Optimal U-Net for Image Segmentation

EPFL Center for Imaging

Antoine Bonnet
EPFL
antoine.bonnet@epfl.ch

Daniel Demko
EPFL
daniel.demko@epfl.ch

Louca Gerussi
EPFL
louca.gerussi@epfl.ch

***Abstract*—We offer an empirical analysis of the effects of U-Net architecture parameters; initial channels, pooling layers and skip connections as well as object radius and texture on the model performance for image segmentation. We found that objects of larger radius and high texture similarity (between the zones to be classified) were easier to accurately segment, while higher amounts of initial channels generally improved performance. However, adding skip connections only offered slight increase in accuracy and increasing the number of pooling layers carried no impact on model performance. Some results, particularly related to receptive field size and object recognition, were inconclusive. Our findings provide insight for optimizing the U-Net for object segmentation.**

I. Introduction

The aim of this project is to give an empirical analysis on the effect of various U-Net architecture parameters on the model's ability to segment synthetic biological cellular images. This general aim can be decomposed in three specific goals. In Experiment A, we determine the relationship between the number of pooling layers, the number of initial channels and the prediction accuracy of the network on datasets with varying properties. In Experiment B, we determine a relationship between the number of pooling layers, the size of the receptive field, the distribution of object size in the labelled images and the resulting model performance. In Experiment C, we ascertain the benefits of skip connections at different depths of the model.

Motivation

In recent years, there has been a growing interest in using deep learning techniques for medical image analysis tasks, such as segmentation of cellular images [1]. In this paper, we focused on finding optimal values for the U-Net architecture that would allow for good performance in the task of segmenting synthetic biological cellular images, while also minimizing computational cost. The U-Net is a popular architecture for image segmentation tasks due to its ability to effectively capture both local and global features in the input data. However, neural network architectures with higher complexity tend to produce better results, but at the expense of increased computational resources. By finding the optimal balance between performance and efficiency, we aimed to improve the effectiveness of the U-Net for this specific application.

II. Datasets

As it is costly to label authentic cellular images, the first part of our research was focused on generating synthetic monochrome cell images. These artificial images are intended to reproduce cells on a brightfield microscopy background (the most common microscopy modality, which is known to be difficult to segment). The idea behind this decision is for one to be able to generate highly specific datasets whose properties are under total control by stripping down the cell images to their most basic features: an image with deformed circle shapes over a background. Thus, through image generation, one has total control over the properties of each dataset. The two most important data generation parameters that will be analysed are the cells' average size and the "similarity" of the cell texture to the background texture.

The images created are 512 x 512 pixels wide with an average cell density of 40%. Each cell is an oval that is given a uniformly random orientation, while its width and length are generated from a normal distribution $\mathcal{N}(r, 1)$ centred around the average radius parameter r.

The background and cell zones are then respectively textured with a difference of Gaussians applied on normal random noise with different standard deviations σ_{bg} and σ_{cell}. Throughout our experiments, $\sigma_{cell} = 1$ is kept constant while σ_{bg} varies. The difference in texture between background and cell is quantified by the similarity measure $s = \frac{\sigma_{cell}}{\sigma_{bg}}$. Similar values of standard deviations yield high similarity, making it difficult for the U-Net to distinguish the cells from the background, whereas a lower similarity is linked with stronger texture discrepancy between classes. Datasets were generated for several combinations of average cell radius r and texture similarity s. Each dataset contains 100 training and 100 testing images and their associated labels. Examples of these generated images and their corresponding labels are available in the Appendix.

III. Models

A U-Net is composed of two paths: the encoder and the decoder. The encoder (or *contractive*) path of the network processes the input image and extracts high-level features, while the decoder (or *expansive*) path of the network uses these features to generate a detailed output image. Every pooling layer in the encoding path is matched by an upsampling layer in the decoding path.

Architecture parameters

1) **Number of pooling layers**: A pooling layer is used in the encoding path to downsample the spatial resolution of the input data. This is typically done by applying a pooling function, such as maximum pooling or average

pooling, to the input data within a local region. The use of pooling layers can help to access to visual features at different size/scales. This is a common practice in image-processing to a multi-resolution processing.

2) **Number of initial channels**: A channel refers to a dimension of the input data that corresponds to a specific feature or attribute. The number of initial channels is the number of channels at the first layer, then for every layer in the contractive path we downsample by 2 (divide resolution of image by 2) but double the number of channels.
3) **Skip connections**: Each layer of a U-Net may contain a skip connection by adding a direct connection between the encoder and decoder parts of the network. Skip connections allow the decoder to directly access the high-level features extracted by the encoder, rather than having to learn these features on its own. They make learning more efficient, by allowing the decoder to make use of information that has already been learned by the encoder. This can also help the model to produce more accurate results, by allowing it to incorporate both low-level and high-level information from the input image.

Convolution and pooling parameters

Other parameters were not considered in our experiment but could be made the subject of further study. We kept their value fixed during all experiments.

- **Kernel size** (set to 3): Size of the filters used in the convolutional layer.
- **Convolution stride** (set to 2): Number of input image pixels over which the kernel moves during each convolutional operation.
- **Pooling type** (set to *max*): Pooling layers are used to downsample the input image by aggregating several pixels into one. It can be performed using various techniques, such as max pooling or average pooling.
- **Pooling stride** (set to 2): Width of the window over which the pooling operation is performed.

Receptive field

The receptive field is the region of the input image that is used by a convolutional neural network to produce the output for a specific location in the output feature map. It is the region of the input that a particular filter in the network "sees" when producing its output. The size of the receptive field determines how much of the input image is used to predict one pixel of the output at a specific location, and therefore has an impact on the ability of the network to capture spatial context and long-range dependencies in the input.

IV. Methods

Experiment A: Architecture parameters

Goal: Determine the effect of U-Net architecture parameters (number of initial channels and number of pooling layers) on the performance of the image segmentation model on datasets with various parameters (cell radius and texture similarity).

We generated a dataset for every combination of average cell radius $r \in \{20, 30, 40, 50\}$ and texture similarity $s = \{0.1, 0.133, 0.2, 0.4\}$ for a total of 16 dataset parameter combinations. We limited the experiment to a maximum of 5 pooling layers, 5 different initial channel widths $\{2, 4, 8, 16, 32\}$ for all for a total of 25 architecture parameter combinations. We then trained each U-Net architecture with full skip connections on each dataset. The resulting 400 trained models were then applied on their respective test sets to produce 100 prediction images. These images form the estimations of the model where each pixel is labeled as 0 for background or 1 for cell.

The most common performance measure in image segmentation is the Intersection over Union (IoU) measure, also called Jaccard index. We therefore computed the average IoU value for each class (background and cell) over all 100 test predictions. These two IoU values were combined into a single IoU value by taking a weighted average based on the fixed cell density of 40%:

$$IoU = 0.4 \cdot IoU_{cell} + 0.6 \cdot IoU_{bg}{}^{1}$$

Experiment B: Receptive field

Goal: Determine the relationship between the number of pooling layers, the size of the receptive field, the distribution of object size in the labelled images and the resulting model performance.

The size of the receptive field is dependent on several parameters, including the convolution kernel size, the pooling stride, the convolution stride and the number of pooling layers. All of them were kept fixed except the latter, which is therefore mapped in one-to-one correspondence to the size of the receptive field [2]. We then trained several U-Nets with fixed 16 initial channels, full skip connections and from 1 to 5 pooling layers on different datasets with fixed similarity $s = 20\%$ and varying cell radii $r \in \{20, 30, 40, 50\}$.

Experiment C: Skip connections

As skip connections can be enabled or disabled at every level of the model, a U-Net with p pooling layers will have 2^p possible skip connection activations. We decided to build one model per skip connection combination for every number of layers ranging from 1 to 4 with 16 initial channels. We obtain respectively 2, 4, 8, and 16 combinations according to the number of layers. We train the model for each combination on 4 different datasets, with average radius of $r \in \{20, 30, 40, 50\}$ and similarity $s = 20\%$.

V. Results

Experiment A

As seen in the resulting IoU values of all architecture-dataset combinations shown in Figure 1, the accuracy of the model increases in relation with the number of initial channels. We observe however that there are runs yielding very low accuracy (dark shade values). We tried to investigate where the error came from, if it was from an error in the U-Net implementation, a too low number of training epochs or an issue related to early stopping. But none of these seemed to improve these outlying values. We believe that these are possibly the result of a too "shallow" processing done by the

[1] See the Further work section for further expansion

U-Net. Hence, based on the observation, we assumed that it is an error due to the low number of initial channels.

According to our intuition, an increase in the similarity between the background and the cells leads to a decrease in the accuracy. We notice nonetheless that the U-Net manages to classify with high accuracy all sets up to a similarity of 0.4.

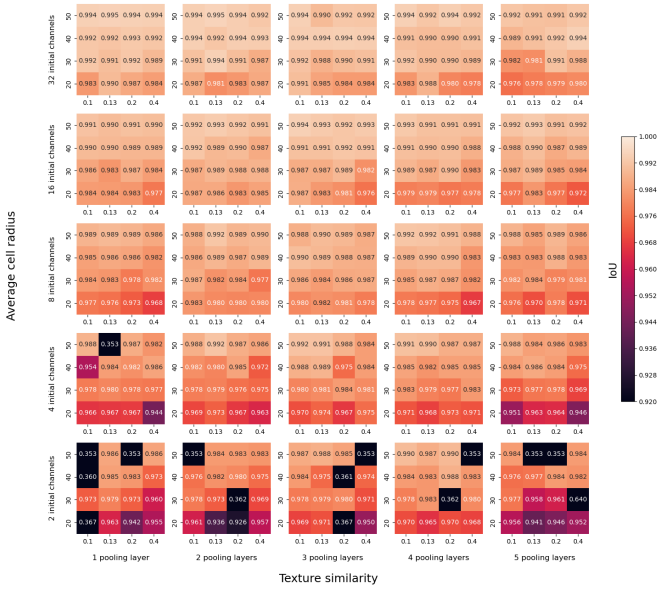

Figure 1: **Experiment A** — U-Net Model performance based on the number of pooling layers and the number of initial channels, for cell with different properties.

Experiment B

The resulting density-weighted test IoU value of each model is shown in Figure 2. The receptive field should be big enough to recognize bigger object, so we are expecting lower accuracy when the receptive field is small but the average cell radius is big. When the objects becomes smaller, a receptive field that is too big might misclassify the objects. When see this trend on the graph, however since the models are already efficient, the differences are minor.

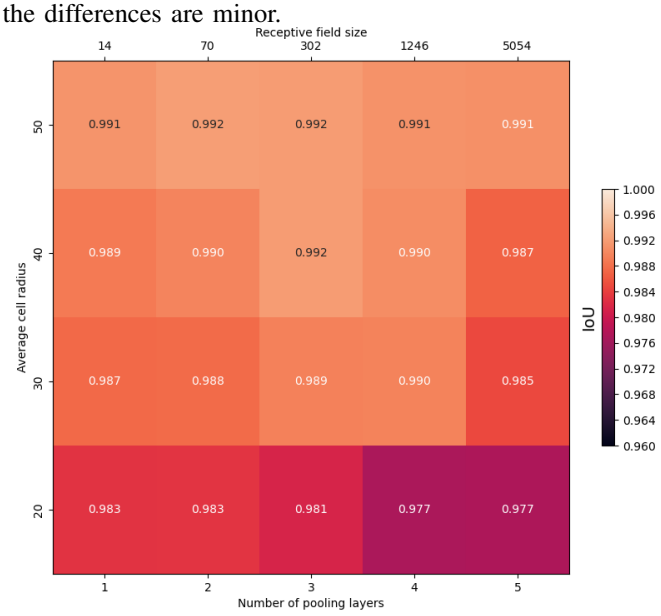

Figure 2: **Experiment B** — U-Net Model performance based on the number of pooling layers, the size of the receptive field and the distribution of object size.

Experiment C

The density-weighted test IoU values of each skip connection combination are shown in Figure3. The intention behind skip connections would be to help deep models converge faster and take care of problems such as vanishing gradient[3]. Our results however show no significant improvements of having skip connections. We discuss potential reason in the following section.

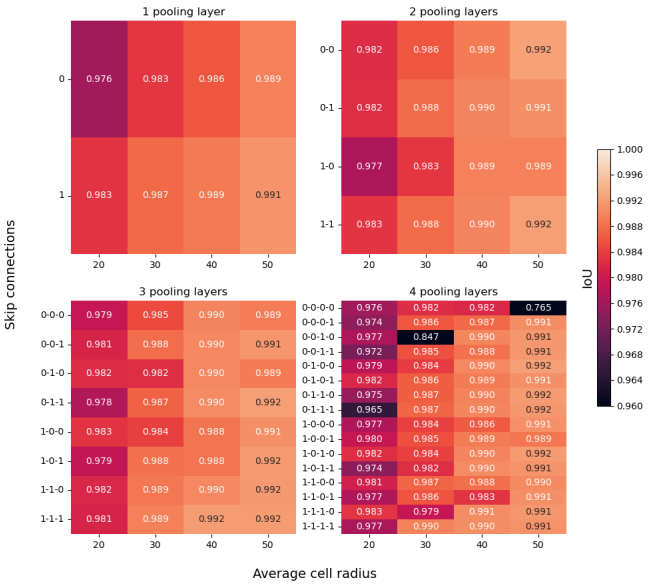

Figure 3: **Experiment C** — U-Net Model performance with respect to the number of pooling layers given different skipping layers combinations.

VI. Discussion

Experiment A

Hypothesis: Our intuition was that the optimization landscape would contain more local minima and be harder to optimize if there was higher similarity, leading to a decrease in the final Intersection over Union (IoU) score. This was slightly observed in our results. Additionally, we hypothesized that using more pooling layers with larger receptive fields and larger context would be better for detecting large objects. However, this was not observed and there was no gain from using more pooling layers. On the other hand, we found that using a higher number of channels to extract deeper information from the image led to a higher IoU. This was observed as a low number of channels resulted in frequent failures (models that have an unexpectedly low IoU). However, there is a trade-off that must be struck between efficiency and depth when determining the optimal number of channels to use.

Findings: The following observations were made:

- There was a medium effect of radius on IoU, with larger radius generally resulting in more accurate segmentation.
- The effect of similarity on IoU was low, but there was a slight improvement for low similarity.

- Using a lower number of initial channels made model training more difficult, often resulting in stagnation at local optima or complete failure. Some cases with a 0.35 IoU result were due to full-background prediction. In general, better results were obtained as the number of initial channels increased.
- There were more frequent failures when the radius was set to 50 and the number of initial channels was low.
- There was no noticeable improvement in the results as the number of pooling layers increased.

Experiment B

Hypothesis: We expected that larger objects would require a larger receptive field in order to be appropriately segmented and not split into several sub-objects. It follows that a higher number of pooling layers should favor the IoU, as these layers increase the receptive field size and may allow the model to better identify and segment larger objects.

Findings: The results of the empirical study did not provide a clear conclusion. While it was observed that smaller radius objects were harder to segment than large radius objects, this was already known prior to the study. The receptive field size was found to grow exponentially and surpass the cell size after 2 or 3 pooling layers. However, no other significant findings were made and it may be necessary to differ the different approach in order to evaluate the factors influencing the performance of the model more effectively.

Experiment C

Hypothesis: We expected that more skip connections would improve the IoU metric by mixing high-level and low-level contexts. This was generally observed in the study, as there was a slight increase in the IoU as the number of skip connections increased.

Findings: For 1 pooling layer, we observed an improvement from no skip to skip over all radii. For more layers on the other hand, increasing the numebr of skip connections leads to no clear improvement, as our accuracies are already near-perfect.

Further work

Future improvements: One potential future improvement for this study would be to create a perform correlation analysis on the parameters radius, similarity, number of channels, number of pools, and corresponding IoU to identify the parameters that are most and least correlated with good performance. By understanding the relationships between these variables and the IoU, it may be possible to identify key areas for optimization and improve the overall performance of the model. It was also brought to our attention that the formula we used to compute the IoU is unnecessarily complicated, and could be reduced by simply taking into account the cells' IoU. If one was to run the experiments again, we would recommend this approach as the class-weighted IoU only offers an approximation of the true model performance.

Reflection on results: The results of Experiments B and C did not meet our expectations and proved to be unsatisfactory. One potential reason for this may be that our datasets were too simplistic and easy to learn. Thus led trained models to be either near-perfect or completely wrong; there was no moderately efficient model. This made it difficult to clearly identify the factors influencing performance and draw meaningful conclusions from the results. To improve the relevance and usefulness of our results, it may be necessary to train the model on more complex datasets, such as real-world data rather than synthetic data. This could provide a more diverse range of test cases and allow us to better evaluate the effectiveness of different configurations and techniques.

VII. Summary

In this project, we studied the effects of U-Net architecture parameters on model performance for object segmentation. We evaluated the impact of object radius, similarity, number of initial channels, number of pooling layers, and number of skip connections on the IoU metric. Our results showed that larger radius objects were easier to accurately segment, and that low similarity led to a slight improvement in the IoU. The number of initial channels had a significant impact on the model's performance, with better results obtained as the number of initial channels increased. Increasing the number of pooling layers did not lead to any noticeable improvement in the IoU. We also observed a slight increase in the IoU as the number of skip connections increased. Some of our results, particularly the relationship between receptive field size and object recognition, were not as convincing as we had hoped. Overall, our findings provide insight into the factors influencing the performance of the U-Net for object segmentation and may inform future optimization efforts.

Acknowledgements

The authors thank Professor Daniel Sage from the EPFL Center for Imaging for his generous help and guidance throughout the project. This report was submitted in December 2022 to the ML 4 Science challenge as the final project of the *CS-433: Machine Learning* course at Ecole Polytechnique Fédérale de Lausanne (EPFL) [4].

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[2] A. Araujo, W. Norris, and J. Sim, "Computing receptive fields of convolutional neural networks," *Distill*, 2019, https://distill.pub/2019/computing-receptive-fields.

[3] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Deep Learning and Data Labeling for Medical Applications*, G. Carneiro, D. Mateus, L. Peter, A. Bradley, J. M. R. S. Tavares, V. Belagiannis, J. P. Papa, J. C. Nascimento, M. Loog, Z. Lu, J. S. Cardoso, and J. Cornebise, Eds. Cham: Springer International Publishing, 2016, pp. 179–187.

[4] "ML4Science," 2022, interdisciplinary Machine Learning Projects Across Campus. [Online]. Available: https://www.epfl.ch/labs/mlo/ml4science/

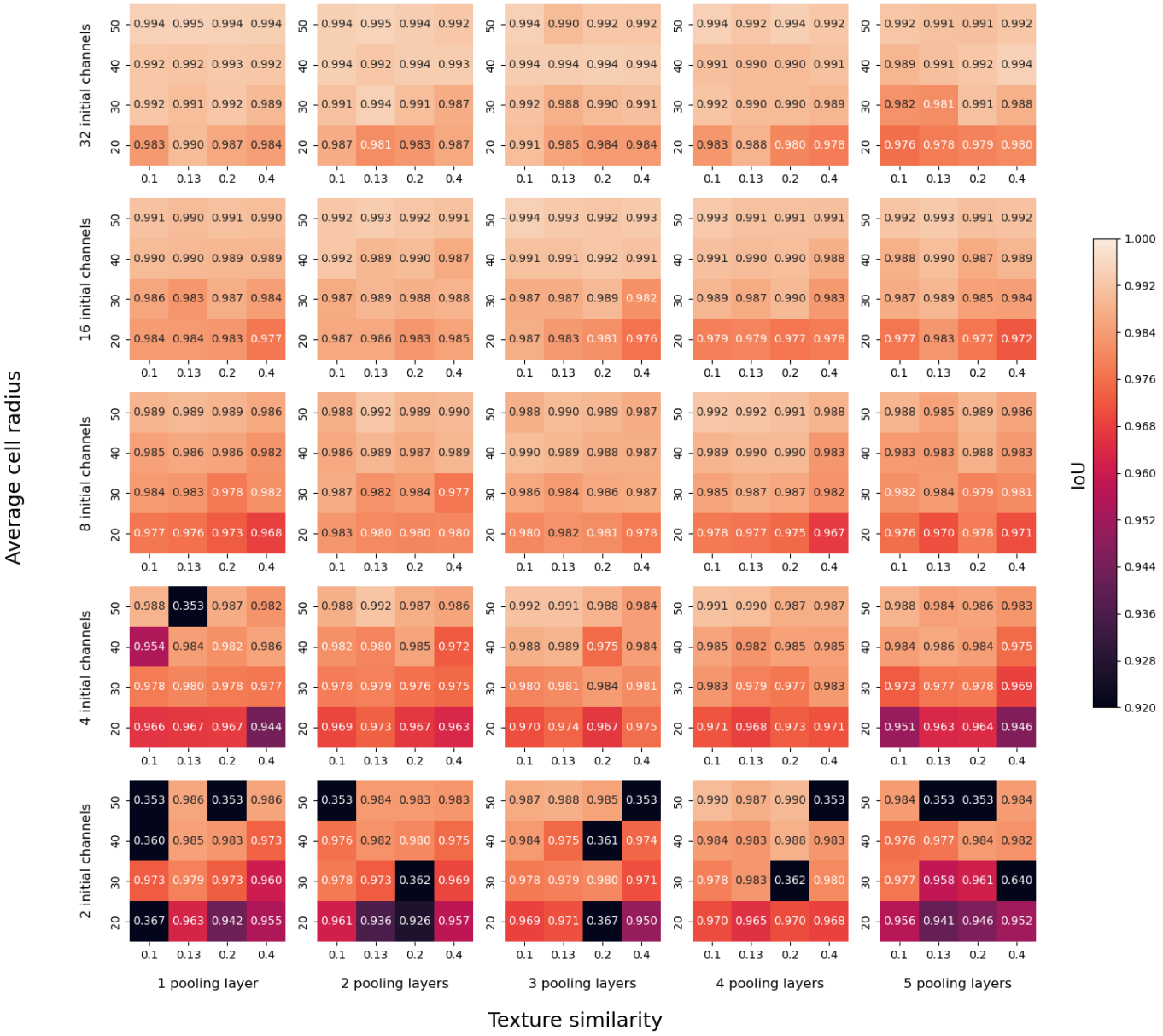

Figure 1: **Experiment A** — U-Net Model performance based on the number of pooling layers and the number of initial channels, for cell with different properties.

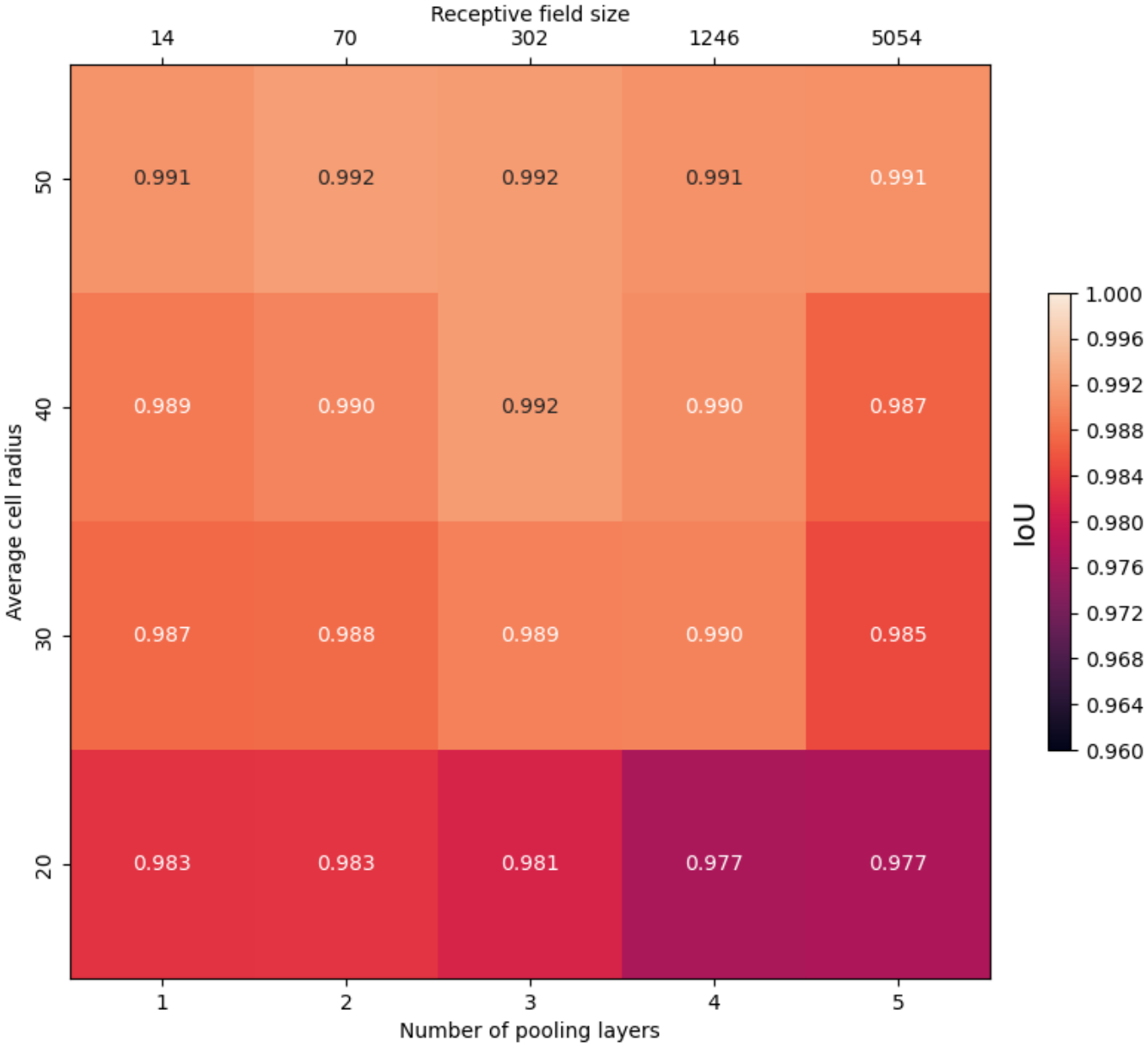

Figure 2: **Experiment B** — U-Net Model performance based on the number of pooling layers, the size of the receptive field and the distribution of object size.

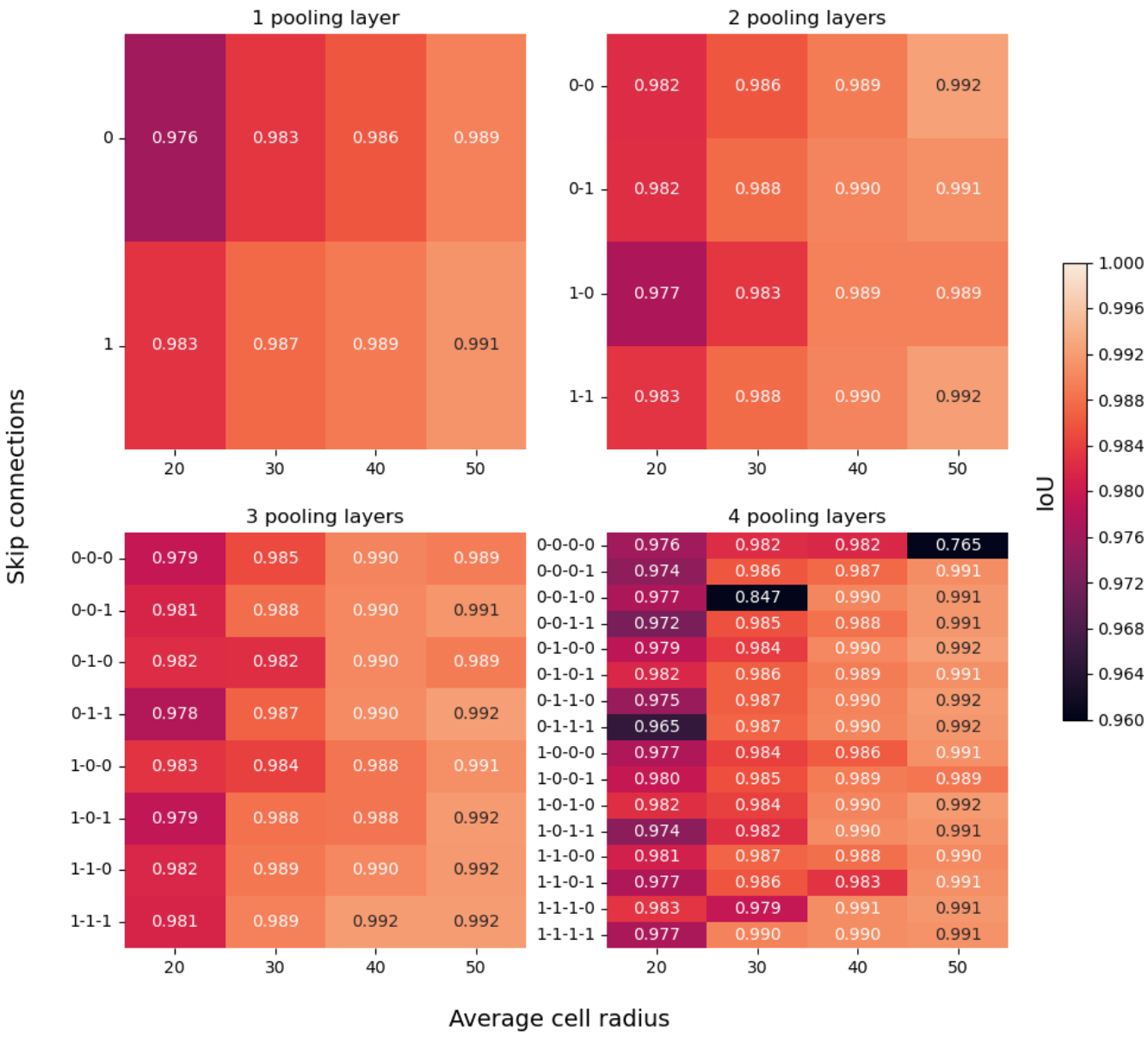

Figure 3: **Experiment C** — U-Net Model performance with respect to the number of pooling layers given different skipping layers combinations.

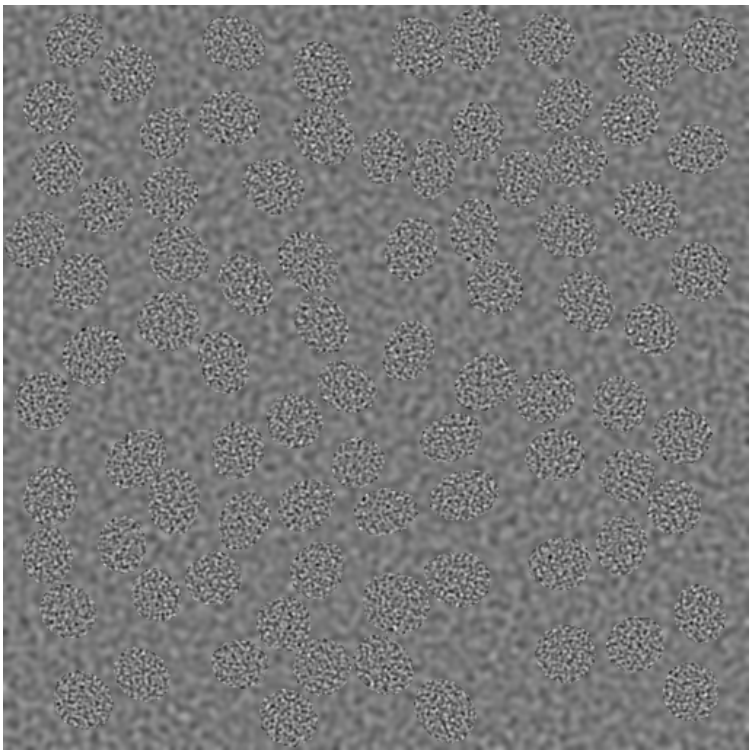

Figure 4: Example of a training image with radius 20 and $\sigma_{back} = 2.5$.

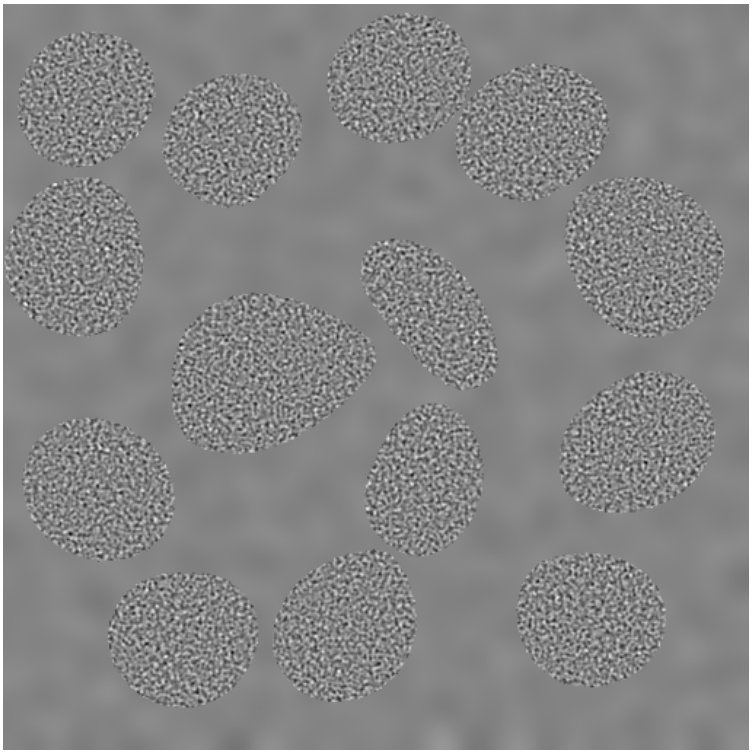

Figure 5: Example of a training image with radius 50 and $\sigma_{back} = 10$.