SUPPORTING INFORMATION

CrossMicroNet: A Cross-Scale Small-Sample Image Restoration Framework for Two-dimensional Material Microscopy Imaging

Mingwei Feng[a,b], Xilu Zou[c], Lei Liu[d], Shengqiang Wu[d], Haotian Zhang[e], Silin Chen[e], Zikang Zeng[e], Yiru Wang[e], Xiaotian Zhang[d], Xuping Zhang[a,b], Taotao Li[e,f] and Ningmu Zou*[e,f]*

[a] College of Engineering and Applied Sciences, Nanjing University, Nanjing 210023, China

[b] Key Laboratory of Intelligent Optical Sensing and Manipulation, Ministry of Education, Nanjing University, Nanjing 210093, China

[c] National Laboratory of Solid-State Microstructures, School of Electronic Science and Engineering and Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing 210023, China

[d] Suzhou Laboratory, Suzhou 215123, China

[e] School of Integrated Circuits, Nanjing University, Suzhou 215163, China

[f] Interdisciplinary Research Center for Future Intelligent Chips (Chip-X), Nanjing University, Suzhou 215163, China

E-mail: nzou@nju.edu.cn

* To whom all correspondence should be addressed.

Table of Contents

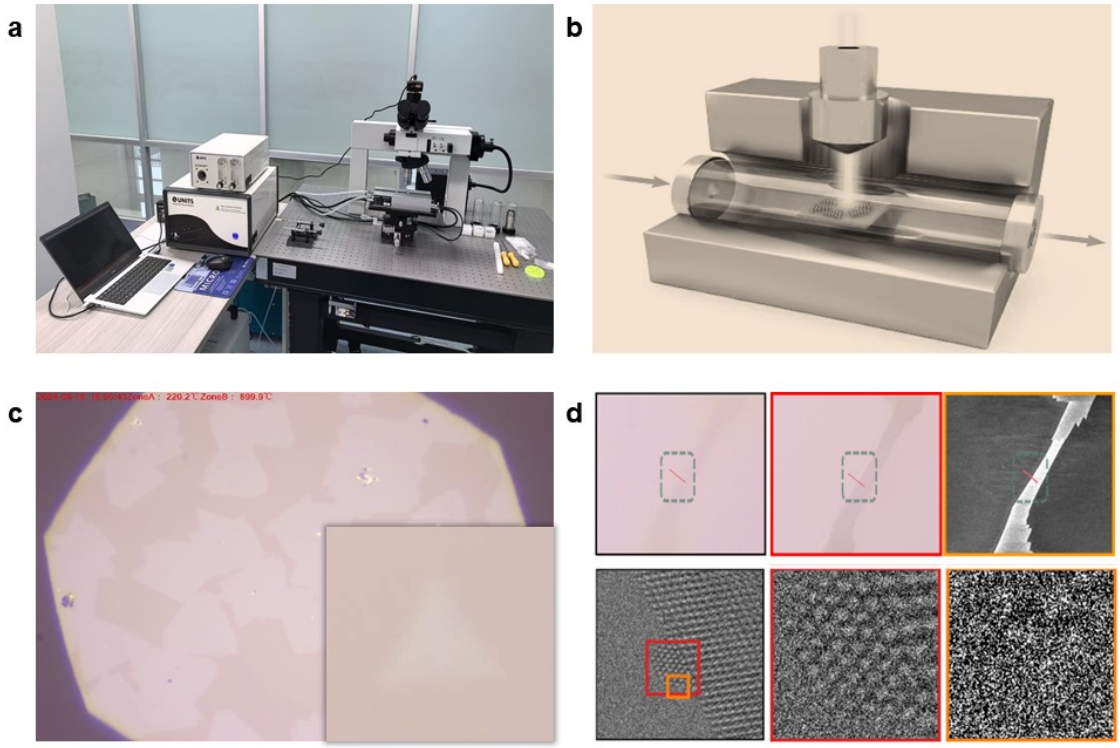

Fig.S 1 Device schematic diagram and task description. (a) Photograph of the in-situ optical microscopy setup integrated with a CVD growth chamber. The optical microscope (center) is mounted on the reactor, allowing real-time imaging of the 2D materials growth on the substrate. This setup provided the raw video frames for enhancement. Important components (microscope, sample stage, and control computer) are visible; (b) Schematic diagram of a reaction tube furnace and wall mounted optical microscope, with argon gas flowing into the left side. The tube contains sulfur powder, molybdenum oxide, and catalyst (NaCl, etc.); (c) The in-situ growth video recorded by the optical microscope, which can reveal the growth kinetics laws; (d) Up: Optical in-situ micrograph of a 2D materials: original frame (left, black outline) is blurry with low contrast edges; after enhancement (middle, red outline), the sheet's triangular edges become sharper and more defined; an SEM image of the same region (right, orange outline) serves as a ground-truth reference with highest clarity. The green dashed box highlights an edge region and the green line indicates where the edge spread function was measured for spatial resolution analysis. Down: Atomic-resolution STEM images demonstrating the carbon contamination removal branch algorithm: a raw STEM image with carbon contamination (left, black outline) shows a faint atomic lattice obscured by noise; the processed image after U-Net dehazing (middle, red outline) has significantly reduced haze, revealing the atomic lattice; a reference "clean" image (right, orange outline) is provided for comparison (e.g., either after plasma cleaning or simulation). The red square in the left image denotes a region of interest, and the subsequent orange square zooms into a single atomic cluster. The bottom row images illustrate how the algorithm preserves actual atomic features while removing diffuse contamination.

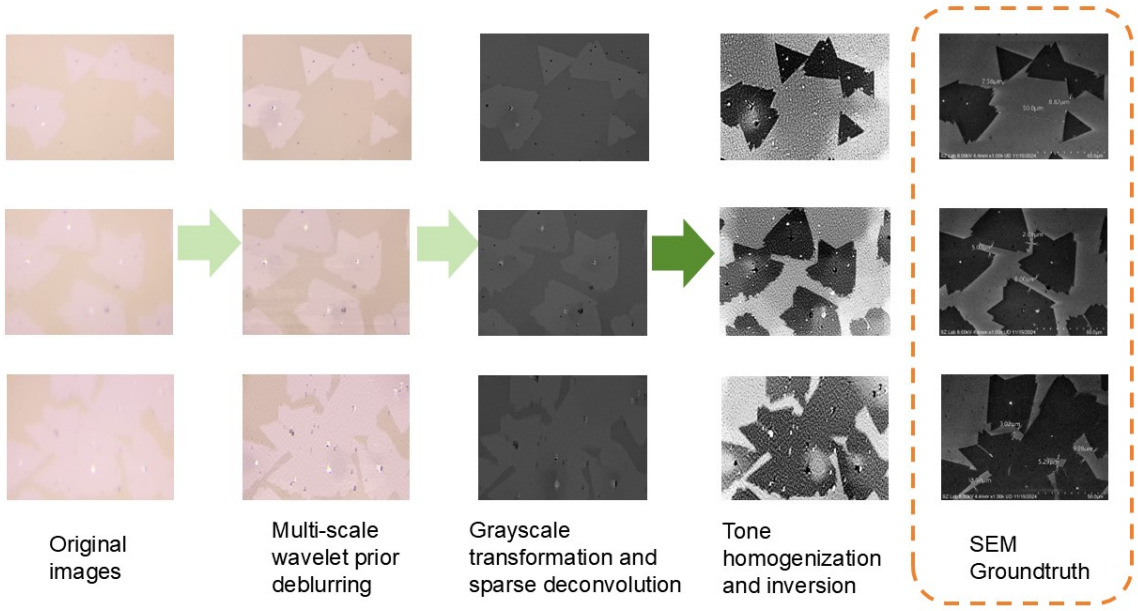

Fig.S 2 Comparison of image quality before and after gradually applying enhancement methods.

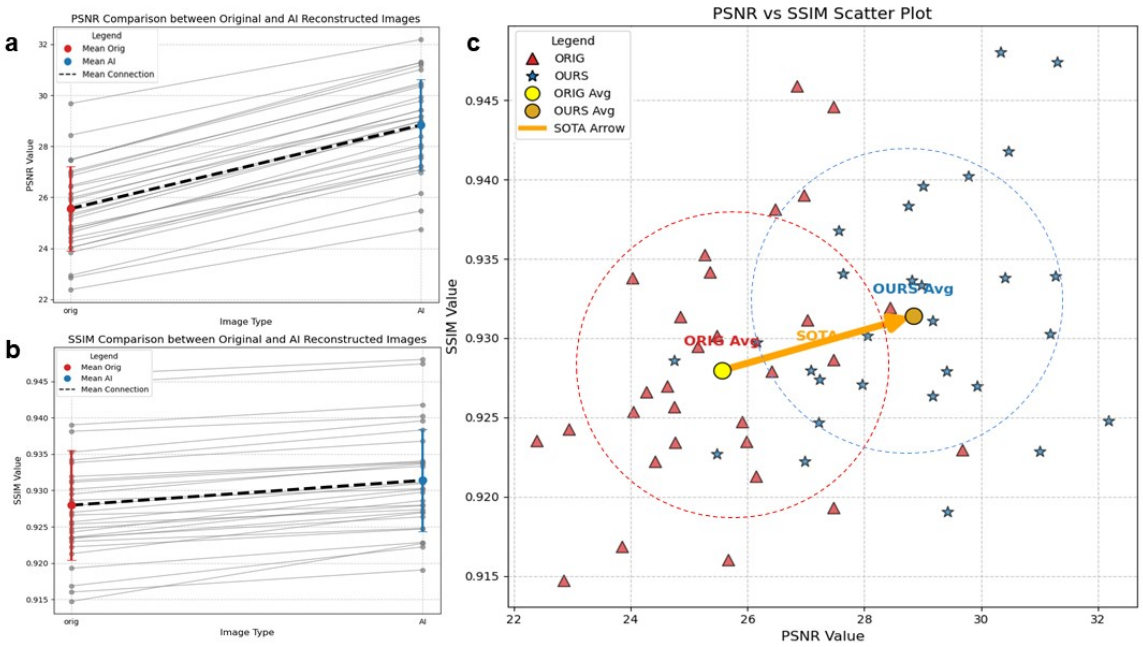

Fig.S 3 Comparison of PSNR and SSIM between original and AI-reconstructed images. Algorithm enhancement and grayscale inversion are compared with the original image, and processing results are evaluated relative to SEM images as a reference. (a) PSNR Comparison: Distribution of PSNR values across a set of image frames comparing original images (red) versus AI-enhanced images (blue). Data points along the x-axis represent individual image frames, and PSNR values are shown on the y-axis. The blue points consistently lie above the red, indicating that the AI reconstruction consistently improves PSNR. The variation in PSNR values reflects the differences in image content and initial quality, but the overall upward shift in PSNR is clear; (b) SSIM Comparison: Distribution of SSIM values for original and AI-reconstructed images. Red points represent the SSIM of the original images, and blue points represent the enhanced images. The enhanced images show higher SSIM values than the original images, indicating that the enhancement brings the reconstructed images closer to the structural characteristics of the SEM reference. Despite modest absolute SSIM values (0.92–0.94 range), the enhancement improves the structural similarity of the images; (c) PSNR vs SSIM Scatter Plot: This plot compares PSNR (x-axis) and SSIM (y-axis) values

for 30 pairs of original and AI-reconstructed images. The red triangles represent original images (ORIG), while the blue stars represent the AI-enhanced images (OURS). Yellow and orange circles mark the average PSNR and SSIM values for the original and enhanced images, respectively. The plot also highlights a "SOTA" (state-of-the-art) reference. The scatter plot clearly shows that the AI-enhanced images (OURS) achieve higher PSNR and SSIM compared to the original images, with the averages for the enhanced images (OURS Avg) closer to the state-of-the-art benchmark. The orange arrow indicates the direction from the original to the enhanced images, with a notable improvement in both PSNR and SSIM. The red and blue circles represent areas of the scatter plot where the original and enhanced images respectively fall, indicating the relative performance improvement.

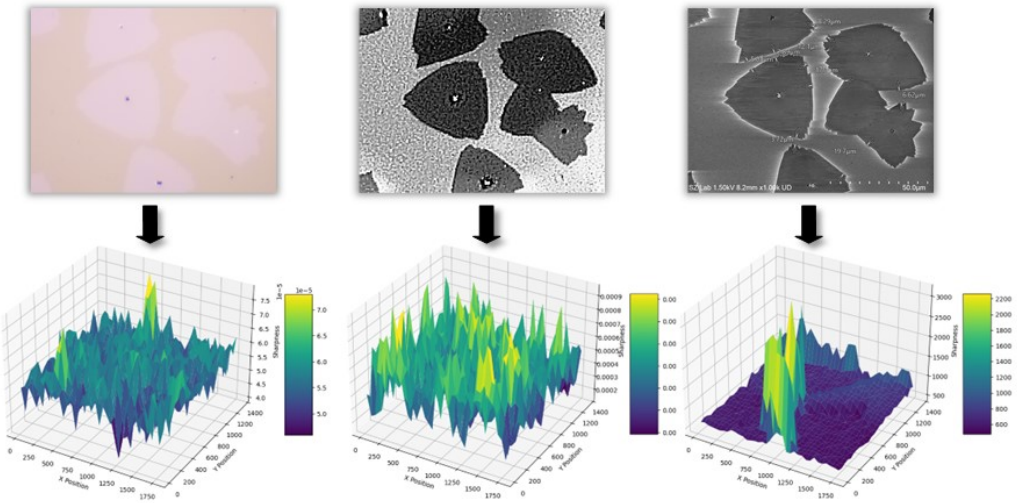

Fig.S 4 Clarity assessment results by calculating the variance of the Laplacian operator. left: original image, middle: image with algorithmic clarity enhancement and grayscale inversion that approaches SEM style, right: SEM reference image.

Note S1 In-situ Optical Microscopy Data Acquisition

2D material growth was observed in real-time using an optical microscope integrated over a chemical vapor deposition (CVD) chamber. The microscope (with adjustable magnification, exposure, and frame rate) captured video of the growing 2D material in situ, ensuring stable imaging conditions. The video stream was segmented into sequential frames to serve as input images for enhancement. This provided a time-resolved dataset of blurry, low-contrast micrographs reflecting the 2D material's evolving microstructure.

Note S2 OM Dataset Preprocessing and Contrast Enhancement

Each captured frame was first preprocessed to improve visibility of features. If the image had an alpha transparency channel (RGBA format), it was split into separate RGB color and alpha channels. The RGB channels were converted to the LAB color space to isolate luminance (L) from color information. Local contrast was then enhanced on the luminance channel using Contrast-Limited Adaptive Histogram Equalization (CLAHE)S1. CLAHE (clip limit ≈ 2.0, grid size 8×8) amplifies local contrast (especially in darker regions) while preventing over-amplification of noiseS2. After CLAHE, the adjusted luminance was recombined with the original A and B color channels and converted back to RGB, then merged with the preserved alpha channel. This yielded an RGB image with improved contrast (particularly revealing dark features) but retained overall color fidelity. Finally, the enhanced color image was converted to grayscale (8-bit) for subsequent deblurring, preserving structural and textural details in a single intensity channel. A mild denoising was applied (e.g. bilateral filtering or non-local means) to suppress high-frequency noise while maintaining edges.

Note S3 Wavelet-Based Edge Enhancement

To further restore high-frequency details, a multiscale wavelet transform was applied to the grayscale image. Using a 2-level 2D discrete wavelet decomposition (Haar wavelet), the image was split into low-frequency (approximation) and high-frequency (detail) sub-bands. An adaptive gain was then applied to

the high-frequency sub-bands: if the image's average brightness was low (indicating an overall darker image), a larger gain (e.g. 2.0×) was used to strongly boost fine edges; for brighter images, a moderate gain (~1.2×) was used to avoid overshooting. This adaptively amplifies edge contrast and texture details. The modified high-frequency sub-bands were then recombined with the original low-frequency sub-band using inverse wavelet transform (pywt.waverec2), reconstructing an edge-enhanced image. The pixel intensities of the reconstructed image were clipped to [0,255] to remove any out-of-range values introduced by enhancement. The result was an RGBA image with noticeably sharper edges and finer detail visibility than the original.

Note S4 Blind Deconvolution (Blind Deblurring)

The edge-enhanced grayscale image was next processed with a blind deconvolution algorithm to correct blurring without prior knowledge of the point-spread function (PSF). We employed a maximum a posteriori (MAP) estimation approach to simultaneously estimate the blur kernel (PSF) and the latent sharp image. The algorithm iteratively alternated between two updates: (1) Kernel Estimation – using the current estimated sharp image to infer the blur kernel that, when convolved, best produces the observed image; and (2) Image Restoration – using the current kernel estimate to deblur the image, enforcing prior constraints. The MAP objective included a data fidelity term (minimizing the difference between the blurred image and the convolution of the estimated sharp image with the kernel) and regularization terms on the image gradients and on image intensity sparsity to favor sharp edges and suppress noise. We initialized with a rough estimate (e.g. a small Gaussian kernel or uniform kernel) and ran a fixed number of iterations (e.g. 5 iterations). Regularization parameters were tuned (e.g. gradient sparsity weight ~4×10^−3) to balance sharpness and smoothness. During each iteration, the blur kernel and image estimate were updated in an alternating minimization scheme. The final output of the blind deconvolution step was an estimated blur kernel (PSF) and an intermediate deblurred image. The estimated kernel was normalized and saved for analysis. This blind deblurring step significantly reduced global blur and revealed finer structural details, but sometimes introduced mild ringing artifacts around sharp features.

Note S5 Non-blind Deblurring and Artifact Removal

Given the estimated blur kernel from the previous step, a non-blind deconvolution was performed to further refine the image and remove artifacts like ringing. Since the presence of saturated regions can affect deconvolution, the algorithm first checked for saturation in the intermediate image (e.g. pixel intensity clipping). In our dataset, images were not saturated (no overexposed areas), so a standard deconvolution method was selected (if saturation were detected, an alternative de-ringing algorithm would be used to specifically suppress ringing artifacts) S3. We employed an iterative deconvolution with combined regularizations – total variation (TV) and an L_0 gradient sparsity term – using the known kernel. Regularization parameters were set to moderate values (e.g. $\lambda_{TV} \approx 10^{-3}$, $\lambda_{L0} \approx 5 \times 10^{-4}$) to smooth noise and remove ringing, without over-smoothing fine details. The algorithm (implemented as a custom ringing_artifacts_removal function) iteratively optimized the image: reducing oscillatory ringing patterns while preserving true edges. After convergence, the image was intensity-normalized and converted back to 8-bit. This step yielded a cleaner, sharper image with suppressed halos or ringing near edges (result from aggressive deblurring) and reduced background noise. (At this stage, an example of the image before and after deblurring is shown in Figure S2.)

Note S6 Sparse Domain Reconstruction

As a final enhancement, a sparse reconstruction step was applied to exploit the image's sparsity in a transform domain and further boost clarity. We assumed that in an appropriate domain (such as wavelet or gradient domain), the image has a sparse representation (i.e., most coefficients are near zero, with only a few significant ones corresponding to real features) S4. We constructed an optimization problem combining a data fidelity term (ensuring the result remains close to the current deblurred image) with sparsity and smoothness priors. Specifically, we imposed an L_1 penalty on a transform-domain representation (to encourage sparsity in, e.g., multi-scale wavelet coefficients or image gradients) and possibly a continuity prior to avoids isolated artifacts. In our implementation, parameters were set (for example, fidelity weight ~150, sparsity weight ~15, continuity weight ~1) to balance these objectives. We solved this optimization using an accelerated iterative shrinkage-thresholding algorithm (ISTA/FISTA) implemented in PyTorch and leveraged GPU acceleration for

efficiency. We ran on the order of 100 iterations for convergence. The output of this sparse optimization was a high-quality reconstructed image (denoted SHVideo internally) with enhanced fine textures and edge continuity beyond the conventional deconvolution result.

Note S7 Post-Enhancement and Output

To ensure no remaining subtle blurring, we applied a final touch-up by again performing a wavelet-based sharpening on the reconstructed image. Using the same Haar wavelet approach (2-level), the high-frequency components were slightly boosted (with the same adaptive gain criteria based on brightness). The high-frequency–boosted image was then recombined with its low-frequency part and merged with the original alpha channel. This produced the final enhanced image with improved edge crispness. The pixel intensities were clipped to [0,255] and cast to 8-bit. The final enhanced frames were saved (e.g. as TIFF or PNG with an "_enhanced" suffix). The normalized blur kernel from the blind deconvolution stage was also saved as a PNG image for reference.

Note S8 Image Quality Evaluation

We quantitatively evaluated the improvement in image quality using both reference-based and no-reference metrics. Since ground-truth images for the in-situ optical frames are not directly available, we utilized ex-situ high-resolution images for reference comparisons. In particular, after the growth experiment, the same sample regions were imaged by scanning electron microscopy (SEM), providing a high-resolution "ground truth" for structural details. For a set of corresponding regions, we computed the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) of the original optical images and our enhanced images, using the SEM image as the reference. As expected, both PSNR and SSIM improved notably after enhancement (Figures S3): for example, PSNR increased from ~6–8 dB (original) to ~9–11 dB with our method for various frames, indicating the enhanced images more closely resemble the high-resolution reference. SSIM values similarly showed improvement (from very low values ~0.1–0.2 up to ~0.2–0.3 in many cases), although they remained relatively low in absolute terms due to inherent differences between optical and electron microscopy contrast (Figure S3).

For a no-reference assessment of sharpness, we analyzed the spatial resolution improvement via edge sharpness. Because the light source is a point light source, we extracted Point Spread Function (PSFs) from well-defined feature edges (e.g. the boundary of a MoS2 domain) in the images. Each PSF (intensity profile across an edge) was smoothed with a small Gaussian filter (σ = 1 px) to reduce noise. We then fit each PSF to the cumulative distribution of a Gaussian (error function fit) to model the intensity transition across the edge. From this fit, the σ parameter (standard deviation of the underlying point-spread function) was obtained as a quantitative measure of edge blurriness – a smaller σ indicates a sharper, higher-resolution edge. We compared σ for the same edge region imaged under three conditions: the original optical image, the image after our enhancement (AI-enhanced), and the high-res SEM image. For example, in one representative edge (Figure S1d), the original optical image had the largest σ (blurriest edge), the SEM reference had the smallest σ (sharpest edge), and our enhanced image's σ was intermediate, much closer to SEM than the original. This indicates that our method significantly improves spatial resolution, narrowing the gap toward electron microscopy clarity. We also evaluated a few classical image enhancement approaches for comparison (such as simple CLAHE-only processing and standard deblurring algorithms). These methods provided some improvement in contrast or resolution, but none achieved the level of detail recovery of our integrated approach. In terms of the edge σ metric, the images processed by conventional methods remained more blurred (higher σ) than those processed by our method (data included in Figure 2 analysis), confirming the superiority of the proposed enhancement pipeline.

Note S9 Algorithm Description: STEM Carbon Contamination Removal

In addition to optical image enhancement, we developed a branch algorithm to remove carbon contamination artifacts from high-resolution Scanning Tunneling Microscope (STEM) images. Carbon contamination accumulates as a hazy, amorphous layer on STEM samples, degrading image clarity. The algorithm employs a deep-learning approach (U-Net convolutional network) to "dehaze" contaminated atomic-resolution images, combined with post-processing to refine atomic details. Pseudocode and key components of this algorithm are outlined below:

Note S10 STEM Dataset Preparation

Construct a training dataset of contaminated vs. clean STEM images. If ground-truth clean images are not directly available, simulate contamination by overlaying STEM images with synthetic "haze" or noise layers that mimic carbon buildup (e.g. low-frequency intensity gradients or blurring). Alternatively, use pairs of experimentally obtained images before and after plasma cleaning. All images are normalized and patched into manageable sizes (e.g. 512×512) for training. Data augmentation (rotations, flips) is applied to increase robustness. The network learns to map a contaminated input patch to a cleaned output patch.

Note S11 Multi scale small sample U-Net Model Structure

Implement a U-Net deep convolutional neural network to learn contamination removal. The U-Net consists of an encoder path (progressively down sampling the input via convolution and pooling layers to capture context) and a decoder path (up sampling via transposed convolutions and skip-connecting high-resolution features from the encoder). This architecture preserves fine spatial details (important for atomic lattices) while learning the global haze patternsS5. We use a depth of several layers (e.g. 4 down/up-sampling levels) with increasing feature channels (e.g. doubling channels at each down sampling). Activation functions (e.g. ReLU) and batch normalization are used to stabilize training. The output layer produces a residue or cleaned image. Training: The U-Net is trained using a mean squared error or L_1 loss between the network output and the target clean image. Optionally, we include a perceptual loss term computed on image features to ensure atomic lattice structures are accurately recovered. The network is optimized (Adam optimizer) until the validation loss converges, typically over dozens of epochs.

Note S12 Patch-Based Dehazing with Blending

Due to the large size of STEM images (often several thousand pixels wide), the trained U-Net is applied in a patch-wise fashion during inference. The contaminated STEM image is divided into overlapping patches (for example, 512×512 with 50% overlap) to fit into GPU memory. Each patch is fed through the U-Net to predict a preliminary dehazed patch. To avoid seam artifacts at patch borders, we use a blending strategy: overlapping regions of adjacent patches are averaged (or smoothly weighted) when stitching the output patches back together. This ensures a seamless reconstruction of the full decontaminated image without boundary discontinuities. The patch blending can involve Gaussian weighting near edges of patches or simple averaging since the network predictions are generally consistent across overlaps.

Note S13 Atomic Mask Refinement

After the initial U-Net dehazing, the image is much clearer but we apply an additional refinement to ensure atomic features (e.g. atomic columns or lattice fringes) are crisp. We generate an atomic mask by identifying the positions of atomic columns in the image. For instance, we can band-pass filter the dehazed image to highlight periodic atomic lattice frequencies and then threshold or apply a blob detection to find bright spots corresponding to atomic columns. This yields a binary mask locating atomic structure. We also locate these features in the original contaminated image (some may be faint or obscured). Using the mask, we refine the U-Net output: wherever an atomic column is expected (mask = true), we ensure the intensity and contrast in the dehazed image align with the original image's high-frequency content. In practice, this can mean adding back some high-frequency components from the original image at atomic sites or sharpening those regions. The mask can also guide a mild unsharp masking specifically on atomic features. This step preserves genuine atomic details and prevents the network from over-smoothing or altering the lattice.

Note S14 Inference and Post-Processing

The final inference pipeline combines the above steps. Given a new contaminated STEM image, we apply preprocessing if necessary (normalization and tiling into patches). Each patch goes through the U-Net model to obtain a cleaned patch, and patches are merged with overlap blending to form the full cleaned imageS6. Next, the atomic mask is computed on the cleaned image (optionally informed by the original) and used to touch up the cleaned image as described, thereby reinstating any attenuated atomic spots. Finally, post-processing adjusts the image contrast and removes any residual low-frequency shading: a polynomial background subtraction or low-pass filtering can be applied to the cleaned image to flatten any remaining contamination gradient (tone homogenization). The output is a high-clarity STEM image

with carbon contamination effectively removed and atomic lattice contrast restored. Any minor artifacts introduced by the network (such as small speckles) can be removed with a median filter or by referencing the atomic mask (to avoid affecting atomic sites). The result is suitable for accurate analysis of atomic structure without the interference of contamination haze.

References

S1. K. J. Zuiderveld, in Graphics Gems IV, ed. P. S. Heckbert, Academic Press Professional, Inc., San Diego, CA, 1994, pp. 474–485.

S2. A. Buades, B. Coll and J.-M. Morel, Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, IEEE, 2005, vol. 2, pp. 60–65.

S3. A. Mittal, R. Soundararajan and A. C. Bovik, IEEE Signal Process. Lett., 2013, 20, 209–212.

S4. L. Xu, S. Zheng and J. Jia, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 1107–1114.

S5. O. Ronneberger, P. Fischer and T. Brox, in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, ed. N. Navab, J. Hornegger, W. M. Wells and A. F. Frangi, Springer, Cham, 2015, vol. 9351, pp. 234–241.

S6. R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, Proceedings - 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, IEEE Computer Society, 2018, pp. 586–595.