Machine Learning-Assisted Grain Segmentation on Scanning Electron Microscopy Images of High-Entropy Alloys

Matija Budja
mb32544@student.uni-lj.si
Faculty of Mathematics and Physics,
University of Ljubljana
Ljubljana, Slovenia

Anton Gradišek
anton.gradisek@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Andreja Jelen
andreja.jelen@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

The microstructure of a material holds the key to understanding its properties, performance and failure. Among the many parameters defining this microstructure, grains play a fundamental role. While grains can be characterized using various microscopic techniques—each requiring specific sample preparation—scanning electron microscopy (SEM) utilizing backscattered electron (BSE) imaging offers a significant advantage as BSE imaging completely eliminates the need for chemical etching of the sample. However, identifying individual grains on SEM-BSE images of unconventional metallic alloys, such as high-entropy alloys (HEAs), remains a major challenge. Due to their complex and peculiar microstructure, grains cannot be intuitively distinguished, rendering the microscope's standard built-in software insufficient for accurate analysis. Therefore, we fine-tuned and evaluated three pre-trained machine learning models (*ResNet-UNet*, *MiT-UNet*, and *DINOv2 with a linear segmentation head*) on an augmented dataset of manually annotated grain boundary images. To further improve boundary detection, we applied watershed segmentation as post-processing to model inferences. After testing, the best model was *MiT-UNet* trained with a combined *Dice and Focal loss*, yielding a final mean Intersection over Union (*mIoU*) of 50.9% and a grain boundary Intersection over Union of 20.7%. The results at this point are moderately useful for further analysis, however, we identified types of images where additional training will be required to improve performance.

1 Introduction

High-Entropy Alloys (HEAs) are a trendy topic in materials science with interesting physics, combining contradictory properties to achieve desired performance, such as superior thermal stability, mechanical strength, and oxidation resistance. These unique materials hold significant potential for advanced applications across leading industries, including aerospace, defense, military, biomedical, and chemical sectors [1]. High-entropy alloy processing is a relatively new approach in metallic alloy design with multiple principal elements in equimolar or close to equimolar ratios. According to this concept, high entropy of mixing can stabilize disordered solid solution phases (HEA phase) in competition with intermetallic phases and phase-segregated solid mixtures [2]. In order to better understand their properties, it is crucial to gain detailed insight into the microstructure on the scale of individual crystal domains, called grains. For detailed investigation, we need scanning electron microscopy (SEM), which produces images of the surface, and the BSE imaging technique which shows chemical contrast based on the atomic number (Z) of constituting chemical elements. Obtaining complete grain information is not straightforward, but requires an additional complex powerful technique called Electron Backscatter Diffraction (EBSD), which uses Bragg diffraction to determine crystal structure and orientation in consecutive points across the sample. Since many HEAs are composed of two or more crystallographic-chemical phases, which may be detected with a less complex SEM-BSE imaging technique, the grains could be recognized with the naked eye by an experienced observer. In SEM-BSE images, the chemical contrast becomes visible. Fig.1a) shows a typical SEM-BSE image of eutectic high-entropy alloy $CoCrFeNiZr_{0.5}$. It is composed of two phases, where the brighter one contains heavier elements - Zr - containing C15 Laves-phase intermetallic compound, and the darker lighter elements - Co, Cr, Fe, Ni - comprise the true solid solution high-entropy alloy [3]. In this figure we can also see that it consists of individual domains, which are parts of the sample that crystallized in the same way and share the same orientation. In Fig. 1 b) and c) we compare the images of the same sample, but a different site of interest, done with EBSD imaging. Fig. 1 b) shows the overlay of SEM-BSE contrast with grains and c) color coded grains. Pixels of the same color belong to the same grain. The EBSD system defines grains by measuring the Kikuchi diffraction pattern (DP) in each pixel. Based on those DPs and user-specific requirements, the algorithm defines grains. Here, the colors represent only different grains, and not their orientation. That information can be obtained by EBSD as well, but this is beyond the scope of the present paper. As there can be many SEM-BSE images created in a short time, 1000× faster than EBSD analysis, and the grains could be extracted using image analysis alone, this calls for automated methods for segmentation of images that will allow us to identify the domains and further study them. Current state-of-the-art commercial image analysis solutions, including both built-in microscope software and standalone packages, all fail to determine the grains in eutectic alloys [4]. The main reason is an inability to distinguish internal lamellae from colony boundaries, leading to severe over-segmentation. Furthermore, these systems struggle with multi-scale geometries in off-eutectic alloys and are highly sensitive to etching artifacts,



Figure 1: Micrographs of the $CoCrFeNiZr_{0.5}$ sample. a) SEM-BSE image, b) EBSD BSE image + grains, c) EBSD color coded grains, where b) and c) are from the same site of interest.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.skui.81

resulting in inaccurate phase and grain analysis. The same problem occurs for our HEA eutectic samples. Fortunately, many studies have already been done, focusing on using deep learning for their own eutectic systems [5, 6, 7]. The study in [5] shows the superiority of deep learning (FCNN) over commercial software based on hand-engineered algorithms. Their AI model achieved an accuracy of 94%, whereas traditional methods—utilized by the majority of commercial tools for analyzing complex metallographic textures—reached only 49%. Research activity is also very vivid, or even more developed, in other related fields, especially in the biomedical sciences. The failure of hard-coded commercial algorithms to properly characterize complex, multi-phase structures is not unique to materials science, but also well-documented across biomedical imaging, where traditional thresholding and watershed workflows yield inadequate accuracy due to low contrast and overlapping features. Across medical diagnostics and high-throughput cell profiling, deep learning architectures have consistently outpaced conventional rule-based pipelines by utilizing semantic feature representation and custom, domain-specific loss functions [8, 9, 10].

Here, we demonstrate the initial results of a customized algorithm that identifies individual domains in SEM-BSE images of polycrystalline eutectic high-entropy metallic alloys $CoCrFeNiZr_x$ (x = 0.4, 0.45, and 0.5) at different magnifications. We fine-tuned a series of conventional algorithms on a dedicated dataset, assisted with data augmentation methods. We evaluated the algorithms on a test set comprising held-out regions from eight out of nine source images, representing variations in alloy composition and microscope magnification. We discuss next steps in research.

2 Materials and Methods

2.1 Micrographs

The SEM-BSE images were taken with the high-resolution field-emission-gun scanning electron microscope Carl ZEISS Merlin with a high-resolution BSE detector: Energy Selective Backscattered (ESB) detector. EBSD measurements were conducted using integrated Bruker QUANTAX EBSD.

2.2 Dataset

An initial dataset of nine images was selected from studies of several eutectic HEAs, varying the concentration of crucial element, determining the mechanical properties of the alloy [3]. The images were manually segmented using Procreate on an Apple iPad, splitting them into 3 classes - *grain interior* (marked as white), *grain boundary* (black) and *impurities* (green). The grain boundary lines were drawn to be 4 px wide when straight. Seven source images were spatially divided into separate training/validation and test regions. One additional source image was used exclusively for training/validation, while another was used exclusively for testing. The images that were designated to be passed into the training pipeline were further split into two parts - ≈80% going to the training part of the dataset and ≈20% to the validation dataset. The split is not exact, as not all image sizes were divisible by 5. The training part of the labeled images was then split into tiles of size either 128 × 128px or 112 × 112px (depending on the machine learning model used) with a 50% overlap between succeeding tiles and the validation dataset into tiles of the same size, but with no overlap. All of the images were also normalized to the *ImageNet* mean and standard deviation values of [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225] for all three RGB channels (which are in this case just three repeated grayscale channels, as the pre-trained models expect three color channels).

The final step of the data preparation phase was the augmentation of the training dataset. The splitting and tiling of the full-size images left us with 754 training tiles and 59 validation tiles for the tile size of 128 × 128 px and 1003 training and 78 validation for the tile size of 112 × 112 px. The augmentation pipeline was applied *online* (meaning that non-deterministic augmentations were applied each time the dataset was loaded – inside each batch for every epoch; *offline* would mean that all of the augmented images are created before the training loop is initialized). The transformations used are presented in Table 1.

Table 1: Training data augmentation pipeline.

Technique	Parameters	Prob. (p)
Random Rotate 90°	/	0.50
Horizontal flip	/	0.50
Vertical flip	/	0.50
Affine Scale	Scale ∈ [0.85, 1.15]	0.40
CLAHE	Clip limit: 3.0, Grid: 8 × 8	0.40
Brightness & Contrast	ΔB: ± 15%, ΔC: ± 20%	0.40
Sharpening	$\alpha \in [0.1, 0.3]$	0.30
Blur / Noise (One-of)	Kernel [3, 5] / Mult. [0.99, 1.01]	0.40
Padding (if needed)	Size: 128 × 128 / 112 × 112 px	1.00

In addition, a 5× multiplier was applied to the training dataset, resulting in a final image count of 3770 training tiles and 59 validation tiles for tiles of size 128 × 128 px and 5015 training and 78 validation for tiles of size 112 × 112 px.

Some examples of manually segmented masks overlaid on the original SEM-BSE images are shown in the rightmost column of Fig. 2. For better visibility, the color scheme used in these overlays differs from the one described above: *grain interiors* are shown as transparent rather than white (leaving the original SEM-BSE image visible below), *grain boundaries* are shown in red rather than black, and *impurities* are shown in green.

2.3 Methods

The objective of each algorithm is to take the raw SEM-BSE images as input and produce corresponding segmented images in which the individual grains are identified. The raw SEM-BSE images constitute the input to the models, while the manually segmented masks are used only as ground-truth labels during training and evaluation. Thus, the white grain interiors, black grain boundaries, and green impurities from the annotated masks are not present in the input images supplied to the model.

The segmentation task was split into two phases: first, a number of different pre-trained machine learning models were fine-tuned on the training dataset comprising training regions from eight source images, as described in Section 2.2, using the manually segmented masks as ground-truth targets for backpropagation. The trained models were then applied to the raw SEM-BSE images in the testing dataset to perform semantic segmentation, outputting a pixel-wise prediction for three distinct classes: *grain*, *grain boundary* and *impurity*. In the second phase, a watershed algorithm was applied to these semantic predictions, obtaining the final segmented image. These segmentations were then compared with the manually segmented ground-truth masks of the testing dataset.

We evaluated three model variants based on combinations of feature extraction backbones and segmentation decoders:

- **ResNet-UNet – UResNet**: ResNet34 convolutional backbone pre-trained on *ImageNet-1k* with a UNet decoder [11, 12, 13]
- **MiT-UNet**: Mix Vision Transformer (MiT, specifically *mit_b2*) backbone pre-trained on *ImageNet-1k* with a UNet decoder [13, 14]
- **DINOv2-Linear**: DINOv2-Base self-supervised pre-trained on *LVD-142M* with a simple linear segmentation head [15, 16]

Each of the models was fine-tuned on the dataset from Section 2.2. *UResNet* and *MiT-UNet* used a split learning rate of 10^{-5} for the encoder and 10^{-3} for the decoder and segmentation head. For the *DINOv2*-based model only the segmentation head was trained, with the encoder parameters remaining frozen. This was done because experiments with the encoder parameters unfrozen failed to converge. *DINOv2-Linear* does not even contain a decoder, instead relying only on a simple 1×1 convolutional head to turn the encoder's patch embeddings directly into predicted logits. Because DINOv2 uses a patch size of 14×14 px, an input image of 112×112 px yields an 8×8 feature grid (64 total patches). The low-resolution logit map (8×8 px) produced by the 1×1 convolution is subsequently scaled back to the original 112×112 px resolution via bilinear interpolation. The segmentation head learning rate in this case was also set to 10^{-3}.

Additionally, each model was fine-tuned using three different loss functions, *Dice loss*, *Weighted Cross-Entropy + Dice loss*, and *Focal + Dice loss*. For the Weighted Cross-Entropy loss, class weights were set to inverse class frequencies and rescaled to a mean value of 1 for numerical stability. The composite loss functions (*Weighted Cross-Entropy + Dice* and *Focal + Dice*) were defined as unweighted linear combinations of their respective components.

The fine-tuning was run for 15 epochs, in each of which the backpropagation algorithm was run on the training dataset and the final epoch loss and the Intersection over Union for the boundary class (Eq. 1) were calculated on the validation dataset and saved along with the model weights. After all 15 epochs were completed, the checkpoint with the highest boundary Intersection over Union was selected as the best and saved.

In the segmentation pipeline, full-resolution input images were processed using a sliding-window approach. The images were first cropped into overlapping tiles matching the tile size used during training. The model then segmented each tile individually, after which the predictions were stitched back together into a continuous output of the same size as the input image using Gaussian-weighted blending along the overlapping boundaries to prevent edge artifacts. Following model inference, a distance transform was calculated for all pixels relative to the predicted boundary class, with the resulting distance map normalized to the range [0, 1]. After this, a peak detection algorithm was used to determine grain seeds for the watershed algorithm. To mitigate over-segmentation caused by prediction noise, an h-maxima filter was incorporated as an intermediate step, utilizing a fixed threshold of h = 0.125. The next step was to run the watershed algorithm, returning segmented grains. Finally, an edge detection algorithm was run on the segmented grains and the detected boundaries were dilated from 1 px to 4 px, in order for the comparison with the 4 px wide manually annotated boundaries to be fair.

2.4 Testing

The test set comprised held-out regions from eight source images, as described in Section 2.2. Seven of these source images also contributed spatially separate regions to the training/validation set, while one image was reserved exclusively for testing. The first two metrics, used to compare the predicted and manually annotated segmentations, are Intersection over Union (*IoU*) for the *grain boundary* class and mean Intersection over Union (*mIoU*). The *IoU* is defined as:

$$\text{IoU} = \frac{\text{Area of intersection}}{\text{Area of Union}} \tag{1}$$

This equation is applied independently to each of the three classes, but boundary IoU is additionally reported separately because accurate grain-boundary detection is particularly important for the subsequent watershed segmentation. The *mIoU* is then calculated as:

$$\text{mIoU} = \frac{\text{IoU}_{\text{boundary}} + \text{IoU}_{\text{grain interior}} + \text{IoU}_{\text{impurity}}}{3} \tag{2}$$

These metrics were evaluated both on the raw model predictions before running watershed and on the final watershed grain segmentation. Because the ground-truth grain boundaries were drawn as 4px wide lines, the final (after-watershed) grain boundary predictions were also diluted from 1 px to 4 px, for a fair comparison.

The other two metrics compared the number of predicted grains after watershed to the number in the manually annotated image, specifically the Mean Absolute Error (*MAE*):

$$\text{MAE} = \frac{1}{M} \sum_{i=1}^{M} |N^{(i)}_{\text{prediction}} - N^{(i)}_{\text{ground-truth}}| \tag{3}$$

and Mean Bias Error (*MBE*):

$$\Delta N = \frac{1}{M} \sum_{i=1}^{M} \left(N^{(i)}_{\text{prediction}} - N^{(i)}_{\text{ground-truth}} \right) \tag{4}$$

For the *MBE*, positive values indicate systematic over-segmentation and negative values systematic under-segmentation. M is the number of test images ($M = 8$).

3 Results and Discussion

Because the primary objective is accurate identification of final grain boundaries after watershed segmentation, post-watershed boundary IoU was used as the primary criterion for selecting the best-performing model. The results are presented in Table 2. For each of the tested models and losses (ResNet-UNet, MiT-UNet, DINOv2+Linear segmentation head; Dice, Weighted Cross Entropy + Dice, Focal + Dice), all of the IoU variations discussed in Section 2.4 were calculated for each of the eight held-out test regions. The mean and standard deviation were then calculated across these eight test regions. In addition, the two grain-count metrics (*MAE* and *MBE*) were calculated for each model–loss combination.

As we see in Table 2, the best model/loss combination is **MiT-UNet** with **Dice + Focal loss**. The worst model in all of the *IoU*-based metrics was *DINOv2-Linear*, though this is to be expected, as the simple 1×1 convolution segmentation head is not adequate for a complex segmentation task such as this. A model utilising the same DINOv2 backbone with a more advanced segmentation head would likely perform much better.

Weighted Cross-Entropy + Dice generally produced larger grain-count errors and stronger positive grain-count bias, indicating a greater tendency toward over-segmentation.

All of the final results are systematically prone to over-segmentation, as characterized by the large *positive* grain counting biases in Table 2. This cannot be attributed solely to the models themselves, as a large part of the error stems from the simplistic nature of the watershed algorithm applied to their predictions. It tends to over-segmentation, partly caused by noise in the models' predictions and partly by the irregular shapes of the grains themselves. In either case, a possible improvement of the segmentation architecture would be training the model to directly predict distance maps from grain boundaries, which can then be used by watershed, instead of the grain boundaries themselves, as discussed in [17].

Pertaining to the specific final segmentation results of the *MiT-UNet* model with *Dice + Focal* loss (with only a representative subset shown in Fig. 2), the quality of the results varied. Sample (4) provides an example of a good result. This is attributable to the borders being clearly visible and the grain interiors being relatively *homogeneous* (relative to the other samples). Others were not as good, for example sample (7) was under-segmented, even though the model tends to over-segmentation – probably a consequence of its different contrast from the other samples in the training dataset. For an example of over-segmentation as a fault of the watershed algorithm and not the model itself, consider sample (5). While the model has no difficulty segmenting the very homogeneous gray *"blobs"*, the watershed algorithm tends to split them into multiple grains on account of their irregular shape.

We also see from Fig. 2 that the model tends to predict grain boundaries that are not closed contours, which affects the watershed algorithm's performance. This is not unique to *MiT-UNet*, instead it is observed for all of the tested models. This is a result of the samples being processed tile-by-tile, instead of the entire image at once, reducing the context available to the models.

The relatively low IoU values shown in Table 2 are not to be viewed as the models inherently performing poorly. First, the segmentation pipeline targets reliable grain alignment where visual accuracy triumphs over pixel-perfect accuracy. Second, the ground-truth annotations on such complex structures contain inherent boundary ambiguity, artificially lowering strict pixel-matching metrics, such as *IoU*, while the model yields usable segmentations.

4 Conclusion

SEM-BSE imaging is a great technique to visualize the microstructure in eutectic HEAs. In combination with machine learning assisted grain segmentation it can compete with the cumbersome EBSD grain size determination. This preliminary study on grain size determination from SEM-BSE images by machine learning yields functionally sufficient segmentations, with room for improvement. Next steps will be attaching a more capable decoder and segmentation head onto the *DINOv2* backbone or training the models to directly predict the distance transform used in the watershed algorithm, instead of predicting the boundaries themselves. Another idea might be to use some kind of adaptive h-suppression instead of the hard cutoff used here. Lastly, a larger training dataset would surely improve model performance.

Keywords: high-entropy alloys, microstructure, scanning electron microscopy (SEM), grain characterization, deep learning, semantic segmentation, transfer learning

Acknowledgments

Authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P1-0125, P2-0209 and research project No. PR-20000-1). AJ acknowledges Dr. Hae Jin Kim and Dr. Hwanuk Guim at KBSI Daejeon, South Korea for the opportunity to use high resolution SEM Carl ZEISS Merlin.

References

[1] Akriti Menon, Dipen Kumar Rajak, and Dwesh Kumar Singh. 2025. High entropy alloys: next-generation material for space exploration. *Materials Chemistry and Physics*, 329, 130091.

[2] Jien-Wei Yeh. 2013. Alloy design strategies and future trends in high-entropy alloys. *Jom*, 65, 12, 1759–1771.

[3] S Vrtnik et al. 2018. Magnetism of cocrfenizrx eutectic high-entropy alloys. *Intermetallics*, 93, 122–133.

[4] Xiang Li, Linyi Cui, Jikang Li, Ying Chen, Wei Han, Sara Shonkwiler, and Sara McMains. 2022. Automation of intercept method for grain size measurement: a topological skeleton approach. *Materials Design*, 224, 111358. doi:https://doi.org/10.1016/j.matdes.2022.111358.

[5] Seyedmajid Azimi, Dominik Britz, Michael Engstler, Mario Fritz, and Frank Mücklich. 2017. Advanced steel microstructure classification by deep learning methods. *CoRR*, abs/1706.06480. http://arxiv.org/abs/1706.06480 arXiv: 1706.06480.

[6] Abdul Mueez and Shruti Vyas. 2026. Bridging foundation models and astm metallurgical standards for automated grain size estimation from microscopy images. (2026). https://arxiv.org/abs/2604.18957 arXiv: 2604.18957 [cs.CV].

[7] Haiyan Liu, Penghua Zhu, and Chenyu Tan. 2026. Deep learning-based image classification for microstructural analysis in computational materials science. *Frontiers in Materials*, Volume 12 - 2025. doi:10.3389/fmats.2025.1648653.

[8] Xiaohang Fu et al. 2024. Bidcell: biologically-informed self-supervised learning for segmentation of subcellular spatial transcriptomics data. *Nature Communications*, 15, 1, 509. doi:10.1038/s41467-023-44560-w.

[9] Yan Xu, Rixiang Quan, Weiting Xu, Yi Huang, Xiaolong Chen, and Fengyuan Liu. 2024. Advances in medical image segmentation: a comprehensive review of traditional, deep learning and hybrid approaches. *Bioengineering*, 11, 10. doi:10.3390/bioengineering11101034.

[10] Rongrong Yang et al. 2026. Systematic evaluation of computational methods for cell segmentation. *Briefings in Bioinformatics*, 27, 1, (Jan. 2026), bbag066. eprint: https://academic.oup.com/bib/article-pdf/27/1/bbag066/67092491/bbag066.pdf. doi:10.1093/bib/bbag066.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385. http://arxiv.org/abs/1512.03385 arXiv: 1512.03385.

[12] Wenkai Li, Zhe Gao, and Yaqing Song. 2025. U-resnet, a novel network fusion method for image classification and segmentation. *Sensors*, 25, 17. doi:10.3390/s25175600.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597. http://arxiv.org/abs/1505.04597 arXiv: 1505.04597.

[14] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José M. Álvarez, and Ping Luo. 2021. Segformer: simple and efficient design for semantic segmentation with transformers. *CoRR*, abs/2105.15203. https://arxiv.org/abs/2105.15203 arXiv: 2105.15203.

[15] Maxime Oquab et al. 2024. Dinov2: learning robust visual features without supervision. (2024). https://arxiv.org/abs/2304.07193 arXiv: 2304.07193 [cs.CV].

[16] Niels Rogge. 2025. Tutorials. (2025). %5Bhttps://github.com/NielsRogge/tutorials%5D(https://github.com/NielsRogge/Transformers-Tutorials).

[17] Min Bai and Raquel Urtasun. 2016. Deep watershed transform for instance segmentation. *CoRR*, abs/1611.08303. http://arxiv.org/abs/1611.08303 arXiv: 1611.08303.

Table 2: Segmentation performance and grain counting accuracy across models and loss functions (Mean ± Standard Deviation). Both IoU metrics are calculated on the raw model output (pre-watershed) and on the final segmentation (post-watershed). *WCE* is short for Weighted Cross Entropy. Best results are bold and underlined.

Model	Loss	Boundary IoU (%)		Mean IoU (%)		Grain Count Error	
		Pre-Watershed	Post-Watershed	Pre-Watershed	Post-Watershed	MAE	Bias (ΔN)
UResNet	WCE + Dice	14.8 ± 5.5	13.0 ± 6.8	42.3 ± 9.6	43.5 ± 9.7	19.50	+19.50
	Dice	18.8 ± 11.8	14.1 ± 10.0	49.9 ± 8.1	48.3 ± 7.7	4.88	+1.88
	Focal + Dice	20.6 ± 11.3	16.6 ± 8.6	51.6 ± 9.7	50.1 ± 9.1	4.00	**+1.00**
MiT-UNet	WCE + Dice	21.4 ± 6.6	17.7 ± 6.3	51.4 ± 8.0	**51.3 ± 7.6**	8.88	+8.12
	Dice	**26.0 ± 9.5**	19.8 ± 7.5	52.1 ± 8.6	50.1 ± 8.0	4.12	+2.62
	Focal + Dice	25.3 ± 9.5	**20.7 ± 7.5**	**52.4 ± 9.1**	50.9 ± 8.6	4.25	+2.25
DINOv2-Linear	WCE + Dice	10.6 ± 4.0	7.3 ± 3.5	40.5 ± 7.1	40.1 ± 7.7	9.88	+8.38
	Dice	11.3 ± 4.6	8.7 ± 2.9	39.5 ± 4.5	39.4 ± 4.4	5.00	+4.25
	Focal + Dice	11.0 ± 5.4	8.0 ± 3.8	43.8 ± 6.3	42.9 ± 6.3	**3.38**	+2.62

Figure 2: Segmented images from the testing dataset using the MiT-UNet model, fine-tuned with Focal + Dice loss (best-performing configuration according to post-watershed boundary IoU). Seven (out of the full eight) representative test regions are shown. For improved visualization, *grain boundaries* that are black in the annotation masks are displayed in red in the overlays, while *impurities* are displayed in green. The *grain interiors* are omitted so that the original SEM-BSE scan remains visible. Column (a) is the original image, (b) the raw model predictions, (c) the final segmentation after running watershed, and (d) the manually annotated ground truth. Scale bars are shown only in the first column, as images in the same row have the same scale.

