Supplementary Material - Implementation Details

The architecture is based on the work of Glow. Our adaptations are as follows:

- Input Size: $16 \times 16 \times 3$ (RGB patches sampled from Places dataset)
- Batch Size: 256
- L (number of flows): 1
- K (number of transformations in a flow): 32
- Channels: 512
- Optimizer: Adam
  - Learning Rate: 1e-4
  - Betas: (0.9, 0.999)
- No use of LU decomposition

During training, we randomly sample a batch of images from the dataset and first resize the images such that the smaller edge of the image is of size 256 pixels. Then, we randomly crop each image to a single 16×16 patch, leading to the same batch size of patches. We also apply random horizontal flip.

For image manipulation, we extract all overlapping patches (sliding window) of the source image and input them in batches to the forward pass of the network. After performing the gradient step on the latent variable, we input it (same batch size) to the reverse pass of the network. Then, the image is composed such that every pixel is the mean of its values in all the patches it appears in.

Our code is available at `https://github.com/Eladhi/VI-Glow`.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.