Data-efficient U-Net for Segmentation of Carbide Microstructures in SEM Images of Steel Alloys

Alinda Ezgi Gerçek
Helmholtz AI Team Matter (FWCC-A)
Helmholtz-Zentrum Dresden-Rossendorf HZDR
01328 Dresden Germany

Till Korten
Helmholtz AI Team Matter (FWCC-A)
Helmholtz-Zentrum Dresden-Rossendorf HZDR
01328 Dresden Germany

Paul Chekhonin
Institute of Resource Ecology, Structural Materials Department
Helmholtz-Zentrum Dresden-Rossendorf HZDR
01328 Dresden Germany

Maleeha Hassan
Helmholtz AI Team Matter (FWCC-A)
Helmholtz-Zentrum Dresden-Rossendorf HZDR
01328 Dresden Germany

Peter Steinbach
Helmholtz AI Team Matter (FWCC-A)
Helmholtz-Zentrum Dresden-Rossendorf HZDR
01328 Dresden Germany
`t.korten@hzdr.de`

Abstract

Understanding reactor-pressure-vessel steel microstructure is crucial for predicting mechanical properties, as carbide precipitates both strengthen the alloy and can initiate cracks. In scanning electron microscopy images, gray-value overlap between carbides and matrix makes simple thresholding ineffective. We present a data-efficient segmentation pipeline using a lightweight U-Net (30.7 M parameters) trained on just **10 annotated scanning electron microscopy images**. Despite limited data, our model achieves a **Dice-Sørensen coefficient of 0.98**, significantly outperforming the state-of-the-art in the field of metallurgy (classical image analysis: 0.85), while reducing annotation effort by one order of magnitude compared to the state-of-the-art data efficient segmentation model. This approach enables rapid, automated carbide quantification for alloy design and generalizes to other steel types, demonstrating the potential of data-efficient deep learning in reactor-pressure-vessel steel analysis.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Machine Learning and and The Physical Sciences.

1 Introduction

The mechanical performance of reactor pressure-vessel (RPV) steels – and ferritic steel components in general – is influenced by secondary phases within the matrix. Among these, carbide precipitates have a dual effect: they hinder dislocation motion, increasing yield strength, whereas the larger carbides (about $0.5\,\mu\mathrm{m}$ equivalent circle diameter or above) may act as preferential sites for crack nucleation, promoting brittle failure and reducing fracture toughness [1–4]. Quantitative descriptors such as carbide number density, size distribution, and spatial arrangement are thus essential for physically-based models of strength and toughness.

Scanning electron microscopy (SEM) of mechanically polished cross-sections remains the standard for visualizing carbides, providing high-resolution, contrast-rich images over large fields of view. In practice, however, gray-level intensities often overlap with surrounding ferritic grains and particle edges may appear blurred. Classical image-analysis pipelines (e.g., denoising, background removal, thresholding and watershed instance segmentation combined with morphological operations) therefore produce often fragmented or spurious segmentation, while manual delineation of each carbide – though accurate – requires many hours of expert labor and impractical for the thousands of carbides needed for statistically robust analyses.

Advances in deep convolutional neural networks, especially the encoder-decoder U-Net architecture, have demonstrated state-of-the-art performance on biomedical and materials-science segmentation tasks when large annotated datasets are available [5] and even relatively small datasets with hundreds of samples can lead to satisfactory results [6]. Unfortunately, generating extensive pixel-wise labels, meaning manually delineated markings of carbides, for SEM micrographs is prohibitively expensive, and there is currently no systematic study investigating how few labeled examples are sufficient to achieve reliable carbide segmentation.

In this work we address the data-efficiency gap by training a lightweight U-Net (30.7 M trainable parameters; see Figure 1 for the network architecture) on only 10 of the 13 available manually annotated SEM images (image size 2048 x 1404 pixels; example in Figure 1). Our contributions are fourfold: (i) a data-efficient training strategy reduces annotation effort by an order of magnitude, (ii) provide uncertainty estimates by calibrating the output of the network to represent the model's confidence, using temperature scaling, (iii) benchmarking against a handcrafted classical image-analysis baseline, and (iv) demonstrating that the trained network generalizes to SEM images of a different steel (ANP-3).

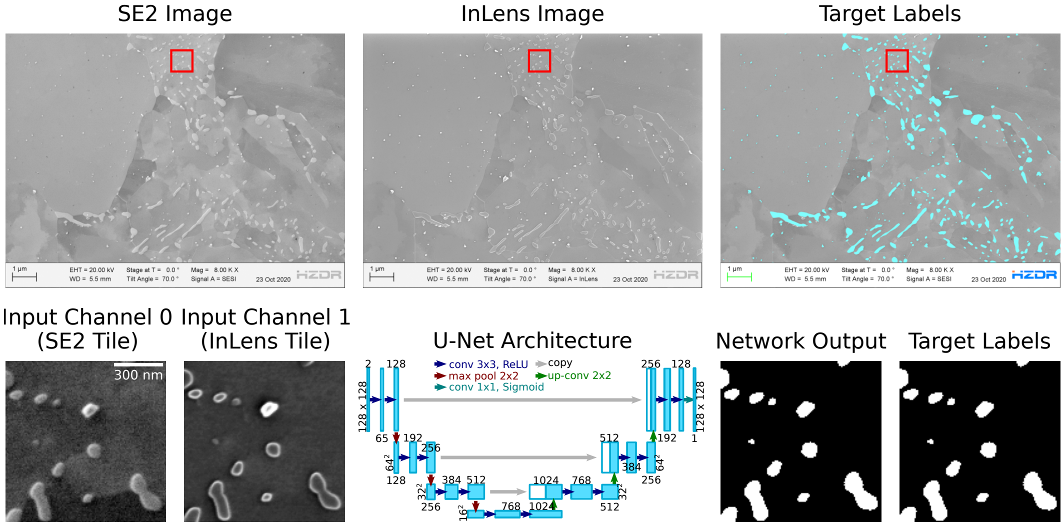

Figure 1: **Data and model architecture.** SEM images (top) of a RPV steel (JFL) acquired with SE and InLens detectors (top left and top center, respectively) and the corresponding manually annotated target label mask is overlaid in cyan onto the SEM image (top right). Red squares indicate the example tile shown in the bottom row as the two input channels (bottom left) for the U-net (bottom center; adapted from [5]) and the network output and corresponding labels (bottom right).

2 Methodology

Dataset The dataset consists of 13 pairs of SEM images of three RPV steels (material codes: JFL, ANP-10 and ANP-3). Images were acquired using Zeiss NVision 40 (JFL) and Zeiss Ultra 55 (ANP-10) microscopes with two secondary electron detectors (termed SE and InLens from now on) at 2048 x 1404 pixel resolution. The image width corresponds to $14.3\,\mu\text{m}$ and $11.5\,\mu\text{m}$ for JFL and ANP-10, respectively. More details on these steels can be found in [7, 4].

Target label masks were created by merging SE and InLens images (ratio 0.5), applying an initial gray-value threshold, and manually correcting to ensure accurate carbide outlines. The annotation process took approximately 20 hours total.

Preprocessing comprised cropping metadata bars, normalizing pixel intensities to [0, 1], and applying data augmentation (random rotations, flips, Gaussian noise, and blur). 12 images were tiled into 1920 non-overlapping 128 x 128 pixel tiles (randomly split into 80% training, 10% validation, 10% testing). One complete image acquired for a different steel type (ANP-3) was held out to evaluate generalization performance. Code and data to reproduce the results are available at [8, 9].

Classical Baseline A classical image-analysis pipeline was implemented using the python wrapper for the image processing library simpleitk [10, 11]. The pipeline merges the two imaging modalities (ratio 0.5), applies Gaussian denoising ($\sigma = 1.0$ pixels), performs background removal via top-hat filtering (radius=30 pixels), and segments carbides using Otsu's thresholding method that minimizes the combined variance of foreground and background [12]. Post-processing includes filling holes that are not connected to the boundary and removing small components (< 3 pixels) to eliminate noise.

Network Architecture and Training The segmentation model uses a U-Net architecture [5] with an encoder-decoder structure and skip connections. Our implementation features a 3-block encoder with 2D convolutions, batch normalization and ReLU activations, connected to a matching decoder via a bottleneck layer. The network has 30.7 Million parameters, making it relatively lightweight and suitable for small datasets. We used a dice loss function:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i} \tag{1}$$

where y_i and $\hat{y}_i$ are target and predicted labels for pixel i, respectively.

The network was trained using the Adam optimizer with initial learning rate 0.0002 and learning rate decay (factor 0.5, patience 7 epochs). Early stopping was applied with 14 epochs patience, and training used batch size 32, using $\approx$ 33 GB of RAM on one NVIDIA A100 GPU.

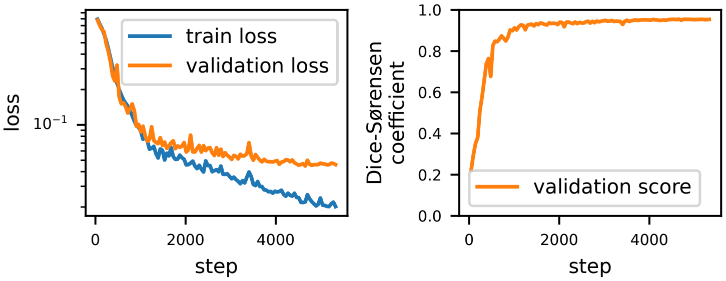

Figure 2: **Model training.** (left) Training and validation loss curves over 110 epochs. (right) Dice-Sørensen coefficient on the validation set.

Hyperparamter optimization was performed using optuna [13] over the following parameters: Starting learning rate, early stopping patience, number of features in the first encoder block and number of encoder blocks. The best hyperparameters were selected based on validation set performance (see Appendix A.2).

In total (including hyperparameter optimization) the training consumed approximately 30 hours on a single NVIDIA A100 GPU; corresponding to a power consumption of approximately 12 kWh.

Uncertainty Estimation Particularly for a small dataset – which may have gaps in distribution coverage – it is important to be able to assess how certain the model is about it's predictions. This uncertainty information can be useful for downstream tasks, such as active learning or decision-making based on the model's confidence. To calibrate the model's output probabilities, we applied temperature scaling [14] on the validation set. It is a post-hoc method where the single scalar temperature parameter T is learned on a held-out validation set by minimizing the negative log-likelihood, thereby softening the predicted probabilities. The temperature parameter T scales the logits before sigmoid activation: $\hat{y}_i = \sigma\left(\frac{z_i}{T}\right)$, where z_i is the logit for pixel i. The optimal $T = 1.87117$ was determined by minimizing negative log-likelihood on the validation set using L-BFGS [15], an optimization algorithm that uses an estimate of the inverse Hessian matrix with a lower memory profile than BFGS. The use of L-BFGS leverages the fact that T is a single scalar and the objective is smooth and deterministic, giving fast convergence with minimal hyperparameter tuning: From a starting value of 1, the algorithm converged to the optimal temperature in just 7 steps. See Appendix A.1 for details.

Statistics For statistical hypothesis testing we used the nonparametric Wilcoxon signed-rank test [16] with a significance level of $\alpha = 0.001$ for rejecting the null hypothesis that both samples were drawn from the same distribution.

Code and Data Availability Code and data to reproduce the results are available at [8, 9].

3 Results

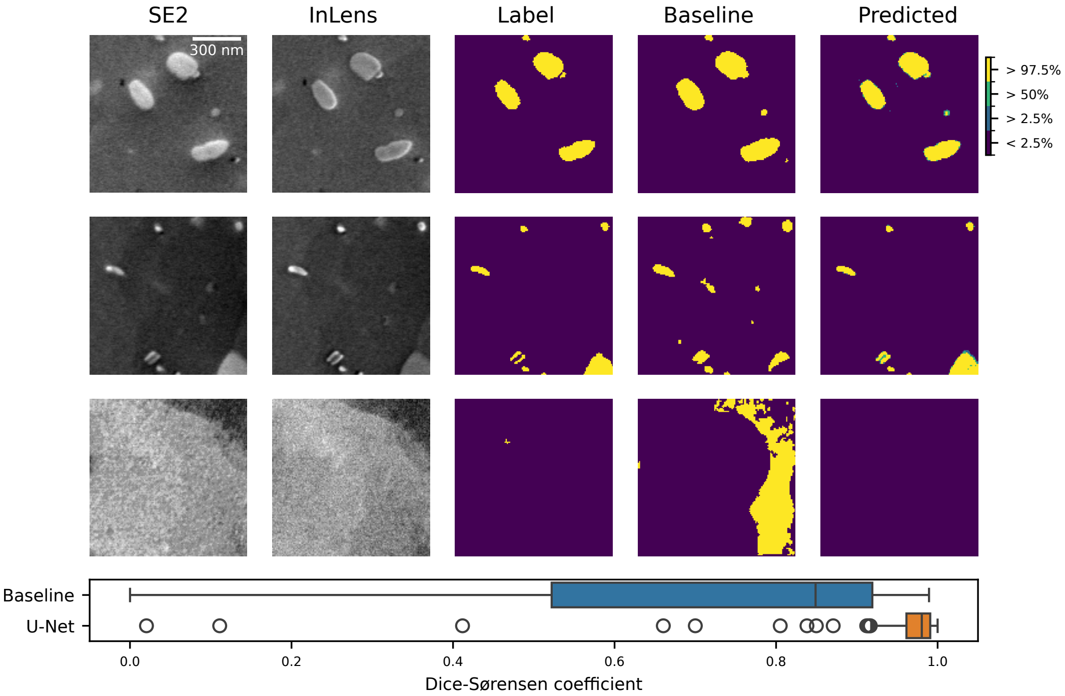

Figure 3: **Segmentation results.** (top) Example segmentation results on the test set. The model accurately segments carbides of varying sizes and shapes. (bottom) Box plot of Dice-Sørensen coefficients on the test set for the classical image-analysis baseline and our U-Net model. The U-Net significantly outperforms the baseline (Wilcoxon signed-rank test $p < 0.001$).

The model was evaluated on a held-out test set of 192 image tiles that were not used during training or validation. Randomly chosen example tiles are shown in Figure 3. Segmentation performance was quantified using the Dice-Sørensen coefficient [17, 18] defined as $\frac{2TP}{2TP+FP+FN}$ where TP is true positive, FP is false positive and FN is false negative. The U-Net model achieved a median (and interquartile range) Dice-Sørensen coefficient of 0.98 (0.964 - 0.991) on the test set, significantly outperforming the classical image-analysis baseline, which attained a median Dice-Sørensen coefficient of 0.85 (0.522 - 0.919) (Wilcoxon signed-rank test $p < 0.001$; see Figure 3). The model's confidence was lower in areas where it was wrong (green pixels in the predicted confidence maps in Figure 3), indicating that the uncertainty estimates are meaningful (see also Appendix A.1)).

To test how the model performs on a dataset acquired of a different sample, it was applied to a SEM image of a different steel type (ANP-3) acquired on a separate experiment day (Figure 4). The U-Net successfully segmented carbides in this image, still outperforming the baseline (Dice-Sørensen coefficient of 0.94, vs 0.90), demonstrating robustness and generalization to previously unseen experimental conditions.

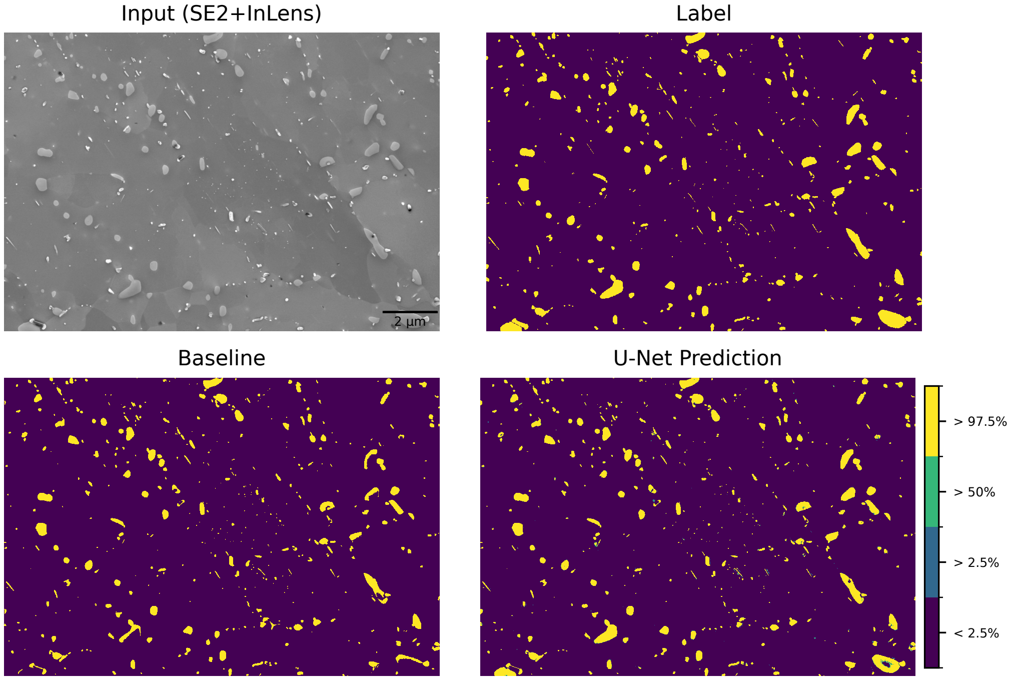

Figure 4: **Model generalization.** Application of the trained U-Net to SEM images from a different steel (ANP-3). Accurate segmentation of carbides confirms that the model generalizes beyond the training dataset. The baseline achieved a Dice-Sørensen coefficient of 0.90, while the U-Net achieved 0.94.

4 Discussion and Outlook

The results show that high-fidelity segmentation of carbide precipitates in SEM images of RPV steels can be achieved with a lightweight U-Net trained on only 10 annotated images, a 10 fold reduction in labeling effort compared to the state-of-the-art data efficient image segmentation model [6]. Extensive data augmentation and a carefully designed training strategy mitigated over-fitting, allowing the model to generalize well to unseen data. The U-Net substantially outperformed the classical image-analysis baseline in both median Dice-Sørensen coefficient and consistency across tiles. Unlike the classical image-analysis baseline, the model does not require manual corrections, demonstrating the model's potential to automate labor-intensive microstructural quantification tasks. The model maintained high accuracy on SEM images from a different steel (ANP-3), highlighting robustness and practical generalization to other steel types.

Limitations While the model generalized well to a different imaging session, further validation on larger, more diverse datasets is needed to confirm robustness. Additionally, the current approach relies on fully supervised learning with manually annotated labels, which are time-consuming to generate.

Outlook Future work could explore semi-supervised or self-supervised methods to leverage unlabeled data and reduce annotation effort further.

Acknowledgments and Disclosure of Funding

We thank the entire Team of Helmholtz AI Matter for invaluable discussions and a great working atmosphere.

Author contributions

All authors contributed to writing the manuscript. In addition:

AG and TK conducted data processing, trained the machine learning model, and prepared the experimental figures.

PC prepared the samples, acquired the SEM images and created label data

MH performed uncertainty quantification

PS coordinated and provided guidance on machine learning and uncertainty quantification techniques

Funding

The work of TK and PS was funded by Helmholtz Incubator Platform Helmholtz AI. The work of TK and MH was funded by the trustPAIKON SAB project.

References

[1] D. A. Curry and J. F. Knott. Effects of microstructure on cleavage fracture stress in steel. *Metal Science*, 12(11):511–514, November 1978. ISSN 0306-3453. doi: 10.1179/msc.1978.12.11.511. URL `https://doi.org/10.1179/msc.1978.12.11.511`. Publisher: Taylor & Francis _eprint: https://doi.org/10.1179/msc.1978.12.11.511.

[2] X. Z. Zhang and J. F. Knott. Cleavage fracture in bainitic and martensitic microstructures. *Acta Materialia*, 47(12):3483–3495, September 1999. ISSN 1359-6454. doi: 10.1016/S1359-6454(99)00200-1. URL `https://www.sciencedirect.com/science/article/pii/S1359645499002001`.

[3] S Lee, S Kim, B Hwang, B. S Lee, and C. G Lee. Effect of carbide distribution on the fracture toughness in the transition temperature region of an SA 508 steel. *Acta Materialia*, 50(19): 4755–4762, November 2002. ISSN 1359-6454. doi: 10.1016/S1359-6454(02)00313-0. URL `https://www.sciencedirect.com/science/article/pii/S1359645402003130`.

[4] Paul Chekhonin, Aniruddh Das, Frank Bergner, and Eberhard Altstadt. Microstructural characterisation of brittle fracture initiation sites in reactor pressure vessel steels. *Nuclear Materials and Energy*, 37:101511, December 2023. ISSN 2352-1791. doi: 10.1016/j.nme.2023.101511. URL `https://www.sciencedirect.com/science/article/pii/S2352179123001503`.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. doi: 10.1007/978-3-319-24574-4_28.

[6] Michelle Bardis, Roozbeh Houshyar, Chanon Chantaduly, Alexander Ushinsky, Justin Glavis-Bloom, Madeleine Shaver, Daniel Chow, Edward Uchio, and Peter Chang. Deep Learning with Limited Data: Organ Segmentation Performance by U-Net. *Electronics*, 9(8):1199, August 2020. ISSN 2079-9292. doi: 10.3390/electronics9081199. URL `https://www.mdpi.com/2079-9292/9/8/1199`. Publisher: Multidisciplinary Digital Publishing Institute.

[7] Libang Lai, Paul Chekhonin, Shavkat Akhmadaliev, Jann-Erik Brandenburg, and Frank Bergner. Microstructural Characterization of Reactor Pressure Vessel Steels. *Metals*, 13 (8):1339, August 2023. ISSN 2075-4701. doi: 10.3390/met13081339. URL `https://www.mdpi.com/2075-4701/13/8/1339`. Publisher: Multidisciplinary Digital Publishing Institute.

[8] Paul Chekhonin, Till Korten, Alinda Ezgi Gerçek, Maleeha Hassan, and Peter Steinbach. Training Data and Models for the paper: Data-efficient U-Net for Segmentation of Carbide Microstructures in SEM Images of Steel Alloys, November 2025. URL `https://rodare.hzdr.de/record/4124`.

[9] Till Korten, Alinda Ezgi Gerçek, and Maleeha Hassan. Automated carbide detection, August 2025. URL `https://github.com/thawn/carde/`.

[10] Bradley Christopher Lowekamp, David T. Chen, Luis Ibanez, and Daniel Blezek. The Design of SimpleITK. *Frontiers in Neuroinformatics*, 7, December 2013. ISSN 1662-5196. doi: 10.3389/fninf.2013.00045. URL `https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2013.00045/full`. Publisher: Frontiers.

[11] Richard Beare, Bradley Lowekamp, and Ziv Yaniv. Image Segmentation, Registration and Characterization in R with SimpleITK. *Journal of Statistical Software*, 86:1–35, September 2018. ISSN 1548-7660. doi: 10.18637/jss.v086.i08. URL `https://doi.org/10.18637/jss.v086.i08`.

[12] Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, January 1979. ISSN 2168-2909. doi: 10.1109/TSMC.1979.4310076. URL `https://dspace.tul.cz/server/api/core/bitstreams/36abcc1c-cd72-4569-90ed-607017063124/content`. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics.

[13] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2623–2631, New York, NY, USA, July 2019. Association for Computing Machinery. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330701. URL `https://doi.org/10.1145/3292500.3330701`.

[14] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, July 2017. URL `https://proceedings.mlr.press/v70/guo17a.html`. ISSN: 2640-3498.

[15] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, August 1989. ISSN 1436-4646. doi: 10.1007/BF01589116. URL `https://doi.org/10.1007/BF01589116`.

[16] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6): 80–83, 1945. ISSN 0099-4987. doi: 10.2307/3001968. URL `https://www.jstor.org/stable/3001968`. Publisher: [International Biometric Society, Wiley].

[17] Lee R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, 1945. ISSN 1939-9170. doi: 10.2307/1932409. URL `https://onlinelibrary.wiley.com/doi/abs/10.2307/1932409`. _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.2307/1932409.

[18] Thorvald Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab*, 5(4):1–34, 1948. URL `https://cir.nii.ac.jp/crid/1370302864784225926`. Pages: 1 Publication Title: Biol Skrifter Volume: 5.

[19] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html`.

A Appendix / supplemental material

A.1 Uncertainty Estimation

Temperature scaling To estimate the uncertainty of the model's predictions, we applied temperature scaling [14] to calibrate the output probabilities. The temperature parameter T was optimized on the validation set by minimizing the negative log-likelihood using the L-BFGS optimization algorithm [15]. The optimal temperature was found to be $T = 1.87117$. The calibration process ensures that the predicted probabilities better reflect the true likelihood of a pixel belonging to the carbide class. Figure 5 shows the reliability diagram before and after temperature scaling, demonstrating improved calibration of the model's output probabilities (smaller difference to perfect calibration).

Mean-variance estimation To estimate the aleatoric (data) uncertainty, we applied mean variance estimation by adding another output layer representing the variance of the predicted logits. Following the approach of Kendall and Gal [19], we used a Monte-Carlo approach to estimate the Dice loss of a set of predictions drawn randomly from a Gaussian distribution with the predicted bias and variance of the respective logit. See the code [9] for implementation details.

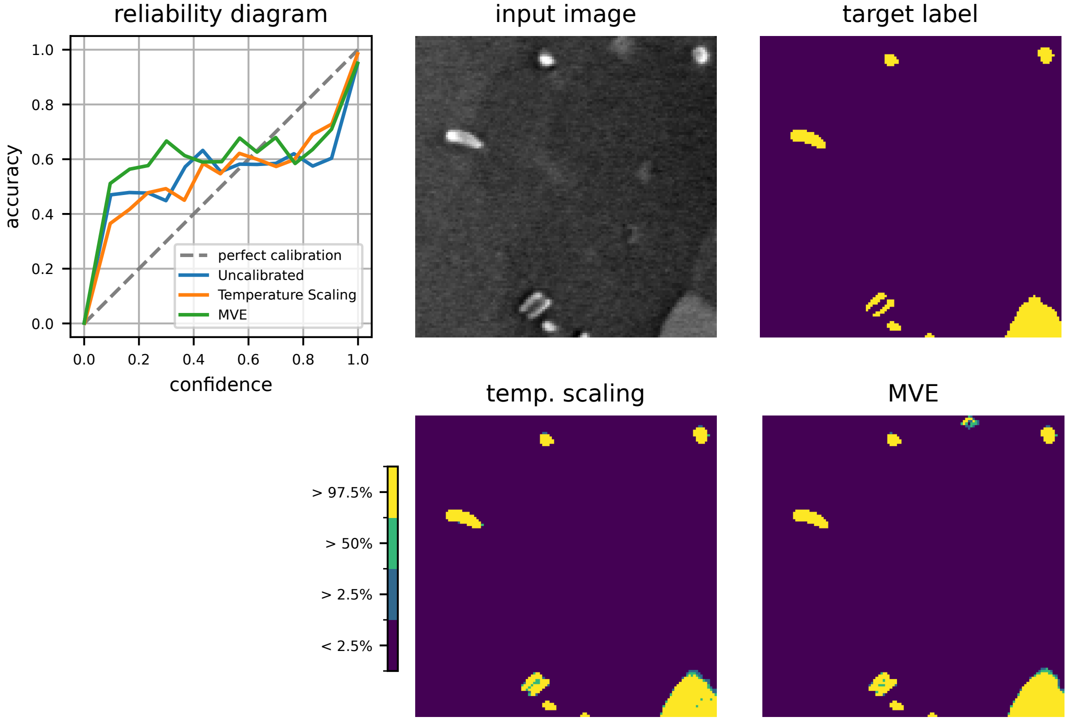

Figure 5: **Uncertainty estimation** (top right) Reliability diagram before and after temperature scaling. The reliability diagram shows the relationship between predicted probabilities and observed frequencies of the positive class (carbide pixels). The diagonal line represents perfect calibration, where predicted probabilities match observed frequencies. After applying temperature scaling with $T = 1.87117$ (orange line), the predictions are better calibrated than without any calibration (blue line) and also better calibrated than with mean-variance estimation (green line), as shown by the curve being closer to the diagonal line. (top center) input image, (top right) target label, (bottom center) model output after temperature scaling, (bottom right) model output with mean-variance estimation. The colors represent the model's confidence in its predictions, with dark green representing pixels classified as true with low confidence, light green medium confidence and yellow high confidence (as indicated in the color bar in the bottom right).

The model's confidence was lower in areas where it was wrong (see Figure 5), indicating that the uncertainty estimates are meaningful. This uncertainty information can be useful for downstream tasks, such as active learning or decision-making based on the model's confidence. Notably, the reliability diagram (Figure 5 bottom left) shows that even after calibration, the model is underconfident in a confidence-range of 0.1 – 0.5 and over-confident in a confidence-range of 0.5 – 0.9. This is likely caused by high labeling noise around the boundaries of objects. This noise stems from the fact that the labels were generated by a thresholding operation, which is affected by truly random imaging noise and hence inherently unpredictable.

A.2 Hyperparamter Optimization

We used the optuna framework [13] to perform hyperparameter optimization over the following parameters: Starting learning rate, early stopping patience, number of features in the first encoder block and number of encoder blocks. The optimization objective was to minimize the Dice-Sørensen coefficient for the validation set. The optimal hyperparameters found were: Starting learning rate = 0.0002, early stopping patience = 14 epochs, number of features in the first encoder block = 128 and number of encoder blocks = 3.

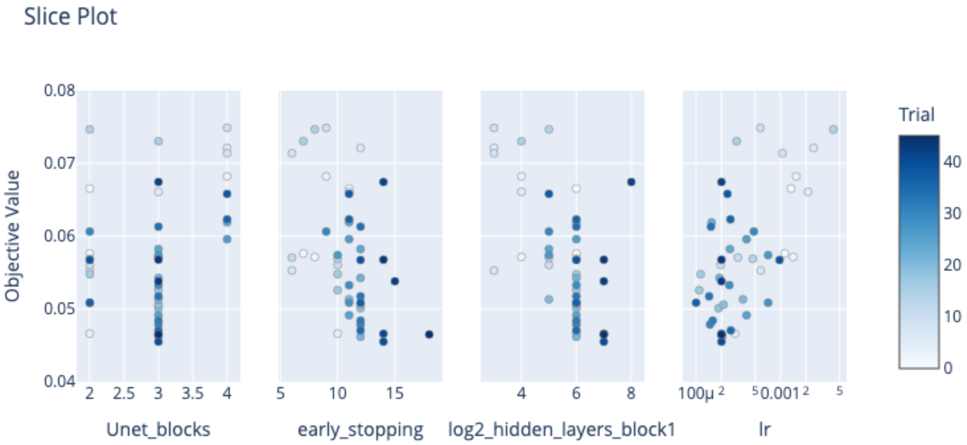

Figure 6: **Hyperparameter optimization.** Slice plot showing the relationship between hyperparameters and validation loss. Each point represents a trial with a specific combination of hyperparameters, and the color indicates the trial number.