# Generative Adversarial Networks Conditioned by Brain Signals

S. Palazzo, C. Spampinato, I.Kavasidis, D. Giordano
PeRCeiVe Lab - Department Electrical Electronics and Computer Engineering
University of Catania - Italy
{palazzosim, kavasidis, dgiordan, cspampin}@dieei.unict.it
M. Shah
Center for Research in Computer Vision
University of Central Florida - USA
shah@crcv.ucf.edu

## Abstract

*Recent advancements in generative adversarial networks (GANs), using deep convolutional models, have supported the development of image generation techniques able to reach satisfactory levels of realism. Further improvements have been proposed to condition GANs to generate images matching a specific object category or a short text description. In this work, we build on the latter class of approaches and investigate the possibility of driving and conditioning the image generation process by means of brain signals recorded, through an electroencephalograph (EEG), while users look at images from a set of 40 ImageNet object categories with the objective of generating the seen images. To accomplish this task, we first demonstrate that brain activity EEG signals encode visually-related information that allows us to accurately discriminate between visual object categories and, accordingly, we extract a more compact class-dependent representation of EEG data using recurrent neural networks. Afterwards, we use the learned EEG manifold to condition image generation employing GANs, which, during inference, will read EEG signals and convert them into images. We tested our generative approach using EEG signals recorded from six subjects while looking at images of the aforementioned 40 visual classes. The results show that for classes represented by well-defined visual patterns (e.g., pandas, airplane, etc.), the generated images are realistic and highly resemble those evoking the EEG signals used for conditioning GANs, resulting in an actual reading-the-mind process.*

## 1. Introduction

Reading the mind is such an ambitious and dreamed-upon capability that is widely — and reasonably — regarded as closer to science fiction than real science. However, little steps are constantly being made by the scientific community to push the limits of our understanding of the brain's workings and of our probing technology. For example, research on brain-computer interfaces for direct-actuated control of machines for disabled people is a very active and relatively successful field, which can make an actual impact on users' lives [14, 26, 8].

But what if, instead of being able to detect a limited set of simple executive commands from the brain, we could generate something more inspiring, more meaningful, more complex — like images?

While cognitive neuroscience studies [11, 19, 21] have attempted — with yet uncertain results — to identify which parts of the human visual cortex and brain are responsible for visual cognitive processes, it has been acknowledged that brain activity recordings contain information about visual object categories [4, 31, 27, 2, 1]. This consideration makes one wonder whether patterns of such brain activity may be identified in order to extract useful information on the content of an observed scene.

This kind of information could, then, be used in conjunction with conditional generative models to reconstruct a meaningful and realistic image from the informative content decoded from brain activity. Luckily, such generative frameworks already exist, and one of them in particular — generative adversarial networks (GANs) [6] — is currently very popular thanks to its simplicity in concept and effectiveness in practice (although some aspects related to reliable training approach are still unclear). Hence, assuming that a GAN approach intrinsically contains the complexity required to model the image generation process, the main problem to solve is how to extract visually-content–representative information.

Recently, EEG has been increasingly used to capture

brain activity signals and process them for visually-related uses, e.g., visual object classification [28, 10]. Although promising results have been shown, the techniques employed to process this kind of multi-dimensional, noisy, temporal data are still very simple, and mostly ignore local temporal dynamics, processing the full EEG signal as a whole.

In this work, we combine a GAN approach with a model based on recurrent neural networks (RNNs) to process EEG signals captured while users look at images on a screen. The recurrent model temporally analyzes input signals and learns to encode them into a compact and visually-content–descriptive representation; in turn, this representation is used to condition the image generation process by a GAN model, with the objective of producing output images depicting objects semantically- related to those shown to users while the original EEG signals have been recorded. The objective is to learn a representation of brain signals which conveys enough meaning for a generative model to capture the visual category associated to it, and to be able to reproduce a relevant sample.

## 2. Related Work

In a typical experimental scenario attempting to study the brain responses and dynamics associated to visual processes, a human subject looks at a series of images, while a recording device interfaced to or scanning the brain is employed to register the appropriate feedback signals for further analysis. Currently, there exists a variety of non-invasive methods that allow us to acquire such brain responses (fMRI, EEG, MEG) with different grades of sensitivity, but still, there is a profound lack in understanding what exactly the acquired data means and, even more importantly, how to interpret it.

In a pioneering work [18], the authors try to generate impressions of what the subjects see based on fMRI images by imposing a prior built on a large image dataset extracted from YouTube. Essentially, this work tries to maximize the *a posteriori* probability that a certain visual stimulus evoking a specific cerebral response corresponds to an image drawn from a large pool of images [16] by exploiting the high sensitivity that fMRI signals offer. However, such advantage is countered by the objective difficulty of setting up and operating an fMRI scanner and by the considerably higher utilization costs.

To alleviate these drawbacks much research effort is concentrated on electrophysiological responses, rather than brain imaging and, especially, EEG, which features a lower spatial resolution with respect to almost all other methods, but has a very high temporal one. An EEG data acquisition session costs also less and is simpler to execute, but the quality of the gathered data often suffers from unwanted environmental noise and artifacts, making the challenge of reconstructing the initial stimuli much harder. It is known that EEG signals encode basic responses to visual stimuli [3, 15], and recently the authors of this paper, in [29], were able to "decode" such information and use it for automated visual classification. This paper builds on this recent discovery and aims at reconstructing the initial stimulus from learned latent space. However, reconstructing the visual stimuli is not trivial.

Indeed, the human visual cortex covers around 30% of the total cortical area [7], which makes it far larger than the other sensory cortices, meaning that visual information representation in the brain is clearly the most complex among all sensory processes. For example, some previous works have managed to recreate the stimuli from other senses than sight. In [20], the authors describe an approach to recreate (partially) speech stimuli based on human auditory cortex data, acquired by cortical surface electrode arrays. Brain signals acquired by this method have the advantage of being affected by noise in a lesser extent than EEG signals, and also require a simpler generative model. However, replicating such procedure is prohibitive because it requires open-skull surgery to be performed on the subject.

Reconstructing human vision, however, is different as it requires to understand if and how brain signals recorded through existing devices convey visual content. There exist a few works that attempt to address that, e.g., methods for identifying visual classes of the visual stimuli. In [10], a classifier is trained to recognize object classes based on topographic maps generated by EEG signals. However, the obtained accuracy is low (29% over 12 classes), mainly because the employed linear classifier cannot represent adequately the spatio-temporal dynamics contained in the EEG signals. A similar work is presented in [30], but this time raw EEG data is first processed by Independent Component Analysis and then fed to a Support Vector Machine classifier, which has the task to distinguish between only 2 classes. While these works are undoubtedly interesting, they present a number of limitations (relatively simple classification models, low number of object classes) that do not permit the investigation at a deeper level of the temporal and spatial dynamics of the EEG signals.

On the other hand, deep learning methods are able to handle large, diverse and noisy datasets with exceptional results. Moreover, recently, there was an explosion in the number of works that employ deep learning methods for image generation, and more specifically, generative adversarial networks [6]. In general, a GAN is a deep convolutional neural network comprised of two parts: the generator, which has the task of creating images starting from pure noise, and the discriminator, which assesses whether an input image is real or fake. While, initially, GANs could generate images based on a single type of images (i.e., a simple GAN can only be trained and used only for

a single object class), conditional GANs [13] introduced the ability to generate images based on specific attributes. Such attributes can be in the form of one-hot binary vectors (i.e., a single bit in the vector indicates the class to generate), words [32, 23] or arbitrary real number vectors representing geometric transformations and coordinates in a 3D space [5]. However, the majority of works describing generative models employ clearly defined images as well as conditioning vectors (e.g. hand-written digits, faces) and adequately large datasets (e.g., MNIST[1], CIFAR-10[2], CelebA[3]) for the training process. The performance, instead, deteriorates substantially when small and noisy datasets are used for training as the case we are tackling in this work. Indeed, EEG signals are particularly noisy and it is not trivial to collect large data.

## 3. Method

Our method consists of a "EEG-in/image-out" processing pipeline, where EEG signals are recorded while showing images to human subjects and output images are obtained by a generator which learns to associate the processed EEG signals to the visual object class observed while those signals were recorded.

The feasability of this approach relies on a few assumptions. First of all, it is necessary that EEG signals intrinsically encode visually-related information, whether these are low-level responses to visual stimuli or high-level cognitive processes associated to more complex activities such as recognition and understanding.

Secondarily, it has to be possible to extract a meaningful representation, suitable for solving visual classification problems, from the high-dimensional and highly-noisy raw EEG signals. For example, a 0.5-second-long 128-channel EEG track at 1 kHz consists of 64,000 data samples, with unclear underlying dynamics, correlations and noise components. Extracting a low-dimensional descriptor encoding visually-relevant information is therefore a critical task for the whole process.

Finally, in order for our image generator to produce images of the correct class given a low-dimensional representation of the EEG signal, the visual information encoded by such representation has to be class-discriminative. Even if EEG signals encode visual information, this does not imply that the level of "detail" of such information allows to distinguish between object categories.

Our design approach takes these hypotheses for granted in the way it processes the input data; the feasibility of the whole process is then verified by the results we obtain. Fig. 1 shows the architecture employed in this work, divided into its basic data-acquisition and processing modules:

- **EEG recording protocol**: each subject in the experiment undergoes an EEG recording session, where he/she simply has to look at images from different classes on a computer monitor.
- **EEG manifold learning**: raw EEG signals are processed by an RNN-based encoder, which is trained to output a vector of what we call *EEG features*, containing visually-relevant and class-discriminative information extracted from the input signals.
- **EEG-conditioned image generation**: a generator network is trained in a conditional GAN framework to produce images from EEG features, so that the visual class of the output image matches that of the conditioning vector.

### 3.1. EEG data acquisition

Six subjects partecipated in the experiment and were shown images of objects while EEG data was recorded. All subjects were evaluated by a professional physician in order to exclude possible health conditions or medication that could alter normal cerebral activity.

The subjects were shown 50 images from 40 different object classes[4] for a total of 2,000 images per subject. Each image class was presented in bursts of 25 seconds (0.5 second per image) followed by a 10 seconds pause where a black image was shown. The black image was used to "flush" any high-level class information present from the previous one. The total running time of each experiment was 1,400 seconds (23 minutes and 20 seconds). Details of the experimental protocol are shown in Table 1.

We used the actiCAP[5] cap with 128 active low-impedance, low-noise electrodes. Four 32-channel Brainvision[6] high-precision, low-latency signal amplifiers were used (exact model: BrainAmp DC) and a qualified technician was present during the experiments' execution, ensuring that skin impedance remained under 10 kOhm at all times by using conductive abrasive gel. The acquired EEG signals were filtered in run-time (i.e. during the acquisition phase) by the integrated hardware notch filter (49-51 Hz) and a second order Butterworth (band-pass) filter with frequency boundaries 14-70 Hz. This frequency range contains the necessary bands (Alpha, Beta and Gamma) that are most meaningful during the visual recognition task [17]. The sampling frequency was set to 1000 Hz and the quantization resolution to 16 bit.

[1] http://yann.lecun.com/exdb/mnist/
[2] https://www.cs.toronto.edu/ kriz/cifar.html
[3] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[4] A subset of the ImageNet dataset [24] was used consisting of the 40 classes that are shown in Tab. 3.
[5] http://www.brainproducts.com/
[6] http://www.brainvision.com/

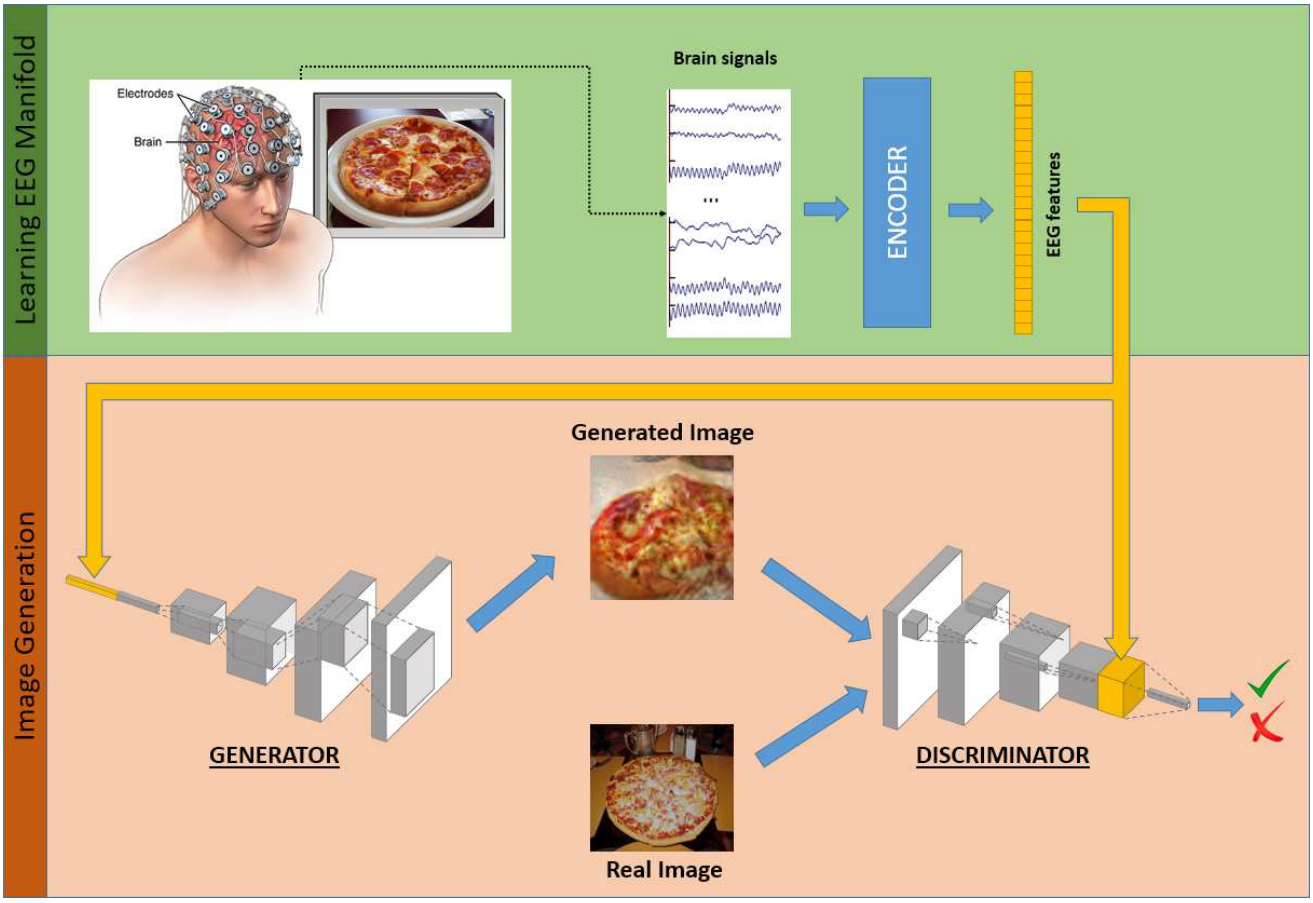


Figure 1. Overview of the architecture design of the proposed EEG-driven image generation approach.

The histogram of the acquired signals over the different values presented with a high density near the zero value and a much lower density at the extremities. In order to reduce input space sparsity, non-uniform quantization was applied for data compression.

Deep-learning networks need constant length input sequences both for training and validation, however, data coming from analog devices may present variable data size due to different factors. Indeed, by acquiring data at a 1000 Hz sampling rate for 500 ms, 500 samples of data should be acquired per image. Given that the systems involved are not real-time (Operating system process scheduler, DAQ hardware etc...), variable length EEG sequences were dealt with by discarding those with less than 480 samples. Data sequences whose length was between 480 and 500 samples where padded with zeros until reaching 500 samples. Sequences longer than 500 samples were tail trimmed.

From each recorded EEG sequence, the first 40 samples were discarded in order to minimize any possible interference from the previously shown image (i.e., to give the necessary time for the stimulus to clear its way through the optical tract [9]). The following 440 samples (440 ms) were used for the experiments.

| | |
|---|---|
| Number of classes | 40 |
| Number of images per class | 50 |
| Total number of images | 2,000 |
| Visualization order | Sequential |
| Time for each image | 0.5 s |
| Pause time between classes | 10 s |
| Number of sessions | 4 |
| Session running time | 350 s |
| Total running time | 1,400 s |

Table 1. The parameters of the experimental protocol.

By using the protocol in Table 1, we acquired $12,000$ ($2,000$ images for 6 subjects) 128-channel EEG sequences. $534$ samples did not satisfy the minimum data length criteria described above, resulting in $11,466$ valid samples.

### 3.2. Learning EEG visual descriptors

Although previous works have attempted to work directly with the multi-channel temporal EEG sequences

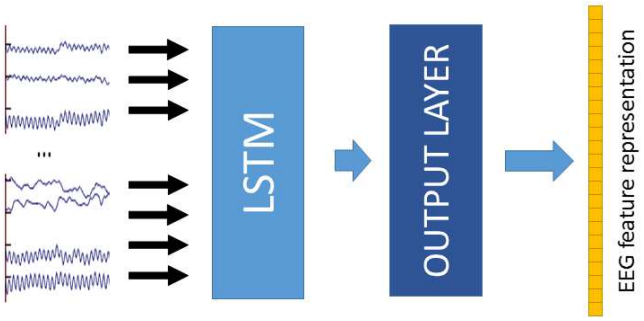


Figure 2. EEG feature encoder architecture.

[10, 30], by simply concatenating time sequences into a single feature vector (albeit with a smaller number of channels than 128), this kind of methods ignores local temporal dynamics. To account for time dependencies, we employ LSTM recurrent neural networks inspired by previous results in [29].

Our EEG feature encoder is illustrated in Fig. 2, and consists of a standard LSTM layers followed by a nonlinear layer. At each time step, input $s(\cdot, t)$ (i.e., the set of values from all channels at time $t$) is fed into the LSTM layer; when all time steps have been processed, the final output state of the LSTM goes into a fully-connected layer with ReLU non-linearity. The resulting output is what we refer to as "EEG features", and should ideally be a compact representation of visual class–discriminative brain activity information. We append a softmax classification layer and perform gradient descent optimization (supervised by the class of the image shown when the input signal had been recorded) to train the encoder and the classifier end-to-end.

### 3.3. Brain Signal–Conditioned GANs for Image Generation

We train our generator network in a conditional GAN framework [13]. In the original formulation, a generative model $G(z|y)$ maps random input, from a $p_z(z)$ noise distribution and from a condition $y$ to the target data distribution $p_{\text{data}}(x)$. A discriminative model $D(x|y)$ then predicts the probability that a data point belongs to the target distribution, given the condition. The generator and the discriminator are trained simultaneously, so that the discriminator tries to maximize the probability of assigning the correct label to "real" data (from $p_{\text{data}}(x)$) and "fake" data (from $p_G(z|y)$), while the generator tries to maximize the probability that the discriminator mistakes generated samples for "real" ones. In other words, the two models play the following minimax game defined by value function $V(D, G)$:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \in p_{\text{data}}}(x) \left[\log D\left(x|y\right)\right] + \mathbb{E}_{z \in p_z(z)} \left[\log\left(1 - D\left(G\left(z|y\right)|y\right)\right)\right]$$

In practice, from a training point of view, this means that, given a correct sample $s_c = (x_c, y_c)$, consisting of real data with correct condition and a fake sample $s_w = (x_w, y_w)$, consisting of fake data (with arbitrary condition), the negative log-likelihood discriminator loss is computed as:

$$\mathcal{L}_D = -\log D\left(x_c|y_c\right) - \log\left(1 - D\left(x_w|y_w\right)\right), \quad (1)$$

, while the generator loss, for an analogous $s_w$ sample, is:

$$\mathcal{L}_G = -\log D\left(x_w|y_w\right). \quad (2)$$

In our case, the condition vector associated to each image is the average EEG feature vector (as computed by the encoder described in the previous section) over all images of each class and all subjects. Our expectation is that, if the encoder has been correctly trained to produce distinguishable features for different classes, the generator and the discriminator will be able to capture this separability and behave accordingly.

Both $D$ and $G$ are convolutional networks are illustrated in the bottom part of Fig. 1 and their architecture is inspired by DCGANs [22].

In the generator, the condition $y$ is appended to the random noise vector $z$, and a cascade of transposed convolutions upsample the concatenated input to an output color image. The discriminator takes as input a same-size image (either a real one or a generated one). After going through a few convolution layers which reduce the size of the feature maps, the condition $y$ associated to the input image, is spatially replicated and appended to the set of feature maps from the second-to-last convolutional layer, on which the final probability estimation is made.

During learning, we modify the discriminator loss function previously presented in Eq. 1 by following the approach described in [23]: instead of training the discriminator with real images employing correct conditions and fake images with arbitrary conditions (which forces the discriminator to learn how to distinguish between real images with correct conditions and real images with wrong conditions without any explicit supervision), we also provide a wrong sample consisting of a real image and a wrong condition, randomly chosen as the representative EEG feature vector from a different class. Hence, given a correct sample $s_c = (x_c, y_c)$ and wrong samples $s_{w_1} = (x_c, y_w)$ and $s_{w_2} = (x_w, y_w)$, the discriminator loss becomes:

$$\begin{aligned}\mathcal{L}_D = &-\log D\left(x_c|y_c\right) \\ &-\log\left(1 - D\left(x_c|y_w\right)\right) \\ &-\log\left(1 - D\left(x_w|y_w\right)\right).\end{aligned} \quad (3)$$

| Model | Max VA | TA at max VA |
|---|---|---|
| LSTMs + nonlinear | 86.1% | 83.9% |

Table 2. Maximum validation accuracy ("Max VA") and corresponding test accuracy ("TA at max VA") for the LSTM-based EEG feature encoder shown in Sect. 3.2.

## 4. Performance Analysis

Performance analysis consists of two main parts: first, we evaluate how our EEG feature encoding architecture has learnt to extract meaningful representations from visual-stimuli–evoked raw EEG signals; secondly, we analyze the performance of our EEG-driven image generator. The latter is not a trivial task: apart from a purely qualitative judgment (i.e., "do the images look good?"), no quantitative evaluation practice exists for evaluating GAN models. In this work, we employ the Inception score [25], which estimates the realism and diversity of a batch of generated images by analyzing the softmax distribution of the Inception network, and a criterion based on its classification accuracy on the considered batch of generated images: this last test is meant to estimate whether the generated images are of good enough quality for Inception to still classify them correctly (which is not taken into account in the Inception score).

### 4.1. Learning EEG features: classification accuracy

The evaluation of our LSTM-based approach has been reported in [29], where we split our EEG signal dataset into training, validation and test sets, with respective fractions 80% (1,600 images), 10% (200), 10% (200). Splitting by images, rather than by EEG signals (which, for each image, are as many as the number of participant subjects), makes sure that the signals generated by all subjects for a single image are not spread over different splits.

Training was performed by using the Adam gradient descent method (learning rate initialized to 0.001), with mini-batches of size 16. All layer sizes in the model (the stacked LSTMs and the following non-linear layer) were set to 128. Model and training hyperparameters were tuned on the validation set.

Table 2 reports the best classification accuracy achieved by our EEG signal classifier.

### 4.2. GAN model and training details

The generator takes as input a concatenated vector of 100-dimensional random noise and 128-dimensional EEG features. Such input then goes through 5 transposed convolutional layers: the first layer spatially upsamples the vector by four times, while each of the other layers double the size at every step, so that the output image size is 64×64. The number of features maps starts at 512 at the first layer, and is halved for each layer before the last one, which outputs a 3-channel (color) image.

The discriminator is made up of four convolutional layers and two fully-connected layers. It takes as an input 64×64 images, and analogously halves the feature map size at every convolutional step. After the final convolutional layer, where the feature map size is 4×4 (to which the condition vector is spatially appended), two fully-connected layers reduce the number of features to 1024 and 1, the latter being the sigmoidal probability estimate on the input image/condition pair. The number of feature maps in the convolutional layers starts at 64 at the first layer, and is doubled at every layer before the fully-connected ones. Both the generator and the discriminator include batch normalization modules and ReLU nonlinearities.

Training GANs is notoriously difficult, due to the difficult design choices required to make the generator and the discriminator balanced. In our case, the low number of images for which we had recorded EEG tracks made it impossibile to train directly on those images, as either the generator or the discriminator would overfit.

However, the images used for the EEG data acquisition protocol were subsets of 50 elements taken from 40 ImageNet classes, with each one containing about 1,200 images. In order to make use of all images in the selected classes, we trained our GAN network in two stages, making use of both EEG-available images and EEG-unavailable ones. In the first stage, we trained the generator and the discriminator as a regular (non-conditional) GAN using only images for which no EEG data was available. All condition vectors $y$ were set to the zero vector, and the loss term related to real images and wrong conditions (i.e. the second term of Eq. 3) was ignored. After 100 epochs, we re-trained the models for 50 more epochs on the images with EEG data available, providing the correct condition vectors and applying the full discriminator loss function.

During training, data augmentation was performed by resizing images at 96×96 pixels, and extracting random 64×64 (horizontally flipped with 50% chance).

### 4.3. Image generation: qualitative and quantitative analysis

Fig. 3 and 4 show samples for some of the 40 classes in our dataset. While the generator is generally able to capture the basic distinguishing patterns, which confirms that the generator and the discriminator were able to make use of the conditioning EEG features in order to distinguish between different input/output classes, it can be noticed that for some classes (Fig. 3) the level of realism is markedly higher than others (Fig. 4). This can be explained by analyzing the complexity of the dataset. Unlike typical benchmarking datasets such as the CelebA face dataset or LSUN Bedroom, the selected 40 ImageNet classes exhibit high intra-class variance in object appearance and low size (about 1,200): to make a

comparison, CIFAR-10 has a lower number of classes (10), a larger number of images per class (6,000) and a relatively low intra-class variance for many classes (e.g., "airplane", "horse", "car").

We computed the Inception score both globally (across all classes) and on a per-class base. In the first case, we generated a sample of 50,000 images (1,250 per class); in the second case, we generated a sample of 50,000 images for each class, and computed per-class Inception scores. The results are shown in Table 3. To the best of our knowledge, Inception score results have not been published on ImageNet (or subsets thereof, as in our case); on CIFAR-10, the current best published result is 8.07 [25]. The results we obtain on our dataset approximate the capability of the network to better understand the structure of certain classes with respect to others. While the achieved Inception scores are not at the same level as those computed on CIFAR-10, it should be noted that several factor impact these results:

1. Higher resolution: 64×64 in our case, 32×32 in CIFAR-10;

2. More classes: 40 in our case, 10 in CIFAR-10;

3. Fewer images per class: 1,200-1,300 in our case, 6,000 in CIFAR-10;

4. Higher intra-class variability;

5. Noisy conditioning vectors. Indeed, conditional GANs are often trained with image class labels, while in our case GANs are conditioned using a learned EEG manifold that allows for separation among image classes but in some case may fail (EEG classification results are about 84%Table 2).

Since the Inception score does not measure the correctness of the generated images in terms of correspondence with the condition vectors, we performed an evaluation aimed at verifying that the generated images for a given condition (expressed as the average EEG features for each class) were actually similar to images of the correct class.

To do so, we re-used the previously generated sample of 50,000 images (1,250 images per class) to compute the class probability distribution through the Inception network, whose classification layer was pruned by keeping only the 40 classes in our dataset. The correct classification rate was 0.43, which, albeit relatively low (though it should be noted that random guess on 40 classes in 2.5%) shows that the generated images are realistic enough to make automatic classification meaningful. Table 3 shows per-class correct classification rate. As expected, similar to the qualitative visual analysis, the lowest classification accuracy are related to the classes whose internal visual appearance variance is higher, which — concurrently with the small number of images — made it difficult to the generator to learn to reproduce the correct patterns.

| **Class** | **IS** | **IC** |
|---|---|---|
| German shepherd (n02106662) | 4.91 | 0.23 |
| Egyptian cat (n02124075) | 4.45 | 0.29 |
| Lycaenid butterfly (n02281787) | 5.03 | 0.37 |
| Sorrel (n02389026) | 5.86 | 0.62 |
| Capuchin (n02492035) | 4.99 | 0.41 |
| Elephant (n02504458) | 5.35 | 0.57 |
| Panda (n02510455) | 6.35 | 0.72 |
| Anemone fish (n02607072) | 6.11 | 0.81 |
| Airliner (n02690373) | 6.20 | 0.86 |
| Broom (n02906734) | 4.76 | 0.35 |
| Canoe (n02951358) | 4.59 | 0.24 |
| Cellphone (n02992529) | 5.17 | 0.31 |
| Mug (n03063599) | 4.62 | 0.23 |
| Convertible (n03100240) | 4.54 | 0.34 |
| Desktop PC (n03180011) | 5.81 | 0.61 |
| Digital watch (n03197337) | 4.54 | 0.51 |
| Electric guitar (n03272010) | 4.91 | 0.32 |
| Electric locomotive (n03272562) | 4.88 | 0.24 |
| Espresso maker (n03297495) | 5.33 | 0.32 |
| Folding chair (n03376595) | 4.88 | 0.27 |
| Golf ball (n03445777) | 5.06 | 0.28 |
| Piano (n03452741) | 4.47 | 0.22 |
| Iron (n03584829) | 4.32 | 0.23 |
| Jack-o'-lantern (n03590841) | 6.64 | 0.91 |
| Mailbag (n03709823) | 5.51 | 0.49 |
| Missile (n03773504) | 5.87 | 0.54 |
| Mitten (n03775071) | 5.10 | 0.36 |
| Mountain bike (n03792782) | 4.86 | 0.33 |
| Mountain tent (n03792972) | 4.70 | 0.30 |
| Pyjama (n03877472) | 4.21 | 0.20 |
| Parachute (n03888257) | 4.59 | 0.38 |
| Pool table (n03982430) | 4.68 | 0.35 |
| Radio telescope (n04044716) | 5.08 | 0.37 |
| Reflex camera (n04069434) | 4.64 | 0.29 |
| Revolver (n04086273) | 4.55 | 0.26 |
| Running shoe (n04120489) | 4.31 | 0.22 |
| Banana (n07753592) | 6.28 | 0.83 |
| Pizza (n07873807) | 5.87 | 0.79 |
| Daisy (n11939491) | 5.81 | 0.74 |
| Bolete (n13054560) | 5.37 | 0.60 |
| *All* | *5.07* | *0.43* |

Table 3. Inception scores (IS) and Inception classification accuracies (IC) for each class of the dataset (specified by their ImageNet synset identifier and by a short description), and overall.

(a) Airliner

(b) Jack-o'-Lantern

(c) Panda

Figure 3. Good results

(a) Banana

(b) Capuchin

(c) Bolete

Figure 4. Bad results

## 5. Conclusions

Although reading the mind may still be something which humanity will not be able to achieve for a while, in this work we showed that brain activity signals can be successfully analyzed to drive the generation of images depicting similar objects as those being observed by a subject when those signals were recorded. Our approach, combining an LSTM recurrent neural network for extracting visual-class–discriminative descriptors from raw EEG signals, and a conditional GAN for generating images from those very descriptors, is able to produce realistic and diverse images which match the expected object classes, thus demonstrating the goodness of the method and the validity of the initial assumptions.

Of course, improvements can be made: the method suffers in presence of classes with high internal variability in appearance, which, combined with the relatively small size of the employed dataset (if compared to other typical benchmarks for GANs), causes the image generator not being able to create targeted and clearly recognizable images.

In the future, we aim at pushing the limits of this approach by attemping not just at generating an image depicting the same visual category as the one from which an EEG signal was generated, but at reconstructing the original image. Of course, this is a much more complicated task. Indeed, the sensitivity and resolution of the EEG technology may not be sufficient, and we will need to resort to higher-resolution modalities such as fMRI [18]. In turn, this will require an adaptation of the models employed: for example, given the volumetric nature of fMRI data, our brain encoding module may become a 3D recurrent-convolutional hybrid. Additionally, to compensate for the low temporal resolution of the fMRI scanners, we are going to investigate methods to combine fMRI data with EEG data [12].

### Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of two Titan X Pascal GPUs used for this research. We also acknowledge Dr. Martina Platania for carrying out EEG data acquisition.

## References


[1] T. Carlson, D. A. Tovar, A. Alink, and N. Kriegeskorte. Representational dynamics of object vision: the first 1000 ms. *Journal of Vision*, 13(10), 2013. 1

[2] T. A. Carlson, H. Hogendoorn, R. Kanai, J. Mesik, and J. Turret. High temporal resolution decoding of object position and category. *Journal of Vision*, 11(10), 2011. 1

[3] H. Cecotti and A. Graser. Convolutional neural networks for p300 detection with application to brain-computer interfaces. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):433–445, 2011. 2

[4] K. Das, B. Giesbrecht, and M. P. Eckstein. Predicting variations of perceptual performance across individuals from neural activity using pattern classifiers. *Neuroimage*, 51(4):1425–1437, Jul 2010. 1

[5] A. Dosovitskiy, J. Tobias Springenberg, and T. Brox. Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1538–1546, 2015. 3

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 2

[7] D. Grady. The vision thing: Mainly in the brain. *Discover*, 14(6):56–66, 1993. 2

[8] A. M. Green and J. F. Kalaska. Learning to move machines with the mind. *Trends in neurosciences*, 34(2):61–75, 2011. 1

[9] J. R. Heckenlively and G. B. Arden. *Principles and practice of clinical electrophysiology of vision*. MIT press, 2006. 4

[10] B. Kaneshiro, M. Perreau Guimaraes, H.-S. Kim, A. M. Norcia, and P. Suppes. A Representational Similarity Analysis of the Dynamics of Object Processing Using Single-Trial EEG Classification. *Plos One*, 10(8):e0135697, 2015. 2, 5

[11] Z. Kourtzi and N. Kanwisher. Cortical regions involved in perceiving object shape. *J. Neurosci.*, 20(9):3310–3318, May 2000. 1

[12] Z. Liu and B. He. fmri–eeg integrated cortical source imaging by use of time-variant spatial constraints. *Neuroimage*, 39(3):1198–1214, 2008. 8

[13] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3, 5

[14] G. R. Muller-Putz and G. Pfurtscheller. Control of an electrical prosthesis with an ssvep-based bci. *IEEE Transactions on Biomedical Engineering*, 55(1):361–364, 2008. 1

[15] M. Nakanishi, Y. Wang, Y.-T. Wang, Y. Mitsukura, and T.-P. Jung. A high-speed brain speller using steady-state visual evoked potentials. *International journal of neural systems*, 24(06):1450019, 2014. 2

[16] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009. 2

[17] E. Niedermeyer and F. L. da Silva. *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005. 3

[18] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011. 2, 8

[19] H. P. Op de Beeck, K. Torfs, and J. Wagemans. Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *J. Neurosci.*, 28(40):10111–10123, Oct 2008. 1

[20] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang. Reconstructing speech from human auditory cortex. *PLoS Biol*, 10(1):e1001251, 2012. 2

[21] M. V. Peelen and P. E. Downing. The neural basis of visual body perception. *Nat. Rev. Neurosci.*, 8(8):636–648, Aug 2007. 1

[22] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. 5

[23] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 3, 2016. 3, 5

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 3

[25] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2226–2234, 2016. 6, 7

[26] A. B. Schwartz, X. T. Cui, D. J. Weber, and D. W. Moran. Brain-controlled interfaces: movement restoration with neural prosthetics. *Neuron*, 52(1):205–220, 2006. 1

[27] P. Shenoy and D. Tan. Human-aided computing: Utilizing implicit human processing to classify images. In *CHI 2008 Conference on Human Factors in Computing Systems*, 2008. 1

[28] I. Simanova, M. van Gerven, R. Oostenveld, and P. Hagoort. Identifying object categories from event-related EEG: Toward decoding of conceptual representations. *PLoS ONE*, 5(12), 2010. 2

[29] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah. Deep Learning Human Mind for Automated Visual Classification . *CVPR 2017*, 2017. 2, 5, 6

[30] A. X. Stewart, A. Nuthmann, and G. Sanguinetti. Single-trial classification of EEG in a visual object task using ICA and machine learning. *Journal of Neuroscience Methods*, 228:1–14, 2014. 2, 5

[31] C. Wang, S. Xiong, X. Hu, L. Yao, and J. Zhang. Combining features from ERP components in single-trial EEG for discriminating four-category visual objects. *J Neural Eng*, 9(5):056013, Oct 2012. 1

[32] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016. 3