Supplementary Material for "A Deep Generative Deconvolutional Image Model"

Yunchen Pu Duke University

Xin Yuan Bell Labs

Andrew Stevens Duke University

Chunyuan Li Duke University

Lawrence Carin Duke University

A More Results

A.1 Generated images with random weights

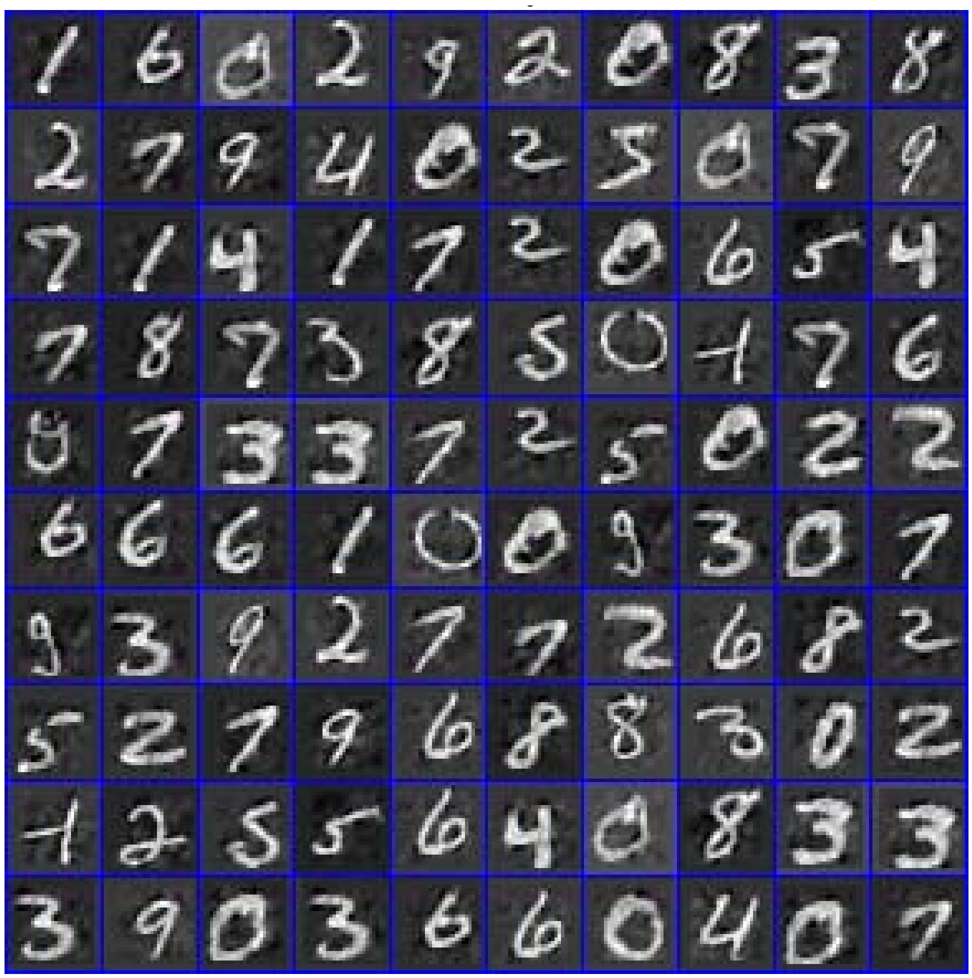

Figure 1: Generated images from the dictionaries trained from MNIST with random dictionary weights at the top of the two-layer model.

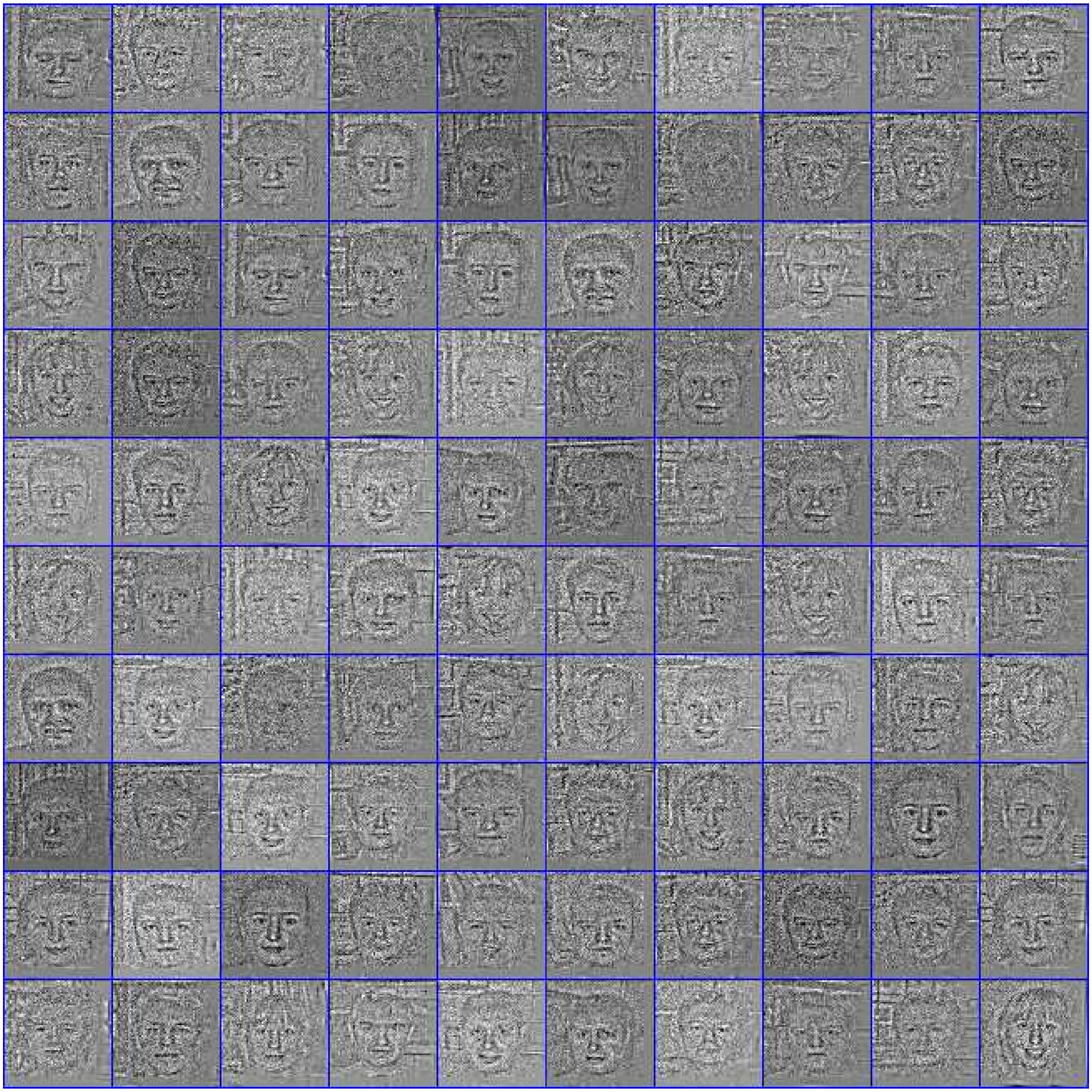

Figure 2: Generated images from the dictionaries trained from "Faces_easy" category of Caltech 256 with random dictionary weights at the top of the three-layer model.

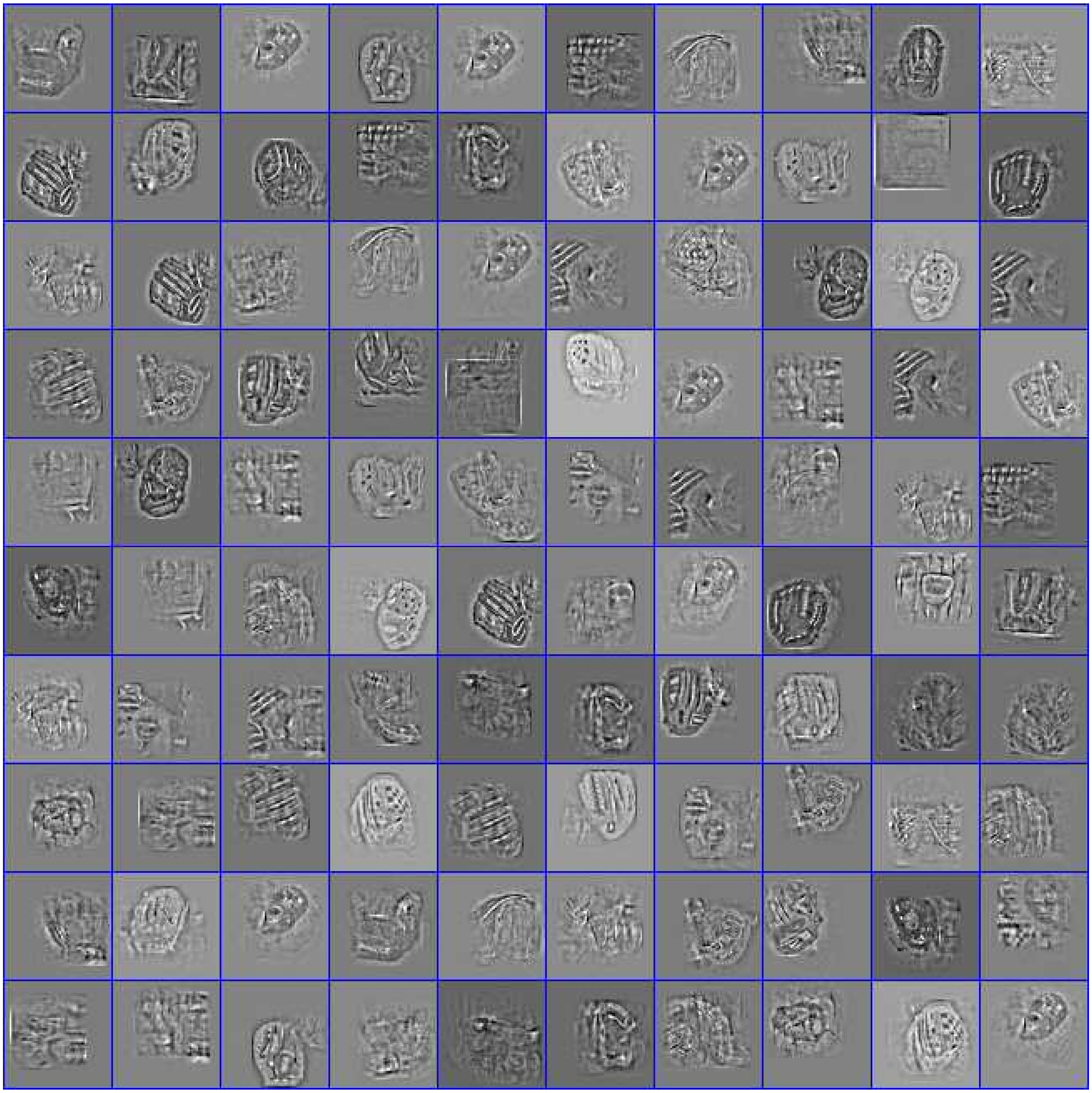

Figure 3: Generated images from the dictionaries trained from ‘baseball-glove” category of Caltech 256 with random dictionary weights at the top of the three-layer model.

A.2 Missing data interpolation

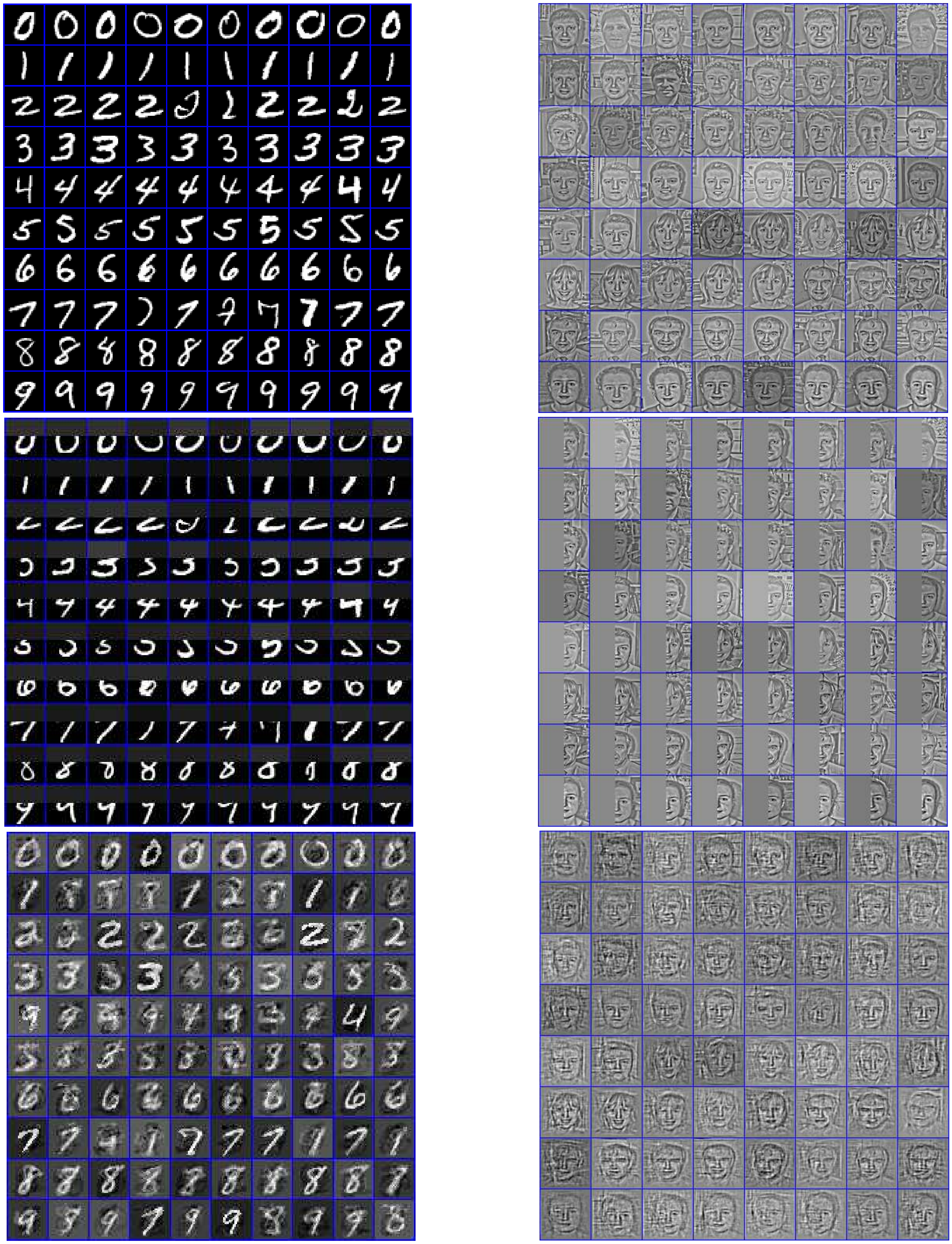

Figure 4: Missing data interpolation of digits (left column) and Face easy (right column). For each column: (Top) Original data. (Middle) Observed data. (Bottom) Reconstruction.

B MCEM algorithm

Algorithms 1 and 2 detail the training and testing process. The steps are explained in the next two sections.

Algorithm 1 Stochastic MCEM Algorithm

Require: Input data $\{\mathbf{X}^{(n)}, \ell_n\}_{n=1}^N$.
  for $t = 1$ **to** ∞ **do**
    Get mini-batch $(\mathbf{Y}^{(n)}; n \in \mathcal{I}_t)$ randomly.
    for $s = 1$ **to** N_s **do**
      Sample $\{\boldsymbol{\gamma}_e^{(n,k_0)}\}_{k_0=1}^{K_0}$ from the distribution in (34);
      sample $\{\gamma_s{}^{(n,k_L)}\}_{k_L=1}^{K_L}$ from the distribution in (33);
      sample $\{\{\mathbf{Z}^{(n,k_l,l)}\}_{k_l=1}^{K_l}\}_{l=1}^{L}$ from the distribution in (24);
      sample $\{\mathbf{S}^{(n,k_L,L)}\}_{k_L=1}^{K_L}$ from the distribution in (31).
    end for
    Compute $\bar{Q}(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(t)})$ according to (41)
    for $l = 1$ **to** L **do**
      Update $\{\boldsymbol{\delta}^{(n,k_{l-1},l,t)}\}_{k_{l-1}=1}^{K_{l-1}}$ according to (46).
      for $k_{l-1} = 1$ **to** K_{L-1} **do**
        for $k_l = 1$ **to** K_L **do**
          Update $\mathbf{D}^{(k_{l-1},k_l,l,t)}$ according to (47).
        end for
        Update $\bar{\mathbf{X}}^{(n,k_{l-1},l,t)} := \sum_{k_l=1}^{K_l} \mathbf{D}^{(k_{l-1},k_l,l,t)} * \bar{\mathbf{S}}^{(n,k_l,l,t)}$.
        Update $\bar{\mathbf{S}}^{(n,k_{l-1},l-1,t)} = f(\bar{\mathbf{X}}^{(n,k_{l-1},l,t)}, \bar{\mathbf{Z}}^{(n,k_{l-1},l-1,t)})$.
      end for
    end for
    for $\ell = 1$ **to** C **do**
      Sample $\lambda_n^{(\ell)}$ from the distibution in (39) and compute the sample average $\bar{\lambda}_n^{(\ell,t)}$.
      Update $\boldsymbol{\beta}^{(\ell,t)}$ according to (48).
    end for
  end for
  return A point estimator of $\mathbf{D}$ and $\boldsymbol{\beta}$.

Algorithm 2 Testing

Require: Input test images $\mathbf{X}^{(*)}$, learned dictionaires $\{\{\mathbf{D}^{(k_l,l)}\}_{k_l=1}^{K_L}\}_{l=1}^{L}$
  for $t = 1$ **to** T **do**
    for $s = 1$ **to** N_s **do**
      Sample $\{\boldsymbol{\gamma}_e^{(n,k_0)}\}_{k_0=1}^{K_0}$ from the distribution in (34);
      sample $\{\gamma_s{}^{(n,k_L)}\}_{k_L=1}^{K_L}$ from the distribution in (33);
      sample $\{\{\mathbf{Z}^{(n,k_l,l)}\}_{k_l=1}^{K_l}\}_{l=1}^{L-1}$ from the distribution in (24);
    end for
    Compute $\bar{Q}_{test}(\boldsymbol{\Psi}_{test}|\boldsymbol{\Psi}_{test}^{(t)})$ according to (55)
    for $l = 1$ **to** L **do**
      Update $\{\boldsymbol{\delta}^{(*,k_{l-1},l,t)}\}_{k_{l-1}=1}^{K_{l-1}}$ according to (46).
    end for
    for $k_L = 1$ **to** K_L **do**
      Update $\mathbf{Z}^{(*,k_L,L)}$ according to (61).
      Update $\mathbf{W}^{(*,k_L,L)}$ according to (60).
    end for
  end for
  Compute $\{\mathbf{S}^{(*,k_L,L)}\}_{k_L=1}^{K_L}$ and get its vector verstion $\boldsymbol{s}_*$.
  Predict label ℓ^*=$\arg\max_\ell \boldsymbol{\beta}_\ell^\top \boldsymbol{s}_*$.
  return the predicted label ℓ^* and the decision value $\boldsymbol{\beta}_\ell^\top \boldsymbol{s}_*$.

C Gibbs Sampling

C.1 Notations

In the remainder of this discussion, we use the following definitions.

(1) **The ceiling function**:
$ceil(x) = \lceil x \rceil$ is the smallest integer that is not less than x.

(2) **The summation and the quadratic summation of all elements in a matrix**:
if $\mathbf{X} \in \mathbb{R}^{N_x \times N_y}$,

$$\mathsf{sum}(\mathbf{X}) = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} X_{ij}, \qquad \|\mathbf{X}\|_2^2 = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} X_{ij}^2. \tag{1}$$

(3) **The unpooling function**:
Assume $\mathbf{S} \in \mathbb{R}^{N_x \times N_y}$ and $\mathbf{X} \in \mathbb{R}^{N_x/p_x \times N_y/p_y}$. Here $p_x, p_y \in N$ are the pooling ratio and the pooling map is $\mathbf{Z} \in \{0,1\}^{N_x \times N_y}$. Let $i' \in \{1, ..., \lceil N_x/p_x \rceil\}$, $j' \in \{1, ..., \lceil N_y/p_y \rceil\}$, $i \in \{1, ..., N_x\}$, $j \in \{1, ..., N_y\}$, then $f : \mathbb{R}^{Nx/p_x \times Ny/p_y} \times \{0,1\}^{N_x \times N_y} \to \mathbb{R}^{N_x \times N_y}$.
If $\mathbf{S} = f(\mathbf{X}, \mathbf{Z})$

$$S_{i,j} = X_{\lceil i/p_x \rceil, \lceil j/p_y \rceil} Z_{i,j}. \tag{2}$$

Thus, the unpooling process (equation(6) in the main paper) can be formed as:

$$\mathbf{S}^{(n,k_l,l)} = \mathsf{unpool}(\mathbf{X}^{(n,k_l,l+1)}) = f(\mathbf{X}^{(n,k_l,l+1)}, \mathbf{Z}^{(n,k_l,l)}). \tag{3}$$

(4) **The 2D correlation operation**:
Assume $\mathbf{B} \in \mathbb{R}^{N_{Bx} \times N_{By}}$ and $\mathbf{C} \in \mathbb{R}^{N_{Cx} \times N_{Cy}}$. If $\mathbf{A} = \mathbf{B} \circledast \mathbf{C}$, then $\mathbf{A} \in \mathbb{R}^{(N_{Bx} - N_{Cx}+1) \times (N_{By} - N_{Cy}+1)}$ with element (i, j) given by

$$A_{i,j} = \sum_{p=1}^{N_{Cx}} \sum_{q=1}^{N_{Cy}} B_{p+i-1, q+j-1} C_{p,q}. \tag{4}$$

(5) **The "error term" in each layer**:

$$\delta_{i,j}^{(n,k_{l-1},l)} = \frac{\partial}{\partial \mathbf{X}_{i,j}^{(n,k_{l-1},l)}} \left\{ \frac{\gamma_e^{(n)}}{2} \sum_{k_0=1}^{K_0} \|\mathbf{E}^{(n,k_0)}\|_2^2 \right\}. \tag{5}$$

(6) **The "generative" function**:
This "generative" function measures how much the k^{th} band of l^{th} layer feature is "responsible" for the of input image $\mathbf{X}^{(n)}$ in the current model:

$$g(\mathbf{X}, n, k, l) = \begin{cases} \mathbf{D}^{(k,1)} * f(\mathbf{X}, \mathbf{Z}^{(n,k,1)}) & \text{if } l = 2, \\ \sum_{m=1}^{K_{l-1}} g\left(\mathbf{D}^{(m,k,l-1)} * f(\mathbf{X}, \mathbf{Z}^{(n,k,l-1)}), n, m, l-1\right) & \text{if } l > 2. \end{cases} \tag{6}$$

It can be considered as if k^{th} band of l^{th} layer feature changes $\mathbf{X}$ (*i.e.* $\mathbf{X}^{(n,k,l)} \to \mathbf{X}^{(n,k,l)} + \mathbf{X}$), the corresponding data layer representation will change $g(\mathbf{X}, n, k, l)$ (*i.e.* $\mathbf{X}^{(n)} \to \mathbf{X}^{(n)} + g(\mathbf{X}, n, k, l)$). Thus, for $l = 2, \ldots, L$, we have

$$\mathbf{X}^{(n)} = \sum_{k=1}^{K_l} g(\mathbf{X}, n, k, l) + \mathbf{E}^{(n)}. \tag{7}$$

Note that $g()$ is a *linear* function for $\mathbf{X}$, which means:

$$g(\mu_1 \mathbf{X}_1 + \mu_2 \mathbf{X}_2, n, k, l) = \mu_1 g(\mathbf{X}_1, n, k, l) + \mu_2 g(\mathbf{X}_2, n, k, l). \tag{8}$$

For convenience, we also use the following notations:

- We use $\mathbf{Z}^{(n,k_l,l)}$ to represent $\{\boldsymbol{z}_{i,j}^{(n,k_l,l)}; \forall i,j\}$, where the vector version of the $(i,j)^{\text{th}}$ block of $\mathbf{Z}^{(n,k_l,l)}$ is equal to $\boldsymbol{z}_{i,j}^{(n,k_l,l)}$.
- $\mathbf{0}$ denotes the all 0 vector or matrix. $\mathbf{1}$ denotes the all one vector or matrix. $\boldsymbol{e}_m$ denotes a "one-hot" vector with the m^{th} element equal to 1.

C.2 Full Conditional Posterior Distribution

Assume the spatial dimension: $\mathbf{X}^{(n,l)} \in \mathbb{R}^{N_x^l \times N_y^l \times K_{l-1}}$, $\mathbf{D}^{(k_l,l)} \in \mathbb{R}^{N_{dx}^l \times N_{dy}^l \times K_{l-1}}$, $\mathbf{S}^{(n,k_l,l)} \in \mathbb{R}^{N_{Sx}^l \times N_{Sy}^l}$ and $\mathbf{Z}^{(n,k_l,l)} \in \{0,1\}^{N_{Sx}^l \times N_{Sy}^l}$. For $l = 0,\dots,L$, we have $k_l = 1,\dots,K_l$. The (un)pooling ratio from $l-$th layer to $(l+1)-$layer is $p_x^l \times p_y^l$ (where $l = 1,\dots,L-1$). We have:

$$N_x^l = N_{dx}^l + N_{Sx}^l - 1, \qquad N_{Sx}^l = p_x^l \times N_x^{(l+1)}, \tag{9}$$

$$N_y^l = N_{dy}^l + N_{Sy}^l - 1, \qquad N_{Sy}^l = p_y^l \times N_y^{(l+1)}. \tag{10}$$

Recall that, for $l = 2,\dots,L$:

$$\mathbf{X}^{(n,k_{l-1},l)} = \sum_{k_l}^{K_l} \mathbf{D}^{(k_{l-1},k_l,l)} * \mathbf{S}^{(n,k_l,l)}. \tag{11}$$

Without loss of generality, we omit the superscript (n, k_{l-1}, l) below. Each element of $\mathbf{X}$ can be represent as:

$$\begin{aligned} X_{i,j} &= \sum_{p=1}^{N_{dx}} \sum_{q=1}^{N_{dy}} D_{p,q} S_{(i+N_{dx}-p,j+N_{dy}-q)} \\ &= D_{p,q} S_{(i+N_{dx}-p,j+N_{dy}-q)} + X_{i,j}^{-(p,q)} \end{aligned} \tag{12}$$

where $X_{i,j}^{-(p,q)}$ is a term which is independent of $D_{p,q}$ but related by the index (i,j,p,q); so is $S_{(i+N_{dx}-p,j+N_{dy}-q)}$. Following this, for every elements in $\mathbf{D}$, we can represent $\mathbf{X}$ as:

$$\mathbf{X} = \mathbf{X}_{-(p,q)} + D_{p,q} \mathbf{S}_{-(p,q)} \tag{13}$$

where matrices $\mathbf{X}_{-(p,q)}$ and $\mathbf{S}_{-(p,q)}$ are independent of $D_{p,q}$ but related by the index (p,q) (and the superscript (n, k_{l-1}, l)). Therefore:

$$\mathbf{E}^{(n)} = \mathbf{X}^{(n)} - \sum_{k=1}^{K_l} g(\mathbf{X}, n, k, l) \tag{14}$$

$$= \mathbf{X}^{(n)} - \sum_{k=1,\neq k_{l-1}}^{K_l} g(\mathbf{X}, n, k, l) - g(\mathbf{X}, n, k_{l-1}, l) \tag{15}$$

$$= \mathbf{X}^{(n)} - \sum_{k=1,\neq k_{l-1}}^{K_l} g(\mathbf{X}, n, k, l) - g\left(\mathbf{X}_{-(p,q)} + D_{p,q}\mathbf{S}_{-(p,q)}, n, k_{l-1}, l\right) \tag{16}$$

$$= \mathbf{X}^{(n)} - \sum_{k=1,\neq k_{l-1}}^{K_l} g(\mathbf{X}, n, k, l) - g\left(\mathbf{X}_{-(p,q)}, n, k_{l-1}, l\right) + g\left(\mathbf{S}_{-(p,q)}, n, k_{l-1}, l\right) D_{p,q} \tag{17}$$

$$= \mathbf{C}_{p,q} - D_{p,q}\mathbf{F}_{(p,q)} \tag{18}$$

If we add the superscripts back, we have:

$$\mathbf{E}^{(n)} = \mathbf{C}_{p,q}^{(n,k_l,l)} + D_{p,q}^{(n,k_l,l)} \mathbf{F}_{p,q}^{(n,k_l,l)}, \tag{19}$$

where matrices $\mathbf{C}_{p,q}^{(n,k_l,l)}$ and $\mathbf{F}_{i,j}^{(n,k_l,l)}$ are independent of $D_{p,q}^{(n,k_l,l)}$ but related by the index (n, k_l, l, p, q).

Similarly, for every elements in $\boldsymbol{z}$, we have

$$\mathbf{E}^{(n)} = \mathbf{A}_{i,j,m}^{(n,k_l,l)} + z_{i,j,m}^{(n,k_l,l)} \mathbf{B}_{i,j,m}^{(n,k_l,l)}. \tag{20}$$

1. The conditional posterior of $\mathbf{D}_{i,j}^{(k_{l-1},k_l,l)}$:

$$D_{i,j}^{(k_{l-1},k_l,l)}|- \quad \sim \quad \mathcal{N}(\mu_{i,j}^{(k_{l-1},k_l,l)}, \sigma_{i,j}^{(k_{l-1},k_l,l)}), \tag{21}$$

where

$$\sigma_{i,j}^{(k_{l-1},k_l,l)} \quad = \quad \left(\frac{\gamma_e^n}{2}\|\mathbf{F}_{i,j}^{(n,k_l,l)}\|_2^2 + 1\right)^{-1}, \tag{22}$$

$$\mu_{i,j}^{(k_{l-1},k_l,l)} \quad = \quad \sigma_{i,j}^{(k_{l-1},k_l,l)} \mathsf{sum}(\mathbf{C}_{i,j}^{(n,k_l,l)} \circ \mathbf{F}_{i,j}^{(n,k_l,l)}). \tag{23}$$

2. The conditional posterior of $\boldsymbol{z}_{i,j}^{(n,k_l,l)}$:

$$\boldsymbol{z}_{i,j}|- \;\sim\; \hat{\boldsymbol{\theta}}_0[\boldsymbol{z}_{i,j} = \mathbf{0}] + \sum_{m=1}^{p_x^l p_y^l} \hat{\boldsymbol{\theta}}_m[\boldsymbol{z}_{i,j} = \boldsymbol{e}_m], \tag{24}$$

where

$$\hat{\theta}_m = \frac{\theta_{i,j}^{(m)} \eta_{i,j}^{(m)}}{\theta_{i,j}^{(0)} + \sum_{\hat{m}=1}^{p_x p_y} \theta_{i,j}^{(\hat{m})} \eta_{i,j}^{(\hat{m})}}, \tag{25}$$

$$\hat{\theta}_0 = \frac{\theta_{i,j}^{(0)}}{\theta_{i,j}^{(0)} + \sum_{\hat{m}=1}^{p_x p_y} \theta_{i,j}^{(\hat{m})} \eta_{i,j}^{(\hat{m})}}, \tag{26}$$

$$\eta_{i,j}^{(m)} = \exp\left\{-\frac{\gamma_e}{2}\left(\|\mathbf{A}_{i,j}^{(m)} - \mathbf{B}_{i,j}^{(m)}\|_2^2 - \|\mathbf{A}_{i,j}^{(m)}\|_2^2\right)\right\}. \tag{27}$$

For notational simplicity, we omit the superscript (n, k_l, l). We can see that when $\eta_{i,j}^{(m)}$ is large, $\hat{\theta}_m$ is large, causing the m^{th} pixel to be activated as the unpooling location. When all of the $\eta_{i,j}^{(m)}$ are small the model will prefer not unpooling – none of the positions m make the model fit the data (*i.e.*, $\mathbf{B}_{i,j}^{(m)}$ is not close to $\mathbf{A}_{i,j}^{(m)}$ for all m); this is mentioned in the main paper.

3. The conditional posterior of $\boldsymbol{\theta}^{(n,k_l,l)}$

$$\boldsymbol{\theta}^{(n,k_l,l)}|- \quad \sim \quad \text{Dir}(\alpha^{(n,k_l,l)}), \tag{28}$$

where

$$\alpha_m^{(n,k_l,l)} = \frac{1}{p_x^l p_y^l + 1} + \sum_i \sum_j Z_{i,j,m}^{(n,k_l,l)} \qquad \text{for } m = 1, ..., p_x^l p_y^l, \tag{29}$$

$$\alpha_0^{(n,k_l,l)} = \frac{1}{p_x^l p_y^l + 1} + \sum_i \sum_j \left(1 - \sum_m Z_{i,j,m}^{(n,k_l,l)}\right). \tag{30}$$

4. The conditional posterior of $S_{i,j}^{(n,k_L,L)}$:

$$S_{i,j}^{(n,k_L,L)}|- \sim (1 - Z_{i,j}^{(n,k_L,L)})\delta_0 + Z_{i,j}^{(n,k_L,L)} \mathcal{N}(\Xi_{i,j}^{(n,k_L,L)}, \Delta_{i,j}^{(n,k_L,L)}), \tag{31}$$

where

$$\Delta_{i,j}^{(n,k_L,L)} = \left(\gamma_e^{(n)} \|\mathbf{F}_{i,j}^{(n,k_L,L)} Z_{i,j}^{(n,k_L,L)}\|_2^2 + \sum_\ell \frac{\gamma}{\lambda_n^{(\ell)}} y_n^{(\ell)} (Z_{i,j}^{(n,k_L,L)} \hat{\beta}_{i,j}^{(k_L,\ell)})^2 + \gamma_s^{(n,k_L)}\right)^{-1},$$

$$\Xi_{i,j}^{(n,k_L,L)} = \Delta_{i,j}^{(n,k_L,L)} Z_{i,j}^{(n,k_L,L)} \left(\mathsf{sum}(\mathbf{F}_{i,j}^{(n,k_L,L)} \circ \mathbf{C}_{i,j}^{(n,k_L,L)}) + \sum_\ell y_n^{(\ell)} \hat{\beta}_{i,j}^{(k_L,\ell)} (1 + \lambda_n^{(\ell)})\right). \tag{32}$$

Here we reshape the long vector $\boldsymbol{\beta}_\ell \in \mathbb{R}^{N_{sx}^l N_{sy}^L K_L \times 1}$ into a matrix $\hat{\boldsymbol{\beta}}_\ell \in \mathbb{R}^{N_{sx}^l \times N_{sy}^L \times K_L}$ which has the same size of $\mathbf{S}^{(n,L)}$.

5. The conditional posterior of $\gamma_s^{(n,k_L)}$:

$$\gamma_s^{(n,k_L)}|- \sim \text{Gamma}\left(a_s + \frac{N_{Sx}^L \times N_{Sy}^L}{2}, b_s + \frac{1}{2}\|\mathbf{S}^{(n,k_L,L)}\|_2^2\right). \tag{33}$$

6. The conditional posterior of $\gamma_e^{(n)}$:

$$\gamma_e^{(n)}|- \sim \text{Gamma}\left(a_0 + \frac{N_x \times N_y \times K_0}{2}, b_0 + \frac{1}{2}\sum_{k_0=1}^{K_0}\|\mathbf{E}^{(n,k_0)}\|_2^2\right). \tag{34}$$

7. The conditional posterior of $\boldsymbol{\beta}_\ell$:

   Reshape the long vector $\boldsymbol{\beta}_\ell \in \mathbb{R}^{N_{sx}^l N_{sy}^L K_L \times 1}$ into a matrix $\hat{\boldsymbol{\beta}}_\ell \in \mathbb{R}^{N_{sx}^l \times N_{sy}^L \times K_L}$ which has the same size as $\mathbf{S}^{(n,L)}$. We have:

$$\begin{aligned}
\hat{\beta}_{i,j}^{(k_L,\ell)}|- &\sim \mathcal{N}(\mu_{i,j}^{(k_L,\ell)}, \sigma_{i,j}^{(k_L,\ell)}), &(35)\\
\sigma_{i,j}^{(k_L,\ell)} &= \left(\sum_n \frac{\gamma}{\lambda_n^{(\ell)}} y_n^{(\ell)} (S_{i,j}^{(n,k_L,L)})^2 + \frac{1}{\omega_{i,j}^{(k_L,\ell)}}\right)^{-1}, &(36)\\
\mu_{i,j}^{(k_L,\ell)} &= \sigma_{i,j}^{(n,\ell)} \sum_n \left[y_n^{(\ell)} S_{i,j}^{(n,k_L,L)} (1 + \lambda_n^{(\ell)} - \Gamma_{-(k,i,j)}^{(n,k_L,L)})\right], &(37)\\
\Gamma_{-(k,i,j)}^{(n,k_L,L)} &= \sum_{\substack{k'\\ k'\neq\ell}} \sum_{\substack{i'\\ i'\neq i}} \sum_{\substack{j'\\ j'\neq j}} S_{i',j'}^{(n,k',L)} \beta_{i',j'}^{(k',\ell)}. &(38)
\end{aligned}$$

8. The conditional posterior of $\lambda_n^{(\ell)}$

$$(\lambda_n^{(\ell)})^{-1} \sim \mathcal{IG}(|1 - \boldsymbol{y}_n^\ell \boldsymbol{s}_n^\top \boldsymbol{\beta}^{(\ell,t)}|^{-1}, 1), \tag{39}$$

   where $\mathcal{IG}$ denotes the inverse Gaussian distribution.

D MCEM algorithm Details

D.1 E step

Recall that we consolidate the "local" model parameters (latent data-sample-specific variables) as $\boldsymbol{\Phi}_n = \left(\{\boldsymbol{z}^{(n,l)}\}_{l=1}^L, \mathbf{S}^{(n,L)}, \boldsymbol{\gamma}_s^{(n)}, \mathbf{E}^{(n)}, \{\lambda_n^{(\ell)}\}_{\ell=1}^C\right)$, the "global" parameters (shared across all data) as $\boldsymbol{\Psi} = \left(\{\mathbf{D}^{(l)}\}_{l=1}^L, \boldsymbol{\beta}\right)$, and the data as $\mathbf{Y}_n = (\mathbf{X}^{(n)}, \ell_n)$.. At t^{th} iteration of the MCEM algorithm, the exact Q function can be written as:

$$\begin{aligned}
Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(t)}) &= \ln p(\boldsymbol{\Psi}) + \sum_{n\in\mathcal{I}_t} \mathbb{E}_{(\boldsymbol{\Phi}_n|\boldsymbol{\Psi}^{(t)},\mathbf{Y},\boldsymbol{y})}\{\ln p(\mathbf{Y}_n, \boldsymbol{\Phi}_n|\boldsymbol{\Psi})\}\\
&= -\mathbb{E}_{(\mathbf{Z},\boldsymbol{\gamma}_e,\mathbf{S}^{(L)},\boldsymbol{\gamma}_s,\boldsymbol{\lambda}|\mathbf{Y},\mathbf{D}^{(t)},\boldsymbol{\beta}^{(t)})}\left\{\sum_{n\in\mathcal{I}_t}\left[\frac{\gamma_e{}^{(n)}}{2}\sum_{k_0=1}^{K_0}\|\mathbf{E}^{(n,k_0)}\|_2^2 + \sum_{\ell=1}^C \frac{(1+\lambda_n^\ell - y_n^\ell \boldsymbol{\beta}_\ell^T \boldsymbol{s}_n)^2}{2\lambda_n^\ell}\right]\right\}\\
&\quad -\frac{1}{2}\sum_{l=1}^L \sum_{k_{l-1}=1}^{K_{l-1}} \sum_{k_l=1}^{K_l} \|\mathbf{D}^{(k_{l-1},k_l,l)}\|_2^2 + const,
\end{aligned} \tag{40}$$

where $const$ denotes the terms which are not a function of $\boldsymbol{\Psi}$.

Obtaining a closed form of the exact Q function is analytically intractable. We here approximate the expectations in (40) by samples collected from the posterior distribution of the hidden variables developed in Section C.2.
The Q function in (40) can be approximated by:

$$\begin{aligned}
\bar{Q}(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(t)}) &= -\frac{1}{N_s}\sum_{s=1}^{N_s}\left\{\sum_{n\in\mathcal{I}_t}\left[\frac{\bar{\gamma_e}^{(n,s,t)}}{2}\sum_{k_0=1}^{K_0}\|\bar{\mathbf{E}}^{(n,k_0,s,t)}\|_2^2 + \sum_{\ell=1}^C \frac{(1+\bar{\lambda_n}^{(\ell,s,t)} - y_n^\ell \boldsymbol{\beta}_\ell^T \bar{\boldsymbol{s}_n}^{(s,t)})^2}{2\bar{\lambda}_n^{(\ell,s,t)}}\right]\right\}\\
&\quad -\frac{1}{2}\sum_{l=1}^L \sum_{k_{l-1}=1}^{K_{l-1}} \sum_{k_l=1}^{K_l} \|\mathbf{D}^{(k_{l-1},k_l,l)}\|_2^2 + const,
\end{aligned} \tag{41}$$

where

$$\bar{\mathbf{E}}^{(n,k_0,s,t)} = \mathbf{X}^{(n,k_0)} - \sum_{k_1=1}^{K_1} \mathbf{D}^{(k_0,k_1,1)} * \bar{\mathbf{S}}^{(n,k_1,1,s,t)}, \tag{42}$$

and for $l = 2, \ldots, L$

$$\bar{\mathbf{X}}^{(n,k_{l-1},l,s,t)} = \sum_{k_l=1}^{K_l} \mathbf{D}^{(k_{l-1},k_l,l)} * \bar{\mathbf{S}}^{(n,k_l,l,s,t)}, \tag{43}$$

$$\bar{\mathbf{S}}^{(n,k_{l-1},l-1,s,t)} = f(\bar{\mathbf{X}}^{(n,k_{l-1},l,s,t)}, \bar{\mathbf{Z}}^{(n,k_{l-1},l-1,s,t)}), \tag{44}$$

where $\bar{\mathbf{S}}^{(L,s,t)}$, $\bar{\gamma_e}^{(s,t)}$, $\bar{\boldsymbol{\lambda}}^{(s,t)}$ and $\bar{\mathbf{Z}}^{(s,t)}$ are a sample of the corresponding variables from the full conditional posterior at the t^{th} iteration. N_s is the number of collected samples.

D.2 M step

We can maximize $\bar{Q}(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(t)})$ via the following updates:

1. For $l = 1, \ldots, L,\ k_{l-1} = 1, \ldots, K_{L-1}$ and $k_l = 1, \ldots, K_L$, the gradient wrt $\mathbf{D}^{(k_{l-1},k_l,l)}$ is:

$$\frac{\partial \bar{Q}}{\partial \mathbf{D}^{(k_{l-1},k_l,l,t)}} = \sum_{n \in \mathcal{I}_t} \boldsymbol{\delta}^{(n,k_{l-1},l,t)} \circledast \bar{\mathbf{S}}^{(n,k_l,l,t)} + \mathbf{D}^{(k_{l-1},k_l,l,t)}, \tag{45}$$

where

$$\begin{aligned} \boldsymbol{\delta}^{(n,k_0,1,t)} &= \bar{\gamma}_e^{(n,k_0,t)} \left[\mathbf{X}^{(n,k_0)} - \textstyle\sum_{k_1=1}^{K_1} \mathbf{D}^{(k_0,k_1,1)} * \bar{\mathbf{S}}^{(n,k_1,1,t)} \right], \\ \boldsymbol{\delta}^{(n,k_{l-1},l,t)} &= f\left(\textstyle\sum_{k_{l-2}=1}^{K_{l-2}} (\boldsymbol{\delta}^{(n,k_{l-2},l-1,t)} \circledast D^{(k_{l-2},k_{l-1},l-1,t)}), \bar{Z}^{(n,k_{l-1},l-1,t)} \right). \end{aligned} \tag{46}$$

Following this, the update rule of $\mathbf{D}$ based on RMSprop is:

$$\begin{aligned} \boldsymbol{v}^{t+1} &= \alpha \boldsymbol{v}^t + (1-\alpha)\left(\tfrac{\partial \bar{Q}}{\partial \mathbf{D}^{(k_{l-1},k_l,l,t)}}\right)^2, \\ \mathbf{D}^{(k_{l-1},k_l,l,t+1)} &= \mathbf{D}^{(k_{l-1},k_l,l,t)} + \tfrac{\epsilon}{\sqrt{\boldsymbol{v}_{t+1}}} \tfrac{\partial \bar{Q}}{\partial \mathbf{D}^{(k_{l-1},k_l,l,t)}}. \end{aligned} \tag{47}$$

2. For $\ell = 1, \ldots, C$, the update rule of $\boldsymbol{\beta}^{\ell}$ is:

$$\boldsymbol{\beta}^{(\ell,t+1)} = \left[(\Omega^{(\ell,t)})^{-1} + \bar{\boldsymbol{s}}_{(\ell,t)}^{\top} (\Lambda^{(\ell,t)})^{-1} \bar{\boldsymbol{s}}_{(\ell,t)} \right]^{-1} \bar{\boldsymbol{s}}_{(\ell,t)}^{\top} (\mathbf{1} + (\Lambda^{(\ell,t)})^{-1}), \tag{48}$$

where

$$(\Lambda^{(\ell,t)})^{-1} = diag((\bar{\lambda}_n^{(\ell,t)})^{-1}), \tag{49}$$

$$(\Omega^{(\ell,t)})^{-1} = diag(|\boldsymbol{\beta}^{(\ell,t)}|^{-1}). \tag{50}$$

and $\bar{\boldsymbol{s}}_{(\ell,t)}$ denotes a matrix with row n equal to $\boldsymbol{y}_n^{\ell} \bar{\boldsymbol{s}}_n^{(t)}$.

D.3 Testing

During testing, when given a test image $\mathbf{X}^{(*)}$, we treat $\mathbf{S}^{(*,L)}$ as model parameters and use MCEM to find a MAP estimator:

$$\mathbf{S}^{(*,L)} = \underset{\mathbf{S}^{(*,L)}}{\operatorname{argmax}} \ \ln p(\mathbf{S}^{(*,L)}|\mathbf{X}^{(*)}, \mathbf{D}). \tag{51}$$

Let $\mathbf{S}^{(*,k_L,L)} = \mathbf{W}^{(*,k_L,L)} \circ \mathbf{Z}^{(*,k_L,L)}$, where $\mathbf{W}^{(*,k_L,L)} \in \mathbb{R}^{N_{sx}^L \times N_{sy}^L}$. The marginal posterior distribution can be represented as:

$$p(\mathbf{S}^{(*,L)}|\mathbf{X}^*, \mathbf{D}) = p(\mathbf{W}^{(*,L)}, \mathbf{Z}^{(*,L)}|\mathbf{Y}^{(*)}, \mathbf{D}) \quad (52)$$

$$\propto \int \sum_{/\mathbf{Z}^{(L)}} p(\mathbf{X}^{(*)}|\mathbf{W}^{(*,L)}, \mathbf{Z}, \mathbf{E}^{(*)}, \mathbf{D}) p(\mathbf{W}^{(*,L)}|\gamma_s^{(*)}) p(\mathbf{Z}) p(\gamma_s^{(*)}) p(\mathbf{E}^{(*)}) d\mathbf{E}^{(*)} d\gamma_s^{(*)}, \quad (53)$$

where $/\mathbf{Z}^{(L)} = \{\mathbf{Z}^{(l)}\}_{l=1}^{L-1}$. Let $\mathbf{\Psi}_{test} = \{\mathbf{W}^{(*,L)}, \mathbf{Z}^{(*,L)}\}$ and $\mathbf{\Phi}_{test} = \{\{\mathbf{Z}^{(l)}\}_{l=1}^{L-1}, \boldsymbol{\gamma}_s^*, \mathbf{E}^*\}$. The Q function for testing can be represented as:

$$Q_{test}(\mathbf{\Psi}_{test}|\mathbf{\Psi}_{test}^{(t)}) = \mathbb{E}_{(\mathbf{\Phi}_{test}|\mathbf{\Psi}_{test}^{(t)}, \mathbf{Y}^{(*)}, \mathbf{D})} \left\{ \ln p(\mathbf{X}^{(*)}, \mathbf{D}, \mathbf{\Phi}_{test}, \mathbf{\Psi}_{test}) \right\}. \quad (54)$$

The testing also follows EM steps:

E-step: In the E-step we collect the samples of $\boldsymbol{\gamma}_e, \boldsymbol{\gamma}_s$ and $\{\mathbf{Z}^{(l)}\}_{l=1}^{L-1}$ from conditional posterior distributions, which is similar to the training process. Q_{test} can thus be approximated by:

$$\bar{Q}_{test}(\mathbf{\Psi}_{test}|\mathbf{\Psi}_{test}^{(t)}) = -\sum_{s=1}^{N_s} \left\{ \frac{\bar{\gamma}_e^{(*,s,t)}}{2} \sum_{k_0=1}^{K_0} \| \sum_{k_1=1}^{K_1} \mathbf{D}^{(k_0,k_1,1)} * \bar{\mathbf{S}}^{(*,k_1,1,s,t)} \|_2^2 + \frac{1}{2} \sum_{k_L=1}^{K_L} \bar{\gamma}_s^{(*,k_L,s)} \|\mathbf{W}^{(*,k_L,L)}\|_2^2 \right\} \quad (55)$$

where

$$\bar{\mathbf{X}}^{(*,k_{L-1},L,t)} = \sum_{k_L=1}^{K_L} \mathbf{D}^{(k_{L-1},k_L,L)} * \left(\mathbf{W}^{(*,k_L,L)} \circ \mathbf{Z}^{(*,k_L,L)} \right), \quad (56)$$

and for $l = 2, \ldots, L-1$

$$\bar{\mathbf{S}}^{(*,k_{l-1},l-1,s,t)} = f(\bar{\mathbf{X}}^{(*,k_{l-1},l,t)}, \bar{\mathbf{Z}}^{(*,k_{l-1},l-1,s,t)}), \quad (57)$$

$$\bar{\mathbf{X}}^{(*,k_{l-1},l,s,t)} = \sum_{k_l=1}^{K_l} \mathbf{D}^{(k_{l-1},k_l,l)} * \bar{\mathbf{S}}^{(*,k_l,l,s,t)}. \quad (58)$$

M-step: In the M-step, we maximize $\bar{Q}_{test}$ via the following updates:

1. The gradient $w.r.t.$ $\mathbf{W}^{(*,K_L,L)}$ is:

$$\frac{\partial \bar{Q}_{test}}{\partial \mathbf{W}^{(*,k_L,L,t)}} = \left[\sum_{k_{L-1}}^{K_L} \boldsymbol{\delta}^{(*,k_{L-1},L,t)} \circledast \mathbf{D}^{(k_{L-1},k_L,L)} \right] \circ \mathbf{Z}^{(*,k_L,L)} + \bar{\gamma}_s^{(*,k_L)} \mathbf{W}^{(*,k_L,L,t)}, \quad (59)$$

where $\boldsymbol{\delta}^{(*,k_{L-1},L,t)}$ is the same as (46). Therefore, the update rule of $\mathbf{W}$ based on RMSprop is:

$$\begin{aligned} \boldsymbol{u}^{t+1} &= \alpha \boldsymbol{u}^t + (1-\alpha)\left(\tfrac{\partial \bar{Q}_{test}}{\partial \mathbf{W}^{(*,k_L,L,t)}}\right)^2 \\ \mathbf{W}^{(*,K_L,L,t+1)} &= \mathbf{W}^{(*,K_L,L,t)} + \tfrac{\epsilon}{\sqrt{\boldsymbol{u}_{t+1}}} \tfrac{\partial \bar{Q}_{test}}{\partial \mathbf{W}^{(*,k_L,L,t)}} \end{aligned} \quad (60)$$

2. The update rule $\mathbf{Z}^{(*,k_L,L)}$ is

$$\mathbf{Z}_{i,j}^{(*,k_L,L)} = \begin{cases} 1 & \text{if } \theta^{(*,k_L,L)} \eta_{i,j}^{(*,k_L,L)} > 1 - \theta^{(*,k_L,L)} \\ 0 & \text{otherwise} \end{cases} \quad (61)$$

where $\eta_{i,j}^{(*,k_L,L)}$ is the same as (27).

E Bottom-Up Pretraining

E.1 Pretraining Model

The model is pretrained sequentially from the bottom layer to the top layer. We consider here pretraining the relationship between layer l and layer $l+1$, and this process may be repeated up to layer L. The basic framework of this pretraining process is closely connected to top-down generative process, with a few small but important modifications. Matrix $\mathbf{X}^{(n,l)}$ represents the pooled and stacked activation weights from layer $l-1$, image n (K_{l-1} "spectral bands" in $\mathbf{X}^{(n,l)}$, due to K_{l-1} dictionary elements at layer $l-1$). We constitute the representation

$$\mathbf{X}^{(n,l)} = \sum_{k_l=1}^{K_l} \mathbf{D}^{(k_l,l)} * \mathbf{S}^{(n,k_l,l)} + \mathbf{E}^{(n,l)}, \tag{62}$$

with

$$\mathbf{D}^{(k_l,l)} \sim \mathcal{N}(0, \mathbf{I}_{N_D^{(l)}}) \qquad \mathbf{E}^{(n,l)} \sim \mathcal{N}(0, (\boldsymbol{\gamma}_e^{(n,l)})^{-1}\mathbf{I}_{N_X^{(l)}}) \qquad \boldsymbol{\gamma}_e^{(n,l)} \sim \text{Gamma}(a_e, b_e). \tag{63}$$

The features $\mathbf{S}^{(n,k_l,l)}$ can be partitioned into contiguous blocks with dimension $p_x^l \times p_y^l$. In our generative model, $\mathbf{S}^{(n,k_l,l)}$ is generated from $\mathbf{X}^{(n,k_l,l+1)}$ and $\boldsymbol{z}^{(n,k_l,l)}$, where the non-zero element within the $(i,j)-$th pooling block of $\mathbf{S}^{(n,k_l,l)}$ is set equal to $X_{i,j}^{(n,k_l,l+1)}$, and its location within the pooling block is determined by $\boldsymbol{z}_{i,j}^{(n,k_l,l)}$, a $p_x^l \times p_y^l$ binary vector (Sec. 2.2 in the main paper). Now the matrix $\mathbf{X}^{(n,k_l,l+1)}$ is constituted by "stacking" the spatially-aligned and pooled versions of $\mathbf{S}_{k_l=1,K_l}^{(n,k_l,l)}$. Thus, we need to place a prior on the $(i,j)-$th pooling block of $\mathbf{S}^{(n,k_l,l)}$:

$$\mathbf{S}_{i,j,m}^{(n,k_l,l)} = z_{i,j,m}^{(n,k_l,l)} W_{i,j,m}^{(n,k_l,l)}, \quad m = 1, \dots, p_x^l p_y^l \tag{64}$$

$$\boldsymbol{z}_{i,j}^{(n,k_l,l)} \sim \boldsymbol{\theta}_0^{(n,k_l,l)}[\boldsymbol{z}_{i,j}^{(n,k_l,l)} = \mathbf{0}] + \sum_{m=1}^{p_x^l p_y^l} \boldsymbol{\theta}_m^{(n,k_l,l)}[\boldsymbol{z}_{i,j}^{(n,k_l,l)} = \boldsymbol{e}_m], \qquad \boldsymbol{\theta}^{(n,l,k_l)} \sim \text{Dir}(1/p_x^l p_y^l, \dots, 1/p_x^l p_y^l), \tag{65}$$

$$W_{i,j,m}^{(n,k_l,l)} \sim \mathcal{N}(0, \gamma_{wl}^{-1}), \qquad \gamma_{wl} \sim \text{Gamma}(a_w, b_w). \tag{66}$$

If all the elements of $\boldsymbol{z}_{i,j,k_l}^{(n,l)}$ are zero, the corresponding pooling block in $\mathbf{S}_{i,j}^{(n,k_l,l)}$ will be all zero and $X_{i,j}^{(n,k_l,l+1)}$ will be zero.

Therefore, the model can be formed as:

$$\mathbf{X}^{(n,l)} = \sum_{k_l=1}^{K_l} \mathbf{D}^{(k_l,l)} * \underbrace{\left(\mathbf{Z}^{(n,k_l,l)} \odot \mathbf{W}^{(n,k_l,l)}\right)}_{=\mathbf{S}^{(n,k_l,l)}} + \mathbf{E}^{(n,l)}, \tag{67}$$

where the vector version of the (i,j)-th block of $\mathbf{Z}^{(n,k_l,l)}$ is equal to $\boldsymbol{z}_{i,j}^{(n,k_l,l)}$ and $\odot$ is the Hadamard (element-wise) product operator. The hyperparameters are set as $a_e = b_e = a_w = b_w = 10^{-6}$.

We summarize distinctions between pretraining, and the top-down generative model.

- A pair of consecutive layers is considered at a time during pretraining.
- During the pretraining process, the residual term $\mathbf{E}^{(n,l)}$ is used to fit each layer.
- In the top-down generative process, the residual is only employed at the bottom layer to fit the data.
- During pretraining, the pooled activation weights $\mathbf{X}^{(n,l+1)}$ are sparse, encouraging a parsimonious convolutional dictionary representation.
- The model parameters learned from pretraining are used to initialize the model when executing top-down model refinement, using the full generative model.

E.2 Conditional Posterior Distribution for Pretraining

- $D_{i,j}^{(k_{l-1},k_l,l)}|- \sim \mathcal{N}(\mathbf{\Phi}_{i,j}^{(k_{l-1},k_l,l)}, \mathbf{\Sigma}_{i,j}^{(k_{l-1},k_l,l)})$

$$\mathbf{\Sigma}^{(k_{l-1},k_l,l)} = \mathbf{1} \oslash \left(\sum_{n=1}^{N} \gamma_e^{(n,l)} \|\mathbf{Z}^{(n,k_l,l)} \odot \mathbf{W}^{(n,k_l,l)}\|_2^2 + \mathbf{1} \right) \tag{68}$$

$$\mathbf{\Phi}^{(k_{l-1},k_l,l)} = \mathbf{\Sigma}^{(k_{l-1},k_l,l)} \odot \left\{ \sum_{n=1}^{N} \gamma_e^{(n,l)} \Big[\mathbf{X}^{-(n,k_{l-1},l)} \circledast (\mathbf{Z}^{(n,k_l,l)} \odot \mathbf{W}^{(n,k_l,l)}) + \|\mathbf{Z}^{(n,k_l,l)} \odot \mathbf{W}^{(n,k_l,l)}\|_2^2 \mathbf{D}^{(k_{l-1},k_l,l)} \Big] \right\} \tag{69}$$

- $W_{i,j}^{(n,k_l,l)}|- \sim \mathcal{N}(\mathbf{\Xi}_{i,j}^{(n,k_l,l)}, \mathbf{\Lambda}_{i,j}^{(n,k_l,l)})$

$$\mathbf{\Lambda}^{(n,k_l,l)} = \mathbf{1} \oslash \left(\sum_{k_{l-1}=1}^{K_{l-1}} \gamma_e^{(n,l)} \|\mathbf{D}^{(k_{l-1},k_l,l)}\|_2^2 \mathbf{Z}^{(n,k_l,l)} + \gamma_w^{(n,k_l,l)} \mathbf{1} \right) \tag{70}$$

$$\mathbf{\Xi}^{(n,k_l,l)} = \mathbf{\Lambda}^{(n,k_l,l)} \odot \mathbf{Z}^{(n,k_l,l)} \odot \left\{ \sum_{k_{l-1}=1}^{K_{l-1}} \gamma_e^{(n,l)} \Big[\mathbf{X}^{-(n,k_{l-1},l)} \circledast \mathbf{D}^{(k_{l-1},k_l,l)} + \|\mathbf{D}^{(k_{l-1},k_l,l)}\|_2^2 \mathbf{W}^{(n,k_l,l)} \odot \mathbf{Z}^{(n,k_l,l)} \Big] \right\} \tag{71}$$

- $\gamma_w^{(n,k_l,l)}|- \sim \text{Gamma}\left(a_w + \frac{N_{sx}^l \times N_{sy}^l}{2}, b_w + \frac{\|\mathbf{W}^{(k_{l-1},k_l,l)}\|_2^2}{2} \right)$

- $\boldsymbol{z}_{i,j}^{(n,k_l,l)}$:

  Let $m \in \{1, ..., p_x^l p_y^l\}$; from

$$\mathbf{Y}^{(n,k_l,l)} = \sum_{k_{l-1}=1}^{K_{l-1}} \gamma_e^{(n,l)} \left[\|\mathbf{D}^{(k_{l-1},k_l,l)}\|_2^2 \odot \left(\mathbf{W}^{(n,k_l,l)}\right)^2 - 2\left(\mathbf{X}_{-k_l}^{(n,k_{l-1},l)} \circledast \mathbf{D}^{k_{l-1},k_l,l}\right) \odot \mathbf{W}^{(n,k_l,l)} \right] \tag{72}$$

  and

$$\hat{\theta}_{i,j,m}^{(n,k_l,l)} = \frac{\theta_m^{(n,k_l,l)} \exp\left\{-\frac{1}{2} Y_{i,j,m}^{(n,k_l,l)}\right\}}{\theta_0^{(n,k_l,l)} + \sum_{\hat{m}=1}^{p_x^l p_y^l} \theta_{\hat{m}}^{(n,k_l,l)} \exp\left\{-\frac{1}{2} Y_{i,j,\hat{m}}^{(n,k_l,l)}\right\}}, \tag{73}$$

$$\hat{\theta}_{i,j,0}^{(n,k_l,l)} = \frac{\theta_0^{(n,k_l,l)}}{\theta_0^{(n,k_l,l)} + \sum_{\hat{m}=1}^{p_x^l p_y^l} \theta_{\hat{m}}^{(n,k_l,l)} \exp\left\{-\frac{1}{2} Y_{i,j,\hat{m}}^{(n,k_l,l)}\right\}}, \tag{74}$$

(75)

  we have

$$\boldsymbol{z}_{i,j}^{(n,k_l,l)}| \sim \hat{\boldsymbol{\theta}}_0^{(n,k_l,l)}[\boldsymbol{z}_{i,j}^{(n,k_l,l)} = \mathbf{0}] + \sum_{m=1}^{p_x^l p_y^l} \hat{\boldsymbol{\theta}}_m^{(n,k_l,l)}[\boldsymbol{z}_{i,j}^{(n,k_l,l)} = \boldsymbol{e}_m]. \tag{76}$$

- $\boldsymbol{\theta}^{(n,k_l,l)}|- \sim \mathrm{Dir}(\alpha^{(n,k_l,l)})$

$$\alpha_m^{(n,k_l,l)} = \frac{1}{p_x^l p_y^l + 1} + \sum_i \sum_j Z_{i,j,m}^{(n,k_l,l)} \qquad \text{for } m = 1, ..., p_x^l p_y^l, \tag{77}$$

$$\alpha_0^{(n,k_l,l)} = \frac{1}{p_x^l p_y^l + 1} + \sum_i \sum_j \left(1 - \sum_m Z_{i,j,m}^{(n,k_l,l)}\right) \tag{78}$$

- $\gamma_e^{(n,l)}|- \sim \mathrm{Gamma}\left(a_e + \frac{N_x^l \times N_y^l \times K_{l-1}}{2}, b_e + \sum_{k_{l-1}=1}^{K_{l-1}} \frac{\|\mathbf{X}^{-(n,k_{l-1},l)}\|_2^2}{2}\right)$