Brain2Pix: Fully convolutional naturalistic video reconstruction from brain activity

Lynn Le, Luca Ambrogioni, Katja Seeliger,
Yağmur Güçlütürk, Marcel van Gerven, Umut Güçlü
Radboud University, Donders Institute for Brain, Cognition and Behaviour
Nijmegen, Netherlands
l.le@donders.ru.nl

Abstract

Reconstructing complex and dynamic visual perception from brain activity remains a major challenge in machine learning applications to neuroscience. Here we present a new method for reconstructing naturalistic images and videos from very large single-participant functional magnetic resonance data that leverages the recent success of image-to-image transformation networks. This is achieved by exploiting spatial information obtained from retinotopic mappings across the visual system. More specifically, we first determine what position each voxel in a particular region of interest would represent on the visual field based on its corresponding receptive field location. Then, the 2D image representation of the brain activity on the visual field is passed to a fully conventional image-to-image network trained to recover the original stimuli using VGG feature loss with an adversarial regularizer. In our experiments we show that our method offers a significant improvement over existing techniques.

1 Introduction

A great interest of systems neuroscience is understanding the information that lies in neural activity. Decoding visual stimuli from neural activity using deep learning is a promising approach, providing more research questions and bringing us closer to understanding neural patterns. Recent advancements allow the successful decoding of simple static images from brain data [44, 43, 28, 30, 16, 23, 9, 39]. Reconstructing novel natural movies is significantly more challenging [31]. The difficulty with reconstructing natural movies is in large part due to the limited temporal information provided by fMRI measurements as well as the complex dynamics of the natural world that the model must learn.

Convolutional image-to-image models have recently achieved unprecedented results in multiple tasks such as semantic segmentation [27, 33, 32, 26, 49], style transfer [52, 10, 40, 21], colorization [47, 19, 48] and super-resolution [22, 5, 50]. Convolutional image-to-image networks have the great advantage of preserving the topography of input images throughout all the layers of the network. Consequently, the network does not need to learn a remapping between locations and can focus on processing local features. The reconstruction of perceived natural images from brain responses can be considered as a form of image-to-image problem since visual cortices process information in a topographically organized manner [13, 14, 20] such that the topology of the input images is preserved within each visual area. The retinotopic mapping of visually responsive neurons reveals relationships between the visual field and its cortical representation in individual subjects and has uncovered many important aspects of the visual cortex across different species [17, 6]. However, it is not straightforward to exploit this in a image-to-image ConvNet architecture. The cortex itself can be roughly seen as a pair of topological spheres embedded in a 3D space. Several separate visual representations are embedded in this cortical space, corresponding to several visual areas (e.g.

Preprint. Under review.

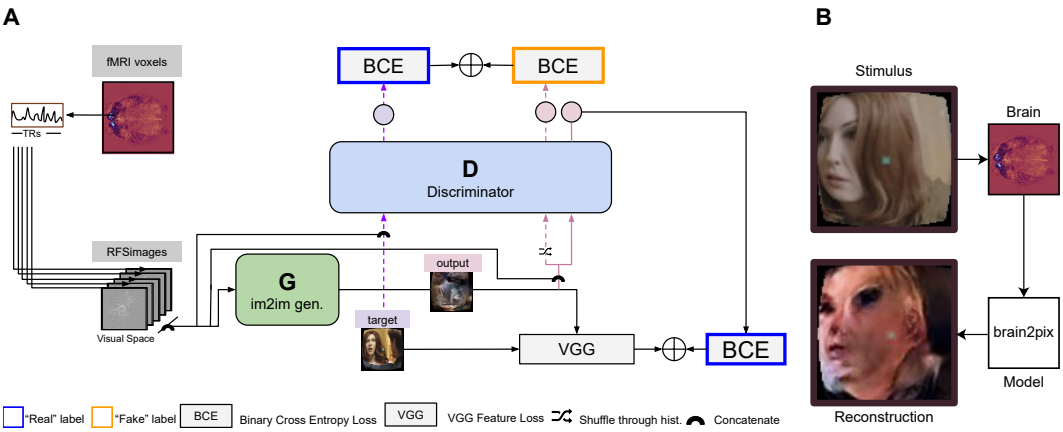

Figure 1: A) Visualization of the brain2pix architecture. First, each individual fMRI brain voxel is extracted and sorted by its corresponding region of interest (ROI). Each voxel is mapped onto the visual space based on the retinotopic mapping of that ROI, to become RFSimages. The input for each sample consists of the response at 5 timepoints (0.7s), thus the input channels are obtained by concatenating the time dimension. The **generator** receives as input the RFSimages, and then outputs the reconstruction. Then the loss between this reconstruction and the target is calculated with a vgg loss. The reconstruction also goes through the discriminator, concatenated with the RFSimages. The BCE loss of the discriminator's output is summed with the feature loss, which is then backpropagated to update the parameters of the generator. Training of the **discriminator** is done by comparing the output of the discriminator based on the reconstructed image and the target image (concatenated with RFSimages), using a BCE loss. B) Example of test set results. Reconstruction of a frame from the brain signal of a participant watching an episode of Dr. Who in the fMRI scanner.

V1,V2,V3). These representations are furthermore distorted by the geometry of the cortex and by the uneven sampling of different parts of the visual field. Therefore, there is not a natural way of constructing a convolutional architecture that exploits the image-to-image nature of the problem by preserving the topography between voxel responses and pixel brightness and color.

In this paper, we exploit the receptive field mapping of visual areas to convert voxel responses defined in the brain to activations in pixel-space. Functional visual areas are identified using functional localizers. The voxel activations of each area are then converted to images by applying a receptive field mapping. Importantly, these images (visual representations) do have a pixel-to-pixel correspondence with the images used as stimuli. We then transform these visual representations into realistic images using an image-to-image U-network trained using a combination of pixelwise, feature and adversarial losses.

2 Related work

Recent work on image reconstruction from fMRI data has demonstrated the success of employing deep neural networks (DNNs) and generative adversarial networks (GANs) in neural decoding [15, 9, 46, 39, 11, 41, 42]. For instance, [39] used a GAN to reconstruct grayscale natural images as well as simpler handwritten characters with reasonable accuracy. More recently, [41] showed that even with a limited set of data – in the order of thousands compared to millions that the field is accustomed to – it was possible to train an end-to-end model for natural image stimulus reconstruction by training a GAN with an additional high-level feature loss. Their reconstructions matched several high-level and low-level features of the presented stimuli. However, a comparable performance has not yet been achieved for naturalistic video stimuli. The most recent notable video reconstruction study by [11] made use of a variational auto-encoder and was only able to reconstruct very low-level properties of the images, where the reconstructions resembled shadows or silhouettes of the stimulus images at best.

The simplest way to apply ConvNets on fMRI voxel responses is to treat fMRI slices as separate images stacked on the channel dimension [35]. However, these images do not respect the topography

of neural representations and contain a large fraction of non-responsive voxels corresponding to white matter and cerebrospinal fluid. This results in most of the contrast of the images depending on irrelevant anatomical factors. Another possibility is to use spatial 3D convolutions on the brain volume [1]. This method has the benefit of preserving the topography of the neural responses but otherwise has the same issues as the 2D approach. These shortcomings make such methods unsuitable for brain decoding and reconstruction. A more viable strategy is to map the voxel responses on a mesh representing the cortical surface [8] and apply a geometric deep learning technique [29, 7, 4, 24].

3 Brain2Pix

Our brain2pix architecture has two components: 1) a receptive field mapping that transforms the brain activity of visual regions to a tensor in pixel space, exploiting the topographical organization of the visual cortex; 2) a pix2pix network that converts the brain responses in pixel space to realistic looking natural images. In the following, we describe the two components in detail.

3.1 From voxels to pixels

A receptive field mapping is a (potentially many-to-one) function that maps the 3D coordinate of the voxels of a visual area to Cartesian coordinates in the stimulus space. This coordinate is defined as the region of the image that elicits the highest response in the voxel. Given a visual ROI, we can refer to these mappings using the following notation:

$$\mathrm{RF}(r_1, r_2, r_3) = (x, y) \,, \tag{1}$$

where (r_1, r_2, r_3) are the voxel coordinates and (x, y) is a pair of coordinates in the image space. Since visual areas are topographically organized, this map can be seen as an approximate homeomorphism (i.e. a function that preserves the topology). Note that RF does not respect the metric structure of the image since the representation of the fovea is inflated while the periphery is contracted. We denote the function associating a measured neural activation (BOLD response) to each voxel as $n(r_1, r_2, r_3)$. Using the receptive field mapping, we can transport this activation map to pixel space as follows:

$$n(x', y') = \frac{1}{M(x', y')} \sum_{\substack{r_1, r_2, r_3 \\ \mathrm{RF}(r_1, r_2, r_3) = (x', y')}} n(r_1, r_2, r_3) \,, \tag{2}$$

where $M(x', y')$ is the number of voxels that map to the coordinates (x', y'). Eq. 1 is limited to the case of point-like receptive fields. More generally, the RF transport map can be written as a linear operator:

$$n(x', y') = \sum_{r_1, r_2, r_3} W^{x', y'}_{r_1, r_2, r_3} n(r_1, r_2, r_3) \,, \tag{3}$$

where the weight tensor W is a (pseudo-)inverse of the linear response function of the cortex under single pixel simulations. This second formulation has the benefit of allowing each voxel to contribute to multiple pixels and to be suitable to gradient descent training.

In this paper we use two strategies for determining W. The first approach, is to apply an off-the-shelf receptive field estimator and to use Eq 1. The second, more machine learning oriented approach is to learn to weight matrix together with the network. In order to preserve the topographical organization, we include the learnable part as a perturbation of the receptive field estimation:

$$n(x', y') = \sum_{r_1, r_2, r_3} \left(\frac{\delta^{(x', y')}_{\mathrm{RF}(r_1, r_2, r_3)}}{M(x', y')} + V^{x', y'}_{r_1, r_2, r_3} \right) n(r_1, r_2, r_3) \,, \tag{4}$$

where δ^y_x is the discrete delta function and the wights $V^{x', y'}_{r_1, r_2, r_3}$ are learnable parameters.

3.2 Image-to-image network

If brain responses were linear and the distribution of natural images were Gaussian, an equation of the form given in Eq. 2 would produce an accurate reconstruction [36]. However, brain responses to naturalistic images have strong nonlinearities since large-scale features influence low-level responses.

Furthermore, the space of natural images is highly non-Gaussian, living in a lower-dimensional manifold with a complex geometry [51]. The problem of using a proper "prior" (i.e. regularization) on the space of natural images is particularly important as the fMRI signal has a low SNR and it therefore does not contain enough information to fully reconstruct the presented image. Therefore, we cannot expect a simple linear Gaussian solution to provide meaningful reconstructions. However, Eq. 1 and Eq. 2 contain the relevant topographical organization of the brain responses since the topography estimated under the linear approximation is well-preserved in the non-linear regime [45]. In other words, non-local non-linear effects can be interpreted as perturbations on the local linear responses. In order to account for these effects while also accounting for the structure of the natural images, we transform the activation maps in pixel space using a pix2pix network trained with a combination of feature, pixel and adversarial loss. The pix2pix convolutional architecture can exploit the low-level topography while introducing the global features necessary for generating natural videos.

The input to the pix2pix network is a tensor obtained by stacking the activation maps, one map for each combination of ROI and time lag. In fact, the network needs to integrate the topographically organized information contained in several layers of the visual hierarchy (V1, V2 and V3 in our case) but also the responses at different time lags as the BOLD response introduces a time shift.

3.2.1 Architecture

The architecture of the brain2pix model is inspired by the pix2pix architecture [21] which comprises a convolutional U-Net-based generator [33] and a convolutional PatchGAN-based discriminator (Figure 1). The first and the last layers of the generator are respectively convolutional layer and devoncolutional with four standard U-net skip blocks in between. All five layers of the discriminator are convolutional with batch normalization and leaky ReLU activation function.

The discriminator was trained to distinguish stimuli from their reconstructions by iteratively minimizing a loss function with a sole adversarial loss (binary cross-entropy) and using a history buffer to encourage the discriminator to remember past errors.

The generator was trained for converting brain responses to stimulus reconstructions by iteratively minimizing a loss function with three weighted components: i) pixel-loss, which was taken to be the absolute difference between ground-truths and predictions, ii) feature loss, which was taken to be the Euclidean distance between pretrained layer 10 VGG features of ground-truths and predictions and iii) adversarial loss, which was taken to be the "inverse" of the adversarial loss that was used to train the discriminator.

All models were implemented in Python with the MXNet framework [3]. They were trained and tested on Nvidia GeForce 2080 Ti GPUs.

3.2.2 Receptive field estimation

Receptive fields for dorsal and ventral visual regions V1, V2 and V3 were estimated in a data-driven way using *neural information flow* [37]. Grayscale video sections were passed through three 3D convolutional neural network layers corresponding to the visual ROIs. Before the ROI-specific layers a linear layer with a single $1 \times 3 \times 3$ channel was used to allow learning retinal and LGN preprocessing steps. Average pooling was applied after each layer to account for increasing receptive field sizes, the temporal dimension was average pooled to a TR of 700ms before applying the observation models, and spatio-temporal receptive fields were constrained to be positive. For training this neural network, low-rank tensor decomposition was applied to estimate voxel-wise spatial, temporal and channel observation (readout) vectors, which were used to predict voxel-wise activity from the neural network activity tensors. The receptive field location (x, y) for every voxel was then estimated as its center of mass of the low-rank receptive field maps.

4 Experimental procedure

4.1 Data acquisition

We made use of a large fMRI dataset from single-participant responses to naturalistic stimuli, which was published by [38]. The exact experiments performed to obtain the data are explained in detail

in the original study [38]. In short, the participant fixated on a fixation cross on the screen while watching 30 episodes of BBC's Doctor Who. The videos were presented while the BOLD response was measured from the brain in multiple runs; 121 runs were used for the training set and 7 runs for the test set. Each trial of the test set were repeated 10 times and were ultimately averaged across repetitions for model evaluation.

4.2 Data preprocessing

Prior to utilizing the inputs for training the model, 3D brain matrices were transformed to 2D receptive field signal images (RFSimages) in two main steps. First, regions of interests (ROIs) were selected from the brain (V1, V2, V3), based on their corresponding masks. Second, each voxel in that brain region was mapped onto its corresponding visual space based on the retinotopic map. The 2D RFSimages were adjusted to 96×96 pixels (width $\times$ height) and separated by time channels (5) and brain regions (N ROIs). For each input, five time-points were selected, resulting in 15 input channels into the model per sample (in the case of V1, V2, V3).

The videos were downsampled spatially ($96 \times 96 \times 3$) and temporally to match the TRs of the fMRI recordings (one frame every 0.7 s). This resulted in a total of 7459 video frames for training and 1034 video frames for model evaluation. Considering the haemodynamic delay, we realigned the stimuli and brain signals such that the current signals correspond to the stimuli that were presented at 4 timesteps before, allowing a time window of 2.8s – 5.4s delay. Finally, each frame underwent a fish-eye transformation, which mimics biological retinal eccentricity [2]. The receptive field centers we used for mapping brain signals onto the visual space were based on images that underwent this transformation.

5 Experiments

Once the brain signals were mapped onto visual space and the model was assembled, we ran it and compared the performance with alternative reconstruction models. This included a baseline comparison where our model is compared with traditional models. Since we wanted to focus on early visual areas, we also trained our model on V1, V2, and V3 individually (which we called the ROI experiment). Finally, we tested whether our model was robust to various ablations. The same four evaluation metrics were used in all experiments: Pearson's product-moment correlation coefficient (corr.) and Euclidean distance (dist.) between the features of test stimuli and their reconstructions. Features were extracted with both the first dense layer of an AlexNet model [25] and the only global average pooling layer of a ResNet18 model [12]. Both models were pretrained on ImageNet [34]. Additional details of the experiments, additional results and a link to the source code are provided in the supplementary materials.

5.1 Simulation experiments

Before experimenting on real data, we used synthetic data to test feasibility and tune hyperparameters of brain2pix. Instead of mapping brain signals per ROI onto visual space, we mapped target stimuli onto visual space using the same exact method. This filtering of target images with the RF centers gave us the same amount of input pixels for the model all at the same location. However, their activations were not based on actual brain signals, but rather the target image itself. We got very clear reconstruction images from this, which confirmed that the number of RF pixels provided by V1 + V2 + V3 ROIs could theoretically carry enough spatial information for the model to generate realistic and accurate results.

5.2 Baseline experiments

In our first experiment on real data, we evaluated two variants of our model as well as comparing them with two established decoding models from the literature.

The brain2pix variants differed only in how they transformed brain responses from volumetric representation to image representation. The first variant (referred to as fixed RF) used the RF estimates alone. The second variant (referred to as learned RF) used a dense layer to perturb the image representation in the fixed RF variant as a function of the volumetric representation.

Figure 2: Baseline experiment: Comparison between the reconstructions of the brain2pix and the baseline models. The reconstructions of the brain2pix are shown in columns 2 and 4, with their corresponding ground truth (GT) (column 3) and inputs (columns 1 and 5). The reconstruction of the baseline models are shown in columns 6 and 8 with their corresponding GT (column 7).

The first (simple) baseline reconstructed the stimuli by inverting a linear Gaussian encoding model with map estimation [36]. The second (more complex) baseline reconstructed the stimuli by maximizing the likelihood of a nonlinear-linear encoding model with a SqueezeNet component [18] as the nonlinear feature extractor (second max-pooling layer outputs of the SqueezeNet V2 architecture pretrained on ImageNet). All densities were assumed to be Gaussian except for the prior which was an empirical natural image prior constructed from the training set [30, 31]

We found that the learned RF variant has better qualitative and quantitative (except for one metric) reconstructions (see Table 1). However, the improvement was not large, suggesting that the RF model captures the correct topographical structure. Both brain2pix variants had significantly above chance level performance ($p < 0.05$; t-test) and significantly outperformed both baselines ($p < 0.05$; binomial test) (Figure 2).

5.3 ROI experiments

In order to isolate the role of the regions of interest, we performed a series of follow up experiments where only one ROI was given to the network. We used a fixed receptive field matrix (Eq. 1). All the experimental details are identical to the main experiment. Figure 3 shows the ROI-specific

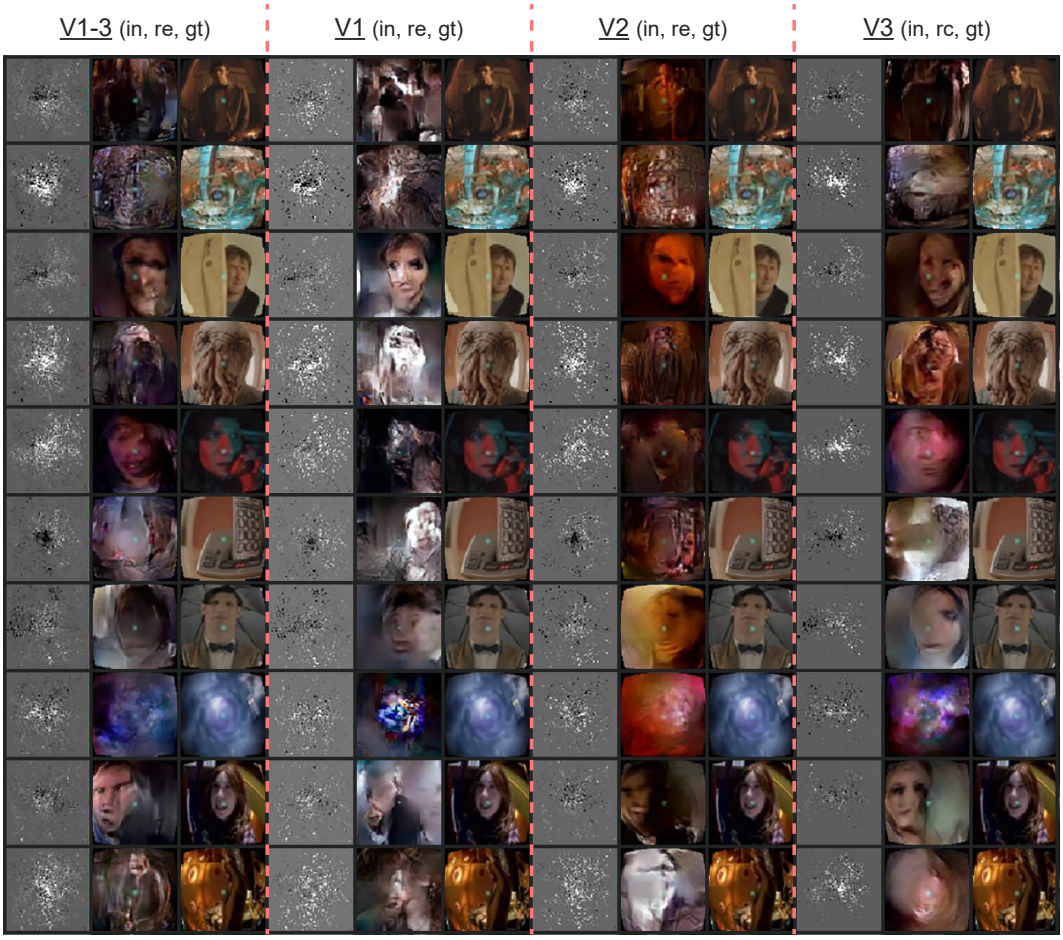

Figure 3: ROI experiment: Reconstructions from the brain2pix fixed RF method trained on various brain regions. Columns 1–3 show the inputs (in), reconstructions (re) and ground truths (gt) of all combined regions (V1–V3), respectively. Columns 4–6 show these results for only V1, columns 7–9 for V2, and finally columns 10–12 are results belonging to the model trained only on V3.

Table 1: Baseline experiment: Correlation and distance values between reconstruction and target features obtained from passing through a pretrained network (Alexnet and Resnet). Here comparisons are done for the brain2pix models and baseline models. Bold values indicate the highest correlation or lowest distance values.

	brain2pix (learned RF)	brain2pix (fixed RF)	DNN	Linear
Alexnet corr.	**0.4684 ± 0.0032**	0.4608 ± 0.0031	0.3806 ± 0.0071	0.3535 ± 0.0104
Resnet corr.	0.4456 ± 0.0038	**0.4603 ± 0.0036**	0.3652 ± 0.0097	0.3518 ± 0.0103
Alexnet dist.	**0.0852 ± 0.0004**	0.0899 ± 0.0004	0.1014 ± 0.0011	0.1279 ± 0.0013
Resnet dist.	**0.0808 ± 0.0003**	0.0814 ± 0.0003	0.1224 ± 0.0011	0.1651 ± 0.0013

Table 2: ROI experiment: Correlation and distance values between Alexnet and Resnet features for the ROI experiment. V1, V2, V3 are the individual ROIs and V1–V3 are all three ROIs combined.

	V1-3	V1	V2	V3
Alexnet corr.	**0.4608 ± 0.0031**	0.2597 ± 0.0032	0.4419 ± 0.0029	0.4320 ± 0.0032
Resnet corr.	**0.4603 ± 0.0036**	0.2438 ± 0.0036	0.4413 ± 0.0031	0.4599 ± 0.0035
Alexnet dist.	**0.0899 ± 0.0004**	0.1094 ± 0.0004	0.0939 ± 0.0003	0.0949 ± 0.0004
Resnet dist.	**0.0814 ± 0.0003**	0.0904 ± 0.0002	0.0840 ± 0.0002	0.0825 ± 0.0003

Table 3: Ablation experiments: Correlations and distances between Alexnet and Resnet features of the test stimuli and their reconstructions. The brain2pix is compared with three models with either no feature loss, no adversarial loss or no training.

	brain2pix	No feature	No adversarial	No loss
Alexnet corr.	**0.4608 ± 0.0031**	0.4266 ± 0.0028	0.1381 ± 0.0022	0.0313 ± 0.0018
Resnet corr.	**0.4603 ± 0.0036**	0.4484 ± 0.0036	0.1808 ± 0.0032	- 0.0275 ± 0.0019
Alexnet dist.	**0.0899 ± 0.0004**	0.1337 ± 0.0004	0.0913 ± 0.0004	0.1912 ± 0.0002
Resnet dist.	0.0814 ± 0.0003	**0.0792 ± 0.0003**	0.1463 ± 0.0003	0.1188 ± 0.0002

reconstructions. Reconstructions based on V1 tend to have sharper pixelwise correspondence but some more global features such as the overall color were not captured very well. Combined brain2pix model with all ROIs and V3 reconstructions on the other hand were able to capture the color profile of the scenes very well. These two models generated images that captured further readily interpretable high-level information such as the existence of a person in the scene and even the expressions on the faces of individuals in the scenes. It is interesting to note that the ROIs did not contain higher-level brain regions, such as lateral occipital cortex that play a large role object perception and fusiform face area that specializes in face processing. The quantitative results are given in Table 2. The combined model performs substantially better than the individual models with the V1 model having the worst performance. V2 and V3 were similar to each other in quantitative performance. The poor performance of V1 is likely due to the fact that the adversarial and feature losses had a larger weight in the training process, biasing the model towards using higher-order features for reconstruction.

5.4 Ablation experiments

The ablation studies were performed to test the impact of VGG-loss and adversarial loss on the performance of the model. We ran the model with the removed components and compared the final results in Table 3. "No adversarial" refers to the brain2pix without a discriminator loss, using only the VGG-feature loss to optimize the model. In this ablation case, the model did not learn to reconstruct images but rather outputting square patterns that repeated across all images. The second model is the "no feature" model which was trained without the VGG-loss, only making use of the adversarial loss. This resulted in images that look like reconstructions but did not approximate the target.

6 Conclusions

In this paper, we introduced a new brain reconstruction method, brain2pix, that exploits the topographic organization of the visual cortices by mapping brain activation to a linear pixel space where it is then processed with a fully convolutional image-to-image network. To the best of our knowledge, this is the first approach capable of generating semantically accurate reconstructions from a naturalistic video stream. In our current experiment we only used responses from the early visual regions V1, V2, V3. A natural extension of the current work is including higher level areas in the temporal and parietal cortex in the analysis pipeline. Since these areas process coarse-grained semantic information, experiments feeding their responses to deeper layers of the network could reveal even further progress and more meaningful reconstructions.

Broader Impact

Neural decoding studies are crucial for understanding the functioning of the human brain, broadly benefiting the field of neuroscience. Furthermore, neural decoding algorithms make up a major component of brain computer interfaces. Brain computer interfaces enable disabled people to perform tasks that they would not be able to perform otherwise, by substituting their lost faculties. These technologies can range from a communication interface for a locked-in patient, to a neuroprosthetic limb, and more. While the algorithms that we develop and study in this paper are specialized to reconstruct visual stimuli from brain responses, we foresee that the suggested principles can be applied to different applications, with some adaptations. For instance, we use a relatively slow signal (BOLD response), which reflects the neural responses that take place several seconds prior to them. A time-critical BCI system would need to make use of a signal with no such delays to perform well. While admittedly the promises of these algorithms to reconstruct the viewed stimuli is yet to be fully achieved, scientists that attempt to extract information from the brain should make sure the safety and privacy of the users. Our experiments were performed on prerecorded data that strictly followed data safety regulations (GDPR), and experimental procedures were approved by the relevant ethical committees. Future studies should make sure to follow similar strict regulations, ensuring only a positive impact of these fascinating methods that allow us to peek into the human mind.

References

[1] K Bäckström, Md Nazari, I Y Gu, and A S Jakola. An efficient 3d deep convolutional network for Alzheimer's disease diagnosis using MR images. In *2018 IEEE 15th International Symposium on Biomedical Imaging*, pages 149–153. IEEE, 2018.

[2] Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):9436, 2019.

[3] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. 2015.

[4] T S Cohen, M Geiger, J Köhler, and M Welling. Spherical CNNs. In *International Conference on Learning Representations*, 2018.

[5] C Dong, C C Loy, and X Tang. Accelerating the super-resolution convolutional neural network. In *European Conference on Computer Vision*, pages 391–407. Springer, 2016.

[6] Serge O Dumoulin and Brian A Wandell. Population receptive field estimates in human visual cortex. *Neuroimage*, 39(2):647–660, 2008.

[7] M Fey, Jan Eric L, F Weichert, and H Müller. Splinecnn: Fast geometric deep learning with continuous B-spline kernels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 869–877, 2018.

[8] B Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[9] Yağmur Güçlütürk, Umut Güçlü, Katja Seeliger, Sander Bosch, Rob van Lier, and Marcel AJ van Gerven. Reconstructing perceived faces from brain activations with deep adversarial neural decoding. In *Advances in Neural Information Processing Systems*, pages 4246–4257, 2017.

[10] Yağmur Güçlütürk, Umut Güçlü, Rob van Lier, and Marcel AJ van Gerven. Convolutional sketch inversion. In *European Conference on Computer Vision*, pages 810–824. Springer, 2016.

[11] Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. Variational autoencoder: An unsupervised model for encoding and decoding fMRI activity in visual cortex. *NeuroImage*, 198:125–136, sep 2019.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. 2015.

[13] Salomon Eberhard Henschen. On the visual path and centre. *Brain*, 16(1-2):170–180, 1893.

[14] Gordon Holmes and WT Lister. Disturbances of vision from cerebral lesions, with special reference to the cortical representation of the macula. *Brain*, 39(1-2):34–73, 1916.

[15] Tomoyasu Horikawa and Yukiyasu Kamitani. Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features. *Frontiers in Computational Neuroscience*, 11:4, 2017.

[16] Tomoyasu Horikawa, Masako Tamaki, Yoichi Miyawaki, and Yukiyasu Kamitani. Neural decoding of visual imagery during sleep. *Science*, 340(6132):639–642, 2013.

[17] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.

[18] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. 2016.

[19] S Iizuka, E Simo-Serra, and H Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics*, 35(4):1–11, 2016.

[20] Tatsuji Inouye. Die sehstorungen bei schussverletzungen der kortikalen sehsphare. *Nach Beobachtungen an Verwundeten der letszten Japanischen Kriege*, 1909.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[22] J Kim, J Kwon L, and K Mu L. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1645, 2016.

[23] Peter Kok, Janneke FM Jehee, and Floris P De Lange. Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*, 75(2):265–270, 2012.

[24] R Kondor, Zhen Lin, and S Trivedi. Clebsch–gordan nets: a fully Fourier space spherical convolutional neural network. In *Advances in Neural Information Processing Systems*, pages 10117–10126, 2018.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[26] Y Li, H Qi, J Dai, X Ji, and Y Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2359–2367, 2017.

[27] J Long, E Shelhamer, and T Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[28] Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008.

[29] F Monti, D Boscaini, J Masci, E Rodola, J Svoboda, and M M Bronstein. Geometric deep learning on graphs and manifolds using mixture model CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5115–5124, 2017.

[30] Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

[31] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011.

[32] H Noh, S Hong, and Bg Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[35] S Sarraf and G Tofighi. Classification of Alzheimer's disease using fMRI data and deep learning convolutional neural networks. *arXiv preprint arXiv:1603.08631*, 2016.

[36] S Schoenmakers, M Barth, T Heskes, and M Van Gerven. Linear reconstruction of perceived images from human brain activity. *NeuroImage*, 83:951–961, 2013.

[37] K Seeliger, L Ambrogioni, U Güçlü, and MAJ van Gerven. Neural information flow: Learning neural information processing systems from brain activity. *BioRXiv*, 2019.

[38] K Seeliger, RP Sommers, U Güçlü, SE Bosch, and MAJ van Gerven. A large single-participant fMRI dataset for probing brain responses to naturalistic stimuli in space and time. *BioRxiv*, page 687681, 2019.

[39] Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and Marcel AJ van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, 2018.

[40] A Selim, M Elgharib, and L Doyle. Painting style transfer for head portraits using convolutional neural networks. *ACM Transactions on Graphics*, 35(4):1–18, 2016.

[41] Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, 13:21, 2019.

[42] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS Computational Biology*, 15(1):e1006633, jan 2019.

[43] Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4):1104–1116, 2006.

[44] Marcel AJ van Gerven, Floris P de Lange, and Tom Heskes. Neural decoding with hierarchical generative models. *Neural Computation*, 22(12):3127–3142, 2010.

[45] A L Vazquez and D C Noll. Nonlinear aspects of the bold response in functional MRI. *Neuroimage*, 7(2):108–118, 1998.

[46] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, 28(12):4136–4160, oct 2017.

[47] R Zhang, P Isola, and A A Efros. Colorful image colorization. In *European Conference on Computer Vision*, pages 649–666. Springer, 2016.

[48] R Zhang, J Zhu, P Isola, X Geng, A S Lin, T Yu, and A A Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017.

[49] Y Zhang, Z Qiu, T Yao, D Liu, and T Mei. Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6810–6818, 2018.

[50] Y Zhang, Y Tian, Y Kong, B Zhong, and Y Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018.

[51] J Zhu, P Krähenbühl, E Shechtman, and A A Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.