A. Dataset Details

Dataset Generation ICSD contains around 258,000 entries that are experimentally verified. We choose only the crystal structures with lattice constants (a, b, c) less than 20 Å,because crystal structures with large lattice constants are relatively rare in nature. This produces a list of $\approx$ 211,000 materials. The CIF files for these selected materials are imported into abTEM, an electron microscopy simulation package, to render the crystal structure into 3-dimensional (3D) datacube sized (16, 100, 100) with the voxel sampling rate of (1.6 Å, 0.2 Å, 0.2 Å) along depth, height, and width dimensions. This anisotropic voxel sampling is chosen to match with the electron ptychography resolution because typically the depth resolution is much poorer compared to lateral (height and width) resolution. We also randomly remove 1% of the atoms to emulate vacancies that are commonly observed in real materials, this enhance the applicability of the generative model although we did not incorporate the vacancy-induced strain to the crystal. The dataset is further augmented 3 times by orienting crystals along the 3 major a-, b-, and c-axis, resulting in a full dataset of $\approx$ 633,000 materials. The rendered 3D datacubes are converted from atomic electrostatic potential (volts) to phase change angle (radians) by taking the angle of the complex function $\mathbf{O}(\mathbf{r}) \approx exp(i\sigma_e V(\mathbf{r}))$, so $\mathbf{x} := \sigma_e V(\mathbf{r}) = angle(\mathbf{O}(\mathbf{r}))$, which is consistent with our electron ptychography forward model. The value range of such phase change angle is bounded between [-π, π] but commonly within [0, 1] because both σ_e and $V(\mathbf{r})$ are larger than 0 and $\sigma_e V$ is usually less than 1 radian given our voxel sampling.

Training Set Augmentation During training, we first apply random in-plane rotation (depth dimension being the rotating axis), and then randomly crop a sub-cube sized (8, 64, 64). This produces a training datacube that spans across (12.8 Å, 12.8 Å, 12.8 Å) in space and covers a couple repeating units cells.

Diffraction Pattern Simulation For MEP-DIFFUSION sampling, we simulate diffraction patterns from test set materials using abTEM. The diffraction patterns are simulated with optical parameters comparable to actual experimental conditions, including 300 kV acceleration voltage for electrons, 21.4 semi-convergence angle, 200 Å overfocus, 500 nm spherical aberration, 0.512 Å scan step size, 26 $\times$ 26 scan patterns, and electron dose of 10^6 e^-/Å^2. The diffraction patterns are simulated with maximum collection angle of 2.5 Å^{-1} and resampled to 128 by 128 pixels. We did not include partial coherence or phonon vibration because the effect will be quite limited at these experimental conditions.

B. Data Rescaling

We demonstrate how data rescaling can be interpreted as up-weighting the "x"-prediction component in the "v"-prediction task. Consider the ground truth $\mathbf{v} = \alpha_t \boldsymbol{\epsilon} - \sigma_t \mathbf{x}$ and network prediction $\mathbf{v}_\theta(\mathbf{z}_t, t)$. The original "v"-prediction objective can be decomposed as:

$$\begin{aligned}\min_\theta ||\mathbf{v} - \mathbf{v}_\theta||_2^2 &= \min_\theta ||(\alpha_t \boldsymbol{\epsilon} - \sigma_t \mathbf{x}) - (\alpha_t \boldsymbol{\epsilon}_\theta - \sigma_t \mathbf{x}_\theta)||_2^2 \\ &= \min_\theta ||\alpha_t \boldsymbol{\epsilon} - \sigma_t \mathbf{x} - \alpha_t \boldsymbol{\epsilon}_\theta + \sigma_t \mathbf{x}_\theta||_2^2 \\ &= \min_\theta ||\alpha_t (\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta) + \sigma_t (\mathbf{x}_\theta - \mathbf{x})||_2^2\end{aligned}$$

When we rescale the data by a factor c such that $\mathbf{x}' = c\mathbf{x}$, the new objective becomes:

$$\min_\theta ||\mathbf{v}' - \mathbf{v}'_\theta||_2^2 = \min_\theta ||\alpha_t (\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta) + c\sigma_t (\mathbf{x}_\theta - \mathbf{x})||_2^2.$$

This shows that the "x"-prediction component is up-weighted by a factor of c in the revised objective.

C. PtyRAD Solver for Ptychographic Reconstruction

We use PtyRAD [26], an open-source ptychographic reconstruction package, to implement iterative gradient descent algorithms for computing gradients required by DPS and to reconstruct baselines with Adam and L-BFGS optimizers. PtyRAD leverages PyTorch's automatic differentiation engine to efficiently compute gradients of optimizable tensors. For all experiments, we first learn a single fixed probe using Adam, and hold it constant for all reconstruction methods using the physical forward model because the test data are simulated with the same probe condition. The probe is fit on 32 examples from our validation set. We minimize the *electron ptychography reconstruction objective*, except we minimize both the 32 atomic structures and the probe simultaneously. For PtyRAD[Adam] and PtyRAD[L-BFGS] baselines, we adopt the mini-batch update scheme and use a batch size of 32 diffraction patterns for each update step. A learning rate of 5e-4 was used for Adam,

while L-BFGS was ran three times with the following learning rates: 1, 1e-1, 1e-2. We select the best result from L-BFGS according to the loss as our baseline. The Adam baseline is reconstructed with 200 iterations, while the L-BFGS baseline is reconstructed with 5 iterations because it converges faster. Note that each iteration in the Adam baseline corresponds to a full pass of all 676 diffraction patterns per example, while in the L-BFGS baseline each iteration is done by evaluating 20 randomly chosen mini-batches to get the estimation of the Hessian matrix. The total number of diffraction patterns seen by each optimizer for each iteration is roughly the same.

D. Network Architecture Details

The UNet architecture is composed of five macro structures, illustrated in Figure 8.

- Stage 1 Down: Residual Block, Residual Block, 2D Downsampling/Convolutional Block
- Stage 2 Down: Residual Block, Residual Block, Attention, 3D Downsampling/Convolutional Block
- Bottleneck: Residual Block, Attention, Residual Block
- Stage 2 Up: 3D Upsampling/Convolutional Block, Residual Block, Residual Block, Attention
- Stage 1 Up: 2D Upsampling/Convolutional Block, Residual Block, Residual Block

The architecture only downsamples or upsamples the spatial dimensions when the channel count changes. Specifically, we downsample when increasing channel dimensions and upsample when decreasing them. Each down structure maintains two connections to its corresponding up structure. The channel progression for each stage is as follows:

- Stage 1 Down: 16, 16, 32, 32, 64, 64
- Stage 2 Down: 128, 128, 128, 128, 256, 256, 256, 256
- Bottleneck: 256
- Stage 2 Up: 256, 256, 256, 256, 128, 128, 128, 128
- Stage 1 Up: 64, 64, 32, 32, 16, 16

The network's spatial transformation is significant: after Stage 1 Down, the input tensor of $8 \times 64 \times 64$ is compressed to $8 \times 8 \times 8$, with voxel sampling evolving from (1.6 Å, 0.2 Å, 0.2 Å) to (1.6 Å, 1.6 Å, 1.6 Å). Stage 2 Down performs uniform downsampling across all dimensions, with Stage 1 Up ultimately restoring the original tensor size and sampling rate. The model has 100M total parameters.

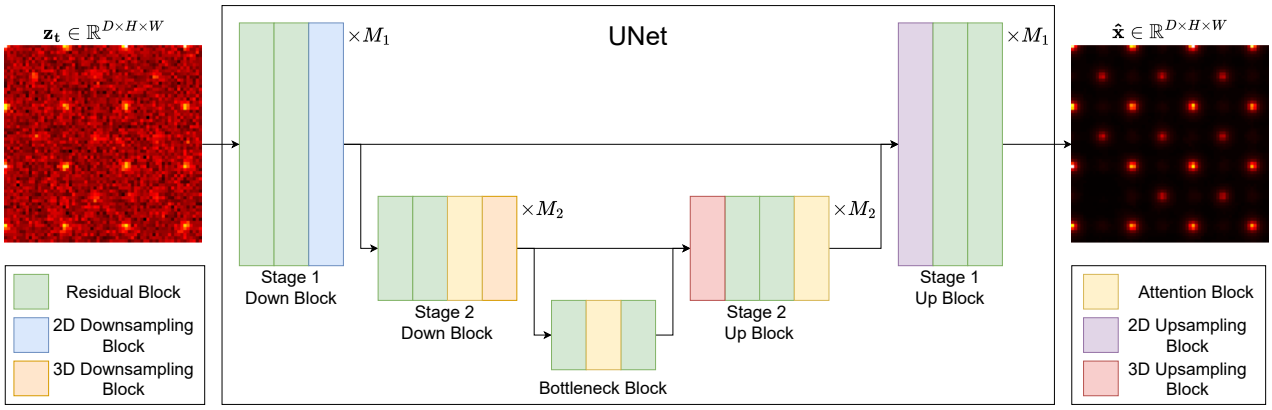

Figure 8. 3D UNet architecture with anisotropic processing. The network first downsamples along height and width to match the coarser spatial sampling along depth, then performs downsampling across all dimensions. Self-attention layers are applied in low resolution blocks.

E. Hyperparameters

See Table 3 for training hyperparameters, and Table 4 for sampling hyperparameters.

Hyperparameter	Setting
Optimizer	AdamW
Batch Size	256
Learning Rate	0.0001
Weight Decay	0.01
λ_1	-20
λ_0	20
$w(\lambda_t)$	$\frac{\mathcal{N}(\lambda_t;-7,3)}{Z}$

Table 3. Table of Training Hyperparameters.

Hyperparameter	Setting
Sampler	SDE DPMSolver++
Schedule	PolyExponential
Min Sigma	0.1
Max Sigma	800
ρ	1.0
$g(\lambda_t)$	$5000 * \text{sigmoid}(4 - \lambda_t)^{1/2}$

Table 4. Table of Sampling Hyperparameters.

F. Periodicity and logSNR

One of the critical adjustments required to fit a diffusion model on this periodic data was to tune our loss weighting to focus the model on very low logSNR regions. This is because as soon as the faintest crystal emerges the model immediately can recognize it. In order to unconditionally sample periodic crystals with our model we need the network to be capable of hallucinating periodicity from Gaussian noise and therefore we focus training on the region where the periodicity is beginning to emerge. We provide an example of the trained network predicting the crystal from diffusion latents at different logSNR in Figure 3.

G. Additional Qualitative Comparisons

We include additional qualitative comparisons in Figure 9 and Figure 10. We do this primarily to illustrate the limitations discussed in section 6. In Figure 9, subfigure (A) shows a near perfect reconstruction, subfigure (B) shows a generation where the structure is blurred, subfigure (C) shows a generation where many slices are correct but misordered, subfigure (D) shows a generation with incorrect structure. In Figure 10, all subfigures show partially correct structures of varying degrees.

H. Comparison of diffraction patterns

We include additional qualitative comparisons of diffraction patterns in Figure 11. The simulated diffraction patterns are calculated by feeding the reconstructed crystals (generated by MEP-DIFFUSION) into the physical forward model. We observe good agreement between ground truth and simulated patterns, indicating that the reconstructed crystals are similar with the ground truth crystal structures.

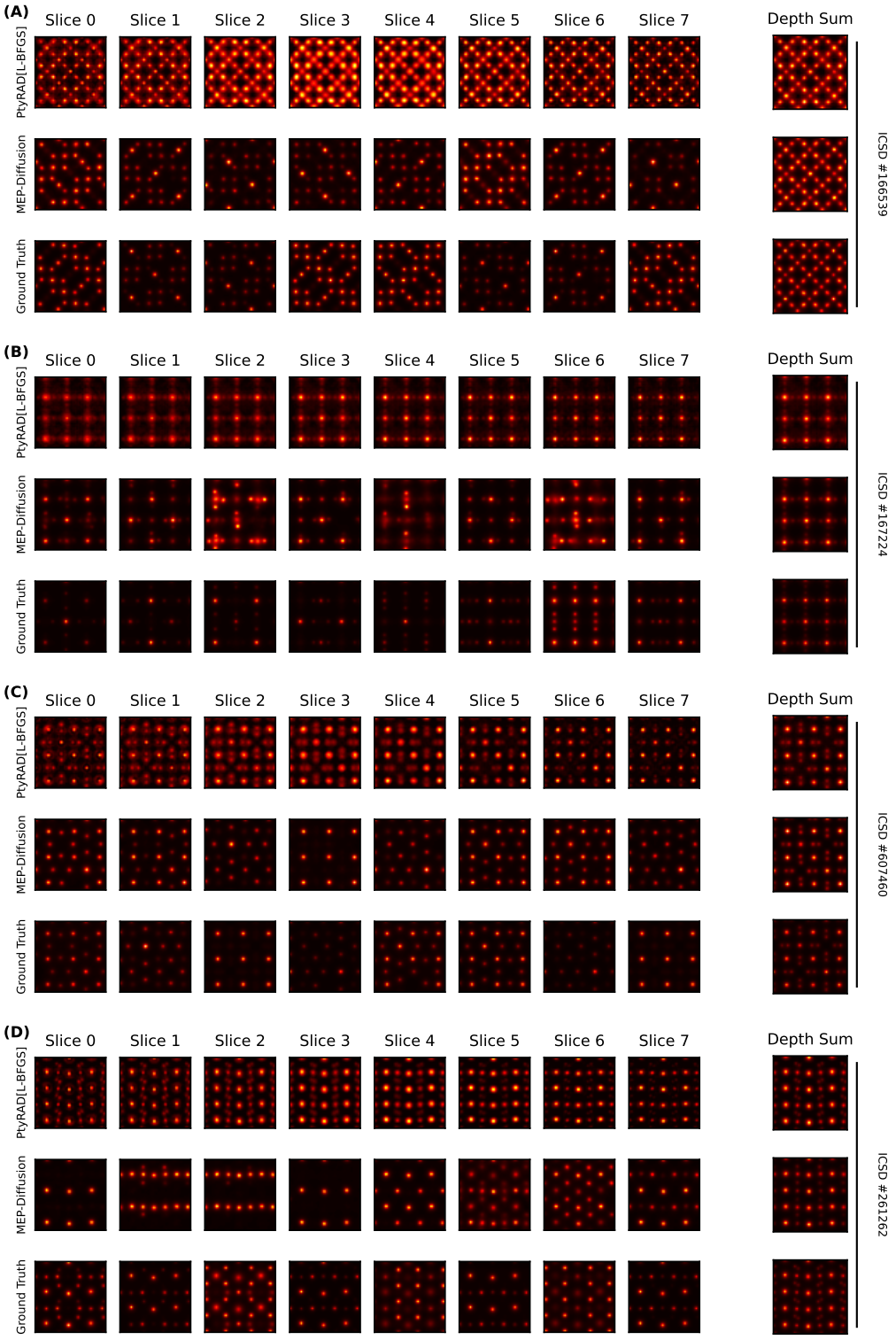

Figure 9. Additional qualitative comparisons. The slice indices denote the crystal structure at specific depth. The slice thickness is 1.6 Å and the image width is 12.8 Å.

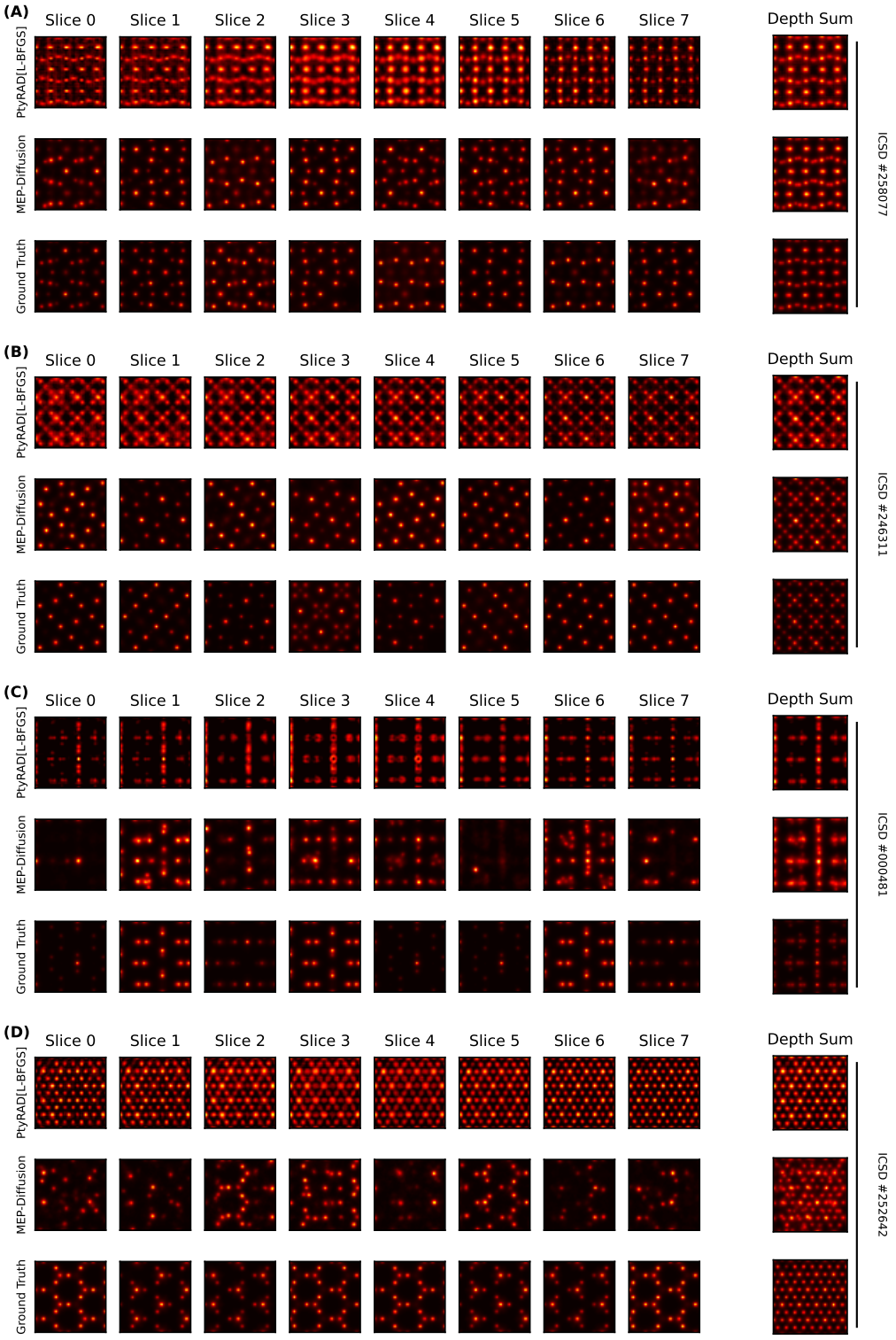

Figure 10. Additional qualitative comparisons. The slice indices denote the crystal structure at specific depth. The slice thickness is 1.6 Å and the image width is 12.8 Å.

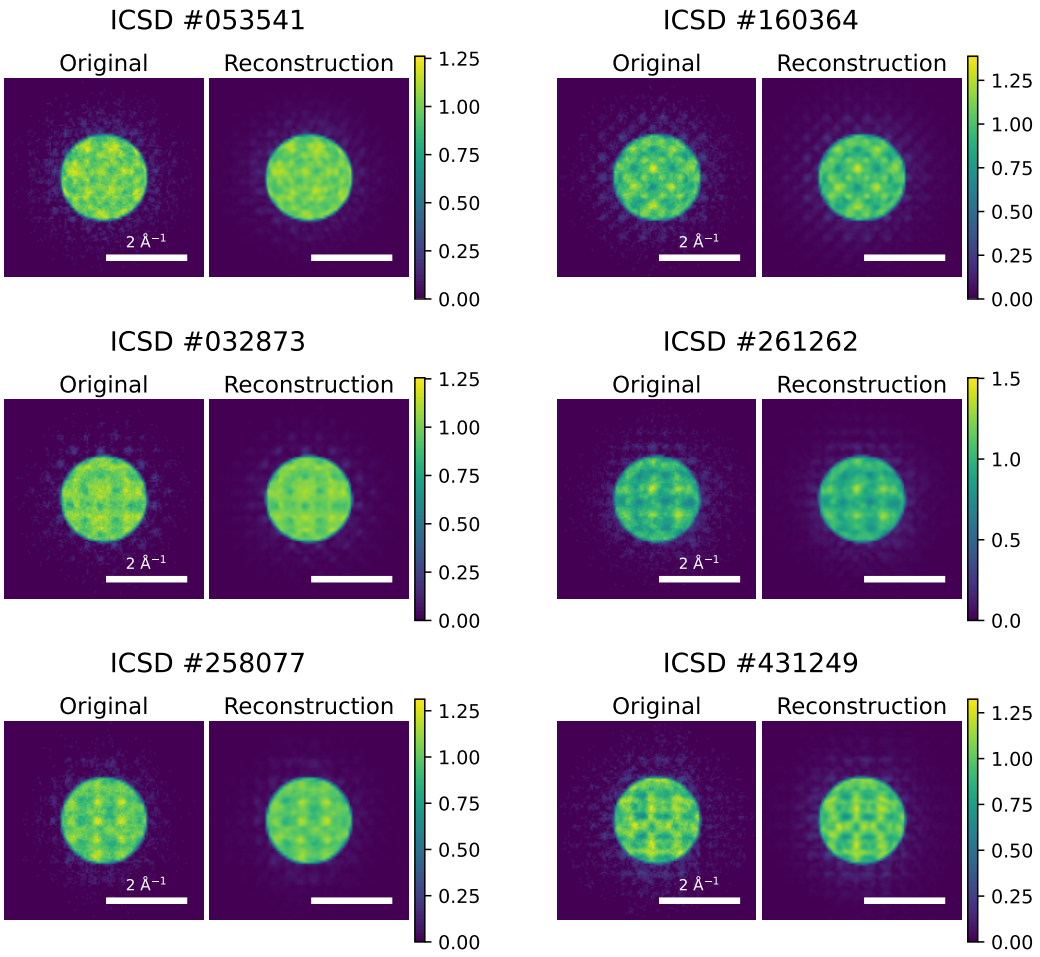

Figure 11. Ground truth and simulated diffraction patterns from reconstructed crystals using MEP-DIFFUSION. The corresponding crystals are shown in Figure 1. We take the square root of the diffraction pattern intensity for better visualization.