Contrastive Image-Metadata Pre-Training for Materials Transmission Electron Microscopy

Georgia Channing [*1 2] **Debora Keller** [3] **Marta D. Rossell** [3] **Philip Torr** [2] **Rolf Erni** [3] **Stig Helveg** [4] **Henrik Eliasson** [*4]

Abstract

The vast majority of transmission electron microscopy (TEM) data never gets published and ends up on a backup drive until deleted to free up space. These left-over datasets are rich in detail and variation, often paired with automatically saved metadata of instrument state and acquisition parameters. In this work, we introduce a dataset of 7,330 high-angle annular dark-field scanning-TEM (HAADF-STEM) images from a single instrument to learn a joint embedding space between image metadata and HAADF image. These embeddings link image style with acquisition parameters, which allows us to train a generative style transfer network that can convert experimental images into the style they would have had if they were recorded with different instrument parameters. We evaluate the performance of the network and explore the usefulness of the technique for physical denoising.

1. Introduction

The transmission electron microscope is the most accessible and versatile instrument for direct visualization of materials and biological samples on the nano and atomic scale. Yet for every published image, a massive volume of data, more than 90% by some estimates (Ivanova et al., 2025), remains unused. This is not because the quality of the data is bad, but because of the nature of TEM experiments, where exploration and tuning are necessary to identify and record the golden data that eventually makes publication. Over time, TEM labs typically find themselves with large amounts of leftover data from different users, of different materials samples, different operating modes, and acquisition settings. This data is rich in detail and variation, and often comes with automatically saved metadata that describe the state of the microscope at acquisition. Although the electron microscopy community is aware that these datasets could be very valuable for data-driven research, how, and for what applications, remains an open question.

Modern TEM is no longer limited by spatial resolution, which has been the main driver of development in the last century. Today, the main issue is instead dose efficiency, minimising the number of electrons that hit the sample while retaining an interpretable signal. Limiting the electron dose is crucial to avoid sample damage and ensure physical relevance of observed phenomena, but naturally leads to very noisy data (Egerton, 2025; Reimer, 1984). For dealing with noisy data, modern analysis workflows rely heavily on machine learning. However, machine learning models for materials science TEM are typically trained from scratch, specialised for a specific noise level and material sample. In comparison to cryo-TEM for biology where analysis workflows are well-established, the large heterogeneity in materials samples, the variation in length scales that are studied in materials, and the diversity of operating conditions and modalities used for materials, require more robust ML solutions that generalise across these parameters to be broadly applicable in practice.

For example, networks for real-time denoising and analysis would be a great help to microscope operators for decision making in low-signal scenarios. But as acquisition parameters that affect the noise profile and quantitative meaning of pixel intensities in recorded images are often changed during standard operation, such models must be robust to changes such as magnification and detector settings. One possible application of leftover datasets could be to learn data characteristics as a function of instrument parameters, to inform networks about the state of the microscope at acquisition. Implicitly learning the relationship between images and their paired metadata could then be the foundation of generalisable ML workflows for TEM by providing physically meaningful embeddings on which to condition

[1]Hugging Face, London, United Kingdom [2]University of Oxford [3]Swiss Federal Laboratories for Materials Science and Technology, Dübendorf, Switzerland [4]Department of Physics, Technical University of Denmark, Kgs. Lyngby, Denmark. Correspondence to: Henrik Eliasson <haoel@dtu.dk>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

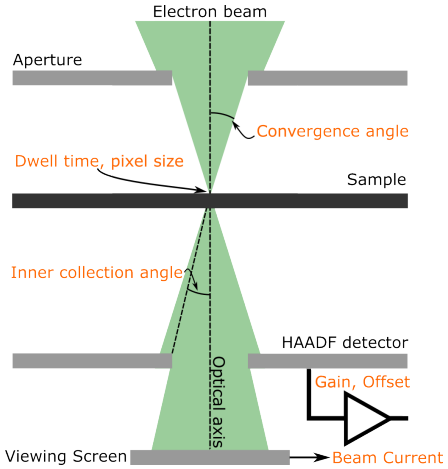


Figure 1. STEM diagram, selected metadata is shown in orange text.

downstream networks.

In this work, we operationalise the idea above and make four contributions. First, we release a paired image-metadata dataset of 7,330 HAADF-STEM images from the Titan microscope at Empa Dübendorf, each associated with seven acquisition-metadata parameters: pixel size, dwell time, beam convergence angle, beam current, detector gain, detector offset, and inner collection angle (Figure 1). Second, we propose Contrastive Image-Metadata Pre-training (CIMP), a CLIP-style method (Radford et al., 2021) that learns a joint embedding space over HAADF-STEM images and their acquisition metadata, and we treat the resulting embeddings as a foundation layer for TEM-ML workflows rather than as infrastructure for any single downstream task. Third, we evaluate CIMP on cross-modal retrieval and linear-probe recovery of the seven acquisition parameters from the visual embedding alone, and use its embeddings as both a perceptual loss and a conditioning signal for a generative metadata-conditioned style-transfer network that translates an experimental image into the style it would have had under different acquisition parameters. Finally, virtually scaling dwell time and beam current at generation time turns the style-transfer network into a physics-informed denoiser whose outputs are more grounded to the input image than popular self-supervised baselines.

2. Related Work

Open-Source STEM Data While petabytes of HAADF-STEM images are collected each year (Ivanova et al., 2025), very few images ever make it to the open source. This is partially because publications from microscopists usually focus on specific findings from the highest quality images, but also because the culture of data sharing is not well-established in the electron microscopy community. Nevertheless, some large datasets, both experimental and synthetic, have been made available to the community. On the experimental side, the DRYAD release of Sytwu et al. (2023) provides 407 HRTEM images of Au, Ag, and CdSe nanoparticles (2.2–20 nm) with acquisition metadata including magnification and electron dose, which is the closest publicly available precedent for an experimental image corpus for materials science with paired instrument settings, though at a much smaller scale and in HRTEM rather than HAADF-STEM. On the simulated side, Atomagined (Schwenker et al., 2020) offers atomic-resolution HAADF-STEM images generated with the PRISM imaging software and post-processed to emulate noise and distortion typical of the modality, covering a wide range of ICSD structure prototypes. Eliasson & Erni (2025) release a large-scale (>100.000 images) multislice-simulated HAADF-STEM corpus (Eliasson, 2025) covering Pt nanoparticles of varying size and shape designed for nanoparticle size and morphology prediction, generated by a surrogate model for the multislice algorithm. TEMImageNet (Lin et al., 2021) provides a simulated training library for atom-resolution STEM analysis (segmentation, localisation, denoising, deblurring) and shows that models trained purely on simulated data can transfer to experimental images. None of these datasets, however, provide the scale of pairing between experimental HAADF-STEM images and rich instrument metadata that we target in this work. The few experimental datasets that are publicly available tend to be small, narrow in material coverage, and rarely include the acquisition metadata needed to train models that generalise across instrument settings.

Embedding Models CIMP adapts the CLIP image–text pretraining framework (Radford et al., 2021) to pairs of HAADF-STEM images and their acquisition metadata. The closest structural analogue in any microscopy domain is CLOOME (Sanchez-Fernandez et al., 2023), which contrastively aligns fluorescence cell-painting images with chemical-structure embeddings and demonstrates cross-modal retrieval across a large bioimaging database. ContIG (Taleb et al., 2022) is the closest conceptual parallel: a medical-imaging model that contrastively aligns retinal images with a structured physical-annotation vector (SNP arrays, polygenic risk scores). AstroCLIP (Parker et al., 2024) provides the cleanest methodological template with separate per-modality encoders, pretrained and then aligned via a symmetric contrastive loss across galaxy images and spectra. To our knowledge, however, no prior work applies this or any similar framework to electron microscopy.

Image-to-Image Translation in Electron Microscopy Prior work has applied image-to-image networks to EM data in two main directions: bridging the simulation-to-experiment gap so that synthetic training data gains realism, and denoising experimental acquisitions. On the simulation-

to-experiment side, Khan et al. (2023) employ a CycleGAN with a reciprocal-space discriminator to add realistic high-frequency content to simulated STEM images, then use the augmented data to train a downstream segmentation FCN that transfers to real experiments. Eliasson et al. (2024) use a CycleGAN to translate between simulated and experimental HAADF-STEM of Pt nanoparticles, showcasing both denoising and realistic noise addition. Wang et al. (2020) train an unsupervised GAN-based STEM denoiser that denoises noisy acquisitions of sub-nm catalytic clusters. And in another work, Lobato et al. (2024) train a GAN denoiser on synthetic image pairs with intricate modeled noise. These approaches are all one-to-one and agnostic to microscope settings (i.e., a dedicated network is trained per source/target domain pair, with no natural way to parameterise the target style). Simulating a different dwell time or detector gain with any previous methods requires training a new CycleGAN with a new paired or unpaired corpus at that setting. The style-transfer network we present downstream of CIMP is instead one-to-many: a single generator is conditioned on continuous acquisition metadata through CIMP embeddings such that any target instrument setting can be queried at inference without retraining.

STEM Denoisers Stability improvements of the instrument and direct electron cameras and detectors that approach quantum efficiencies of 1 allow for ever smaller electron doses to be used reliably (Gatan, Inc., 2023; Thermo Fisher Scientific, 2024; DECTRIS Ltd., 2024). At the same time, developments in denoising models and noise-robust analysis models enable work at higher noise levels. Self-supervised denoising frameworks based on Noise2Noise (Lehtinen et al., 2018), and Noise2Void (Krull et al., 2018) have recently become popular for transmission electron microscopy, and produce impressive results (Sheth et al., 2021; Kim et al., 2025; Thornley et al., 2026). However, these models are typically heavily specialised for single noise profiles and materials samples, with Noise2Noise and Unsupervised Deep Video Denoising (UDVD) also requiring time-series data where noise changes but image content does not. Noise2Void-based techniques that can work on single images instead introduce checkerboard artefacts when trained to convergence, and may also lose image detail if pixel size is not significantly smaller than the atomic features we are interested in. GAN-based denoisers (discussed in the image-to-image paragraph above) have also been explored but have not been adopted to the same extent. No STEM denoisers are to our knowledge general in the sense that they are informed about the acquisition settings and pixel size related to the images, and can produce good results across instrument settings.

3. Data Contribution

In this work, we introduce 7,330 HAADF-STEM images with paired metadata from the Titan microscope at Empa Dübendorf. The dataset covers more than 30 distinct materials systems, predominantly heterogeneous catalysts (Pt/CeO_2, Pd/ZrO_2, Ru/TiO_2, In_2O_3/ZrO_2, etc.), with more than 130 material samples synthesized by different techniques and with different metal loading, imaged with a variety of instrument settings. The full size images have variable sizes between 256×256 and 4096×4096, with 83% at 2048×2048. In Figure 2, we show the distribution of metadata parameters in our dataset, and a description of each of the metadata variables is provided in Appendix A. We also solicited the creation of a few "evaluation sets", which are series of images taken of the same sample in quick succession with varied metadata. Our "evaluation set" includes one series for dwell time, one series for gain, and one series for offset. We denote this set the "evaluation set" as we do not train on any of this data, but retain it for qualitative comparison with our model output, as shown in Figure 5.

4. Experiments & Results

We evaluate CIMP in three stages, each reported as its own subsection below. We first tune the contrastive pre-training setup across crop size, backbone, metadata-encoder capacity, and effective batch size, selecting the strongest configuration by cross-modal retrieval accuracy on the validation set. We then characterise the learned representation by asking how much of the raw acquisition metadata can be recovered from the frozen visual embedding alone through linear probing. Finally, we use the embeddings as a conditioning signal and perceptual loss for a metadata-conditioned style-transfer generator. We show both that the representation transfers to a non-trivial generative task and that virtually scaling dwell time and beam current turns the same network into a physics-informed denoiser.

4.1. Learning Embeddings

To train our embedding model, CIMP, we split our dataset into training and validation splits of size 6,597 and 733, respectively, with a fixed seed. For training, we subsample each image into 512×512 patches following experiments shown in Table 1. We sample 1 random crop per image per training epoch, which yields effectively unlimited unique patches across our training. In addition, the patches are augmented by random 90 degree rotations and horizontal and vertical flips.

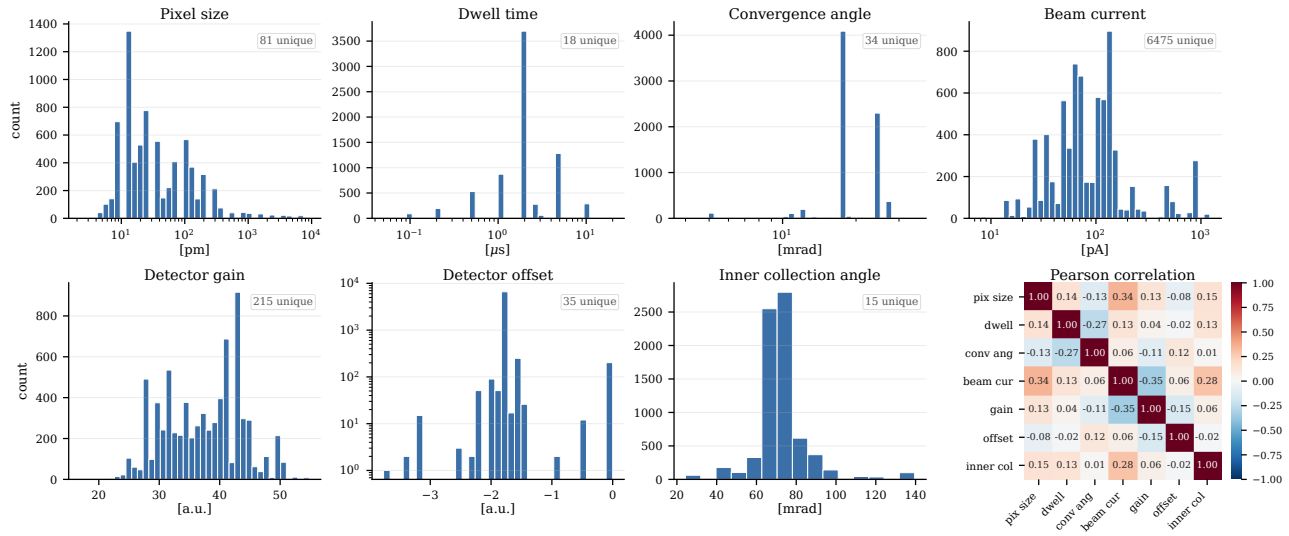


Figure 2. Distribution of acquisition metadata across 7,330 HAADF-STEM images.

4.1.1. Crop Size Sweep

To determine the optimal image crop size for CIMP training, we swept over four crop sizes while holding the effective batch size fixed at 512. All runs were trained for 1000 epochs on the same train/val split (6,597/733 images). Table 1 summarises the best-epoch retrieval metrics.

Table 1. Crop size sweep for CIMP contrastive pre-training (1000 epochs each). ViT-pretrained backbone, metadata encoder 64×2, effective batch 512. For crop 1024, the per-GPU batch was reduced to 8 and gradients were accumulated over 8 steps to fit GPU memory.

Crop	Batch	Ep.	Top-1	Top-5	Top-10
128	512	1000	0.781	0.906	0.953
256	512	1000	0.766	**0.969**	**0.984**
512	512	1000	**0.844**	0.953	0.969
1024	512	1000	0.738	0.913	0.950

Crop 512 achieves the best Top-1 retrieval accuracy across the sweep, with smaller and larger crops both underperforming it. Crop 256 remains competitive on Top-5 and Top-10, while crop 1024 drops off noticeably, possibly because the reduced per-GPU batch and 8-way gradient accumulation weaken the per-step signal. We use crop 512 in subsequent experiments unless otherwise noted.

4.1.2. Metadata Encoder Sweep

Scaling the metadata encoder beyond 64×3 yields modest returns: Top-1 improves from 0.797 at hidden dim 64 to 0.828 at 256, and saturates there. Top-10 reaches 1.000 at hidden dim 128 and above. We adopt 256×3 as the default metadata encoder configuration, which matches the Pareto frontier while keeping parameter count modest.

Table 2. Metadata encoder capacity ablation. Best-model Top-k retrieval accuracy. ViT-pretrained backbone, crop 256, InfoNCE, effective batch 512, 1000 epochs. Width varied at depth 3.

Hidden	Layers	Top-1	Top-5	Top-10
64	3	0.797	0.953	0.984
128	3	0.766	0.953	**1.000**
256	3	**0.828**	0.969	**1.000**
512	3	**0.828**	0.969	**1.000**
1024	3	**0.828**	**0.984**	**1.000**

4.1.3. Batch Size Sweep

Table 3. Effective batch size sweep. Best-model Top-k retrieval accuracy. ResNet-18 backbone, crop 512, InfoNCE, metadata encoder 256×3, 1000 epochs. As the validation set of our data contains only 733 samples, we do not explore batch sizes beyond 512. See Appendix C for the corresponding ViT-pretrained sweep.

Eff. batch	Top-1	Top-5	Top-10
128	0.775	0.988	**1.000**
256	**0.859**	**1.000**	**1.000**
512	**0.859**	**1.000**	**1.000**

Retrieval accuracy improves noticeably from effective batch 128 (Top-1: 0.775) to 256 (Top-1: 0.859) and saturates thereafter. Since our validation set contains only 733 samples, batch sizes beyond 512 would exceed reasonable negative-to-positive ratios; we use 512 as the default.

4.1.4. BACKBONE SWEEP

To select the image encoder, we compared four backbone configurations trained for 1000 epochs on the same train/val split (6,597/733 images) with a crop size of 256, effective batch size of 512, and metadata encoder 256×3. Table 4 summarises the best-epoch retrieval metrics.

Table 4. Backbone sweep for the CIMP image encoder (1000 epochs, crop size 256, effective batch 512, metadata encoder 256×3). "ViT pretrained" and "ViT frozen" use ImageNet-pretrained weights; the former is finetuned end-to-end while the latter keeps the backbone frozen and only trains the projection head.

Backbone	**Top-1**	**Top-5**	**Top-10**
ResNet-18	**0.844**	**0.969**	0.984
ViT (scratch)	0.750	0.953	0.984
ViT (pretrained)	0.828	**0.969**	**1.000**
ViT (pretrained, frozen)	0.234	0.438	0.641

Perhaps surprisingly, ResNet-18 trained from scratch matches or exceeds the ImageNet-pretrained ViT (Top-1: 84.4% vs. 82.8%). This parity suggests that the distribution of HAADF-STEM micrographs differs substantially from natural images in ImageNet: the pretrained transformer features provide little transfer benefit for this domain, while a lightweight convolutional encoder with strong locality priors suffices to learn the relevant structure directly from the STEM data. Training a ViT from scratch lags both (Top-1: 75.0%), as transformers tend to be more data-hungry when initialized randomly. Freezing the pretrained ViT and only training the projection head yields the worst results (Top-1: 23.4%), which again suggests a large domain gap between STEM and natural images. Given this result, along with its smaller parameter count and simpler training pipeline, we adopt ResNet-18 as our default backbone for downstream style-transfer and denoising tasks.

4.2. Downstream Task: Metadata Recovery

Retrieval accuracy tells us whether a visual embedding can be matched back to its own metadata vector, but not whether individual acquisition parameters are recoverable from the image alone. To probe this, we extract 128-d visual embeddings from the frozen CIMP encoder for every image in the train and validation splits (6,597 / 733 images, the same split as CIMP pre-training). We then fit seven independent Ridge regressors ($\alpha = 1.0$) on the training embeddings, one per metadata dimension, each predicting a single z-scored target value. We report per-dimension coefficient of determination (R^2) and symmetric mean absolute percentage error (SMAPE) on the held-out validation split. SMAPE is evaluated in physical units after inverse-transforming predictions (inverting the z-score, and the log10 transform for the four dimensions that span several orders of magnitude). Since Ridge has a closed-form solution, the probe measures how much metadata information is linearly present in the embedding without additional nonlinear decoding capacity.

Table 5. Linear-probe recovery of each metadata dimension from the frozen CIMP visual embedding. Ridge regression ($\alpha = 1.0$) fit on the train split, evaluated on the val split. Arrows indicate the preferred direction for each metric. Dimensions marked † are stored in log10 space in the dataset; their SMAPE is computed after re-exponentiating predictions to physical units.

	ResNet-18		**ViT pretrained**	
Dimension	$R^2 \uparrow$	SMAPE$\downarrow$	$R^2 \uparrow$	SMAPE$\downarrow$
Pixel size†	0.720	41.9%	0.675	47.1%
Dwell time†	0.782	28.4%	0.780	28.6%
Convergence angle†	0.537	13.0%	0.610	11.7%
Beam current†	0.655	34.7%	0.662	34.6%
Gain	0.841	5.3%	0.821	5.6%
Offset	0.802	9.2%	0.812	9.0%
Inner coll. angle	0.499	9.5%	0.655	8.2%
Mean	0.691	20.3%	0.716	20.7%

Both backbones achieve a mean SMAPE of roughly 20% across parameters. Detector gain is the most accurately recovered dimension (SMAPE $\approx 5\%$, $R^2 > 0.82$), followed by offset and inner collection angle at 8 to 10%, and convergence angle at 11 to 13%. These settings modulate global image statistics (contrast, brightness, scattering geometry) in ways that a visual encoder can pick up directly, and their limited dynamic range keeps percentage errors low. The three log-transformed dimensions with wide dynamic range (pixel size, dwell time, beam current) show R^2 between 0.66 and 0.78, but inflated SMAPE of 28 to 47%. This is likely because a modest residual in log10 space becomes a multiplicative factor after re-exponentiation, so recovering a value to within 0.1 decades in log space maps to $\approx 25\%$ error on the physical scale.

The two backbones behave similarly overall, with ViT slightly ahead on mean R^2 (0.716 vs. 0.691) and essentially tied on mean SMAPE ($\approx 20\%$). The R^2 gap is concentrated in the two weakest dimensions: ViT recovers inner collection angle noticeably better ($R^2 = 0.655$ vs. 0.499). We hypothesize that this is because its global receptive field picks up long-range detector cues that the locality-biased ResNet misses. Importantly, the contrastive objective never supervises metadata values directly; it only asks visual embeddings to align with the MLP projection of the metadata vector. Nevertheless, each of the seven parameters is recoverable from the 128-d embedding, which suggests that the encoder has learned the seven distinct features rather than collapsing them onto a single axis.

4.3. Downstream Task: Style Transfer

We utilise a GAN-based architecture for metadata-conditioned style transfer. While other newer methods have become popular in recent years, they are less obvious to adapt to our setting. For example, Palette-style image-to-image diffusion (Saharia et al., 2021) is trained on paired source–target examples, which we do not have, and flow matching (Lipman et al., 2022) is natively a noise-to-image generative method rather than an image-to-image translator. Instead, we rely on GANs as they train reliably on our ∼7,000-image dataset, integrate naturally with the multi-task LPIPS-based objective we use, admit one-shot inference without iterative ODE solving, and allow us to jointly optimise for the correct metadata style, content preservation, and image realism without true image-metadata pairs.

Our style transfer interface takes an input image x and two metadata embeddings, a source e_{id} and a target e_{tgt}, and produces an image that preserves the content of x but matches the style associated with e_{tgt}. We implement this with a U-Net generator G whose convolutional blocks are conditioned on the embeddings via FiLM layers (Perez et al., 2018), allowing the metadata signal to modulate features at every scale while the U-Net skip connections carry content information through unchanged. Both e_{id} and e_{tgt} are produced by the frozen CIMP metadata encoder from raw acquisition-metadata vectors; e_{id} is the embedding of the image's original metadata, and e_{tgt} is the embedding of the shuffled target metadata we want the output to match. FiLM layers receive the concatenation $[e_{\text{tgt}}; e_{\text{id}}]$ as conditioning. We train G against four objectives. The adversarial loss $\mathcal{L}_{\text{LSGAN}}$, in the Least Squares GAN formulation (Mao et al., 2017), drives outputs toward the distribution of real micrographs. The remaining three components are defined as

$$\begin{aligned}
\mathcal{L}_{\text{cyc}} &= \text{LPIPS}\big(G(G(x, [e_{\text{tgt}}; e_{\text{id}}]), [e_{\text{id}}; e_{\text{tgt}}]),\ x\big),\\
\mathcal{L}_{\text{id}} &= \text{LPIPS}\big(G(x, [e_{\text{id}}; e_{\text{id}}]),\ x\big),\\
\mathcal{L}_{\text{emb}} &= \big\|e_{\text{tgt}} - \text{CIMP}_{\text{vis}}(G(x, [e_{\text{tgt}}; e_{\text{id}}]))\big\|_2^2,
\end{aligned}$$

where $\mathcal{L}_{\text{cyc}}$ (Zhang et al., 2018) enforces content preservation by requiring a round trip (source to target, then target back to source) to recover x; $\mathcal{L}_{\text{id}}$ discourages G from modifying images whose target matches their source; and $\mathcal{L}_{\text{emb}}$ anchors the output to e_{tgt} in the CIMP visual-embedding space, ensuring the generated image matches the target style. The full objective combines them:

$$\mathcal{L}_{\text{GAN_total}} = \mathcal{L}_{\text{LSGAN}} + \lambda_1 \mathcal{L}_{\text{cyc}} + \lambda_2 \mathcal{L}_{\text{id}} + \lambda_3 \mathcal{L}_{\text{emb}} \quad (1)$$

with $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.5$ set empirically.

To verify that style transfer works, the metadata in the validation set is shuffled to assign a target style for each image. Style-transferred images are then passed through the CIMP visual encoder to generate embeddings, and we calculate the cosine similarities between these and the target metadata embeddings. If the generator has learned the metadata-to-style mapping, the cosine similarities between output images and their target style should be high while the similarity with other metadata embeddings should be low. As shown in Figure 3, the mean similarity of correct pairs exceeds that of incorrect pairs by 0.533, demonstrating that the generator aligns image style with the provided conditioning embedding.

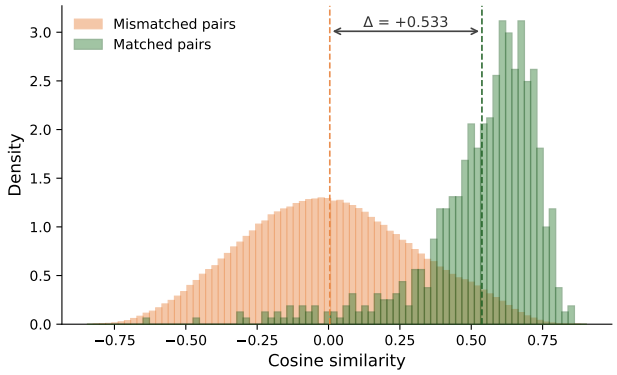


Figure 3. Cosine similarity distribution for style transfer.

Since we have no paired data, further evaluation is qualitative. We evaluate style transfer for each metadata parameter individually, excluding pixel size (as motivated in Appendix A). For each parameter, Figure 5 shows the generator's output at five target values (-3σ, -1σ, 0, $+1\sigma$, and $+3\sigma$ of the distribution of that metadata in the entire dataset), with all other metadata held at the source image's original values. The general observed trends for each metadata are as follows:

1. **Dwell time**: From low to high dwell time, the output image goes from noisy to smooth as expected.

2. **Convergence angle**: When virtually increasing the convergence angle, the resulting effect is a brighter image. This is not what you would expect in theory, and likely happens because of the way convergence angle was controlled in the recording of the dataset. For the vast majority of images (see Figure 2), the convergence angle was set to two different values, controlled by changing the size of an aperture. A smaller aperture gave a smaller convergence angle but also blocked more of the electron beam, leading to a lower beam current which leads to a darker image. This bias is likely what is picked up by the model.

3. **Beam Current**: For beam current, the trend is clear, more electrons lead to a brighter image. Eventually, this should also lead to clipping of the image, but as discussed in section 5, that boundary condition is not

represented in the training data.

4. **Gain**: Increased gain also leads to a brighter image. However, in contrast to the beam current sweep, the noise is also amplified.

5. **Offset**: There is no clear trend for offset except that the background appears slightly darker for lower values.

6. **Inner collection angle**: Making the inner collection angle smaller results in a clear trend of increased image intensity. With a smaller collection angle, more of the scattered electrons end up on the HAADF detector, which should make the image brighter. However, electrons scattered to different angles differ, and with a much smaller collection angle we would expect the contrast mechanism to change as more Bragg-diffracted electrons reach the detector. Such intricacies are likely not learned by the model, especially since the vast majority of images in the dataset were recorded with large collection angles relative to beam convergence angle.

We compare the generator output to the real target for every (source, target) pair in each evaluation series, using SSIM, PSNR, and LPIPS. As a baseline we report the same metrics between the unmodified source image and the real target, which tells us how similar an unedited image already is to its target without any style transfer. Table 6 shows the results. The generator improves on the baseline most clearly for gain, where it wins on all three metrics. Dwell ties with the baseline, which is reasonable given that consecutive images in the dwell series are already very similar (baseline SSIM $= 0.324$, PSNR $= 24.07$) and leave little room to improve. Offset is also close to tied. Overall, LPIPS drops from 0.289 to 0.273.

Table 6. Pixel-level style-transfer metrics on the three held-out experimental sweeps. SSIM / PSNR / LPIPS between the generator output $G(x_{\mathrm{src}}, e_{\mathrm{tgt}}, e_{\mathrm{id}})$ and the real target image, averaged across every ordered (source, target) pair within each series. The "No-edit baseline" columns report the same metrics between the unmodified source image and the real target.

Series	**Model**	SSIM↑	PSNR↑	LPIPS↓
Dwell	CIMP GAN	0.311	22.82	0.188
	No-edit baseline	0.324	24.07	0.187
Gain	CIMP GAN	0.118	11.03	0.322
	No-edit baseline	0.100	10.48	0.369
Offset	CIMP GAN	0.153	10.77	0.364
	No-edit baseline	0.151	10.30	0.367
Overall	CIMP GAN	0.211	16.13	0.273
	No-edit baseline	0.210	16.38	0.289

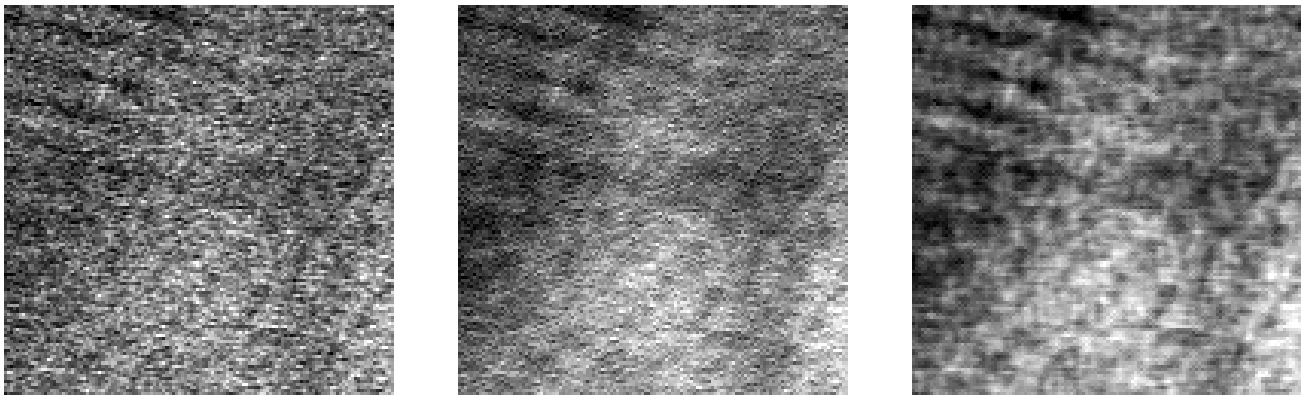
We also test distributional realism with the Kernel Inception Distance (KID) (Bińkowski et al., 2018). We report four comparisons in Table 7. They are generator output against the real eval-set images (a direct realism check), generator output against the training corpus, the 21 real eval-set images against the training corpus (a content-gap floor that isolates the distribution shift between the narrow eval-series content and the wider training corpus, independent of any generator contribution), and two disjoint halves of a training pool against each other (a within-corpus sampling-noise floor). The last three comparisons are computed across five disjoint training-image pools of 500 images each with pinned seeds, so they share the same right-hand reference within each seed. Subset size is 21 for the first three comparisons (bounded by the eval set size) and 50 for the noise floor.

Table 7. Distributional style-transfer realism. Kernel Inception Distance (KID). The first row compares the generator output directly to the real eval-set images, with errors from the internal subset bootstrap. The last three rows average across 5 disjoint training pools of 500 images each, with pinned seeds so they share the same right-hand reference within each seed. Subset size 21 for the first three comparisons (eval set has 21 images), 50 for the noise floor.

Comparison	**KID↓**
Generator vs. eval set	-0.016 ± 0.009
Generator vs. training	0.185 ± 0.009
Eval set vs. training (content floor)	0.189 ± 0.010
Training vs. training (noise floor)	0.0003 ± 0.0004

The within-corpus floor (3×10^{-4}) confirms that Inception features discriminate cleanly on this dataset. As expected, two random samples of real training images are indistinguishable. Most directly, the generator's KID against the real eval-set images is near zero (-0.016 ± 0.009), and we conclude that the generator's output distribution almost perfectly matches the distribution of the real evaluation images. Against the training corpus, the generator's KID (0.185) matches the content-gap floor (0.189) within one standard error. The common ~ 0.19 value therefore reflects the shift between the narrow eval-series content and the wider training distribution, not a generator artefact.

We also compare synthetically changing dwell time, detector gain, and detector offset to the same experimental changes on the microscope. As ground truth, we utilise the experimentally acquired "evaluation set" of gold particles on a carbon film recorded at five different dwell times, five different gains, and five different offsets. We choose to focus on these parameters as they are simple parameters to change experimentally, have a large impact on the image style, and are independent of other system components and thus do not affect the other metadata. Figures 7a, 7b, and 7c show comparisons for dwell time, gain, and offset, respectively. In general, increasing dwell time does make the



(a) Input, $t_{\text{dwell}} = 500$ ns, $I_{\text{beam}} = 74$pA *(b)* CIMP: $t_{\text{dwell}} = 24\,\mu$s, $I_{\text{beam}} = 1.2$ nA *(c)* Noise2Void

Figure 4. Comparison between a 128×128 image crop from the input image, the image denoised by virtually changing dwell time and beam current with the CIMP GAN, and the same image denoised by Noise2Void.

image smoother as expected, but also makes the background brighter which does not happen in the experimental series. For gain, our model correctly predicts the behaviour of lowering and increasing brightness in the image. However, it does not succesfully model the boundary condition where the experimental image is being clipped. The CIMP GAN models change in offset the least well and there is almost no difference between generated images at different offset while there are significant differences in the experimental series.

Finally, as an application, we exploit the trends visualized in Figure 5 and utilize the model as a denoiser by virtually changing dwell time and beam current of experimental images. We compare our model to unconditioned Noise2Void trained on our dataset. A representative image patch showing pixel level denoising performance is plotted in Figure 4, and four more are displayed in Figure 6. Our CIMP GAN denoiser consistently outputs more realistic looking denoised images than Noise2Void, which is overly aggressive in its denoising, suffers from checkerboard artifacts (Figure 4c), and tends to hallucinate content in background regions (see bottom row of Figure 6).

5. Limitations

Boundary conditions are missing from the training data. Operators do not routinely record images with clipped intensities, saturated detectors, or other extreme boundary conditions, because such acquisitions are considered bad data and are typically discarded. As a result, the upper and lower bounds of each acquisition parameter are underrepresented in the dataset, and the model has no examples from which to learn the corresponding image behaviours. This is visible in the gain sweep (Figure 7b), where the model correctly reproduces the trend of increasing brightness and noise but fails to saturate at the high end, and in the offset sweep (Figure 7c), where extreme values are poorly reproduced.

Scale relative to computer-vision datasets. To our knowledge, the 7,330 images released here constitute the largest paired image–metadata HAADF-STEM dataset published to date, and far exceeds any comparable precedent in materials electron microscopy. It is nonetheless modest by computer-vision standards: ImageNet (Deng et al., 2009) contains 1.28 million natural images and has supported three orders of magnitude of model scaling beyond what our corpus can support.

6. Discussion & Conclusion

We introduced CIMP, a CLIP-style contrastive framework that learns a joint embedding space over HAADF-STEM images and their acquisition metadata from 7,330 paired examples collected on a single instrument. The headline finding, in our view, is not the style transfer demonstration but the linear decodability result in Section 4.2: seven physically distinct acquisition parameters are individually recoverable from the frozen 128-d visual embedding through Ridge regression, despite the contrastive objective never supervising any metadata value directly. This indicates that aligning images with a structured metadata vector produces a factored representation as a byproduct of retrieval training, rather than collapsing to a single "noisy-vs-clean" style axis. To our knowledge, this has not been demonstrated before in electron microscopy, and it is the main reason CIMP embeddings are useful as a conditioning signal downstream.

The style transfer network is one demonstration of this usefulness, but the embeddings are not tied to that application. The same representation could condition segmentation or analysis networks that must be robust to changes in dwell time, beam current, or detector settings; serve as a retrieval

index for curating training data by instrument state; or act as a perceptual loss for other generative models in the TEM domain.

Some of the most interesting behaviour we observed is the model's sensitivity to dataset structure rather than to physics. The convergence-angle sweep in Section 4.3 brightens the image as the angle increases, which is not what theory predicts, but does reflect how convergence angle and beam current are correlated in our data through the choice of aperture. Future work will extend the dataset across instruments, investigate how CIMP embeddings transfer to downstream analysis tasks, and quantify the denoising application against stronger baselines on paired reference data.

Software and Data

All code for reproducing the experiments in this paper is available on GitHub: https://github.com/benke97/CIMP. To lower the barrier for other laboratories and members of the electron microscopy community to adopt the method on their own instruments, we will additionally publish a companion repository that streamlines fine-tuning of our best configuration on experimental data collected elsewhere. The paired image–metadata dataset introduced in this work is released on Hugging Face alongside our best pre-trained CIMP and style-transfer: https://huggingface.co/buckets/Stemson-AI/CIMP.

Impact Statement

We hope this work opens the door to higher-throughput electron microscopy and to data-driven methods in experimental materials science more broadly, by showing that leftover, metadata-rich acquisition archives can be turned into general-purpose representations rather than being discarded. More generally, we see this effort as part of the ongoing expansion of machine learning into domains of science that have so far been under-served by the field. To our knowledge, the paired image–metadata HAADF-STEM dataset we release is the first of its kind, and we hope that publishing it encourages other electron-microscopy laboratories to share comparable corpora so that the community can collectively build toward scales at which modern learning methods realise their full potential.

References

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018.

DECTRIS Ltd. QUADRO Hybrid-Pixel Electron Detector. https://www.dectris.com/detectors/electron-detectors/for-materials-science/quadro/, 2024. Accessed: 2026-03-30.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Egerton, R. Two- and three-dimensional electron imaging of beam-sensitive specimens. *Micron*, 194:103819, 2025. ISSN 0968-4328. doi: https://doi.org/10.1016/j.micron.2025.103819. URL https://www.sciencedirect.com/science/article/pii/S096843282500037X.

Eliasson, H. Synthetic data from: Improving nanoparticle size estimation from scanning transmission electron micrographs with a multislice surrogate model, 2025. URL https://doi.org/10.5281/zenodo.14608502. Dataset.

Eliasson, H. and Erni, R. Improving nanoparticle size estimation from scanning transmission electron micrographs with a multislice surrogate model. *Nano Letters*, 25(6):2474–2479, 2025. doi: 10.1021/acs.nanolett.4c06025. URL https://doi.org/10.1021/acs.nanolett.4c06025.

Eliasson, H., Lothian, A., Surin, I., Mitchell, S., Pérez-Ramírez, J., and Erni, R. Precise size determination of supported catalyst nanoparticles via generative ai and scanning transmission electron microscopy. *Small Methods*, pp. 2401108, 2024. ISSN 2366-9608. doi: https://doi.org/10.1002/smtd.202401108. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/smtd.202401108.

Gatan, Inc. K3 Direct Detection Camera. https://www.gatan.com/products/tem-imaging-spectroscopy/k3-camera, 2023. Accessed: 2026-03-30.

Ivanova, N. M., Kashin, A. S., and Ananikov, V. P. Lost data in electron microscopy. *Chemistry*, 7(5), 2025. ISSN 2624-8549. doi: 10.3390/chemistry7050160. URL https://www.mdpi.com/2624-8549/7/5/160.

Khan, A., Lee, C.-H., Huang, P. Y., and Clark, B. K. Leveraging generative adversarial networks to create realistic scanning transmission electron microscopy images. *npj Computational Materials*, 9(1):85, may 2023. ISSN 2057-3960. doi: 10.1038/s41524-023-01042-3. URL https://doi.org/10.1038/s41524-023-01042-3.

Kim, J., Rhee, J., Kang, S., Jung, M., Kim, J., Jeon, M., Park, J., Ham, J., Kim, B. H., Lee, W. C., Roh, S.-H., and Park, J. Self-supervised machine learning framework for high-throughput electron microscopy. *Science Advances*, 11(14):eads5552, 2025. doi: 10.1126/sciadv.ads5552. URL https://www.science.org/doi/abs/10.1126/sciadv.ads5552.

Krull, A., Buchholz, T.-O., and Jug, F. Noise2void - learning denoising from single noisy images. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2124–2132, 2018. URL https://api.semanticscholar.org/CorpusID:53751136.

Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., and Aila, T. Noise2Noise: Learning image restoration without clean data. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2965–2974. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/lehtinen18a.html.

Lin, R., Zhang, R., Wang, C., Yang, X.-Q., and Xin, H. L. 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):5386, 2021. doi: 10.1038/s41598-021-84499-w. URL https://doi.org/10.1038/s41598-021-84499-w.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling, 2022. URL https://arxiv.org/abs/2210.02747.

Lobato, I., Friedrich, T., and Van Aert, S. Deep convolutional neural networks to restore single-shot electron microscopy images. *npj Computational Materials*, 10(1):10, jan 2024. ISSN 2057-3960. doi: 10.1038/s41524-023-01188-0. URL https://doi.org/10.1038/s41524-023-01188-0.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Smolley, S. P. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2794–2802, 2017.

Parker, L., Lanusse, F., Golkar, S., Sarra, L., Cranmer, M., Bietti, A., Eickenberg, M., Krawezik, G., McCabe, M., Morel, R., Ohana, R., Pettee, M., Régaldo-Saint Blancard, B., Cho, K., Ho, S., and The Polymathic AI Collaboration. Astroclip: a cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011, 2024. doi: 10.1093/mnras/stae1450. URL https://doi.org/10.1093/mnras/stae1450.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. FiLM: Feature-wise linear modulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Reimer, L. *Specimen Damage by Electron Irradiation*, pp. 421–453. Springer Berlin Heidelberg, Berlin, Heidelberg, 1984. ISBN 978-3-662-13553-2. doi: 10.1007/978-3-662-13553-2_10. URL https://doi.org/10.1007/978-3-662-13553-2_10.

Saharia, C., Chan, W., Chang, H., Lee, C. A., Ho, J., Salimans, T., Fleet, D. J., and Norouzi, M. Palette: Image-to-image diffusion models, 2021. URL https://arxiv.org/abs/2111.05826.

Sanchez-Fernandez, A., Rumetshofer, E., Hochreiter, S., and Klambauer, G. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1):7339, 2023. doi: 10.1038/s41467-023-42328-w. URL https://doi.org/10.1038/s41467-023-42328-w.

Schwenker, E., Sen, F., Wolverton, C., Ophus, C., and Chan, M. K. Y. A simulated atomic-resolution haadf stem imaging dataset containing unique icsd structure prototypes, 2020. URL https://doi.org/10.18126/szeq-yde5. Dataset.

Sheth, D. Y., Mohan, S., Vincent, J., Manzorro, R., Crozier, P. A., Khapra, M. M., Simoncelli, E. P., and Fernandez-Granda, C. Unsupervised deep video denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

Sytwu, K., Rangel DaCosta, L., and Scott, M. Segmented high-resolution transmission electron microscopy images of nanoparticles, 2023. URL https://doi.org/10.7941/D1SP93. Dataset.

Taleb, A., Kirchler, M., Monti, R., and Lippert, C. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20876–20889. IEEE, 2022. doi: 10.1109/CVPR52688.2022.02024. URL https://doi.org/10.1109/CVPR52688.2022.02024.

Thermo Fisher Scientific. Falcon 4i Direct Electron Detector Datasheet. https://documents.thermofisher.com/TFS-Assets/MSD/Datasheets/falcon-detector-datasheet-ds0371.pdf, 2024. Accessed: 2026-03-30.

Thornley, W., Sullivan-Allsop, S., Cai, R., Clark, N., Gorbachev, R., and Haigh, S. J. Noise2void for denoising atomic resolution scanning transmission electron microscopy images. *npj Computational Materials*, 12 (1):68, Jan 2026. ISSN 2057-3960. doi: 10.1038/s41524-025-01939-1. URL https://doi.org/10.1038/s41524-025-01939-1.

Wang, F., Henninen, T. R., Keller, D., and Erni, R. Noise2atom: unsupervised denoising for scanning transmission electron microscopy images. *Applied Microscopy*, 50(1):23, 2020. ISSN 2287-4445. doi: 10.1186/s42649-020-00041-8. URL https://doi.org/10.1186/s42649-020-00041-8.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

A. Brief explanation of the metadata parameters

1. **Pixel Size:** This is the physical size of a pixel (typically around 15-50 pm for high resolution imaging), and the lateral step between probe positions in the STEM raster scan. Changing the pixel size is the result of changing magnification, which in normal operation would result in a larger field of view. It does not make physical sense to ask a model to change pixel size of the image while keeping the image size fixed. Therefore it is ignored and also not shuffled in the metadata shuffling within batches nor in the validation set for evaluation.

2. **Dwell Time:** Dwell time determines how long the electron beam stays at each pixel position. A longer dwell time leads to more electrons hitting the sample at each pixel position and consequently leads to a higher signal image. A shorter dwell time leads to reduced signal and a significant increase of poisson noise.

3. **Beam Convergence Angle:** The convergence angle of the beam is the incident angle of the convergent STEM probe as it hits the sample. The convergence angle is a key parameter of the diffraction-limited achievable resolution of the microscope. Changing the convergence angle can be done in multiple ways, but in this dataset, it was mostly controlled by the size of an aperture.

4. **Beam Current** This is the current of the electron probe, as measured by the fluorescent viewing screen, which is placed after the sample and HAADF detector. While a larger beam current indicates that more electrons hit the sample, the effect on signal strength depends on how many electrons hit the HAADF detector which is affected by other metadata like inner collection angle described below. Beam current should in general be kept as low as possible to avoid damaging the sample and inducing phenomena that do not occur naturally.

5. **Detector Gain:** The ”Gain” or ”Contrast” is a parameter that determines how much the signal on the detector is amplified.

6. **Detector Offset:** The “Offset” or “Brightness” is a parameter that sets the black level of the detector reading.

7. **Inner Collection Angle:** The HAADF detector is annular, and the inner collection angle defines the minimum scattering angle of electrons that contribute to the recorded signal. This parameter is important both because it determines the fraction of the incident beam collected and because it governs the imaging contrast mechanism. At high scattering angles, the signal is dominated by incoherent Rutherford scattering, which forms the basis of HAADF-STEM contrast and gives rise to the approximate $Z^{1.7}$ dependence and mass–thickness contrast. At lower collection angles, coherently scattered electrons are also detected, introducing diffraction contrast and altering the imaging mechanism. For HAADF-STEM, including the majority of images in our dataset, the inner collection angle is typically chosen to be at least ~ 3 times the probe convergence semi-angle.

B. Qualitative Analysis

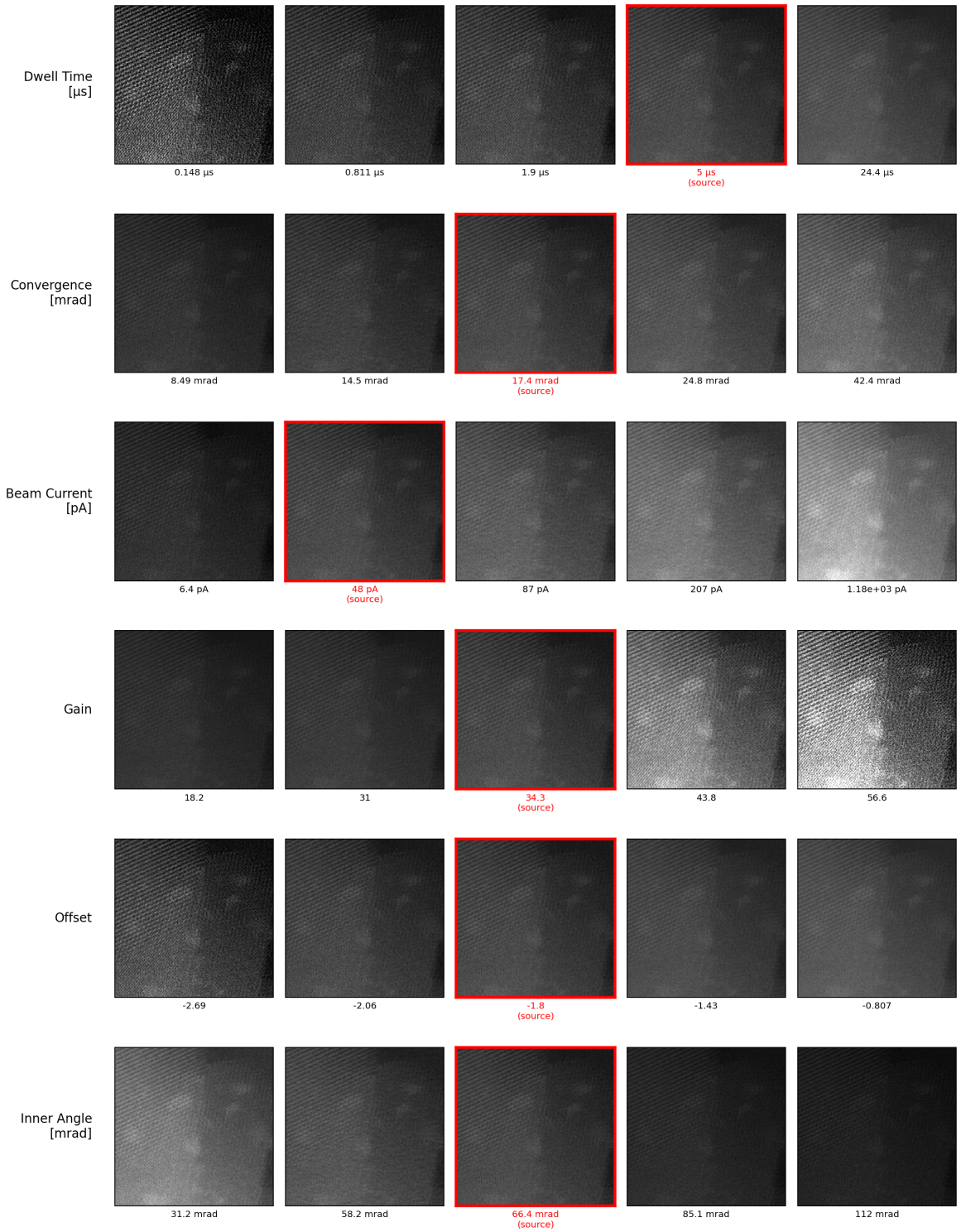


Figure 5. Our CIMP GAN style transfer model synthetically varying the metadata parameters of a validation image, one by one. Style-transferred images were generated at five target values (-3σ, -1σ, 0, $+1\sigma$, and $+3\sigma$ of the distribution of that metadata in the full dataset). The image in the red frame is the original input image and it replaces whichever of the generated images that it has the closest metadata to. Collecting the metadata of all images with a red frames gives you the original metadata. *Note: Gain and Offset are also denoted as Contrast and Brightness by some other STEM instruments/software, and their ranges can be different from Gain and Offset.*

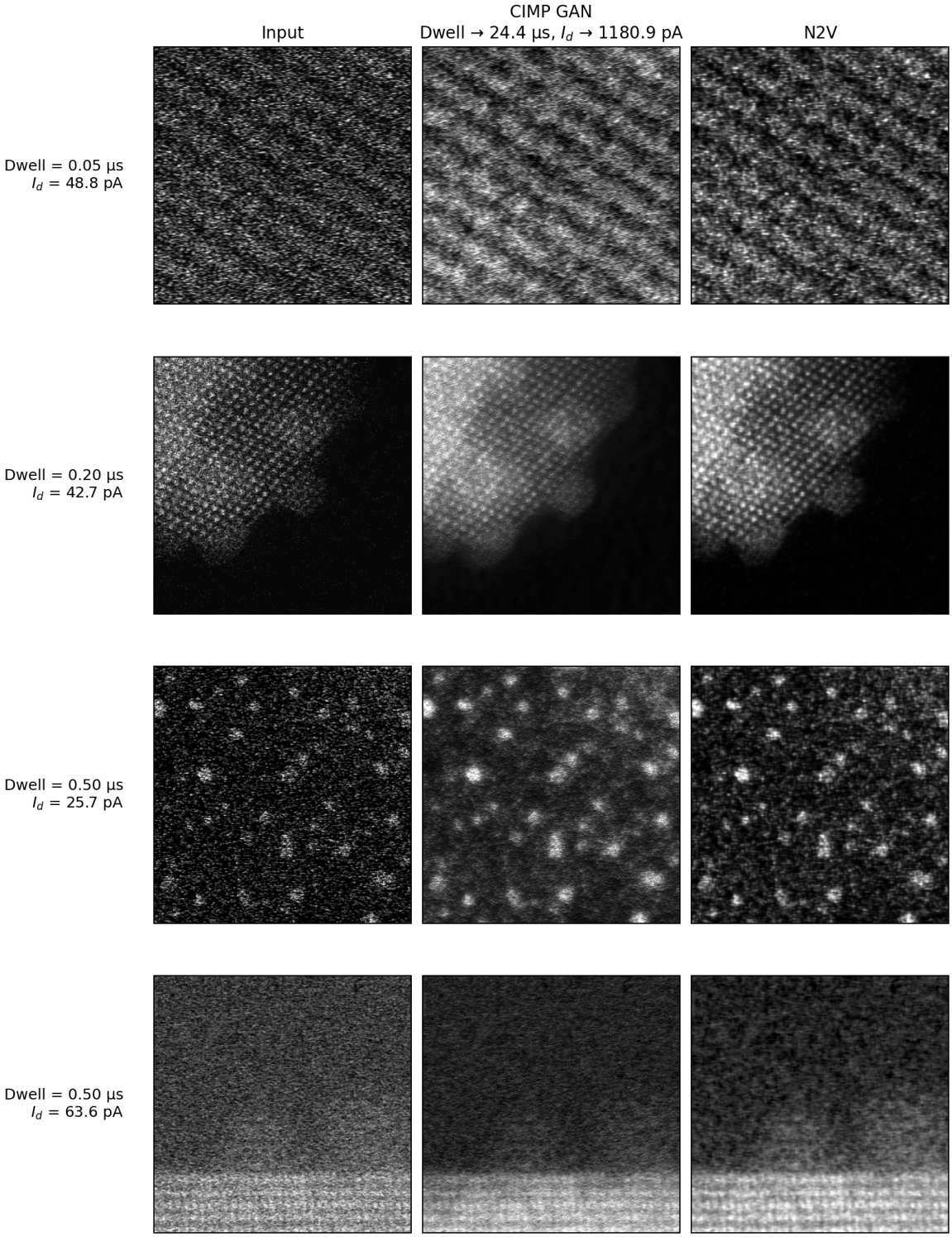
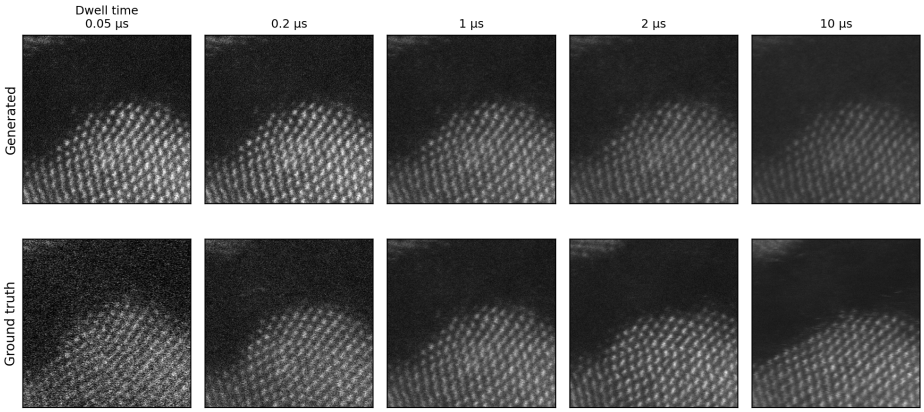
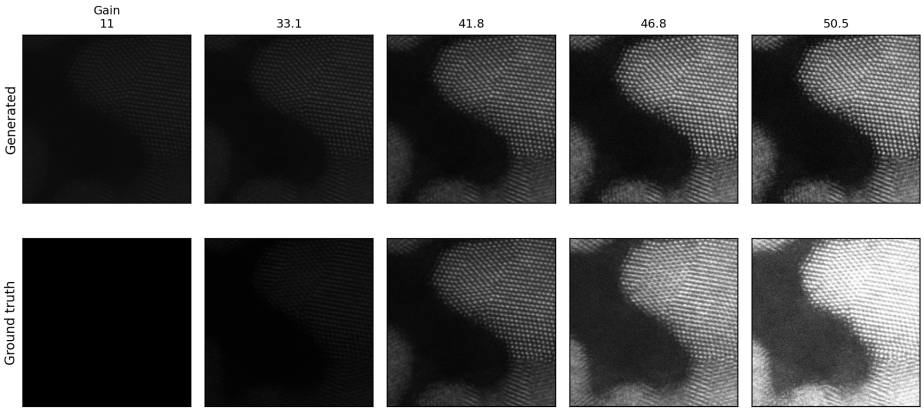


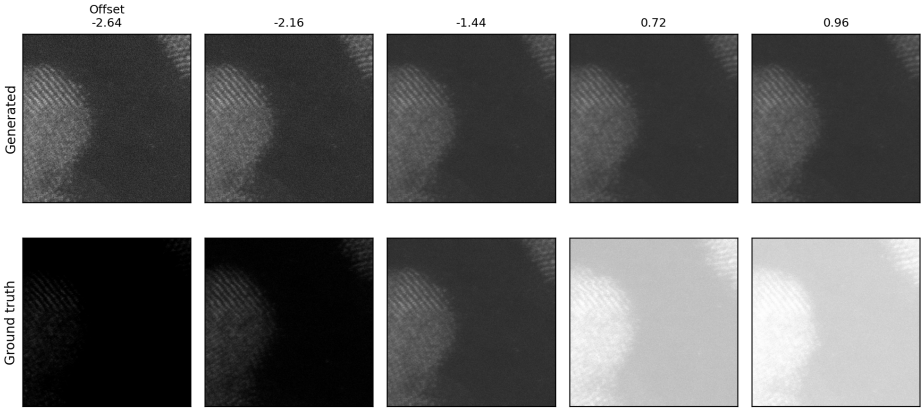
Figure 6. Comparison between 512×512 image crops from input image, image denoised by virtually changing dwell time and beam current with the CIMP GAN, and image denoised by Noise2Void.



(a) Results of varying the dwell time synthetically (top row) with our style transfer model and experimentally (bottom row). The expected behaviour of increasing the dwell time is a smoother image with less poisson noise but with a maintained pixel intensity distribution. We observe that our model indeed makes the image smoother, but also alters the intensity distribution by making background slightly brighter and object slightly darker. Furthermore, reducing dwell time virtually does not add as much noise as we observe experimentally.



(b) Results of varying the gain (or contrast) synthetically (top row) with our style transfer model and experimentally (bottom row). The expected behaviour of increasing the gain is an amplification of both signal and noise and a brighter image. We observe that our model can successfully model this trend but fails to model the upper bound where the the image would be clipped (further discussed in section 5)



(c) Results of varying the offset (or brightness) synthetically (top row) with our style transfer model and experimentally (bottom row). The expected behaviour of increasing the offset on this instrument is that the black level is lowered, pushing the visible content into a smaller part of the dynamic range, eventually clipping the image. Decreasing the offset leads to real signal being treated as black and parts of the object disappearing. We observe that a virtual change of offset does not closely mimic the behaviour of experimentally changing offset, we discuss possible reasons for this in section 5.

Figure 7. Comparison between changing dwell time, gain, and offset virtually and experimentally

C. Batch Size Sweep (ViT)

For completeness we reproduce the effective batch size sweep with the ViT-pretrained backbone. ResNet-18 saturates at Top-1 $= 0.859$ from batch 256 onwards; ViT saturates at Top-1 $= 0.828$ at the same batch sizes, consistent with the backbone comparison in Table 4.

Table 8. Effective batch size sweep with ViT-pretrained backbone. Best-model Top-k retrieval accuracy. Crop 256, InfoNCE, metadata encoder 256×3, 1000 epochs.

Eff. batch	Top-1	Top-5	Top-10
128	0.775	0.950	0.975
256	**0.828**	**0.969**	**1.000**
512	**0.828**	**0.969**	**1.000**

D. Training Histories

Figures 8 and 9 show the per-epoch training histories for the two learned models reported in the main text: the best CIMP encoder (ResNet-18, crop 512, metadata encoder 256×3, effective batch 512, 1000 epochs) and the metadata-conditioned style-transfer GAN trained on top of it (StyleUNet with FiLM conditioning, base filters 32, crop 512, batch 8, 250 epochs). Bold curves show a rolling mean (25 epochs for CIMP, 10 epochs for the GAN); raw per-epoch values are overlaid in faint.

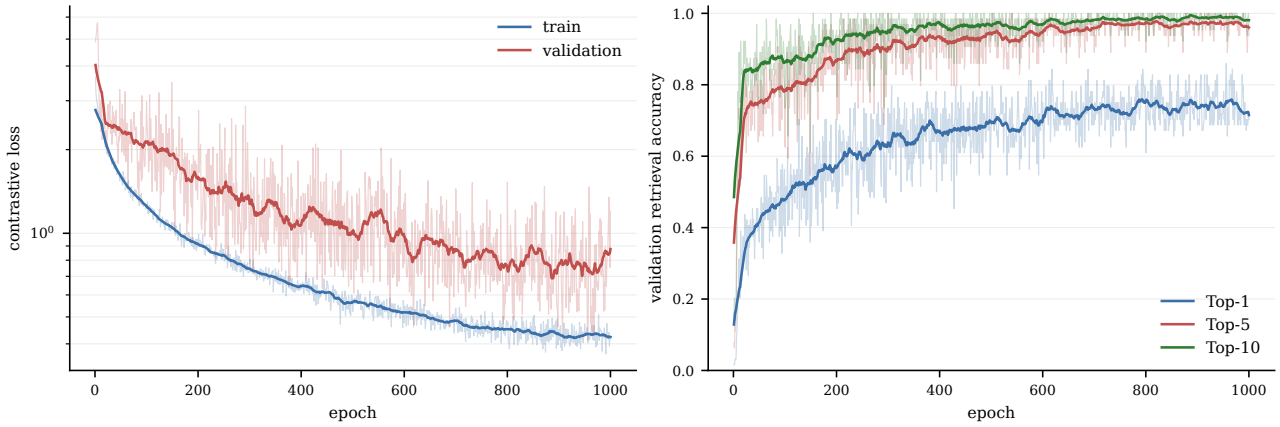


Figure 8. Training history of the best CIMP encoder. Left: train and validation contrastive loss (log scale). Right: validation retrieval accuracy (Top-1, Top-5, Top-10).

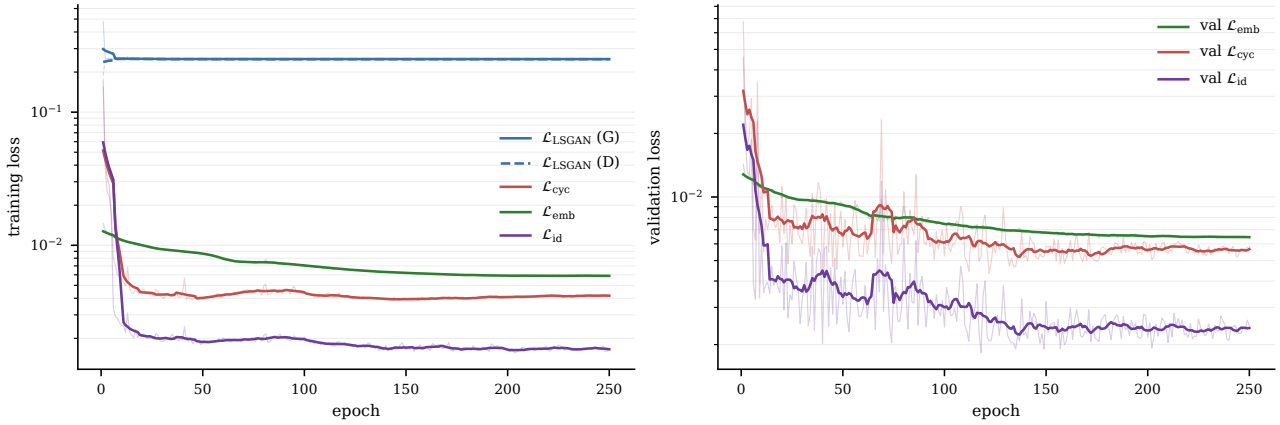


Figure 9. Training history of the style-transfer GAN. Left: training losses per term, including the LSGAN generator / discriminator pair and the three non-adversarial components. Right: validation losses for the three non-adversarial components.