

CS-UNet: A Flexible Segmentation Algorithm for Microscopy Images

Khaled Alrfou, Tian Zhao
University of Wisconsin – Milwaukee
3200 N Cramer St., Milwaukee, WI, 53211
{kalrfou, tzhao}@uwm.edu

Amir Kordijaz
University of Southern Maine
135 John Mitchell Center, Gorham, ME, 04038
Amir.kordijazi@maine.edu

Abstract

CS-UNet is a U-shaped image-segmentation algorithm with parallel CNN and Transformer encoders. This algorithm leverages the relative strength of CNN and Transformers, and enables flexible combination of encoders pre-trained on different datasets to extract the maximum benefit of transfer-learning. CS-UNet is evaluated for its segmentation accuracy on microscopy images of materials science. The performance of CS-UNet is comparable or better than state-of-the-art algorithms based on CNN or Transformer encoders. Pre-training the encoders of CS-UNet on microscopy images further improves its performance in out-of-distribution learning and one-shot learning. The Intersection over Union (IoU) accuracy of nickel-based super-alloy images is improved from 77.89% to 82.13% for out-of-distribution learning and IoU accuracy of environmental-barrier-coating images is improved from 65.9% to 70.45% for one-shot learning. This suggests that Transformer and CNN complement each other and pre-training on images with similar attributes is beneficial to the downstream tasks. The implementation is freely available¹.

1. Introduction

Deep Learning (DL) has been widely applied to complex systems because of its ability to extract important information automatically. Researchers have applied DL algorithms to image analysis to identify structures and to determine the relationship between microstructure and performance [1]. DL has been demonstrated to complement physics-based methods for materials design. However, DL requires large amount of training data while the limited number of microscopy images tends to reduce its effectiveness. Learning techniques, such as transfer-learning, multi-fidelity modeling, and active learning, were developed to make DL applicable to smaller datasets [1, 2]. Transfer-learning uses

the parameters of a model pre-trained on a larger dataset to initialize a model trained on a smaller dataset for a downstream task. For example, a Convolutional Neural Network (CNN) pre-trained on natural images can be used to initialize a neural network for image segmentation such as UNet to improve its precision and reduce the training time.

In recent years, attention-based neural networks called Transformers are widely adopted in computer vision. While CNN extracts features from local regions of images using convolution filters to capture the spatial relation between the pixels, Transformer divides an image into patches and feeds them into a Transformer-based encoder to capture the long-range relation between pixels across the images [3, 4]. Thus, it is possible that a combination of CNN and Transformer may be more effective in transfer-learning than either of the models alone.

In this paper, we present a segmentation algorithm called CS-UNet that includes parallel CNN and Transformer encoders in a U-shaped encoder-decoder architecture. The parameters of the encoders are initialized from models pre-trained on natural or microscopy images. Each encoder transforms the input image into a latent representation vector to extract semantic information. Each decoder maps the extracted information back to each pixel in the input image to generate a pixel-wise classification of the image [1, 5]. The output of the CNN and Transformer encoders are fused before connecting to the decoder. CS-UNet allows great flexibility in combining different types of CNN and Transformer encoders pre-trained on different types of data to allow optimal choices of encoders for the segmentation tasks.

Encoder-decoder architecture allows pre-training to improve segmentation accuracy. Pre-training with in-domain images should improve microscopy image segmentation since natural images has high-level features that do not exist in microscopy images. Recent work by Stuckner *et al.* [6] confirmed the benefit of pre-training CNN encoders on a microscopy dataset called MicroNet with over 100,000 images. They evaluated the CNN encoders with the segmentation of nickel-based super alloy (Super) and environmental barrier coating (EBC) images. Their tests showed higher

¹<https://github.com/Kalrfou/SwinT-pretrained-microscopy-models>

accuracy in Intersection over Union (IoU) for one-shot and few-shot learning and for out-of-distribution images that have different compositions, etching, and imaging conditions than the training images.

To evaluate the performance of CS-UNet, we pre-trained CNN and Transformer encoders on different types of datasets and performed segmentation on the same test sets used by Stuckner *et al.* [6]. We chose the tiny version of Swin-Transformer – Swin-T [7] as our Transformer encoder. While we can initialize the CNN encoders using the CNN models of Stuckner *et al.* [6], we are unable to obtain their dataset MicroNet to train our Swin-T encoder. To this end, we created a similar pre-training dataset with about 50,000 microscopy images in 74 classes, which we will refer to as MicroLite.

Our experiments showed that CS-UNet has similar or better accuracy than the state-of-the-art algorithms based on CNN or Transformer encoders including the CNN encoders evaluated in Stuckner *et al.* [6].

2. Related Work

CNN uses the convolution operators to provide translational equivariance but its local receptive field has limitation in capturing long-range relation between pixels [4]. In recent years, Transformer [8] has been used in place of CNN for computer vision (CV) tasks to overcome this limitation [3] in areas such as image recognition, image segmentation [9], object detection [10, 11], image super-resolution, and image generation [12].

In recent years, many variations of U-shaped networks have been used in image segmentation. U-Net is a Fully Convolutional Network (FCN) [13, 14], which is a symmetric, U-shaped, encoder-decoder neural network for semantic image segmentation. U-Net typically consists of a down-sampling encoder and an up-sampling decoder structure and a “skip connection” between them. These connections copy feature maps from the encoder and concatenate them with the feature maps in the decoder. Transformer encoder was used in SegFormer [15], which is a semantic segmentation framework that combines Transformer encoders with lightweight MultiLayer Perceptron (MLP) decoders.

CNN and Transformer were combined in TransUNet [16], which is a U-shaped architecture that employs a hybrid CNN-Transformer encoder followed by multiple up-sampling layers in the CNN decoder. This method leverages both detailed high-resolution spatial information from CNN features and the global context encoded by Transformers. The TransUNet architecture includes 12 ViT [3] layers in the encoder, which encodes tokenized image patches obtained from the CNN layers. These encoded features are then up-sampled in the decoder to generate the final segmentation map, with skip-connections incorporated. TransUNet achieved high performance compared with the CNN-

based models.

Swin-UNet [17] uses only Transformer encoders in its U-shaped encoder-decoder architecture for medical image segmentation. Swin-UNet includes skip-connections for local-global semantic-feature learning by feeding the tokenized image patches into the model. Both the encoder and decoder structures of Swin-UNet were inspired by the hierarchical Swin-Transformer [7] with shifted windows.

In summary, the hybrid architectures mentioned above either replace CNN with a Transformer in the encoder [18] or stack a CNN with a Transformer sequentially to form a new encoder [16]. Replacing CNN with a Transformer in the encoder gives the ability to model long distance dependency in the network. However, it results in a lack of detailed texture feature extraction due to the removal of CNN in the encoder. Stacking CNN with a Transformer to form a new encoder fails to account for the complementary relationship between the global modeling capability of self-attention and the local modeling capability of convolution. Instead, they treat the convolution operation and self-attention as two separate and unrelated operations [19, 20].

3. Methodology

In this section, we give details on how CS-UNet is implemented, how the dataset MicroLite is created, and how the encoders of CS-UNet are pre-trained.

3.1. CS-UNet Architecture

CS-UNet, as shown in Figure 1, is such a hybrid model that consists of CNN encoders, Transformer encoders, bottlenecks, Transformer decoders, and skip connections. The CNN encoders extract low-level features and the Swin-T encoders extract global contextual features. Each Swin-T encoder operates on the input image divided into non-overlapping patches, applying self-attention mechanisms to capture global dependencies. The Swin-T encoders capture long-range dependencies and contextual information from the entire image at different scales.

The decoder is similar to that of Swin-UNet[17], which employs the patch-expanding layer to up-sample the extracted deep features by reshaping the feature maps of adjacent dimensions to form a higher-resolution feature map, which effectively achieves a $2\times$ up-sampling. Additionally, it reduces the feature dimension to half of the original dimension. This allows the decoder to reconstruct the output with increased spatial resolution while reducing the feature dimension for efficient processing. The final patch-expanding layer further performs a $4\times$ up-sampling to restore the resolution of the feature maps to match the input resolution ($W \times H$). Finally, a linear projection layer is applied to these up-sampled features to generate pixel-level segmentation predictions. Different CNN families can be

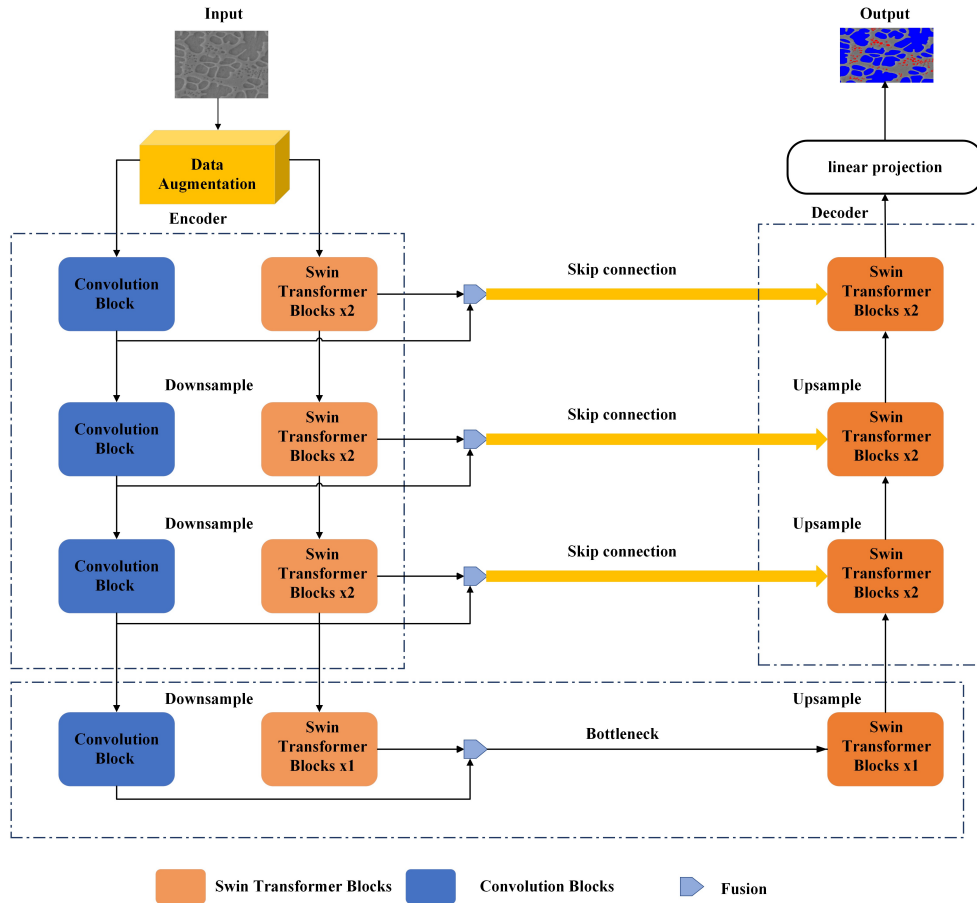


Figure 1. CS-UNet architecture includes CNN and Swin-T encoders, bottlenecks, skip connections, and Swin-T decoder.

used in the encoder part such as EfficientNet, ResNet, MobileNet, DenseNet, VGG, and Inception. We initialize CNN weights using MicroNet and the transformer weights using MicroLite.

3.2. Pre-training Dataset

The MicroLite images were collected from multiple sources including images from different materials and compounds using several measurement techniques such as light microscopy, SEM, TEM, and X-ray. MicroLite aggregates the Aversa dataset [21], UltraHigh Carbon Steel Micrograph [22], SEM images from the Materials Data Repository, and the images from some recent publications [23–30]. The Aversa dataset includes over 25,000 SEM microscopy images in 10 classes, where each class consists of images in different scales (including 1, 2, 10, 20 μm and 100, 200 nm) and contrast. To properly classify these images, we used a pre-trained VGG-16 model to extract feature maps from these images and used a K-means algorithm to cluster the feature maps so that images with similar feature maps are grouped in the same class. After the pre-processing step,

we obtained 53 classes. The authors of Aversa dataset manually classified a small set of the images (1038) into a hierarchical dataset, where the 10 classes are further divided into 27 subclasses [21]. Our classification of these 1038 images is largely consistent with the manually assigned subclasses. Note that we have more classes since we processed the entire Aversa dataset. In total, MicroLite includes about 50,000 microscopy images labelled in 74 classes.

3.3. Pre-train Swin-T Encoders

We pre-trained Swin Transformers on microscopy images on classification tasks so that it can be transferred to segmentation tasks. The classification tasks use Swin-T, which is the tiny version of the Swin Transformer. Swin-T has two architectures: the original Swin-T with [2,2,6,2] transformer blocks and the intermediate network with [2,2,2,2] transformer blocks. We speculate that intermediate network may be enough for microscopy analysis tasks since the earlier layers learn corner edges and shapes, the intermediate layers learn the texture or patterns, and deeper network layers in the original models learn the high-level features. The

Table 1. The top / average performance (IoU) of UNet++ / UNet pre-trained on MicroNet [6], Transformer-based algorithms pre-trained on MicroLite, and CS-UNet pre-trained on MicroNet and MicroLite. The highest IoU percentages are shown in bold font.

Test Set	UNet++ / UNet	Transformer	top Transformer algo.	CS-UNet	CNN of top CS-UNet
Super-1	96.4% / 95.89%	95.72% / 94.84%	Swin-Unet	96.43% / 96.14%	InceptionV4
Super-2	94.2% / 95.24%	95.16% / 93.53%	TransDeepLabV3+	96.06% / 95.7%	VGG16_bn
Super-3	93.0% / 72.2%	92.23% / 76.39%	TransDeepLabV3+	93.5% / 77.99%	EfficientNet-b4
Super-4	78.5% / 74.57%	78.91% / 73.57%	Swin-Unet	82.13% / 75.08%	EfficientNet-b3
EBC-1	97.6% / 95.17%	96.59% / 93.01%	TransDeepLabV3+	97.66% / 95.98%	SE_ResNet152
EBC-2	93.3% / 84.6%	91.11% / 84.3%	TransDeepLabV3+	92.82% / 86.73%	SE_ResNet-101
EBC-3	65.9% / 42.58%	82.13% / 56.72%	Swin-Unet	70.46% / 45.69%	SE_ResNeXt-101_32x4d

original and intermediate Swin-T models were pre-trained on MicroLite from scratch, where the model weights are randomly initialized. The two models were also pre-trained on ImageNet and fine-tuned on MicroLite.

The pre-training step uses an AdamW optimizer for 30 epochs with a cosine-decay learning-rate scheduler with 5 epochs of linear warm-up and batch size of 128. The initial learning rate is 10^{-3} and weight decay is 0.05. The fine-tuning step also uses an AdamW optimizer for 30 epochs with a batch size of 128 but the learning rate is reduced to 10^{-5} and the weight decay is reduced to 10^{-8} . Models were trained until there was no improvement to the validation score using an early stopping criterion with a patience of 5 epochs. Training data had been augmented using albumenations library, which included random changes to the contrast and the brightness, vertical and horizontal flips, photometric distortions, and added noise. Swin-T models were trained by classifying microscopy images into 74 different classes. Swin-T models were either pre-trained on ImageNet and fine-tuned on MicroLite, or trained with MicroLite with randomized parameters. The training stops when the validation accuracy does not improve after 5 epochs. The model accuracy is evaluated using the top-1 and top-5 accuracy. The top-1 accuracy measures the percentage of test samples for which the correct label is predicted while the top-5 accuracy measures the percentage of correct labeling in the top five predictions. All segmentation models in this study were trained using PyTorch [31].

4. Result

We evaluated CS-UNet by comparing its performance with the results of Stuckner *et al.* [6]’s 7 microscopy datasets derived from two materials: nickel-based super-alloys (Super) and Environmental Barrier Coatings (EBC). Super datasets have 3 classes: matrix, secondary, and tertiary. EBC datasets have two classes: the oxide layer and the background (non-oxide) layer. Super-1 and EBC-1 contain the full dataset labeled for their respective materials. Super-2

and EBC-2 have only 4 images in the training set to evaluate the performance of few-shot learning. Super-3 and EBC-3 have only 1 image in the training set to evaluate performance of one-shot learning. Super-4 have test images taken under different imaging and sample conditions to evaluate the performance of out-of-distribution learning. EBC and Super datasets were augmented in ways similar to Stucker *et al.* [6], which includes random cropping, random changes to contrast, brightness, and gamma, and added blurring or sharpening. EBC dataset was horizontally flipped and Super dataset was randomly flipped and rotated.

Table 1 compares the top and average performance of UNet++ / UNet pre-trained on MicroNet (Figure 3–5 in [6]), segmentation algorithms using Transformer (Swin-Unet, TransDeepLabV3+, and HiFormer) pre-trained on MicroLite, and CS-UNet pre-trained on MicroNet and MicroLite. The top results for each test are shown in bold font. CS-UNet has the best performance in most experiments except EBC-2 and EBC-3. For experiments with ample training data such as Super-1 and EBC-1, the difference between UNet++/UNet, Transformer, and CS-UNet is small. For few-shot learning experiments such as Super-2 and EBC-2, the accuracy gain of CS-UNet is modest. For one-shot learning experiments, the result is mixed, where CS-UNet has modest improvement in Super-3 while significant gain in EBC-3. For out of context learning, CS-UNet shows significant improvement over UNet or Transformer.

5. Conclusion

Our results show that CS-UNet is more consistent than the prior algorithms. CS-UNet is similar or better than UNet++/UNet in all experiments and it is better than Transformers for most experiments. Though MicroLite has less than half the number of images in MicroNet, CS-UNet and Transformer models pre-trained on MicroLite has comparable or better performance than that of UNet++/UNet models pre-trained on MicroNet.

References

- [1] M. Ge, F. Su, Z. Zhao, and D. Su, "Deep learning analysis on microscopic imaging in materials science," *Materials Today Nano*, vol. 11, p. 100087, 2020. [1](#)
- [2] K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. W. Park, A. Choudhary, A. Agrawal, S. J. Billinge, *et al.*, "Recent advances and applications of deep learning methods in materials science," *npj Computational Materials*, vol. 8, no. 1, p. 59, 2022. [1](#)
- [3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. [1](#), [2](#)
- [4] K. Alrfou, A. Kordijazi, and T. Zhao, "Computer vision methods for the microstructural analysis of materials: The state-of-the-art and future perspectives," *arXiv preprint arXiv:2208.04149*, 2022. [1](#), [2](#)
- [5] G. Jacquemet, "Deep learning to analyse microscopy images," *The Biochemist*, vol. 43, no. 5, pp. 60–64, 2021. [1](#)
- [6] J. Stuckner, B. Harder, and T. M. Smith, "Microstructure segmentation with deep learning encoders pre-trained on a large microscopy dataset," *npj Computational Materials*, vol. 8, no. 1, p. 200, 2022. [1](#), [2](#), [4](#)
- [7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021. [2](#)
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. [2](#)
- [9] L. Ye, M. Roohan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10502–10511, 2019. [2](#)
- [10] J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," [2](#)
- [11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 213–229, Springer, 2020. [2](#)
- [12] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*, pp. 7354–7363, PMLR, 2019. [2](#)
- [13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015. [2](#)
- [14] K. Alrfou, A. Kordijazi, P. Rohatgi, and T. Zhao, "Synergy of unsupervised and supervised machine learning methods for the segmentation of the graphite particles in the microstructure of ductile iron," *Materials Today Communications*, vol. 30, p. 103174, 2022. [2](#)
- [15] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077–12090, 2021. [2](#)
- [16] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021. [2](#)
- [17] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European Conference on Computer Vision*, pp. 205–218, Springer, 2022. [2](#)
- [18] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*, pp. 272–284, Springer, 2022. [2](#)
- [19] J. Wang, H. Zhao, W. Liang, S. Wang, and Y. Zhang, "Cross-convolutional transformer for automated multi-organs segmentation in a variety of medical images," *Physics in Medicine & Biology*, vol. 68, no. 3, p. 035008, 2023. [2](#)
- [20] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pp. 14–24, Springer, 2021. [2](#)
- [21] R. Aversa, M. H. Modarres, S. Cozzini, R. Ciancio, and A. Chiusole, "The first annotated set of scanning electron microscopy images for nanoscience," *Scientific data*, vol. 5, no. 1, pp. 1–10, 2018. [3](#)
- [22] B. L. DeCost, M. D. Hecht, T. Francis, B. A. Webler, Y. N. Picard, and E. A. Holm, "Uhcddb: ultrahigh carbon steel micrograph database: tools for exploring large heterogeneous microstructure datasets," *Integrating Materials and Manufacturing Innovation*, vol. 6, pp. 197–205, 2017. [3](#)
- [23] E. Christiansen, C. D. Marioara, B. Holmedal, O. S. Hopperstad, and R. Holmestad, "Nano-scale characterisation of sheared β " precipitates in a deformed al-mg-si alloy," *Scientific reports*, vol. 9, no. 1, p. 17446, 2019. [3](#)
- [24] L. P. Mikkelsen, S. Fæster, S. Goutianos, and B. F. Sørensen, "Scanning electron microscopy datasets for local fibre volume fraction determination in non-crimp glass-fibre reinforced composites," *Data in Brief*, vol. 35, p. 106868, 2021.
- [25] F. B. Salling, N. Jeppesen, M. R. Sonne, J. H. Hattel, and L. P. Mikkelsen, "Individual fibre inclination segmentation from x-ray computed tomography using principal component analysis," *Journal of Composite Materials*, vol. 56, no. 1, pp. 83–98, 2022.
- [26] S. Masubuchi, E. Watanabe, Y. Seo, S. Okazaki, T. Sasagawa, K. Watanabe, T. Taniguchi, and T. Machida,

“Deep-learning-based image segmentation integrated with optical microscopy for automatically searching for two-dimensional materials,” *npj 2D Materials and Applications*, vol. 4, no. 1, p. 3, 2020.

- [27] D. A. Boiko, E. O. Pentsak, V. A. Cherepanova, and V. P. Ananikov, “Electron microscopy dataset for the recognition of nanoscale ordering effects and location of nanoparticles,” *Scientific data*, vol. 7, no. 1, p. 101, 2020.
- [28] P. Creveling, W. Whitacre, and M. Czabaj, “Synthetic x-ray microtomographic image data of fiber-reinforced composites,” 2019.
- [29] M. Klinkmüller, G. Schreurs, M. Rosenau, and H. Kemnitz, “Properties of granular analogue model materials: A community wide survey,” *Tectonophysics*, vol. 684, pp. 23–38, 2016.
- [30] R. Van Stone, J. Low, and J. Shannon, “Investigation of the fracture mechanism of ti-5ai-2.5 sn at cryogenic temperatures,” *Metallurgical Transactions A*, vol. 9, pp. 539–552, 1978. 3
- [31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimeshein, L. Antiga, *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019. 4