Survey of semantic segmentation models for in situ TEM images

Technical Report

Lucas Degeorge
ENSAE Paris - Institut Polytechinque de Paris
lucas.degeorge@ensae.fr

Federico Panciera
Centre for Nanoscience and Nanotechnology
Université Paris-Saclay
federico.panciera@c2n.upsaclay.fr

Nam Hong
Centre for Nanoscience and Nanotechnology
Université Paris-Saclay
nam.hong@c2n.upsaclay.fr

Abstract

The segmentation of in-situ TEM images of the nanowire is a crucial step, as the geometry of the nanowire could impact the physical and chemical properties. However, manually labeling nanowire images is time-consuming, and such images are generally difficult to segment with histographical methods. We compare, in this work, different deep learning models (UNet and PSPNet) and paradigms (supervised and semi-supervised learning) for nanowire segmentation. We extend the focus beyond segmentation accuracy to encompass precision in nanowire's geometry measurement. The results reveal that UNet-based models outperform PSPNet-based ones, with pseudo-labeling emerging as the most effective technique for leveraging unlabeled data. Semi-supervised approaches showed limited advantages in utilizing unlabeled data, challenging prior expectations

Keywords Semantic Segmentation · Semi-supervised learning · Pseudo-labeling · in-situ TEM

1 Introduction

Nanowires, with their unique properties and diverse applications in nanoelectronics, photonics, and sensing, have garnered significant attention in the field of materials science and nanotechnology. The controlled growth and the cristal phase of nanowires remain pivotal for tailoring their properties to specific applications. Recently, [12] showed that, for certain materials, the contact angle between the droplet (a liquid used as a catalyst) and the nanowire could affect the crystal phase in which the nanowire grows (see Figure 1). To accurately track the growth of nanowires, transmission electron microscopy (TEM) is used. in-situ TEM techniques provide real-time insights by offering a wealth of high-resolution image data for analysis.

Segmentation and subsequent geometric analysis of in-situ TEM images are indispensable steps in extracting quantitative information about nanowires. Accurate delineation of nanowire boundaries and the measurement of key geometrical properties, such as diameter, length, and orientation, are critical for measuring the contact angle and then optimising synthesis conditions. However, the segmentation task is highly time-consuming if it is performed manually. The uneven contrast and the high noise levels in images are making traditional image analysis or computer vision methods ineffective, especially for handling the large volumes of data generated by modern TEM instruments.

Recent leaps in deep learning, specifically within the realm of semantic segmentation, have revolutionized image analysis tasks by enabling computers to precisely identify and categorise objects within images. The advent of convolutional neural networks (CNNs) with deep architectures and innovative architectures like U-Net [4] and PSPNet [7] has substantially improved the accuracy and efficiency of semantic segmentation, even in scenarios with intricate and fine-grained structures like nanowires.

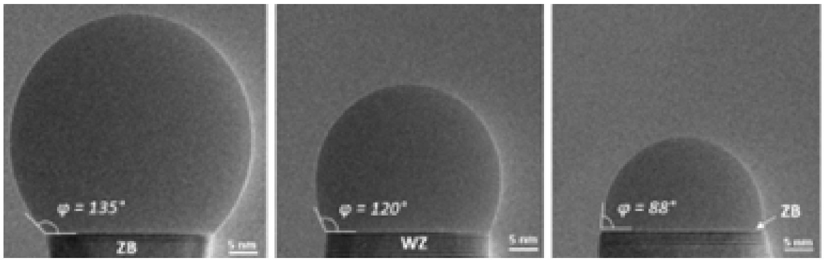

Figure 1: Crystal phase selection at different contact angles

In the context of our study, selecting the most suitable deep learning model transcends mere semantic segmentation accuracy. While high segmentation accuracy, according to established metrics used in computer vision, is undoubtedly crucial, it is at least equally imperative that the chosen model demonstrates high accuracy for measurements of nanowire contact angles. Any inaccuracies introduced during segmentation can propagate into the subsequent geometric analysis, potentially introducing bias and uncertainty in the physical interpretation of the results.

The primary objective of this work is to identify the most effective model that can accurately and robustly segment nanowires from TEM images, while also measuring their geometrical properties with precision and efficiency. In the following sections, we will describe the experimental setup, the dataset used for training and evaluation, the deep learning architectures employed, and the metrics utilized for performance assessment. We will present a detailed analysis of the results, highlighting the strengths and limitations of each model.

2 Related work

Deep learning in material sciences In the realm of material science, the application of deep learning models has emerged as a transformative force, revolutionizing the analysis and segmentation of complex microstructures. The authors of [14] demonstrate the potential of deep learning to address the challenges associated with the time-consuming manual labeling of X-ray CT images and the difficulty in segmenting such intricate datasets. By introducing an asymmetrical depth encode-decoder CNN, this approach achieves remarkable accuracy even with limited labeled data, while quantifying the impact of human bias in the training data. Furthermore, the paper emphasizes the versatility of transfer learning in enhancing model performance across diverse battery materials, ultimately improving material property determinations. Similarly, in [13], the study underscores the value of transfer learning from domain-specific datasets, such as MicroNet, in improving segmentation accuracy for microscopy images, outperforming traditional ImageNet pre-training. The research done by the authors of [10] further contributes to the field by addressing the challenges of applying deep learning to physical material science problems, emphasizing the need for robust, generalizable models. Collectively, these studies illuminate the profound impact of deep learning models in material sciences, offering efficient solutions for segmentation, analysis, and optimization across a spectrum of applications, from battery characterization to microscopy and nanofabrication, ultimately enhancing our understanding of material properties and enabling advancements in material science.

Semantic segmentation One pioneering work in semantic segmentation is the Fully Convolutional Networks, introduced in [3] by Jonathan Long, Evan Shelhamer, and Trevor Darrell. This paper introduced the concept of adapting CNNs for dense pixel-wise classification, laying the foundation for many subsequent models. One of the key innovations here was the development of end-to-end trainable architectures that produce pixel-level predictions directly from input images. Another notable milestone is the U-Net by Olaf Ronneberger, Philipp Fischer, and Thomas Brox [4]. The U-Net architecture has become iconic in the field, particularly in biomedical image segmentation. Its symmetric encoder-decoder structure, featuring skip connections, enables precise segmentation and has been widely adopted in various applications. The introduction of residual networks (ResNets) by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun in [2] significantly improved the depth and performance of CNNs. Researchers soon adapted ResNets for semantic segmentation tasks, leading to the development of models like DeepLab [5], which incorporates dilated convolutions to capture multi-scale contextual information. PSPNet, as presented by Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang and Jiaya Jia. in [7] excels in capturing rich contextual information through its Pyramid Pooling Module. This module, based on dilated convolutions and residual blocks, effectively captures multi-scale contextual features, enabling precise object segmentation within complex scenes.

Semi-supervised learning and pseudo-labeling Semi-supervised learning plays a pivotal role in enhancing the performance of deep learning models for semantic segmentation. It addresses the challenge of limited annotated data by incorporating both labeled and unlabeled examples during training. One common approach, as demonstrated in [1], involves pseudo-labeling unlabeled data by assigning class labels based on the highest predicted probabilities. This method effectively extends the labeled dataset and encourages the model to generalize better, leading to improved segmentation results. Another important principle of SSL is consistency training. It is built on the assumption that the model's predictions should remain relatively stable when realistic perturbations are applied to unlabeled examples. This concept aligns with the idea of favouring models with decision boundaries residing in low-density regions, resulting in consistent predictions for similar inputs. For instance, the Π-Model [6] enforces consistency by comparing predictions over two perturbed versions of the inputs, incorporating different data augmentations and dropout. Other approaches, such as Mean Teacher [9], achieve stability in predictions by utilizing weighted moving averages, either over previous predictions or the model's parameters. Meanwhile, Virtual Adversarial Training [8] approximates perturbations that would induce the most significant changes in the model's predictions without relying on random perturbations. The paper *Semi-Supervised Semantic Segmentation with Cross-Consistency Training* [11] extends this idea of consistency training by enforcing the alignment of predictions between the main decoder and auxiliary decoders over various perturbations. Notably, these perturbations are applied to the encoder's outputs, rather than the inputs, enhancing the model's ability to capture meaningful features and relationships in the data.

3 Approach and configuration

In this section, we outline the methodology employed in our study to compare various semantic segmentation deep learning models. The goal of our research is to comprehensively assess the performance and characteristics of these models, shedding light on their suitability for the efficient measurement of geometric properties of nanowires. We also want to understand the impact of supervised and semi-supervised learning approaches.

3.1 Paradigms and problem definition

To comprehensively assess the performance of the models, we explore two different learning paradigms: supervised and semi-supervised learning. One of the primary motivations for adopting semi-supervised learning in our study is the scarcity of labeled images and the abundance of unlabeled data. This is a common scenario in material science applications, where the cost and time required for manual annotation can be prohibitively high.

One promising technique in semi-supervised learning is cross-consistency training. This approach leverages the idea that predictions made by a model on unlabeled data should be consistent regardless of the modality or augmentation applied to the input. It encourages the model to produce coherent predictions and learn meaningful representations from both labeled and unlabeled data. To that purpose, a consistency loss term is introduced. It quantifies the similarity between predictions generated from different views of the same unlabeled data, thus promoting consistency in the model's output. Cross-consistency training involves training a model in a dual fashion: one path uses labeled data with ground truth annotations, while the other path processes unlabeled data.

Formally, we consider two datasets $\mathcal{D}_l$ and $\mathcal{D}_u$. $\mathcal{D}_l = \{(x_1^l, y_1^l), \dots, (x_m^l, y_m^l)\}$ is the small labeled dataset, containing m labeled images x_i^l (of nanowire) and their corresponding annotation masks y_i^l. $\mathcal{D}_l = \{x_1^u, \dots, x_p^u\}$ is the unlabeled dataset. We have $m << p$.

In our work, the models have the encoder-decoder architecture (see Section 3.2). The semi-supervised architecture is composed of a shared encoder h, a main decoder g and one (or more) auxiliary decoder g_a. The network $f = g \circ h$ is trained on the labeled dataset $\mathcal{D}_l$ in a traditional supervised way. A loss, denoted $\mathcal{L}_s$ ("supervised loss") is computed between the output of the main decoder g and the ground-truth mask. The auxiliary network $g_a \circ h$ is trained on the unlabeled dataset $\mathcal{D}_l$ by enforcing consistency of predictions between the main decoder and the auxiliary decoder (see Figure 2). The auxiliary decoder g_a takes as input a perturbed version of the encoder's output. A loss, denoted $\mathcal{L}_u$ ("unsupervised loss"), is computed between the output of the auxiliary decoder g_a and the output of the main decoder g (considered as a pseudo-ground truth). The total loss, denoted $\mathcal{L}$, is then: $\mathcal{L} = \mathcal{L}_s + w_u \cdot \mathcal{L}_u$ where w_u is an unsupervised loss weighting function. In our study, $\mathcal{L}_l$ is the DICE-Cross Entropy Loss, and $\mathcal{L}_u$ is the Mean Square Error (MSE).

This cross-consistency approach was first introduced in the work of Ouali *et al* in [11].

3.2 Architectures and models

Our study focuses on two prominent architectures for semantic segmentation: UNet [4] and PSPNet [7]. These architectures have gained significant attention in recent years due to their remarkable performance in pixel-wise image segmentation tasks. UNet, known for its skip connections and symmetric architecture (see Figure 3), offers an effective

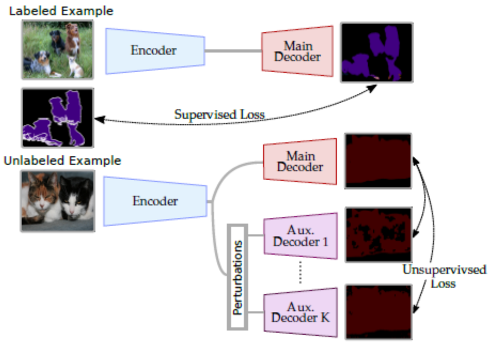

Figure 2: Cross consistency training for the semi-supervised models. From [11]

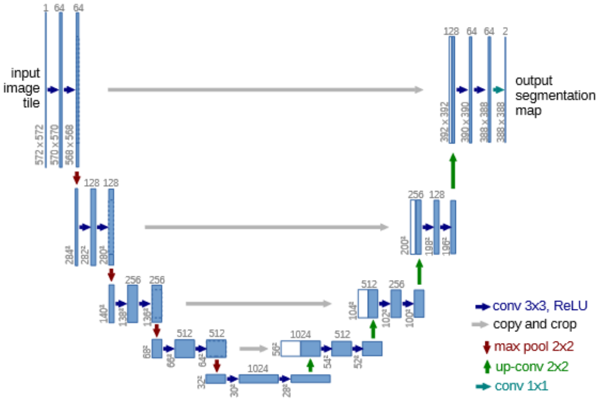

Figure 3: Archtiecture of the UNet. From [4]

framework for capturing fine-grained details, while PSPNet, with its pyramid pooling module (see Figure 4), excels at modelling contextual information across various scales.

We have trained six different networks:

Model A: supervised PSPNet The encoder links a ResNet-34 and a PSP module (from [7]). last two strided convolutions of ResNet are replaced with dilated convolutions. The main (and only here) decoder contains an initial 1×1 convolution and a series of three sub-pixel convolutions with ReLU non-linearities to upsample the outputs to the original input size. This model is only trained on the labeled dataset $\mathcal{D}_l$.

Model B: semi-supervised PSPNet This model has the same encoder and main decoder as Model A. The auxiliary decoder has the same architecture as the main decoder. The perturbation functions are detailed in Section 3.3 This model is trained with the cross-consistency training approach detailed in Section 3.1. This model is the one developed and tested in the work of Ouali *et al.* in [11].

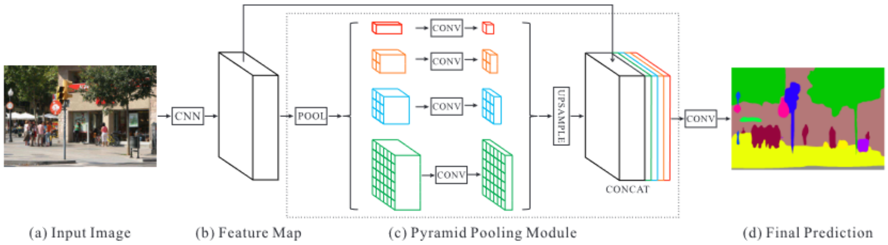

Figure 4: Archtiecture of the PSPNet. From [7]

Model C: semi-supervised PSPNet with pre-trained ResNet The architecture is the same as Model B. The ResNet is first trained in a supervised manner on $\mathcal{D}_l$ and the weights are then frozen. The PSP module's and the decoders' weights are then learnt with the cross-consistency approach.

Model D: supervised UNet This is the original UNet introduced in [4]. It is trained only on $\mathcal{D}_l$.

Model E: semi-supervised UNet This semi-supervised version of the UNet is inspired from the semi-supervised PSPNet. The encoder (the left part of the "U") and the bottleneck are not modified. The main and the auxiliary decoders are the same as the right part of the "U". The perturbation functions are applied after the encoder and the bottleneck. However, the skip connections are not affected by the perturbations.

Model F: supervised UNet with pseudo-labeling A first UNet (with the same architecture as Model D is trained on the labeled dataset $\mathcal{D}_l$. Some prediction is done on a subset of the unlabeled dataset $\mathcal{D}_u$. The outputs are considered as pseudo-labels for the corresponding images. We now have a new pseudo-labeled dataset $\mathcal{D}_{pl}$. Model F is trained on the reunion of the labeled and pseudo-labeled datasets: $\mathcal{D}_l \cup \mathcal{D}_{pl}$.

3.3 Perturbation functions

The perturbation functions used in our work are the ones used by Ouali et al. in their study [11]. There are three kinds of perturbations:

- Feature-based perturbations: They consist of either injecting noise into or dropping some of the activations of the encoder's output feature map.
- Prediction based perturbations: They consist of adding perturbations based on the main decoder's prediction $\hat{y} = g(z)$ or that of the auxiliary decoders. It could be masking based perturbation or adversarial perturbation.
- Random perturbations: a spatial dropout.

All the perturbations were used for the semi-supervised PSPNet (Model B and C). However, due to the skip connections, only the feature-based and random perturbation were used for the semi-supervised UNet (Model E).

Section 3.2.3 of [11] gives more details about the perturbation functions.

4 Measurement of geometric properties

After the segmentation of an image, some post-processing steps are required to get the value of the contact angle.

The first step involves the extraction of object edges. A 3x3 square kernel is used with the operations erosion and dilation of OpenCV to get the edges of the nanowire and the droplet. Then, the interface between the droplet and the nanowire could be isolated (see Figure 5). the Hough Line Transform technique is then applied to identify lines along the interface. The coordinates of the endpoints and the middle point of the interface are computed. The perpendicular lines are calculated. Finally, the coordinates of the droplet endpoints are determined. The result of this process is shown in Figure 6.

The coordinates of those points allow us to compute some distances. h, the height of the droplet, is computed using the endpoints of the droplet. a, the length of the interface is computed using the endpoints of the interface. Given these two values, the contact angle is computed: $\varphi = \pi - 2\arctan\left(\frac{a}{h}\right)$.

When the drop does not appear entirely inside the frame, h could not be obtained using the previous method. In this case, the Hough Circle Transform is applied on the droplet edges. Given the radius and the center of the circle outputted byt the Hough Transform, h is computed.

5 Implementation and training details

The implementation is based on the PyTorch 2.0 and OpenCV 4.7 frameworks. During the training of models, we train for 50 epochs using the stochastic gradient descent (SGD) optimizer. The initial learning rate is 0.01. We use the polynomial learning rate scheduler. At iteration t, the learning rate is defined by: $lr_{t+1} = lr_t \cdot \left(1 - \left(\frac{t}{max\_iter}\right)^{0.9}\right)$.

The datasets are composed of $(512, 512)$ one-channel grayscale images. The size of the datasets are reported in Table 1.

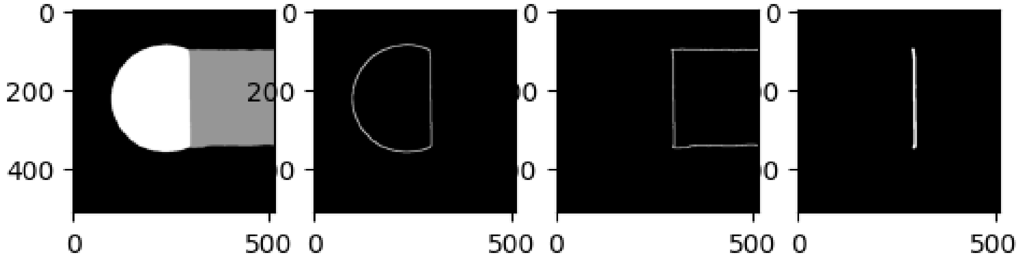

Figure 5: Left: segmentation outputted. Centre: edges of the droplet and the nanowire. Right: Interface isolated

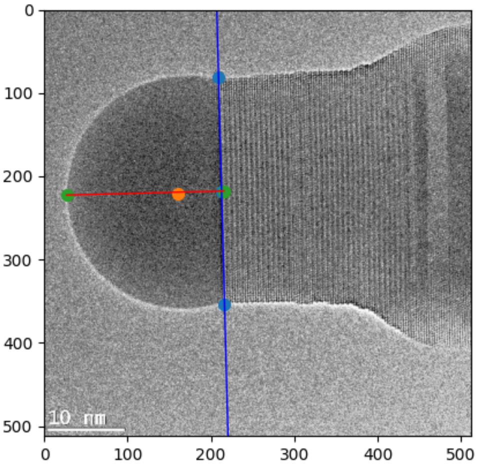

Figure 6: Points and lines detected overlaid on the input image

Dataset		Size
Labeled dataset $\mathcal{D}_l$	Training set	200 images + 1125 augmented images
	Evaluation set	25 images
	Test set	20 images
Unlabeled dataset $\mathcal{D}_u$		10 000 images
Pseudo-labeled dataset $\mathcal{D}_{pl}$		2350 images

Table 1: Size of the datasets used

All the experiments are conducted on a Nvidia Quadro P5000 GPUs. The implementation is available at: https://github.com/lucasdegeorge/NW_SemSeg

6 Results

In this section, we present the comprehensive results of our study, which aims to evaluate and compare the performance the six models A to F. Our evaluation encompasses traditional metric scores and the accuracy of contact angle measurement.

6.1 Usual metric scores

The metrics used for the experiments are the DICE score, the Jaccard index (or the mean Intersection over Union (mIoU)), and pixel accuracy. Figure 7 shows the score of the best epoch of each model on the test dataset.

First, we note that the UNet-based models (Models D to F) have better scores than the PSPNet-based models (Models A to C). This can also be seen when looking at the segmentations outputted by the models for some images (see Section **??**).

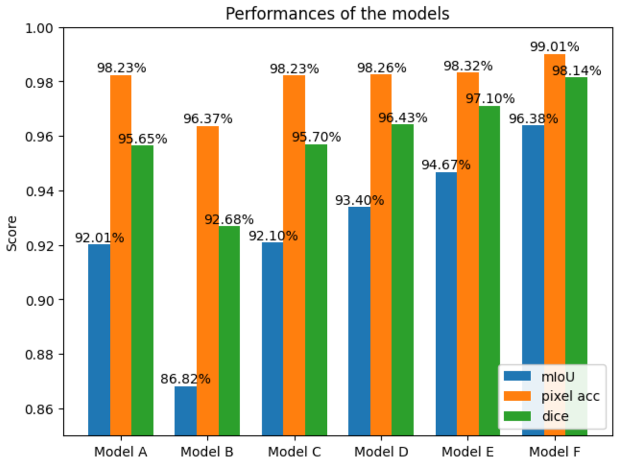

Figure 7: Scores of the six models

Model	MSE
Model A	28.30
Model B	61.46
Model C	30.68
Model D	3.88
Model E	6.01
Model F	3.26

Table 2: MSE between the models' and human annotator's values

For Models A to C (based on PSPNet), the semi-supervised model (B) performs less well. Both models A and C have similar results, indicating that using the unlabeled dataset does not bring a gain. This is contrary to the expectations we had when reading the conclusions of [11].

For models D to F (based on UNet), the semi-supervised model (E) performs slightly better than the supervised model (D) (at least for the DICE and Jaccard metrics). Adding all the unlabeled images therefore seems to be of interest. However, the best performances are obtained by Model F (using pseudo-labeling). It therefore seems that the Cross Consistency training approach is less effective than the pseudo-labeling one in taking advantage of the unlabeled dataset.

6.2 Contact angle measurement

Moving beyond traditional metrics, we extended our investigation to explore the practical implications of the segmentation results. In material science applications, the final value measured (the contact angle value in our case) matter more than the score obtained by the model on some usual computer vision metric. We hypothesized that the accuracy of object segmentation could directly influence the computation of geometrical properties of objects within the images. To test this hypothesis, we employed the segmented outputs from the six models to compute the contact angle between the nanowire and the droplet.

We computed the Mean Squared Error (MSE) between the values obtained from model-based geometric properties and those determined manually by human annotators. This novel approach allowed us to quantify the models' ability to accurately capture object geometry.

We segmented frames of one video showing the growth of a nanowire and computed the contact angle value. The MSE scores of the six models are given in Table 2.

Table 3 shows the evolution of the contact angle, only for the frames annotated by a human. Figure 8 shows the evolution of the contact angle measured by Model F for each frame.

The results are broadly similar to those obtained in the previous section. UNet-based models make far fewer errors than PSPNet-based models. These PSPNet-based models therefore don't seem to be suited to our problem. Semi-supervised architectures (B, C and E) don't seem to be able to take advantage of unlabeled images. In this case, too, the most effective method is pseudo-labeling.

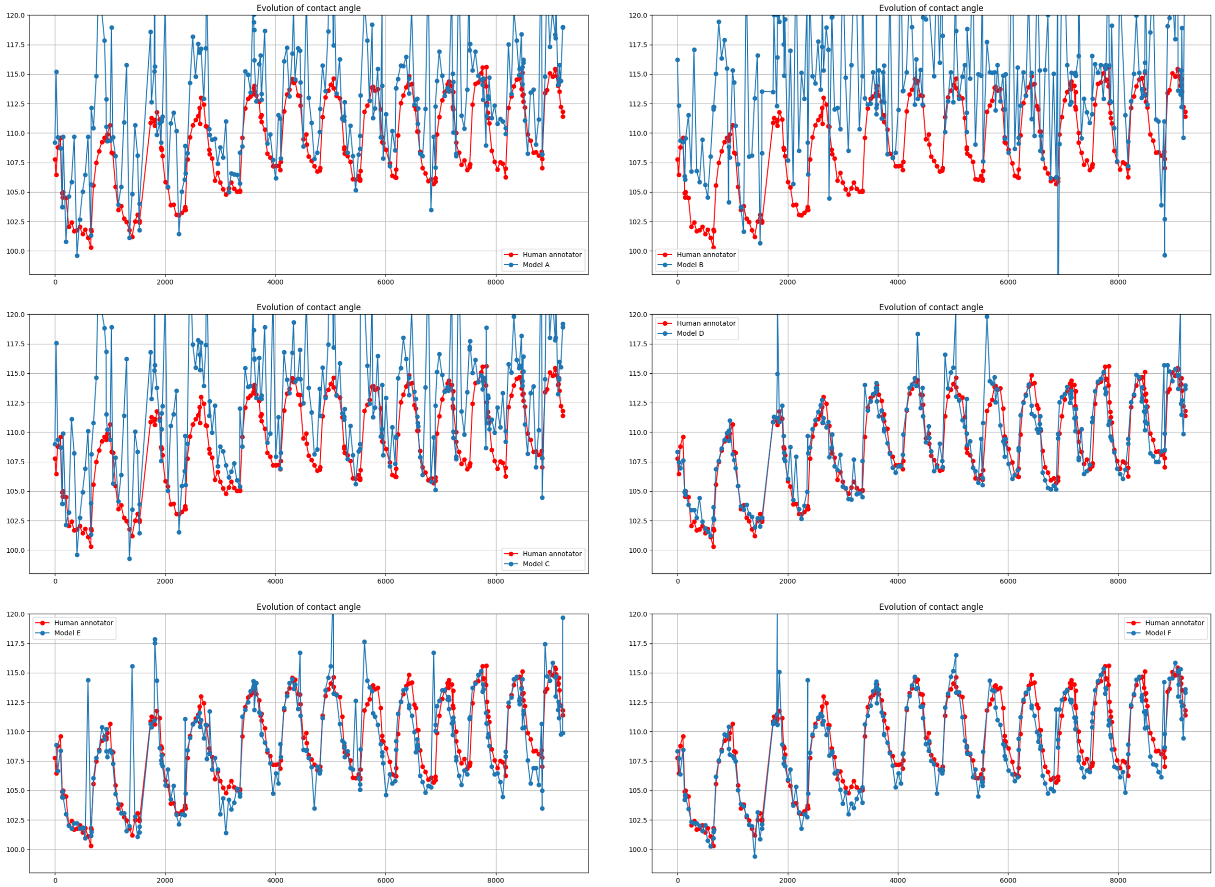

Table 3: Evolution of the contact angle, frames annotated by a human only

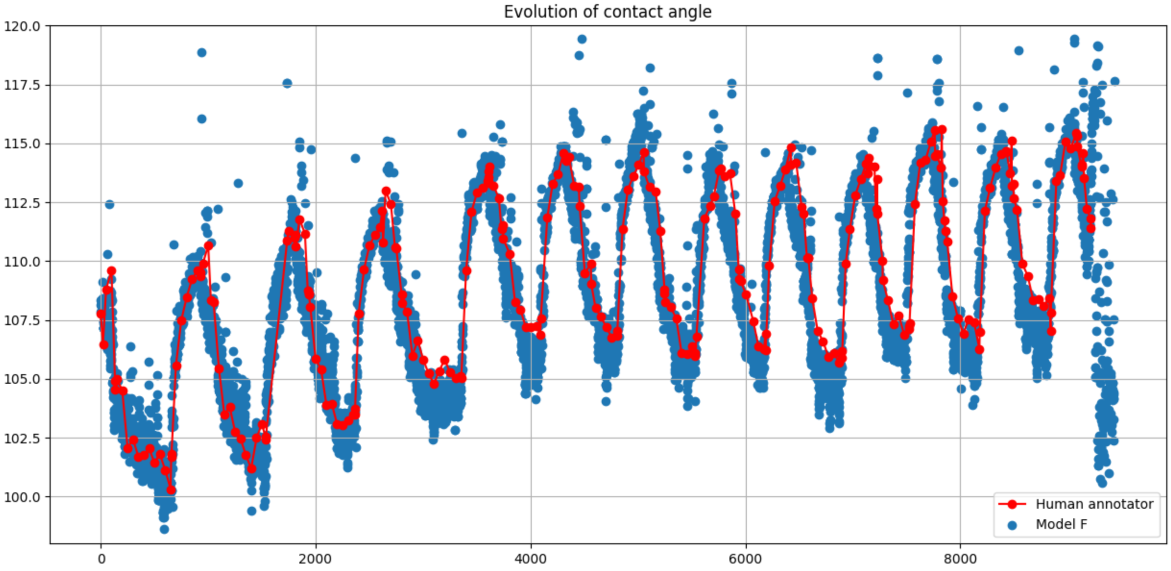

Figure 8: Evoluation of the contact angle, computed by Model F and a human annotator

7 Conclusion

In this work, we compare the performances of two paradigms and 2 architectures at segmenting and measuring nanowire in-situ TEM images. For future works, a possible direction is exploring the usage of labeled data generation, as a way to increase the size of the labeled dataset. It would also be interesting to adapt this study to other applications in material sciences.

References

[1] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, Jul. 2013.

[2] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV].

[3] J. Long, E. Shelhamer, and T. Darrell, *Fully convolutional networks for semantic segmentation*, 2015. arXiv: 1411.4038 [cs.CV].

[4] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015. arXiv: 1505.04597 [cs.CV].

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs*, 2017. arXiv: 1606.00915 [cs.CV].

[6] S. Laine and T. Aila, *Temporal ensembling for semi-supervised learning*, 2017. arXiv: 1610.02242 [cs.NE].

[7] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, *Pyramid scene parsing network*, 2017. arXiv: 1612.01105 [cs.CV].

[8] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, *Virtual adversarial training: A regularization method for supervised and semi-supervised learning*, 2018. arXiv: 1704.03976 [stat.ML].

[9] A. Tarvainen and H. Valpola, *Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results*, 2018. arXiv: 1703.01780 [cs.NE].

[10] J. P. Horwath, D. N. Zakharov, R. Megret, and E. A. Stach, *Understanding important features of deep learning models for transmission electron microscopy image segmentation*, 2019. arXiv: 1912.06077 [eess.IV].

[11] Y. Ouali, C. Hudelot, and M. Tami, *Semi-supervised semantic segmentation with cross-consistency training*, 2020. arXiv: 2003.09005 [cs.CV].

[12] F. Panciera, Z. Baraissov, G. Patriarche, *et al.*, "Phase selection in self-catalyzed gaas nanowires," *Nano Letters*, vol. 20, no. 3, pp. 1669–1675, 2020, PMID: 32027145. DOI: 10.1021/acs.nanolett.9b04808. eprint: https://doi.org/10.1021/acs.nanolett.9b04808. [Online]. Available: https://doi.org/10.1021/acs.nanolett.9b04808.

[13] J. Stuckner, B. Harder, and T. Smith, "Microstructure segmentation with deep learning encoders pre-trained on a large microscopy dataset," *npj Computational Materials*, vol. 8, Sep. 2022. DOI: 10.1038/s41524-022-00878-5.

[14] Z. Su, E. Decencière, N. Tun Tú, *et al.*, "Artificial neural network approach for multiphase segmentation of battery electrode nano-ct images," *npj Computational Materials*, vol. 8, Dec. 2022. DOI: 10.1038/s41524-022-00709-7.