Supplementary Material of Hierarchically Structured Neural Decoding with Matrix-variate Gaussian Prior for Perceived Image Reconstruction

A Proof of Proposition 1

To prove Proposition 1, we need the following facts about tensor product:

Fact A.1 *Let* $\mathbf{A}$ *be a matrix. Then* $||\mathbf{A}||_F = ||\mathrm{vec}(\mathbf{A})||_2$.

Fact A.2 *Let* $\mathbf{A} \in \mathbb{R}^{m_1 \times n_1}$, $\mathbf{B} \in \mathbb{R}^{n_1 \times n_2}$ *and* $\mathbf{C} \in \mathbb{R}^{n_2 \times m_2}$. *Then* $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^\top \otimes \mathbf{A})\mathrm{vec}(\mathbf{B})$.

Fact A.3 *Let* $\mathbf{A} \in \mathbb{R}^{m_1 \times n_1}, \mathbf{B} \in \mathbb{R}^{n_1 \times p_1}$, $\mathbf{C} \in \mathbb{R}^{m_2 \times n_2}$ *and* $\mathbf{D} \in \mathbb{R}^{n_2 \times p_2}$. *Then* $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AB}) \otimes (\mathbf{CD})$.

We can show the following result by transforming $\mathbf{W}$ into its isomorphic counterpart:

$$
\begin{aligned}
\mathcal{L}_{\mathbf{W}} &= \mathrm{tr}\big((\mathbf{H} - \mathbf{XW} - \mathbf{1b}^\top)\mathbf{\Omega}^{-1}(\mathbf{H} - \mathbf{XW} - \mathbf{1b}^\top)^\top\big) + \lambda \mathrm{tr}(\mathbf{WW}^\top) + \lambda_1 \mathrm{tr}(\mathbf{\Sigma}_r^{-1}\mathbf{W}\mathbf{\Sigma}_c^{-1}\mathbf{W}^\top) \\
&= ||(\mathbf{H} - \mathbf{XW} - \mathbf{1b}^\top)\mathbf{\Omega}^{-\frac{1}{2}}||_F^2 + \lambda||\mathbf{W}||_F^2 + \lambda_1||\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{W}\mathbf{\Sigma}_c^{-\frac{1}{2}}||_F^2 \\
&= ||\mathrm{vec}[(\mathbf{H} - \mathbf{XW} - \mathbf{1b}^\top)\mathbf{\Omega}^{-\frac{1}{2}}]||_2^2 + \lambda||\mathrm{vec}(\mathbf{W})||_2^2 + \lambda_1||\mathrm{vec}(\mathbf{\Sigma}_r^{-\frac{1}{2}}\mathbf{W}\mathbf{\Sigma}_c^{-\frac{1}{2}})||_2^2 && \text{(By Fact A.1)} \\
&= ||\mathrm{vec}[(\mathbf{H} - \mathbf{1b}^\top)\mathbf{\Omega}^{-\frac{1}{2}}] - (\mathbf{\Omega}^{-\frac{1}{2}} \otimes \mathbf{X})\mathrm{vec}(\mathbf{W})||_2^2 + \lambda||\mathrm{vec}(\mathbf{W})||_2^2 + \lambda_1||(\mathbf{\Sigma}_c^{-\frac{1}{2}} \otimes \mathbf{\Sigma}_r^{-\frac{1}{2}})\mathrm{vec}(\mathbf{W})||_2^2 && \text{(By Fact A.2)} \\
&= \mathrm{vec}(\mathbf{W})^\top \Big((\mathbf{\Omega}^{-\frac{1}{2}} \otimes \mathbf{X})^\top(\mathbf{\Omega}^{-\frac{1}{2}} \otimes \mathbf{X}) + \lambda \mathbf{I}_K \otimes \mathbf{I}_D + \lambda_1(\mathbf{\Sigma}_c^{-\frac{1}{2}} \otimes \mathbf{\Sigma}_r^{-\frac{1}{2}})^\top(\mathbf{\Sigma}_c^{-\frac{1}{2}} \otimes \mathbf{\Sigma}_r^{-\frac{1}{2}})\Big)\mathrm{vec}(\mathbf{W}) \\
&\quad - 2\mathrm{vec}(\mathbf{W})^\top(\mathbf{\Omega}^{-\frac{1}{2}} \otimes \mathbf{X}^\top)\mathrm{vec}[(\mathbf{H} - \mathbf{1b}^\top)\mathbf{\Omega}^{-\frac{1}{2}}] + \mathrm{vec}[(\mathbf{H} - \mathbf{1b}^\top)\mathbf{\Omega}^{-\frac{1}{2}}]^\top\mathrm{vec}[(\mathbf{H} - \mathbf{1b}^\top)\mathbf{\Omega}^{-\frac{1}{2}}] \\
&= \mathrm{vec}(\mathbf{W})^\top \big(\mathbf{\Omega}^{-1} \otimes \mathbf{X}^\top\mathbf{X} + \lambda \mathbf{I}_{KD} + \lambda_1(\mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r)\big)\mathrm{vec}(\mathbf{W}) \\
&\quad - 2\mathrm{vec}(\mathbf{W})^\top(\mathbf{\Omega}^{-\frac{1}{2}} \otimes \mathbf{X}^\top)\mathrm{vec}[(\mathbf{H} - \mathbf{1b}^\top)\mathbf{\Omega}^{-\frac{1}{2}}] + \mathrm{vec}[(\mathbf{H} - \mathbf{1b}^\top)\mathbf{\Omega}^{-\frac{1}{2}}]^\top\mathrm{vec}[(\mathbf{H} - \mathbf{1b}^\top)\mathbf{\Omega}^{-\frac{1}{2}}] && \text{(By Fact A.3)} \\
&= \mathrm{vec}(\mathbf{W})^\top \underbrace{\big(\mathbf{\Omega}^{-1} \otimes \mathbf{X}^\top\mathbf{X} + \lambda \mathbf{I}_{KD} + \lambda_1(\mathbf{\Sigma}_c \otimes \mathbf{\Sigma}_r)\big)}_{\mathbf{U}}\mathrm{vec}(\mathbf{W}) - 2\mathrm{vec}(\mathbf{W})^\top \underbrace{\mathrm{vec}[\mathbf{X}^\top(\mathbf{H} - \mathbf{1b}^\top)\mathbf{\Omega}^{-1}]}_{\mathbf{V}} \\
&\quad + \mathrm{vec}[(\mathbf{H} - \mathbf{1b}^\top)\mathbf{\Omega}^{-\frac{1}{2}}]^\top\mathrm{vec}[(\mathbf{H} - \mathbf{1b}^\top)\mathbf{\Omega}^{-\frac{1}{2}}] && \text{(By Fact A.2)}
\end{aligned}
$$

The last equation above is a quadratic function of $\mathrm{vec}(\mathbf{W})$, from which we can read off that the optimal solution should satisfy:

$$
\mathrm{vec}(\hat{\mathbf{W}}) = \mathbf{U}^{-1}\mathbf{V} = \big[\mathbf{\Omega}^{-1} \otimes (\mathbf{X}^\top\mathbf{X}) + \lambda \mathbf{I}_{KD} + \lambda_1 \mathbf{\Sigma}_c^{-1} \otimes \mathbf{\Sigma}_r^{-1}\big]^{-1}\mathrm{vec}(\mathbf{X}^\top(\mathbf{H} - \mathbf{1b}^\top)\mathbf{\Omega}^{-1}). \tag{1}
$$

$\hat{\mathbf{W}}$ can then be obtained simply by reformatting $\mathrm{vec}(\hat{\mathbf{W}})$ into a $D \times K$ matrix. The computational bottleneck in the above procedure is in solving an $KD \times KD$ system of equations, which scales as $\mathcal{O}(K^3D^3)$ if no further structure is available. The overall computational complexity is $\mathcal{O}(K^3D^3 + K^2ND^2)$.

B Algorithms

Algorithm 1 Structured Multi-output Regression (SMR)

Input: $\mathbf{X} \in \mathbb{R}^{N \times D}$, $\mathbf{H} \in \mathbb{R}^{N \times K}$, $\lambda, \lambda_1, \lambda_2, \lambda_3$ and η.

Initialize $\mathbf{W} = \mathbf{0}$, $\mathbf{\Omega}^{-1} = \mathbf{I}_K$, $\mathbf{\Sigma}_r^{-1} = \mathbf{I}_D$, $\mathbf{\Sigma}_c^{-1} = \mathbf{I}_K$ and $\mathbf{b} = \frac{1}{N}\mathbf{H}^\top\mathbf{1}$
while not converged **do**
  while not converged **do**
    Update $\mathbf{W}$ by $\mathbf{W} \leftarrow \mathbf{W} - \eta\nabla_{\mathbf{W}}\mathcal{L}_{\mathbf{W}}$
  end while
  Update $\mathbf{b}$ by $\hat{\mathbf{b}} = \frac{1}{N}(\mathbf{H} - \mathbf{XW})^\top\mathbf{1}$
  Update $\mathbf{\Omega}^{-1}$ by Eq. (13)
  Update $\mathbf{\Sigma}_r^{-1}$ by Eq. (16)
  Update $\mathbf{\Sigma}_c^{-1}$ by Eq. (17)
end while
Output: $\mathbf{W}, \mathbf{b}, \mathbf{\Omega}^{-1}, \mathbf{\Sigma}_r^{-1}, \mathbf{\Sigma}_c^{-1}$

Algorithm 2 Introspective Conditional Generation (ICG)

Input: Training images $\mathbf{Y}$, conditions $\mathbf{H}$, hyperparameters α and β.

1: Initialize network parameters θ and ϕ
2: **while** not converged **do**
3: Update the generator by Eq. (18) and Eq. (19) : $\hat{\theta} = \arg\min_{\theta}[L_{AE} + \alpha D_{KL}(\mathbf{y}_g)]$
4: Update the encoder by Eq. (18) and Eq. (19) : $\hat{\phi} = \arg\min_{\phi}[L_{AE} + \beta D_{KL}(\mathbf{y}) - \alpha D_{KL}(\mathbf{y}_g)]$
5: **end while**
6: **Output**: $\hat{\theta}$ and $\hat{\phi}$

C Datasets

C.1 Vim-1

The Vim-1 dataset is a publicly available fMRI dataset, which contains the blood-oxygen-level dependent (BOLD) responses of two subjects when they are presented with grayscale natural images [Kay *et al.*, 2008]. The fMRI voxels come from the visual brain areas V1, V2, V3, V4, V3A, V3B and LO. The dataset is partitioned into distinct training and test sets which consist of 1750 and 120 instances, respectively. This dataset was acquired using a 4T INOVA MR scanner (Varian, Inc.) at a spatial resolution of $3 \times 3 \times 4$ mm^3 and a temporal resolution of 1 Hz. During the acquisition, subjects viewed sequences of $15^\circ \times 15^\circ$ grayscale natural photographs (500×500 pixels) while fixating on a central white square. Photographs were presented for 1 s with a delay of 3 s between successive photographs. The dataset is available online at https://crcns.org/data-sets/vc/vim-1. All results in the current study use data from subject 1.

C.2 FaceBold

We collected a new fMRI dataset, which comprises grayscale face stimuli and the corresponding BOLD responses. The stimuli (330×380 pixels) used in the fMRI experiment were drawn from the NimStim set, the California Facial Expressions of emotion dataset, the Japanese Female Facial Expression Database, the Karolinska Directed Emotional Faces, and the Radboud Faces Database. The whole dataset consists of photographs of front-facing individuals with four (happy, fear, disgust, and neutral) kinds of expressions. We recorded the BOLD responses (TR = 2 s, voxel size = $3 \times 3 \times 4$ mm^3, whole-brain coverage) of six healthy adult subjects (3 female and 3 males, 20 to 30 years old) as they were fixating on a small dot superimposed on the stimuli ($15^\circ \times 15^\circ$). Each face was presented at 5 Hz for 2 s and followed by a middle gray background presented for 6 s. In total, 720 faces were presented once for the training set, and 80 faces were presented twice for the test set. The brain activity for each stimuli was estimated using the general linear model (GLM) denoise pipeline [Kay *et al.*, 2013]. All results in the current study use data from subject 1.

C.3 ImageNet-1k

The ICG model for natural images were trained on a downsampled (128×128) variant of Vim-1 together with the ImageNet-1k [Deng *et al.*, 2009] dataset. The image size of ImageNet-1k was downsampled to 128×128 using the method proposed in [Chrabaszcz *et al.*, 2017]. Before training, all images were converted to gray scale and contrast-enhanced, similar to the transformation described in [Kay *et al.*, 2008] (remapping the pixel values to a new range and saturating the bottom and top 1% of the pixel values). To make them look like the images in Vim-1, we finally applied the circular mask [Kay *et al.*, 2008] to them. This resulted in approximately 1,280,000 grayscale natural images used for training in total.

C.4 CelebA

The ICG model for human faces were trained on a downsampled (128×128) variant of FaceBold together with the CelebA [Liu *et al.*, 2015] dataset. The CelebA dataset is available online at http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html. This dataset comprises 202599 in-the-wild portraits of 10177 people, which were drawn from online sources. We use the aligned and cropped version (i.e., the "img_align_celeba.zip" file) in our experiments. The images in CelebA were further center-cropped to 128×128 to align them with the images in FaceBold. Finally, all images were converted to gray scale. This resulted in approximately 203,000 grayscale human face images used for training in total.

D Network architectures

Network architectures of the proposed ICG model are listed in Table 1 and Table 2 for reconstructing the natural images and human faces, respectively. The latent variable $\mathbf{z}$ is randomly drawn from a $\mathcal{N}(\mathbf{0}, \mathbf{I})$ distribution, with the dimension set to 512 and 256 on the Vim-1 and FaceBold datasets, respectively. For optimizing the proposed ICG model, we used the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) [Kingma and Ba, 2014] with a batch size of 32 and a fixed learning rate 0.0001.

Table 1: Network architectures of the proposed ICG model for generating 128×128 natural images on the Vim-1 dataset.

Encoder		
Layer	Output shape	Filter
FC (Condition)	32	$5000 \rightarrow 512 \rightarrow 32$
Conv2D (Image)	$32 \times 128 \times 128$	$3 \rightarrow 32$
Concat	$64 \times 128 \times 128$	-
Resnet-Block	$64 \times 128 \times 128$	$64 \rightarrow 64 \rightarrow 64$
Resnet-Block	$128 \times 128 \times 128$	$64 \rightarrow 64 \rightarrow 128$
Avg-Pool2D	$128 \times 64 \times 64$	-
Resnet-Block	$128 \times 64 \times 64$	$128 \rightarrow 128 \rightarrow 128$
Resnet-Block	$256 \times 64 \times 64$	$128 \rightarrow 128 \rightarrow 256$
Avg-Pool2D	$256 \times 32 \times 32$	-
Resnet-Block	$256 \times 32 \times 32$	$256 \rightarrow 256 \rightarrow 256$
Resnet-Block	$512 \times 32 \times 32$	$256 \rightarrow 256 \rightarrow 512$
Avg-Pool2D	$512 \times 16 \times 16$	-
Resnet-Block	$512 \times 16 \times 16$	$512 \rightarrow 512 \rightarrow 512$
Resnet-Block	$1024 \times 16 \times 16$	$512 \rightarrow 512 \rightarrow 1024$
Avg-Pool2D	$1024 \times 8 \times 8$	-
Resnet-Block	$1024 \times 8 \times 8$	$1024 \rightarrow 1024 \rightarrow 1024$
Resnet-Block	$1024 \times 8 \times 8$	$1024 \rightarrow 1024 \rightarrow 1024$
Avg-Pool2D	$1024 \times 4 \times 4$	-
Resnet-Block	$1024 \times 4 \times 4$	$1024 \rightarrow 1024 \rightarrow 1024$
Resnet-Block	$1024 \times 4 \times 4$	$1024 \rightarrow 1024 \rightarrow 1024$
FC	1024	$1024 \cdot 4 \cdot 4 \rightarrow 1024$
Split	$512, 512$	-
Generator		
Layer	Output shape	Filter
Latent vector	512	reparameterization
FC (Condition)	512	$5000 \rightarrow 512$
Concat	1024	-
FC	$1024 \cdot 4 \cdot 4$	$1024 \rightarrow 1024 \cdot 4 \cdot 4$
Reshape	$1024 \times 4 \times 4$	-
Resnet-Block	$1024 \times 4 \times 4$	$1024 \rightarrow 1024 \rightarrow 1024$
Resnet-Block	$1024 \times 4 \times 4$	$1024 \rightarrow 1024 \rightarrow 1024$
Upsampling	$1024 \times 8 \times 8$	-
Resnet-Block	$1024 \times 8 \times 8$	$1024 \rightarrow 1024 \rightarrow 1024$
Resnet-Block	$1024 \times 8 \times 8$	$1024 \rightarrow 1024 \rightarrow 1024$
Upsampling	$1024 \times 16 \times 16$	-
Resnet-Block	$512 \times 16 \times 16$	$1024 \rightarrow 512 \rightarrow 512$
Resnet-Block	$512 \times 16 \times 16$	$512 \rightarrow 512 \rightarrow 512$
Upsampling	$512 \times 32 \times 32$	-
Resnet-Block	$256 \times 32 \times 32$	$512 \rightarrow 256 \rightarrow 256$
Resnet-Block	$256 \times 32 \times 32$	$256 \rightarrow 256 \rightarrow 256$
Upsampling	$256 \times 64 \times 64$	-
Resnet-Block	$128 \times 64 \times 64$	$256 \rightarrow 128 \rightarrow 128$
Resnet-Block	$128 \times 64 \times 64$	$128 \rightarrow 128 \rightarrow 128$
Upsampling	$128 \times 128 \times 128$	-
Resnet-Block	$64 \times 128 \times 128$	$128 \rightarrow 64 \rightarrow 64$
Resnet-Block	$64 \times 128 \times 128$	$64 \rightarrow 64 \rightarrow 64$
Conv2D	$3 \times 128 \times 128$	$64 \rightarrow 3$

Table 2: Network architectures of the proposed ICG model for generating 128×128 human faces on the FaceBold dataset.

Encoder		
Layer	Output shape	Filter
FC (Condition)	32	$5000 \rightarrow 512 \rightarrow 32$
Conv2D (Image)	$32 \times 128 \times 128$	$3 \rightarrow 32$
Concat	$64 \times 128 \times 128$	-
Resnet-Block	$64 \times 128 \times 128$	$64 \rightarrow 64 \rightarrow 64$
Resnet-Block	$128 \times 128 \times 128$	$64 \rightarrow 64 \rightarrow 128$
Avg-Pool2D	$128 \times 64 \times 64$	-
Resnet-Block	$128 \times 64 \times 64$	$128 \rightarrow 128 \rightarrow 128$
Resnet-Block	$256 \times 64 \times 64$	$128 \rightarrow 128 \rightarrow 256$
Avg-Pool2D	$256 \times 32 \times 32$	-
Resnet-Block	$256 \times 32 \times 32$	$256 \rightarrow 256 \rightarrow 256$
Resnet-Block	$512 \times 32 \times 32$	$256 \rightarrow 256 \rightarrow 512$
Avg-Pool2D	$512 \times 16 \times 16$	-
Resnet-Block	$512 \times 16 \times 16$	$512 \rightarrow 512 \rightarrow 512$
Resnet-Block	$512 \times 16 \times 16$	$512 \rightarrow 512 \rightarrow 512$
Avg-Pool2D	$512 \times 8 \times 8$	-
Resnet-Block	$512 \times 8 \times 8$	$512 \rightarrow 512 \rightarrow 512$
Resnet-Block	$512 \times 8 \times 8$	$512 \rightarrow 512 \rightarrow 512$
Avg-Pool2D	$512 \times 4 \times 4$	-
Resnet-Block	$512 \times 4 \times 4$	$512 \rightarrow 512 \rightarrow 512$
Resnet-Block	$512 \times 4 \times 4$	$512 \rightarrow 512 \rightarrow 512$
FC	512	$512 \cdot 4 \cdot 4 \rightarrow 512$
Split	$256, 256$	-
Generator		
Layer	Output shape	Filter
Latent vector	256	reparameterization
FC (Condition)	256	$5000 \rightarrow 256$
Concat	512	-
FC	$512 \cdot 4 \cdot 4$	$512 \rightarrow 512 \cdot 4 \cdot 4$
Reshape	$512 \times 4 \times 4$	-
Resnet-Block	$512 \times 4 \times 4$	$512 \rightarrow 512 \rightarrow 512$
Resnet-Block	$512 \times 4 \times 4$	$512 \rightarrow 512 \rightarrow 512$
Upsampling	$512 \times 8 \times 8$	-
Resnet-Block	$512 \times 8 \times 8$	$512 \rightarrow 512 \rightarrow 512$
Resnet-Block	$512 \times 8 \times 8$	$512 \rightarrow 1024 \rightarrow 512$
Upsampling	$512 \times 16 \times 16$	-
Resnet-Block	$512 \times 16 \times 16$	$1024 \rightarrow 512 \rightarrow 512$
Resnet-Block	$512 \times 16 \times 16$	$512 \rightarrow 512 \rightarrow 512$
Upsampling	$512 \times 32 \times 32$	-
Resnet-Block	$256 \times 32 \times 32$	$512 \rightarrow 256 \rightarrow 256$
Resnet-Block	$256 \times 32 \times 32$	$256 \rightarrow 256 \rightarrow 256$
Upsampling	$256 \times 64 \times 64$	-
Resnet-Block	$128 \times 64 \times 64$	$256 \rightarrow 128 \rightarrow 128$
Resnet-Block	$128 \times 64 \times 64$	$128 \rightarrow 128 \rightarrow 128$
Upsampling	$128 \times 128 \times 128$	-
Resnet-Block	$64 \times 128 \times 128$	$128 \rightarrow 64 \rightarrow 64$
Resnet-Block	$64 \times 128 \times 128$	$64 \rightarrow 64 \rightarrow 64$
Conv2D	$3 \times 128 \times 128$	$64 \rightarrow 3$

E Additional results

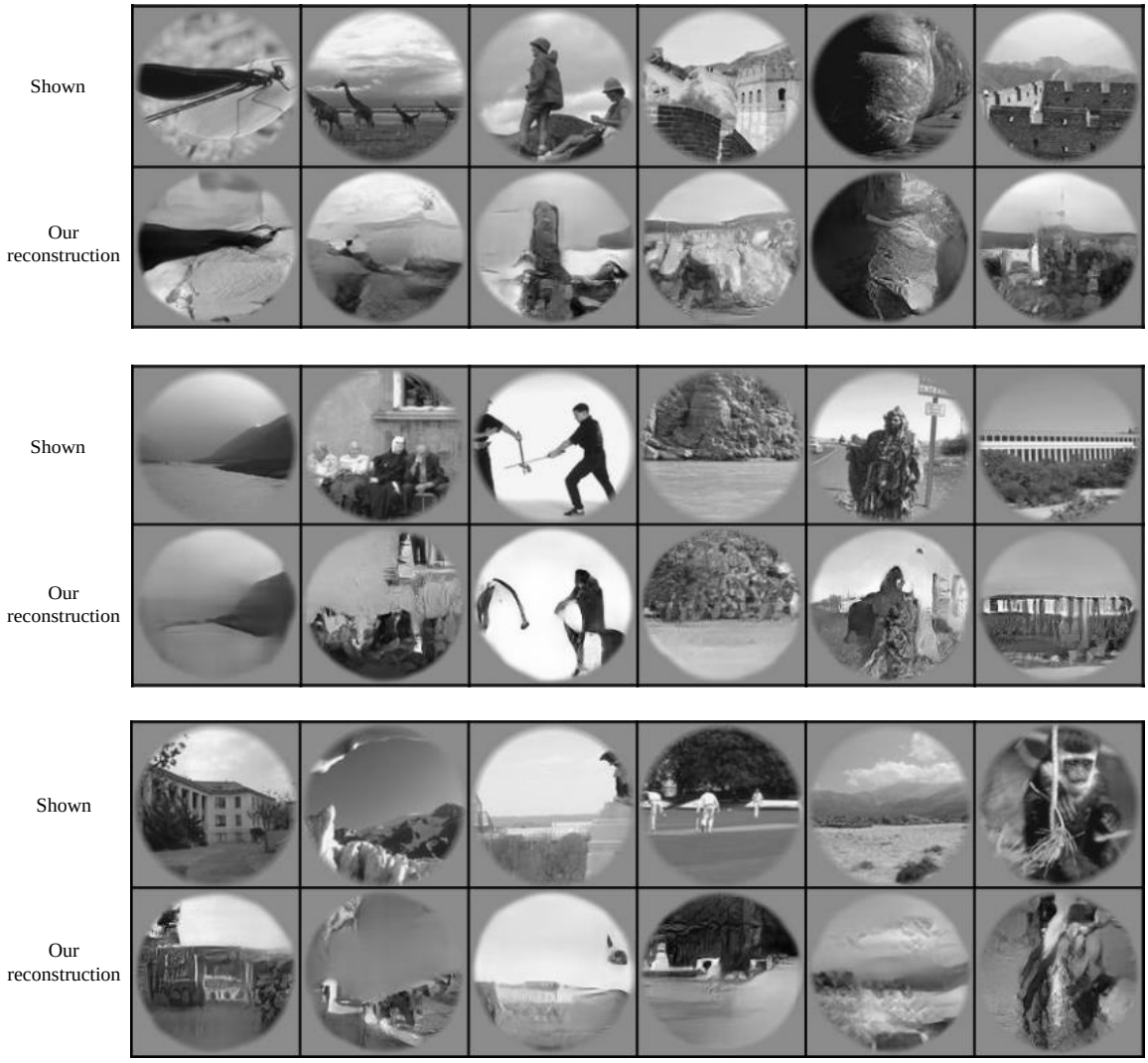

Figure 1: Additional examples of reconstructed 128×128 natural images on the Vim-1 dataset.

References

[Chrabaszcz *et al.*, 2017] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[Kay *et al.*, 2008] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352, 2008.

[Kay *et al.*, 2013] Kendrick Kay, Ariel Rokem, Jonathan Winawer, Robert Dougherty, and Brian Wandell. Glmdenoise: a fast, automated technique for denoising task-based fmri data. *Frontiers in neuroscience*, 7:247, 2013.

[Kingma and Ba, 2014] D. P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.