# A deep generative model explaining tuning properties of monkey face processing patches

**Haruo Hosoya (hosoya@atr.jp)**
Cognitive Mechanisms Laboratories, ATR International
2–2–2 Hikaridai, Keihanna Science City, Kyoto, Japan, 619–0288

## Abstract

Recent monkey studies have revealed a face processing network in the IT cortex that consists of multiple face-selective patches and forms a putative functional hierarchy. Although a number of computational models accounting for this have been proposed, they have been mostly feedforward, ignoring the reciprocal nature of the visual system. Here, we present a two-layer deep generative model based on variational autoencoder (VAE), which provides a Bayesian probabilistic framework with explicit feedforward and feedback processing. While the lower layer of our model uses a standard VAE, the upper layer uses our recently developed algorithm called group-based VAE, which is capable of learning invariant representations from inputs with grouping information. After training with multi-view face images, the upper layer encoded view-invariant facial identities while the lower layer showed facial feature tuning, both in a way quantitatively similar to the observations in patches AM and ML, respectively, as shown in Freiwald and Tsao (2010) and Freiwald et al. (2009). Taken together, we have found a novel deep generative model that might have some computational relevance with the monkey face processing system.

**Keywords:** invariant representation; variational autoencoder

## Introduction

Recent experimental studies have revealed a face-processing network in the monkey IT cortex that consists of multiple face-selective patches (Moeller, Freiwald, & Tsao, 2008). Those patches have specific tuning properties, forming a functional hierarchy with progressively increase of both facial identity selectivity and view invariance from the posterior to the anterior patches (Freiwald & Tsao, 2010). Although a number of computational studies have been conducted to explain these findings, their models have been mostly feedforward, thus precluding any further speculation of possible roles of the reciprocal visual processing (Einhäuser, Hipp, Eggert, Körner, & König, 2005; Farzmahdi, Rajaei, Ghodrati, Ebrahimpour, & Khaligh-Razavi, 2016; Leibo, Liao, Anselmi, Freiwald, & Poggio, 2017). Thus, this poses the question: is there a hierarchical generative model that can account for the properties of the face-processing system?

In this study, we take a first step to address this question and propose a two-layer deep generative model that targets at face-processing patches AM and ML. In this model, we use learning algorithms based on variational autoencoder (VAE) (Kingma & Welling, 2014). VAE generally provides a Bayesian probabilistic framework, along the lines of previous theoretical approaches for visual modeling (Olshausen & Field, 1997; Hyvärinen, Hurri, & Hoyer, 2009; Hosoya & Hyvärinen, 2017; Hosoya, 2012), but formalizes explicit feedforward and feedback processing in inference and learning. In our model, the lower layer uses a standard variational autoencoder (VAE), while the upper layer uses our recently developed algorithm called group-based variational autoencoder (GVAE) (Hosoya, 2019), an extension of VAE capable of learning invariant representations. In GVAE, the model assumes that training images of the same identity are grouped together (inspired by the classical temporal coherence principle (Földiák, 1991)) and thereby separately learns the identity representation as the factor common within a group and the view representation as the factor specific to each image. We have trained the model with multi-view face images with grouping information and tested it by simulating two past monkey experiments (Freiwald & Tsao, 2010; Freiwald, Tsao, & Livingstone, 2009). As a result, we found that the model showed view-invariant coding of facial identities in the upper layer and more specific facial feature tuning in the lower layer, in a way similar to the monkey face-processing patches AM and ML, respectively, as shown in the physiological experiments.

Two other prior studies have taken a generative approach but different from our work here. In (Yildirim, Kulkarni, & Freiwald, 2015), the generative part of their model is a fixed inverse-graphics algorithm that can generate a face image from given parameters of view (pose/light) and identity (shape/texture), whereas the forward part is a convolutional network trained to infer those parameters from an image input (and compared to physiology); thus, view invariance is hard-coded in their model. In (Hosoya & Hyvärinen, 2017), a mixture of sparse coding models is introduced to account for selectivity and tuning properties of ML, but does not explain view invariance in AM.

## Model

Our model consists of two-layers with the architecture depicted in Figure 1A. In the lower layer (layer-I), we have one latent variable $x$ and a forward network $e$ to infer $x$ as well as a generative network $d$ to infer back the input $t$; we use convolutional and deconvolutional neural networks for $e$ an $d$, respectively. In the upper layer (layer-II), we have two latent variables $y$ (view) and $z$ (identity) and forward networks $g$ and $h$ each to infer $y$ and $z$ as well as a generative network $f$ to infer back $x$; all networks are fully-connected.

Formally, layer-I is a (standard) VAE model. That is, we assume the following probabilistic generative model:

$$p(x) = \mathcal{N}(0, I) \tag{1}$$

$$p(t|x) = \mathcal{N}(d(x), \sigma^2 I) \tag{2}$$

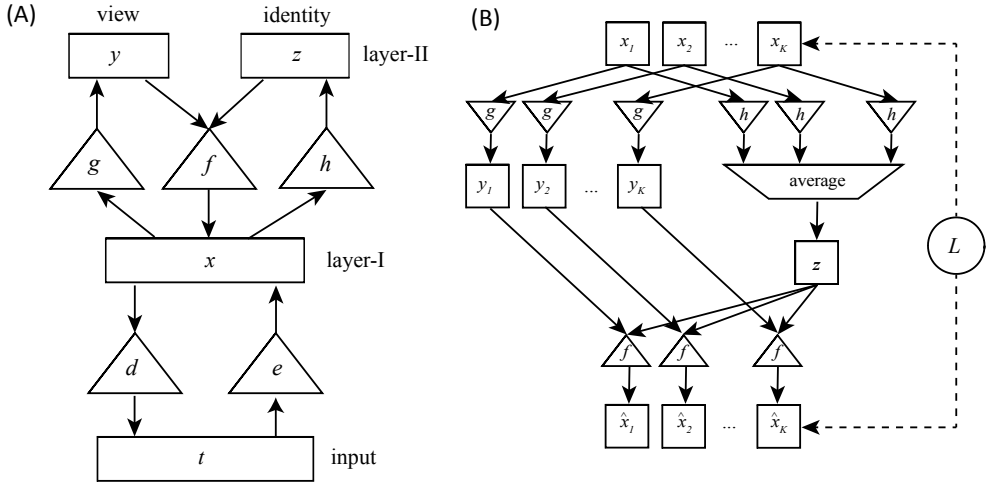


Figure 1: (A) Architecture of our two-layer model. In layer-I, the forward network $e$ infers the intermediate variable $x$ from the input $t$, while the generative network $d$ infers back the input. In layer-II, the forward networks $g$ and $h$ each infer the view variable $y$ and the identity variable $z$, while the generative network $f$ infers back the intermediate variable $x$. (B) Algorithmic outline of GVAE. Given a group of inputs $x_k$, the corresponding individual views $y_k$ are computed by $g$ and the common identity $z$ is computed by the average of $h$. Then, each reconstructed input $\hat{x}_k$ is computed by $f$ from the combination of $y_k$ and $z$. The learning of the network weights is based on the variational autoencoder scheme (see text).

This defines the generative process that first draws $x$ from the standard Gaussian prior and then draws $t$ from the Gaussian distribution whose mean is given by the generative network $d$ applied to $x$. For inference of the posterior distribution of $x$ for a given input $t$, we assume the following inference model:

$$q(x|t) = \mathcal{N}(e(t), e^v(t)) \tag{3}$$

which is the Gaussian distribution whose mean is given by the forward network $e$ applied to $t$ and whose variance is given by an additional neural network $e^v$.[1] The weights of all the neural networks are parameters of the model and are determined by a learning algorithm based on variational Bayes. The algorithm is essentially to minimize the reconstruction error of the autoencoding loop with a certain regularization constraint. See (Kingma & Welling, 2014) for more details of VAE.

For layer-II, a naive application of VAE would not work since there would be no clue on which latent dimension corresponds to view or identity. Thus, we instead use GVAE, an extension of VAE taking inputs with grouping information. In this, we assume that inputs of the same identity are grouped together. From such groups of inputs, we learn to extract the identity code as the factor common within each group and the view code as the factor differentiating each group member. Formally, each input group has $K$ members, $(x_1, \ldots, x_K)$, with $x_k$ indexed by the member number $k$. We assume independence between groups but not members within a group. For an input group, we consider the following probabilistic generative model with the (member-specific) view variables $y_1, \ldots, y_K$ and the (group-common) identity variable $z$:

$$p(z) = \mathcal{N}(0, I) \tag{4}$$

$$p(y_k) = \mathcal{N}(0, I) \tag{5}$$

$$p(x_k|y_k, z) = \mathcal{N}(f(y_k, z), \rho^2 I) \tag{6}$$

for $k = 1, \ldots, K$. For inference of posteriors, we again assume the following inference models:

$$q(y_k|x_k) = \mathcal{N}(g(x_k), g^v(x_k)) \tag{7}$$

$$q(z|x_1, \ldots, x_K) = \mathcal{N}\left(\frac{1}{K}\sum_{k=1}^{K} h(x_k), \frac{1}{K}\sum_{k=1}^{K} h^v(x_k)\right) \tag{8}$$

($g^v$ and $h^v$ are additional neural networks to infer the variances of $y_k$ and $z$, respectively.) Here, the inference of $z$ is slightly more complicated since we need to estimate the group-common identity code. Our approach here is to first compute the individual identity codes and then take the average. Again, the weights in the neural networks are determined by a variational Bayes learning algorithm similarly to the VAE method. Figure 1B illustrates the algorithmic outline of GVAE; see (Hosoya, 2019) for more details of GVAE.

We have constructed a concrete two-layer model with 100 units for intermediate variable $x$, 3 units for view variable $y$, and 100 units for identity variable $z$ (with $\sigma = \rho = 0.1$). For training, we used a dataset of synthetic face images generated from a 3D morphable head model (Dai, Pears, Smith, & Duncan, 2017). In this, each image had an identity with random 100 shape and 100 texture dimensions and had a view of random horizontal (from $-90^\circ$ to $+90^\circ$) and vertical angles (from $-30^\circ$ to $+30^\circ$). We first trained layer-I by the VAE method with this dataset. To train layer-II, we randomly grouped 5 face

[1] Since we consider only Gaussians with diagonal covariances, we specify a vector of variances in the second parameter to Gaussian distribution as convention.

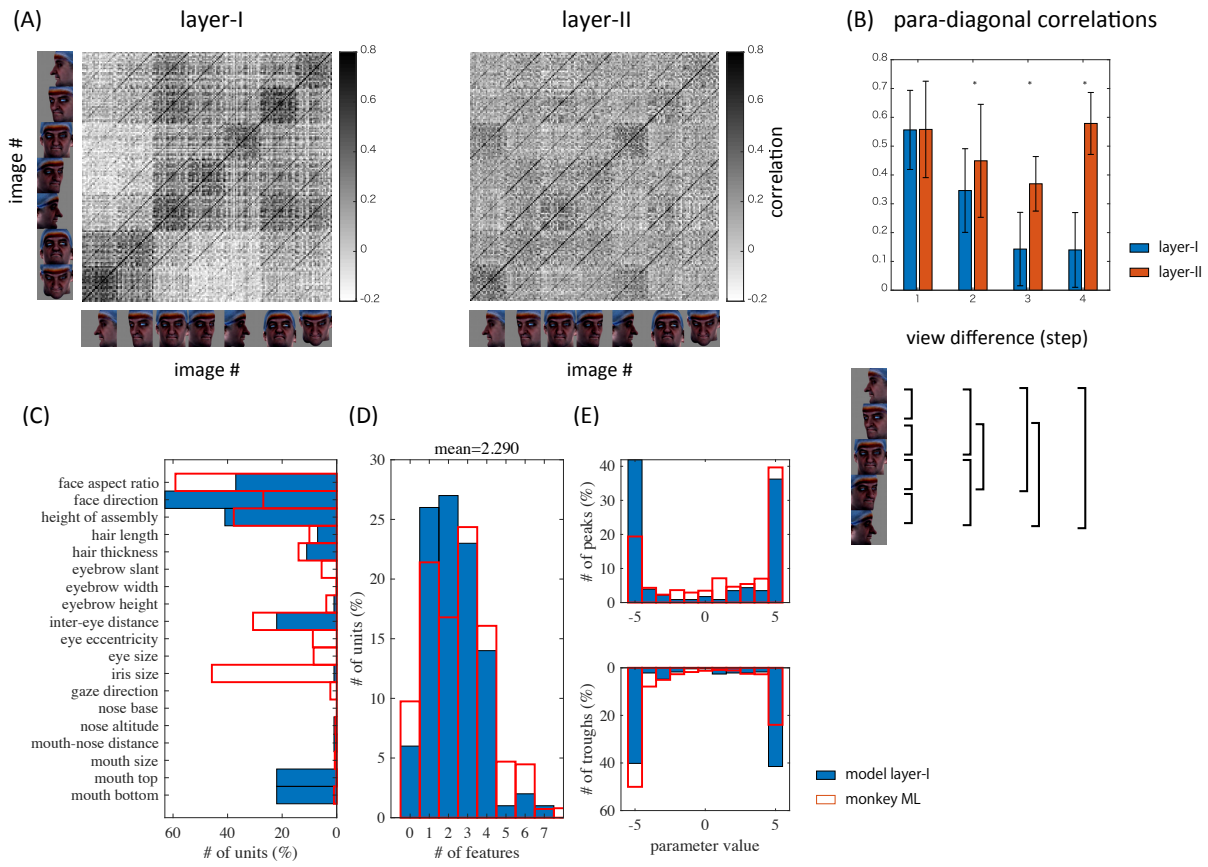


Figure 2: (A) Response similarity matrices for layer-I and layer-II. Each shows the correlation coefficients between the population responses for each pair of images, where the images of the same view are grouped together. (B) Para-diagonal correlations. Each bar shows the mean (with s.d.) of the correlations between the images of the same identity with different views (only horizontal views). View difference ranges from nearest views (1) to farest views (4). Blue: layer-I, red: layer-II. $*$: statistical significance (t-test; $p < 0.01$). (C-E) Comparison of layer-I and monkey ML area in terms of the experiment using cartoon face images (Freiwald et al., 2009). Shown are (C) the number of tuned units for each feature, (D) distribution of the number of tuned features per unit, and (E) distributions of peak (top) and trough (bottom) feature values. Blue: layer-I, red: monkey ML.

images having the same identity but possibly with different views; we fed groups of such images to layer-I and trained layer-II on the outputs of layer-I. We do not specify here the precise architecture of each forward or generative network, but the details of the architecture generally do not affect much the result.

In the sequel, when testing our trained model with a new input $t$, we refer to the value of $x = e(t)$ as the response of layer-I and to the value of $z = h(x)$ as the response of layer-II (the identity variable). However, to compare with actual neural responses, we apply the soft half-wave rectifier $\phi(a) = \log(1 + e^a)$ to each unit response to ensure non-negativity.

## Results

We have investigated the property of our trained two-layer model with respect to two physiological findings on the macaque face processing system (Freiwald & Tsao, 2010; Freiwald et al., 2009). We compared layer-I with area ML and layer-II with area AM.

To simulate the experiment in (Freiwald & Tsao, 2010), we presented, to our model, a set of test face images consisting of 25 identities and 7 views (eliding the back view) that were generated from the aforementioned 3D head model, but separately from the training set (thus, the identities are new). Then, we calculated the correlation between the population activities for each pair of test images at layer-I or II. Figure 2A shows the correlation matrix for each layer, where the image numbers are grouped according to the view. We can see a block-diagonal structure as the most prevalent feature in layer-I, indicating view-specificity. In layer-II, on the other hand, such block diagonal disappears while a para-diagonal structure becomes much more prominent, showing identity-selectivity. These results are similar to areas ML and AM, respectively, as shown in Fig. 4D and F of (Freiwald & Tsao, 2010). Although layer-I also shows

para-diagonal lines, the magnitudes are relatively weak in particular for the correlations between distant views (Figure 2B). Although the experimental study also showed mirror-symmetric view tuning in area AL (Freiwald & Tsao, 2010), we could not find such property in any layer of our network including intermediate layers of the forward or generative networks.

We have also simulated the experiment in (Freiwald et al., 2009) using cartoon face images that are composed of 19 feature parameters (all ranging between $-5$ and $+5$). We measured responses of each unit at layer-I while presenting randomly chosen cartoon face images. Following the analysis method described in (Freiwald et al., 2009), we estimated, for each unit, a tuning curve for each feature parameter and determined its statistical significance using their criterion. Figures 2C and D show the number of units tuned to each feature parameter and the distribution of the number of tuned features per unit, respectively. The results from the model (blue) and from the macaque ML area (red) are generally similar: most units represent a small number of geometrically large features. The discrepancy in the iris size representation is, however, easily noticeable. Figure 2E shows the distributions of the peak (top) or trough (bottom) parameter values. In both the model and the experiment, the tuning curves are maximum or minimum for the extreme feature parameter. Thus, layer-I exhibited tuning properties quite similar to area ML in terms of this experiment. We also tested layer-II with the same experiment, but did not observe such tuning.

## Conclusion

We have presented a novel deep generative model that explains some major properties of face-processing areas ML and AM in the monkey IT. Unlike most prior studies, our model here is based on a Bayesian probabilistic framework with explicit feedforward and feedback computations, which would lead to future investigation of the multi-node reciprocal processing in the face processing network.

Our model lacks a layer corresponding to area AL, even in any intermediate layer of neural networks. In (Freiwald & Tsao, 2010), they raised two possibilities for how the view invariance might emerge in AM, either gradually from ML through AL, or directly from ML. In one sense, our model favors the second possibility, although the role of AL then remains unclear.

## Acknowledgments

This work has been supported by the Commissioned Research of National Institute of Information and Communications Technology (1940201), the New Energy and Industrial Technology Development Organization (P15009), and Grants-in-Aid for Scientific Research (18H05021, 18K11517, and 19H04999).

## References

Dai, H., Pears, N., Smith, W., & Duncan, C. (2017). A 3D Morphable Model of Craniofacial Shape and Texture Variation. *ICCV*, 3104–3112.

Einhäuser, W., Hipp, J., Eggert, J., Körner, E., & König, P. (2005, July). Learning viewpoint invariant object representations using a temporal coherence principle. *Biological cybernetics*, *93*(1), 79–90.

Farzmahdi, A., Rajaei, K., Ghodrati, M., Ebrahimpour, R., & Khaligh-Razavi, S.-M. (2016). A specialized face-processing model inspired by the organization of monkey face patches explains several face-specific phenomena observed in humans. *Scientific Reports*, *6*, 25025.

Földiák, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*(2), 194–200.

Freiwald, W. A., & Tsao, D. Y. (2010, November). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–851.

Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009, August). A face feature space in the macaque temporal lobe. *Nature Neuroscience*, *12*(9), 1187–1196.

Hosoya, H. (2012, August). Multinomial Bayesian learning for modeling classical and nonclassical receptive field properties. *Neural Computation*, *24*(8), 2119–2150.

Hosoya, H. (2019). Group-based learning of disentangled representations with generalizability for novel contents. In *International joint conference on artificial intelligence.*

Hosoya, H., & Hyvärinen, A. (2017, July). A mixture of sparse coding models explaining properties of face neurons related to holistic and parts-based processing. *PLoS Computational Biology*, *13*(7), e1005667.

Hyvärinen, A., Hurri, J., & Hoyer, P. O. (2009). *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.* (Vol. 39). Springer.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In *International conference on learning representations.*

Leibo, J. Z., Liao, Q., Anselmi, F., Freiwald, W. A., & Poggio, T. (2017, January). View-Tolerant Face Recognition and Hebbian Learning Imply Mirror-Symmetric Neural Tuning to Head Orientation. *Current biology : CB*, *27*(1), 62–67.

Moeller, S., Freiwald, W. A., & Tsao, D. Y. (2008, June). Patches with Links: A Unified System for Processing Faces in the Macaque Temporal Lobe. *Science*, *320*(5881), 1355–1359.

Olshausen, B. A., & Field, D. J. (1997, December). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research*, *37*(23), 3311–3325.

Yildirim, I., Kulkarni, T. D., & Freiwald, W. A. (2015). Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. *Annual Conference of the Cognitive Science Society.*.