MINER: Multiscale Implicit Neural Representation (supplementary)

Vishwanath Saragadam, Jasper Tan, Guha Balakrishnan,
Richard G. Baraniuk, Ashok Veeraraghavan

Rice University, Houston TX 77005, USA
vishwanath.saragadam@rice.edu

1 MINER Algorithm Details

Algorithm 1 shows the overall MINER flow with initialization, pruning, and parameter update. The function `DOWNSAMPLE` implements a domain-specific downsampling operator. For images it is,

$$\mathcal{D}_j(I) = I_j(\mathbf{x}) = \int_0^{2^j/H} \int_0^{2^j/W} \mathrm{I}(2^j\mathbf{x} + d\mathbf{x}) dx dy. \tag{1}$$

For three-dimensional signals such as videos $(I(x, y, t))$ and 3D occupancy volumes (I(x, y, z)) we have,

$$\mathcal{D}_j(I) = I_j(\mathbf{x}) = \int_0^{2^j/H} \int_0^{2^j/W} \int_0^{2^j/T} \mathrm{I}(2^j\mathbf{x} + d\mathbf{x}) dx dy dz. \tag{2}$$

2 Experimental Results

Baselines. We compare MINER against three competing baselines for 2D images, and three for 3D volumes.

1. SIREN [5] fits a single large MLP at a single scale and utilizes a sinusoidal activation function for accelerated training. We varied the number of hidden units for each experiment to ensure that the number of parameters matched that of MINER.
2. KiloNeRF [4] fits multiple small MLPs at a single scale instead of a single large MLP. The number of hidden units for each MLP was chosen to be the same as that for MINER.
3. ACORN [2] fits a single large MLP at a single scale with adaptive coordinate decomposition.
4. Convolutional occupancy network [3] utilizes convolutions to capture local correlations. We used this only for 3D volume comparisons. While one of the key strengths of the convolutional occupancy network is its ability to generalize to unseen data points, we only compare it with respect to its capability of overfitting to a single training data point.

Algorithm 1 MINER algorithm.

```
Require: $I(\mathbf{x})$, number of scales $J$, block size $b$, number of features per layer $N_{\text{feat}}$,
  number of layers $N_{\text{layers}}$
  for $j = J-1, J-2, \dots, 0$ do                              ▷ Loop over spatial scales
      if $j$ is $J-1$ then
          $R_j(\mathbf{x}) \leftarrow \text{DOWNSAMPLE}(I(\mathbf{x}), 1/2^{J-1})$
      end if
      $Q_j \leftarrow \frac{HW}{2^j b^2}$                                ▷ Number of blocks
      $A_j \leftarrow \{1, 2, \dots, Q_j\}$                              ▷ Active set
      for $q = 1, 2, \dots, Q_j$ do                                  ▷ Loop over blocks
          if $||R_j(\mathbf{x}^q)|| \leq \tau_j$ then
              $A_j \leftarrow A_j \backslash q$                          ▷ Remove converged blocks
          else
              $\mathcal{N}_j^q \leftarrow \text{MLP}(N_{\text{feat}}, N_{\text{layers}})$           ▷ MLPs for each block
          end if
      end for
      for $i = 1, 2, \dots, N_{\text{iter}}$ do
          for $q$ in $A_j$ do
              $\widehat{I}(\mathbf{x}^q) \leftarrow \mathcal{N}_j^q(\mathbf{x}^q)$                  ▷ Compute MLP output
              $\epsilon_j^q \leftarrow ||\widehat{I}(\mathbf{x}^q) - I(\mathbf{x}^q)||^2$          ▷ Compute MSE loss
              Backpropagate $\epsilon_j^q$ to update $\theta_j^q$
              if $\epsilon_j^q \leq \tau_j$ then
                  $A_j \leftarrow A_j \backslash q$                      ▷ Prune converged blocks
              end if
          end for
      end for
  end for
```

5. Screened Poisson Surface Reconstruction (SPSR) [1] utilizes local normals to construct the mesh. SPSR does not utilize a neural network but requires additional information in the form of normals at each voxel.

We used code from the respective authors and optimized the training parameters to ensure a fair comparison.

Fitting 3D point clouds. Figure 1 visualizes the meshes fit with various reconstruction approaches for a fixed duration. MINER has superior reconstruction quality compared to ACORN and convolutional occupancy networks [3], and comparable performance to Screened Poisson surface reconstruction (SPSR) [1]. We do note that the quality of reconstruction specifically for the engine model is superior to SPSR. Since the engine model has a large number of sharp edges, SPSR tends to oversmooth the result. Since MINER combined with marching cubes relied only on local information for reconstruction, the resultant mesh was more accurate.

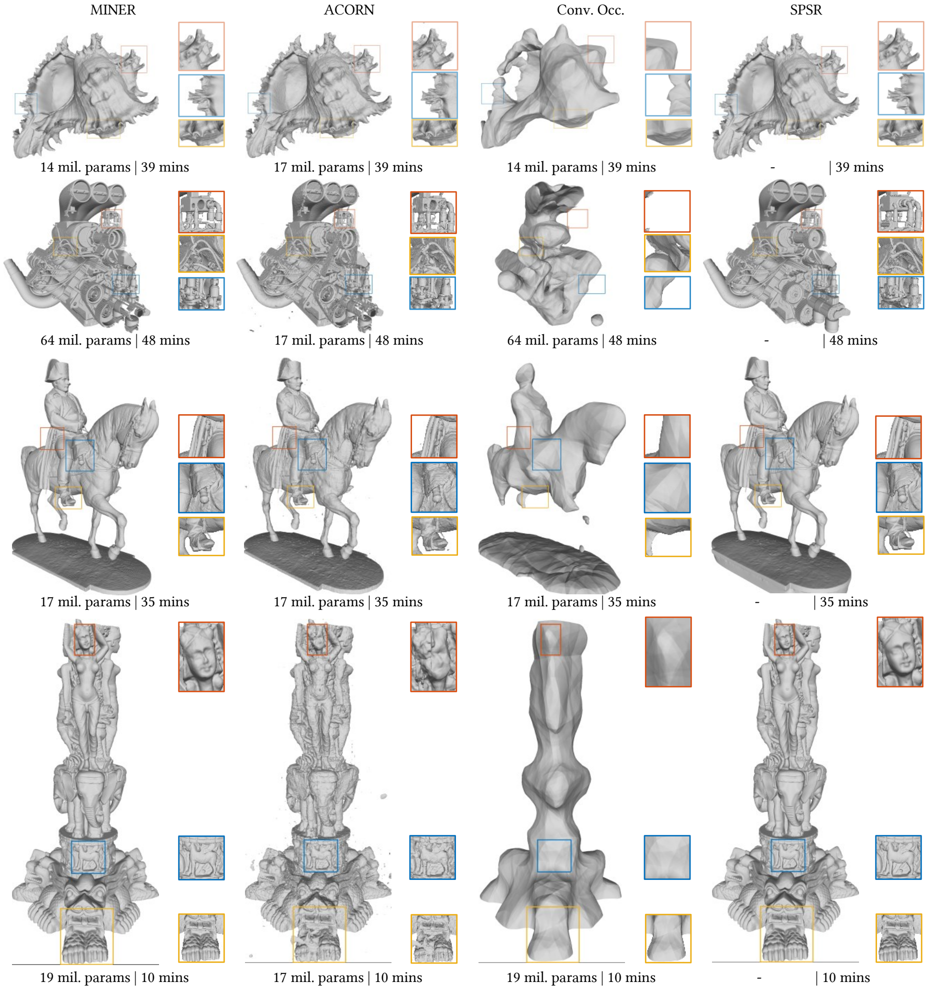

Fig. 1: **Comparisons against state-of-the-art for 3D volume fitting.** The figure compares 3D occupancy fitting for a **fixed duration** with MINER, ACORN, Convolutional occupancy [3], and screening poisson reconstruction [1]. The number of parameters of MINER was chosen automatically according to model complexity. MINER achieves high accuracy in a very short duration for arbitrarily complex shapes, which is not possible with prior works, even though some models such as the engine (second row) require significantly larger number of parameters.

2.1 Analysis of parameter space

The training time of MINER is affected by the number of scales, the size of each patch, the stopping criteria when switching to a finer scale, and the parameters of each MLP including the number of layers, the number of features per layer, and the type of non-linearity. We now provide a thorough analysis of the parameters.

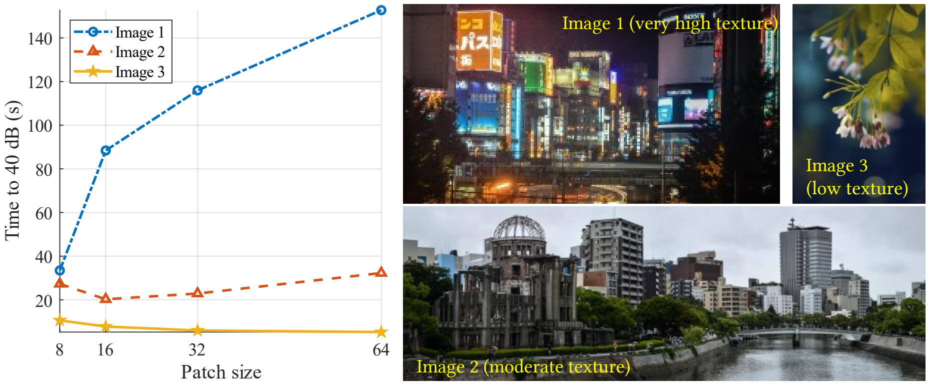

Fig. 2: **Effect of patch size.** The plot shows the time taken to achieve 40 dB for three different images. We notice that the optimal patch size is highly dependent on the image. For images with high texture content (image 1), a smaller patch size of 8 pixels is optimal. For medium texture with large flat areas (image 2), a medium patch size of 32 pixels is optimal. For images such as macro photography (image 3) which has strong low frequency content, larger patch size of 64 is optimal.

Performance with varying patch size. The optimal patch size is highly dependent on the signal itself. To understand the empirical relationship, we fit three types of images with low, medium, and high texture content using MINER to achieve 40dB accuracy. In each case, we varied the patch size from 8 pixels to 64 pixels. We proportionally increased the number of features per layer for each patch to keep the total number of parameters approximately the same. Figure 2 shows the plot of time taken to achieve 40 dB as a function of patch size for the three images. We notice that the optimal patch size for least training time increases with reducing texture content – which can be used as a guideline when choosing the appropriate patch size.

Performance with number of scales. Figure 3 shows the time taken to achieve 40dB for various images with varying number of spatial scales. The optimal number of scales is strongly dependent on the texture content – highly textured images benefit from fewer scales, while images with low texture benefit from larger scales. In practice, we found 3 - 4 scales sufficed for optimal results in terms of training time and number of parameters.

Effect of stopping threshold. The number of parameters, and ultimately the training time are effected by the threshold at which each spatial block is terminated. Fig. 4 compares the convergence time to 40dB, and number of parameters for varying thresholds for fitting a 4MP Pluto image. The trends are as expected – increase the block threshold terminates blocks at the state of each scale, which leads to fewer parameters. However, this may adversely affect convergence time, fewer blocks are required to achieve the targeted MSE. We found that a threshold

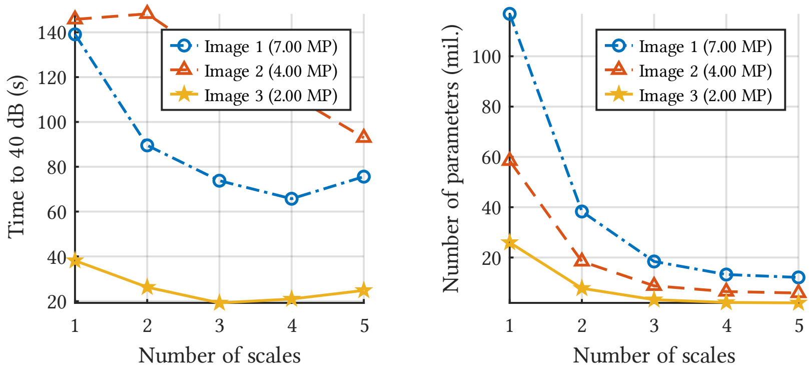

Fig. 3: **Effect of number of scales.** The plots show the time taken to achieve 40dB, and the total number of parameters for the three images shown in Fig. 2. The optimal number of scales is image-dependent, but we found 4 scales to work well for most images.

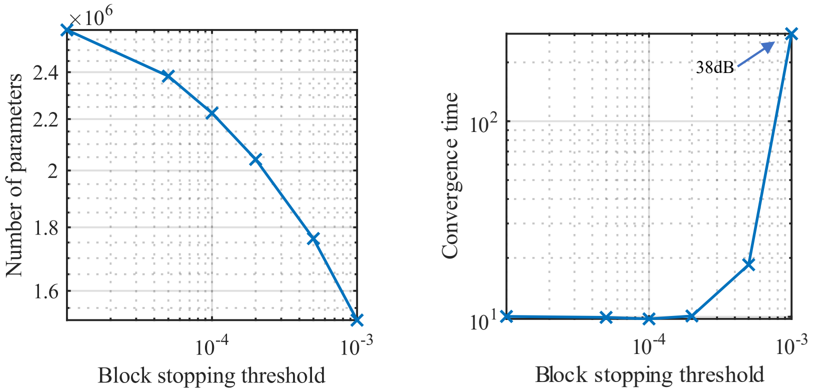

Fig. 4: **Effect of block stopping threshold.** The plots show the time taken to achieve 40dB, and the total number of parameters for a 4MP Pluto image. Increasing the stopping threshold reduces the number of parameters, but leads to increased convergence time. For very high threshold, sufficient blocks may not be active to achieve 40dB.

of 0.1 to 2× the desired MSE enabled best results in terms of convergence times and number of parameters.

References

1. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. ACM Trans. Graphics **32**(3), 1–13 (2013)
2. Martel, J.N., Lindell, D.B., Lin, C.Z., Chan, E.R., Monteiro, M., Wetzstein, G.: Acorn: Adaptive coordinate networks for neural scene representation. arXiv preprint arXiv:2105.02788 (2021)

3. Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., Geiger, A.: Convolutional occupancy networks. In: IEEE European Conf. Computer Vision (ECCV) (2020)
4. Reiser, C., Peng, S., Liao, Y., Geiger, A.: Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. arXiv preprint arXiv:2103.13744 (2021)
5. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. Adv. Neural Info. Processing Systems (2020)