represents the original image (i.e. ground truth). By applying two random sequences of augmentations on each image x in the given batch of size b, two augmented data sets $C = \{a_1, a_2, ...a_b\}$ and $D = \{a_1, a_2, ...a_b\}$ are generated of size b where a represents the augmented image of x. The two augmented sets are combined into one augmented dataset $E = \{a_1, ..., a_{2b}\}$ of size $2b$. Let $a_i, i \in I = \{1..2b\}$ be the index of an augmented image in the augmented set E. For each image a_i, j(i) represents the index of the other augmented image in augmented image set E that was derived from the same image set B as a_i. Hence, a_i and $a_j(i)$ form a positive pair. The formula for the global contrastive loss is

$$L_g = -\frac{1}{|A|}\sum_{i\in I}\log\frac{e^{sim(z_i,z_{j(i)})}/\tau}{\sum_{k=1\ [k\neq i]}^{2b} e^{sim(z_i,z_k)}/\tau} \quad (3.1)$$

where τ is the temperature and z is the normalized output of the MLP, i.e. $z_i = g(f(a_i))$ and $z_{j(i)} = g(f(a_{j(i)}))$, $sim()$ represents the dot similarity between the vectors, f is the U-Net encoder output and g is the MLP.

3.4.2 Second stage: Local supervised contrastive learning

The second stage of the semi-supervised learning approach concentrates on local supervised contrastive learning. This stage follows a structure similar to the first stage, with the difference that now the full U-Net is employed to learn about the local features of the images. Transfer learning is applied by loading the saved encoding weights from the U-Net of the first stage into the U-Net of this stage. Figure 3.6 visualizes the structure of this stage. The green blocks resemble the full U-Net and the head is the MLP. The augmentations of this stage consist of:

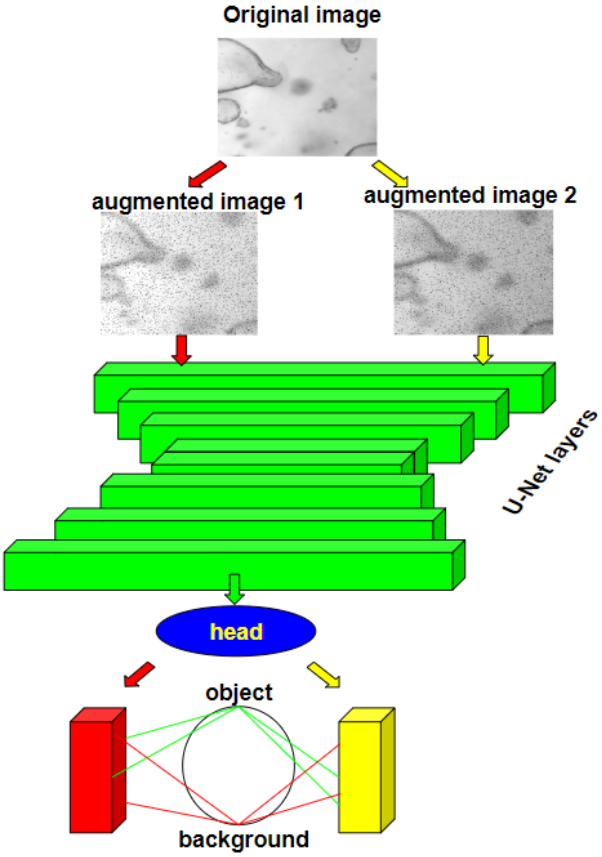

Figure 3.6: The structure of the local supervised contrastive learning stage. The green blocks resemble the full U-Net layers. The head is the MLP. The object represents the organoid pixels and the background represents the background pixels

1. Random brightness adjustment

2. Random Gaussian blurring

3. Random Gaussian noise or random salt and pepper noise

The positive- and negative pairs are different from the previous stage. A pair is considered positive when two pixels share the same value (i.e. both "organoid" pixels or both "non-organoid" pixels). Conversely, a pair is considered negative when two pixels contain different values. Figure 3.7 shows an example of both a positive pair and a negative pair. Every pixel in the batch is being compared to all other pixels in the batch. This means that both the positive- and negative set consist of within pairs (i.e. positive- and negative pairs within an image) and between pairs (i.e. positive- and negative pairs between images). Formally, the positive set P can be defined as

$$P = \{(p_1, p_2) \mid p_1, p_2 \in I, \text{loc}(p_1) \neq \text{loc}(p_2), \text{val}(p_1) = \text{val}(p_2)\} \quad (3.2)$$

where p_1 and p_2 represent all pixels in image set I, *loc()* represents the pixel location and *val()* represents the pixel annotation. Since the pixels can only contain either zero (background) or one (organoid), the positive set consists of positive "background" pairs and positive "non-background" pairs. Similarly, the negative set N can be defined as

$$N = \{(p_1, p_2) \mid p_1, p_2 \in I, \text{loc}(p_1) \neq \text{loc}(p_2), \text{val}(p_1) \neq \text{val}(p_2)\} \quad (3.3)$$

The negative set N will therefore only consist of pairs with one "background pixel" and one "non-background" pixel. Translation and zooming are not used in this stage for the reason that positive non-background pixels are already sparse (the images mainly consist of background pixels) and translation and zooming can cause less "organoid" pixels.

Similar to the first stage, each image is undergoing a random sequence of the augmentations twice. a_i represents the augmented image at index i in A and $h(x_i) = g(f(a_i)))$ represents the output feature map after going through the U-Net network f and the MLP g. The formula for the local contrastive loss of feature map $h(a_i)$ is

$$Loss(a_i) = -\frac{1}{|\Omega|} \sum_{(u,v)\in\Omega} \frac{1}{P(u,v)} log \frac{\sum_{(u_p,v_p)\in P(u,v)} exp(h_{u,v} \cdot h_{u_p,v_p} / \tau)}{\sum_{(u_f,v_f)\in N(u,v)} exp(h_{u,v} \cdot h_{u_f,v_f} / \tau)} \quad (3.4)$$

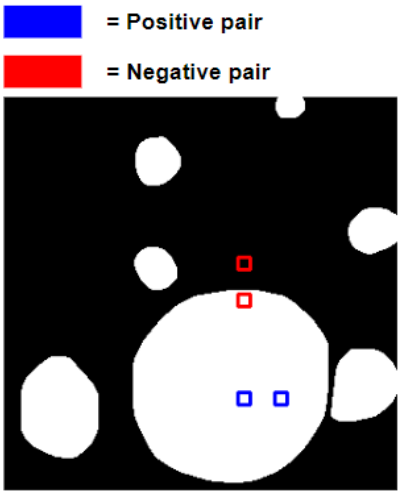

Figure 3.7: Example of a positive pair and a negative pair in the local supervised contrastive stage.

where (u,v) represents the column and row indices of the feature map, respectively. τ is the temperature. P(u,v) and N(u,v) denote the sets of similar and dissimilar features, respectively. Ω is the total number of non-background pixels used in the loss computation. Since the goal of the neural network is to learn about the structure of the organoids, positive pairs that consist of "background" pixels are filtered out of the positive set. The formula for the total local loss can be defined as

$$L_l = \frac{1}{|A|} \sum_{a_i \in A} loss(a_i) \quad (3.5)$$

where A is the augmented image set.

3.4.3 Third stage: Segmentation fine-tuning

The final stage of the semi-supervised learning approach involves segmentation fine-tuning, which utilizes the pre-trained U-Net model from the pre-training stage. Figure 3.8 visualizes the structure of this stage. Again, the green blocks resemble the U-Net and the head is the MLP. This fine-tuning process is employed for the purpose of performing semantic segmentation tasks on an organoid image dataset. In this stage, no augmentation techniques are used. The augmentation techniques can cause distortions/alterations that compromise the

semantic integrity of the ground truth images and accurate segmentation is needed in this stage.

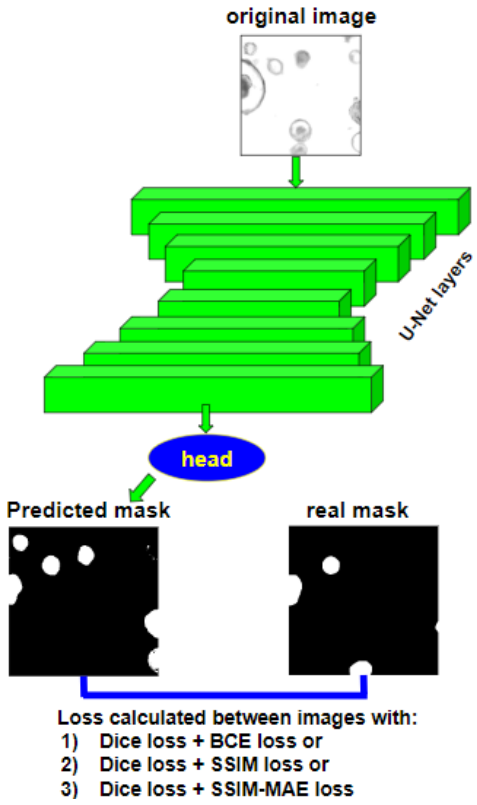

Figure 3.8: The structure of the segmentation fine-tuning stage. The green blocks resemble the full U-Net layers. The head is the MLP.

3.5 Loss-functions

To optimize the performance of the model for semantic segmentation, it is trained using three different loss functions. The first loss function consists of a combination of the Binary Cross-Entropy (BCE) loss and the Dice loss. Additionally, two other loss functions are employed with the combination of the Dice loss, namely the structural similarity (SSIM) loss and the SSIM loss combined with mean absolute error (MAE). After training the model with the three aforementioned loss functions, the resulting three models are compared to each other using the F1-score metric. This comparison allows for the determination of the loss function that most effectively optimizes the model's performance for the segmentation tasks.

3.5.1 BCE loss

The Binary Cross-Entropy loss is a widely used loss function for classification tasks. Cross-entropy is defined as the measure of the difference between two probability distributions (Jadon, 2020). The formula for the BCE loss is defined as

$$L_{BCE}(y, \hat{y}) = -(y \cdot log(\hat{y}) + (1-y) \cdot log(1-\hat{y})) \quad (3.6)$$

where y is the ground truth and $\hat{y}$ is the predicted mask.

3.5.2 Dice loss

The Dice coefficient is widely used for tasks where the similarity needs to be calculated between the predicted image and its real mask (Jadon, 2020). The formula for the dice loss is defined as

$$L_{dice}(y, \hat{y}) = 1 - 2 \cdot \frac{y \cap \hat{y}}{y + \hat{y}} \quad (3.7)$$

where the numerator represents the union of the ground truth and the predicted mask, and the denominator represents the sum of the ground truth and the predicted mask.

3.5.3 SSIM loss

The Structural Similarity Index Measure is a method to calculate the structural similarity of the ground truth and the predicted mask (Nilsson & Akenine-Möller, 2020). The formula for the SSIM loss is defined as

$$L_{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (3.8)$$

where μ_x and μ_y denote the mean of the input image x and target image y respectively, σ_x and σ_y denote the standard deviations of x and y and σ_{xy} is the covariance of x and y. C1 and C2 are small constants added to avoid division by zero.

3.5.4 MAE

The Mean Absolute Error measures the error between the ground truth and the predicted mask. The formula for the MAE is defined as

$$L_{MAE}(x, y) = \frac{\sum_{i=1}^{N} |y_i - x_i|}{n} \quad (3.9)$$

where y is the prediction, x is the true value and n is the total number of data points.

4 Experiment

4.1 Dataset

The dataset contains a collection of grey images, each of which contains one or more organoids. The images are accompanied by masks that identify the pixels corresponding to the organoids and non-organoids. The dataset consists of approximately 80,000 crops and masks. The dataset of 80,000 crops and masks was partitioned into three distinct sets for different stages of training and testing. Specifically, around 35,000 crops were set aside for the pre training phase, while another 35,000 crops were reserved for the segmentation fine-tuning stage. The remaining 10,000 crops were used for assessing the performance of the model. The resolutions of the crops are 320 x 320 pixels.

4.2 Pre-training stage

33,680 images and the corresponding masks were reserved for the pre-training stage. Since the pre-training stage itself consists of two stages, 16,840 images were set aside for each stage. To manage the computational costs and ensure reasonable training times, a maximum of 60% of the images (i.e. 10104 images) was used for both stages.

4.2.1 Global unsupervised contrastive learning stage

In this stage, the neural network consists of the U-Net encoder and the MLP. During this stage, the neural network was trained six times, each time using a different percentage of the dataset. Specifically, the training process involved training the neural network with subsets of increasing size, ranging from 10% (1684 images) to 60% (10,104 images) of the entire dataset. Each time the neural network was trained, 5-fold cross validation was used to obtain a more reliable estimate of the model's performance on unseen data. This resulted in 6 x 5 = 30 models. Table 4.1 shows all the values of the different parameters used in this stage.
The neural network receives two tensors as input, denoted as $x_i s$ and $x_j s$. Each element x_i and x_j, where $i == j$, are derived from the same image, and thus form a positive pair. Both tensors possess a shape of (50,1,320,320), where 50 corresponds to the batch size, 1 corresponds to the greyscale channel, and (320 x 320) represents the image size. After being processed through the encoding component of the U-Net architecture, the output dimensions of both $x_i s$ and $x_j s$ become (50,512,320,320). Subsequently, these tensors are directed to the MLP, which employs a 2D average pooling layer followed by two fully connected layers to transform the neural network outputs into vector representations of the images. The final output of both $x_j s$ and $x_j s$ will be (50,256) and these vector representations are used for computing the global contrastive loss.

Parameter	Value
Learning rate	0.0001
Batch size	50
Epochs	50
Temperature	0.5

Table 4.1: Model parameters of the global unsupervised contrastive learning stage with their values

4.2.2 Local supervised contrastive learning stage

All 6 models from the first stage needed to be trained on the second stage. The best fold was chosen for each of the six models of the first stage. The optimal fold was determined by selecting the model with the lowest average validation loss across the epochs. The saved weights of each best fold were loaded into the encoding path of the full U-Net. Then, the model was trained again on the six models from the local supervised contrastive learning stage, each with five folds. As a result, 6x6x5 = 180 models were created. Table 4.2 shows all values of the different parameters used in this stage.

4.3 Segmentation fine-tuning stage

The best folds from each of the 36 models of the second stage were chosen for the last training-stage. Again, the optimal fold was determined by selecting the model with the lowest average validation loss across the epochs. In the last stage, all 36 models were trained using the 3 losses defined in section

Parameter	Value
Learning rate	0.0001
Batch size	8
Epochs	8
Block size	16
Temperature	0.07

Table 4.2: Model parameters of the local supervised contrastive learning stage with their values

3.5, namely Dice loss + BCE loss, Dice loss + SSIM loss and Dice + SSIM-MAE loss. Hence, 36 x 3 = 108 models were created. The resulting three models were compared to each other using the F1-score metric.

5 Results

5.1 Semi-supervised learning approach vs supervised learning approach

The **first** research question that needs to be answered is whether the semi-supervised learning approach outperforms the supervised learning approach. And if so, what is the minimum amount of data required in the pre-training stage to develop a model that outperforms the supervised learning approach?

In order to test the effectiveness of the pre-training stage on the segmentation performance, the F1-scores of the pre-trained models (i.e. semi-supervised approach) are compared to the non pre-trained models (supervised approach). Figures 5.1, 5.3 and 5.5, each trained with a different loss combination, contain six plots that show the progression of the F1-scores with increasing data proportions for the global stage. The data proportion for the local stage remains fixed. By keeping the data proportions for the local stage constant, it is possible to determine whether increasing the data proportion for the global stage has a positive impact on the F1-score. Figures 5.1, 5.3 and 5.5 all show the same pattern, in which more data used in the global stage does not increase the F1-score. Instead the F1-score remains unchanged when more data is reserved for the global stage. This is likely due to the fact that in a "global sense" the organoid images in the dataset are already quite similar to each other. Therefore, by allocating more data to the global stage, the data set will not be more diverse and will not lead to a better F1-score.

Figures 5.2, 5.4 and 5.6, each trained with a different loss combination, contain six plots that show the progression of the F1-scores with increasing data proportions for the local stage. The data proportion for the global stage is held constant to be able to determine whether increasing data proportions for the local stage has a positive impact on the F1-scores. In these Figures, it can be seen that as more data is used for the local stage there is an increase in the F1-score. Thereby, most pre-trained models outperform the supervised model when 60% of the data is used for the local stage as opposed to 10%. Since the local stage is trained with supervision, adding more data will lead to a more generalized model, which is more robust to over-fitting.

To conclude, it is difficult to determine the specific data amounts needed for the pre-training
stage to outperform the supervised model. However, the local stage seems to have a more positive impact on the F1-scores as more data is used compared to the global stage.

5.2 SSIM(-MAE) loss vs BCE-loss

The **second** research question is to examine whether the SSIM loss and the SSIM-MAE loss lead to better organoid segmentation as opposed to the BCE loss when being used for fine-tuning the model. Table 5.1 shows the mean F1-score and the mean standard deviation of all 36 pre-trained models trained with the three different loss combinations. Based on the F1-scores, the BCE and Dice loss combination outperforms the SSIM(-MAE) and Dice loss combinations. However, the F1-scores alone do not provide a comprehensive understanding of the overall performance. Figure 5.7 shows three randomly selected organoid images from the test set, their masks and the predictions of the three fine-tuned models with the highest F1-scores. Although the three predicted masks

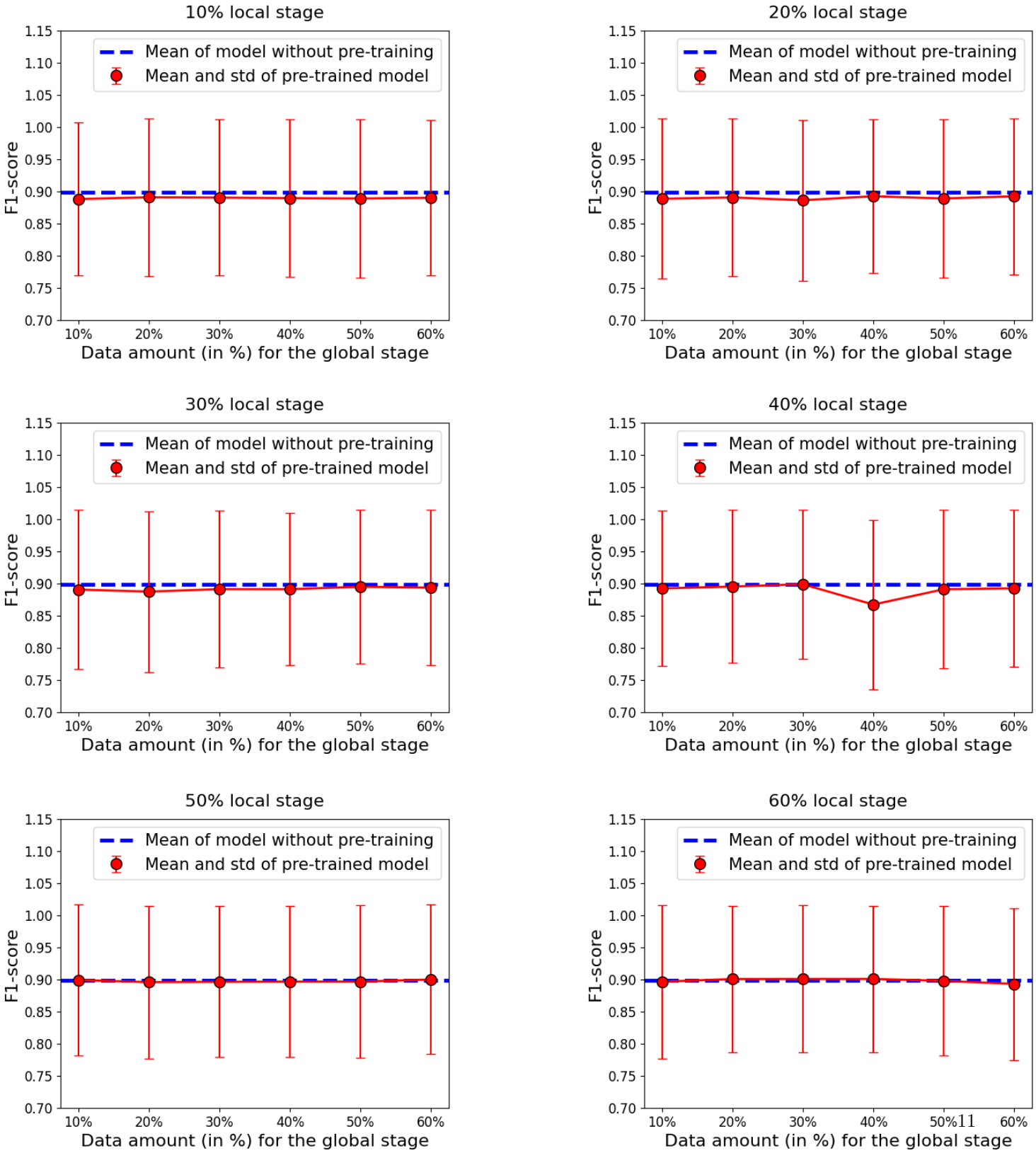

Figure 5.1: Progression of F1-score with increasing data usage in the global stage, with constant proportion for the local stage, and trained with the combination of the BCE loss and Dice loss. The blue lines indicate the mean of the supervised model without the pre-training stage. The red lines indicate the mean of the semi-supervised models and the vertical lines indicate their corresponding standard deviations.

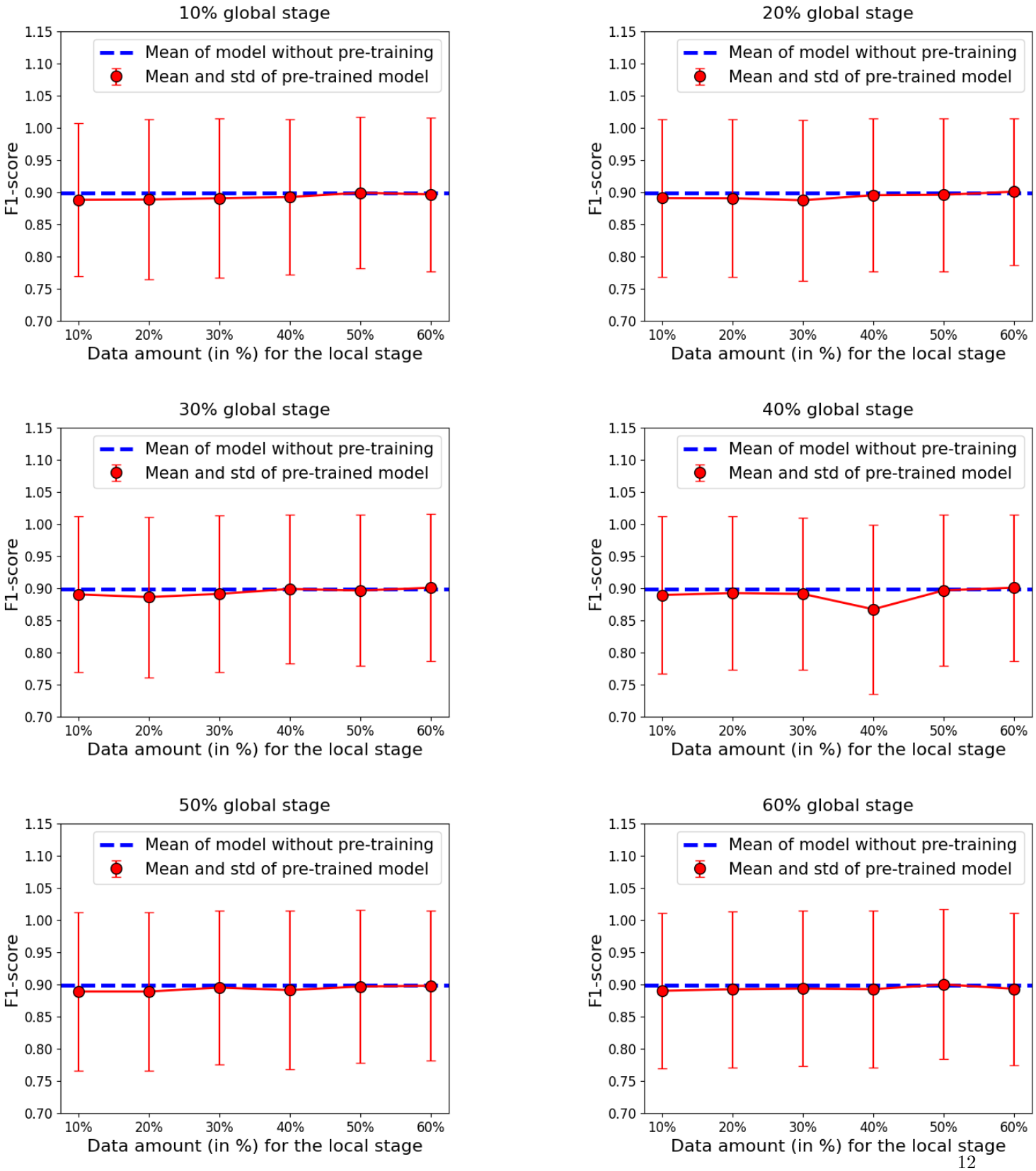

Figure 5.2: Progression of F1-score with increasing data usage in the local stage, with constant proportion for the global stage, and trained with the combination of the BCE loss and Dice loss. The blue lines indicate the mean of the supervised model without the pre-training stage. The red lines indicate the mean of the semi-supervised models and the vertical lines indicate their corresponding standard deviations.

(Semi-)supervised results: SSIM loss and Dice loss

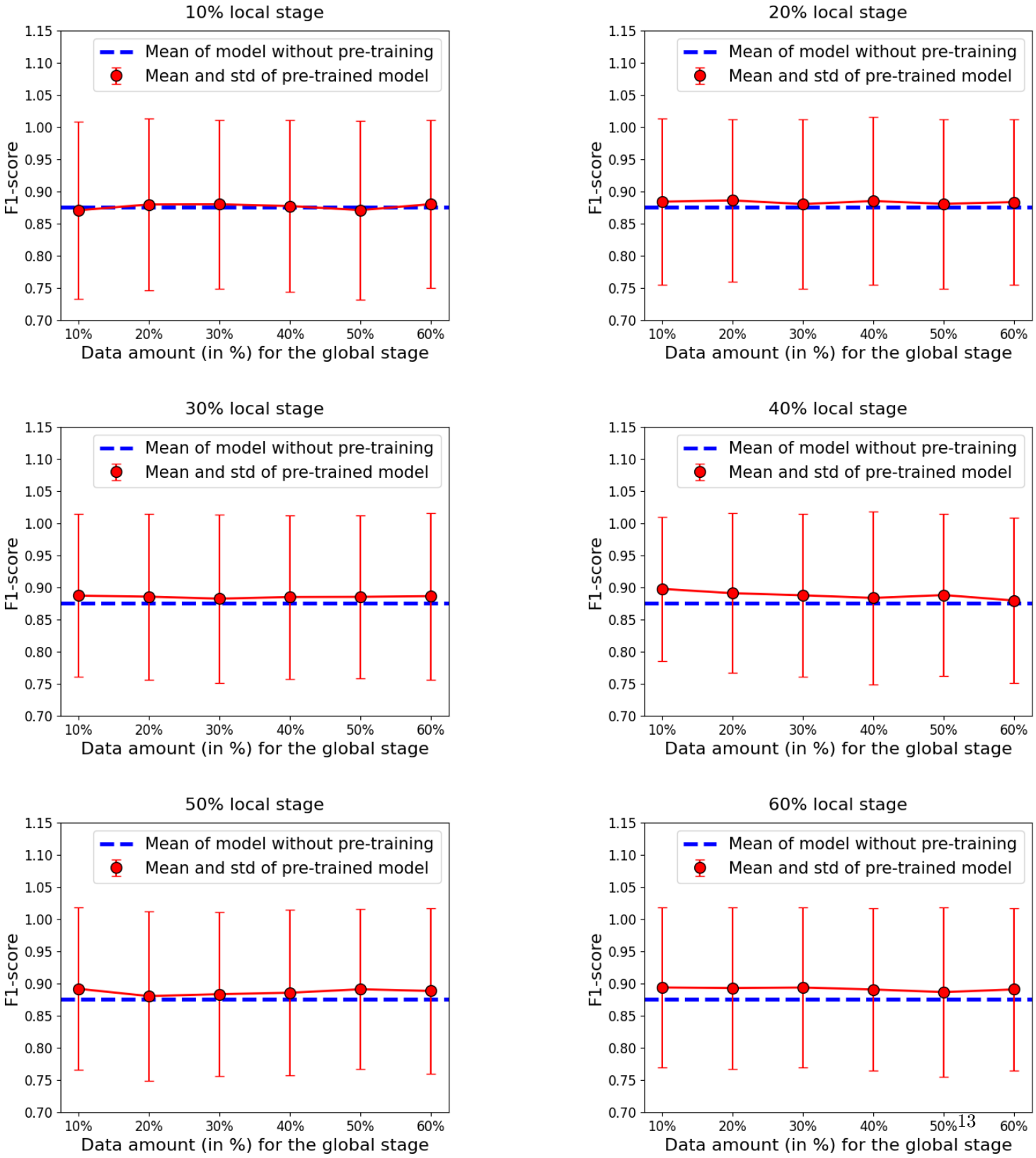

Figure 5.3: Progression of F1-score with increasing data usage in the global stage, with constant proportion for the local stage, and trained with the combination of the SSIM loss and Dice loss. The blue lines indicate the mean of the supervised model without the pre-training stage. The red lines indicate the mean of the semi-supervised models and the vertical lines indicate their corresponding standard deviations.

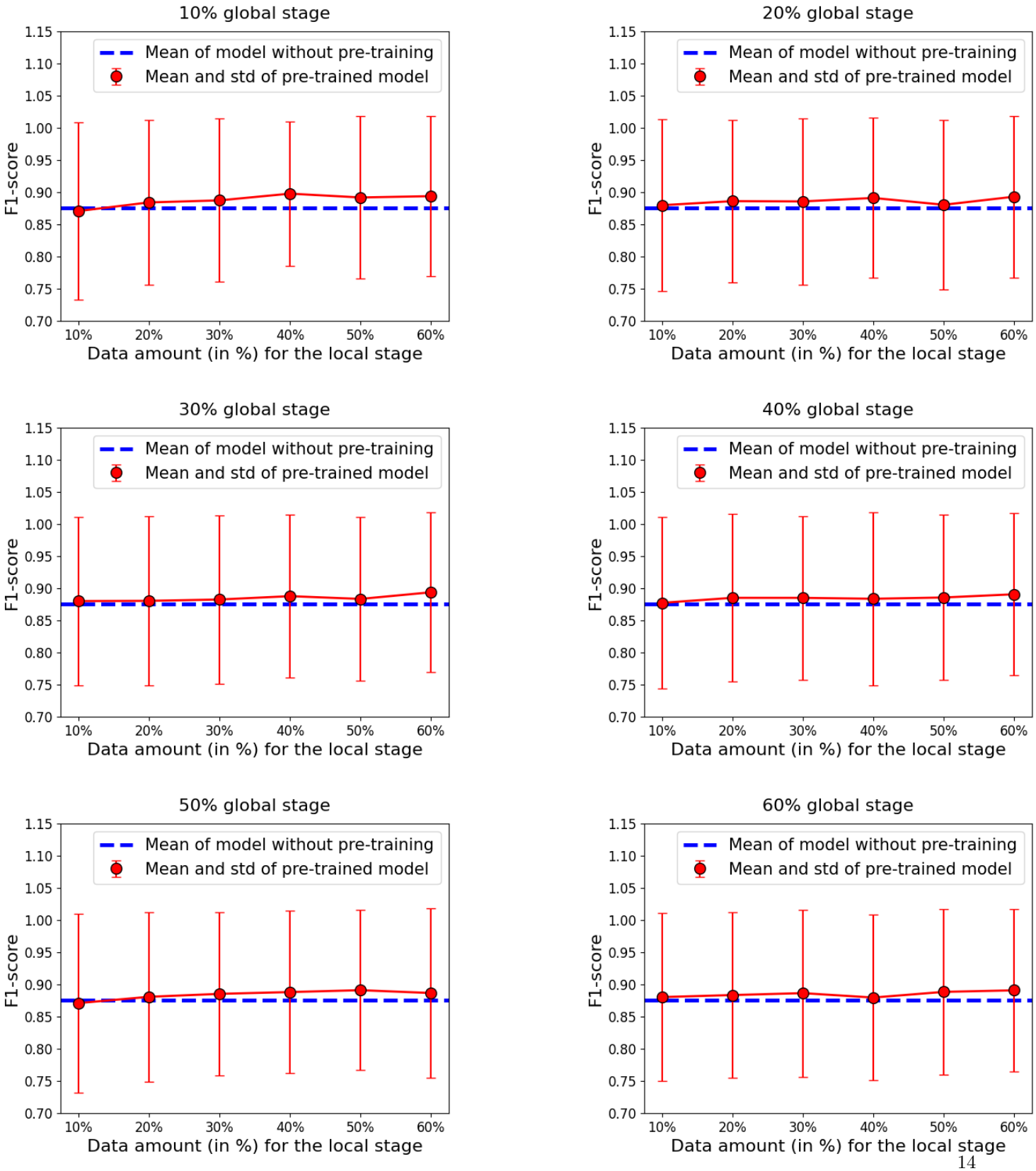

Figure 5.4: Progression of F1-score with increasing data usage in the local stage, with constant proportion for the global stage, and trained with the combination of the SSIM loss and Dice loss. The blue lines indicate the mean of the supervised model without the pre-training stage. The red lines indicate the mean of the semi-supervised models and the vertical lines indicate their corresponding standard deviations.

(Semi-)supervised results: SSIM-MAE loss and Dice loss

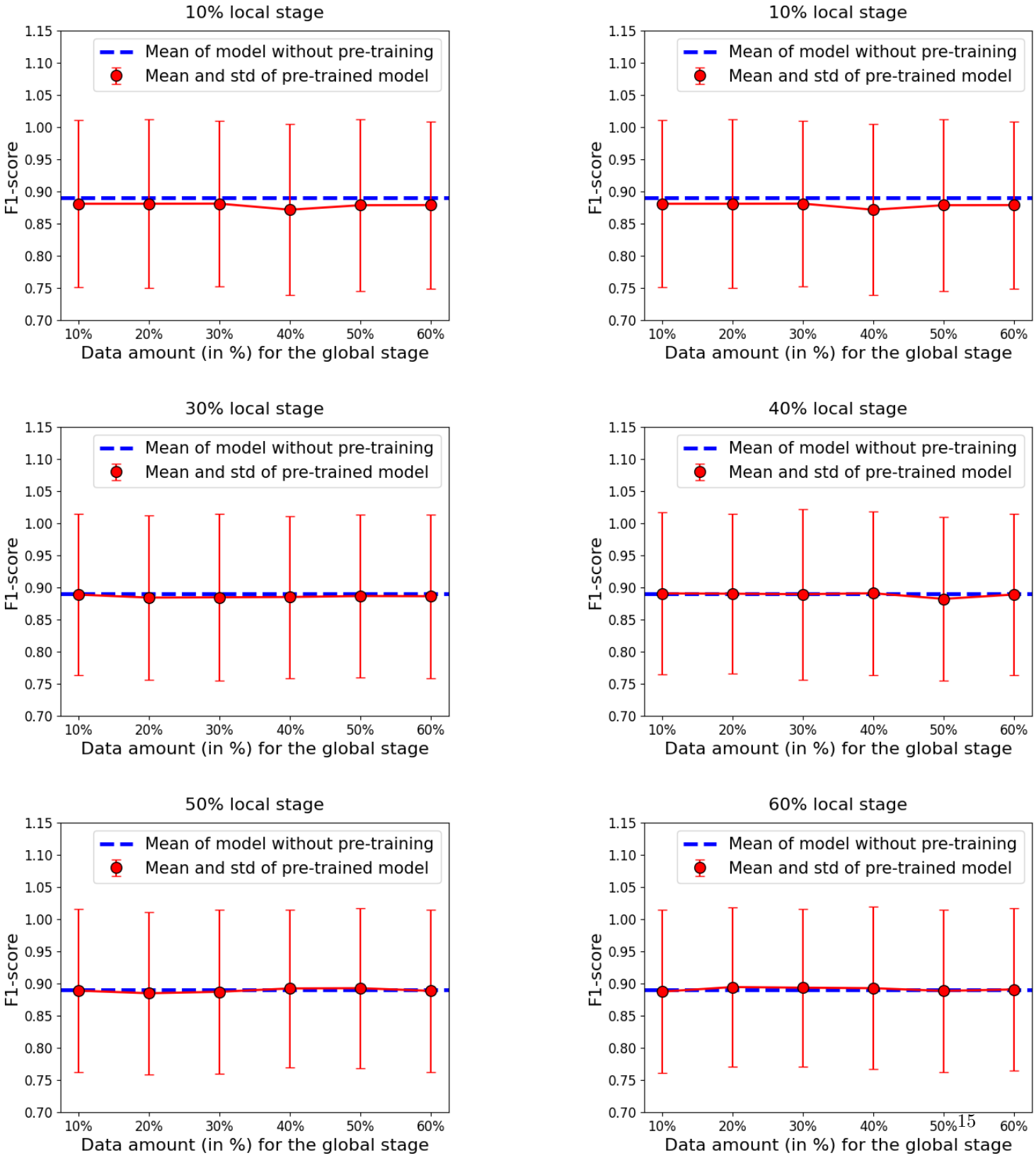

Figure 5.5: Progression of F1-score with increasing data usage in the global stage, with constant proportion for the local stage, and trained with the combination of the SSIM loss and Dice loss. The blue lines indicate the mean of the supervised model without the pre-training stage. The red lines indicate the mean of the semi-supervised models and the vertical lines indicate their corresponding standard deviations.

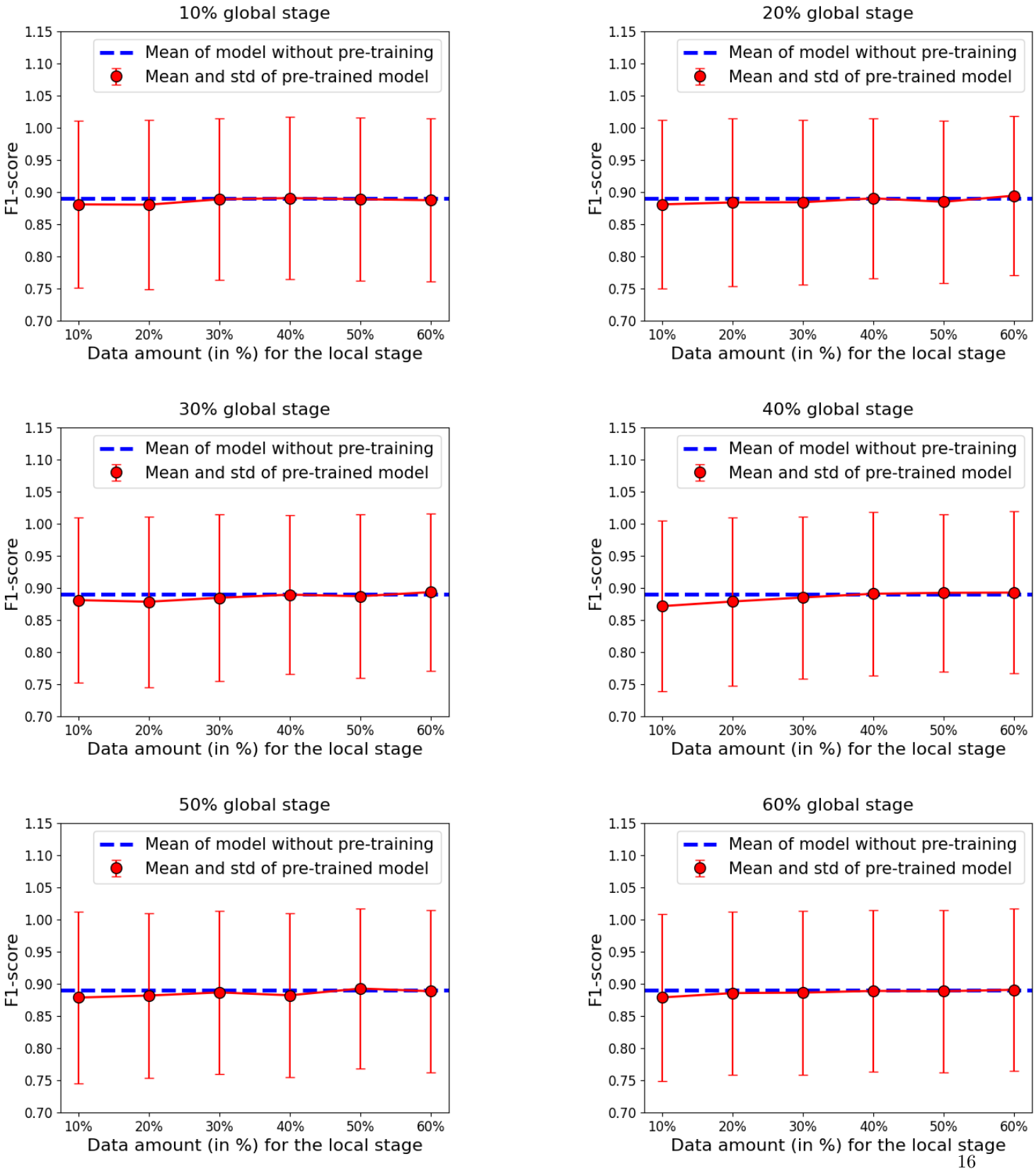

Figure 5.6: Progression of F1-score with increasing data usage in the local stage, with constant proportion for the global stage, and trained with the combination of the SSIM loss and Dice loss. The blue lines indicate the mean of the supervised model without the pre-training stage. The red lines indicate the mean of the semi-supervised models and the vertical lines indicate their corresponding standard deviations.

	BCE & Dice	SSIM & Dice	SSIM-MAE & Dice
F1-score	0.893 ± 0.120	0.885 ± 0.128	0.886 ± 0.127

Table 5.1: Mean F1-score and standard deviation of all 36 pre-trained models trained with different loss combinations.

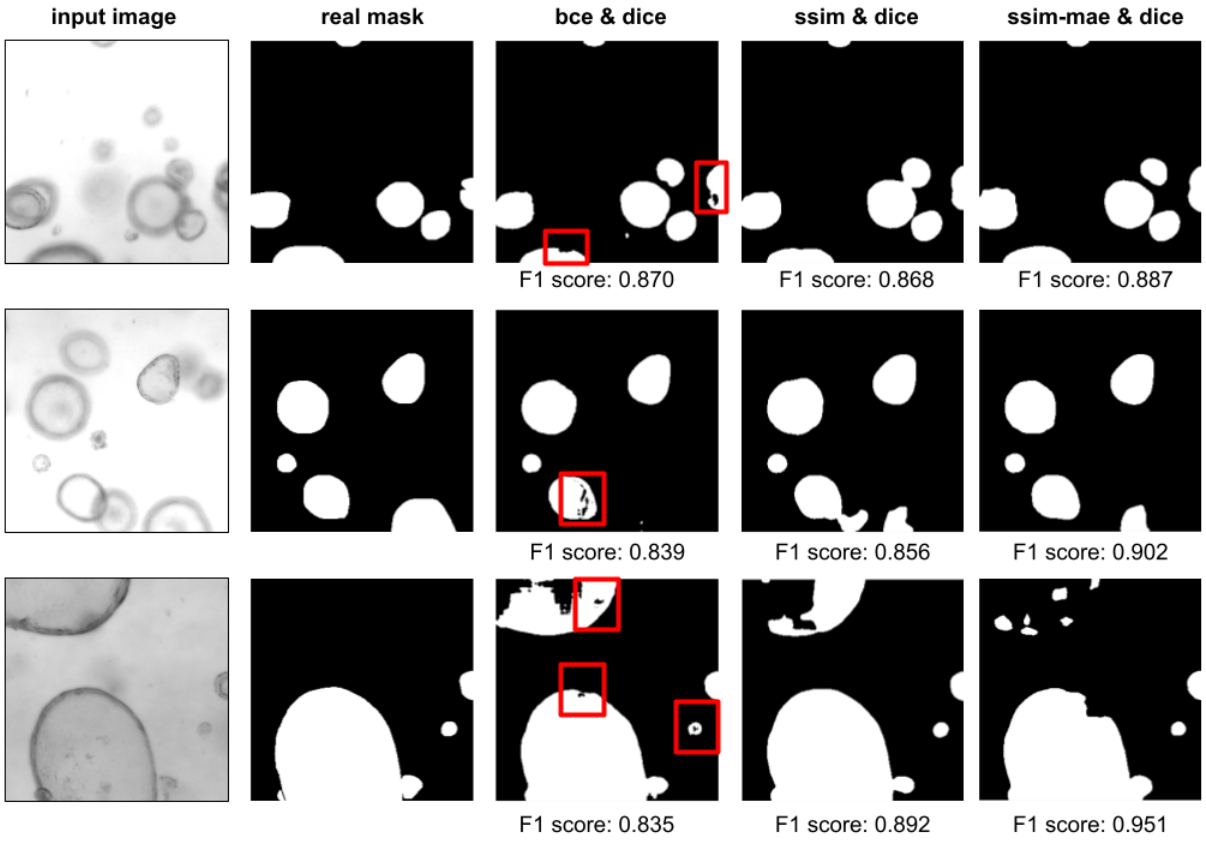

Figure 5.7: Three randomly selected organoid images from the test set, attached with the real masks and the predictions of the three fine-tuned models. The predictions of each of the three fine-tuned models are derived from the model with the highest F1-score of all 36 models.

look quite similar for each organoid image, there are some differences between them. Upon observing the predictions obtained from training with the BCE and dice loss combination, irregularities become apparent in some organoids, indicated by the red boxes. The predictions of the other loss combinations also contain irregularities in certain organoids but is more evident in the predictions of the BCE and dice loss combination. Some organoids contain random black pixels and other organoids exhibit irregular boundaries. This can best be explained by the formula of the BCE loss (see formula 3.6). The BCE loss function treats each pixel independently, aiming to minimize the difference between the predicted values and the ground truth labels on a pixel-wise basis. Since the BCE loss function does not take into account the relationship between pixels and the spatial structure of the image, it may result in inconsistencies and irregularities in the segmented regions. On the other hand, the SSIM loss (see formula (3.8)) does take into account the luminance, contrast and structural components of the images, which allows the model to better capture the spatial structure of the image and the relationships between pixels. Figure 5.8 provides a visualization that clarifies the explanation. The left image shows a circle, intended to represent a real organoid object. The middle- and right images exhibit similar structural patterns although the middle image has a smaller radius while the right image contains random black pixels. Both images have an equal number of misclassified pixels.

Figure 5.8: Three 320x320 pixel images: left - real organoid, middle - similar shape with smaller radius, right - similar shape with random black pixels.

As expected, the BCE loss values are similar for the middle image and right image as the BCE loss does not consider pixel relationships and the spatial structure. Conversely, the SSIM loss values differ between the images. The SSIM loss is lower for the middle image compared to the right image, indicating that the SSIM prioritizes coherent shapes over incoherent shapes.

To conclude, based on the F1-score, the BCE and Dice loss combination outperforms the SSIM(-MAE) loss combinations. However, in terms of the organoids' coherent structure, the SSIM loss might be a better choice than the BCE loss.

5.3 Frozen models vs non-frozen models

The **third** research question focuses on examining whether freezing (vs not freezing) the encoder of the global stage when training on the local stage led to better final segmentation results. Table 5.2 shows the F1-scores of the frozen models and the non-frozen models. All non-frozen models exhibit higher F1-scores than their corresponding frozen models. Figure 5.9 visualizes three random organoid images from the test set, where the F1-score for the frozen and non-frozen models can be discussed. As can be seen in the Figure, the F1-scores of the frozen models are (far) lower than the F1-scores of the non-frozen models. However, the frozen models actually provide better predictions of the input image than the non-frozen predictions. The reason that the F1-score is lower for the frozen models is because the real mask does not contain some large organoids, indicated with the red boxes. Therefore, given that the masks are not always perfectly aligned with the actual organoids, the F1-scores should not be fully trusted. The larger organoids are better captured by the frozen models because the global stage parameters aims to learn the high-level features that can help the network differentiate between images with different characteristics, such as variations in shape and size. The local stage aims to learn the low-level features by capturing the fine details. For the non-frozen models, the local stage parameters diminish the efficiency of the global stage parameters and therefore the larger organoids are detected worse.

	frozen			non-frozen		
F1-score	BCE & Dice	SSIM & Dice	SSIM-MAE & Dice	BCE & Dice	SSIM & Dice	SSIM-MAE & Dice
	0.886 ± 0.119	0.867 ± 0.138	0.864 ± 0.139	**0.901** ± 0.114	**0.891** ± 0.126	**0.893** ± 0.126

(a) 40% global stage, 60% local stage

	frozen			non-frozen		
F1-score	BCE & Dice	SSIM & Dice	SSIM-MAE & Dice	BCE & Dice	SSIM & Dice	SSIM-MAE & Dice
	0.875 ± 0.128	0.869 ± 0.130	0.852 ± 0.137	**0.890** ± 0.121	**0.880** ± 0.130	**0.879** ± 0.130

(b) 60% global stage, 10% local stage

Table 5.2: The F1-scores of the six trained frozen models attached with their corresponding non-frozen models.

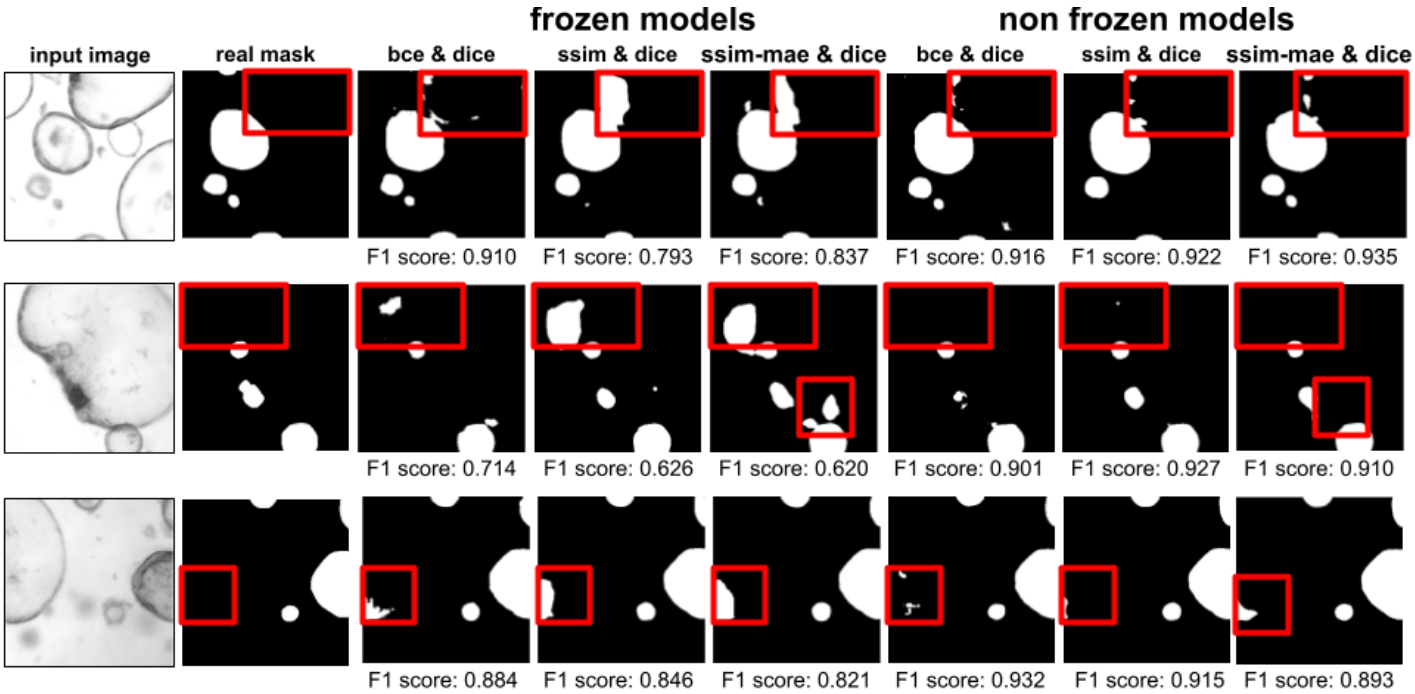

Figure 5.9: Three randomly selected organoid images from the test set, attached with the real masks and the predictions of three frozen models and three non-frozen models.

To sum up, based on the F1-scores, the non-frozen models outperform the frozen models. However, the real masks are not always perfectly aligned with the actual organoids and the frozen models seem to better capture the larger organoids.

6 Conclusion

6.1 Discussion

This work evaluates the performance of the semi-supervised learning approach by implementing a two-step pre-training procedure. The performance is measured by comparing the F1-scores of the semi-supervised learning approach with the F1-scores of the supervised learning approach. Contrary to previous research (Berthelot et al., 2019) (Zhai et al., 2019), the results indicate that increasing the amount of unlabeled data does not result in improved segmentation performances for the semi-supervised models compared to the supervised models. The images of the dataset show a certain degree of similarity, which is why more data does not necessarily lead to better segmentation performances.

Additionally, the three different loss combinations are evaluated by comparing the F1-scores and by comparing the predicted masks. The results indicate that the SSIM(-MAE) (vs BCE) loss is better at segmenting the organoids' coherent structure. This result is evident when looking at their respective formulas (Nilsson & Akenine-Möller, 2020) (Jadon, 2020).

Finally, image visualization and F1-score comparison is done for the frozen (vs non-frozen) models. The results indicate better segmentation performances of the large organoid structures for the non-frozen (vs frozen) models. Due to the larger influence of the global stage parameters on the final model, the model is better at capturing the larger organoid structures.

The study is confronted with two major issues. First, the F1-scores are not fully reliable since the real masks are not always perfectly aligned with the actual organoids present in the original images. Therefore, drawing conclusions on the F1-scores only becomes problematic. The only way to discover the performance is by visually analysing all of the predicted masks. Unfortunately, analysing all predicted masks is impractical due to the substantial workload.

Another potential problem is the class imbalance of the image dataset. Most of the dataset is made up of background pixels and only a small amount consists of organoid pixels. This can lead to poor generalization on new seen datasets (Ali et al., 2013). Since the background class dominates the training data, the model may not effectively capture the patterns and characteristics of the organoid pixels.

6.2 Future work

For further research, the focus should be placed on two key aspects: improving the reliability of the segmented masks and resolving the class imbalance of the organoid- and background pixels. By improving the reliability of the segmented masks, perfect segmentation performance can be achieved and the evaluation scores can be more heavily relied on. By minimizing the differences between the organoid- and background pixels, the model will receive equal exposure to both of the classes, making it better at learning the patterns and characteristics of the organoids.

References

Ahamed, M. A., & Imran, A. A. Z. (2022). Joint learning with local and global consistency for improved medical image segmentation. In *Annual conference on medical image understanding and analysis* (pp. 298–312).

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2013). Classification with class imbalance problem. *Int. J. Advance Soft Compu. Appl*, *5*(3), 176–204.

Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., ... Rueckert, D. (2017). Semi-supervised learning for network-based cardiac mr image segmentation. In *Medical image computing and computer-assisted intervention- miccai 2017: 20th international conference, quebec city,*

qc, canada, september 11-13, 2017, proceedings, part ii 20 (pp. 253–260).

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, *32*.

Bian, X., Li, G., Wang, C., Liu, W., Lin, X., Chen, Z., ... Luo, X. (2021). A deep learning model for detection and tracking in high-throughput images of organoid. *Computers in Biology and Medicine*, *134*, 104490.

Cai, W., Huang, J., Deng, A., & Wang, Q. (2021). Volumetric reconstruction for combustion diagnostics via transfer learning and semi-supervised learning with limited labels. *Aerospace Science and Technology*, *110*, 106487.

Chaitanya, K., Erdil, E., Karani, N., & Konukoglu, E. (2020). Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, *33*, 12546–12558.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).

Deperlioglu, O. (2018). Classification of phonocardiograms with convolutional neural networks. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *9*(2), 22–33.

de Souza, N. (2018). Organoids. *Nature Methods*, *15*(1), 23–23.

Gavali, P., & Banu, J. S. (2019). Deep convolutional neural network for image classification on cuda platform. In *Deep learning and parallel computing environment for bioengineering systems* (pp. 99–122). Elsevier.

Haja, A., Horcas-Nieto, J. M., Bakker, B. M., & Schomaker, L. (2023). Towards automatization of organoid analysis: A deep learning approach to localize and quantify organoid images. *Computer Methods and Programs in Biomedicine Update*, *3*, 100101.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the ieee international conference on computer vision* (pp. 2961–2969).

Hu, X., Zeng, D., Xu, X., & Shi, Y. (2021). Semi-supervised contrastive learning for label-efficient medical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 481–490).

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., ... Wu, J. (2020). Unet 3+: A full-scale connected unet for medical image segmentation. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1055–1059).

Jadon, S. (2020). A survey of loss functions for semantic segmentation. In *2020 ieee conference on computational intelligence in bioinformatics and computational biology (cibcb)* (pp. 1–7).

Kretzschmar, K., & Clevers, H. (2016). Organoids: modeling development and the stem cell niche in a dish. *Developmental cell*, *38*(6), 590–600.

Nigam, K. P. (2001). *Using unlabeled data to improve text classification*. Carnegie Mellon University.

Nilsson, J., & Akenine-Möller, T. (2020). Understanding ssim. *arXiv preprint arXiv:2006.13846*.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... others (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Seeger, M. (2000). *Learning with labeled and unlabeled data* (Tech. Rep.).

Suzuki, K. (2017). Overview of deep learning in medical imaging. *Radiological physics and technology*, *10*(3), 257–273.

Tao, L., Zhu, C., Xiang, G., Li, Y., Jia, H., & Xie, X. (2017). Llcnn: A convolutional neural network for low-light image enhancement. In *2017 ieee visual communications and image processing (vcip)* (pp. 1–4).

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600–612.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, *3*(1), 1–40.

Wu, X., Hong, D., & Chanussot, J. (2022). Uiu-net: U-net in u-net for infrared small object detection. *IEEE Transactions on Image Processing*, *32*, 364–376.

Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, *33*, 6256–6268.

Zhai, X., Oliver, A., Kolesnikov, A., & Beyer, L. (2019). S4l: Self-supervised semi-supervised learning. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 1476–1485).

Zhang, J., Li, C., Kosov, S., Grzegorzek, M., Shirahama, K., Jiang, T., ... Li, H. (2021). Lcu-net: A novel low-cost u-net for environmental microorganism image segmentation. *Pattern Recognition*, *115*, 107885.

Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2015). Loss functions for neural networks for image processing. *arXiv preprint arXiv:1511.08861*.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, dlmia 2018, and 8th international workshop, ml-cds 2018, held in conjunction with miccai 2018, granada, spain, september 20, 2018, proceedings 4* (pp. 3–11).

Zhu, X. J. (2005). Semi-supervised learning literature survey.