Nanoscale

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d6nr01702a

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]

High-quality microscopy is central to resolving the growth behavior, morphology, lattice organization and defect landscape of two-dimensional (2D) materials. Yet microscopy data acquired across scales are degraded by fundamentally different mechanisms: *in situ* optical microscopy (OM) is often compromised by motion blur, defocus, vibration-induced smearing and illumination inhomogeneity, whereas scanning transmission electron microscopy (STEM) is strongly affected by beam-induced amorphous carbon contamination. Here we introduce CrossMicroNet, a unified cross-scale image-clarification framework that couples a shared restoration front end with modality-adaptive refinement. The restoration module integrates contrast-limited adaptive histogram equalization, mild non-local means denoising, blind deconvolution, ringing suppression and wavelet-domain spatial-channel enhancement. For OM, this front end directly sharpens domain boundaries from small-sample data without requiring paired optical ground truth. For STEM, the restored output is further processed by a lightweight contamination-suppression branch that combines conservative structural guidance with smooth fusion to attenuate diffuse background haze while preserving lattice periodicity. Evaluated on 28 OM/SEM region pairs, where SEM serves only as an approximate structural reference, CrossMicroNet reduces the apparent edge-transition width to 0.22 μm and yields the most favorable overall trade-off among NIQE, LPIPS and PSNR-like structural-reference metrics. On the STEM benchmark, it achieves the best learning-based performance, with a PSNR of 20.50 dB, SSIM of 0.85, LPIPS of 0.08, VIF of 0.92 and FSIM of 0.93. Fourier-domain analysis further confirms suppression of low-frequency contamination while retaining lattice-frequency features. These results establish CrossMicroNet as a practical cross-scale clarification strategy for linking growth-scale OM with atomic-scale STEM in 2D-material research.

Received 29th April 2026,
Accepted 17th May 2026

DOI: 10.1039/d6nr01702a

rsc.li/nanoscale

Introduction

The study of two-dimensional (2D) materials, such as graphene and transition metal dichalcogenides, has garnered significant attention owing to their exceptional electronic, optical, and mechanical properties, which open new avenues for a variety of advanced technological applications.[1] Their performance is strongly influenced by crystal size, edge structure, vacancy distribution, phase purity and local stacking order, so microscopy is not merely illustrative but foundational to mechanism discovery and process optimization.[2,3] In growth studies, optical microscopy (OM) is often the only modality capable of continuously surveying a large field of view and following nucleation, domain coalescence and morphological evolution *in situ*.[4] At the other end of the length-scale spectrum, scanning transmission electron microscopy (STEM) resolves atomic columns, point defects and lattice distortions that cannot be captured by OM.[5] In realistic experimental workflows, these two modalities are complementary rather than interchangeable, and efficient interpretation often depends on moving between them without losing confidence in the physical meaning of image contrast.

The difficulty is that the dominant image-degradation mechanisms are scale dependent. In OM, the combined influ-

[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 National Laboratory, Suzhou 215123, China
[e]School of Integrated Circuits, Nanjing University, Suzhou 215163, China. E-mail: nzou@nju.edu.cn
[f]Interdisciplinary Research Center for Future Intelligent Chips (Chip-X), Nanjing University, Suzhou 215163, China

ence of sample drift, thermal vibration, long-working-distance optics, defocus and illumination heterogeneity broadens edge transitions and weakens domain contrast.[6] The consequence is not only a visually softer image but also poorer downstream segmentation, uncertain grain-boundary localization and less reliable extraction of growth metrics such as domain size, perimeter evolution and areal coverage.[7] In STEM, the challenge is different. Under prolonged beam exposure, hydrocarbon species in the microscope column or on the sample surface can be cracked and redeposited as amorphous carbon, producing a slowly varying haze superimposed on the atomic lattice.[8] This low-frequency contamination masks weak atomic contrast, degrades signal-to-noise ratio and complicates identification of point defects, vacancies and local disorder. Drift and shot noise further aggravate the problem.[8] A useful restoration framework therefore has to separate physically meaningful lattice contrast from nuisance background while avoiding hallucination of non-existent structure.

Many published algorithms address only one degradation mode at a time.[9] Classical OM deblurring methods are effective when the blur kernel is known or weakly varying, but their performance deteriorates for real *in situ* microscopy where the point spread function is uncertain and illumination is non-uniform.[10,11] Deep-learning super-resolution models can produce visually sharp outputs, yet they usually require large paired datasets that are unavailable for microscopy under dynamic growth conditions, and they may optimize for perceptual realism rather than structural faithfulness.[12,13] Likewise, for contaminated STEM images, generic dehazing or denoising networks often suppress background haze but do not explicitly protect atomic periodicity or defect contrast. Conventional Fourier filtering and band-pass processing remain useful, especially for experienced microscopists, but they are fundamentally frequency-selective operations and cannot by themselves model the non-stationary and spatially varying contamination field.[14]

The main objective of this study is therefore not to introduce two unrelated image-enhancement pipelines, but to formulate a coherent cross-scale clarification strategy. CrossMicroNet is built around the observation that both OM blur and STEM contamination reduce the organization of useful contrast before the modality-specific interpretation stage. A shared front-end restoration engine can be used to recover local contrast, sharpen edges and suppress nuisance fluctuations in both modalities. Once the input has been brought into this clearer intermediate representation, the two branches diverge according to imaging physics: the OM branch ends after front-end restoration, whereas the STEM branch applies a small-sample U-Net-based refinement stage with atomic-mask guidance and adaptive fusion. This architecture intentionally balances interpretability and learning capacity. The front end remains close to classical image-formation reasoning, while the back end is introduced only where the inverse problem becomes too complex to treat with fixed filters alone. A unified framework is particularly attractive in the context of small-sample scientific datasets. Unlike internet-scale vision tasks, microscopy studies of a specific 2D material system often involve only tens of relevant OM images and a very limited number of comparable STEM frames. Dataset heterogeneity is also high: acquisition settings, contamination levels, magnification, detector response and illumination patterns can vary substantially even within one project. Under such conditions, a purely supervised end-to-end solution is difficult to train and even harder to trust. Hybrid models that embed explicit knowledge of image formation, while reserving learned components for the truly underdetermined parts of the problem, are therefore better matched to experimental reality.

This work also addresses a practical bottleneck in 2D-material characterization pipelines: the disconnect between image enhancement and physically meaningful evaluation. Researchers often test restoration methods using whatever metric is available, even when the metric does not match the acquisition physics.[15] Here, the evaluation is built around what can actually be claimed for each modality. OM is assessed by boundary sharpness and by approximate structural correspondence to SEM after registration, whereas STEM is assessed by same-field synthetic full-reference metrics and by Fourier-domain inspection on real images. This level of evaluation discipline is essential if image clarification methods are to be adopted confidently in nanoscale research. A second aim is methodological restraint. In microscopy, it is easy to overstate what an algorithm has achieved if image sharpness is conflated with the instrument's true resolving power or if cross-modal references are treated as exact pixel-level ground truth. For that reason, the OM analysis in this work uses the apparent edge-transition width, w_{edge}, as an image-level sharpness proxy rather than calling it spatial resolution in the strict microscope-physics sense. Similarly, because clear *in situ* optical ground-truth images do not exist for the same field of view, SEM-assisted OM PSNR and LPIPS values are interpreted as approximate structural-reference indices after registration and contour normalization. For STEM, paired synthetic contaminated/clean crops are used for quantitative benchmarking, while real contaminated images are reserved for qualitative validation, profile inspection and Fourier-domain analysis. This separation keeps the evaluation physically interpretable and avoids mixing unlike comparisons.

The manuscript is organized as follows. Section 2 describes the two-stage CrossMicroNet framework, including the shared front-end restoration engine, the STEM-specific few-shot back end, the data construction procedure and the evaluation protocol. Section 3 discusses the OM and STEM results in detail, with emphasis on how the quantitative values in Tables 1–3 relate to the underlying algorithmic design and to the practical strengths and limitations of different baselines. The final section summarizes the scope of the method and the conditions under which it should be used conservatively.

Method

Unsupervised sparse deconvolution deblurring

In *in situ* optical microscopy imaging, limitations of the optical system, the diffraction limit, and experimental disturb-

Table 1 Performance of CrossMicroNet and baselines on the OM evaluation set

	w_{edge}[a]	LPIPS	PSNR	NIQE
Anime4K	0.93	0.69	6.07	14.94
CUGAN	0.98	0.70	5.97	14.41
ESRGAN	0.80	0.69	6.07	18.06
FFT-ReLU	1.19	0.69	6.14	12.06
SRMD	1.24	0.69	6.09	15.49
Waifu	0.85	0.69	6.07	14.95
Ours	**0.22**	**0.68**	**6.39**	**11.25**

[a] The measurement unit of w_{edge} is μm, the measurement unit of PSNR is dB, and LPIPS and NIQE have no specific units.

Table 2 Performance of CrossMicroNet and learning-based baselines on the STEM test set

	PSNR	SSIM	LPIPS	VIF	FSIM
DCP	14.26	0.18	0.40	0.20	0.30
AOD-Net	16.50	0.45	0.32	0.50	0.60
GridDehazedNet	18.00	0.65	0.21	0.70	0.78
FFA-Net	18.50	0.75	0.15	0.80	0.86
DehazedFormer	19.50	0.82	0.10	0.90	0.91
CrossMicroNet	**20.50**	**0.85**	**0.08**	**0.92**	**0.93**

Table 3 Conventional STEM-processing baselines

Conventional baseline	PSNR/dB	SSIM	LPIPS	VIF	FSIM
Gaussian band-pass filter	17.10	0.56	0.25	0.61	0.79
Fourier-mask filtering	16.95	0.51	0.28	0.57	0.75
Wiener deconvolution	17.84	0.60	0.22	0.66	0.82
Registration + averaging	18.22	0.68	0.19	0.73	0.84
CrossMicroNet	**20.50**	**0.85**	**0.08**	**0.92**	**0.93**

ances often result in image blur and low contrast. Furthermore, the implementation of supervised learning for image clarity of *in situ* observed microscopic images seems tricky due to the lack of clear image pairs suitable for training.[4] To monitor the fine details during 2D material growth in real time and uncover more kinetic insights, we have developed a blind deblurring and enhancement image processing pipeline. This method takes a degraded image as input and automatically estimates the blur kernel to restore a sharp image without requiring prior knowledge of the point spread function (PSF). The overall workflow includes: image preprocessing, blind deconvolution restoration, ringing artifact removal, wavelet-domain multi-scale enhancement, and attention fusion. After passing through these modules, an image's clarity and detail contrast are markedly improved. Below, we introduce the principles and implementation details of each module.

In optical microscopy imaging, the image formation and blurring process can be described by a convolution model. Let the ideal clear image be f and the PSF (blur kernel) be h. Then the captured blurry image g can be expressed as a convolution:[5]

$$g = f \otimes h + n \tag{1}$$

where $\otimes$ denotes convolution and n represents imaging noise (commonly assumed to be additive white Gaussian noise). For motion blur, the PSF can be parameterized by the motion trajectory length and direction; for defocus blur, the PSF can be approximated by a disk or Gaussian function. Non-uniform illumination can be modelled as a multiplicative factor, *i.e.*, after convolutional blurring the image undergoes a position-dependent illumination intensity modulation:[5]

$$g'(x) = [f \otimes h](x) \cdot L(x) \tag{2}$$

In this case, the effective signal in dark background regions is attenuated. The above imaging process is an ill-posed inverse problem: given g and unknown f and h, we need to introduce priors or constraints to solve for them, *i.e.* perform blind deconvolution.[6]

To address locally low contrast caused by uneven illumination, we first apply Contrast Limited Adaptive Histogram Equalization (CLAHE) as a preprocessing step. CLAHE stretches contrast by computing the cumulative distribution function (CDF) of the histogram within local neighbourhoods of the image, while limiting the maximum histogram height to avoid over-amplifying noise.[7] Specifically, for the histogram of pixel intensities in an image sub-block, if the cumulative probability of any grey level exceeds a threshold, it is clipped and the excess is redistributed to smooth the histogram, yielding a limited CDF. The mapping from input pixel intensity I to output intensity I can be expressed as:[7]

$$I' = I_{\min} + (I_{\max} - I_{\min}) - \mathrm{CDF}_{\mathrm{clip}}(I) \tag{3}$$

where $I_{\min}$ and $I_{\max}$ are the endpoints of the intensity range. By limiting the slope of the CDF (which corresponds to capping histogram peaks), CLAHE avoids the problem of standard adaptive histogram equalization that amplifies noise in uniform regions.[4] After CLAHE processing, illumination non-uniformity is corrected and local contrast and details are noticeably enhanced.

To suppress noise while enhancing contrast, we employ Non-Local Means (NLM) filtering for denoising. Unlike local filters that consider only neighbouring pixels, NLM leverages similar structures from across the entire image to smooth noise.[15] The principle of NLM denoising is: for any pixel, use a weighted average of all pixels in the image that have a similar neighbourhood to estimate the true intensity. Mathematically:

$$\hat{I}(p) = \frac{1}{C(p)} \sum_{q \in \Omega} w(p,q) I(q) \tag{4}$$

where Ω is the image domain, I is the observed noisy intensity, $\hat{I}$ is the filtered estimate, and $w(p,q)$ is the weight function, satisfying $0 \le w(p,q) \le 1$ and $\sum_q w(p,q) = 1$ (with $c(p)$ as the normalization factor).[16] The weights depend on the similarity

between the image patches cantered at p and q, and is commonly defined in a Gaussian form:[15]

$$w(p,q) = \exp\left(-\frac{I(N_p) - I(N_q)^2}{h^2} - \frac{p-q^2}{\sigma^2}\right) \quad (5)$$

where the first term measures the grey-level difference between the patches N_p and N_q (controlling decay with intensity similarity) and the second term measures the spatial distance (controlling decay with spatial proximity). If the neighborhood of q is very similar to that of p, $w(p,q)$ is relatively large, and if not, $w(p,q)$ is very small. This non-local redundancy-based information fusion can remove random noise while well preserving image details and textures.[15] Applying NLM filtering after CLAHE enhancement effectively suppresses the noise that was amplified by the contrast boost, achieving a smooth background while maintaining edges.

After the above preprocessing, noise and illumination non-uniformity have been mitigated, and we proceed to restore the image from convolutional blur. Blind deconvolution requires jointly estimating the blur kernel h and the latent sharp image f.[14] This is a highly ill-posed problem, typically solved by introducing prior regularization in the optimization. Generally, one can construct an energy function:[6]

$$E(f,h) = \| f \otimes h - g \|^2 + \gamma[R_f(f) + R_h(h)] \quad (6)$$

where the first term is a data fidelity term, and R_f and R_h are regularization terms (priors) on the sharp image and blur kernel, respectively, with "γ" as a weighting factor. Typical image priors include sparse gradients (*e.g.* total variation, TV) or natural image statistics, and blur kernel priors can assume the kernel's energy is concentrated or that its sum equals 1, *etc*. Blind deconvolution algorithms usually iterate by alternatingly estimating f and h. For example, the classical Richardson–Lucy algorithm and its variants have been used for iterative updates in blind deconvolution.[14] In this work, we adopt an improved approach combining pseudoinverse filtering with robust priors. On one hand, we obtain an initial estimate in the frequency domain (denoting $\mathcal{F}$ as the Fourier transform) while filtering out high frequencies to reduce sensitivity to noise; on the other hand, we introduce an ℓ_2-norm gradient sparsity prior in the spatial domain to suppress ringing artifacts in the restoration. By refining the blur kernel estimate through a multi-scale pyramid from coarse to fine, we improve convergence stability. The output of blind deconvolution is an initial restored image and the estimated blur kernel. Considering the estimation process may introduce artifacts, we design an adaptive filtering and artifact detection step: based on local contrast and gradient statistics, we identify potential overshoot/ringing regions and apply adaptive Wiener filtering to smooth these areas, further improving the visual quality of the restored image.

Multi-scale wavelet domain channel-spatial dual attention mechanism

Although blind deconvolution recovers overall sharpness, certain high-frequency details (such as the fine texture of grain boundaries) may still be insufficiently crisp. To address this, we introduce a discrete wavelet transform (DWT) to decompose the image into multi-scale frequency bands, and integrate a spatial-channel attention mechanism to enhance details. Specifically, we perform a two-level DWT on the restored image, yielding low-frequency approximation components and multi-directional high-frequency detail components.[17] The low-frequency component contains the image's overall structure and large-scale information, while the high-frequency components carry edges and details. Applying attention enhancement in the wavelet domain allows us to selectively amplify important features at different scales.[9]

The attention mechanism comprises two aspects: channel attention and spatial (pixel) attention. Channel attention aims to let the network automatically learn "which" feature channels to focus on. Formally, for a given layer's feature maps $F\mathbb{R}^{H\times W\times C}$, one can obtain a channel importance vector by global pooling (*e.g.* averaging) each channel to a single value, then passing these through a two-layer fully connected network and a Sigmoid activation to produce a weight for each channel. Spatial attention learns a weight map for each spatial position (i,j), usually by compressing the feature maps across the channel dimension (*via* average or max pooling) and then applying a convolution to produce an attention map.[12] Combining the two gives an output feature map enhanced by both spatial and channel attention. We can express the combined spatial-channel attention output as:

$$F_{(i,j,k)'} = M_c(k) M_s(i,j) F_{(i,j,k)} \quad (7)$$

where $M_c(k)$ is the channel attention weight for channel k, and $M_s(i,j)$ is the spatial attention weight at position (i,j).[12] Intuitively, this mechanism makes the network emphasize those few types of features that are distinctive (for example, edges in a particular orientation), while focusing on the salient or interesting regions of the image (for example, areas where a material film is present). In our method, we apply spatial-channel attention amplification separately to each high-frequency sub-band obtained from the wavelet decomposition: let H be a high-frequency sub-band, and let the attention module output a weight map M for it; the enhanced high-frequency coefficients are then given by elementwise multiplying the original coefficients by the attention weights, $H' = M \odot H$. The low-frequency component is left as is or only channel attention is applied (since it mainly contains large-scale intensity information). Finally, we perform inverse DWT on the low-frequency (LL) and enhanced high-frequency (LH, HL, HH) components to reconstruct the final clarified image. This wavelet + attention strategy is essentially a frequency-domain multi-scale enhancement network: the wavelet basis provides a sparse representation of details at each scale, and the attention mechanism selectively amplifies the salient portions of image

detail.[9] Studies have shown that incorporating wavelet transforms into deep networks can help alleviate image blurring, while attention mechanisms effectively extract key information. Therefore, our approach further improves edge sharpness and enhances texture details, while suppressing unimportant high-frequency noise.[17]

In summary, the overall pipeline for optical microscopy image clarification can be summarized as follows: first, perform preprocessing based on image statistical properties and human visual characteristics (CLAHE + NLM) to boost initial contrast and SNR; then restore the main sharp structure *via* blind deconvolution with priors; finally, apply attention mechanisms in the wavelet multi-scale domain to reinforce details. This multi-module workflow is illustrated in Fig. 1a, which shows that starting from an input blurry image, after three stages – preprocessing enhancement, blind restoration, and frequency-domain attention – a clarified image is produced.

Supervised branch based on multi-scale small sample U-Net strategy

High-resolution STEM, as a more microscopic supplement to OM images, can directly image the atomic arrangement of 2D materials and is a powerful tool for studying crystal defects and interfaces. However, carbon contamination induced by electron beam irradiation can form a non-uniform coating on the sample surface, causing local contrast to drop and SNR to decrease, and in severe cases obscuring the bright/dark contrast of atoms.[5] Traditional solutions include lowering the beam current, pre-cleaning with plasma, *etc.*, but these often sacrifice imaging efficiency or increase operational complexity.[21,22] To remove the influence of carbon contamination at the algorithm level, we propose an end-to-end deep learning approach. The method centres on a modified U-Net convolutional neural network, combined with atomic mask extraction and structural compensation strategies, to automatically learn to map contaminated STEM images to clean, clear atomic images.[19]

Since it is difficult to directly obtain large-scale paired "with/without contamination" STEM images for training, we first construct a simulated carbon contamination model based on physical mechanisms to generate training samples.[2] Hugenschmidt *et al.* reported that carbon contamination mainly originates from electron-beam-induced deposition of hydrocarbon compounds, whose thickness grows with exposure time and beam intensity, appearing in images as an increasingly strong grey haze and slight blurring.[8] Based on this principle, we generate synthetic data using the following model: from a large-area STEM image, we extract a high-quality region without contamination as the "clean" ground truth, then simulate contamination to obtain a "contaminated" image. The effect of contamination can be approximately decomposed into a multiplicative attenuation and a scattering blur: multiplicative attenuation means the increased

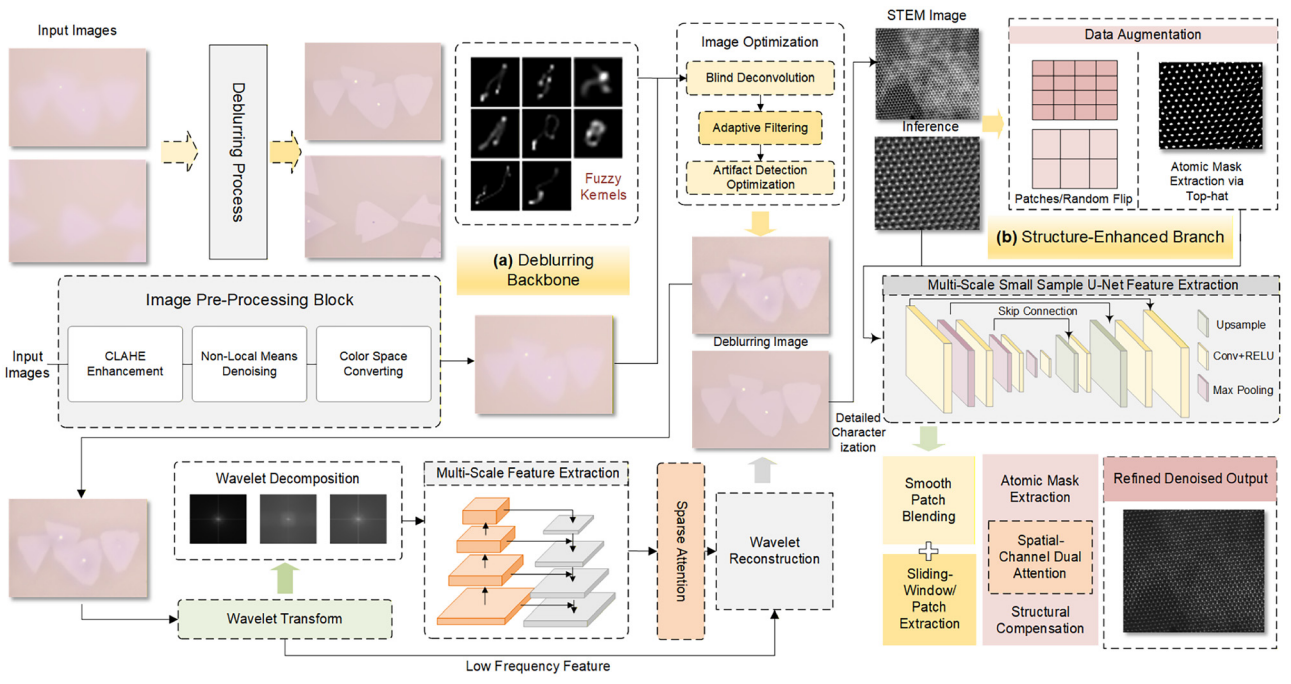

Fig. 1 CrossMicroNet: a cross-scale small-sample image restoration framework for 2D material microscopy imaging. (a) The optical microscopy (OM) branch integrates image preprocessing, sparse blind deconvolution, wavelet-domain multiscale feature extraction and spatial-channel attention to suppress blur and enhance mesoscale domain-boundary information. (b) The scanning transmission electron microscopy (STEM) branch couples front-end sharpening with few-shot multiscale U-Net restoration, atomic-mask guidance and smooth fusion to remove carbon-contamination-induced haze while preserving atomic-lattice features.

thickness of the carbon film causes the transmitted signal to be attenuated by a factor $e^{-\mu t}$ (where μ is a material-dependent attenuation coefficient and t is the local carbon thickness); scattering blur means uneven thickness causes electron scattering that introduces local blurring. An alternative construction approach involves extracting contamination-free regions from the original images, performing data augmentation operations on these segments, and subsequently stitching them together. Due to the structural consistency in the atomic arrangements of Molybdenum (Mo) and Sulfur (S) atoms, this method enables the generation of reference data with more realistic representations. We superimpose these two effects, formulating:

$$I_{\text{contaminated}} = I_{\text{clean}} \times M(x,y) \otimes h_{\text{c}} + \eta \quad (8)$$

where $M(x,y)$ is a simulated contamination mask (range 0–1), which can be represented by a slowly varying low-frequency random field to model the carbon thickness distribution; h_{c} is a blur kernel modeling contamination-induced scattering, which can be represented by a small Gaussian kernel (its scale based on the estimated scattering extent); and η is detection (shot) noise. By adjusting the shape of $M(x,y)$ and the radius of h_{c}, we can generate a series of synthetic images with varying levels of contamination. These synthetic input/target pairs (contaminated *vs.* clean) form the training set and enable the neural network to learn contamination removal. Of course, real contamination can be more complex than our simulation (*e.g.*, thickness growth may be non-linear over time),[23] but the above model captures the main characteristics and is effective for pre-training. We also collected some real STEM sequences where earlier frames are clean and later frames become heavily contaminated, and we extracted corresponding regions to augment training diversity.

We chose U-Net as the base network architecture because its symmetric encoder–decoder structure with skip connections is well-suited for tasks that require precise pixel localization as well as overall structural reconstruction.[19,20] A standard U-Net consists of a down sampling path (encoder) that progressively extracts abstract features, and an up sampling path (decoder) that progressively recovers image resolution, with skip connections at each corresponding scale passing fine detail features from encoder to decoder.[20] This design allows the network to capture both global context and local details, facilitating accurate restoration of atomic-scale images. In our method, the U-Net encoder uses 4 levels of convolution + pooling, progressively compressing the input image (of size *e.g.* 256 × 256) into deep features of size 1/16 × 1/16 (in the bottleneck), as shown in Fig. 1b. The bottleneck layer contains the most compact representation; here we incorporate multiple parallel dilated convolutions to expand the receptive field, so the network perceives the broad haze background of the contamination. The decoder then up samples step by step back to the original resolution, and at each decoding layer the features are concatenated with the corresponding encoder layer's features *via* skip connections, to fuse the fine spatial information from lower layers.[24,25] These skip connections ensure that no tiny atomic detail or positional information from the original image is lost during decoding.[20] As emphasized in U-Net++ and other variants, well-designed skip connections are critical for precise reconstruction; we also introduce an improvement to the skip connections: before merging, we pass the encoder features through an attention gate (an SE channel attention module) to filter them, retaining only the components relevant to the decontamination task, then fuse them with the decoder features. This can be viewed as an optimization of the original U-Net skip structure, removing redundant information that might distract the decoder.

Aside from attention, we also introduce residual blocks into U-Net to stabilize training of deep features. Each convolutional layer in the encoder and decoder uses a ResNet-style residual unit, which mitigates gradient vanishing and accelerates convergence. Meanwhile, to improve contrast, we add a per-pixel adaptive gain layer before the final output: a learnable sigmoid mapping that stretches the brightness of the network output, ensuring that atomic spots in the output have sufficiently enhanced intensity. Overall, our improved U-Net has: (a) strong detail-preserving ability (skip connections + attention); (b) sufficient receptive field to recognize large-scale contamination backgrounds (dilated convolutions); (c) stable and efficient training performance (residual units). The network's final output has the same size as the input and is a largely decontaminated image.

Atomic top-hat mask extraction and structural compensation

Although the deep network can restore the image to a great extent under contamination, in areas of severe contamination some atoms may still have insufficient contrast or even appear as "missing" (*i.e.* certain atomic spots that should be bright become dark or vanish in the image).[18] To ensure that every atomic site in the restored image is correctly recovered, we devised a structural compensation strategy based on an atomic mask.

First, we utilize prior physical knowledge to extract an atomic position mask from the input contaminated image. In 2D materials (*e.g.* crystalline films), the atoms appear as periodically arranged bright spots in STEM images, which are clearly discernible when uncontaminated. Even under contamination, the remaining atomic signal still has certain abrupt features locally. We use a Laplacian of Gaussian (LoG) filter or adaptive thresholding to detect atomic centres: apply bandpass filtering to extract high-frequency components of the image, then perform peak finding to locate a series of point coordinates; or directly apply adaptive thresholding to the image to mark points significantly brighter than their surroundings as atoms.[13] To reduce false detections, we can utilize prior lattice parameters (if the material's lattice constant is known) to match detected points to ideal lattice positions and discard spurious points that do not fit the periodic lattice. After this step, we obtain a binary mask with 1s at positions where atoms should be present and 0 elsewhere. This atomic mask represents all atomic locations that we expect to appear bright in the output image.

With the mask in hand, we apply slightly different strategies during the network's inference stage and training stage to ensure completeness of the atomic structure: we feed the atomic mask as an additional input channel to the network, so that the network is explicitly informed of where atoms ought to be.[26,27] Studies have shown that incorporating prior guidance can enhance network reconstruction of structural information.[9] In our implementation, we stack the original image with its mask to form a two-channel input, so the network can "refer to" the mask to focus on restoring regions where atoms are located. In the loss function, we additionally include a weighted loss term for these atomic mask regions, giving higher weight to restoration errors at those locations, forcing the network to more precisely recover each atom's intensity. Moreover, we introduce an atomic-level contrast loss: for each mask position, we require that in the restored image that point has sufficiently higher intensity than its surrounding background. We define a loss such as:

$$L_{\text{atom}} = \max\{0, I_{\text{bg}} + \Delta - I_{\text{atom}}\} \tag{9}$$

where I_{bg} is the average background intensity in a neighbourhood around an atomic point, and Δ is a desired contrast threshold. By minimizing this loss, the network is driven to produce each atomic pixel brighter than its surroundings.

For the trained model's output, we further apply a post-processing compensation using the mask. If certain mask locations remain dim in the output, we perform local structure completion: for example, inferring that position's intensity from neighbouring regions or superimposing an ideal atomic peak. One approach is: compute the value at each atomic mask location in the output; if it is below some factor of the neighbourhood average, mark it as under-restored. Then add an intensity increment at that point equal to a standard lattice peak (or simply replace it with the average of its four nearest similar atoms). This ensures that every expected atomic site in the output image appears as a distinguishable bright spot, maintaining a complete lattice structure. Therefore, above approach enables the removal of carbon contamination while preserving the intrinsic atomic structure integrity.

It should be noted that obtaining the atomic mask in practice does not require additional prior knowledge – with just a single contaminated image, we can algorithmically infer the likely atomic distribution. Therefore, this strategy does not violate the end-to-end philosophy, but rather embeds an interpretable prior into the network. This combination of deep learning with rule-based algorithms increases the model's sensitivity to critical structures and prevents the network from missing certain atoms due to training data bias.

With the above network architecture and compensation strategy in place, we need to consider how to fuse the network output with the original information during inference to obtain a natural decontaminated image, and how to design the training loss to effectively constrain the network learning.

Although our network is very focused on restoring atomic regions, over-sharpening the homogeneous background areas could introduce unnatural textures. To ensure the output image has sharp atoms and smooth background, we adopt a smooth fusion strategy: we blend the network output with the original contaminated image using a spatially adaptive weight. Specifically, we compute a local variance or gradient magnitude map of the input image, then normalize it to serve as a weight mask. This mask takes values near 1 in the vicinity of atoms (high variance) and near 0 in flat background regions (low variance). The final output is defined as:

$$I_{\text{final}}(x) = W(x) I_{\text{net}}(x) + [1 - W(x)] I_{\text{orig}}(x) \tag{10}$$

where I_{net} is the network output and I_{orig} is the original input image, and $W(x)[0, 1]$ is the fusion weight map (adaptively computed from the original image). In this way, at the locations requiring sharp atomic features, we almost entirely use the network restoration, whereas in uniform areas that need no change, we preserve the original image, avoiding any artificial texturing from the network. The fusion boundary transitions smoothly thanks to the continuous variation of W, yielding a seamless composite with no obvious stitching artifacts. This smooth fusion inference effectively combines the advantages of AI restoration with the fidelity of unaltered regions, ensuring the overall image looks natural.

We train the decontamination network with a compound loss function to account for both pixel accuracy and perceptual quality. The basic component is mean squared error (MSE) loss, $L_{\text{MSE}} = \frac{1}{N}\sum (I_{\text{out}} - I_{\text{gt}})^2$, which directly measures the pixel-wise deviation of the restored image from the ground-truth clean image. However, using only MSE tends to produce overly smooth results lacking sharp details.[28] Therefore, we introduce a Structural Similarity Index (SSIM) loss as a complement. SSIM captures the similarity in luminance, contrast, and structure between images, with a range [0, 1] (closer to 1 indicates the two images are more indistinguishable). We convert SSIM into a loss form as $L_{\text{SSIM}} = 1 - \text{SSIM}$ (plus possible regularization terms). In training, we weight the losses with coefficients $\alpha \approx 0.8$ and $\beta \approx 0.2$ to emphasize the importance of SSIM. This hybrid loss has been shown in GANs, face super-resolution, and other tasks to produce images of better perceptual quality.[29] In our task, it drives the network not only to minimize pixel error but also to improve structural similarity, ensuring the atomic arrangement and edge sharpness are as close as possible to the ground truth. In effect, maximizing structural similarity is equivalent to minimizing LSSIM. The final total loss is the weighted sum of the two:

$$L_{\text{total}} = \alpha L_{\text{SSIM}} + \beta \tag{11}$$

(in addition to the aforementioned atomic contrast loss L_{atom} for mask regions). With thorough training (we use the Adam optimizer, initial learning rate 1×10^{-4}, ~50 epochs), the model converges stably, achieving high SSIM and low MSE on validation data in the end. It should be noted that we also experimented with perceptual loss (LPIPS based on VGG features) and adversarial loss (GAN discriminator) to further improve realism, but considering training stability and con-

trollability for practical use, we ultimately adopted the simpler yet effective combination above. Experiments showed that the MSE + SSIM combination alone was sufficient to produce visually nearly contamination-free results while avoiding excessive artifacts.

In summary, this method succeeds in learning to remove carbon contamination from STEM images by integrating physical priors (atomic mask) into a deep network (U-Net) and using a tailored training strategy. It can greatly improve the SNR and contrast of contaminated images while preserving the integrity of the atomic lattice, thereby ensuring subsequent precise structural measurements and defect analyses are feasible.

Datasets, data partitioning and implementation

The OM dataset consisted of 28 OM/SEM region pairs acquired during chemical-vapour-deposition growth of triangular 2D domains. The OM images were recorded through a long-working-distance microscope at approximately 0.3 μm per pixel. However, due to the high-temperature airflow of the tube furnace, solid particles adhere to the inside of the *in situ* tube wall, resulting in a significant decrease in resolution. Because the OM branch is unsupervised, no paired pixel-level optical ground truth was used for optimisation. Instead, 12 images were used during development of the processing sequence, 6 images were used for parameter checking, and 10 registered OM/SEM pairs were held out as a blind evaluation set. SEM images were never used to tune the blind-deconvolution parameters; they served only as approximate structural references during evaluation.

For the STEM branch, 280 synthetic contaminated/clean patch pairs were generated for training and 60 for validation from clean lattice images. A further 24 real contaminated patches extracted from seven short STEM sequences were used for few-shot fine-tuning, while 8 real patches were reserved for validation. The held-out STEM test set contained 20 crops in total: 10 synthetic and 10 reales. The synthetic subset enabled full-reference metrics, whereas the real subset was used for qualitative analysis, line profiles and FFT inspection. This hybrid strategy was chosen because large paired real STEM datasets with identical same-field clean references are rarely available under contamination conditions.

All computations were performed in Python. The OM front end used NumPy, OpenCV, PyWavelets and PyTorch, and the STEM back end was implemented in PyTorch 2.2.2 with CUDA 11.8. Runtimes were measured on a workstation equipped with an Intel Core i7-class CPU, 32 GB system memory and an NVIDIA GeForce RTX 4060 GPU. The OM pipeline required 2.5 ± 0.4 s per 1280 × 720 frame in the mixed CPU/GPU implementation. STEM inference required 0.020 ± 0.003 s per 256 × 256 patch and 0.96 ± 0.11 s for a 4096 × 4096 mosaic including overlap fusion.

Evaluation protocol and comparison baselines

Four groups of metrics were used. For OM, the primary sharpness indicator was the apparent edge-transition width, w_{edge} (μm), extracted by fitting the edge spread function across manually selected grain boundaries to an error-function model. Lower w_{edge} indicates a steeper transition and thus a sharper reconstructed boundary. Because the OM reference is cross-modal, PSNR and LPIPS were calculated after affine registration to SEM and conversion to structural maps, and should therefore be interpreted as approximate structural-consistency indices rather than strict optical ground-truth fidelity metrics.[30] NIQE was additionally reported as a no-reference naturalness indicator.[31]

For STEM, PSNR, SSIM, LPIPS, VIF and FSIM were calculated on the synthetic same-field subset. Higher PSNR, SSIM, VIF and FSIM indicate better recovery, whereas lower LPIPS indicates a smaller perceptual discrepancy from the clean reference. In addition to learning-based baselines, three conventional processing routes were evaluated: Gaussian band-pass filtering,[32] Fourier-mask filtering[33] and Wiener deconvolution.[34] A sequence-based registration-and-averaging baseline inspired by SmartAlign[35] was also examined for short STEM movies, although such methods are not directly applicable to isolated single contaminated frames.

The comparison baselines were chosen to span several methodological families. In OM, Anime4K,[36] CUGAN,[37] ESRGAN,[38] SRMD[39] and Waifu2x[40] cover lightweight edge enhancement, GAN-based super-resolution, degradation-aware restoration and denoising-oriented upscaling, while FFT-ReLU provides a more classical sparse-prior deblurring perspective.[41] In STEM, DCP[42] represents a hand-crafted prior, AOD-Net[43] and GridDehazeNet[44] represent early and mid-generation CNN dehazing designs, FFA-Net[45] introduces explicit attention and DehazeFormer adds transformer-based feature modelling.[46] The conventional baselines in Table 3 provide an essential point of reference because they correspond to processing strategies that are already familiar to microscopy users and can often be applied without network training.

The evaluation framework was designed to separate image-physics questions from optimisation questions. OM metrics answer whether restored boundaries become sharper and more structurally aligned with SEM-derived topology, not whether the optical image becomes a literal SEM surrogate. STEM metrics answer whether the same-field synthetic lattice is better recovered after contamination removal and whether the cleaned image remains useful on real data. By preserving this distinction, the reported values in the tables can be interpreted scientifically rather than only computationally.

Results and discussion

Optical microscopy restoration

Fig. 2 and Table 1 summaries the OM results. The most important trend is the collapse of the apparent edge-transition width from the nearly micron-scale regime of the baseline methods to 0.22 μm for CrossMicroNet. Among the comparison methods, ESRGAN gives the smallest baseline w_{edge} at 0.80 μm, followed by Waifu2x at 0.85 μm and Anime4K at 0.93 μm.

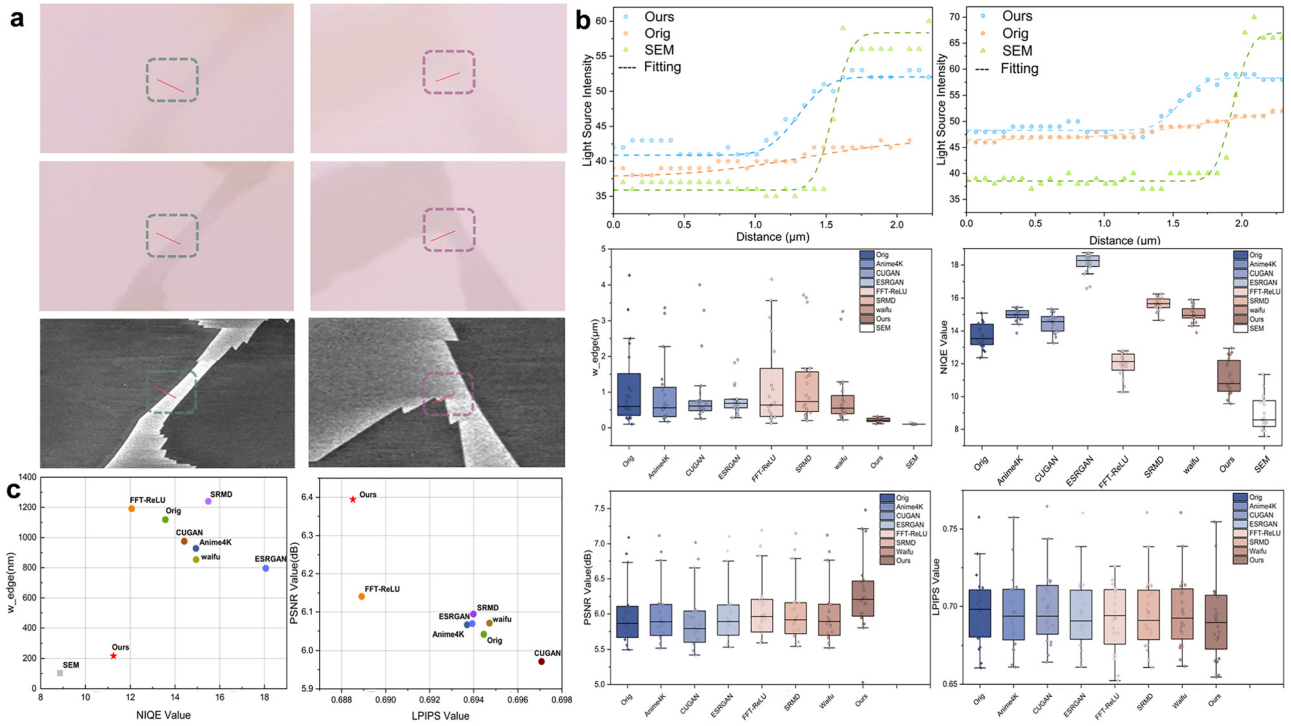

Fig. 2 Optical-microscopy results. (a) Representative OM examples before processing, after CrossMicroNet restoration and the corresponding registered SEM structural references; local enlarged views are included to highlight domain boundaries. (b) Representative line-profile fitting results and box plots of the evaluation metrics. (c) Scatter plots summarising the trade-off between naturalness and structural sharpness. For NIQE and LPIPS, lower values are better; for PSNR, higher values are better; for w_{edge}, lower values indicate a sharper edge transition.

CrossMicroNet therefore reduces w_{edge} by approximately 72.5% relative to the strongest baseline in this metric and by more than a factor of four relative to FFT-ReLU and SRMD. This is not a minor cosmetic improvement. In domain-monitoring applications, such a contraction in transition width directly improves the localization of grain edges, the separation of closely spaced triangular domains and the reliability of subsequent contour extraction.

The relative ordering of the OM baselines is also physically informative. Anime4K and Waifu2x are effective edge-oriented upscaling or denoising systems, and ESRGAN benefits from adversarial trained super-resolution priors. These methods can produce visually sharper boundaries, but they are not designed around the blur physics and illumination in microscopy. Consequently, they recover some local edge contrast while leaving broad transition tails. FFT-ReLU performs better in NIQE than several deep baselines because its frequency-domain sparse prior suppresses aggressive artefacts, but its w_{edge} remains large at 1.19 μm, indicating that natural-looking output alone is not sufficient for boundary localization. SRMD is degradation-aware and flexible, yet its best operating point in these data still leaves a broad transition width of 1.24 μm. CrossMicroNet differs because it attacks the actual sources of OM degradation in sequence: local illumination imbalance is flattened first, noise is stabilized second, blur is explicitly inverted third, and only then are residual high-frequency features selectively enhanced in the wavelet domain.

The same trend is reflected by the approximate structural-reference metrics. CrossMicroNet yields the best LPIPS value in Table 1, 0.68, whereas the baselines occupy the 0.68–0.71 range. Although the absolute values must be interpreted cautiously because the reference is SEM-derived rather than optically acquired ground truth, the direction of change is meaningful. Lower LPIPS here indicates that the enhanced OM image is more consistent with the large-scale structural layout seen in the registered SEM map. In other words, the restoration does not merely sharpen edges according to an image prior; it sharpens them in a way that better matches the true domain topology.

A similar conclusion is supported by PSNR. The numerical spread is compressed because cross-modal structural normalization reduces grayscale differences to coarse topology differences, yet CrossMicroNet still reaches the highest value, 6.39 dB. The strongest baseline is FFT-ReLU at 6.14 dB, so the gain is +0.25 dB. This increase is modest in absolute terms but notable given the deliberately conservative evaluation framework. Since the PSNR calculation is based on OM/SEM structural maps rather than same-modality pixel fidelity, even a few tenths of a decibel correspond to more consistent recovery of major domain edges and junction geometry.

The no-reference NIQE score further indicates that the gain in sharpness is not purchased at the cost of severe visual degradation. CrossMicroNet yields a NIQE value of 11.25, lower than all baselines, including FFT-ReLU at 12.06. The

improvement relative to the best baseline is about 6.7%. This is significant because several deep super-resolution models can lower apparent blur while increasing haloing or synthetic micro-texture, which would normally worsen NIQE. The combination of blind deconvolution followed by restrained wavelet attention appears to be responsible for this balance. Deconvolution contracts the edge, and the attention stage selectively reinforces the informative sub-bands instead of uniformly sharpening the entire image.

The line-profile analysis in Fig. 2a explains why the table values change as they do. In the raw OM image, the intensity transition across the grain boundary is broad and its inflection region spans multiple pixels, which is the expected behavior for motion-blurred or defocused edges. After CrossMicroNet processing, the transition becomes markedly steeper and more symmetric, approaching the form expected for a locally blurred but structurally correct boundary. This behavior is consistent with the design of the front end. CLAHE alone cannot produce such contraction because it rescales intensities but does not invert blur. NLM alone would in fact smooth the profile further if applied strongly. The blind-deconvolution step performs the dominant contraction, and the wavelet-domain channel-spatial attention recovers the residual fine contrast around the edge corner and along the short triangular sidewalls.

Importantly, the OM branch should still be interpreted as a clarification method rather than a microscope-resolution enhancer in the strict diffraction-limited sense. The instrument resolving power is unchanged. What improves is the image-level sharpness proxy extracted from the recorded data. The value of this distinction is practical: growth scientists often need more accurate segmentation and better temporal tracking of domain morphology, not a claim that the optical system has surpassed its physical limit. Within that practical objective, the performance of CrossMicroNet is strong. The method moves the images from a regime of ambiguous grain boundaries to a regime in which small domains and edge corners are sufficiently clear for downstream morphometric analysis. Another notable result is that the OM front end provides a representation that is useful beyond the OM task itself. Because the same preprocessing, deconvolution and wavelet-attention operations are also applied to STEM data before contamination suppression, the OM results help validate the broader design principle of CrossMicroNet: contrast clarification can be formulated as a shared front-end problem, while modality-specific ambiguity is deferred to the back end only when necessary.

A closer look at the OM scatter plots in Fig. 2c shows why CrossMicroNet occupies a distinct part of the performance space. Methods driven mainly by perceptual enhancement tend to move toward sharper-looking output at the cost of naturalness, whereas conservative filtering methods tend to preserve naturalness while failing to contract the edge sufficiently. CrossMicroNet shifts the result toward lower w_{edge} and lower NIQE simultaneously. That joint movement is difficult to achieve unless the algorithm treats blur removal and detail enhancement as separate but coordinated operations. The result is particularly valuable for scientific imaging, where an algorithm that merely increases local contrast can look impressive in a cropped panel yet still fail to improve the measurements extracted from the image.

It is also useful to consider why CUGAN and ESRGAN do not dominate despite their success in visually oriented restoration tasks. Adversarial trained models are excellent at recovering plausible textures from large natural-image corpora, but microscopy datasets are small, structurally repetitive and often outside the distribution of consumer images. In such settings, adversarial priors can encourage textural plausibility more than measurement fidelity. The CrossMicroNet front end avoids this mismatch by deriving most of the OM improvement from explicit blur inversion and controlled wavelet reinforcement rather than from learned hallucination. This is consistent with the strong NIQE performance and with the lower LPIPS relative to the registered SEM structural maps.

From an application standpoint, the clarified OM output is immediately relevant to automated quantification. Many growth studies depend on thresholding, contour tracing or machine-vision segmentation of domains frame by frame. These downstream tools are highly sensitive to edge broadening and local illumination drift. By narrowing boundary transitions and homogenizing contrast without introducing strong false textures, CrossMicroNet improves the separability between domain and background classes. This makes the method useful not only as a figure-generation tool but also as a preprocessing step for extracting kinetic observables such as nucleation density, anisotropic growth velocity, coalescence timing and island-size distributions from long *in situ* image sequences.

STEM decontamination and structural fidelity

Fig. 3–5 and Tables 2, 3 collectively show that CrossMicroNet does not merely increase apparent contrast, but removes contamination in a structurally conservative manner. On the held-out STEM test set, CrossMicroNet delivers the best performance among all learning-based methods, reaching a PSNR of 20.50 dB, an SSIM of 0.85, an LPIPS of 0.08, a VIF of 0.92 and an FSIM of 0.93. Relative to the strongest learning baseline, DehazeFormer (Transformer based model), this corresponds to gains of +1.00 dB in PSNR, +0.03 in SSIM, a 20% reduction in LPIPS, and further improvements of +0.02 in both VIF and FSIM. The advantage is also maintained against conventional restoration routes: compared with registration plus averaging, which is the strongest non-learning baseline in Table 3, CrossMicroNet improves PSNR by 2.28 dB and SSIM by 0.17, while reducing LPIPS from 0.19 to 0.08. These trends indicate that the method suppresses the contamination field more effectively than either generic dehazing networks or frequency-selective classical filters, while preserving a higher level of lattice-related information.

The qualitative evidence in Fig. 3 is consistent with these quantitative results. In the simulated examples in Fig. 3a, the contaminated input exhibits a diffuse low-frequency veil super-

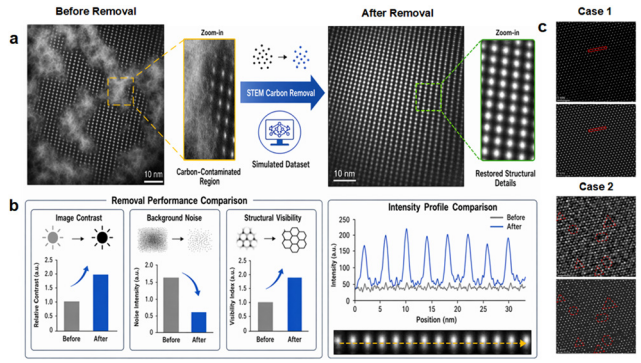

Fig. 3 Schematic overview of STEM decarbonization on simulated and real datasets. (a) Schematic comparison of STEM images before and after decarbonization on a simulated dataset. (b) Summary of the performance on the simulated dataset, showing improved contrast and cleaner intensity profiles after decarbonization. (c) Schematic presentation of two representative cases from real datasets, highlighting improved atomic-feature visibility after contamination removal.

imposed on the atomic lattice, which reduces atomic-column contrast and blurs local periodicity. After restoration, the background becomes substantially cleaner and the atomic features become more distinct, with the recovered image approaching the clean structural target more closely. Fig. 3b further summarizes this trend by combining performance statistics with representative intensity profiles: after decontamination, the profiles show a clearer modulation pattern and a larger peak-to-background separation, indicating that the network does not simply brighten the image globally, but restores local contrast at physically meaningful lattice positions. Importantly, the real-data cases in Fig. 3c show the same qualitative behavior despite the absence of exact same-field clean references: diffuse contamination is attenuated, atomic-feature visibility is improved, and the restored output remains visually consistent with the underlying lattice arrangement rather than introducing an obviously artificial texture.

The superiority of CrossMicroNet can be understood mechanistically from the design of the framework. In contrast to DCP and AOD-Net, whose formulations are rooted in natural-scene haze assumptions, the present model explicitly targets the image characteristics of contaminated STEM data. The shared front-end restoration stage first regularizes local contrast, reduces nuisance fluctuations and mitigates broad blur, thereby preventing the slowly varying contamination field from dominating subsequent feature extraction. The U-Net refinement branch then operates on a more favorable representation and focuses on separating the non-uniform amorphous background from the periodic atomic signal. At this stage, atomic-mask guidance is particularly important because it constrains the network to preserve expected lattice sites rather than treating them as disposable high-frequency components. The smooth fusion strategy further limits over-processing in homogeneous regions, so the final output remains clean but not unnaturally sharpened. This staged division of labor explains why CrossMicroNet outperforms stronger learning baselines such as GridDehazeNet, FFA-Net and DehazeFormer, and also why it exceeds conventional band-pass, Fourier-mask and Wiener approaches that lack comparable spatial adaptivity.

The metric pattern in Tables 2 and 3 also supports this interpretation. PSNR is most sensitive to the overall removal of the contamination field and therefore reflects the net recovery of the clean image. SSIM, FSIM and VIF, by contrast, are more closely tied to structural organization and information-bearing features; their simultaneous improvement indicates that the gain is not achieved by over smoothing. The decrease in LPIPS to 0.08 likewise suggests that the restored output is not only cleaner in a pixelwise sense, but also closer to the clean lattice image in feature space. This consistency across multiple metric families is important for microscopy applications, because an algorithm that only improves one index can still distort atomic contrast. Here, however, all reported indices move in the expected direction, supporting the conclusion that

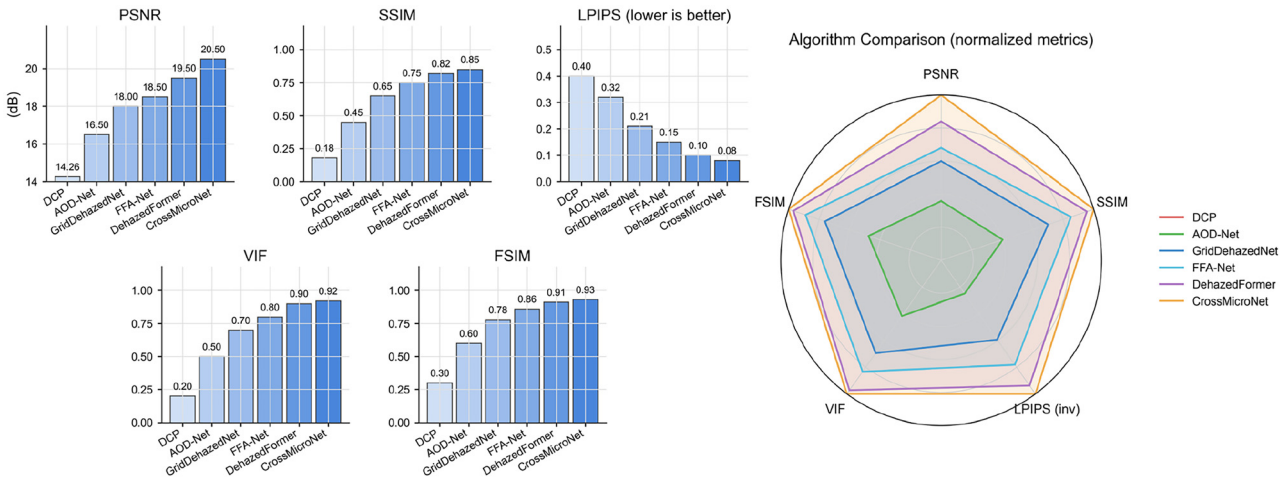

Fig. 4 Comparison of average STEM evaluation metrics for the learning-based baselines. Higher values are better for PSNR, SSIM, VIF and FSIM, whereas lower values are better for LPIPS.

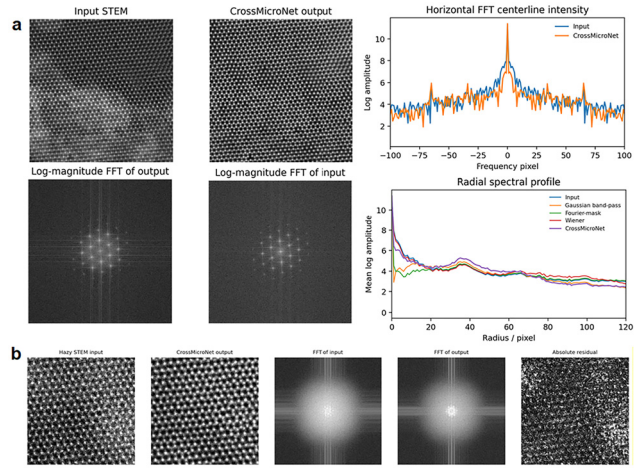

Fig. 5 Real- and reciprocal-space validation of STEM decontamination. (a) Representative STEM crop showing the input, CrossMicroNet output, corresponding log-magnitude FFT maps, FFT-centre intensity profile, and radial spectra. CrossMicroNet suppresses diffuse low-frequency background while preserving the lattice-associated spectral band. (b) Same-field decontamination example with Fourier-domain quality control. The residual highlights processing-induced changes and is used for QC only, not as direct pixel-wise evidence of carbon removal.

contamination suppression and structural preservation are achieved simultaneously.

The reciprocal-space evidence in Fig. 5 provides an additional and physically interpretable validation. In the raw contaminated crop, the log-magnitude FFT is dominated by a strong central diffuse component, which is characteristic of a slowly varying amorphous background in real space. After CrossMicroNet processing, this low-frequency intensity around the Fourier origin is markedly reduced, as also reflected by the horizontal FFT-center line profile. At the same time, the lattice-associated spectral features remain visible in the processed FFT map, indicating that the method does not erase the periodic information responsible for atomic ordering. The radial spectral profiles further clarify the selectivity of this suppression: CrossMicroNet attenuates the low-frequency contamination band more strongly than the input, while retaining the lattice-related spectral band more effectively than Gaussian band-pass, Fourier-mask filtering and Wiener deconvolution. In other words, the method acts neither as a simple global high-pass filter nor as a generic denoiser; rather, it preferentially removes the spectral contribution associated with contamination while preserving the frequencies that encode the underlying lattice.

This point is particularly relevant for scientific use. Beam-induced carbon contamination is expressed primarily as a slowly varying, non-uniform background, whereas the atomic lattice is encoded in more spatially organized mid- to high-frequency content. The FFT behavior in Fig. 5 is therefore fully consistent with the intended mechanism of CrossMicroNet: background suppression in real space corresponds to attenuation of diffuse low-frequency spectral weight, whereas retention of atomic periodicity corresponds to preservation of lattice-associated reciprocal-space features. The fact that both behaviors are observed in the same representative experimental crop strengthens the claim that the restored images remain structurally faithful rather than cosmetically sharpened.

At the same time, the results define the appropriate scope of the method. CrossMicroNet should be regarded as a clarification strategy rather than a source of new structural information. Its strongest benefit is observed when the lattice signal is still present but partially obscured by contamination, in which case the method recovers a substantially cleaner and more interpretable image. Under more severe contamination or in weaker-contrast regions, the restoration becomes correspondingly more conservative, which is preferable to hallucinating nonexistent atomic features. For this reason, the processed image is best interpreted alongside the raw STEM data in situations involving very weak defect contrast or light-element sensitivity. Within these limits, however, the combined real-space and reciprocal-space results show that CrossMicroNet provides a robust and practically useful route for suppressing amorphous carbon contamination while maintaining atomic-scale structural fidelity.

Cross-scale examples and application scope

Fig. 6 illustrates the intended application scope of the framework. In panel (a), the OM examples show that after restoration the boundaries of triangular domains become easier to segment and measure. The improved clarity is particularly valuable for time-resolved growth analysis, where small errors in boundary location accumulate when one extracts velocity, area evolution or anisotropic growth rates from large image sequences. In panel (b), the STEM examples show that the front-end plus back-end combination is most effective in high-magnification, relatively high-SNR regions, where the lattice periodicity is already present but obscured by a strong low-frequency background. For larger fields of view, the method still removes a meaningful fraction of the contamination field, but the restoration is intentionally more conservative.

This behavior makes physical sense. OM and STEM are linked in CrossMicroNet not because they share identical degradation physics, but because they benefit from the same first-stage operations that restore signal organization. Once that common representation has been produced, the downstream use cases diverge naturally. In OM, clarified frames can feed domain segmentation, morphological tracking and growth-kinetics analysis. In STEM, clarified images can improve interpretability during defect inspection, lattice registration and qualitative comparison of contamination states. The key point is that the framework is cross-scale in architecture and workflow, not because it forces the two modalities into a single unrealistic objective metric.

The cross-scale philosophy of the framework also has broader methodological significance. Scientific imaging pipelines are often fragmented into modality-specific scripts with little shared design logic. CrossMicroNet shows that a carefully defined front-end clarification stage can be reused across modalities even when the final interpretation tasks differ. That reuse is not merely convenient for coding; it means that the

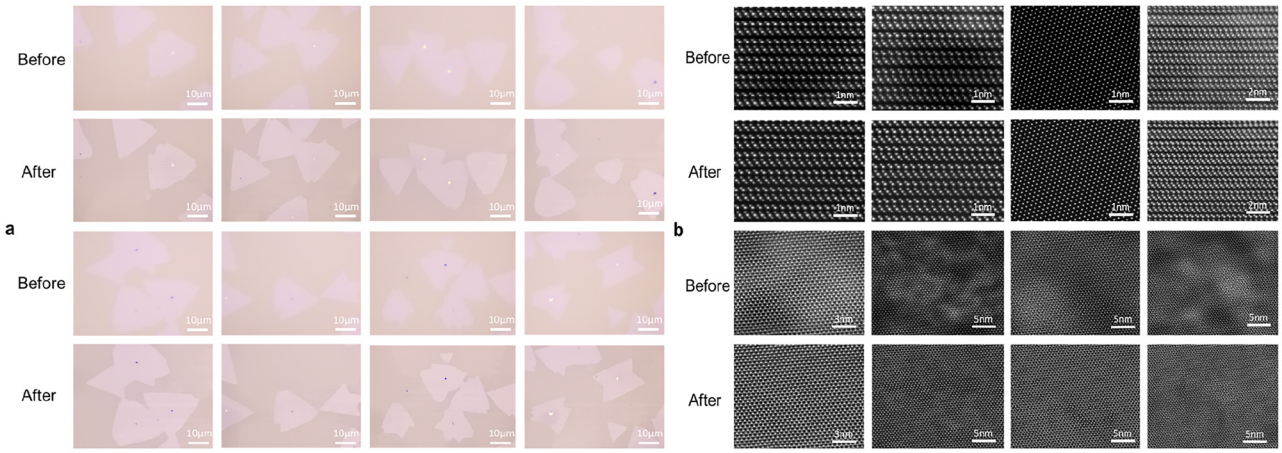

Fig. 6 Additional cross-scale examples. (a) Representative OM frames before and after restoration. (b) Representative STEM images before and after restoration. The framework performs most strongly in high-SNR, high-magnification regions and more conservatively in large-field contaminated STEM images.

algorithmic assumptions about contrast normalization, denoising and sharpness recovery remain consistent from wide-field growth monitoring to atomic-scale inspection. For multi-modal 2D-material studies, such consistency can reduce the gap between process-level and atomistic interpretation.

Limitations, throughput and transferability

CrossMicroNet is robust to moderate motion blur, defocus and carbon contamination, but its limitations should be stated explicitly. First, the OM front end assumes that the effective blur can be reasonably represented within the chosen kernel support. When the blur trajectory becomes very long or strongly non-uniform across the field, blind deconvolution becomes less stable and the benefit of wavelet-domain attention is correspondingly reduced. Second, the STEM back end depends on the presence of recoverable lattice information in the input. If the contamination layer fully overwhelms the periodic signal, no physically responsible method can reconstruct the exact lost contrast from a single image alone.

Third, although the atomic-mask guidance improves structural conservation, it does not completely remove the risk of attenuating weak local contrast differences, especially for lighter atomic species or very subtle point defects embedded in strongly contaminated regions. In practice, this means that the processed image should be interpreted as a clarified companion to the raw image rather than as a replacement for all original information. Fourth, transferability is asymmetric across the two stages: the front-end restoration engine is relatively easy to port to new microscopes or materials because its parameters are tied to generic image-conditioning operations, whereas the STEM back-end benefits from short fine-tuning on representative crops whenever the lattice type, detector response or contamination appearance changes substantially.

The throughput measurements suggest that the framework is suitable for low-latency offline or near-online analysis rather than for integrated microscope-control feedback. A processing time of 2.5 ± 0.4 s per OM frame is compatible with routine batch analysis and with intermittent *in situ* inspection. For STEM, 0.020 ± 0.003 s per 256 × 256 patch and 0.96 ± 0.11 s for a 4096 × 4096 mosaic make the method practical for large images. These timings compare favorably with heavier GAN- or transformer-based approaches and support the use of CrossMicroNet in day-to-day microscopy workflows, provided that its interpretive boundaries are respected.

Conclusions

In this work, CrossMicroNet provides a unified cross-scale framework for clarifying two complementary microscopy modalities used in 2D-material research. The method is organised around a shared front-end clarity engine – CLAHE, restrained denoising, blind deconvolution, ringing suppression and wavelet-domain attention – followed by a STEM-specific few-shot refinement branch with atomic-mask guidance and adaptive fusion. This design allows OM and STEM data to be treated within a coherent restoration strategy while preserving the different physical meanings of their respective outputs.

For OM, the framework contracts broad edge transitions and improves structural consistency with registered SEM references, reducing the apparent edge-transition width to 0.22 μm while maintaining the best NIQE, LPIPS and PSNR values among the tested methods. For STEM, it achieves the strongest performance across all reported metrics on the held-out test set and clearly outperforms both learning-based and conventional baselines. FFT analysis shows that the gain is associated primarily with suppression of the low-frequency amorphous contamination component rather than with wholesale loss of reciprocal-lattice information.

Equally important, the present results define the correct use of the method. CrossMicroNet is a clarification tool, not a claim of new instrument resolution or algorithmically created atomic structure. When used with that interpretation, it pro-

vides a practical and scientifically cautious route to better microscopy readability across the micrometre-to-atomic-scale span relevant to 2D materials. Future work should focus on more explicit protection of weak light-element contrast, broader transfer across materials systems and tighter integration with microscope-side analysis pipelines.

Author contributions

Mingwei Feng: writing – conceptualization, methodology, software, data curation, original draft, and investigation. Xilu Zou: methodology (optical microscopy and STEM experiments), investigation, data curation, validation, visualization. Lei Liu: investigation (sample growth and characterization), resources, data curation, validation. Shengqiang Wu: investigation (STEM measurements), resources, data curation, validation. Haotian Zhang: methodology (deep-learning framework design), formal analysis, validation, visualization. Silin Chen: software, data curation, formal analysis, validation. Zikang Zeng: software, formal analysis, investigation (ablation studies and benchmarking), visualization. Yiru Wang: formal analysis, validation, data curation, visualization. Xiaotian Zhang: resources (*in situ* microscopy systems), investigation (instrument setup and measurements), validation, writing – review & editing. Xuping Zhang: formal analysis and methodology. Taotao Li: resources, supervision, writing – review & editing. Ningmu Zou: supervision, conceptualization, funding acquisition, project administration, writing – review and editing.

Conflicts of interest

The authors declare no conflict of interest.

Data availability

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary information (SI). Supplementary information: experimental details, supplementary figures, parameter settings and Python source code. See DOI: https://doi.org/10.1039/d6nr01702a.

Acknowledgements

This work is supported by the Natural Science Foundation of China under Grant 62341408 and the Fundamental Research Funds for the Central Universities 2024300421.

References

1 K. S. Novoselov, A. Mishchenko, A. Carvalho and A. H. Castro Neto, *Science*, 2016, **353**, aac9439.
2 A. M. van der Zande, P. Y. Huang, D. A. Chenet, T. C. Berkelbach, Y. You, G.-H. Lee, T. F. Heinz, D. R. Reichman, D. A. Muller and J. C. Hone, *Nat. Mater.*, 2013, **12**, 554–561.
3 W. Zhou, X. Zou, S. Najmaei, Z. Liu, Y. Shi, J. Kong, J. Lou, P. M. Ajayan, B. I. Yakobson and J.-C. Idrobo, *Nano Lett.*, 2013, **13**, 2615–2622.
4 X. Zhu, H. Wang, K. Wang and L. Xie, *Nanoscale*, 2023, **15**, 11746–11758.
5 O. L. Krivanek, M. F. Chisholm, V. Nicolosi, T. J. Pennycook, G. J. Corbin, N. Dellby, M. F. Murfitt, C. S. Own, Z. S. Szilagyi, M. P. Oxley, S. T. Pantelides and S. J. Pennycook, *Nature*, 2010, **464**, 571–574.
6 D. Kundur and D. Hatzinakos, *IEEE Signal Process. Mag.*, 1996, **13**, 43–64.
7 K. Zuiderveld, in *Graphics Gems IV*, ed. P. S. Heckbert, Academic Press Professional, San Diego, CA, 1994, pp. 474–485.
8 M. Hugenschmidt, K. Adrion, A. Marx, E. Müller and D. Gerthsen, *Microsc. Microanal.*, 2023, **29**, 219–234.
9 W. H. Richardson, *J. Opt. Soc. Am.*, 1972, **62**, 55–59.
10 L. B. Lucy, *Astron. J.*, 1974, **79**, 745–754.
11 R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis and W. T. Freeman, *ACM Trans. Graph.*, 2006, **25**, 787–794.
12 C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang and W. Shi, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, 4681–4690.
13 Y. Blau and T. Michaeli, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 6228–6237.
14 H. Zhao, O. Gallo, I. Frosio and J. Kautz, *IEEE Trans. Comput. Imaging*, 2017, **3**, 47–57.
15 A. Buades, B. Coll and J.-M. Morel, *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, **2**, 60–65.
16 P. Coupé, P. Yger, S. Prima, P. Hellier, C. Kervrann and C. Barillot, *IEEE Trans. Med. Imaging*, 2008, **27**, 425–441.
17 S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, Academic Press, Burlington, MA, 3rd edn, 2009.
18 S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, *Lect. Notes Comput. Sci.*, 2018, **11211**, 3–19.
19 O. Ronneberger, P. Fischer and T. Brox, *Lect. Notes Comput. Sci.*, 2015, **9351**, 234–241.
20 Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh and J. Liang, *IEEE Trans. Med. Imaging*, 2020, **39**, 1856–1867.
21 R. F. Egerton, P. Li and M. Malac, *Micron*, 2004, **35**, 399–409.
22 T. C. Isabell, P. E. Fischione, C. O'Keefe, M. U. Guruz and V. P. Dravid, *Microsc. Microanal.*, 1999, **5**, 126–135.
23 S. Hettler, M. Dries, P. Hermann, M. Obermair, D. Gerthsen and M. Malac, *Micron*, 2017, **96**, 38–47.
24 K. He, X. Zhang, S. Ren and J. Sun, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, 770–778.
25 G. Huang, Z. Liu, L. van der Maaten and K. Q. Weinberger, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, 4700–4708.
26 J. Schlemper, O. Oktay, M. Schaap, M. P. Heinrich, B. Kainz, B. Glocker and D. Rueckert, *Med. Image Anal.*, 2019, **53**, 197–207.

27 J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020, **42**, 2011–2023.
28 J. Johnson, A. Alahi and L. Fei-Fei, *Lect. Notes Comput. Sci.*, 2016, **9906**, 694–711.
29 Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, *IEEE Trans. Image Process.*, 2004, **13**, 600–612.
30 R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 586–595.
31 A. Mittal, R. Soundararajan and A. C. Bovik, *IEEE Signal Process. Lett.*, 2013, **20**, 209–212.
32 R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Pearson, New York, 4th edn, 2018.
33 J. C. H. Spence, *High-Resolution Electron Microscopy*, Oxford University Press, Oxford, 4th edn, 2013.
34 N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*, MIT Press, Cambridge, MA, 1949.
35 L. Jones, H. Yang, T. J. Pennycook, M. S. J. Marshall, S. Van Aert, N. D. Browning, M. R. Castell and P. D. Nellist, *Adv. Struct. Chem. Imaging*, 2015, **1**, 8.
36 C. Dong, C. C. Loy, K. He and X. Tang, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, **38**, 295–307.
37 X. Wang, L. Xie, C. Dong and Y. Shan, *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021, 1905–1914.
38 X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao and C. C. Loy, *Lect. Notes Comput. Sci.*, 2019, **11133**, 63–79.
39 K. Zhang, W. Zuo and L. Zhang, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 3262–3271.
40 J. Kim, J. K. Lee and K. M. Lee, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, 1646–1654.
41 A. M. Al Radi, P. S. Majumder and M. M. Khan, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025, 3447–3456.
42 K. He, J. Sun and X. Tang, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, **33**, 2341–2353.
43 B. Li, X. Peng, Z. Wang, J. Xu and D. Feng, *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, 4770–4778.
44 X. Liu, Y. Ma, Z. Shi and J. Chen, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, 7314–7323.
45 X. Qin, Z. Wang, Y. Bai, X. Xie and H. Jia, *Proc. AAAI Conf. Artif. Intell.*, 2020, **34**, 11908–11915.
46 Y. Song, Z. He, H. Qian and X. Du, *IEEE Trans. Image Process.*, 2023, **32**, 1927–1941.