Relation between Visual Stimuli and Brain Response: Reconstructing image from fMRI data

沈乙晨 (Melody)
110022109
Department of Physics
National Tsing Hua University
Hsinchu, Taiwan
melodyshen152@gmail.com

Jeremias Rück
x1120071
School of Computation Information and Technology
Technical University of Munich
Munich, Germany
jeremias.rueck@tum.de

黃政倫 (Icarus)
110060004
Department of Electrical Engineering and Computer Science
National Tsing Hua University
Hsinchu, Taiwan
jungle920525@gmail.com

歐偉興 (Arithat)
110006214
Department of Electrical Engineering and Computer Science (Global Program)
National Tsing Hua University
Hsinchu, Taiwan
a.ariyanuchitkul@gmail.com

黃明霞 (Michelle)
110006311
Department of Electrical Engineering and Computer Science (Global Program)
National Tsing Hua University
Hsinchu, Taiwan
mchllgrcflc@gmail.com

杜銘勝 (Christian Owen)
110006217
Department of Electrical Engineering and Computer Science (Global Program)
National Tsing Hua University
Hsinchu, Taiwan
chris27owen@gmail.com

***Abstract*—This research delves into the relation between visual stimuli and brain responses by employing functional Magnetic Resonance Imaging (fMRI) data. Through innovative methodologies, we aim to reconstruct visual images based on the patterns of brain activity observed during exposure to specific visual stimuli. This study not only seeks to deepen our understanding of the neural mechanisms underlying visual perception but also holds the potential to pave the way for advanced applications in neuroimaging and cognitive neuroscience.**

Keywords—visual; brain; fmri; image reconstruction; gan

I. Introduction

Visual cognitive learning has long been implemented by humans ever since the birth of mankind. At least 65 percent of the humans now are considered as visual learners [16] which have shown to improve learning rates for both comprehension and memorization. Within the last decades, to show how this visual cognition affects our brain, the human visual decoding data [17][18][19] is extracted from the analysis of multi-voxel functional Magnetic Resonance Imaging (fMRI) [20] patterns which are generated from the blood oxygen levels within the certain regions of interest (ROI) in the human brains [21][22]. Through this breakthrough, neuroscientists, psychologists, and Artificial Intelligence programmers have sought this data to further improve their research regarding the correlation between visual stimuli and brain responses, driven by the importance of understanding how the human brain processes this visual information.

Human visual decoding can be categorized into stimuli category classification, stimuli identification, and visual reconstruction [23]. The stimuli category classification allows the brain to categorize the presented stimuli into discrete object categories [17][19]. The stimuli identification identifies the stimulus that corresponds to the pattern that reacts in the region of interest (ROI) from the ground truth stimuli images [24][23]. Thus, the main goal of stimuli identification and visual reconstruction is to verify the details of given images by their position, size, and angles, but the reconstruction of an image requires the use of a given fMRI data to create a replica of the stimulus which is known to be a challenge as it is needed to create machine learning models to recreate the picture. However, the fMRI-based visual reconstruction will enable us to improve our understanding of the mechanisms behind the brain's visual processing [10].

However, the huge amount of features and the absence of regularization in the regression model may cause the

decoding to be less accurate, causing the reconstruction of the image to be far-fetched from the original one[9].

By referring to existing studies and related work, this project seeks to build upon the previous research by analyzing how visual stimuli and neural responses are connected, allowing us to be able to comprehend human cognition and perception.

II. Related works

In this paper, we refer to Phillip Isola's previous paper, which explores the application of conditional adversarial networks (cGANs) as a general-purpose solution for image-to-image translation tasks. It emphasizes the effectiveness of cGANs in learning a loss function adapted to the specific task and data at hand, making them applicable in a wide variety of settings. The paper also discusses the community-driven research and the diverse applications of the pix2pix codebase, demonstrating its widespread adoption and exploration for various image translation tasks. Additionally, the study presents a simple framework for achieving good results and analyzes the effects of several important architectural choices. The research paper contributes to the understanding of conditional GANs as a versatile solution for image translation tasks and provides insights into the architectural choices for the generator and discriminator. [14]

III. Methods

There are several aspects to be looked upon on the objective of this project which includes the datasets that are used, preprocesses of the data input, the classification of the fMRI data to learn its implementation in this field, and the reconstruction of the image from the input data with the masking from region of interest of the brain. There will also be a few approaches towards the main goal of the topic and it will be elaborated in the following documents. Below are the approaches that are implemented in this work:

A. Collection of Datasets

In this work, the dataset comes from the Natural Scenes Dataset (NSD) (Allen et al., 2022), a massive dataset of 7T fMRI responses to images of natural scenes coming from the COCO database [25]. The COCO image dataset also includes the 133 ground truth categories that classify the labels available in the images. Additionally, the Region of Interest (ROI) mask is provided to improve the model performance by focusing the accuracy of the relevant brain area.

The functional Magnetic Resonance Imaging (fMRI) data contains the data of the left and right hemisphere of the brain with 19004 pixels and 20544 pixels correspondingly, each containing the binary number 0 or 1 that is represented as black and white as an image. Whereas, the ground truth images contain 425 x 425 coloured (RGB) pixels.

The Region of Interest (ROI) mask is used to acquire the human brain activity (voxels) from a specific region on the fMRI data. This allows the input data to be precise as to which part of the brain is to be selected for the visual stimuli tasks. As there are various different parts of masks that are included in the dataset, the research focused its attention only on the visual regions (V1v, V1d, V2v, V2d, V3v, V3d, hV4).

B. Preprocessing data

The datasets that are given are preprocessed into the right shape and rationalized into training and validation data as shown in Figure 1.

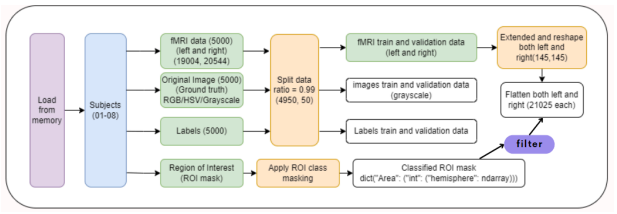

Fig. 1. Data preprocessing diagram

The training and validation data is being split with a 8 to 2 ratio. The reason behind this is to ensure better results by more data training to train both the image reconstruction model and the image classification model.

For the image reconstruction model, the fMRI data and ROI are not flatten to preserve spatial information and pattern across the 2D data. Whereas in the classification model, The fMRI data that is split into validation and training would then be flattened to and reshape into the same shape for both left and right part of the brain. The ROI masks are then applied to the flatten image to act like filters that select the data to be fed to the models independently.

On the other hand, We've noticed that some labels, although present in the images, occupy a very small area, sometimes even challenging for the human eye to discern. This situation can lead to potential misclassifications. Consequently, we've opted to utilize the bounding box annotations for labels provided by the COCO dataset. This information helps us to calculate the actual proportion of a label within an image. Labels with disproportionately low coverage are then excluded, allowing the model to concentrate on information that aligns more closely with what the human brain is likely to process.

C. Image Classification

Before starting with our main model, the image reconstruction, we built an additional independent model to

classify the labels and in order to learn the basics usage of the fMRI data before the reconstruction of the image. With the integration of fMRI and visual images, our goal is to build a classifier model that can define categories of things out of the 133 different classes present in the images. The methods encompass data preprocessing, model architecture designing, model training, and evaluation metrics.

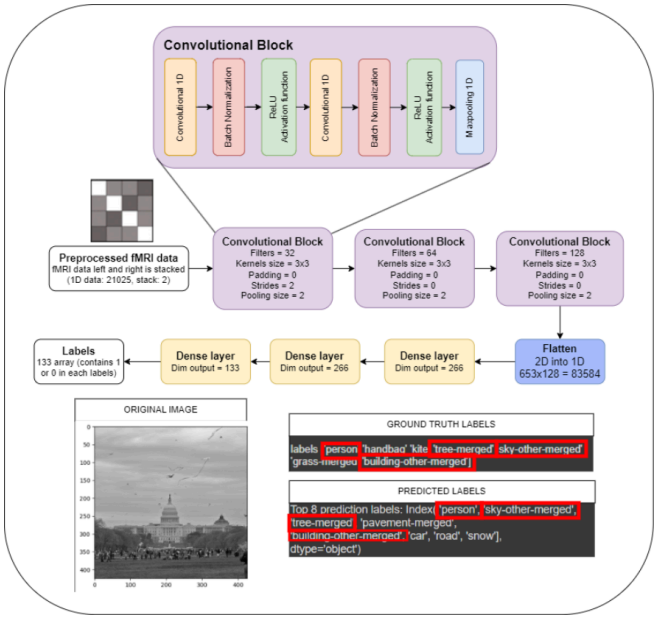

Fig. 2. CNN model for Image Classification

In Figure 2, we implemented the classification with a convolutional neural network (CNN) that convoluted the fMRI data using three convolutional blocks that will be max pooled, flatten and dense into the labels later on. The main activation function that we are using is ReLU and the cross entropy using focal loss. The detailed variables and arguments for the convolution block can be found in Figure 2 above.

There is also the addition of the bounding block, which is a rectangular frame drawn around an object in an image to identify and locate its position, to filter the labels that are too small on the actual ground truth image.

D. Image Reconstruction

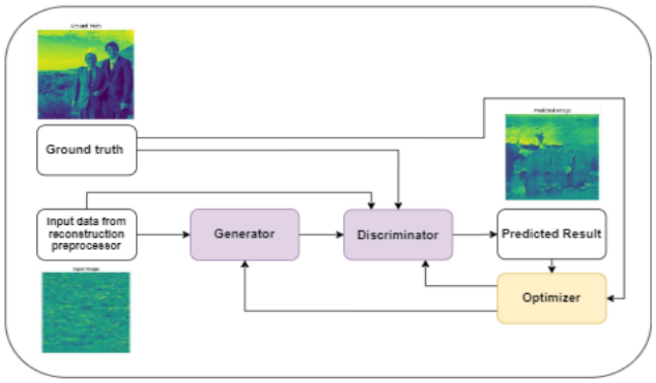

Fig. 3. General GAN diagram

Our main objective's model for image reconstruction is built using the GANs model which works like shown in Figure 3. Generative Adversarial Networks (GANs) are a network architecture consisting of a generator and a discriminator (I. J. Goodfellow et al., 2014). The generator generates data whereas the discriminator decides whether it is the ground truth data or the generated fake data. GAN's model's powerful performance relies on the learned loss function of the discriminator as described in [1]. This is because the generator competes to produce more realistic data, while the discriminator competes to upgrade its skill in distinguishing ground truth data from fake data.

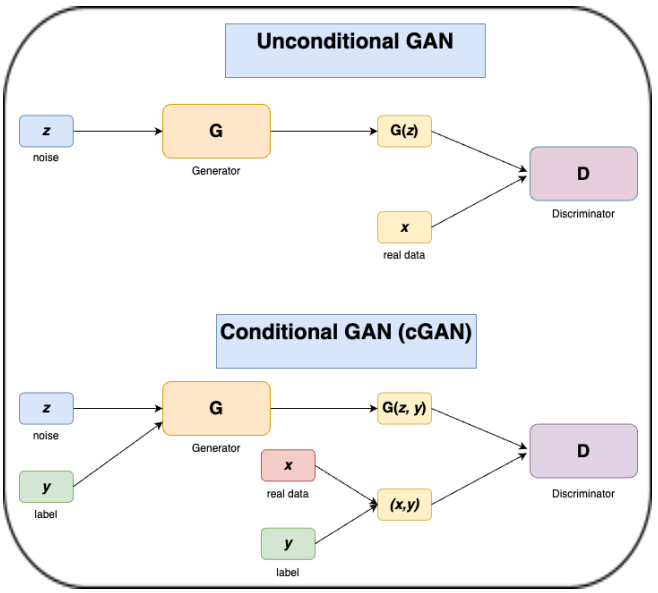

Fig. 4. Unconditional GAN and Conditional GAN (cGAN) model diagram

Albeit, the promising performance of unconditional GAN with the formula:

$$x = G(z) \tag{1}$$

In this project, we used conditional GANs (cGANs) which uses the following formula:

$$x = G(z, y) \tag{2}$$

Unconditional GANs differ with cGANs in terms of that cGANs generate outputs according to the input data. The generator in cGANs takes random noise and condition information (in our case, the labels) to produce the output. Whereas in unconditional GANs, the generator generates the output solely from random noise [4]. We use the cGANs with the label data to improve the accuracy and better control the generated output.

Our cGAN model focuses on the objective of:

$$G^* = arg \min_G \max_D L_{cGAN}(G, D) + \lambda L_{L1}(G) \quad (3)$$

Where the discriminator D aims to maximize the loss and competes with the generator G that minimizes the loss. For this project's loss function, we use both cGAN loss function and L1 loss function since as described in [1], the combination of both reconstructs more accurate images. The L1 and cGAN loss are described as Eqn. 4 and Eqn. 5 respectively.

$$L_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x, z)\|_1] \quad (4)$$

$$L_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log (1 - D(x, G(x, z)))] \quad (5)$$

1. Additional Data Preprocessing

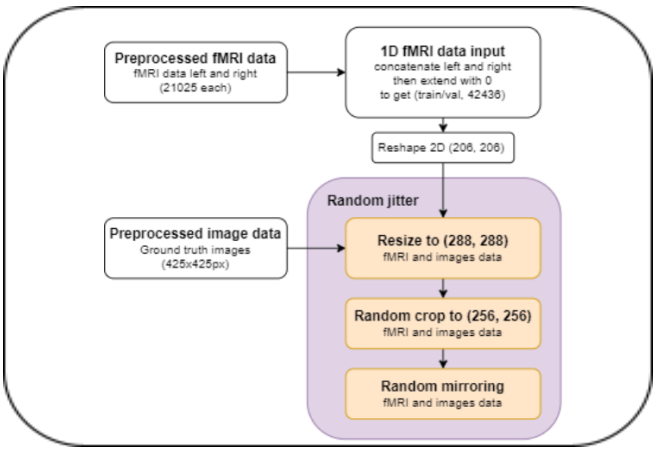

Fig. 5. fMRI data preprocessing using Random Jitter

We did additional data preprocessing for the image reconstruction model by applying the Random Jitter method as shown in Figure 4. The Jitter method can prevent overfitting and introduce more randomness to create a better generator (Z. Cai et al., 2021). This method works by enlarging the fMRI input data to 288x288 pixel size from 256x256, then crop back to its original size randomly.

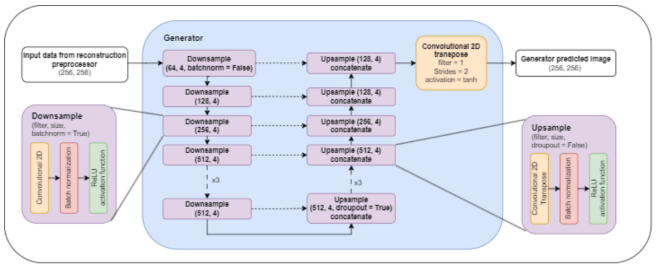

Fig. 6. The cGAN generator to reconstruct the image from fMRI

2. Generator

In general, the generator is made out of encoder-decoder architecture, where the model keeps downsampling the input data until a bottleneck layer, then it gradually upsampling until it reaches the final layer to generate to output. With this architecture, there might be some information loss along the downsampling process.

Therefore, we use U-Net for the generator. The term U-Net defines CNN but with skip connections like shown in Figure 6. It works by enabling connections between *i*th layer to the (n-i)th layer in the n layers generator architecture. These connections concatenate channels from the ith layer to the n-i (n-i)th layer. This U-Net model can help the generator to yield better quality images with higher precision, but with less training data [8].

For the generator, we applied Batch Normalization followed by the ReLu activation function each downsample and upsample using 2D-CNN.

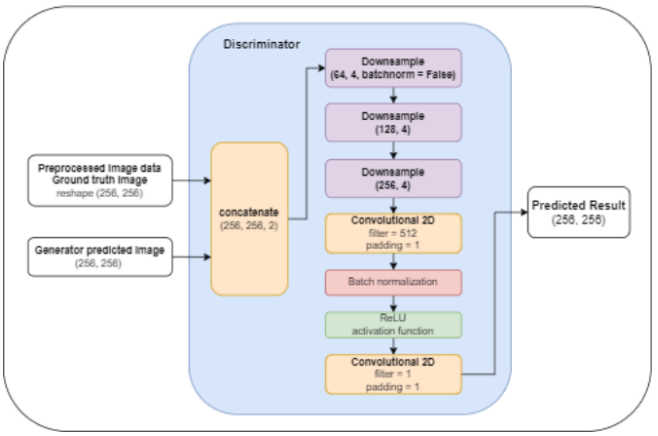

Fig. 7. The cGAN discriminator to evaluates the output

3. Discriminator

The discriminator we use is of type PatchGAN. This term defines a type of GAN discriminator that classifies the authenticity of each NxN patch from the input image and decides whether the image is fake or not. The model architecture can be examined in Figure 7.

Previous paper [1] showed promising results from this method by using 70x70 pixel small patches to evaluate the realness of the input. Therefore, we decided to use 4x4 patches towards the downsampled image with size of 33x33 to increase the accuracy. With 1 stride, the output results in 30x30 size, which each grid shows the decision whether the image is real or generated. The discriminator then averages

the overall output and can decide by using the 0.5 threshold. This means if the results surpass 0.5, then it recognizes the image as real.

The model can run faster even with large images as it relies on the PatchGAN for the high-frequency correctness, and solely uses L1 term to ensure low frequencies (Eqn. 3). This is because, as discussed in [7], PatchGAN may be used as a type of texture loss which works well under the assumption of independence between each patch. Consequently, PatchGAN combined with L1 is already accurate enough and no longer needs L2 loss.

4. Model Strategies

Before being fed up to the model, the data is normalized using min-max normalization. As for the optimization algorithm, our model uses an Adam solver with a learning rate of 0.0002, and momentum parameter of 0.5 for both the generator and the discriminator. Lastly, our model with Pix2Pix GAN architecture, utilizes GAN loss and L1 loss as our loss function. The combination of GAN loss (adversarial loss) and L1 loss (mean absolute error) through the adversarial process are proven to assure structural similarity of the generated image.

E. Experimental Image Reconstruction Approach

Using a self made reconstruction model which is mainly based on the basic Convolutional Neural Network (CNN) model, autoencoders, and simple optimization such as batch normalization and droppings, we are able to bring the result of fMRI reconstruction without ROI masking of the input data. We attempted to do image reconstruction using an encoder-decoder model with 1D CNN. However, the model exhibited underfitting and failed to generate a similar image as shown in Fig. 8. Subsequently, we explored upscaling the dimensions and employing 2D CNN. Unfortunately, this approach still resulted in a blurry image.

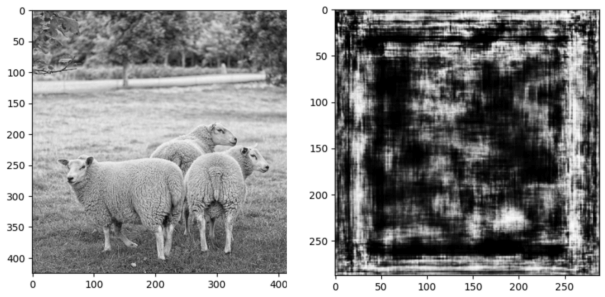

Fig. 8. Image reconstruction from the GAN model

F. Abbreviations and Acronyms

fMRI: Functional Magnetic Resonance Imaging; GAN: Generative Adversarial Network; cGAN: Conditional Generative Adversarial Network; CNN: Convolutional Neural Network; NN: Neural Network; 1D: one dimension; 2D: two dimension; RGB: Red Green Blue; ROI: Region of Interest; SSIM: Structural Similarity Index;

IV. RESULTS

1. Image classification model

The learning curve of the image classification model is shown in Fig.9. The training and validation dataset both converged.

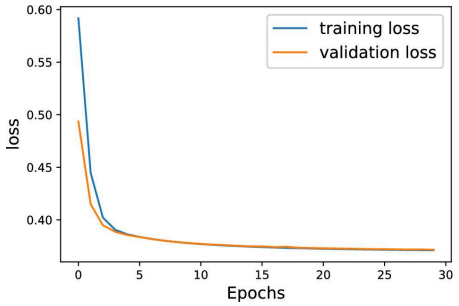

Fig. 9. The loss curve the classification model

In the confusion matrix in Fig. 10, we can see a clear diagonal line, which indicates that the label was correctly predicted. The original accuracy was around 30%. After we used the size of the bounding boxes to exclude less important labels, the accuracy went up to 54.7%. We think that was a nice approach.

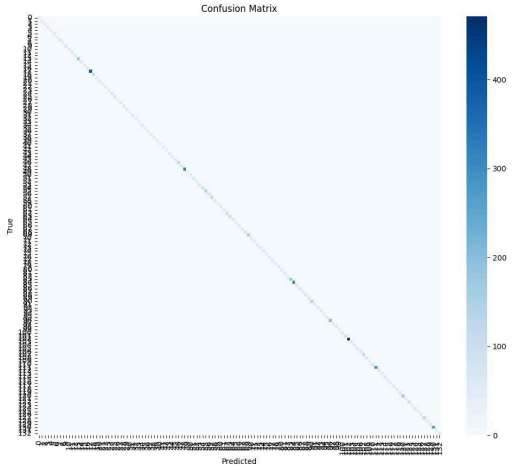

Fig. 10. The confusion matrix of the classification modelSD

2. Image reconstruction model

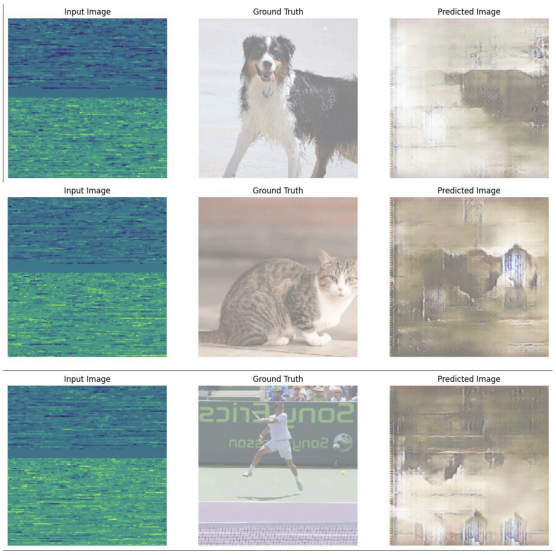

Fig. 11. Image reconstruction from the GAN model

Fig. 11 shows the reconstructed image results from our model. As shown in Fig. 11, there is a low-level similarity and a slight similar in color in some parts. This proves the performance of the L1 loss function. However, Fig. 11 demonstrates the low accuracy on the image reconstruction model.

TABLE 1

Image Reconstruction Model

Method	PixCorr	SSIM
Pix2Pix GAN	0.010	0.012
1D CNN	-0.689	-0.551
2D CNN	-0.675	-0.609
Scotti et al. [27]	**0.210**	0.265
Ozcelik et al. [28]	0.140	0.288
Gu et al. [29]	0.194	**0.304**

Fig. 12. Image reconstruction from the GAN model

Table 2 shows quantitative comparison between our model to previous models where bold indicates the best performance among them. PixCorr evaluation metric depicts the pixel-wise correlation of the original image and the reconstructed image whereas SSIM measures the spatial similarity such as textures, color, and pattern that are more meaningful than pixel-wise pattern. Previously, we also tried to build the model using 1D CNN with encoder-decoder architecture. However, the result is not satisfying. Then, we came up with cGAN model after trying to add the dimension to 2D CNN but to find no noticeable improvement in the PixCorr and SSIM score.

Our image reconstruction model, using Pix2Pix GAN or cGAN, managed to reach 0.010 in PixCorr and 0.012 for the SSIM displayed Table 2. However, according the PixCorr metric still behind the previous work's model, MindEye in [27] with 0.210 score and the SSIM metric's result below the work [29], Brain-Diffuser, with score of 0.304.

V. DISCUSSION AND CONCLUSION

The exploration of the relationship between visual stimuli and brain responses, particularly in reconstructing images from fMRI data, is confronted with several challenges. The structuring of datasets poses an issue, often lacking the granularity and limited resolution required for in-depth analysis. Additionally, the curation of datasets plays a crucial role, influencing the effectiveness of decoding models. Loading numerous samples into memory raises RAM-related problems, necessitating efficient memory management for comprehensive analyses.

Transforming data into meaningful brain activity images is a complex task requiring careful consideration of preprocessing techniques. The identification and tuning of hyperparameters present challenges, impacting the accuracy of image reconstruction. Memory issues further compound these challenges, demanding innovative solutions for extensive analyses.

To address these issues, improvements can be made by refining the selection of Region of Interest (ROI) masks, allowing a more focused analysis. Incorporating detailed features into training methodologies enhances the accuracy of image reconstruction. Selecting a larger picture of the image, rather than fixating on details, acknowledges the holistic nature of visual perception, potentially leading to more comprehensive and accurate reconstructions.

In summary, the exploration of the link between visual stimuli and brain responses through fMRI-based image reconstruction faces challenges such as hyperparameter selection, memory issues, and the refinement of ROI masks. Future researchers should focus on resolving these challenges to enhance the accuracy of image reconstruction. Improving training methodologies, incorporating detailed features, and adopting a broader perspective in image selection are crucial for advancing our understanding. Additionally, considering alternative models like conditional GANs or stable diffusion provides exciting possibilities for further refinement in reconstructing images from fMRI data, paving the way for deeper insights into the intricate relationship between visual stimuli and brain responses.

VI. REPOSITORY LINK

Our repository can be found in the following github link: https://github.com/A-Ariyanuchitkul/G20_ML2023/tree/main

VII. Author Contribution Statements

Melody (20%):

- Preprocessing: Applying roi masks
- Model: Code the image reconstruction model

Mias (20%):

- Preprocessing: Visualizing dataset / flatten,
- Model: Construct the classification model

Michelle (18%):

- Report: Writing report / presentation slides and present final project

Owen (18%):

- Report: Writing report / presentation slides and present final project

Arithat (15%):

- Preprocessing: Organize the data and calculating B-box of the labels
- Report: Repository management

Icarus (9%):

- Preprocessing: Organize the data

VIII. References

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," Berkeley AI Research (BAIR) Laboratory, UC Berkeley, 2018.

[2] H. Huang, P. S. Yu, and C. Wang, "An Introduction to Image Synthesis with Generative Adversarial Nets," arXiv:1803.04469v2, 2018.

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," arXiv:1406.2661v1, 2014.

[4] H. Madokoro et al., "Semantic Segmentation of Agricultural Images Based on Style Transfer Using Conditional and Unconditional Generative Adversarial Networks," *Appl. Sci.*, vol. 12, no. 15, p. 7785, 2022. [Online]. Available: https://doi.org/10.3390/app12157785

[5] Z. Cai, C. Peng, and S. Du, "Jitter: Random Jittering Loss Function," arXiv:2106.13749v2 , 2021.

[6] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. ECCV, 2016.

[7] A. A. Efros and T. K. Leung. Texture synthesis by nonparametric sampling. In ICCV, 1999.

[8] E. Schonfeld, B. Schiele, and A. Khoreva, "A U-Net Based Discriminator for Generative Adversarial Networks," 2020.

[9] Yunfeng Lin, Jiangbei Li, and Hanjing Wang, "DCNN-GAN: Reconstructing Realistic Image from fMRI," 2019.

[10] Zarina Rakhimberdina1, Quentin Jodelet1, Xin Liu, and Tsuyoshi Murata, "Natural Image Reconstruction From fMRI Using Deep Learning: A Survey," 2021.

[11] Guohua Shen1, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani, "End-to-End Deep Image Reconstruction From Human Brain Activity, " 2019.

[12] Kai Qiao, Jian Chen, Linyuan Wang, Chi Zhang, Li Tong, and Bin Yan, "BigGAN-based Bayesian Reconstruction of Natural Images from Human Brain Activity, " 2020.

[13] Thomas Naselaris, R. Prenger, Kendrick Norris Kay, M. Oliver, and J. Gallant, "Bayesian Reconstruction of Natural Images from Human Brain Activity," 2009.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," 2018.

[15] Nilearn, "https://nilearn.github.io/stable/index.html"

[16] Richard Felder, "Learning and teaching styles in engineering education," 1980.

[17] Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. "Distributed and overlapping representations of faces and objects in ventral temporal cortex" Science 293, 2425–2430. doi: 10.1126/science.10 63736, 2001.

[18] Kamitani, Y., and Tong, F. "Decoding the visual and subjective contents of the human brain." Nat. Neurosci. 8, 679–685. doi: 10.1038/nn1444, 2005.

[19] Horikawa, T., and Kamitani, Y. "Generic decoding of seen and imagined objects using hierarchical visual features." Nat. Commun. 8:15037. doi: 10.1038/ncomms15037, 2017.

[20] Poldrack, R. A., and Farah, M. J. "Progress and challenges in probing the human brain." Nature 526, 371–379. doi: 10.1038/nature15692, 2015.

[21] Ogawa, S., Lee, T. M., Kay, A. R., and Tank, D. W. "Brain magnetic resonance imaging with contrast dependent on blood oxygenation." Proc. Natl. Acad. Sci. U.S.A. 87, 9868–9872. doi: 10.1073/pnas.87.24.9868, 1990.

[22] Bandettini, P. A. "Twenty years of functional MRI: the science and the stories." Neuroimage 62, 575–588. doi: 10.1016/j.neuroimage.2012.04.026, 2012.

[23] Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. "Encoding and decoding in fMRI." Neuroimage 56, 400–410. doi: 10.1016/j.neuroimage.2010.07.073, 2011.

[24] Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. "Identifying natural images from human brain activity." Nature 452, 352–355. doi: 10.1038/nature06713, 2008.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár & C. Lawrence Zitnick "Microsoft COCO: Common Objects in Context. ", 2014

[26] Allen, E.J., St-Yves, G., Wu, Y. et al. "A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence." Nat Neurosci 25, 116–126 , 2022.

[27] P. S. Scotti et al., "Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors," arXiv:2305.18274v2 [cs.CV], 2023.

[28] F. Ozcelik and R. VanRullen, "Brain-Diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion," arXiv:2303.05334v2, 2023.

[29] Z. Gu, K. Jamison, A. Kuceyeski, and M. Sabuncu, "Decoding natural image stimuli from fMRI data with a surface-based convolutional network," arXiv:2212.02409v2, 2023.

Acknowledgment

We would like to express our gratitude to Professor Kuo Po Chih for assigning this project and providing a valuable learning experience. Moreover, we extend our gratitude for guiding and assisting throughout this work to Teaching Assistant Li Jia Yun.