A Deep Learning Approach for Complex Microstructure Inference

Ali Riza Durmaz[1,2,*,+], Martin Müller[3,4,+], Bo Lei[5], Akhil Thomas[1], Dominik Britz[3,4], Elizabeth A. Holm[5], Chris Eberl[1], Frank Mücklich[3,4], and Peter Gumbsch[1,2]

[1]Fraunhofer Institute for Mechanics of Materials IWM, Freiburg, 79108, Germany
[2]Karlsruhe Institute of Technology (KIT), Institute for Applied Materials IAM, Karlsruhe, 76131, Germany
[3]Saarland University, Department of Materials Science, Saarbrücken, 66123, Germany
[4]Material Engineering Center Saarland, Saarbrücken, 66123, Germany
[5]Carnegie Mellon University, Department of Materials Science and Engineering, Pittsburgh, PA 15213, USA
[*]ali.riza.durmaz@iwm.fraunhofer.de
[+]these authors contributed equally to this work

ABSTRACT

Automated, reliable, and objective microstructure inference from micrographs is an essential milestone towards a comprehensive understanding of process-microstructure-property relations and tailored materials development. However, such inference, with the increasing complexity of microstructures, requires advanced segmentation methodologies. While deep learning (DL), in principle, offers new opportunities for this task, an intuition about the required data quality and quantity and an extensive methodological DL guideline for microstructure quantification and classification are still missing. This, along with a lack of open-access data sets and the seemingly intransparent decision-making process of DL models, hampers its breakthrough in this field. We address all aforementioned obstacles by a multidisciplinary DL approach, devoting equal attention to specimen preparation, contrasting, and imaging. To this end, we train distinct U-Net architectures with 30–50 micrographs of different imaging modalities and corresponding EBSD-informed annotations. On the challenging task of lath-bainite segmentation in complex-phase steel, we achieve accuracies of 90% rivaling expert segmentations. Further, we discuss the impact of image context, pre-training with domain-extrinsic data, and data augmentation. Network visualization techniques demonstrate plausible model decisions based on grain boundary morphology and triple points. As a result, we resolve preconceptions about required data amounts and interpretability to pave the way for DL's day-to-day application for microstructure quantification.

1 Introduction

Deep learning (DL) is a lasting subject of attention and achieved remarkable results that culminated in a paradigm shift in computer vision. In particular, research fields such as autonomous driving[1], and biomedicine[2,3] acted as driving forces for the development of data-driven approaches, which superseded conventional computer vision (CV) algorithms to a large extent. The introduction of convolutional neural networks (CNN) with versatile architectures was accompanied by substantial improvements in CV tasks[4]. This was rendered possible by the accessibility of source codes and open-access data sets, which enabled the comparability of different modeling approaches and thus steady improvement.

Quality control in materials processing or of safety-critical components, as well as tailored materials development[5] require the segmentation and classification of material microstructures. Segmentation here refers to the pixel/voxel-wise materials phase assignment. It is indispensable when relating microstructure with properties, e.g., the phase morphology with fatigue properties[6]. Predominantly, 2D micrographs of different imaging modalities, such as light optical microscopy (LOM) or scanning electron microscopy (SEM), are utilized for microstructure inference.

However, such micrographs' automated, reliable, and objective segmentation is not established for all desirable material classes. Although DL has more than proven its capability for image segmentation and classification, it is still waiting for its breakthrough in materials science. This can be attributed to DL being frequently associated with a few drawbacks. Namely, the requirement for (very) large data quantities and the black-box nature of CNNs concerning the intransparency of their decisions[7]. Furthermore, microstructure recognition tasks, compared to real-world images, can be very complex regarding the degree of detail and information density in the images. This further impedes the determination of accurate annotations[8] needed for supervised-learning, which may discourage the use of DL, ultimately resulting in a lack of representative annotated open-access data sets.

Hence, there is no practical guide on suitable specimen preparation and contrasting, data acquisition and processing, and no general intuition about the quality and quantity of data required to train a specific DL architecture in the material science domain.

Consequently, material scientists' recurrent questions address the required amount of training data, resolutions, annotation accuracy, model architectures, and training strategies.

The work's primary objective is to tackle former questions and provide a better grasp through an integral approach systematically investigating methodological interdependencies in the whole metallographic and DL process chain. Moreover, a CNN's decision-making process is rendered more transparent by investigating the importance of certain microstructural features for the CNN prediction. Using the microstructure of a complex phase (CP) steel, and particularly its lath-shaped bainitic phase, as a case study, we demonstrate DL's relevance in the field and aim to raise the awareness and acceptance of DL for such tasks. This microstructure class exhibits pronounced importance in engineering, and its constituents can only be segmented to a minimal extent using classic CV approaches[9].

According to the classification scheme suggested by Zajac[10], the microstructure of CP steels, a family of advanced high-strength steels, typically consists of bainite (granular, upper, or degenerate upper bainite), ferrite, and dispersed carbon-rich additional phases like martensite or retained austenite. In micrographs of such heterogeneous microstructures, not all constituents can be distinguished through gray value distribution. Therefore, simple, traditional segmentation methods operating on LOM or SEM quickly reach their limit. Approaches to quantify the separate microstructure constituents using EBSD individually have been reported[11,12]. Müller et al.[13] developed a procedure to segment lath-shaped bainite in CP steel micrographs consisting of lath-shaped and granular bainite by analyzing the microstructure constituents' directionality. Bulgarevich et al.[14] used a trainable segmentation with a random forest classifier to segment ferrite, pearlite, and bainite in light optical micrographs of three-phase steels. Although methods for quantifying multi-phase microstructures have been suggested, the annotation and efficient segmentation of different microstructure constituents solely from LOM or SEM micrographs are not satisfactory.

As opposed to these works, supporting *correlative* electron backscatter diffraction (EBSD) information is used in the LOM and SEM annotation procedure to lay an appropriate foundation for learning. Moreover, the aforementioned conventional CV or ML approaches require complex image processing pipelines and elaborate feature engineering to render predictions robust against variances[13]. In contrast, the applied DL methods are directly based on input and target output image pairs (representation learning). Their application to microstructure recognition demonstrated the potential for quantitative microstructure analysis[15,16], steel type classification[15], crack path prediction[17], and micromechanical damage segmentation[18]. A CNN architecture referred to as U-Net[2] has proven its merit in the latter work and represents a common starting point due to its numerous implementations in different DL frameworks and image processing tools[19,20]. Therefore, this architecture represents a suitable candidate to derive best practices.

2 Methods

2.1 Material

The material used in this study is a low-carbon CP steel, taken from industrially produced heavy plates. Steels were thermo-mechanically rolled, followed by controlled accelerated cooling. The segmentation task is to distinguish between lath-shaped bainite (hereafter called foreground) and the background, consisting of polygonal and irregular ferrite with dispersed granular carbon-rich 2nd phase.

2.2 Specimen preparation

Specimens were taken in the plate's rolling direction, ground using 80–1200 grid SiC papers, and then subjected to 6, 3, and finally, 1 µm diamond polishing. For LOM investigation, metallographic etching was carried out by immersing polished specimen surfaces into a mixture of ethanol and nitric acid (2 vol.-%), also called "Nital" etching. For SEM examination, the specimens were etched electrolytically using Struers electrolyte A2. Nital and electrolytic etching were chosen because they attack and thus reveal grain boundaries. This contrasting step is crucial to identify the boundaries of lath-shaped bainite regions during annotation. For investigation by electron backscatter diffraction, colloidal oxide polishing was additionally performed after diamond polishing.

2.3 Microscopy

Light Optical Microscopy. For imaging, the LOM in an Olympus LEXT OLS 4100 laser scanning microscope was used. Images were taken at a magnification of 1000× with an image size of 1024×1024 pixels, corresponding to an area of 129.6×129.6 µm^2 (pixel size = 126.6 nm). All images were acquired with the same image contrast and brightness settings.
Scanning Electron Microscopy. SEM images were recorded in a Zeiss Merlin FEG-SEM using secondary electron contrast at a magnification of 2000× with an image size of 2048×1536 pixels, equal to 56.7×42.5 µm^2 (pixel size = 27.7 nm). The SEM was operated at an acceleration voltage of 5 kV, a probe current of 300 pA, and a working distance of 5 mm. All images were acquired with the same image contrast and brightness settings in the SEM.

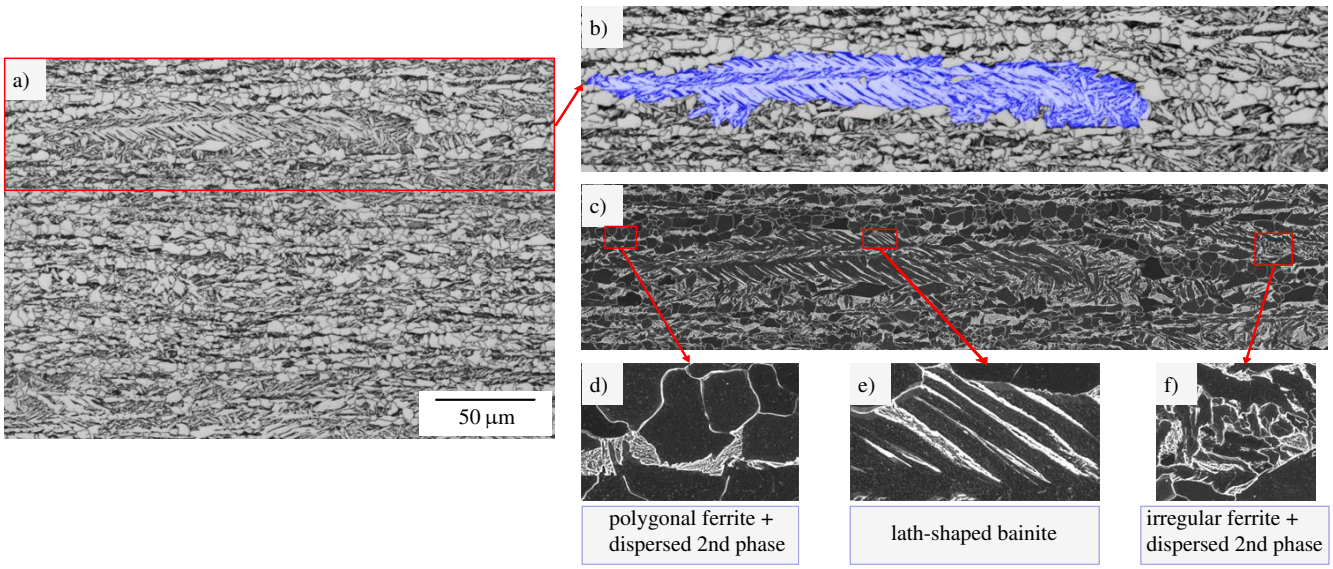

Figure 1. (a) LOM micrograph of CP steel microstructure after Nital etching. (b) enlarged detail from (a), showing an annotated lath-shaped bainite region (blue). (c) correlative SEM micrograph of (b). The enlarged detail figures (d), (e) and (f) depict polygonal ferrite with dispersed carbon-rich 2nd phase, lath-shaped bainite and irregular ferrite with dispersed carbon-rich 2nd phase, respectively. In the SEM modality the carbon-rich 2nd phase particles appear bright.

Correlative Microscopy. In a correlative approach, LOM and SEM were combined with EBSD characterization. The specimen regions of interest were marked by hardness indents for consistent imaging in different modalities. EBSD measurements were performed in a Zeiss Merlin FEG-SEM with an acceleration voltage of 25 kV, a probe current of 10 nA, and 15 mm working distance. Scans were done at a magnification of 200× with a step size of 0.35 µm using a hexagonal grid. Data were analyzed using software OIM TSL Analysis. As cleanup, neighbor confidence index (CI) correlation (CI $\geq$ 0.01) and neighbor orientation correlation (5° grain tolerance angle) were applied. After EBSD measurements, specimens were etched, and micrographs from the same regions of interest were taken in LOM and SEM, as described above. Several such micrographs were stitched together using Microsoft Image Composite Editor to match the EBSD scanned region.

When combining different imaging techniques, the different micrographs must be aligned. This process is referred to as multi-modal image registration and is accompanied by challenges including different specimen states, viewpoints, contrasts, and fields of view[21]. For a general explanation of challenges during correlative characterization and image registration in metallography, the authors refer to Britz et. al.[22].

For registering EBSD maps with LOM and SEM images, the open-source tool ImageJ and its plugins SIFT feature extraction and bUnwarpJ registration were used[22]. First, the Scale-Invariant-Feature Transform (SIFT)[23] algorithm was used to find the same features in both the EBSD map and the LOM/SEM image. For this purpose, the EBSD image quality map[24] was chosen due to its pronounced similarity to the other modalities. The common features extracted by SIFT facilitate the registration using the bUnwarpJ[25] algorithm. Thereby a transformation matrix is determined that is applied to register other EBSD-derived maps, e.g., misorientation maps.

2.4 Data Set Preparation

Annotations for deep learning segmentation. Labeling of images was done manually by human experts on a Wacom Tablet. Since human labeling based solely on the microstructure's visual appearance in topography-sensitive LOM or SEM images can be subjective, parameters from correlative EBSD measurements were used as additional information to annotate the micrographs more objectively. Reasonable EBSD-derived information that assisted the annotation included grain structure visualizations as well as intergranular and intragranular misorientation metrics. Namely, unique grain color, grain boundary, grain orientation spread (GOS), grain average misorientation (GAM), and kernel average misorientation (KAM, with 3rd order neighbors) maps were considered.

Because of time constraints, it is typically not feasible to obtain high-fidelity annotations through correlative EBSD measurements for the comparatively large image sets required for DL. Therefore, correlative EBSD measurements were collected only for a subset of images, and the knowledge and experience gained from the fused data was translated to regular LOM and SEM images. Therefore, the correlative measurements can be regarded as references for the whole data set. Under

these circumstances, well-founded and more objective annotations can be accomplished by human experts also without the EBSD data.

The final LOM image set consists of 51 micrographs with corresponding masks for the segmentation (1024×1024 pixels, approx. 28% lath-shaped bainite on average per image) and the final SEM image set of 36 micrographs with corresponding masks (2048×1433 pixels due to cropping of the SEM annotation bar, approx. 60% lath-shaped bainite on average per image).

Data pre-processing for model training. To comply with network architectural and computational memory constraints, the raw input and derived label images were cropped into tiles for both imaging modalities. Before extracting tiles with a fixed resolution, an optional downscaling step by a factor of 0.5 in both spatial directions was performed to study the image context-dependence.

Since convolutions can not be computed at the immediate image border, tile images are often artificially extended at the border by few pixels (*padding*) at each convolution layer. This is used in the U-Net VGG16 variant by repeating the tile border pixels. In contrast, by virtue of the vanilla U-Net and its unpadded convolutions, the prediction exhibits a reduced tile resolution compared to the network input. To counteract this, the tiles were extracted with an offset that amounted to 62 pixels at all tile edges to take the whole data into account and ensure data efficiency. As a consequence, adjacent input tiles had an overlap, but the network's predicted tiles were adjoining[2]. For tiles originating from the raw image border, this was facilitated by applying mirror padding at the border of full frame images before extracting tiles. In case of previous downscaling or uneven raw image resolutions, such as in SEM, the aforementioned border padding was combined with a minor additional padding that ensured resolution conformity throughout the forward pass of the vanilla U-Net network. In total, four DL data sets were derived from the LOM and SEM raw data sets —- native and scaled versions of both image modalities. Since the SEM images were acquired with higher magnification, the raw images covered a substantially smaller field of view. Tiles with 380×380 pixels and 636×636 pixels were extracted for the LOM and SEM modality, respectively, to assimilate their contained image context. Input images of both modalities were converted into grayscale. A summary of the data sets, including some characteristic metrics, can be found in the Supplemental. The resulting tiles were randomly sampled in five data portions for k-fold cross-validation (unstratified) with k = 5. For each of the five folds the test set was altered to be one of the data portions. Hence, the five distinct data sets contained 80% and 20% of the total tile amount for training and testing, respectively.

2.5 Deep learning methodology

Two research groups collaborated on this segmentation task using their respective approaches and best practices. Both approaches are based on the U-Net architecture, but still contain various differences. By comparing their results, important conclusions regarding the universal applicability and robustness of deep learning techniques for the segmentation of CP steels can be drawn.

Deep learning segmentation approach 1 – Vanilla U-Net. A vanilla U-Net model with an architecture implemented in the PyTorch framework[26] that included few adjustments with respect to[18] was trained from scratch. Only two class channels in the output were used since the work at hand covers a binary segmentation problem. Furthermore, batch normalization was incorporated after convolutions to accelerate the training procedure by smoothing the optimization function[27]. The U-Net architecture has four levels, discards padding for the non-dilated 3×3 convolutions in the encoder, utilizes 2×2 max-pooling, applies "same" padding for transposed convolutions in the decoder, and contains skip-connections[28] between the corresponding encoder and decoder levels. A schematic of the architecture showing the designation of the individual layers is accessible in the Supplemental since network visualization techniques shown hereafter are referring to specific layers. Different data augmentation techniques from the Albumentations package[29] were applied to investigate their impact on the performance. In contrast to our prior study[18], a more systematic approach to data augmentation was taken by applying grid, and random search for optimization of relevant hyperparameters in Tune[30]. The set of optimized augmentation parameters for both image modalities is outlined in the Supplemental. For training, the focal loss function[31], and an Adam optimizer[32] was used. Each model was trained for 250 epochs on a NVIDIA Tesla V100 GPU with 32GB memory and CUDA (v10.0) acceleration. The α and γ parameters of the focal loss function were also considered during hyperparameter optimization to account for data set class imbalances.

Deep learning segmentation approach 2 – U-Net with VGG16 backbone. A U-Net model variant with a VGG16 encoder that was pre-trained on ImageNet[33] was applied. The model was implemented in PyTorch, and the pre-trained weights were from torchvision. A schematic of the architecture is depicted in the Supplemental. The initial five convolution blocks of VGG16 represent the four encoder levels and center level of the U-Net. The decoder contains four upsampling blocks. Each upsampling block contains a bilinear upsampling and two convolutional layers with batch normalization and Relu activation. Skip-connections are applied identically to the regular U-Net. In Table 1 relevant architectural differences are listed. Note that due to padded convolutions in the encoder, the image shape is maintained by this model. Therefore, border mirror padding is not required, and tile images were center-cropped to 256×256 pixels for LOM images and 512×512 pixels for SEM images. The major difference compared to segmentation approach #1 in the training procedure is that the U-Net VGG16 leveraged a

pre-trained encoder. For performance optimization in the training process, data augmentation (see Supplemental), cross-entropy loss, and an Adam optimizer[32] with cosine annealing schedule are utilized. No measures for the correction of class imbalance were taken.

Table 1. Architectural differences between the vanilla U-Net and the U-Net VGG16.

	Vanilla U-Net	**U-Net VGG16**
Upsampling approach	Transposed convolutions (learnable) halving channels	Bilinear Interpolation maintaining channels
Convolution approach encoder	Unpadded, contains two convolution layers in each level	Padded, contains three convolution layers in later encoder and center levels
Batch normalization	Encoder and decoder	Decoder

The model performances are evaluated in terms of the *Accuracy* metric and the intersection over union (*IoU*), also referred to as Jaccard index, defined as follows.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$IoU = \frac{TP}{TP+FP+FN} \tag{2}$$

Here, TP, TN, FP and FN are the true positive, true negative, false positive, and false negative pixel amounts, respectively. Both metrics are defined in the range from zero to unity (or 0–100%), where latter corresponds to an ideal model prediction. The accuracy metric measures the correctly predicted pixel percentage, while the IoU measures the ratio between intersection and union of predicted and labeled pixel areas. We provide the accuracy metric due to its intuitiveness and despite its limited sensitivity in case of notable class imbalances, such as in the LOM case. In contrast, the IoU captures the model differences more adequately for data sets skewed towards the negative class, which is why we focus on it for the comparison between the individual models.

Network visualization techniques In order to render model decisions explainable, the *Network Dissection*[34] and *Gradient Weighted Class Activation Maps (Grad-CAM)*[35] visualization methods were used. The objective of the Network Dissection method is to visualize concepts that were learned by individual filters in specific layers. This is achieved by evaluating activation maps, i.e., single channels of the activation function output, with regard to its spatial attention for regions in the input image. In particular, activation maps were thresholded such that the largest 1.0% of the activation map is obtained. In the original implementation, the thresholded activation maps were then resized to input image resolution and subsequently superimposed. However, since the encoder used unpadded convolutions, in the vanilla U-Net case, a combination of resizing and padding was required.

Grad-CAM, on the other hand, aims to shed light on the decision-making process of models. This technique originally focused on providing a class discriminative localization map for the output convolutional layer for a given input image, highlighting the important regions in the image for a particular class prediction. However, it is also applicable for any convolutional layer in a network. The localization map for a convolutional layer is constructed by a weighted combination of feature maps of that layer for a given input image. The weights for feature maps are computed by propagating the gradient of the particular class score with respect to the feature maps and performing a global average pooling over width and height dimensions. Since the method is applicable only for classification problems, we converted our network prediction to a classification output by global average pooling. Both methodologies for network visualization complement each other well and can, when combined, generate detailed insights into the decision-making process of DL architectures[3].

3 Results

Unifying the aforementioned methodologies entailing specimen preparation, image acquisition, multimodal data registration, data fusion, deep learning modeling, and network visualization, facilitates a holistic approach for microstructure inference. Ultimately, this puts us in a position to explore the interdependencies within and optimize this processing pipeline.

3.1 Image sets and corresponding annotations

For creating the annotation masks, correlative EBSD data was used. Figure 2 shows a LOM image (a), different overlays of LOM with suitable EBSD-derived characteristics (b-e), and the resulting annotations of lath-shaped bainite based on the

LOM image and this EBSD-derived information (f). Enlarged details in (g–j) illustrate how unique grain color maps or grain boundary visualizations can be used to precisely define boundaries of the lath-shaped bainite regions. For instance, it is visible which second phase particles belong to the object or are part of the surroundings (red encircled region in Figure 2g–j). This data also helps if grain boundaries are not clearly visible in LOM images due to weak contrasting. Additionally, when the determination of the class affiliation is impeded due to microstructural units with intermediate morphology between lath and granular, complementing EBSD information can provide a remedy. Without EBSD data, this assignment would have to be done solely based on the microstructure's visual appearance, which can lead to disagreement between human experts and inconsistencies during annotation. Enlarged details in k–n show that misorientation parameters, i.e., low values for KAM or GOS (blue color in Figure 2l+m), indicate polygonal ferrite grains adjacent to or even inside lath-shaped regions (red arrows in Figure 2k–n) that should be excluded during annotation. Distinct crystallographic orientation of the embedded grain (see Figure 2c) does not suffice to unambiguously exclude it from the lath-bainite class. However, intragranular misorientation metrics can characterize such marginal cases as ferritic regions since small intragranular misorientation is incompatible with the notion of lath-shaped bainite[36]. These illustrations also clearly show the difficulty of the segmentation task at hand because the different phases are not distinguishable by gray value distribution, show very complex-shaped borders, and can exhibit objects of one class inside objects of the other class.

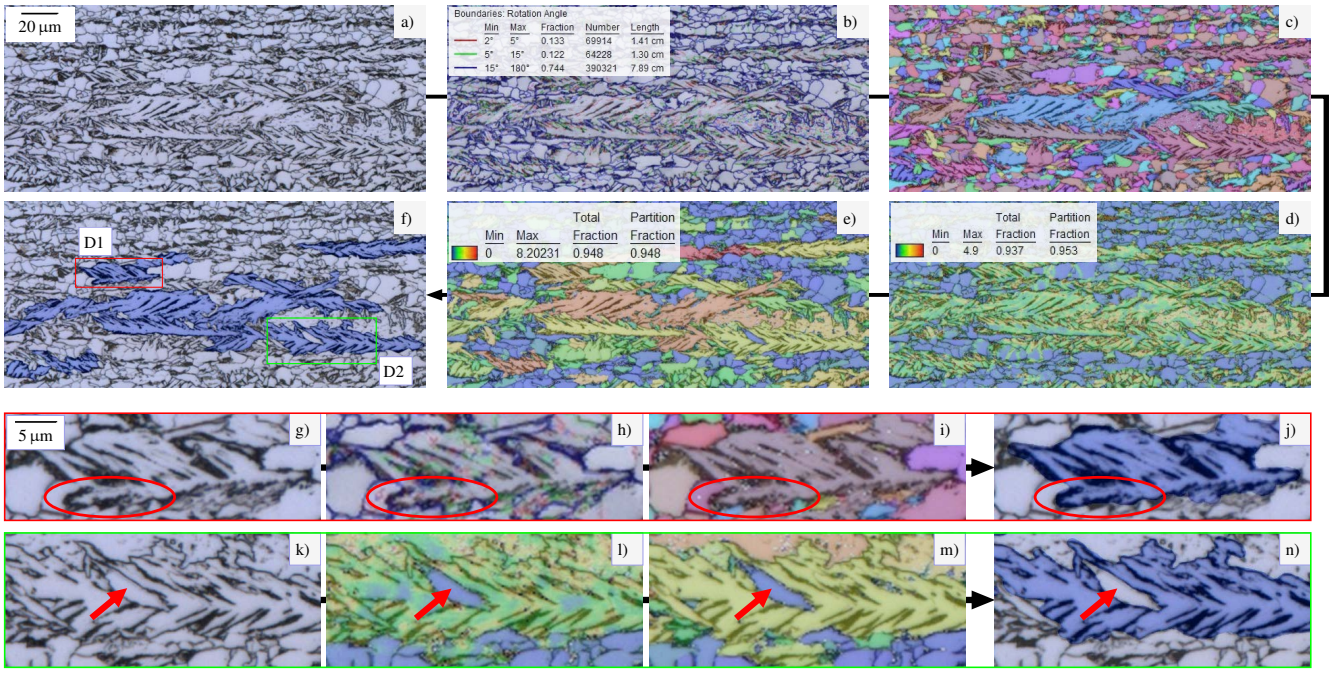

Figure 2. Illustration of correlative microscopy approach for objectively annotating lath-bainite regions. (a) Original LOM micrograph. LOM overlayed with (b) an EBSD-derived grain boundary map, (c) unique grain color map, (d) kernel average misorientation (KAM) map and (e) grain orientation spread (GOS) map. (f) LOM with annotated lath-bainite regions based on EBSD information. Detail views D1 (red frame) and D2 (green frame) are highlighted here. D1 in figures (g–j) displays how (h) grain boundary and (i) grain visualizations are used to correctly annotate the exact boundaries of the lath-bainite. In contrast, D2 in figures (k–n) demonstrates how (l) KAM and (m) GOS indicate polygonal ferrite grains in or adjacent to the lath-bainite region.

3.2 Segmentation results

Both architectures were trained according to Section 2.5 with data from both image modalities using native scale and downscaled image tiles. The segmentation performance is evaluated in terms of intersection over union (IoU) metric for the foreground (fg: lath-shaped bainite regions) and background (bg: polygonal and irregular ferrite with dispersed granular carbon-rich 2nd phase) as positive classes each. Moreover, the accuracy metric is provided, which measures the pixel percentage that is in agreement with the annotation. The given metrics for each model represent the average and standard deviation over all five-fold cross-validation trials. Since the aleatoric uncertainty component[37] introduced during training was previously confirmed to be negligible, the standard deviations given in Tables 2 and 3 are predominately attributed to the k-fold sampling from the low-quantity data sets. This is shown in a diagram in the Supplemental, where the class-averaged IoU is plotted over the fold

number for multiple models. For this reason, and since the same overall data was used to train the models, the average values can be utilized to deduce tendencies between the models within each modality.

3.2.1 LOM Image Set

In Table 2 the LOM image-trained models along with model initialization type, image resizing factor, and the resulting accuracy as well as IoU metrics are listed.

Table 2. Intersection over Union metrics of U-Net-based networks trained on the light optical microscopy data set for different model initializations and downscaling factors. The superscript ^{v} indicates a validation experiment conducted to test a particular hypothesis, as described in Section 3.2.1. The superscripts * and ** indicates that the model used padded convolutions and that downscaling was performed after tiling respectively. The accuracy metric is provided to get an intuition about the model performance, while the IoUs can be used to discriminate the model performance better, especially since they trained on an imbalanced dataset.

#	Model	Model initialization	Downscaling factor	Accuracy	IoU_{bg}	IoU_{fg}
1	Vanilla U-Net	random	native	91.6 ± 0.6	88.9 ± 1.0	74.1 ± 1.1
1^{v}	Vanilla U-Net*	random	native	90.1 ± 0.9	87.0 ± 1.5	70.2 ± 1.2
2	Vanilla U-Net	random	0.5×0.5	90.9 ± 1.1	88.1 ± 1.5	72.1 ± 2.2
2^{v}	Vanilla U-Net	random	0.5×0.5**	90.3 ± 0.7	87.1 ± 1.3	71.6 ± 0.8
3	U-Net VGG16	pre-trained	native	90.6 ± 0.6	87.6 ± 0.9	71.3 ± 1.7
3^{v}	U-Net VGG16	random	native	89.6 ± 0.8	86.3 ± 1.3	69.3 ± 1.5
4	U-Net VGG16	pre-trained	0.5×0.5	90.3 ± 1.1	87.1 ± 1.7	71.6 ± 1.7

The results show that the best models of the two architectures trained on native and downscaled image tiles (#1, #2 ,#3, and #4) attained accuracies of above 90%. This is comparable to the discrepancy in annotation by human experts, when relying solely on topography information. It can also be observed that the conventional U-Net scored better than the VGG16-based model. In an attempt to explain this, an additional ablation study was conducted by gradually assimilating both architectures. To that end, in model #1^{v} a vanilla U-Net with padded convolutions was trained. It is to be noted that in this particular experiment, identical to the VGG16 case, non-overlapping tiles were provided as input thus reducing the context available in the images compared to model #1. On the other hand, a random initialized VGG16 validation model #3^{v} was introduced to facilitate comparability between the VGG16 and vanilla U-Net models. The #1^{v} model performed worse than #1, and converged towards #3^{v}. At the same time, model #3^{v} acted as the random initialized equivalent of the ImageNet pre-trained U-Net VGG16 (#3), therefore enabling assessment of the pre-training dependency. Their comparison showed that utilizing models pre-trained with domain-external data sets such as ImageNet can be beneficial for material scientific challenges.

The downscaling of the images, in general, was performed before tiling, resulting in fewer tile images with sampling-induced and interpolation-induced information loss but comparatively more context in individual tiles. Downscaling for the vanilla U-Net did reduce the lath-bainite segmentation performance by 1–2% IoU on average. The validation experiment 2^{v}, where downscaling after tiling was used for validation purposes, showed an additional slight decrease in IoU. Further, a higher scatter in IoU is observed for models when downscaling before tiling is applied.

In the case of the vanilla U-Net, a foreground weighing factor $\alpha = 1.5$ to correct for the material-inherent class (i.e., phase) distribution imbalance was found during the hyperparameter optimization to improve the overall IoU slightly. While in this study, tiled images were used for both training and evaluation, it was found that using tiled images during evaluation is detrimental for the segmentation of such micrographs.

3.2.2 SEM Image Set

For the SEM image-trained models (Table 3), the difference between foreground and background IoUs is less pronounced than for the LOM. Moreover, the IoU_{bg} shows a strong dependence on the scale, where downscaling before tiling proves advantageous (cf. IoU_{bg} of models #5 and #6 or #7 and #8). While this result seems surprising on a first glance, its plausibility will we validated later. This effect was more pronounced in the U-Net VGG16 case (cf. models #7 and #8). The pre-trained U-Net VGG16 trained on downscaled data (#8) achieved, while not by a large margin, the best IoU for SEM images. Further, the SEM model performances confirm the LOM-case observation that the vanilla U-Net performs better for random initialization. However, while pre-training contributed to only mediocre improvements for LOM (cf. model #3 and #3^{v}), it resulted in a significant foreground and background IoU improvement for SEM, of approximately 8% IoU (cf. model #8 and #8^{v}).

In Figure 3, the resulting segmentation map predictions of the best vanilla U-Net (Figure 3a+b) as well as best random initialized and pre-trained U-Net VGG16 (Figure 3c–d and e–f) models are compared to the annotations for both modalities. Note that the illustrated images are full frame, while for training and testing, tiles of such images were used. The vast majority

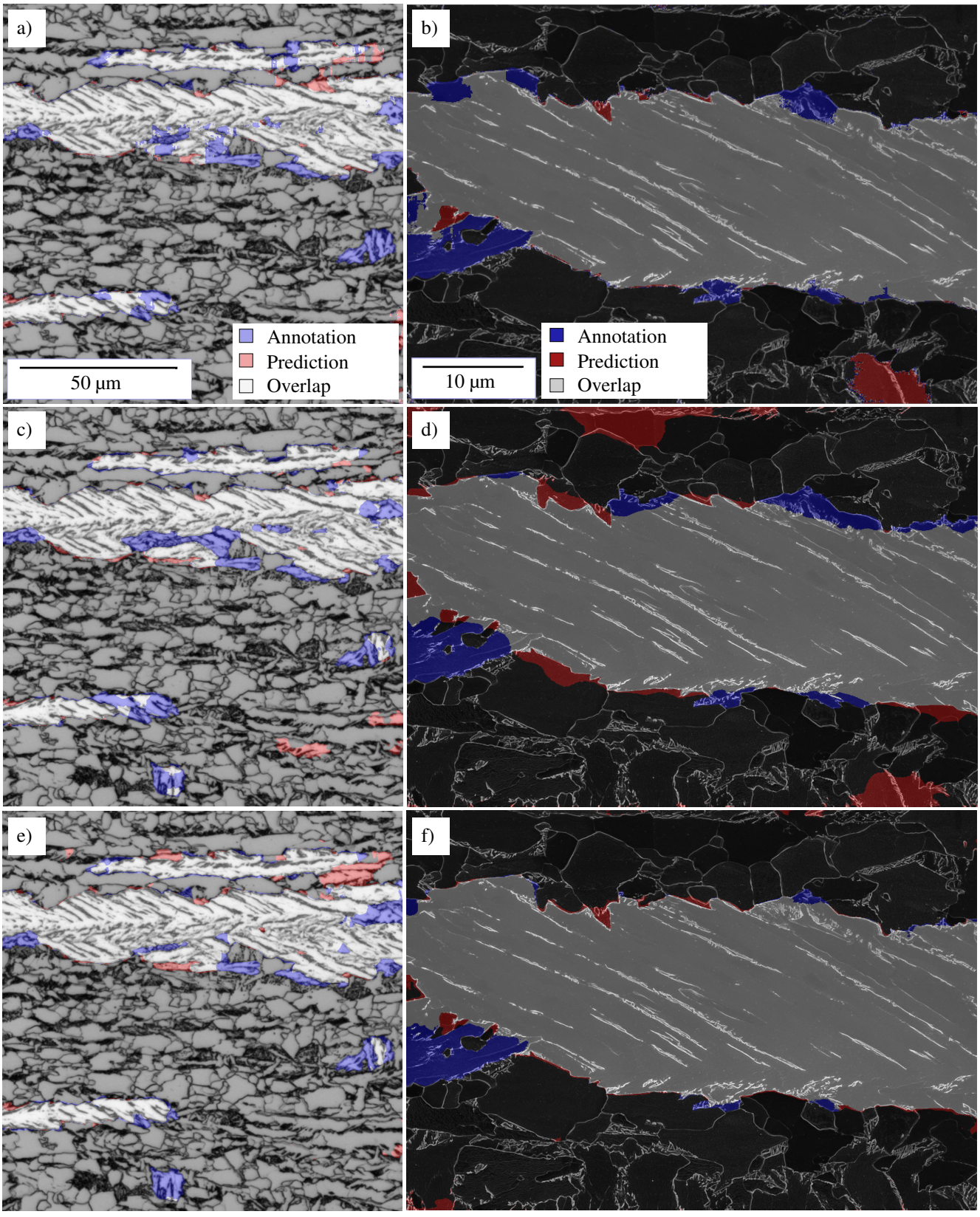

Figure 3. Light optical and scanning electron micrographs superimposed with lath-bainite predictions of different models and annotated regions showing the comparison between model prediction (red) and manual expert annotation (blue). (a, b) random initialized vanilla U-Net (model #1 and #6). (c, d) random intialized U-Net VGG16 (model #3^{v} and #8^{v}). (e, f) pre-trained U-Net VGG16 (model #3 and #8).

Table 3. Intersection of Union metrics of U-Net-based networks trained on the scanning electron microscopy data set for different model initializations and downscaling factors. The superscript ^{v} indicates a validation experiment conducted to test the impact of pre-training.

#	Model	Model initialization	Downscaling factor	Accuracy	IoU_{bg}	IoU_{fg}
5	Vanilla U-Net	random	native	86.5 ± 0.8	71.6 ± 1.0	79.5 ± 2.1
6	Vanilla U-Net	random	0.5×0.5	87.5 ± 1.1	76.2 ± 3.1	78.9 ± 2.3
7	U-Net VGG16	pre-trained	native	86.7 ± 1.0	71.0 ± 2.5	80.1 ± 2.4
8	U-Net VGG16	pre-trained	0.5×0.5	88.4 ± 1.3	77.7 ± 3.2	80.4 ± 2.6
8^{v}	U-Net VGG16	random	0.5×0.5	83.1 ± 0.8	68.4 ± 4.0	73.1 ± 1.5

of lath-shaped bainite regions are well identified and the predictions are widely in accordance for all three models of each modality. Moreover, locations of erroneous predictions match in the models to a large extent. Under-prediction (blue) occurs at individual smaller foreground objects or at the borders of extensive lath-shaped bainite regions, where annotated parts do not exhibit a clear lath structure. Over-prediction (red) tends to arise mostly in smaller areas in which carbides or grain boundaries resemble lath shapes. In both modalities, but especially in the SEM-trained case, the random initialized VGG16 falls short as opposed to the other models, which is mirrored by the performance metrics in Tables 2+3.

4 Discussion

To achieve reliable and objective microstructure inference, an understanding of fabrication, microscopy, DL methodology, and their interdependencies is required. It is important not merely to look at the images and corresponding annotations as an isolated step but also to consider building a DL model as a holistic approach where specimen preparation, reproducible specimen contrasting, and suitable image acquisition techniques are of tremendous importance[8].

Dataset acquisition. In our study, we successfully trained both random initialized networks and pre-trained networks with comparatively small data sets of approximately 50 and 30 images for LOM and SEM, respectively. This invalidates the general claim of DL being only applicable for large-scale data sets. Moreover, it indicates that given reproducible, high-quality imaging can be ensured, a smaller number of images is sufficient for building an accurate DL model.

Special attention was paid to specimen preparation, optimal contrasting during etching, and consistent settings during image acquisition. Moreover, the very reproducible imaging settings, e.g., viewing perspective, brightness, and contrast, lead to a low degree of material-extrinsic variance in the images compared to real-world scenario image sets. The pronounced planarity of the metallographic cross-sections avoids geometry-related image shading and distortions. Class imbalances often pose a challenge for learning. Due to the comparatively lower magnifications during image acquisition, the LOM image data set is representative for the microstructure in terms of phase fractions, where lath-shaped bainite is a minority class (28%). This poses a material-inherent class imbalance. On the other hand, the lath bainite phase was oversampled during SEM image acquisition to correct for the imbalance. If such imbalances were not artificially corrected at the image acquisition stage, post-processing techniques such as sampling or weighting methods could be applied to account for them.

The choice of imaging modality primarily depends on the scale on which relevant microstructural features are to be expected. For instance, while LOM might be well suited to deduce lath-shaped regions in CP steels, SEM was incorporated as it additionally contains information on the exact nature of carbon-rich 2nd phases, which renders the distinction between bainite subclasses possible. Perspectively, when the data quantities and imbalances between both modalities are matched, the more suitable modality for lath-shaped bainite prediction or other tasks can be concluded.

Assigning the ground truth, i.e., correctly annotating the lath-shaped bainitic regions, is challenging, and disagreements between human experts can arise when purely relying on the microstructure's visual appearance in LOM or SEM. By incorporating correlative EBSD data, even though for just a part of the image set, the objectivity and reproducibility for annotating micrographs are improved. However, annotating the foreground regions manually by tracing their perceived object boundaries on a tablet can still lead to some inconsistencies.

Material scientific impact. All segmentation models achieve performances that are comparable to expert segmentations performed in absence of EBSD data. Since both groups applied their DL best practices and the architectures are fairly similar, the changes in performance are not extremely pronounced. Nonetheless, for the first time applying DL best practices of two groups on identical materials data gives important insights. These insights are essential to facilitate subsequent major improvements through adapted pre-processing, architectural choices, and training procedures. The similar segmentation results of U-Net-based models point towards general performance robustness concerning different architectures and training strategies. Namely, no severe performance decrease was observed by different initial network conditions (random initialization as opposed to ImageNet pre-trained model), different internal padding and normalization strategies. The regions at which the models fail

are regions where human experts would primarily make mistakes during manual segmentation.

The segmentation enables the accurate calculation of lath-shaped bainite phase fractions. Reported IoUs for the foreground (lath-shaped bainite, Table 2 and 3) correspond to minor phase fraction errors in the range of 1% compared to human expert annotation. This error is lower than the variance in manual human expert evaluation. These are remarkable results considering the intricacy of the segmentation task at hand. A prerequisite to archive accurate phase fraction predictions is that the training data is not significantly skewed towards a specific class. Skewed data sets would result in models that favor the majority class[38]. Therefore, for accurate phase fraction estimation of relatively uncommon phases, imbalance correction is advised.

Since there are some deviations along the border of lath-shaped bainite objects, localization of phase boundaries is only possible to a limited extent. These border deviations are potentially partly attributed to the aforementioned border annotation inconsistencies and hamper the calculation of individual bainite object morphological parameters associated with the objects' spatial extent. On the other hand, segmentation enables the analysis of *inner* morphology of specific phases in the first place. In this case of CP steel microstructures, it facilitates the detached calculation of the lath-shaped bainite regions' lath-characteristics (e.g., lath-width), instead of calculating these characteristics for the whole image, yielding a more sophisticated microstructure quantification. These morphological parameters are known to impact mechanical properties significantly[39,40]. Furthermore, the also accessible relative spatial distribution of phases in such heterogeneous microstructures affects local fatigue properties severely. Such focused microstructure analyses are the prerequisite to establish processing-microstructure-property correlations. Furthermore, reliable and high-fidelity segmentation has implications for automated and targeted microscopy.

Image context dependency. Considering random initialized models, the vanilla U-Net scoring better than the U-Net VGG16 can be hypothesized to be ascribed to the following factors:

- Inclusion of comparatively more context in each image tile at the training and testing stage. When a complete prediction is to be obtained for an image, tiles of that image passed to the network need to overlap in case of unpadded convolutions.
- The learnable transposed convolution upsampling can recover spatial localization in the decoder more accurately[41].
- In vanilla U-Net training, the imbalance in the data set was corrected by employing a class weight inside the focal loss function.

Indeed, a validation experiment demonstrated converging performances (cf. model #1^{v} and #3^{v}) when the model architecture of the vanilla U-Net was assimilated to that of the U-Net VGG16 such that padded convolutions were utilized and images were supplied accordingly (center-cropped). This indicates that the context contained in the images provided during training and testing is of pronounced importance. This is confirmed by the SEM model associated results, where an increased image field of view through downscaling before tiling resulted in a performance improvement (cf. model #7 and #8). Since even the high-resolved SEM images at native resolution and scale cover a comparatively small field of view, tiles extracted with fixed resolution do not represent the lath structure appropriately but only contain fragments of the lath-shaped regions. Therefore, in this lath-bainite segmentation case, where parallel but distant inter-lath carbide islands are relevant to deduce the foreground, it is valuable to increase the tile image field of view to increase the likelihood of obtaining multiple parallel carbide clusters within a single image. It applies to many microstructure inference tasks that objects of interest in metallographic micrographs (e.g., phases) and features within these objects are comparatively more dispersed than in many real-world scenario images. Therefore, in phase quantification, where long-range features such as the morphology of grain boundary traces are relevant, downscaling before tiling can potentially improve performance and accelerates training. When applying downscaling, it should be used consistently at training and testing time. This holds especially true if no preventive measures, such as scaling data augmentations, are taken at training time. While at training time, tiling and downscaling are often inevitable due to GPU memory constraints, during testing, where the image size is not restricting, avoiding tiling proves to be beneficial. This was observed in a preliminary study for the U-Net VGG16 where an increase of 3.3% IoU over the IoU values reported in Table 2 was achieved when evaluating full (uncropped) and unseen images.

Note that downscaling for increasing image context is dispensable when tile images contain sufficient features and should be avoided when features are likely to disappear by downscaling. For instance, the LOM model performances are assumed to improve only moderately by the increased image context since tiles extracted from native scale images capture sufficient features of the microstructure. Moreover, while segregated carbides at lath interfaces are slender, we believe that these features are still largely preserved when 0.5$\times$ downscaling is applied. The similar model performances #1 and #2 as well as #3 and #4 point towards the validity of both assumptions. A validation model training (#2^{v}) was conducted in the LOM case to confirm that these two competing effects (context increase and information loss) are not individually pronounced effects that compensate each other due to superposition. To this end, data were tiled *before* downscaling as an exception. Both downscaled data sets then had the same physical pixel size (i.e., same information loss) but differed in the image-wise field of view (i.e., image resolution). When tiling before downscaling is applied, a slight reduction of less than 1.0% IoU (cf. model #2 and #2^{v}) is observed compared to tiling after downscaling. This confirms that, at such downscaling factors, the effect of increased image

context and the information loss is negligible for the LOM images. Moreover, this is in accordance with[42], where a plateau of nearly constant performance for a range of downscaling factors (0.2-0.5×) was demonstrated. Downscaling factors below a threshold are typically accompanied by a significant information loss and a decrease in segmentation performance. Such information loss can be ascribed to the image downsampling and non-ideal interpolation. Literature[42,43] suggests that this threshold depends on the specific foreground class. In these works, it was shown that specific classes that have fine features or small object extent profit from discarding downscaling operations.

Network receptive field. Aside from ensuring appropriate feature representation in images, it is important to select a network architecture for the task at hand that takes a sufficient amount of context into consideration. Characteristic image length scales (e.g., phase boundary pixel distance) change depending on the applied magnifications and image resolutions required to resolve relevant features during image acquisition. In such cases, it can be beneficial to adapt the image region that the network considers, also referred to as theoretical receptive field (TRF), accordingly. In Luo et al.[44], the effective receptive field (ERF) metric was proposed for CNNs and was empirically computed for several architectures. The ERF revolves around the notion that not every region within the TRF is taken into account equally. In fact, the predicted ERFs were substantially smaller than the TRF and showed a 2D Gaussian distribution that strongly decays towards border regions of the TRF. This means, the closer a pixel is to a target pixel, the more it influences the target pixels' predicted class[44]. This represents a CNN-based inductive bias appropriate for many scientific segmentation challenges, such as for fatigue damage localization where image features are dense[18]. However, for phase segmentation tasks, where long-range features (parallelism of distant carbide islands) are relevant, Attention-based networks[45] could improve segmentation performance. The observation that the scale of features determines the optimal downscaling factor has led to specialized architectures. Especially in such multi-class segmentation or classification tasks where features are distributed across scales, aggregation of distinctly dilated convolutions is reasonable[46]. In conclusion, it is important that individual tile images comprise sufficient learnable feature information, and the architecture facilitates their appropriate extraction and processing.

Pre-training dependency. The fact that pre-training led only to a minor improvement for the U-Net VGG16 in the LOM case leads us to conclude that the 50 full-frame LOM images with relatively even class distribution suffice for training such a binary segmentation model. While this is dependent on the exact problem and model to be trained, we infer that given such data, U-Net-based models, which score satisfactory results at such ambitious phase segmentation tasks, can be trained from scratch. In contrast, for the SEM case, the pre-training culminates in a significant IoU increase of 7–9% over the random initialized U-Net VGG16 (cf. model #8 and #8^{v}). Therefore, the pre-training dependence is comparatively more pronounced in the SEM modality. This can be ascribed to the fewer amount and higher magnifications of SEM images, hence covering considerably fewer distinct microstructure scenarios. Moreover, the limited SEM data availability can explain that the ImageNet pre-trained VGG16-based network, in contrast to the LOM case, outperformed the randomly initialized vanilla U-Net for SEM models. Therefore, when the data set comprises few images that cover a small field of view, we advise pre-training with readily available data sets. While ImageNet encompasses a wide range of classes, the noise characteristics in microscopic images are different. Potentially, pre-training with other data sets exhibiting a smaller domain gap such as miscellaneous nanoscientific objects in SEM[47] or ultra-high carbon steel phases SEM[48] can be advantageous.

Variances and generalization. The random k-fold sampling of low-quantity data, especially in SEM, results in notable IoU scatter. In such cases, stratified sampling and training can prove beneficial before deployment of the model. The reproducible preparation, mounting, and imaging rendered the data augmentation, and corresponding hyperparameter tuning negligible as the performance improvement associated with it for both modalities was minimal. This implies that data augmentation is not generally essential for small-scale data sets, but only when the applied transformations render the training set more representative of the test set. In instances where such material-extrinsic variance can be ensured to be insignificant, data augmentation through simple spatial (affine and even elastic) or intensity transformations can be evaded. Therefore, such models trained on comparatively small data sets are suitable for tasks with inherently small scatter, such as quality inspection, where recurring tasks and predefined workflows are set. When, for instance, etching-based contrasting methodologies are concerned, reproducibility can be difficult to attain.

To this end, a generalization study was conducted to test the transferability of this model, trained with low variance data, to an alternate data domain. Therefore, the previously best SEM vanilla U-Net model #6 was tested on a SEM image of a surface etched with Nital as an exception. In contrast, another model was trained with dedicated augmentation settings to improve the performance on the alternate etching domain. Figure 4a+b illustrates the comparison of the source domain (electrolytic etch) with the alternate target domain. Moreover, Figure 4c+d depicts the segmentation of Figure 4b using model #6 and a model trained with solely brightness and contrast augmented images of the source domain. The degree of both augmentations was optimized for the target domain.

A substantial improvement of IoU_{bg} and IoU_{fg} of 46% $\rightarrow$ 72% and 54% $\rightarrow$ 62% is achieved, respectively. This corresponds to a change in phase fraction prediction of 66% $\rightarrow$ 37% for a manually labeled phase fraction of 44%. Evidently, the segmentation in Figure 4d is not satisfactory since applied augmentations do not close the domain gap entirely. More elaborate

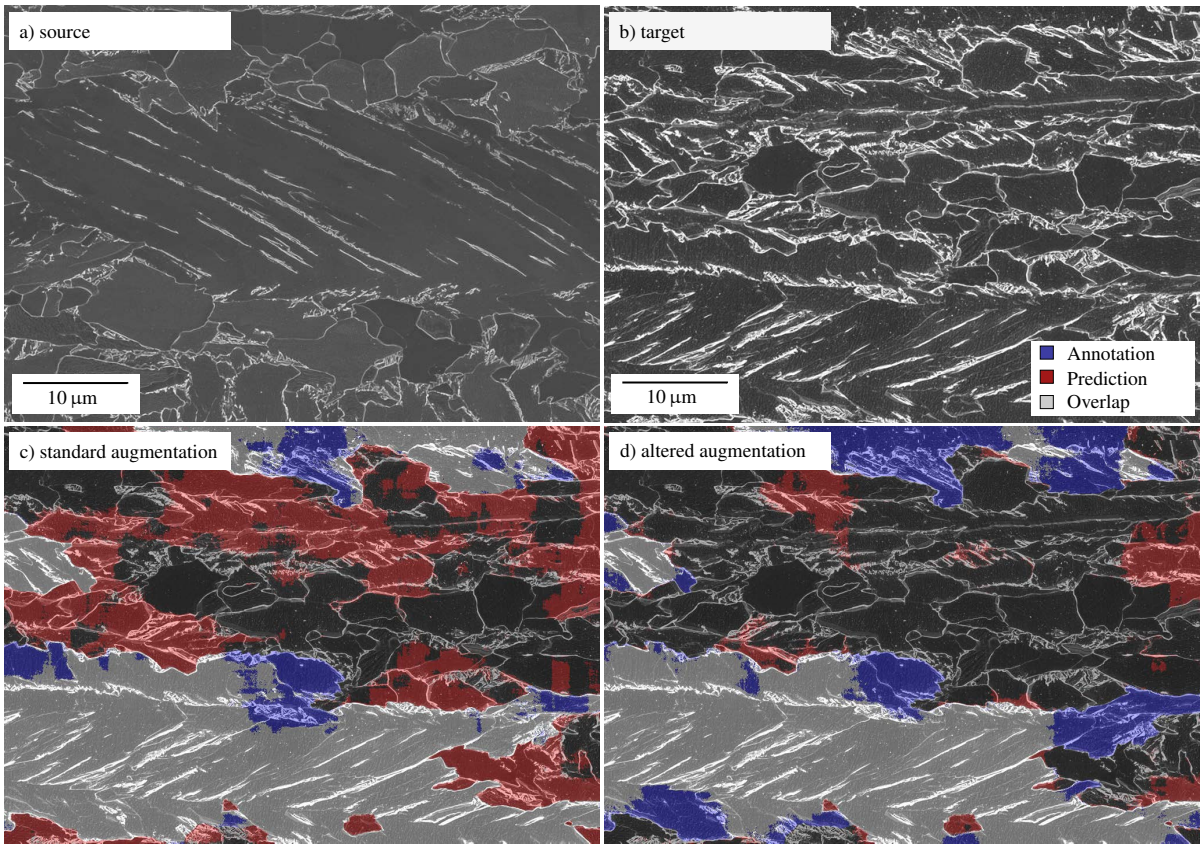

Figure 4. An generalization study for alternatively etched surfaces. a+b) Comparison of the electrolytically-etched source image domain (a) and the alternate Nital-etched specimen (b). c) Segmentation results of model #6. In this case augmentation parameters were altered for the source domain (a). d) Segmentation results of a model like a) but with modified brightness and contrast augmentation to improve performance on the alternately etched domain. The legend for c) and d) and the micron bar for b), c) and d) is in positioned in b) to avoid concealing of relevant regions in the segmentation results.

image transformations would be required to align the domains since the secondary electron image formation is strongly affected by the different topographies. Nonetheless, the fact that even simple targeted optimizations of low-variance training data can cause such improvements, implies that dedicated data augmentation pipelines can presumably render models robust against a large range of perturbations in the specimen preparation or imaging. For instance, in our prior study[18] a substantial improvement was achieved by augmentation of our high-variance data. When training images are acquired from different instruments or at different institutions, such regularization methods become increasingly relevant. In such instances, it is essential to track and store all relevant process parameters along the entire process chain in a structured and ideally semantics-informed database. Moreover, this outcome foreshadows that advanced augmentation with generative adversarial networks (GANs)[49] or domain adaptation[50], potentially can address even more challenging generalization demands of the materials science community in the future.

Network interpretability. Interpretability and explainability of DL models are important to build trust and push for successful implementation in day-to-day applications. Moreover, it can help in finding failure modes of models and give insights on tackling them. To that end, we computed network visualizations that highlight regions or concepts within an image that affected the network's decision. Meaningful examples of Grad-CAM and NetDissect visualization from several network layers are illustrated in Figure 5 and 6, respectively. Furthermore, Grad-CAM visualizations of all network layers for both architectures can be found in the Supplemental.

Grad-CAM masks are generated for a particular convolutional layer of the trained models #1 and #3 for a specific class. These masks are formed by a weighted average of all activation maps originating from all the target layer filters. Hence, the masks highlight those regions in the input image that the specific layer treats as essential for predicting the specified class. Thus, by looking at the Grad-CAM masks of various layers for lath-bainite and background classes, we can deduce how the trained model predicts a segmentation mask for a given input.

Although the following observations are qualitative, they are very helpful in interpretation. Concerning the background in the LOM segmentation, strong activations are caused to a certain extent by particles of the carbon-rich 2nd phase (b, c, e, h) and for the most part by grains and grain boundaries of the polygonal ferrite (b, d, f, g, i). Moreover, in the down4convr1 layer (c), there is a focus on grain boundary junctions, such as triple and quadruple points, which are discriminative features. Activations in the vanilla U-Net and VGG16 U-Net mostly match (note that e and h show similar activations that differ in scaling and the degree of focus on carbon-rich clusters). However, towards the end of the decoder (layer up4convr2), the vanilla U-Net focuses on polygonal ferrite grain boundaries (f) to determine the final output, whereas the VGG16 U-Net focuses on the grains themselves (i). The model during decision-making puts emphasis on image features that correlate with how human experts interpret the image. For instance, the decision for the background will mostly depend on the existence of ferrite grains, which are comparatively more equiaxed. It should be noted that the vertical and horizontal line artifacts visible in Figure 5d are presumably attributed to the checkerboard problem associated with transposed convolutions[51] as such artifacts do not occur in the VGG16 case that used bilinear upsampling.

Lath-bainite activations in some layers are induced by second phase particles and grain boundaries in general or by elongated second phase particles and grain boundaries. However, the strongest reactions are caused by pronounced, more extensive lath regions. An analogy to the human expert examination can be supposed here, too. Pronounced, more extensive lath areas should also be noticed strongly by the human eye because of regular lath structure compared to the surrounding. Significant differences in feature importance between the different U-Net architectures were not found.

In contrast to Grad-CAM, NetDissect enables us to analyze what different filters in the model extracted from an input image, regardless of its contribution to the final segmentation map. This technique offers us the prospect of finding disentangled feature extractors from the model, which make sense to a human expert, see Figure 6. Note that these exemplary images represent only a small portion of filters utilized in the whole network. In the case of LOM images, relevant features include 2nd phase particles plus elongated grain boundaries (a), lath-shaped 2nd phase particles (b), and the area of more extensive grains (c). Thus, an analogy to human expert interpretation can be assumed here as well. Moreover, considering that lath-shaped 2nd phase particles are relevant features, similarities to how feature engineering is performed during conventional ML or CV can be seen, too. For instance, in Müller et al.[9] a sliding window technique which utilizes a Prewitt[52] edge detection filter to calculate directionalities of the 2nd phase particles is applied. Directionalities are used, in combination with a neighborhood analysis, to detect lath-shaped regions.

5 Outlook

Given the large number of different materials and processes, and the time-intensive generation of data sets for many tasks, materials science will always be accompanied by data scarcity. It is all the more important that strategies of model generalization to alternate materials or processing conditions are pioneered. As a consequence of emerging high-speed image acquisition technologies, annotation processes often pose the bottleneck in the creation of statistical data sets. This particularly holds true for the supervised learning of segmentation models in the material scientific domain. By pushing the correlative approach with

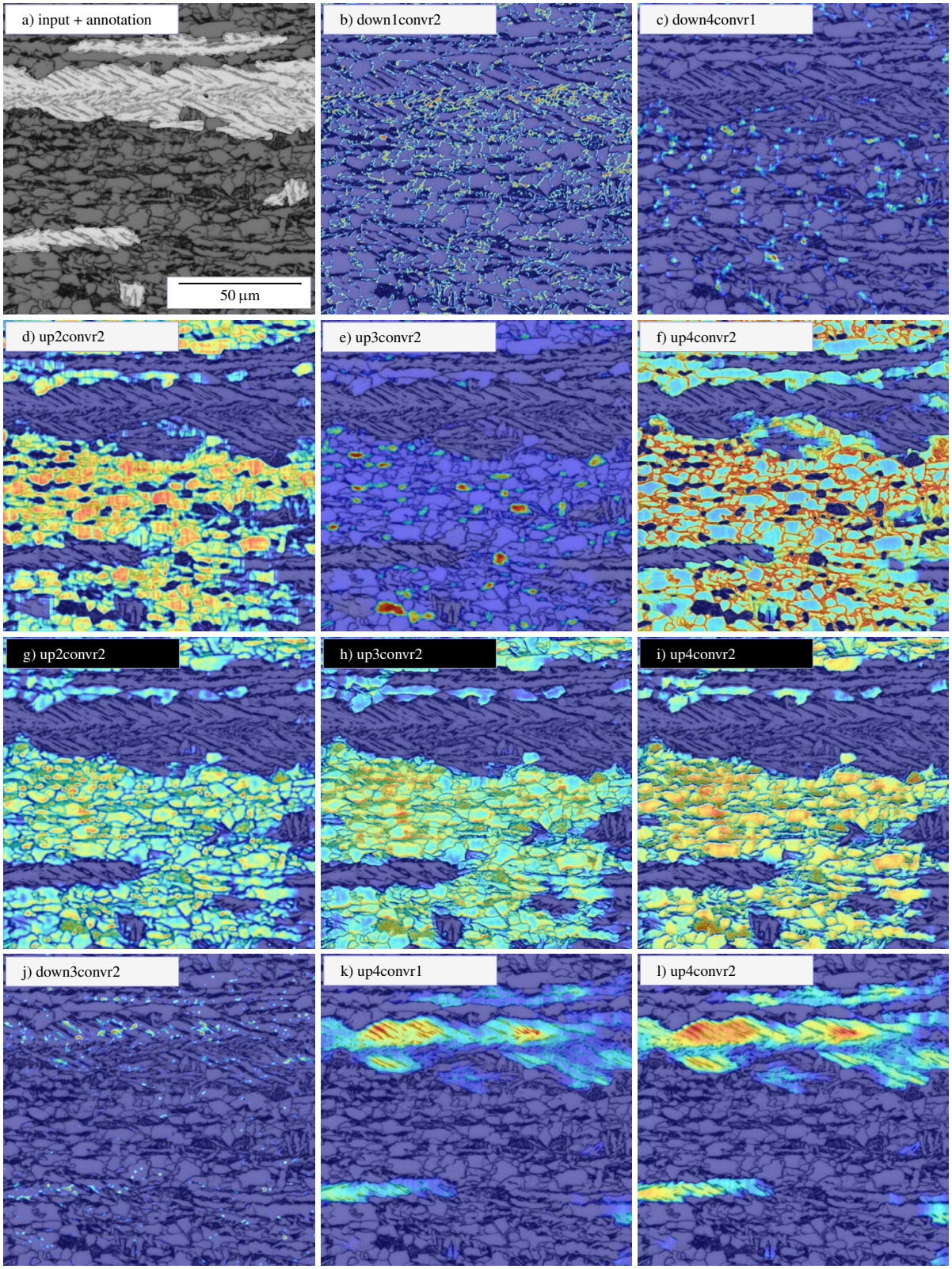

Figure 5. Gradient-weighted class activation maps indicating image regions that dictated the decision of the network with respect to the background class (a–i) and lath-bainite regions (j–l). The Grad-CAM maps are derived from specific, designated layers of light optical microscopy models 1 (black font) and 3 (white font).

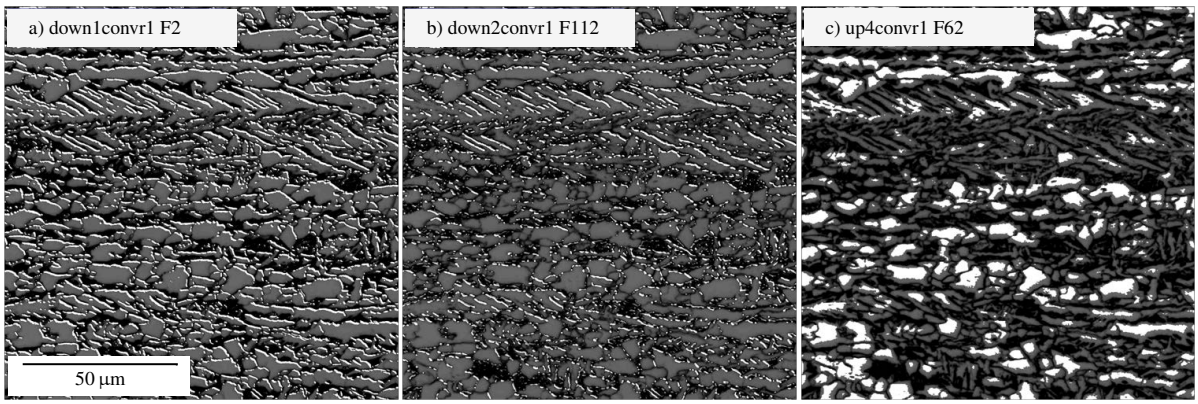

Figure 6. Thresholded activation maps of specific convolution filters (F**X**) using the NetDissect method. The high gray value regions indicate disentangled concepts that were learned in model #1.

EBSD measurements forward, routines for automatically generating annotations based on EBSD data can be developed. This promises to improve the annotation quality and make it less labor-intensive. Moreover, it enables further segmentation tasks to be addressed, e.g., segmenting lath-shaped and granular bainite as well as distinguishing bainitic and pro-eutectoid ferrite.

Nonetheless, generalizing data-driven methodologies and alternate learning strategies will be indispensable to cope with material diversity. In literature different training strategies to tackle the sparsity of annotated data have been developed which rely on comparatively less data. These can be adopted for the segmentation of metallographic phases or the materials community in general.

Rather than providing pixel-wise annotations for training a segmentation network, in a *weakly-supervised learning* setting, e.g., image-wise annotations are used. There are different annotation abstraction levels ranging from bounding boxes[53] to naming the classes present in an image[54]. Typical methodologies rely on classification networks which provide seeds for the segmentation network, and constrained seed region growing to respect object boundaries[54,55]. In recent years a leap in weakly-supervised segmentation performance was achieved[56], rendering it a promising method for phase fractions. This is affirmed, since well contrasted grain boundaries presumably can pose distinct and suitable borders for region growing. In particular for metallographic segmentation tasks in which target phases are often dispersed across the whole image, pixel-wise annotation is cumbersome. Here it can be particularly worthwhile to replace manual pixel-wise annotations by appropriate weak labels.

Alternate techniques called *semi- or unsupervised domain adaption* evolve around the idea that for a specific task (e.g., segmentation) annotated data of one source domain (e.g., material A) can be used together with non-annotated or minimally annotated data of a target domain (e.g., material B) to produce meaningful predictions in latter. The methods achieving this rely on feature matching between both domains, self-training to provide pseudo labels or generative networks to produce target data[50]. The range of materials and processes that can be covered with such techniques in material scientific challenges is yet to be unveiled. Moreover, the materials science domain can profit from its longstanding experience in knowledge-driven, realistic simulation techniques such as phase field simulations. The resulting synthetic data can be exploited in domain adaptation to obtain annotated data in a source domain or for pre-training[57].

Another promising candidate to reduce annotated data requirements are *physics-constrained* DL models[58]. Rather than supplying a multitude of input-output pairs, conditions that represent domain knowledge are imposed on the output space. In such cases the domain knowledge is typically encoded into the loss function. For microstructure inference, laws from thermodynamics including different crystal growth or segregation/precipitate formation models potentially can condition DL models.

6 Conclusion

In this study we demonstrate the applicability of deep learning (DL) for the segmentation of complex phase steel microstructures. Since its individual constituents differ only in shape and arrangement of ferritic and carbon-rich phases rather than image intensity levels, traditional segmentation approaches reach their limits. We propose a holistic approach since the contrasting and imaging has pronounced implications for the DL methodology in terms of data imbalance, variance and spatial feature density. Amongst others, this includes annotations informed by electron backscatter diffraction to alleviate the burden of the manual annotation process based on how the microstructure in topography contrast micrographs visually appears to the expert eye. This allowed to provide a well-founded, objective annotation. While the segmentation models presumably benefit from

more data, the trained U-Net networks achieved a satisfying performance from training with only 30–50 microscopic images. We hope that rebutting the general preconceptions about the large required data quantities, mitigates the reservations towards DL and ultimately encourages more scientists to research in this interdisciplinary field. The results point towards a general robustness of the U-Net with respect to modifications in the training procedure and architecture. Through the experimental design, a general guideline for the application of DL for microstructure inference could be derived. This applies in particular to the appropriate consideration of image context, data augmentation, imaging modalities and pre-training. The network decisions to distinguish lath-bainite from its surroundings are visualized through the Grad-CAM and NetDissect methodologies. These suggest plausible and human comprehensible choices for features such as parallelism of inter-lath carbides, grain boundary junctions, grain aspect ratios and carbon-rich clusters. This is an important step towards the acceptance of DL segmentation in material science community. Finally, we provide an outlook on aspiring and auspicious cutting-edge methodologies from computer science that hold the potential to render microstructure inference from micrographs generalizable across materials and processes. A fundamental requirement to achieve this is the interoperability of diverse data generated across institutes. With the development of materials ontologies and the systematic digitalization of workflows, identifying and unifying relevant data across institutes will come within reach and thus increase the scope of such deep learning techniques substantially.

Data availability

The datasets generated during and/or analysed during the current study are not publicly available because they are part of an ongoing study and subject to third party (AG der Dillinger Hüttenwerke) restrictions.

References

1. Saleh, F. S., Aliakbarian, M. S., Salzmann, M., Petersson, L. & Alvarez, J. M. Effective Use of Synthetic Data for Urban Scene Semantic Segmentation. *Lect. Notes Comput. Sci. (including subseries Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **11206 LNCS**, 86–103, DOI: 10.1007/978-3-030-01216-8_6 (2018). 1807.06132.
2. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *Lect. Notes Comput. Sci. (including subseries Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **9351**, 234–241, DOI: 10.1007/978-3-319-24574-4_28 (2015). 1505.04597.
3. Natekar, P., Kori, A. & Krishnamurthi, G. Demystifying Brain Tumor Segmentation Networks: Interpretability and Uncertainty Analysis. *Front. Comput. Neurosci.* **14**, 1–12, DOI: 10.3389/fncom.2020.00006 (2020).
4. Liu, W. *et al.* NNs Archtectures review. *Elsevier* 1–31 (2017).
5. Koyama, M. *et al.* Bone-like crack resistance in hierarchical metastable nanolaminate steels. 2–5 (2017).
6. Archie, F., Li, X. L. & Zaefferer, S. Damage initiation in dual-phase steels: Influence of crystallographic and morphological parameters. *Mater. Sci. Forum* **879**, 157–163, DOI: 10.4028/www.scientific.net/MSF.879.157 (2017).
7. shi Zhang, Q. & chun Zhu, S. Visual interpretability for deep learning: a survey. *Front. Inf. Technol. Electron. Eng.* **19**, 27–39, DOI: 10.1631/FITEE.1700808 (2018). 1802.00614.
8. Müller, M., Britz, D. & Mücklich, F. Machine Learning for Microstructure Classification - How to Assign the Ground Truth in the Most Objective Way? *ASM Adv. Mater. & Process.* **179**, 16–21 (2021).
9. Müller, M., Stanke, G., Sonntag, U., Britz, D. & Mücklich, F. Segmentation of Lath-Like Structures via Localized Identification of Directionality in a Complex-Phase Steel. *Metallogr. Microstruct. Analysis* 1–12, DOI: 10.1007/s13632-020-00676-9 (2020).
10. Zajac, S., Schwinn, V. & Tacke, K. Characterisation and Quantification of Complex Bainitic Microstructures in High and Ultra-High Strength Linepipe Steels. *Mater. Sci. Forum* **500-501**, 387–394, DOI: 10.4028/www.scientific.net/MSF.500-501.387 (2005).
11. Li, X., Ramazani, A., Prahl, U. & Bleck, W. Quantification of complex-phase steel microstructure by using combined EBSD and EPMA measurements. *Mater. Charact.* **142**, 179–186, DOI: 10.1016/j.matchar.2018.05.038 (2018).
12. Chen, Y. W. *et al.* Phase quantification in low carbon Nb-Mo bearing steel by electron backscatter diffraction technique coupled with kernel average misorientation. *Mater. Charact.* **139**, 49–58, DOI: 10.1016/j.matchar.2018.01.041 (2018).
13. Müller, M., Britz, D., Ulrich, L., Staudt, T. & Mücklich, F. Classification of Bainitic Structures Using Textural Parameters and Machine Learning Techniques. *Metals* **630**, 1–19, DOI: 10.3390/met10050630 (2020).
14. Bulgarevich, D. S., Tsukamoto, S., Kasuya, T., Demura, M. & Watanabe, M. Pattern recognition with machine learning on optical microscopy images of typical metallurgical microstructures. *Sci. Reports* **8**, 3–9, DOI: 10.1038/s41598-018-20438-6 (2018).

15. Azimi, S. M., Britz, D., Engstler, M., Fritz, M. & Mücklich, F. Advanced steel microstructural classification by deep learning methods. *Sci. Reports* **8**, 1–14, DOI: 10.1038/s41598-018-20037-5 (2018). 1706.06480.

16. DeCost, B. L., Lei, B., Francis, T. & Holm, E. A. High throughput quantitative metallography for complex microstructures using deep learning: A case study in ultrahigh carbon steel. *Microsc. Microanal.* **25**, 21–29, DOI: 10.1017/S1431927618015635 (2019). arXiv:1805.08693v1.

17. Pierson, K., Rahman, A. & Spear, A. D. Predicting Microstructure-Sensitive Fatigue-Crack Path in 3D Using a Machine Learning Framework. *Jom* **71**, 2680–2694, DOI: 10.1007/s11837-019-03572-y (2019).

18. Thomas, A., Durmaz, A. R., Straub, T. & Eberl, C. Automated Quantitative Analyses of Fatigue-Induced Surface Damage by Deep Learning. *Materials* **13**, 3298, DOI: 10.3390/ma13153298 (2020).

19. Pawlowski, N. *et al.* DLTK: State of the Art Reference Implementations for Deep Learning on Medical Images (2017). 1711.06853.

20. Falk, T. *et al.* U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* **16**, 67–70, DOI: 10.1038/s41592-018-0261-2 (2019).

21. Zitova, B. & Flusser, J. Image registration methods: a survey. *Image vision computing* **21**, 977–1000 (2003).

22. Britz, D., Webel, J. & Gola, J. A Correlative Approach to Capture and Quantify Substructures by Means of Image Registration. *Pract. Metallogr.* **54**, 685–696 (2017).

23. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**, 91–110, DOI: 10.1023/B:VISI.0000029664.99615.94 (2004).

24. Wright, S. I. & Nowell, M. M. EBSD image quality mapping. *Microsc. Microanal.* **12**, 72–84, DOI: 10.1017/S1431927606060090 (2006).

25. Arganda-Carreras, I. *et al.* Consistent and elastic registration of histological sections using vector-spline regularization. In *Lecture Notes in Computer Science*, vol. 4241 LNCS, 85–95, DOI: 10.1007/11889762_8 (Springer Verlag, 2006).

26. Paszke, A. *et al.* PyTorch : An Imperative Style , High-Performance Deep Learning Library. (2019). arXiv:1912.01703v1.

27. Santurkar, S., Tsipras, D., Ilyas, A. & Madry, A. How does batch normalization help optimization? *Adv. Neural Inf. Process. Syst.* **2018-Decem**, 2483–2493 (2018).

28. Drozdzal, M., Vorontsov, E., Chartrand, G. & Sep, C. V. The Importance of Skip Connections in Biomedical Image Segmentation. (2016). arXiv:1608.04117v2.

29. Buslaev, A. *et al.* Albumentations: Fast and flexible image augmentations. *Inf. (Switzerland)* **11**, 1–20, DOI: 10.3390/info11020125 (2020). 1809.06839.

30. Liaw, R. *et al.* Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118* (2018).

31. Lin, T. Y., Goyal, P., Girshick, R., He, K. & Dollar, P. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis Mach. Intell.* **42**, 318–327, DOI: 10.1109/TPAMI.2018.2858826 (2020). 1708.02002.

32. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization. *3rd Int. Conf. on Learn. Represent. ICLR 2015 - Conf. Track Proc.* 1–15 (2015). 1412.6980.

33. Jia Deng *et al.* ImageNet: A large-scale hierarchical image database. 248–255, DOI: 10.1109/cvprw.2009.5206848 (IEEE, 2009).

34. Bau, D., Zhou, B., Khosla, A., Oliva, A. & Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6541–6549 (2017).

35. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626 (2017).

36. Zaefferer, S., Romano, P. & Friedel, F. Ebsd as a tool to identify and quantify bainite and ferrite in low-alloyed al-trip steels. *J. microscopy* **230**, 499–508 (2008).

37. Hüllermeier, E. & Waegeman, W. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. 1–59, DOI: 10.1007/s10994-021-05946-3 (2019). 1910.09457.

38. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge data engineering* **21**, 1263–1284 (2009).

39. Naylor, J. P. The influence of the lath morphology on the yield stress and transition temperature of martensitic- bainitic steels. *Metall. Transactions A* **10**, 861–873, DOI: 10.1007/BF02658305 (1979).

40. Morito, S., Yoshida, H., Maki, T. & Huang, X. Effect of block size on the strength of lath martensite in low carbon steels. *Mater. Sci. Eng. A* **438**, 237–240 (2006).

41. Wojna, Z. *et al.* The devil is in the decoder. *Br. Mach. Vis. Conf. 2017, BMVC 2017* 1–13, DOI: 10.5244/c.31.10 (2017).

42. Sabottke, C. F. & Spieler, B. M. The Effect of Image Resolution on Deep Learning in Radiography. *Radiol. Artif. Intell.* **2**, e190015, DOI: 10.1148/ryai.2019190015 (2020).

43. Wu, R., Yan, S., Shan, Y., Dang, Q. & Sun, G. Deep Image: Scaling up Image Recognition. *arXiv preprint arXiv:1501.02876* **7** (2015). 1501.02876.

44. Luo, W., Li, Y., Urtasun, R. & Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 4905–4913 (2016). 1701.04128.

45. Wang, H. *et al.* Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, 108–126 (Springer, 2020).

46. Yu, F. & Koltun, V. Multi-scale context aggregation by dilated convolutions. In *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings* (2016). 1511.07122.

47. Aversa, R., Modarres, M. H., Cozzini, S., Ciancio, R. & Chiusole, A. Data descriptor: The first annotated set of scanning electron microscopy images for nanoscience. *Sci. Data* **5**, DOI: 10.1038/sdata.2018.172 (2018).

48. DeCost, B. L. *et al.* UHCSDB: UltraHigh Carbon Steel Micrograph DataBase: Tools for Exploring Large Heterogeneous Microstructure Datasets. *Integrating Mater. Manuf. Innov.* **6**, 197–205, DOI: 10.1007/s40192-017-0097-0 (2017).

49. Huang, S. W. *et al.* AugGAN: Cross domain adaptation with GAN-based data augmentation. *Lect. Notes Comput. Sci. (including subseries Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **11213 LNCS**, 731–744, DOI: 10.1007/978-3-030-01240-3_44 (2018).

50. Vu, T. H., Jain, H., Bucher, M., Cord, M. & Pérez, P. ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation (2018).

51. Gao, H., Yuan, H., Wang, Z. & Ji, S. Pixel Transposed Convolutional Networks. *IEEE Transactions on Pattern Analysis Mach. Intell.* **42**, 1218–1227, DOI: 10.1109/TPAMI.2019.2893965 (2020).

52. Prewitt, J. M. S. Object enhancement and extraction. *Pict. processing Psychopictorics* **10**, 15–19 (1970).

53. Min-cuts, C. P., Member, S. & Sminchisescu, C. CPMC : Automatic Object Segmentation Using. **34**, 1312–1328 (2012).

54. Kolesnikov, A. & Lampert, C. H. Seed , Expand and Constrain : Three Principles for Weakly-Supervised Image Segmentation. (2016). arXiv:1603.06098v3.

55. Huang, Z., Wang, X., Wang, J., Liu, W. & Wang, J. Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 7014–7023, DOI: 10.1109/CVPR.2018.00733 (2018).

56. Lee, J., Kim, E., Lee, S., Lee, J. & Yoon, S. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, 5262–5271, DOI: 10.1109/CVPR.2019.00541 (2019). 1902.10421.

57. Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S. N. & Chellappa, R. Learning from Synthetic Data: Addressing Domain Shift for Semantic Segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3752–3761, DOI: 10.1109/CVPR.2018.00395 (2018). 1711.06969.

58. Stewart, R. & Ermon, S. Label-free supervision of neural networks with physics and domain knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31 (2017).

Acknowledgements

The contribution of A.D. was funded by the Bosch-Forschungsstiftung im Stifterverband grant number T113/30074/17 and Open Access funding provided by Projekt DEAL. The authors wish to acknowledge the EFRE Funds of the European Commission and the State Chancellery of Saarland for support of activities within the ZuMat project. The authors would also like to thank steel manufacturer "AG der Dillinger Hüttenwerke" for providing the sample material and acknowledge support by German Research Foundation (DFG, Deutsche Forschungsgemeinschaft). Work at Carnegie Mellon University was supported by the National Science Foundation under grant CMMI-1826218.

Author contributions statement

Conceptualization: A.D., A.T., B.L., D.B., E.H., M.M; Data curation: A.D., M.M; Formal Analysis: A.D., A.T., B.L.; Investigation: A.D., A.T., B.L., M.M.; Methodology: A.D., A.T., B.L., M.M.; Project administration: A.D., B.L., D.B, M.M.; Resources: C.E., D.B., F.M., E.H., P.G.; Software: A.D., A.T., B.L.; Supervision: C.E., D.B., F.M., L.H., P.G.; Visualization: A.D., A.T., B.L., M.M.; Writing – original draft: A.D., A.T., B.L., D.B., M.M.; Writing – review & editing: all authors;

Additional information

Competing interests The authors declare no competing interests.