Self-Supervised Learning with Generative Adversarial Networks for Electron Microscopy

Bashir Kazimi[1], Karina Ruzaeva[1], Stefan Sandfeld[1,2]
[1]Forschungszentrum Jülich GmbH, Jülich, IAS-9, Germany
[2]RWTH Aachen University, Aachen, Germany
{b.kazimi,k.ruzaeva,s.sandfeld}@fz-juelich.de

Abstract

In this work, we explore the potential of self-supervised learning with Generative Adversarial Networks (GANs) for electron microscopy datasets. We show how self-supervised pretraining facilitates efficient fine-tuning for a spectrum of downstream tasks, including semantic segmentation, denoising, noise & background removal, and super-resolution. Experimentation with varying model complexities and receptive field sizes reveals the remarkable phenomenon that fine-tuned models of lower complexity consistently outperform more complex models with random weight initialization. We demonstrate the versatility of self-supervised pretraining across various downstream tasks in the context of electron microscopy, allowing faster convergence and better performance. We conclude that self-supervised pretraining serves as a powerful catalyst, being especially advantageous when limited annotated data are available and efficient scaling of computational cost is important.

1. Introduction

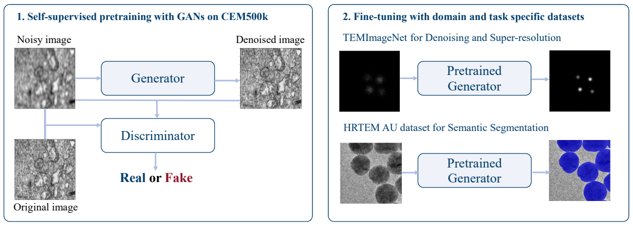
1. Self-supervised pretraining with GANs on CEM500k
Noisy image
Denoised image
Generator
Discriminator
Real or Fake
Original image
2. Fine-tuning with domain and task specific datasets
TEMImageNet for Denoising and Super-resolution
Pretrained Generator
HRTEM AU dataset for Semantic Segmentation
Pretrained Generator

Figure 1. The proposed pertaining pipeline, that includes GAN-based pretrainig on CEM500k dataset [9] followed by fine-tuning for downstream tasks: semantic segmentation of Gold nanoparticles [46], and super-resolution and denoising using the TEMImageNET dataset [26].

Microscopy, a fundamental tool in scientific research for several centuries, encompasses various branches, including optical, electron, scanning probe, and X-ray microscopy [24]. Electron microscopy (EM) is a technique that uses a beam of accelerated electrons to obtain high-resolution images of biological as well as non-biological specimens. Applications of this technique exist across various scientific domains, including biology, materials science, nanotechnology, and physics. In the field of biology, EM has been used for studying a wide range of biological samples such as lungs, muscles, bones, or nerve tissue [21]. In materials science, it has been utilized for visualization of the growth and characterization of nano- and microstructures [27], orientation mapping of semicrystalline polymers [32], and the identification of crystal lattice defects [31].

Imaging techniques and statistical analysis in EM have been instrumental in providing insights into the structure and properties of materials at various scales. Statistical analysis and classical machine learning methods have been used to analyze nanoparticles [11, 23], identify defects in metals [43, 53], and enhancing the quality of superresolution results in correlative tomography [40]. These methods, however, have shortcomings, such as limited resolution, time-consuming sample preparation, and the need for expert interpretation of results.

Deep learning (DL) and computer vision have been increasingly employed to address these limitations, enhance the capabilities of EM, and overcome the limitations of classical imaging and analysis methods by providing automated analysis, improved resolution, and enhanced interpretation of complex data: DL enables the extraction of valuable information from large datasets and offers new opportunities for quantitative image analysis in EM [3]. It has been used for analyzing nanoparticles in TEM images [46], denoising TEM images [49], identifying clean graphene areas [39], automatically segmenting and tracking of crystalline defects, [13, 38], decoding crystallography from high-resolution electron imaging and diffraction datasets [1], understanding important features of DL models for segmentation of high-resolution transmission electron microscopy (TEM) images [17], and segmentation in large-scale cellular EM [4], focused ion-beam scanning EM (FIB-SEM) [22] and high-resolution TEM data [14].

Conventional DL methods, such as convolutional neural networks, require large annotated datasets to be able to learn and generalize well on unseen examples. Manual annotation of datasets, especially EM images, is a time-consuming and labor-intensive task. To alleviate this, techniques such as transfer learning and self-supervised learning can be used. These techniques offer significant advantages for a range of computer vision tasks by enabling models to obtain general features from extensive datasets, facilitating knowledge transfer to specific tasks with limited labeled data. These approaches significantly reduce annotation costs, mitigate the problem of data scarcity, and strongly enhance generalization to unseen scenarios. Pretrained models exhibit faster convergence during fine-tuning, possess broader applicability across tasks, and provide resource-efficient solutions. The robust representations acquired through self-supervised learning contribute to improved performance in real-world scenarios, establishing it as an essential strategy in computer vision tasks across diverse domains [16, 35, 42, 48].

Self-supervised learning aims to learn representations from the data itself without explicit manual supervision. It can be utilized to pretrain a model on a large amount of unlabeled data, allowing it to learn general features and representations from the data. These learned representations can then be transferred and fine-tuned for a specific task, effectively leveraging the knowledge gained from the self-supervised pretraining to improve performance on the target task. Pretraining models on these tasks with unlabeled data and using the pretrained weights to fine-tune models on supervised tasks with limited annotations help improve model performance and reach faster convergence [7, 16, 34]. The first step in self-supervised learning (pretraining on unlabeled data) is called the pretext. The second step (fine-tuning the pretrained models on annotated data) is called downstream.

The main goal of this research is to use Generative Adversarial Networks (GANs) in a self-supervised learning framework (Fig. 1) to pretrain models on large unlabeled EM datasets and use the weights to fine-tune DL models for various supervised downstream tasks such as semantic segmentation of nanoparticles, denoising, super-resolution, and noise & background removal in high-resolution TEM images. We show that such pretraining generalizes well and results in faster convergence and improved performance for different kinds of supervised tasks in EM with limited annotated data. Additionally, pretraining alleviates the need for training complex network architectures and expensive hyperparameter optimization. As a benchmark, results are compared with the work of Sytwu et al., which investigates the impact of receptive field size on the performance of DL models for semantic segmentation of nanoparticles in TEM images. Our results show that pretraining on unlabeled data leads to an improved performance regardless of the receptive field size or network architecture. More specifically, with fine-tuning, simple and smaller models achieve at least similar, often even better performance compared to larger, more complex models with randomly initialized weights.

This work makes the following scientific contributions: (i) demonstrating the substantial performance and convergence improvements in EM tasks through self-supervised (GAN-based) pretraining on unlabeled images; (ii) highlighting the generalization benefits and reduced dependency on hyperparameter optimization across different network architectures and receptive field sizes; (iii) a versatile framework for fine-tuning DL models on various EM tasks is introduced, including semantic segmentation, denoising, noise and background removal, and super-resolution.

2. Related Work

A commonly used approach in deep learning is pretraining models on large labeled datasets and fine-tuning on smaller datasets with limited annotations [29, 36, 54]. Such an approach usually performs well when the source data for pretraining is from a domain similar to the one from which the target data was obtained. In domains where large labeled datasets are scarce but an abundance of unlabeled datasets is available, self-supervised learning proves to be effective. Self-supervised learning leverages unlabeled datasets for pretraining, and the learned knowledge is then transferred to supervised downstream tasks with labeled data. Examples of successful self-supervised learning methods include contrastive learning, jigsaw puzzles, autoencoders, masked image modeling, and generative-based methods. SimCLR [7] is a prominent self-supervised learning method that has demonstrated significant advancements in self-supervised learning on large-scale benchmarks such as ImageNet. It leverages a contrastive pretraining objective, which involves maximizing agreement between differently augmented views of the same data point while minimizing agreement with views from other data points. This approach has been shown to learn semantically meaningful representations from unlabeled data, making it a powerful method for self-supervised learning. Momentum Contrast (MoCo) [16] is another well-known technique that leverages contrastive learning for unsupervised visual representation learning. It has been widely applied in various domains, including remote sensing scene classification [2], chest X-ray model pretraining [41], hand shape estimation [56], and speaker embedding [10]. The method has also been compared with other self-supervised learning techniques, demonstrating its effectiveness in learning representations from images and its potential for various downstream tasks [55]. Conrad and Narayan used this method to pretrain DL models on cellular EM images.

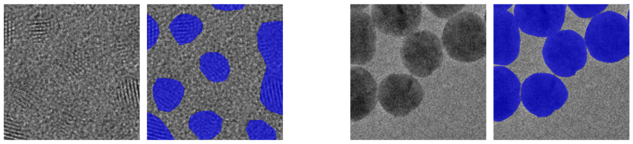

Figure 2. High- (left) and low-resolution (right) TEM image dataset of 2.2 nm and 20 nm Au nanoparticles and their ground truth segmentations. The more ordered structures are the nanoparticles, and the noisy regions are the amorphous matrix.

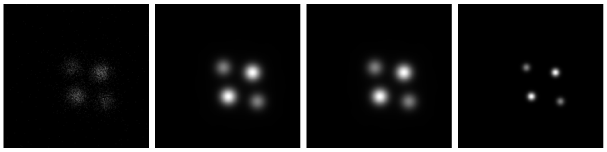

Figure 3. The example TEMImageNet image and corresponding ground truth labels. Left to right: original image, noise reduction, denoising & background removal, and super-resolution

Masked image modeling, another self-supervised learning method, involves training a model to predict the original content of an image from a corrupted or masked version [51]. This approach has been applied in various domains, including medical imaging and spectroscopic data identification. Li et al. proposed RGMIM (Region-Guided Masked Image Modeling) for COVID-19 detection, showcasing the potential of masked image modeling in medical imaging applications. Furthermore, Xue et al. highlighted the success of masked image modeling in self-supervised learning, demonstrating its ability to alleviate data-hungry issues and achieve competitive results. Caron et al. show the prominence of self-supervised learning on various tasks using vision transforms. These examples underscore the significance of masked image modeling in learning robust representations and its applicability across diverse domains.

Using GANs for self-supervised pretraining is also very effective. Chen et al. uses a GAN-based model to pretrain a model that learns image rotation. Guo et al. uses GAN-based pretraining for learning image similarity in remote sensing images. Other notable research in this area includes latent transformation detection [33], GAN-based image colorization for self-supervised visual feature learning [47], and self-supervised learning for semantic segmentation of archaeological monuments [20].

Recently, pretraining methodologies have been explored in the domain of EM. In particular, the microstructure segmentation with DL encoders pretrained on a large microscopy dataset [44], classification of scanning electron microscope images of pharmaceutical excipients using deep convolutional neural networks with transfer learning [19] and the unsupervised pretraining, the Momentum Contrast (MoCoV2) algorithm [8] was used by Conrad and Narayan.

In this paper, we explore the application of self-supervised pretraining based on GANs, specifically the Pix2Pix architecture [18], for EM images. The GAN model is pretrained on a large unlabeled Cellular Electron Microscopy (CEM) dataset called CEM500K [9]. The pretrained generator model can be fine-tuned on a wide range of downstream tasks in EM, including semantic segmentation of nanoparticles in TEMs, denoising, noise & background removal, and super-resolution. We show that such a pretraining approach leads to faster convergence and higher predictive power on all of the mentioned tasks with limited annotated datasets. We also find that fine-tuning with pretrained weights helps smaller and architecturally simpler models achieve similar or even higher scores compared to training with random weight initialization.

3. Materials and Methods

3.1. Datsets

3.1.1 CEM500K

CEM500K [9] is a large-scale, heterogeneous, unlabeled cellular EM image dataset developed for DL applications. The dataset is curated from experiments and various publicly available sources, encompassing 2D and 3D cellular EM images with diverse imaging modalities, sample preparation protocols, resolutions, and cell types. It includes examples from reconstructed FIB-SEM volumes, transmission EM (TEM) images, and EM image volumes and 2D images from various sources.

3.1.2 HRTEM Au Dataset

The Gold nanoparticle image dataset consists of high- and low-resolution TEM images of Gold (Au) nanoparticles with varying sizes (2.2 nm, 5 nm, 10 nm, and 20 nm) and different surface ligands, i.e., citrate (for 2.2 nm) and tannic acid. The images were acquired using an aberration-corrected TEAM 0.5 TEM for 2.2, 5, and 10 nm nanoparticles, while low-resolution images of 20 nm nanoparticles were obtained with a non-aberration-corrected TitanX TEM. The nanoparticles have a different crystalline structure than the embedding matrix (which is amorphous). This is the reason why the atomic arrangements look different in TEM. An important task in materials science is to segment such nanoparticles. The dataset was manually segmented and labeled, followed by preprocessing steps such as outlier removal and image standardization. To optimize memory usage during training, images were divided into 512 × 512-pixel patches, excluding patches consisting solely of amorphous background to address potential class imbalance issues [46]. In this paper, we will refer to the datasets as

Table 1. Number of images for training validation and testing of the datasets, used in the experiments.

DATASET	TRAIN	VALIDATION	TEST
CEM500K	50K/100K/200K	5000	5000
AU10NM	660	220	220
AU5NMV1	144	48	48
AU5NM	1044	348	348
AU20NM	660	220	220
AU2.2NM	1740	580	580
TEMIMAGENET	10377	1832	2155

"Au2.2nm", "Au5nm", "Au10nm","Au20nm". Additionally, the dataset of 5 nm Au nanoparticles [14] was included in our experiments and is referred to as "Au5nmV1". Examples of low and high-resolution TEM images of Gold nanoparticles and the corresponding ground truth segmentation annotations are shown in Figure 2.

3.1.3 TemImageNET

TEMImageNet is an open-source atomic-scale scanning transmission electron microscopy (ADF-STEM) image dataset. The dataset includes ten types of ground truth labels for training and validating DL models for tasks such as segmentation, super-resolution, background subtraction, denoising, and localization. The dataset comprises simulated ADF-STEM images of eight materials projected along multiple orientations with diverse atomic structures and crystallographic orientations. To replicate real-world experimental conditions, the images are augmented with realistic scan and Poisson noise, along with randomized linear and nonlinear low-frequency background patterns [26]. The example simulated image and the ground truth labels are shown in Figure 3. The number of images in each dataset for training, validation and testing is given in Table 1.

3.2. Pretraining Method

We employ a GAN-based approach for pretraining on unlabeled data. GANs are originally designed for generating new data samples that resemble a given dataset. The GAN architecture involves two neural networks: a generator and a discriminator. These are trained simultaneously through adversarial training. The generator takes random noise as input and generates synthetic data samples that are indistinguishable from real data. The discriminator evaluates the real and generated data and distinguishes them from each other. Both networks are simultaneously trained in an adversarial fashion: the generator tries to improve its ability to generate realistic data to fool the discriminator. The discriminator, in turn, strives to become better at distinguishing between real and generated samples [12].

Conditional Generative Adversarial Networks (cGANs) are an extension of the traditional GAN framework, where the generator is conditioned on additional information, typically in the form of class labels or other auxiliary data. The key idea is to guide the generation process based on specific conditions, allowing a more controlled and targeted generation of samples. In a cGAN, both the generator and the discriminator receive additional input information (conditioning) alongside the random noise for the generator and real/fake labels for the discriminator. The conditioning information could be anything relevant to the desired output, e.g., class labels, attributes, or other types of data. Conditional GANs have been used for various tasks, including image-to-image translation, image synthesis with specific attributes, and generating samples from certain classes. They provide a way to control and manipulate the characteristics of the generated data by incorporating additional information during the training process [30]. The same loss function for cGANs as mentioned in [18] is used.

In this paper, we use the cGAN model called Pix2Pix [18] for pretraining on the unlabeled CEM500K dataset of EM images. The images are fed to the generator with added noise, and the goal is to generate output images that are indistinguishable from the original ones. The trained generator model can then be fine-tuned on supervised downstream tasks explained in Section 3.3. To investigate the influence of the size of unlabeled datasets used for pretraining on the results of fine-tuning in the downstream tasks, we run pretraining experiments with different numbers of examples from the CEM500K dataset: 50K, 100K, and 200K examples.

3.3. Downstream Tasks

The pretrained generator model explained in the previous section can be fine-tuned on a wide range of downstream tasks. As case studies in this paper, we have selected tasks including semantic segmentation of nanoparticles, denoising, noise & background removal, and super-resolution in high-resolution TEM images.

Semantic segmentation involves labeling each pixel of an image with a corresponding class of what is being represented. In materials and biological sciences, this imaging task plays a crucial role in analyzing and understanding complex microstructural and elemental features within the images obtained. Sytwu et al. conducted experiments on semantic segmentation of Gold nanoparticles of different sizes and resolutions using the U-Net model [37]. They used a U-Net with different numbers of residual blocks and different receptive field sizes to study the influence of model complexity and receptive field size on the performance of the model. Their findings show that increasing the receptive field increases model performance in high-resolution TEM images. They also conclude that as the model complexity increases, there is a corresponding improvement in performance and prediction confidence. Here, the focus is on an investigation of the influence of pretraining on the performance of DL models in semantic segmentation. Therefore, we selected the same network as that used by Sytwu et al., i.e., the U-Net, and used it as the generator for pretraining on unlabeled data in the previous step. We then fine-tuned it for semantic segmentation on the same dataset, i.e., high-resolution TEM images of Gold nanoparticles. The results were compared to training the same model with randomly initialized weights. These experiments were conducted with different complexities and receptive field sizes. We additionally used a more complex High-Resolution Network (HRNet) [45] as the generator and fine-tuned it on this dataset for comparison.

Furthermore, atom segmentation, localization, noise reduction, and deblurring are crucial tasks in atomic-resolution scanning transmission electron microscopy (STEM). The images captured at the atomic scale often suffer from noise, which can obscure subtle details and compromise the accuracy of atom segmentation and localization [26]. Denoising is important in refining these images, ensuring a clearer representation of the atomic structure. It enhances the level of detail in the images beyond the inherent resolution of the microscope, achieving higher spatial resolution, which allows researchers to discern finer structural features and better characterize atomic arrangements. By reducing noise and enhancing resolution, the method ensures a more accurate and robust analysis of atomic structures, even in challenging imaging conditions with variations in sample thickness. To show the generalizability of pretrained models in our work, we use the TEMImageNet data and do experiments on denoising, noise & background removal, and super-resolution. We use the HRNet model and run the same experiments to compare the results of fine-tuning to those of training with random weight initialization. Details of experiments and results are outlined in Section 4.

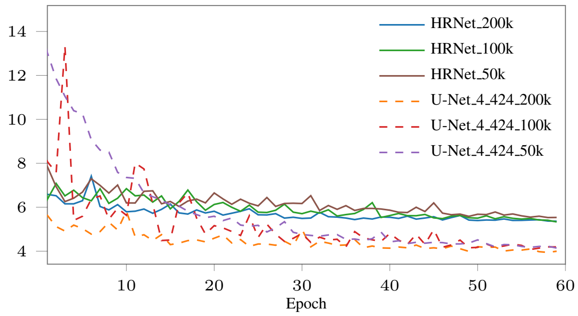

Figure 4. Validation L_1 loss for HRNet and U-Net_4_424 for three different dataset sizes.

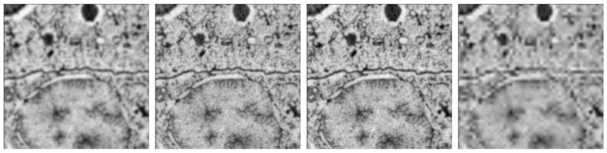

Figure 5. Left to write: input image with added noise, original image, and generated images by U-Net and HRNet.

4. Experiments and Results

4.1. Experiments on CEM500K and GANs

The CEM500K dataset was employed in the pretraining step where cGANs, based on the Pix2Pix model [18], were used. As generator, we used different U-Net architectures with varying numbers of residual blocks and receptive field sizes. Specifically, configurations with 2, 3, and 4 residual blocks were considered, each associated with different receptive field sizes. For two residual blocks, the receptive field sizes were 44, 84, and 116, for three blocks, we used 96, 176, and 240 as well as a receptive field of 200, 360, and 424 for four blocks, in line with the work presented in [46]. For brevity, we refer to these U-Net variations as U-Net_B_RF from now on, where B refers to the number of residual blocks and RF refers to receptive field size. Additionally, a more complex model architecture, HRNet [50], was used to explore its effectiveness in comparison to U-Net variants and, in particular, to understand how far an increase in model complexity results in an increase in expressivity.

For training, random Gaussian noise, blurring, flipping, and rotations were applied to the images used as input to the generator. The generator was trained to predict images indistinguishable from the original "clean" images. The training for each model variation was conducted for 60 epochs with a batch size of 128. Adam optimizer with a learning rate of 2×10^{-4} was used for optimization. In terms of training the whole GAN architecture, it was framed as the Least Square GAN (LSGAN), which adopts the least squares loss function for the discriminator and is more stable than regular GANs. LSGANs are able to generate higher quality images and perform more stably during the learning process [28]. The generator was also trained using the L_1 loss, and the λ in Equation 3 was set to 100.

Each model variation was trained with the same experimental setup for three different subsets of the CEM500K dataset consisting of 50 K, 100 K, and 200 K images. This experiment was conducted to investigate the influence of the dataset size in pretraining on the model performance during fine-tuning on downstream tasks.

The validation plots for the GAN pretraining with HRNet and the most complex U-Net variation are shown in Figure 4. For each of the models, as the dataset size increases, the validation loss decreases. Interestingly, the U-

Net model shows better validation results compared to HRNet. This is also illustrated in the generated images in Figure 5. We believe this is due to the skip connections from the initial residual blocks in the U-Net model to the corresponding upsampling blocks. The U-Net model has direct access to the information in the larger spatial resolutions and is more prone to memorizing, while the HRNet model actually learns the feature representations as it encodes the images into lower spatial resolutions with higher feature maps and then decodes it, without having direct access to the input features in the higher spatial resolution. Additionally, we find that the pretrained HRNet model fine-tuned on a variety of supervised tasks outperforms the U-Net model. For brevity, results for other experiments are included in the supplementary materials.

4.2. Experiments on Semantic segmentation with Au datasets

The segmentation of nanoparticles in the Au datasets was approached using the same variations of the U-Net and the HRNet models that were pretrained on the unlabeled CEM500K dataset. All model variations were trained and systematically compared with respect to weight initialization with random weights and with the pretrained weights from the previous step. All experiments were performed by training for 60 epochs on each dataset. As an objective function, the Binary Cross Entropy (BCE) loss was used and minimized by the Adam optimizer with a learning rate of 2×10^{-4}. The model's performance was evaluated using the dice score. The original image size is 512×512 pixels, but data augmentation techniques, including various types of noise, flips, rotations, resizing, and random cropping to 448×448 pixels, are applied.

In the experiments conducted with different variations of the U-Net, we observe that randomly initialized models with a higher receptive field size perform better than those with a smaller receptive size. However, the fine-tuned models with smaller receptive field sizes initially not only outperform the same models with randomly initialized weights, but even outperform those with bigger receptive fields. Even though the randomly initialized model catches up as training progresses longer, the performance is not stable and the oscillations are large. This is illustrated in the validation plots for the Au5nmV1 dataset in Figure 6. We also observe that the fine-tuned U-Net model (U-Net_2_44_P(100K)) outperforms the randomly initialized models, as illustrated in Figure 7. Suffixes P(50K), P(100K) and P(200K) indicate models that were pretrained on 50 K, 100 K, and 200 K unlabeled images, respectively, in the first step.

In terms of model complexity, the observation is still consistent. As illustrated in Figure 8, a randomly initialized U-Net_4_424 performs worse than the three different combinations of fine-tuned models.

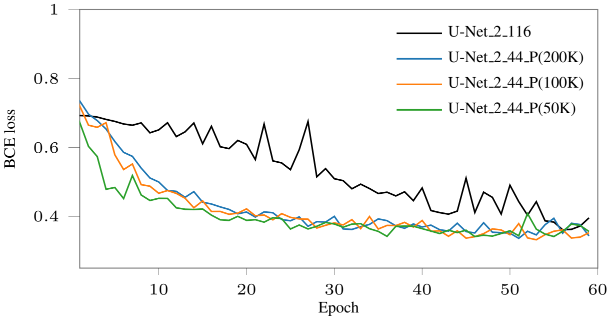

Figure 6. Comparison of validation loss for bigger randomly initialized U-Net and smaller fine-tuned U-Nets.

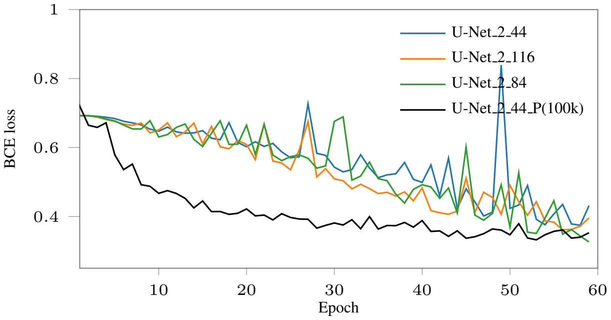

Figure 7. Comparison of validation loss for smaller fine-tuned U-Net and randomly initialized U-Nets of varying complexity.

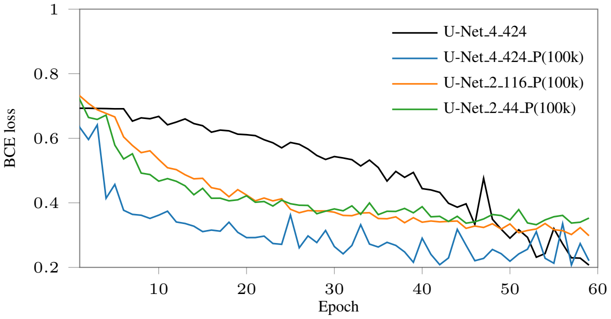

Figure 8. Validation loss for bigger randomly initialized U-Net and fine-tuned U-Nets of varying size.

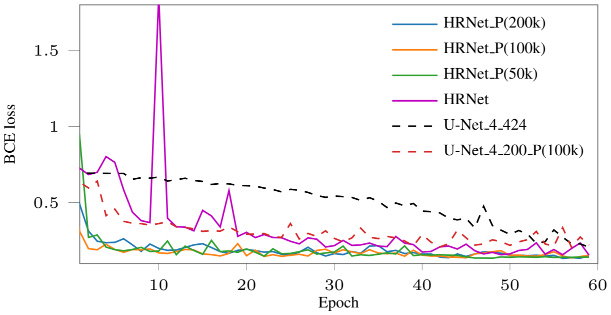

Figure 9. Validation loss for HRNet models compared to U-Net.

Conducting the same experiments with the HRNet model, we find that a fine-tuned U-Net model still outperforms the randomly initialized HRNet model, but all three fine-tuned HRNet models not only outperform the randomly initialized HRNet model but also score higher than the randomly initialized U-Net, as well as all fine-tuned U-Net models. The validation plots are shown in Figure 9. Example predictions are illustrated in Figure 10 for the fine-tuned HRNet and fine-tuned U-Net_4_0, pretrained on 100 K CEM500K images. Moreover, the dice scores on test set are calculated for all models after 5, 30 and 60 epochs. As shown in Table 2, the pretrained models converge faster than randomly initialized models and generally score higher, specially in the beginning epochs. The general trend of the results and observations for all other datasets is consistent with the above-reported ones.

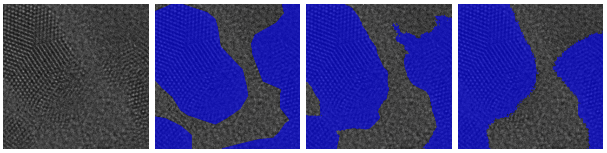

Figure 10. Left to write: original image, ground truth, prediction by fine-tuned HRNet and U-Net_4_0, respectively, both pretrained on 100 K CEM500K images and fine-tuned.

4.3. Experiments on TEMImageNet dataset

As the HRNet model performed better than all U-Net variations regardless of the fine-tuning, we conducted the experiments on the TEMImageNet dataset only with HRNet. On this dataset, we decided to exclusively employ HRNet for denoising, noise & background removal, and super-resolution tasks. For each of these tasks, the HRNet model was trained with random initialization. The same experiments were conducted by fine-tuning the HRNet model pretrained on three separate subsets of the unlabeled CEM500K datasets in the previous step. During the experiments, an image size of 256x256 pixels and a batch size of 64 were used, and the training process continued for 60 epochs. The objective function comprised both L_1 and L_2 terms, and optimization was carried out using the Adam optimizer with a learning rate of 2×10^{-4}. The chosen data augmentation techniques, including noise variations, flipping, rotations, and random resizing, as usual had the goal of enhancing the model's ability to generalize and learn robust features.

The plots for validation loss in all three tasks are shown in Figure 11. We observe that in all three cases, the validation losses for the fine-tuned models are lower than those for the randomly initialized models. During the initial training phase, the validation loss for the model fine-tuned with smaller datasets is larger than that for the models trained with more data. However, already after approximately 10 epochs, this difference vanishes. Additionally, the validation L_1 loss of all randomly initialized models exhibits severe fluctuations, while the fine-tuned models behave much more robustly. Some example predictions for denoising, noise & background removal, and super-resolution are shown in Figures 12, 13, and 14, respectively. Even though all predictions are visually very good and it is difficult to observe any differences, the quantitative results as confirmed by the plots in Figure 11 show that pretraining leads to a better and more stable performance. Similarly, the L_1 metric on test set are calculated for all models after 5, 30 and 60 epochs. As shown in Table 3, the pretrained models converge faster than randomly intialized models and generally score better, specially in the beginning epochs.

Table 2. Comparison of segmentation Dice scores for different training methods (randomly initialized weights (R) and pretrained (P) with GANs on CEM500K using 50 K, 100 K, 200 K images) on the Au5nmV1 test dataset. The experiments were performed with UNets of different sizes (with two and four residual blocks) and receptive fields (two for each U-Net size) and HRNet.

Epochs			5	30	60
HRNet		R	41.05	86.55	90.24
		P(50k)	86.37	91	91.97
		P(100k)	88.49	92.14	92.38
		P(200k)	83.86	91.41	92.07
U-Net 4 blocks	424	R	7.8	86.03	89.95
		P(50k)	72.78	89.81	91.16
		P(100k)	71.8	89	89.72
		P(200k)	75.46	87.34	90.21
	200	R	0.25	80.69	83.61
		P(50k)	82.03	87.05	85.68
		P(100k)	83	86.7	88.53
		P(200k)	55.75	86.56	89
U-Net 2 blocks	176	R	0	80.81	82.49
		P(50k)	61.99	82.3	84.72
		P(100k)	51.97	80.16	83.47
		P(200k)	67.87	80.48	83.18
	44	R	0	74.87	79.75
		P(50k)	65.89	76.9	78.59
		P(100k)	58.62	75.67	79.92
		P(200k)	53.15	75.45	76.99

5. Conclusion

In this paper, we explored the impact of pretraining on various computer vision tasks. Through self-supervised training, GANs, and a pretraining strategy involving unlabeled data followed by fine-tuning on labeled data, our investigation showcased significant advancements in the capabilities of computer vision models, specifically in the context of EM. Self-supervised training enabled the models to extract representations from unlabeled EM data, addressing

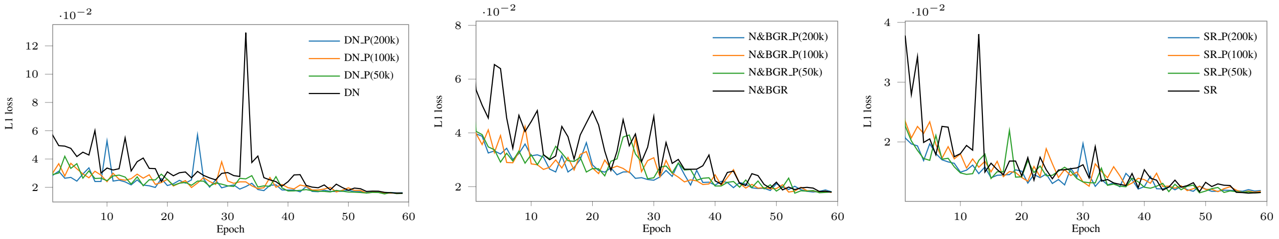

Figure 11. Validation L_1 loss for the downstream tasks on TEMImageNet dataset. Left to right: denoising (DN), noise & background removal (N&BGR), and super-resolution (SR)

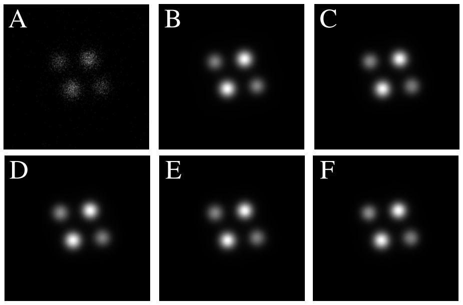

Figure 12. Results for denoising on TEMImageNet. Input (A), ground truth (B), prediction by randomly initialized model (C), and predictions (D, E, F) by fine-tuned models pretrained on 50K, 100K, and 200K images.

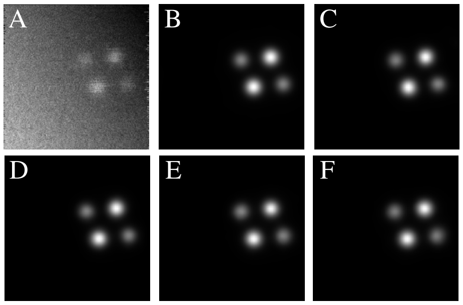

Figure 13. Results for noise & background removal on TEMImageNet. (A)-(F) in analogy to Figure 12.

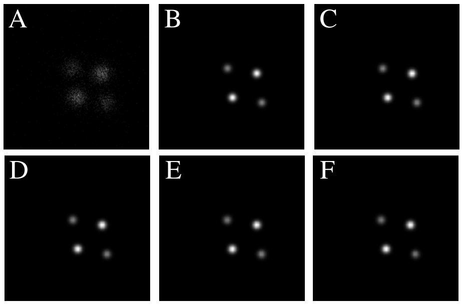

Figure 14. Results for super-resolution on TEMImageNet. (A)-(F) in analogy to Figure 12.

Table 3. Comparison of the L_1 metric for different training methods (randomly initialized weights (R) and pretrained (P) with GANs on CEM500K using $50\,\mathrm{K}$, $100\,\mathrm{K}$, $200\,\mathrm{K}$ images) on each of the downstream tasks: Super-resolution (SR), Noise & Background Removal (N&BGR) and Denoising (DN). The experiments were performed with HRNet.

Epochs		5	30	60
SR	R	0.01972	0.01524	0.01122
	P(50k)	0.01688	0.01359	0.0112
	P(100k)	0.02105	0.01465	0.01142
	P(200k)	0.01639	0.01371	0.01156
N&BGR	R	0.06391	0.03438	0.0176
	P(50k)	0.03332	0.02257	0.01776
	P(100k)	0.03268	0.02913	0.01766
	P(200k)	0.03323	0.0224	0.01792
DN	R	0.04732	0.02996	0.01541
	P(50k)	0.03287	0.02169	0.01542
	P(100k)	0.03693	0.03708	0.01555
	P(200k)	0.02657	0.01907	0.01586

challenges associated with the scarcity and labor intensity of labeled datasets in this domain. Furthermore, the integration of GANs in generative pretraining proved beneficial for improving model generalization.

Pretraining on unlabeled data, followed by fine-tuning on labeled data, enhanced performance and accelerated convergence in several downstream tasks, including segmentation, denoising, and super-resolution. An important outcome of our work for such tasks is that for obtaining a higher predictive accuracy, the model complexity might not be the only or most important factor. The CEM500K dataset, containing SEM images, was used for pretraining and improved performance in TEM image-based downstream tasks, despite the differences between SEM and TEM images. Future research could explore pretraining on TEM images for closer domain relevance. Additionally, while self-supervised learning with GANs enhanced the performance in this study, their training complexities and risk of mode collapse suggest exploring alternative self-supervised methods like contrastive pretraining for potentially better outcomes in the future.

References

[1] J. A. Aguiar, M. L. Gong, R. R. Unocic, T. Taşdizen, and B. Miller. Decoding crystallography from high-resolution electron imaging and diffraction datasets with deep learning. *Science Advances*, 5, 2019. 1

[2] N. Alosaimi, H. Alhichri, Y. Bazi, B. B. Youssef, and N. Alajlan. Self-supervised learning for remote sensing scene classification under the few shot scenario. *Scientific Reports*, 13, 2023. 2

[3] C. Angermueller, T. Pärnamaa, and L. Parts. Deep learning for computational biology. *Molecular Systems Biology*, 12, 2016. 1

[4] Anusha Aswath, Ahmad Alsahaf, Ben NG Giepmans, and George Azzopardi. Segmentation in large-scale cellular electron microscopy with deep learning: A literature survey. *Medical image analysis*, page 102920, 2023. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 3

[6] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised GANs via auxiliary rotation loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12154–12163, 2019. 3

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[8] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *ArXiv*, abs/2003.04297, 2020. 3

[9] Ryan Conrad and Kedar Narayan. Cem500k, a large-scale heterogeneous unlabeled cellular electron microscopy image dataset for deep learning. *eLife*, 10, 2021. 1, 3

[10] Ke Ding, Xuanji He, and Guanglu Wan. Learning speaker embedding with momentum contrast. *ArXiv*, abs/2001.01986, 2020. 2

[11] A. A. Ezzat and M. Bedewy. Machine learning for revealing spatial dependence among nanoparticles: understanding catalyst film dewetting via gibbs point process models. *The Journal of Physical Chemistry C*, 124:27479–27494, 2020. 1

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 4

[13] Kishan Govind, Daniela Oliveros, Antonin Dlouhy, Marc Legros, and Stefan Sandfeld. Deep learning of crystalline defects from TEM images: a solution for the problem of 'never enough training data'. *Machine Learning: Science and Technology*, 5(1):015006, 2024. 1

[14] Catherine K. Groschner, Christina Choi, and Mary C. Scott. Machine learning pipeline for segmentation and defect identification from high-resolution transmission electron microscopy data. *Microscopy and Microanalysis*, 27 (3):549–556, 2021. 2, 4

[15] Dongen Guo, Ying Xia, and Xiaobo Luo. Self-supervised GANs with similarity loss for remote sensing image scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2508–2521, 2021. 3

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[17] James P. Horwath, Dmitri N. Zakharov, Rémi Mégret, and Eric A. Stach. Understanding important features of deep learning models for segmentation of high-resolution transmission electron microscopy images. *npj Computational Materials*, 6(1), 2020. 2

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 3, 4, 5

[19] Hiroaki Iwata, Yoshihiro Hayashi, Aki Hasegawa, Kei Terayama, and Yasushi Okuno. Classification of scanning electron microscope images of pharmaceutical excipients using deep convolutional neural networks with transfer learning. *International Journal of Pharmaceutics: X*, 4:100135, 2022. 3

[20] Bashir Kazimi and Monika Sester. Self-supervised learning for semantic segmentation of archaeological monuments in DTMs. *Journal of computer applications in archaeology 6 (2023), Nr. 1*, 6(1):155–173, 2023. 3

[21] C. Ke. Applications of scanning electron microscopy in biology. *International Review of Cytology*, pages 183–255, 1971. 1

[22] Afshin Khadangi, Thomas Boudier, and Vijay Rajagopal. EM-net: Deep learning for electron microscopy image segmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021. 2

[23] B. Lee, S. Yoon, J. W. Lee, Y. Kim, J. Chang, J. Yun, J. C. Ro, J. Lee, and J. H. Lee. Statistical characterization of the morphologies of nanoparticles through machine learning based electron microscopy image analysis. *ACS Nano*, 14:17125–17133, 2020. 1

[24] Yang Leng. *Materials characterization: introduction to microscopic and spectroscopic methods*. John Wiley & Sons, 2013. 1

[25] Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. RGMIM: Region-guided masked image modeling for COVID-19 detection. *arXiv e-prints*, pages arXiv–2211, 2022. 3

[26] Ruoqian Lin, Rui Zhang, Chunyang Wang, Xiao-Qing Yang, and Huolin L. Xin. TEMImageNet training library and atomsegnet deep-learning models for high-precision atom segmentation, localization, denoising, and deblurring of atomic-resolution images. *Scientific Reports*, 11(1), 2021. 1, 4, 5

[27] O. Lupan, V. Creţu, M. Deng, D. Gedamu, I. Paulowicz, S. Kaps, Y. K. Mishra, O. Polonskyi, C. Zamponi, L. Kienle,

V. Trofim, I. M. Tiginyanu, and R. Adelung. Versatile growth of freestanding orthorhombic α-molybdenum trioxide nano- and microstructures by rapid thermal processing for gas nanosensors. *The Journal of Physical Chemistry C*, 118:15068–15078, 2014. 1

[28] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 5

[29] Dimitrios Marmanis, Mihai Datcu, Thomas Esch, and Uwe Stilla. Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1):105–109, 2015. 2

[30] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 4

[31] Sang Ho Oh, Marc Legros, Daniel Kiener, and Gerhard Dehm. In situ observation of dislocation nucleation and escape in a submicrometre aluminium single crystal. *Nature materials*, 8(2):95–100, 2009. 1

[32] O. Panova, X. C. Chen, K. C. Bustillo, C. Ophus, M. P. Bhatt, N. P. Balsara, and A. M. Minor. Orientation mapping of semicrystalline polymers using scanning electron nanobeam diffraction. *Micron*, 88:30–36, 2016. 1

[33] Parth Patel, Nupur Kumari, Mayank Singh, and Balaji Krishnamurthy. LT-GAN: Self-supervised GAN with latent transformation detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3189–3198, 2021. 3

[34] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[35] R. Paul, S. Hawkins, Y. Balagurunathan, M. B. Schabath, R. J. Gillies, L. O. Hall, and D. B. Goldgof. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. *Tomography*, 2:388–395, 2016. 2

[36] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. ImageNet-21K pretraining for the masses. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 2

[37] Olaf Ronneberger, Philipp Fischer, and Thomas 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*, pages 234–241. Springer, 2015. 4

[38] Karina Ruzaeva, Kishan Govind, Marc Legros, and Stefan Sandfeld. Instance segmentation of dislocations in TEM images. In *2023 IEEE 23rd International Conference on Nanotechnology (NANO)*, pages 1–6. IEEE, 2023. 1

[39] Robbie Sadre, Colin Ophus, Anastasiia Butko, and Gunther H Weber. Deep learning segmentation of complex features in atomic-resolution phase-contrast transmission electron microscopy images. *Microscopy and microanalysis*, 27 (4):804–814, 2021. 1

[40] D. Salas, A. L. Gall, J. Fiche, A. Valeri, Y. Ke, P. Bron, G. Bellot, and M. Nollmann. Angular reconstitution-based 3D reconstructions of nanomolecular structures from superresolution light-microscopy images. *Proceedings of the National Academy of Sciences*, 114:9273–9278, 2017. 1

[41] H. Sowrirajan, J. Yang, A. Y. Ng, and P. Rajpurkar. MoCo-CXR: MoCo pretraining improves representation and transferability of chest X-ray models. 2020. 2

[42] C. Srinivas, N. P. K. S., M. Zakariah, Y. A. Alotaibi, K. Shaukat, B. Partibane, and A. Halifa. Deep transfer learning approaches in performance analysis of brain tumor classification using MRI images. *Journal of Healthcare Engineering*, 2022:1–17, 2022. 2

[43] Dominik Steinberger, Inas Issa, Rachel Strobl, Peter J Imrich, Daniel Kiener, and Stefan Sandfeld. Data-mining of in-situ TEM experiments: Towards understanding nanoscale fracture. *Computational materials science*, 216:111830, 2023. 1

[44] Joshua Stuckner, Bryan Harder, and Timothy M. Smith. Microstructure segmentation with deep learning encoders pre-trained on a large microscopy dataset. *npj Computational Materials*, 8(1), 2022. 3

[45] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 5

[46] Katherine Sytwu, Catherine Groschner, and Mary C Scott. Understanding the influence of receptive field and network complexity in neural network-guided TEM image analysis. *Microscopy and Microanalysis*, 28(6):1896–1904, 2022. 1, 2, 3, 4, 5

[47] Sandra Treneska, Eftim Zdravevski, Ivan Miguel Pires, Petre Lameski, and Sonja Gievska. GAN-based image colorization for self-supervised visual feature learning. *Sensors*, 22(4): 1599, 2022. 3

[48] Juan Miguel Valverde, Vandad Imani, Ali Abdollahzadeh, Riccardo De Feo, Mithilesh Prakash, Robert Ciszek, and Jussi Tohka. Transfer learning in magnetic resonance brain imaging: A systematic review. *Journal of imaging*, 7(4):66, 2021. 2

[49] J. Vincent, R. Manzorro, S. Mohan, B. Tang, D. Y. Sheth, E. P. Simoncelli, D. S. Matteson, C. Fernandez-Granda, and P. A. Crozier. Developing and evaluating deep neural network-based denoising for nanoparticle TEM images with ultra-low signal-to-noise. *Microscopy and Microanalysis*, 27:1431–1447, 2021. 1

[50] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(10):3349–3364, 2021. 5

[51] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 3

[52] Hongwei Xue, Peng Gao, Hongyang Li, Yu Qiao, Hao Sun, Houqiang Li, and Jiebo Luo. Stare at what you see: Masked image modeling without reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22732–22741, 2023. 3

[53] Chen Zhang, Hengxu Song, Daniela Oliveros, Anna Fraczkiewicz, Marc Legros, and Stefan Sandfeld. Datamining of in-situ TEM experiments: On the dynamics of dislocations in cocrfemnni alloys. *Acta Materialia*, 241: 118394, 2022. 1

[54] Rui Zhang, Huimin Xie, Shuning Cai, Yong Hu, Guo-kun Liu, Wenjing Hong, and Zhong-qun Tian. Transfer-learning-based raman spectra identification. *Journal of Raman Spectroscopy*, 51(1):176–186, 2020. 2

[55] Y. Zhang, L. Po, X. Xu, M. Liu, W. Ou, Y. Zhao, and W. Y. Yu. Contrastive spatio-temporal pretext learning for self-supervised video representation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:3380–3389, 2022. 3

[56] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. In *DAGM German Conference on Pattern Recognition*, pages 250–264. Springer, 2021. 2