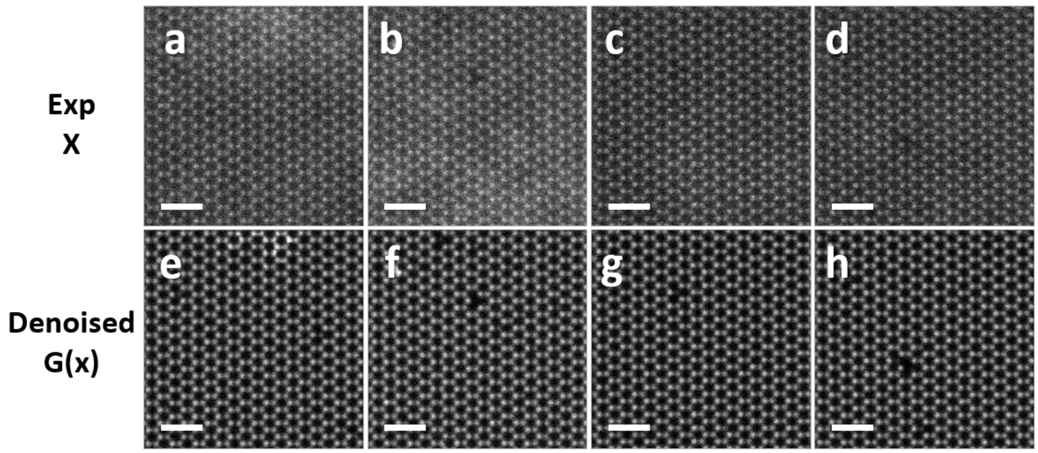
Supplementary Information

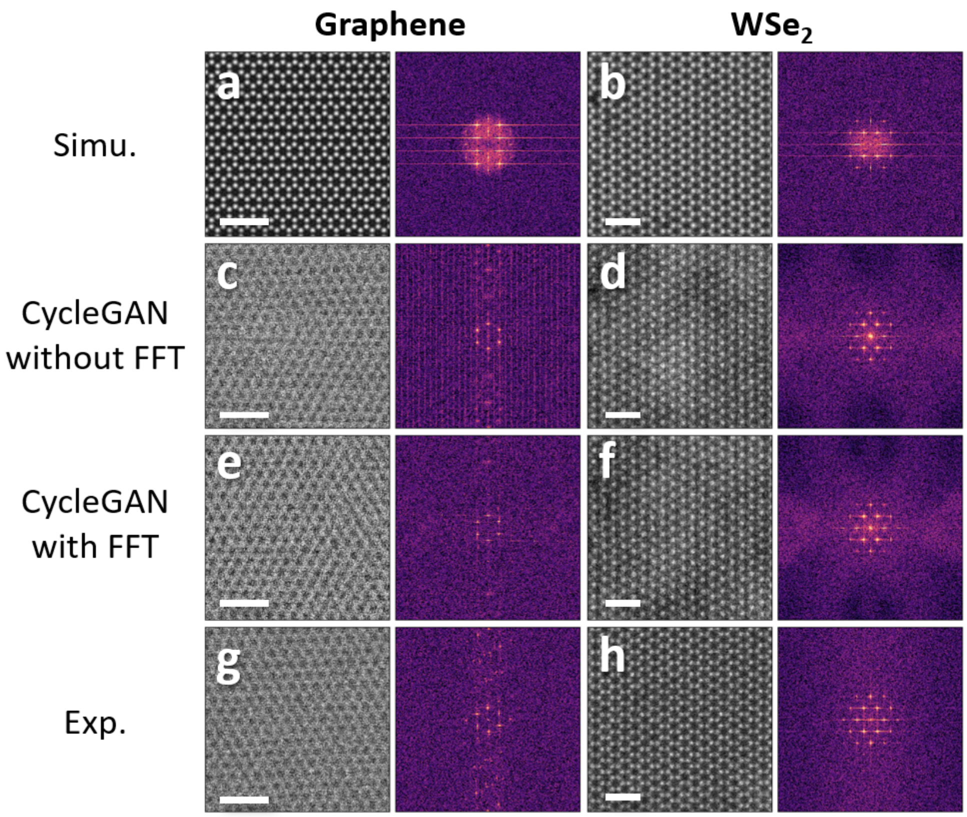
Leveraging Generative Adversarial Networks to Create Realistic Scanning Transmission Electron Microscopy Images

Abid Khan, Chia-Hao Lee, Pinshane Y. Huang,* and Bryan K. Clark*

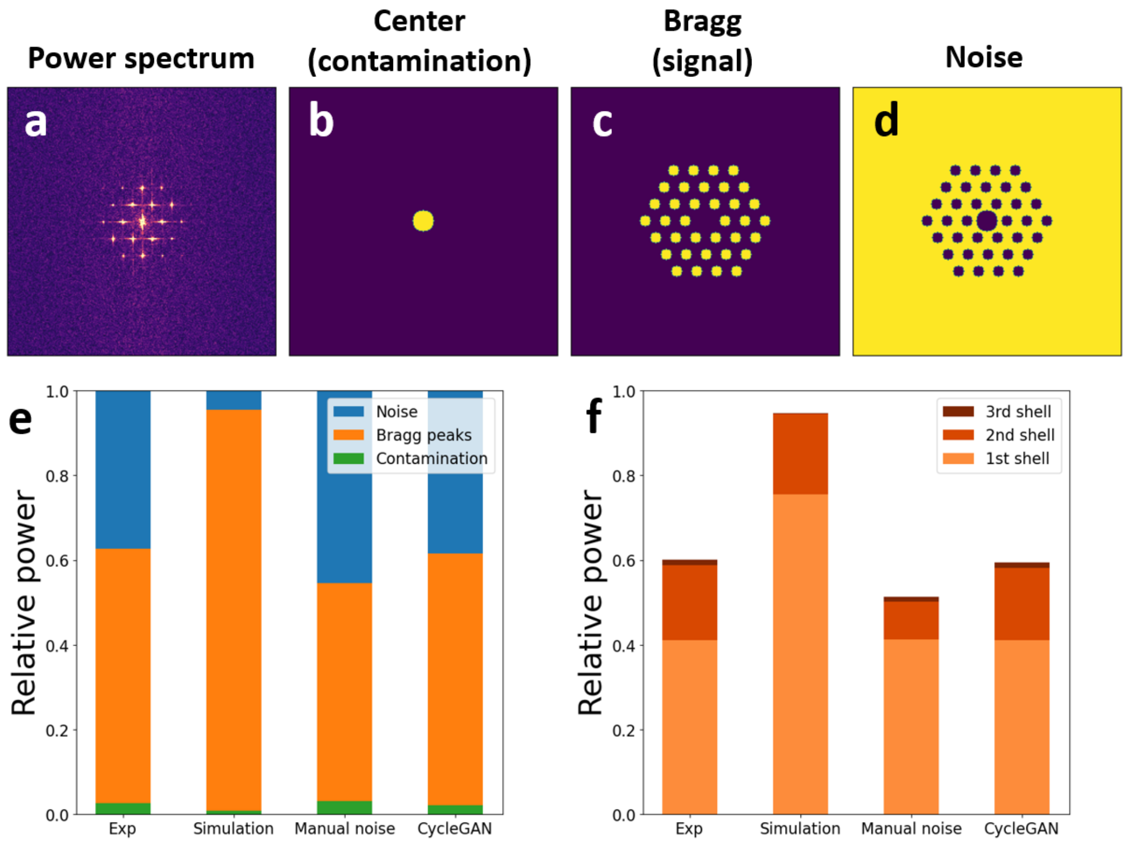
E-mail: pyhuang@illinois.edu; bkclark@illinois.edu



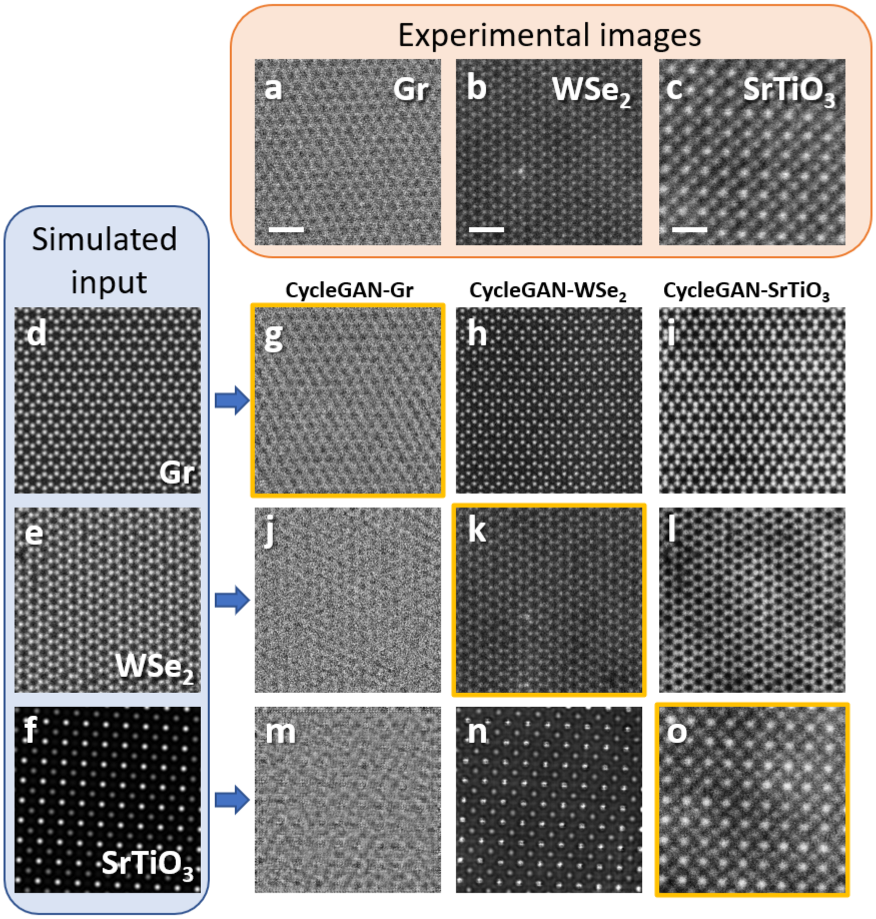
Supplementary Figure 1: Experimental ADF-STEM images of monolayer WSe_2 (a–d) before and (e–h) after denoising by the CycleGAN generator G, where $G : X \to Y$. Note that the denoised images (e–h) still contain small amounts of noise, because Gaussian noise (std = 0.1) are manually added to the out-of-the-box simulated images before CycleGAN training to ensure enough variability in simulation domain Y and the generator G is therefore trying to replicate this Gaussian noise. Scale bars are equal to 1 nm.



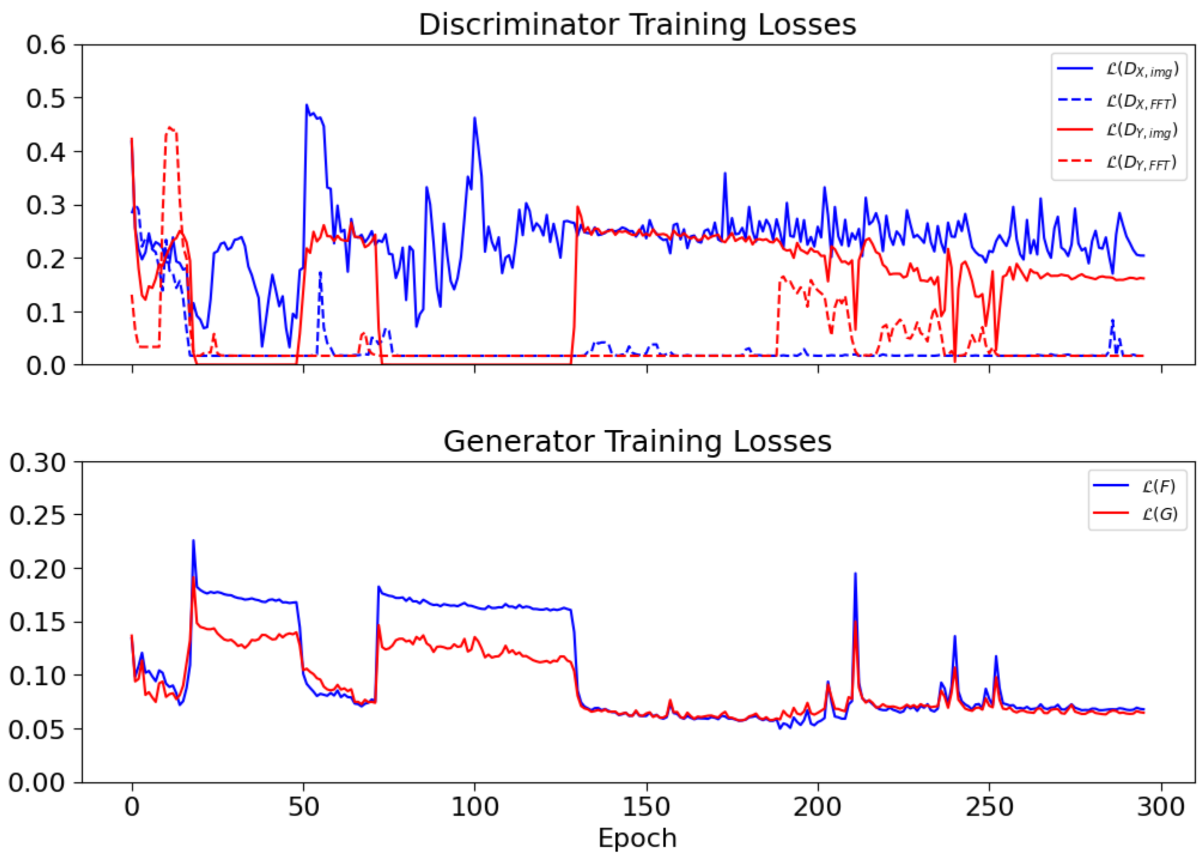
Supplementary Figure 2: Comparison of images of graphene and WSe_2 from (a,b) simulation, (c,d) CycleGAN without FFT discriminator, (e,f) CycleGAN with FFT discriminator, and (g,h) experiment. Each row displays the real space images along with their power spectra. Power spectrum of the image is calculated by the fast Fourier transform (FFT) and displayed in log scale for clarity. While the addition of the FFT discriminators in the CycleGAN may not be easily noticeable in the case of WSe_2, it is critical while processing low signal-to-noise ratio images of graphene. For example, a CycleGAN without FFT discriminator generates graphene images with vertical, streaky artifacts clearly visible in the FFT and to a lesser extent in the real-space image shown in (c). The addition of FFT discriminator significantly reduces such artifacts in (e) and generates images with realistic k-space features. Scale bars are equal to 1 nm.



Supplementary Figure 3: Quantitative comparison of data set similarities using their power spectra. (a) Representative power spectrum of the experimental data set of WSe_2. The intensity is displayed in log scale for clarity. Binary mask for (b) low frequency components (surface contamination), (c) Bragg peaks (signal from crystal structure), and (d) Non-Bragg peaks components (noise). The signal power of each component is integrated within the masked region. (e–f) Bar plots of relative power of each component, while (f) shows the power distribution of individual shells of the 'Bragg peaks' in (e). These data show that the power distribution of the CycleGAN generated data is the best match for the experimental data.



Supplementary Figure 4: Simulated STEM images processed by CycleGANs trained with different materials system. (a–c) Experimental STEM images of graphene, WSe_2, and $SrTiO_3$. Each image has different crystal structure and noise profile. These experimental images provide the targeted 'style' for CycleGAN training. (d–f) Simulated STEM images that are used as the input for the following CycleGAN processing. (g–i) Simulated graphene image being individually processed by CycleGANs trained with different material systems. 'CycleGAN-Gr' denotes the CycleGAN model is trained with experimental and simulated graphene images, so no large geometrical deformation is involved in the domain-to-domain transfer. (j–l) Simulated WSe_2 image being processed by CycleGAN-Gr, CycleGAN-WSe_2, or CycleGAN-$SrTiO_3$, respectively. (m–o) Simulated $SrTiO_3$ image being processed by each CycleGAN. These cross-evaluated images form a 3×3 matrix. The images with matching input and CycleGAN training data (g, k, and o) are marked with yellow outlines. These images are reproduced from Figure 1a–c in the main text. The off-diagonal elements (e.g. unmatch) of this matrix are images processed by CycleGAN trained with different materials system. These images perform poorly as expected, because the features transferred by the CycleGAN, which are the amount and type of noise as well as surface damage and contamination, would typically be different for different material systems. Besides, the translations were trained on dissimilar crystal structures that could have very different domain embeddings. Scale bars are equal to 1 nm.



Supplementary Figure 5: CycleGAN training loss versus number of epochs for experimental training set Day AB. The top plot shows the training losses for the two real-space discriminators $D_{Y,\mathrm{img}}$, $D_{X,\mathrm{img}}$ and the two reciprocal-space discriminators $D_{Y,\mathrm{FFT}}$, $D_{X,\mathrm{FFY}}$. The bottom plot shows training losses for the two generators $G : X \to Y$ and $F : Y \to X$. In the middle of training these losses fluctuate, but reach an equilibrium towards the end as expected since the discriminators and generators are competing against each other.

Supplementary Table 1: FID scores of the 5 different FCN training sets with the 3 FCN test sets A, B, and AB. Smaller FID scores represent a higher similarity with respect to the test sets, hence suggesting better training data. The cells with matching CycleGAN training set and FCN test set are colored in gray. The lowest FID score of each column is bolded.

	FID reference image sets		
FCN training sets	A	B	AB
no noise	31.864	32.256	31.992
manual noise	0.727	1.175	0.930
CycleGAN-A	0.411	1.414	0.990
CycleGAN-B	0.808	1.110	0.914
CycleGAN-AB	**0.361**	**0.722**	**0.499**

Supplementary Table 2: FCN performance metrics evaluated with simulated test sets generated along with each training set. These values are higher than the FCN performance on experimental datasets and are provided for reference because evaluating on simulated data is common in the literature. As shown in this table, the FCN performance on simulated data is quite high, above 96 for each metric.

FCN training sets	P	R	F1
no noise	99.8	99.6	99.7
manual noise	99.2	96.0	97.6
CycleGAN-A	99.5	98.2	98.8
CycleGAN-B	99.5	97.8	98.6
CycleGAN-AB	99.9	98.4	99.1

Supplementary Table 3: FCN performance metrics with training sets processed by different CycleGANs, where each CycleGAN is trained with only 6 experimental 1K-resolution images. CycleGAN-A1 to A3 are the images generated by CycleGANs trained with subsets that are selected from the original Day A data set, while B1 to B3 are generated by CycleGANs trained with subsets from the Day B. Each subset has 6 images that are chosen by the proximity of image acquisition time with respect to the manually labeled test images in each data set. The subset images are selected by acquisition time to emulate the effect of time-dependent microscope instabilities. Images that are acquired around the same time are considered to have similar microscope conditions. The FCN performance are tested with manually labeled experimental test sets (A, B, and AB) and evaluated by precision (P), recall (R), and F1 scores with units in percentage (%). The cells with matching CycleGAN training set and FCN test set are colored in gray; using a matching set between CycleGAN and FCN would be the standard approach for production runs. Unsurprisingly, if you run the CycleGAN and FCN on different sets of data (the non-grayed boxes), which one would not do in practice, the accuracy is much worse. The cells with matching CycleGAN training set and FCN test set are colored in gray. The highest value of each column is bolded.

	FCN test sets								
	A			B			AB		
FCN training sets	P	R	F1	P	R	F1	P	R	F1
CycleGAN-A1	89	90	**89**	22	84	35	32	87	46
CycleGAN-A2	77	**93**	84	10	71	18	18	79	29
CycleGAN-A3	84	92	88	11	46	18	23	64	33
CycleGAN-B1	**99**	40	57	82	81	81	86	65	**74**
CycleGAN-B2	98	28	44	**85**	79	**82**	**87**	59	70
CycleGAN-B3	92	81	86	28	**85**	42	38	**83**	52