frontiers
in Computational Neuroscience

ORIGINAL RESEARCH
published: 19 November 2021
doi: 10.3389/fncom.2021.754587

Photorealistic Reconstruction of Visual Texture From EEG Signals

Suguru Wakita[1]*, *Taiki Orima*[1,2] *and Isamu Motoyoshi*[1]

[1] *Department of Life Sciences, The University of Tokyo, Tokyo, Japan,* [2] *Japan Society for the Promotion of Science, Tokyo, Japan*

Recent advances in brain decoding have made it possible to classify image categories based on neural activity. Increasing numbers of studies have further attempted to reconstruct the image itself. However, because images of objects and scenes inherently involve spatial layout information, the reconstruction usually requires retinotopically organized neural data with high spatial resolution, such as fMRI signals. In contrast, spatial layout does not matter in the perception of "texture," which is known to be represented as spatially global image statistics in the visual cortex. This property of "texture" enables us to reconstruct the perceived image from EEG signals, which have a low spatial resolution. Here, we propose an MVAE-based approach for reconstructing texture images from visual evoked potentials measured from observers viewing natural textures such as the textures of various surfaces and object ensembles. This approach allowed us to reconstruct images that perceptually resemble the original textures with a photographic appearance. The present approach can be used as a method for decoding the highly detailed "impression" of sensory stimuli from brain activity.

Keywords: visual texture, multimodal variational auto encoder (MVAE), DNN (deep neural network), brain decoding, EEG

OPEN ACCESS

Edited by:
Ivan Volosyak,
Rhine-Waal University of Applied Sciences, Germany

Reviewed by:
Ko Sakai,
University of Tsukuba, Japan
Fabiano Baroni,
Autonomous University of Madrid, Spain

****Correspondence:***
Suguru Wakita
suguruwakita.ac@gmail.com

Received: *06 August 2021*
Accepted: *25 October 2021*
Published: *19 November 2021*

Citation:
Wakita S, Orima T and Motoyoshi I (2021) Photorealistic Reconstruction of Visual Texture From EEG Signals.
Front. Comput. Neurosci. 15:754587.
doi: 10.3389/fncom.2021.754587

INTRODUCTION

In the field of neuroscience, an increasing number of studies have been conducted to estimate perceptual content and psychological states by extracting certain statistical patterns from brain activity data (Kamitani and Tong, 2005; Schwartz et al., 2006; Miyawaki et al., 2008; Carlson et al., 2011; Green and Kalaska, 2011; Nishimoto et al., 2011). A number of "brain decoding" techniques that identify the object category of an image from the fMRI-BOLD signal have been reported (Shenoy and Tan, 2008; Das et al., 2010; Wang et al., 2012; Carlson et al., 2013; Stewart et al., 2014; Kaneshiro et al., 2015). In recent years, ambitious attempts have been made to reconstruct the image itself from brain activity (Palazzo et al., 2017; Shen et al., 2019a,b). For instance, Shen et al. (Shen et al., 2019a) proposed a method of decoding visual features for each hierarchical stage of visual information processing from an fMRI signal using a deep neural network (DNN) (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015; He et al., 2016) and successfully reconstructed not only the presented image but also the image that an observer imagined in her/his mind.

While excellent decoding is supported by the big data of fMRI, the scope of application is limited by the high costs and potential invasiveness of fMRI. To overcome this limitation, several studies adopted EEG, which provides an easy, cheap, and non-invasive way to collect brain activity data. Palazzo et al. (2017) introduced a method for reconstructing the image of an object from

EEG signals by converting the EEG signals into features and conditioning generative adversarial networks (GANs) (Goodfellow et al., 2014) by it. This approach allowed them to reconstruct an image that can be correctly classified into the original object category [EEG classification accuracy: 84%, Inception Score (IS): 5.07, Inception Classification accuracy (IC): 0.43]. However, as pointed out by the authors themselves, their result is a product of a generative model conditioned by categorical information extracted from an EEG signal and not the direct reconstruction of the image itself actually given to the observer. It is evident that this method fails to reproduce aspects of the perceptual realism of an image, such as the detailed shape, sharp contours, and textures. This limitation seems unavoidable considering the small data size of EEG signals, especially in terms of spatial resolution.

Against the above background, it is of interest to explore the use of texture images in decoding from EEG signals. The perception of a texture is based on spatially global image statistics (Julesz, 1965; Heeger and Bergen, 1995; Portilla and Simoncelli, 2000; Landy and Graham, 2004; Freeman and Simoncelli, 2011), and it is even possible to synthesize perceptually similar texture images using only those statistics (Portilla and Simoncelli, 2000). Such statistical information is represented in the low- and mid-level visual cortex, such as V1, V2, and V4 (Freeman et al., 2013; Okazawa et al., 2015, 2017; Ziemba et al., 2019), and used in the rapid perception of scenes, objects, and surface materials (Thorpe et al., 1996; Oliva and Torralba, 2001; Motoyoshi et al., 2007; Rosenholtz et al., 2012; Whitney et al., 2014). In convolutional neural network (CNN), which computationally mimics neural processing in the ventral stream of the visual brain, the spatially global information obtained by the Gram matrix transformation of features extracted from each hierarchical layer stage corresponds to texture representation (Gatys et al., 2015, 2016).

According to these findings, it is expected that texture can be reconstructed from EEG signals by estimating the information that correlates with the spatially global statistics for texture representation. In fact, the recent study (Orima and Motoyoshi, 2021) were able to estimate lower-order image statistics from visual evoked potentials (VEPs) using a linear regression model and synthesize the texture images with identical image statistics. Using the Image-VEP dataset collected in that study, the present paper proposes a CNN-based method that allows a high quality of reconstruction of the original texture image from a VEP for a variety of natural textures.

MATERIALS AND METHODS

Texture perception is essentially based on the visual appearance, or impression, of an image according to the continuous perceptual similarity, rather than categorical conceptual knowledge as required for object recognition. From this view, we specifically adopted an MVAE-based approach (Suzuki et al., 2017; Wu and Goodman, 2018; Kurle et al., 2019; Shi et al., 2019; Tsai et al., 2019) that acquires a continuous latent representation shared by a texture image and EEG signal. Using the trained MVAE model, we attempted to reconstruct the texture image from the latent variables obtained when only one-modality information, EEG data, was input.

In our approach, the MVAE model is trained with the texture images and VEP as two-modality information. After training, the latent space shared by the two modalities is acquired in the model. Finally, the test texture image is reconstructed from the latent variable obtained from the corresponding EEG signals input to the trained model.

EEG Measurement

In training the model, we used the dataset obtained by Orima and Motoyoshi (2021). The dataset comprises EEG signals for 166 natural texture images, with each signal measured for a period of 500 ms, 24 times, for each of 15 human observers. **Figure 1** shows examples of texture images used in EEG measurements.

Visual stimuli were images of 166 natural textures subtending 5.7 deg × 5.7 deg (256 × 256 pixels). The images were collected from the Internet and our own image database. Each image was achromatic and had a mean luminance of 33 cd/m^2. In each of 24 measurement blocks, 166 images were presented in random order for 500 ms followed by a 750-ms blank that is equal to a uniform gray background and 15 observers viewed each image with their eyes steadily fixed at the center of the image. During each block, the VEP was measured using 19 electrodes (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, P8, Fz, Cz, and Pz according to the international 10/20 method; BrainVision Recorder, BrainAmp Amplifier, EasyCap; Brain Products GmbH) at a 1,000-Hz sampling rate. All stimuli were presented on a gamma-corrected LCD (BENQ XL2420T). The refresh rate of the LCD was 60 Hz, and the spatial resolution was 1.34 min/pixel at an observation distance of 100 cm. All measurements were conducted in accordance with the Ethics Committee for Experiments on Humans at the Graduate School of Arts and Sciences, The University of Tokyo. The participants completed a written consent form.

Multimodal Variational Auto Encoder for Image Reconstruction From EEG Signals

Considering the continuous and variegated nature of natural textures as visual information, we consider a variational auto encoder (VAE) -based (Kingma and Welling, 2013) approach in which the texture images and the corresponding EEG signals are represented in a continuous latent space.

The VAE is a deep generative model that conducts its generation process by deep learning assuming the existence of a latent variable z when data v are observed (Kingma and Welling, 2013; Kingma et al., 2014; Dai et al., 2015; Krishnan et al., 2015). Here, by assuming that latent variables are represented on a probabilistic distribution space, we can perform continuous representation learning on the observed input data (Equation 1).

$$z \sim p(z) = N(0,\ I),\ v\ \sim p_\theta(v \mid z) \tag{1}$$

In the VAE, the observed input data v are transformed by the encoder into a contractive intermediate representation called latent variable z, and the decoder reconstructs the original input

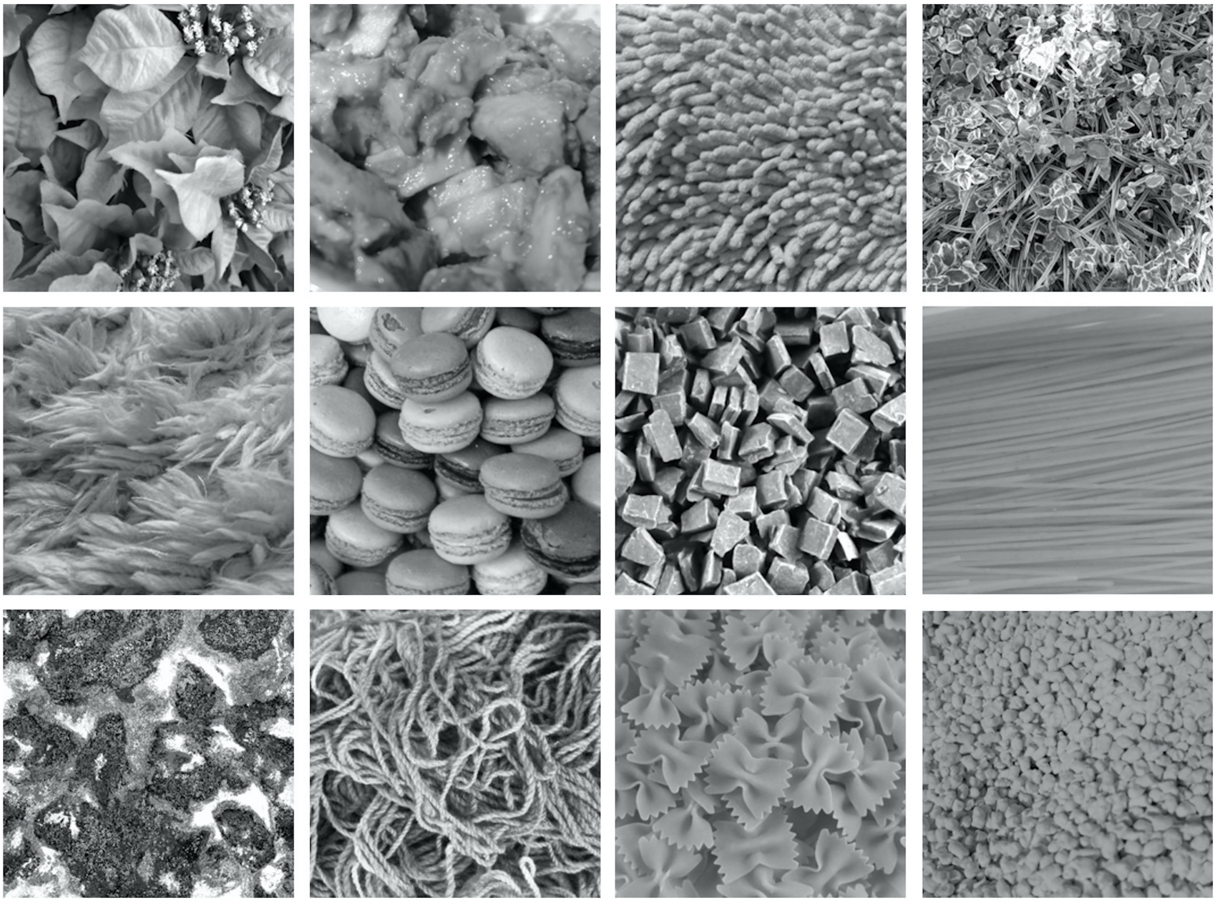

FIGURE 1 | Examples of texture images used in the EEG measurement.

data v' with this latent variable as input. The entire model is trained so as to minimize the difference between the input data v and the reconstructed data v', and the model parameters of the encoder and decoder are updated. The encoder and decoder comprise a neural network. [In the following, Φ and θ refer to the model parameters of the encoder and decoder, respectively, and the multivariate Gaussian distribution is denoted p(z)].

More practically, the target of training is to maximize the marginal likelihood $p_\theta(v)$, but because this cannot be treated directly, we optimize the model parameters of the encoder $q_\phi(z \mid v)$ and decoder $p_\theta(v|z)$ to maximize the evidence lower bound (ELBO) given in Equation 2.

In Equation 2, the first term on the right-hand side is called the regularization term. This term regularizes the latent variable z, which is obtained from the mean vector μ and variance vector σ output by the encoder, to distribute according to prior $p(z)$.

The second term on the right-hand side is the reconstruction error term, which minimizes the difference between the original input data v and v', the input data reconstructed from the decoder using the latent variable z. β and λ are weight parameters.

$$ELBO(v) = -\beta D_{KL}(q_\phi(z|v) \,|p(z)) + E_{q_\phi(z|v)}[\lambda \log p_\theta(v|z)] \tag{2}$$

As an extension of the VAE, the multimodal VAE, which treats multimodal information as input, has been proposed (Suzuki et al., 2017; Wu and Goodman, 2018; Kurle et al., 2019; Shi et al., 2019; Tsai et al., 2019). This extension is inspired by the fact that our cognition in the real world uses multimodal information, not unimodal information (Ngiam et al., 2011; Srivastava and Salakhutdinov, 2012; Kiros et al., 2014; Pandey and Dukkipati, 2016). In fact, it is generally known that learning with multimodal information induces the acquisition of better informative representations compared with the case of unimodal information (Ngiam et al., 2011; Srivastava and Salakhutdinov, 2012).

In this study, we apply the extended method for the MVAE (Wu and Goodman, 2018), which allows inference of latent variables even under the partial observation of multimodal information aiming at reconstructing texture images only from EEG signals.

Here, the texture images and EEG signals are treated as different information modalities, and the latent representation shared by these two modalities is acquired by the learning MVAE. As a result of this training, the stimulus can be reconstructed by decoding the texture image using latent variables acquired by the input of a single modality, the EEG signal. **Figure 2** is

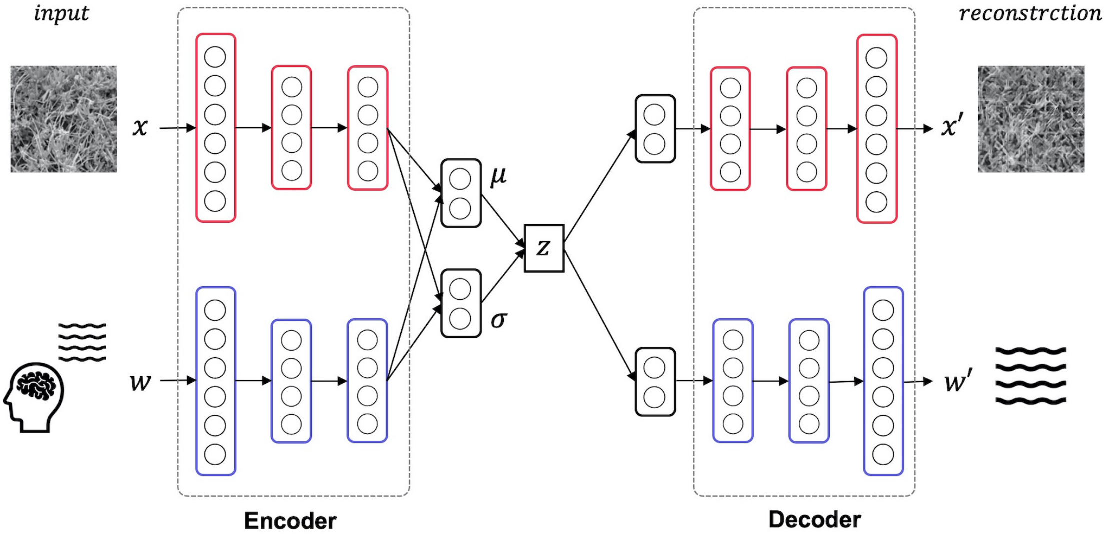

FIGURE 2 | Model overview. The MVAE model comprises an encoder and decoder for the EEG signal modality and an encoder and decoder for the texture stimulus modality. After training this model, the latent space shared by the two modalities is acquired. Thus, the texture stimulus that was presented in the EEG measurement can be reconstructed from the latent variable obtained by inputting only the EEG signals to this trained model.

an overview of the structure of the MVAE model. The MVAE model comprises an encoder and decoder for the EEG signal modality and an encoder and decoder for the texture image modality. By inputting one or both modalities of information into the encoder corresponding to the respective modal information, mean vector μ and variance vector σ can be inferred for the Gaussian distribution. The mean vector μ and variance vector σ obtained here are integrated into, and represented as, a single latent variable using the product of experts (PoE) (Hinton, 2002). Finally, the reconstructed results of EEG signals and texture images are obtained by inputting this latent variable to each of the decoders corresponding to each modal information.

In the training of the MVAE model, there are three possible patterns for the combination of the observable modal information (Here, we denote the texture image modal information as x, the EEG signal modal information as w, and the observed whole or partial modal information as $V = \{v : v_1,\ v_2, ...v_n\}\}$.)

- $V_1 = \{v : x,\ w\}$: Both modalities, the EEG signal and texture image, can be observed.
- $V_2 = \{v : x\}$: One modality, the texture image, can be observed.
- $V_3 = \{v : w\}$: One modality, the EEG signal, can be observed.

In reconstructing texture images from EEG signals, which is the target of this study, it is necessary to obtain a representation in the latent space shared by the two modalities of texture images and EEG signals, and to be able to extract latent variables under the partial modal observation (V_2, V_3) that are as good as or similar to those extracted under the full modal observation (V_1). Considering this point, we maximize the ELBO expressed in Equation 3, proposed by Wu and Goodman (2018), in our training.

$$ELBO\,(V) = -\beta D_{KL}(q_{\phi}\,(z|V)\ |p\,(z)) + E_{q_{\phi}(z|V)} \left[\sum_{v_i \in V} \lambda_i \log p_{\theta}(v_i|z)\right] \tag{3}$$

The loss function of the entire model is thus expressed by Equation 4, and the training proceeds accordingly.

$$loss = \sum_{V_i \in \{V_1, V_2, V_3\}} -ELBO(V_i) \tag{4}$$

One issue that should be considered here is that the image reconstructed using the VAE-based approach is generally blurred. When we tested the reconstruction with the simple VAE using texture images, we found that the reconstruction of fine texture components did not work well, resulting in grayish or blurred images. This is a crucial issue in the present study because we are aiming to realize texture reconstruction with visual similarity to the texture stimuli presented to the observers during EEG measurement. As a solution to such problems, a method combining a VAE and GAN (Rosca et al., 2017) has been proposed to generate natural images and general object images more realistically. However, in the present paper, it is necessary to devise a loss function that improves the reproduction for such texture components when we consider that we use natural texture images in the present study and particularly when failing to reconstruct fine and relatively high spatial frequency components. We thus considered applying precedent knowledge gained in the field of neural style transfer (Gatys et al., 2016; Johnson et al., 2016; Ulyanov et al., 2016; Huang and Belongie, 2017), where texture synthesis is conducted using trained DNN (Gatys et al., 2015).

In general, VGG is used in the implementation of neural style transfer. VGG is a representative DNN that achieved excellent performance in the ILSVRC 2014 (ImageNet Large Scale Visual Recognition Challenge), showing that deepening the layers of the CNN contributes to improved classification accuracy in object recognition tasks (Simonyan and Zisserman, 2015). According to the knowledge in this field, in the trained VGG-19 model (Simonyan and Zisserman, 2015), style information at different levels of abstraction is processed at each stage of the hierarchical processing, and we can extract fine style information at the lower layers and global style information at the higher layers. Additionally, this style information is sufficient for accurate style transferring and texture synthesis (Gatys et al., 2015, 2016; Johnson et al., 2016; Ulyanov et al., 2016; Huang and Belongie, 2017). We therefore use this style information in our approach for more precise texture reconstruction. Specifically, we replace the reconstruction error term in the ELBO with a combination of original reconstruction error term and style error term, which is commonly used in the framework of neural style transfer. The style error is expressed in Equation 5. Here, we denote the input image as x, reconstructed image as x', and set of layers in the trained VGG-19 from which the style information can be extracted as $\mathrm{L} = \{1,\ 2, ..., \mathrm{k}\}$. Style information obtained by Gram matrix transformation of the output from each layer is denoted $\mathrm{G}_x{=}\{\mathrm{G}^{L=1}, \mathrm{G}^2, \ldots, \mathrm{G}^k\}$ and $\widehat{\mathrm{G}}_{x'}{=}\{\widehat{\mathrm{G}}^{L=1}, \widehat{\mathrm{G}}^2, \ldots, \widehat{\mathrm{G}}^k\}$ for the input image and reconstructed image respectively. α is the weighting of style information in each layer, N is the number of filter maps in each layer of VGG-19, and M is the number of elements in each filter map in each layer. i, j denote the index of the vectorized feature map in layer l.

$$\mathrm{StyleLoss}\left(x, x^{'}\right) = \sum_{l=1}^{\mathrm{L}} \alpha_l \frac{1}{4N_l^2 M_l^2} \sum_{i,j} (\mathrm{G}_{ij}^l - \widehat{\mathrm{G}}_{ij}^l)^2 \qquad (5)$$

Applying this style error for the loss function confirmed that the texture pattern can be reconstructed clearly regardless of the spatial frequency of the texture in the input image.

Psychophysical Experiment Setup

In validating the reconstruction results, we carried out a behavioral experiment to examine the relative perceptual similarity of the reconstructed texture to the original. In our display, the original natural texture (2.6 × 2.6 deg, 128 × 128 pixels) was presented at the center, and reconstructed textures were presented on the left and right, 3.5 deg from the center. One reconstructed texture was the target image reconstructed from EEG signals for the central original texture and the other was the non-target image reconstructed from EEG signals for another texture that was chosen randomly from the remaining 165 textures. Six observers with normal or corrected-to-normal vision viewed the stimuli with a free gaze and indicated the texture image (left/right) that was perceptually more similar to the central original texture. Observers were strongly instructed to evaluate the similarity in terms of the visual appearance and not in terms of the categorical meaning. This evaluation was performed on texture images reconstructed using each model trained with the stratified k-fold cross validation (k = 10). For each observer, at least five data for each of the texture images were collected and the probability for each texture image of a response that "the target appeared more similar" was calculated. All experiments were conducted using gamma-corrected LCDs with a refresh rate of 60 Hz (BenQ2720T, SONY PVM-A250, BENQ XL2730Z, BENQ XL2730Z, BENQ XL2730Z, and BENQ XL2735B), each of which was installed in a dark room of the individual observer's home owing to the COVID19 situation. The viewing distance was adjusted so that the spatial resolution was 1.0 min/pixel. Other parameters were the same as those in the EEG measurements.

RESULTS

Multimodal Variational Auto Encoder Model and Training

The MVAE model was trained using a dataset consisting of 166 natural texture images and EEG signals for those images obtained from Orima and Motoyoshi (2021). The dataset was divided into 10 partitions, and stratified k-fold cross-validation (k = 10) was performed. Each partition contained the EEG signals for each of the 166 texture images. For evaluation, we conducted psychological experiments using texture images reconstructed by inputting the test data set into the models trained in each cross-validation. In this experiment, we quantitatively evaluated whether the reconstructed image had a high visual similarity to the original image according to the human eye. Measurements of the EEG signals, which were made every 1 ms for 500 ms after the stimulus onset when the texture stimulus was presented to the observer, were taken as the values of 500-dimensional vector data. When we input the EEG signals to the MVAE model, 25–30 samples of EEG signals corresponding to one particular texture stimuli were selected in random combinations and their average waveforms were obtained. Then, for each average waveform, we normalized the maximum value to be 1 and the minimum value to be zero. Among the electrode channels used in the EEG measurement, the signals measured at Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, and P8 in the international 10/20 method were used as input. Additionally, texture images, as the other information modality, was resized to 128 × 128 on input. At this time, the reconstructed texture image was also output as a 128 × 128 image. In the training, the Adam gradient descent method was used with a learning rate of 1e-4. The batch size was 16. The vector size for the latent variable of the MVAE model was 256. The MVAE model comprises an encoder and decoder that treat the texture images as modal information and an encoder and decoder that treat the EEG signals as modal information. We used 2D-convolution for the encoder and decoder that treat the texture images as modal information, and 1D-convolution for the encoder and decoder that treat the EEG data as modal information. The architectural details of the MVAE model are given in **Table 1**. In the table, Conv{n}d and UpConv{n}d denote the convolution layer and transposition convolution layer, respectively. n refers to the dimension. The parameter of the convolution (Conv, UpConv) layer is denoted by "Up/Conv{n}d- {kernel size}- {number of channels}-{stride}." AvgPool refers to average pooling. FC refers

TABLE 1 | Details of the model architecture.

Texture-image modal		EEG brain signal modal	
Encoder	**Decoder**	**Encoder**	**Decoder**
Input 1 × 128 × 128	UpConv2d-3-256-1	Input 16 × 500	FC-256
Conv2d-1-32-1	UpConv2d-2-128-1	Conv1d-3-128-2	UpConv1d-3-256-1
ResNetBlock-3-32-1	ResNetBlock-3-128-1	Conv1d-1-128-1	UpConv1d-4-128-2
AvgPool2d	UpConv2d-4-128-1	Conv1d-1-128-1	UpConv1d-4-128-2
Conv2d-1-64-1	ResNetBlock-3-64-1	Conv1d-1-128-1	UpConv1d-4-64-2
ResNetBlock-3-64-1	UpConv2d-4-64-1	Conv1d-3-256-2	UpConv1d-4-64-2
Conv2d-1-128-1	ResNetBlock-3-32-1	Conv1d-1-256-1	UpConv1d-4-32-2
ResNetBlock-3-128-1	UpConv2d-1-32-1	Conv1d-1-256-1	UpConv1d-4-32-2
Conv2d-1-256-1	(Sigmoid)	Conv1d-1-256-1	UpConv1d-4-1-2
FC-256 (mean), FC-256(var)	Output 1 × 128 × 128	Conv1d-3-512-2	FC-638
		Conv1d-1-512-1	Output 16 × 500
		Conv1d-1-512-1	
		Conv1d-1-512-1	
		Conv1d-1-1024-1	
		Conv1d-1-256-1	
		GlobalAvgPool1d	
		FC-256 (mean), FC-256(var)	

to a fully connected layer, and the parameter is FC- {size of each input sample}. ResNetBlock is a convolutional module that can be applied to Reflection padding (where the size of the padding is 1), Convolution, Batch Normalization, and ReLU processes are conducted twice. The parameters of ResNetBlock are given as "ResNetBlock- {kernel size}- {number of channels}-{stride}." In actual implementation, except for the final output layer, each convolution layer is followed by batch normalization and ReLU rectifier processing in order.

Reconstruction of the Texture Image

After training the MVAE model, we reconstructed the texture image using the test EEG signals as input. More specifically, the latent variables were extracted from the encoder that treats the EEG signals as modal information, and the texture images were reconstructed by inputting these latent variables to the other decoder that treats the texture image as modal information.

Figure 3 shows examples of reconstructed images. In each row, the upper images show the original textures, and the lower images show the textures reconstructed from EEG. It is seen that most of the reconstructed textures are remarkably photorealistic, and some are similar to the original textures. The quality of reconstruction is much higher than that of texture synthesis based on linear regression reported in our previous study (Orima and Motoyoshi, 2021).

Psychophysical Experiment

We conducted the psychological experiment described in section "Psychophysical Experiment Setup" to validate the reconstruction results. Texture images are evaluated in terms of their continuous perceptual appearance, whereas general object images are evaluated based on their categorical semantic classification. Therefore, in this psychological experiment, we instructed the observers to select the reconstructed texture image that is closer to the original image in pure visual appearance without being confined to the categorical classification.

We prepared 10 samples of reconstructed texture images for each of the 166 different texture images. Each sample for a particular texture image was reconstructed from each of the 10 models in the stratified k-fold cross validation ($k = 10$). Six observers participated in the experiment, and each of the observers followed the experimental procedure five times in evaluating the reconstruction result for each of the 166 textures. After numbering the 10 samples reconstructed from each cross-validation model in order (1, 2, 3,..., 10), we assigned odd-numbered samples to three observers and even-numbered samples to the remaining three observers. Each observer performed 830 trials (166 trials, five sessions), and the total number of trials for all observers was thus 4,980.

The results of the experiment show that the correct identification rate in all trials was 70.1%, which was significantly higher than the chance level (50%) based on the binomial test ($p \ll 0.001$). For all the six individual observers, we found that the correct identification rate for 166 textures was significantly higher than the chance level (50%) in a one-tailed t-test ($t(165) > 9.77,\ p < 4.41e - 18$). Together with the observations presented in **Figure 3**, these results suggest that the reconstruction was successful.

For more analysis, **Figure 4** shows the probability of a response that "the target appeared more similar" averaged across the six observers for each of 166 textures. The horizontal axis is the index of the texture image, sorted from the left in descending order of the proportion correct. The horizontal red line denotes the chance level (50%). The asterisks indicate that the average identification rate across the six observers for that texture is statically significant in a one-tailed t-test ($p < 0.05$).

While our method has performed with a certain degree of success, the reconstruction is limited to the textures used for training the MVAE model. The establishment of a more versatile reconstruction approach requires the consideration of the possibility of reconstruction for unknown novel textures that were not used in the model training phase. We therefore considered conducting a limited test of reconstruction on unknown novel textures. However, it should be explicitly stated at the outset that the validation in this limited test was not sufficient. The dataset used in the present study was collected in our previous study (Orima and Motoyoshi, 2021), which was not carried out for the purpose of brain decoding. Therefore, the dataset was not adequate for considering unknown novel texture reconstruction methods based on sufficient cross-validation. The results presented below are examined under this constraint.

The novel texture reconstruction based on the MVAE-based approach proposed in this study is expected to be

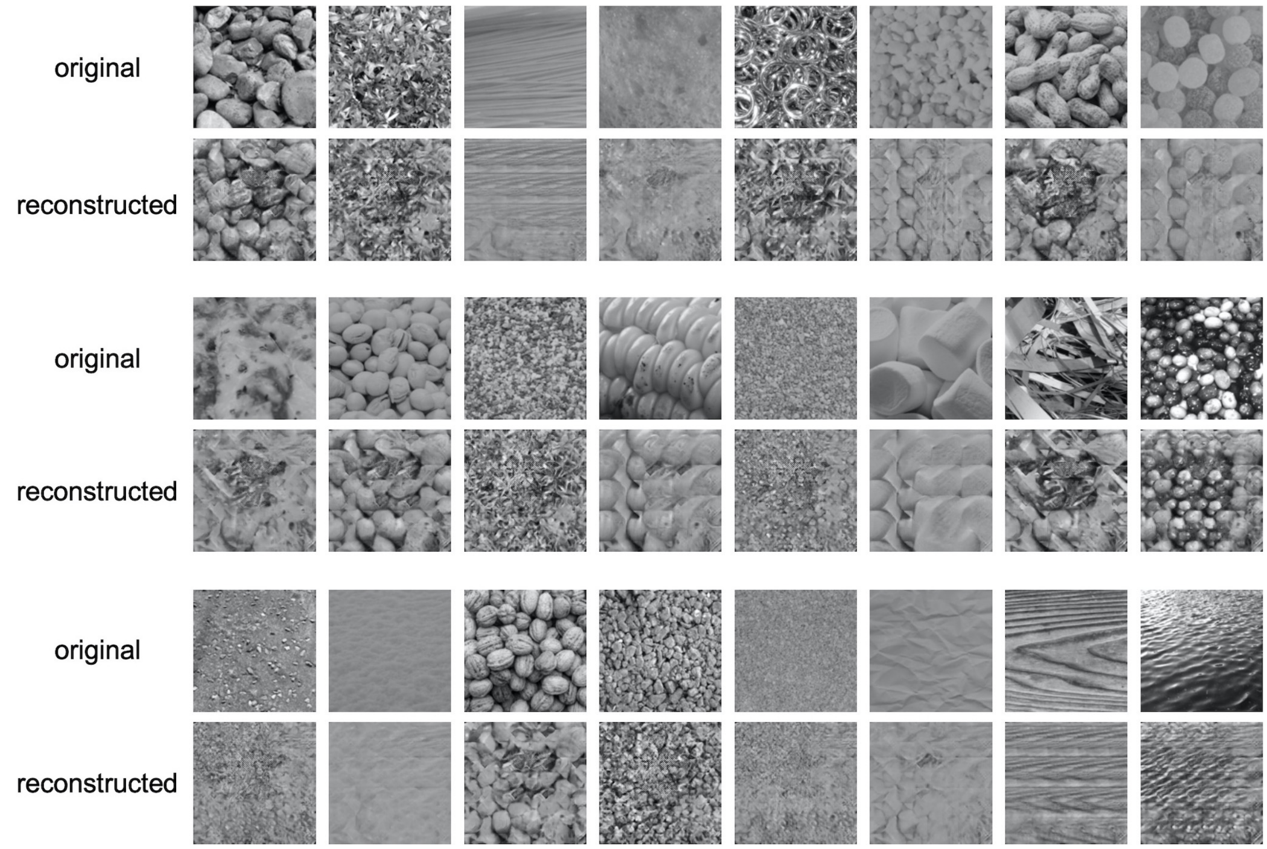

FIGURE 3 | Examples of reconstructed texture images. In each row, the upper images are the original texture images shown to the observer and the lower images are the images reconstructed from EEG.

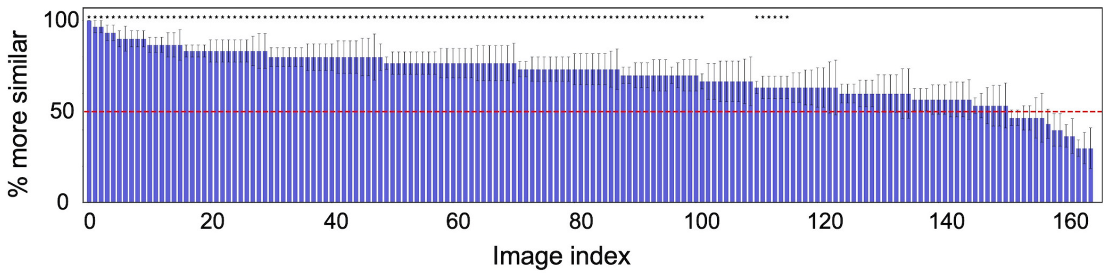

FIGURE 4 | Probability of a response that the image reconstructed from the original texture (target) was more similar than the image reconstructed from another random texture (non-target) to the original image. The horizontal axis shows the index of the texture image, sorted from the left in descending order of the probability. The red line denotes the chance level (50%). The error bars indicate ± 1 s.e.m. across observers. The asterisks indicate that the average identification rate across six observers for that texture is statically significant in a one-tailed *t*-test ($p < 0.05$).

realized using the intermediate representation of other multiple textures on the latent space acquired using a variety of textures according to the nature of the method. Specifically, the acquisition of an internal representation that can represent the new texture is important. This is analogous to the use of visual features in pre-trained DNNs as a proxy for hierarchical visual representation in the brain in visual decoding study with fMRI data (Horikawa and Kamitani, 2017; Shen et al., 2019a). Considering this point, we prepared the dataset in the following manner. We used 140 of the 166 textures for training and 26 for testing in this limited test. In preparing these test textures, we created 83 visually similar pairs from 166 textures based on VGG's style information. Of these pairs, we manually selected 26 pairs that did not overlap in visual impression between the pairs. One of the textures in each of these 26 pairs was picked as the 26 textures for the test dataset. The setting in model training was the same as that in cross-validation training.

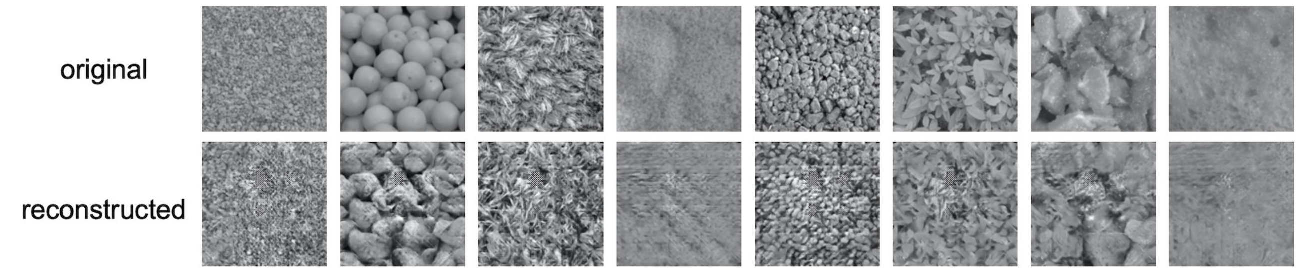

FIGURE 5 | Examples of unknown novel texture reconstruction. The upper images are the original texture images shown to the observer and the lower images are the images reconstructed from EEG.

Figure 5 shows examples of reconstructed images on the limited test. The upper images show the original textures, and the lower images show the textures reconstructed from EEG signals. To evaluate the reconstruction results on novel textures, we conducted a psychological experiment with the same procedure as described in section "Psychophysical Experiment Setup." We prepared five samples of reconstructed texture images for 26 novel texture images in the test dataset. The six observers evaluated the reconstruction results for each of 26 textures (five repetitions for each). We found that the average correct identification rate was 72.8%. For all of six individual observers, the correct identification rate was significantly higher than the chance level (50%) in a one-tailed t-test ($t(25) > 4.18$, $p < 0.0003$). **Figure 6** shows the correct identification rate averaged across the six observers for 26 unknown novel textures, sorted from the left in descending order of the correct identification rate. The horizontal red line denotes the chance level (50%). The asterisks indicate that the identification rate over the six observers for that texture is statically significant in a one-tailed t-test ($p < 0.05$).

DISCUSSION

The present study introduced a method in which an MVAE is used to reconstruct the image of a natural texture from EEG signals alone. Our trained MVAE model successfully reconstructed the original texture with photorealistic quality and greatly outperformed linear regression on the same dataset (Orima and Motoyoshi, 2021).

As mentioned earlier, it is generally challenging to decode neural representations of a natural scene with EEG because of the low retinotopic resolution of EEG as compared with that of fMRI. The present study avoided this limitation by confining the scope to textures for which the perception is determined by spatially global image statistics, and we successfully reconstructed various natural textures from EEG signals. The previous study having a similar scope (Orima and Motoyoshi, 2021) focused on understanding the neural dynamics for image statistics assumed in human texture perception (e.g., Portilla–Simoncelli statistics) and demonstrated a reconstruction of textures using image statistics linearly regressed from EEG signals. In contrast, the present study pursued a technique to reconstruct an image with higher quality and showed that the use of an MVAE allows the reconstruction of textures with high quality.

The previous approach reconstructed natural object images from EEG signals on the basis of the classification of discrete object categories acquired in a supervised network (Palazzo et al., 2017). In contrast, the present study aimed to reconstruct a purely perceptual impression without any dependency on top-down knowledge such as that of categories, by acquiring a continuous representation space of visual textures in a fully unsupervised learning manner. The resulting images duplicated the perceptual impression well. Of course, such success might be possible only for the textures that we used, and it is unclear if the present approach is applicable to a wide range of classes of images, such as images of objects and scenes. However, we believe that the fact that we were able to reproduce images from EEG signals in a highly realistic manner brings a new direction in the decoding of sensory information. We are currently applying the same approach to sounds.

We should also note a limitation of the present approach. **Figure 7** shows the worst examples of texture reconstruction. The upper images show the original texture, and the other images show the image samples reconstructed from EEG signals. These reconstructed images are similar to one of the other textures tested, and there is a large variability among samples for the same original texture. This result is due to the VAE acquiring continuity on visual similarity between the considered textures in the latent space, and therefore, when the proper texture representation was not extracted from the EEG signals, the representation became an intermediate representation that was determined virtually randomly in the space defined by the limited number of textures that we used. As a result, it is highly possible that the reconstructed texture is similar to one of the other original textures. This problem could be avoided with a latent space that is richer with more diverse texture images. However, such a latent space would require many more images and corresponding EEG data.

A more generally applicable texture reconstruction from EEG signals requires consideration of the possibility of reconstruction on unknown novel textures in the model training phase. To investigate this point, we performed a limited study on reconstruction for novel textures by dividing the 166 textures into training and test. While there are challenges for certain textures, such as reconstruction being unstable from sample to sample, the

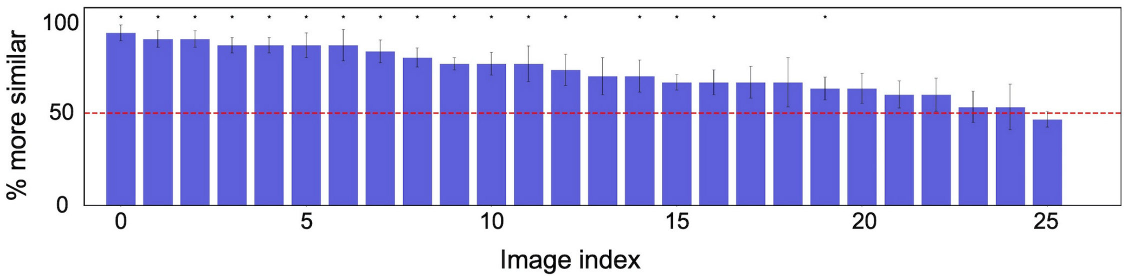

FIGURE 6 | Psychological experimental results on texture images reconstructed from EEG signals corresponding to unknown novel textures in training. The correct identification rate averaged across six observers for 26 unknown novel textures, sorted from the left in descending order of the correct identification rate. The horizontal red line denotes the chance level (50%). The error bars indicate ± 1 s.e.m. across observers. The asterisks indicate that average identification rate across the six observers for that texture is statically significant in a one-tailed *t*-test ($p < 0.05$).

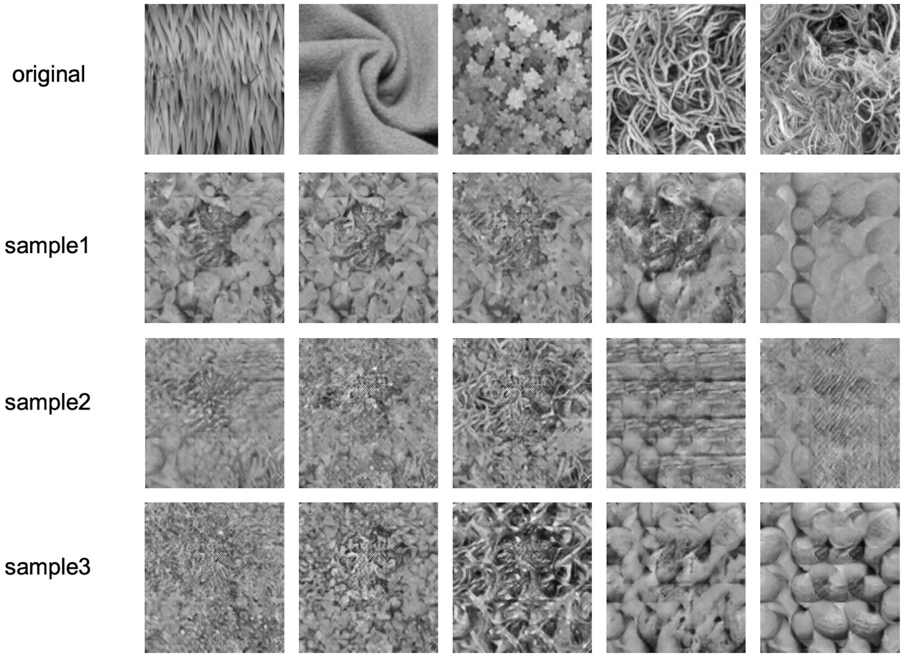

FIGURE 7 | Failure examples of the reconstructed texture images having the lowest "more similar" response rates in the psychophysical experiment. The upper row shows the original textures and the other rows show three sample images reconstructed from EEG signals.

result shows the possibility of reconstruction for novel textures within the framework of the proposed MVAE-based approach. However, it should be noted again that because we used a dataset that measured for a different purpose in our previous study (Orima and Motoyoshi, 2021), we could not examine the results based on sufficient cross-validation in this limited test. This problem needs to be investigated in future work with a sufficiently extended dataset.

As the next advancement in the analysis of texture representation on EEG signals, an investigation of the frequency components on EEG signals may provide new findings. The frequency components that contribute to the reconstruction of a certain texture can be identified by evaluating the reconstruction results with and without stripping certain frequency components in texture reconstruction using the MVAE model. This is expected to be valuable in understanding the correspondence between the frequency bands in texture processing in the brain that correspond to each of the various textural representations.

While the present approach provides an effective tool for reconstructing the visual impression of an image with complex spatial structures from EEG signals, there is still room for improvement. In this study, the model was trained using an EEG signal dataset for 166 texture images, which was divided for cross-validation regardless of the observer. Therefore, we were unable to analyze the characteristics of visual impressions for each observer. As a future development, it would be interesting to conduct research focusing on the differences in visual impressions and visual functions specific to a particular individual, although there are still challenges in measuring sufficient data for each observer and methodological challenges in avoiding inadequate data. In the present study, we focused on the pipeline of reconstructing texture stimuli from EEG signals, but owing to the nature of MVAE-based systems, it is also possible to consider the opposite pipeline through which the EEG signal is reconstructed from an image.

DATA AVAILABILITY STATEMENT

The data supporting the conclusions of this article will be made available by the authors upon reasonable request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee for Experiments on Humans at the Graduate School of Arts and Sciences, The University of Tokyo. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SW, TO, and IM designed the study. TO and IM provided the dataset of images and EEG. SW developed, trained the model for texture reconstruction, performed the psychophysical experiment, and analyzed the data. SW and IM wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This study was supported by the Commissioned Research of NICT (1940101) and JSPS KAKENHI JP20K21803.

REFERENCES

Carlson, T., Tovar, D. A., Alink, A., and Kriegeskorte, N. (2013). Representational dynamics of object vision: the first 1000 ms. *J. Vis.* 13:1. doi: 10.1167/13.10.1

Carlson, T. A., Hogendoorn, H., Kanai, R., Mesik, J., and Turret, J. (2011). High temporal resolution decoding of object position and category. *J. Vis.* 11:9. doi: 10.1167/11.10.9

Dai, Z., Damianou, A., Gonz'alez, J., and Lawrence, N. (2015). Variational auto-encoded deep Gaussian processes. *arXiv* [Preprint]. arXiv:1511.06455,

Das, K., Giesbrecht, B., and Eckstein, M. P. (2010). Predicting variations of perceptual performance across individuals from neural activity using pattern classifiers. *Neuroimage* 51, 1425–1437. doi: 10.1016/j.neuroimage.2010.03.030

Freeman, J., and Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nat. Neurosci.* 14, 1195–1201. doi: 10.1038/nn.2889

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., and Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nat. Neurosci.* 16, 974–981. doi: 10.1038/nn.3402

Gatys, L., Ecker, A. S., and Bethge, M. (2015). "Texture synthesis using convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems* (Montreal, QC: MIT Press), 262–270.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). "Image style transfer using convolutional neural networks," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV: IEEE). doi: 10.1109/cvpr.2016.265

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Proceedings of the Advances in Neural Information Processing Systems*, Montreal, QC.

Green, A. M., and Kalaska, J. F. (2011). Learning to move machines with the mind. *Trends Neurosci.* 34, 61–75. doi: 10.1016/j.tins.2010.11.003

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV: IEEE). doi: 10.1109/cvpr.2016.90

Heeger, D. J., and Bergen, J. R. (1995). "Pyramid-based texture analysis/synthesis," in *Proceedings of the 22nd Annual Conference on Computer graphics and Interactive Techniques– SIGGRAPH '95*, Los Angeles, CA. doi: 10.1145/218380.218446

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.* 14, 1771–1800.

Horikawa, T., and Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nat. Commun.* 8:15037.

Huang, X., and Belongie, S. (2017). "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)* (Venice: IEEE). doi: 10.1109/iccv.2017.167

Johnson, J., Alahi, A., and Li, F. (2016). "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the Computer Vision–ECCV 2016. Lecture Notes in Computer Science*, eds B. Leibe, J. Matas, N. Sebe, and M. Welling (Cham: Springer), 694–711. doi: 10.1007/978-3-319-46475-6_43

Julesz, B. (1965). Texture and visual perception. *Sci. Am.* 212, 38–49. doi: 10.1038/scientificamerican0265-38

Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.* 8, 679–685. doi: 10.1038/nn1444

Kaneshiro, B., Perreau Guimaraes, M., Kim, H.-S., Norcia, A. M., and Suppes, P. (2015). A representational similarity analysis of the dynamics of object processing using single-trial EEG classification. *PLoS One* 10:e0135697. doi: 10.1371/journal.pone.0135697

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). "Semi-supervised learning with deep generative models," in *Proceedings of the Advances in Neural Information Processing Systems* (Montreal, QC: MIT Press), 3581–3589.

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. *arXiv* [Preprint]. arXiv:1312.6114,

Kiros, R., Salakhutdinov, R., and Zemel, R. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv* [Preprint]. arXiv:1411.2539,

Krishnan, R. G., Shalit, U., and Sontag, D. (2015). Deep kalman filters. *arXiv* [Preprint]. arXiv:1511.05121,

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems* (Red Hook, NY: Curran Associates, Inc), 1097–1105.

Kurle, R., Günnemann, S., and Van der Smagt, P. (2019). "Multi-source neural variational inference," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu, HI: AAAI Press), Vol. 33, 4114–4121.

Landy, M. S., and Graham, N. (2004). "73 visual perception of texture," in *The Visual Neurosciences*, eds L. M. Chalupa and J. S. Werner (Cambridge, MA: MIT Press), 1106–1118.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M. A., Morito, Y., Tanabe, H. C., et al. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* 60, 915–929. doi: 10.1016/j.neuron.2008.11.004

Motoyoshi, I., Nishida, S., Sharan, L., and Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature* 447, 206–209. doi: 10.1038/nature05724

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). "Multimodal deep learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, Bellevue, WA.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* 21, 1641–1646. doi: 10.1016/j.cub.2011.08.031

Okazawa, G., Tajima, S., and Komatsu, H. (2015). Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proc. Natl. Acad. Sci. U.S.A.* 112, E351–E360. doi: 10.1073/pnas.1415146112

Okazawa, G., Tajima, S., and Komatsu, H. (2017). Gradual development of visual texture-selective properties between macaque areas V2 and V4. *Cereb. Cortex* 27, 4867–4880. doi: 10.1093/cercor/bhw282

Oliva, A., and Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 145–175.

Orima, T., and Motoyoshi, I. (2021). Analysis and synthesis of natural texture perception from visual evoked potentials. *Front. Neurosci.* 15:876. doi: 10.3389/fnins.2021.698940

Palazzo, S., Spampinato, C., Kavasidis, I., Giordano, D., and Shah, M. (2017). "Generative adversarial networks conditioned by brain signals," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)* (Venice: IEEE). doi: 10.1109/iccv.2017.369

Pandey, G., and Dukkipati, A. (2016). Variational methods for conditional multimodal learning: generating human faces from attributes. *arXiv* [Preprint]. arXiv:1603.01801,

Portilla, J., and Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* 40, 49–70.

Rosca, M., Lakshminarayanan, B., Warde-Farley, D., and Mohamed, S. (2017). Variational approaches for auto-encoding generative adversarial networks. *arXiv* [Preprint]. arXiv:1706.04987,

Rosenholtz, R., Huang, J., and Ehinger, K. A. (2012). Rethinking the role of top-down attention in vision: effects attributable to a lossy representation in peripheral vision. *Front. Psychol.* 3:13. doi: 10.3389/fpsyg.2012.00013

Schwartz, A. B., Cui, X. T., Weber, D. J., and Moran, D. W. (2006). Brain-controlled interfaces: movement restoration with neural prosthetics. *Neuron* 52, 205–220. doi: 10.1016/j.neuron.2006.09.019

Shen, G., Dwivedi, K., Majima, K., Horikawa, T., and Kamitani, Y. (2019a). End-to-end deep image reconstruction from human brain activity. *Front. Comput. Neurosci.* 13:21. doi: 10.3389/fncom.2019.00021

Shen, G., Horikawa, T., Majima, K., and Kamitani, Y. (2019b). Deep image reconstruction from human brain activity. *PLoS Comput. Biol.* 15:e1006633. doi: 10.1371/journal.pcbi.1006633

Shenoy, P., and Tan, D. (2008). "Human-aided computing: utilizing implicit human processing to classify images," in *Proceedings of the CHI 2008 Conference on Human Factors in Computing Systems*, Florence. doi: 10.1145/1357054.1357188

Shi, Y., Siddharth, N., Paige, B., and Torr, P. H. (2019). "Variational mixture-of-experts autoencoders for multi-modal deep generative models," in *Proceedings of the Advances in Neural Information Processing Systems* (Vancouver, BC), 15692–15703.

Simonyan, K., and Zisserman, A. (2015). "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)* (San Diego, CA: ICLR Conference Track Proceedings).

Srivastava, N., and Salakhutdinov, R. (2012). "Multimodal learning with deep Boltzmann machines," in *Proceedings of the Advances in Neural Information Processing Systems* (Lake Tahoe, NV: Curran Associates Inc.), Vol. 2, 2222–2230.

Stewart, A. X., Nuthmann, A., and Sanguinetti, G. (2014). Single-trial classification of EEG in a visual object task using ICA and machine learning. *J. Neurosci. Methods* 228, 1–14. doi: 10.1016/j.jneumeth.2014.02.014

Suzuki, M., Nakayama, K., and Matsuo, Y. (2017). "Joint multimodal learning with deep generative models," in *Proceedings of the International Conference on Learning Representations (ICLR) 2017 Workshop* (Toulon, France: ICLR Conference Track Proceedings).

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381, 520–522. doi: 10.1038/381520a0

Tsai, Y. H. H., Liang, P. P., Zadeh, A., Morency, L. P., and Salakhutdinov, R. (2019). "Learning factorized multimodal representations," in *Proceedings of the International Conference on Learning Representations (ICLR)* (New Orleans, LA: ICLR Conference Track Proceedings).

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2016). Instance normalization: the missing ingredient for fast stylization. *arXiv* [Preprint]. arXiv:1607.08022,

Wang, C., Xiong, S., Hu, S., Yao, L., and Zhang, J. (2012). Combining features from ERP components in single-trial EEG for discriminating four-category visual objects. *J. Neural Eng.* 9:056013. doi: 10.1088/1741-2560/9/5/056013

Whitney, D., Haberman, J., and Sweeny, T. D. (2014). "49 from textures to crowds: multiple levels of summary statistical perception," in *The New Visual Neurosciences*, eds J. S. Werner and L. M. Chalupa (Cambridge, MA: MIT Press), 695–710.

Wu, M., and Goodman, N. (2018). "Multimodal generative models for scalable weakly-supervised learning," in *Proceedings of the Advances in Neural Information Processing Systems* (Montreal, Canada: Curran Associates Inc), 5575–5585.

Ziemba, C. M., Perez, R. K., Pai, J., Kelly, J. G., Hallum, L. E., Shooner, C., et al. (2019). Laminar differences in responses to naturalistic texture in macaque V1 and V2. *J. Neurosci.* 39, 9748–9756. doi: 10.1523/jneurosci.1743-19.2019

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.