Brain-Supervised Image Editing

Keith M. Davis III[1], Carlos de la Torre-Ortiz[1], Tuukka Ruotsalo[1,2]
first.last@helsinki.fi
[1]University of Helsinki, Helsinki, Finland
[2]University of Copenhagen, Copenhagen, Denmark

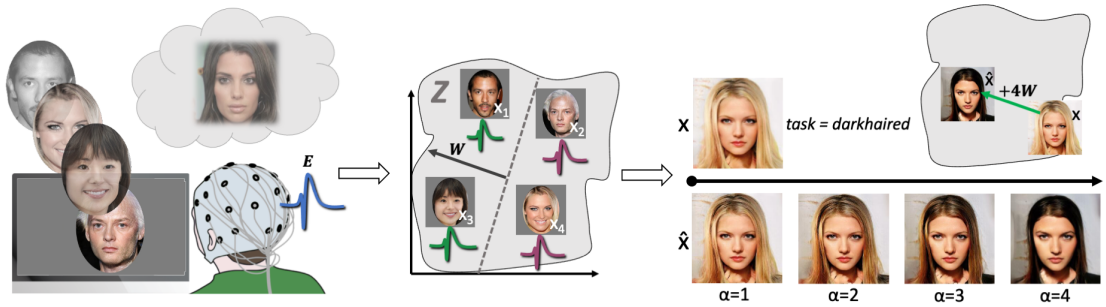


Figure 1: Brain signals are captured via EEG to supervise a semantic editing task. An individual is shown images that have an associated vector representation while they look for a semantic feature of interest (such as dark hair) and their brain responses are recorded. After a model is trained to detect semantic saliency from these brain responses, the classified brain responses and the associated image vector representations are used to model features of the latent space that correlate with semantic salience. Semantic features of a new source image can then be edited using this learned feature representation.

Abstract

Despite recent advances in deep neural models for semantic image editing, present approaches are dependent on explicit human input. Previous work assumes the availability of manually curated datasets for supervised learning, while for unsupervised approaches the human inspection of discovered components is required to identify those which modify worthwhile semantic features. Here, we present a novel alternative: the utilization of brain responses as a supervision signal for learning semantic feature representations. Participants (N=30) in a neurophysiological experiment were shown artificially generated faces and instructed to look for a particular semantic feature, such as "old" or "smiling", while their brain responses were recorded via electroencephalography (EEG). Using supervision signals inferred from these responses, semantic features within the latent space of a generative adversarial network (GAN) were learned and then used to edit semantic features of new images. We show that implicit brain supervision achieves comparable semantic image editing performance to explicit manual labeling. This work demonstrates the feasibility of utilizing implicit human reactions recorded via brain-computer interfaces for semantic image editing and interpretation.

1. Introduction

Semantic editing of images has recently become possible by utilizing models that allow for the smooth manipulation of image representations. However, semantic editing requires real-world conceptualizations of semantic information to be captured by the underlying models to achieve convincing results. Due to their high performance in modeling highly complex features, the most popular techniques involve various approaches built on generative neural networks [28, 2, 11, 42, 41, 27, 32, 16, 7], although other neural architectures [4, 29] have also shown promise.

Recent work has demonstrated that Generative Adversarial Networks (GANs) [15] encode human-interpretable representations of semantic concepts [14, 42, 16, 7, 41], which partially explains their performance in semantic editing tasks. However, GANs suffer from a lack of direct interpretability of their latent representations and do not directly allow accurate semantic control. That is, semantic representations are encoded in a continuous space, but due to their high-dimensional and multivariate representation, mapping from latent features to salient semantic image features is non-trivial. Because of this, identifying and translating learned semantic representations into a usable form remains an unsolved problem.

Supervised approaches, such as conditional GANs, *do*

allow specific features to be controlled, but they do so using extensive manual labor, as they require appropriately labeled data to be available during model training. These approaches are also influenced by the subjective opinions of the labelers themselves. As GANs are typically trained using anywhere from thousands to millions of examples, the crowdsourced labeling of such datasets for specific semantic features to match the personal interests of an individual is unrealistic.

Unsupervised approaches typically involve identifying components within the latent GAN space [7, 16]. Human assessment is then required to filter through these discovered components to determine what is and what is not semantically relevant [7]. While such approaches allow for discovery and control of semantic features, it is by no means guaranteed to find features that are highly subjective or personal, such as faces that an individual finds attractive or scenery that evokes particular emotions, moods, or memories.

Supervised or unsupervised, all methods of semantic editing and how they are assessed are fundamentally informed by the natural human ability to assess semantic relevance and saliency. In other words, they need human judgment of what semantic information is present and how noticeable it is. However, the present approaches are fairly limited as they require manual human involvement to perform. While *replacing* human judgments is not advisable, how these judgments are collected could be significantly improved.

Here, we propose a novel alternative: brain-supervised semantic editing. By obtaining human judgments from natural, immediate responses recorded from the brain while an individual perceives visual stimuli, we demonstrate that it is possible to model semantic features of the latent space using implicit feedback from the brain. Unlike conventional supervised methods, the brain-supervised approach acquires relevant labeling information much more rapidly, does not require labels to be available at the time of training the GAN models, and is not limited to features discoverable by exploratory methods such as those used in unsupervised approaches.

In detail, we ask the following research questions:

RQ1: Can brain responses be used as supervision signals for semantic image editing?

RQ2: How does brain-supervised semantic editing perform compared with editing informed by manual labels?

We show that semantically meaningful decision boundaries within the latent GAN space can be learned using implicit feedback from the brain and that transformations using these decision boundaries offer similar performance as those produced by decision boundaries trained from explicit manually provided labels. More generally, we demonstrate an intriguing new paradigm: utilizing the natural human ability to detect and assess salient semantic information within images using signals recorded directly from the brain. This offers a new methodology for semantic image understanding and processing.

2. Background

2.1. Semantic image editing

The current state of the art for semantic editing features a variety of approaches and techniques that achieve impressive results, such as image in-painting [21], style transformation [23, 44], and the disentangled modification of semantic features [2, 39, 43].

While the methods and techniques can vary significantly in their implementation, how semantic features are learned, discovered, and changed typically involves one or more of the following: aggregating large labeled datasets, manually inspecting the results of exploratory techniques, and/or providing some example with the salient semantic feature of interest to the model. For example, in [32], an image classifier was trained through transfer learning to detect manually labeled facial features that are included in the CelebA dataset[1]. This classifier was then used to automatically label hundreds of thousands of randomly generated images from a GAN, also trained from the CelebA dataset. Images with labels in the top 10% of model confidence were then used to learn decision boundaries within the GAN latent space, which in turn was used to make changes to semantic features.

Alternative methods have involved specially designing the GAN architecture to facilitate modification [24, 39], while others involve utilizing various mathematical techniques such as principal component analysis [16] to identify important dimensions within the latent space or even building customized models that can learn to manipulate the latent space, such as in [41]. Exemplar-based methods, ones in which an image containing the feature of interest is used as input to modify some pre-existing image, have also been demonstrated [40, 38, 26, 31].

The methods in which the desired transformation of the source image are conducted can vary from sliders and checkboxes that add or remove binary features (clouds, no clouds; glasses, no glasses), to categorical or multi-dimensional attributes (hair color, breed of dog) [42], addition and subtraction of scene objects using free-form drawn inputs [27] and semantic segmentation masks [3], as well as modifications using speech or text inputs [11]. While they

[1] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

may vary in form, ultimately these inputs stem from direct human interaction with the system.

2.2. Brain-computer interfacing

Brain-computer interfacing is an interaction paradigm where brain activity is used to control software or a mechanical system. Typically, brain activity is measured using a portable, wearable device located on top of or around a user's head. One of the most popular ways to monitor brain activity is electroencephalography (EEG), which is a non-invasive way to measure, at the surface of the scalp, differences in electrical potential produced by the brain. These data in turn can be used to model brain states and user reactions in real-time.

Historically, many of the applications for brain-computer interfaces (BCIs) have involved replacements to existing interaction paradigms, such as controlling a mouse [37, 45] or keyboard [5]. However, recent work has shown that brain-computer interfacing can be applied to a variety of other areas, such as information retrieval [13], content recommendation [9], graded relevance detection [30], cognitive load estimation [1], and even crowdsourcing tasks [8]. Preliminary work has also demonstrated the possibility of combining BCIs with GANs and other neural architectures [18, 10] and generating images that match personal preferences [34].

Owing to the nature of EEG measurements being taken at the surface of the scalp, rather than inside the cranial cavity, the data collected are noisy and with poor spatial resolution. Thus, while EEG data alone may not be of sufficient quality to monitor very specific cognitive processes, they nonetheless have a high temporal resolution containing features, such as the event-related potential (ERP), that render EEG suitable for use in real-time interaction.

ERPs are changes in voltage produced by brain activity in response to an event, such as viewing an image. The time-locked nature of ERPs renders them particularly useful in BCI applications, as it allows for the relatively easy association between a brain response and a digital event. ERPs may consist of various components, which are identified by their polarity (positive or negative) and time relative to the event. For example, the N200 is a negativity that occurs approximately 200 ms after viewing a face, while the P300 is a positivity occurring around 300 ms after being exposed to a stimulus recognized as relevant or otherwise important to a current task [17].

While there have been previous attempts to pair BCIs with generative models [20, 33, 36], there is growing concern that the results are not from cognitive effects, but from confounds introduced by the block structure of the experimental setup [22]. As a consequence, utilizing brain responses to guide generative models in a computer vision context remains an unsolved problem. In [20, 33, 36], placing target stimuli at the end of an experimental block, rather than randomly throughout the block, produces an artificial positive classification due to the natural temporal properties of EEG. That is, signals collected at the end of the experimental block can be distinguished from signals collected at the beginning of the experimental block, regardless of the content of the stimuli that produced these signals.

In our experiment, we carefully control for the temporal variation of EEG signals using a randomized "oddball" paradigm [35]. In our experimental design, we use complete randomization of both target and non-target stimuli classes within the same experimental block.

3. Neurophysiological Experiment

In this section, we provide a full description of how the neurophysiological experiment was performed. In detail, we describe the participants, the stimuli, the experimental apparatus, the procedure for collecting the EEG data, and how this data was processed and cleaned after collection.

3.1. Participants

Neurophysiological data were collected from thirty-one participants, recruited from the University of Helsinki and Aalto University. The nature and purpose of the experiment were explained to all participants and each participant signed a statement of informed consent to acknowledge understanding of their rights under the Declaration of Helsinki. One participant chose to end the experiment early, and so complete data were obtained for 30 participants, 13 of which self-reported as female and 17 as male, all with normal or corrected-to-normal vision and without any known history of neurological disease. The mean age of the participants was 28 years (SD = 7.14, Min = 18, Max = 45). All participants, regardless of whether or not they completed the full experiment, received compensation for their participation in the form of vouchers to the local cinema.

3.2. Stimuli

Stimuli were generated using a pre-trained GAN architecture[2] using a random process by sampling from 70,000 latent vectors drawn from a 512-dimensional multivariate normal distribution [19]. A human assessor, who did not participate in the neurophysiological experiment, manually screened all stimuli to ensure they appeared human and did not contain unrealistic artifacts. These images and the associated 512-dimensional latent vectors used to produce them were then sorted into one of eight groups based on the following visual features: smiling, not smiling, female, male, young, old, dark hair, and light hair (blond). An elliptic grey frame was applied to all images to mask the background and non-facial features.

[2] https : / / github . com / tkarras / progressive _ growing_of_gans

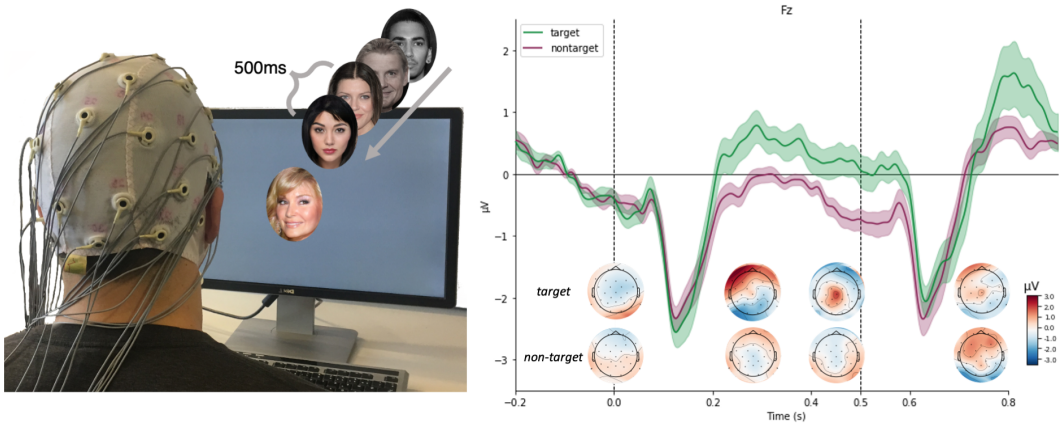


Figure 2: The experimental setup and the visualization of the RSVP task are depicted left. A participant is instructed to look for a semantic feature, such as dark hair, and is presented with a new stimulus every 500 ms. On the right, we plot the average brain response, measured at the Fz electrode, produced by viewing the same stimuli during different tasks. For "target", the responses are from viewing images of dark-haired people during task *dark-haired*. For "non-target" the responses are from viewing images of dark-haired people during task *blond*. Thus, while the stimuli remain the same, the evoked responses from the brain are different, as the nature of a brain response depends on whether or not a given stimulus contained a salient semantic feature of interest.

3.3. Apparatus

The EEG data were recorded using 32 Ag/AgCl electrodes, arranged according to the 10–20 system (exact positions can be found in the supplementary material). A QuickAmp USB (BrainProducts GmbH, Gilching, Germany) amplifier running at 2,000 Hz was used for amplification, filtering, and digitization of the signal. Eye movements were detected using two pairs of bipolar electrodes for artifact detection — one situated 1 cm to the lateral canthi of the left and right eye, and the other 2 cm above and below the right pupil.

3.4. Experimental Procedure

Participants were presented with eight saliency recognition tasks, each task corresponding to a predetermined visual feature of interest as described in section 3.2. Before each iteration, participants completed a demonstration task. For the demonstration task, they were shown four example stimuli images and were asked to manually select the images that contained the semantic feature of interest for the proceeding saliency recognition task. These images were not used as stimuli in the actual task. All stimuli presented during each task were assigned a binary label based on semantic feature saliency. For example, during the task "smile", participants were shown faces that were either smiling (labeled as target) or not smiling (non-target). Participants were instructed to only observe the presented faces and make a mental note whenever they saw a face that matched the task description (smile, target). No other physical inputs were required from the participants during the saliency recognition task. Twenty stimuli of the target class and fifty stimuli of the non-target class were shown in random order during each iteration of the task. Stimuli were presented in rapid serial visual presentation (RSVP) format at a rate of one every 500 ms.

To ensure enough data were collected for each participant and the participant saw each image at least once, the saliency recognition task and demonstration task were conducted a total of four iterations for each image category, for a total of 32 iterations.

3.5. Data Preprocessing

EEG measurements generally contain unwanted artifacts and noise originating from a variety of sources, such as movements of the participant and other electrical equipment. Standard signal cleaning procedures were employed [25] to improve the signal-to-noise ratio.

The preprocessing step was designed to reduce signal noise for real-time applications, so only automated operations which can be done in real-time were used for signal cleaning. To remove slow signal fluctuations caused by respiration and high-frequency background noise produced by electrical equipment, a band-pass filter in the frequency range 0.2–35 Hz was applied to the signal.

After filtering, the data were split into time-locked bands (*epochs*) ranging from -200 to 900 ms relative to stimulus onset. Baseline correction was performed for each epoch based on a pre-stimulus period of -200 to 0 ms. Epochs containing large amounts of transient artifacts, such as those caused by eye blinks, were removed using a threshold-based heuristic. After pre-processing, approximately 11% of each

participants' epochs were removed, with an average of 2239 epochs per participant remaining.

4. Brain-Supervised Semantic Editing

Here, we explain the steps to our method, as visualized in Figure 1.

Preliminaries We assume a generative function $G(\mathbf{z}) \rightarrow \mathbf{x}$ with a latent space $\mathbb{Z}$, where any given vector $\mathbf{z}$ can be transformed into an image representation $\mathbf{x}$ and vice-versa. The recognition tasks described above in section 3.4 yield a tensor representation of the brain signals $\mathbf{E}$ for each generated image and its associated vector representation. When a participant is tasked with identifying images containing some semantic feature, viewing an image produces a response $\mathbf{E}$ containing information related to the saliency s of the target semantic feature.

First, we define a function $SAL(\mathbf{E}) \rightarrow s$, such that using only the brain response e as input, the semantic saliency $\mathbf{E}$ of some image $\mathbf{x}$ can be estimated. Next, we define a function $SEM(Z, S) \rightarrow \mathbf{W}$. Given a set of image vectors Z and associated set of semantic saliency scores S, this function identifies a set of feature transformations $\mathbf{W}$ which, when applied to any given $\mathbf{z}$, correspond to change in s for the image form x. Finally, we use the learned transformations $\mathbf{W}$, scaled by a positive or negative constant α, to modify a given representation $\mathbf{z}$. The result is a transformed vector $\hat{\mathbf{z}}$ with an image form $\hat{x}$ where the saliency score s for a specific semantic has changed while remaining disentangled from other semantic features (i.e, other features are left unchanged). The magnitude of the change is proportional to α. This function can be written $EDIT(z, \mathbf{W}, \alpha) \rightarrow \hat{z}$.

Saliency estimation from brain signals To construct the saliency estimation function SAL, we used a regularized Linear Discriminant Analysis (LDA) classifier [6]. The classifier was trained with vectorized representations of the brain responses using a binary label (target or non-target) indicating if the semantic feature of interest was salient within the associated stimulus image. Brain responses were vectorized by taking time-series voltage data from all 32 channels for a given epoch and concatenating it into a single array. Using leave-one-out cross-validation, the semantic saliency score s for each stimulus image x was estimated from an individual's associated brain responses.

Mapping semantic saliency within the latent space To identify the feature transformations $\mathbf{W}$ within the latent GAN space $\mathbf{Z}$, we made use of a support vector regression (SVR) model with a linear kernel [12]. Here, we wish to satisfy the equation $0 = \mathbf{W}^\top \mathbf{Z} + b$. That is, to find the hyperplane $\mathbf{W}^\top$ for some set of points Z along which the semantic saliency s for a given z and its image form x is equal to zero.

Given a set of vector representations Z and associated saliency scores S, the SVR model was trained for each semantic category. SVR has the convenient property that, upon learning to estimate s given a vector z, the unit normal vector $\mathbf{W}$ of the resulting hyperplane is equivalent to $\mathbf{W}$ from function $EDIT$.

Performing semantic transformations on an image Provided a representation z of an image x, the saliency s for a given semantic can be changed using $\hat{z} = z + \alpha\mathbf{W}$ while remaining disentangled from other semantic features. When $\alpha > 0$, saliency s will increase for the target semantic, and when $\alpha < 0$, s will decrease.

5. Semantic Editing Experiment

After collecting the necessary neurophysiological data, we conducted a modeling experiment using the recorded brain responses to perform semantic transformations on new images. In this section, we provide a detailed explanation of how controls were constructed to quantify the performance of the brain-supervised semantic editing procedure. We then describe the evaluation procedure for quantifying the performance of our method.

5.1. Control Conditions

To assess the performance of the brain-based model, three control conditions were selected. For these controls, all factors remained equal *except for what signals were used to estimate* $\mathbf{W}$.

For the first control, $\mathbf{W}$ was found through an SVR model trained using explicit labels manually assigned to the stimuli for the neurophysiological experiment. This control condition, referred to as the *explicit* model, was selected to compare performance between the brain-labeled model and a model trained with explicitly defined labels.

The second control used randomly shuffled permutations of the brain-derived labels to find $\mathbf{W}$. This was done to gauge how important label accuracy is for producing a good hyperplane for transformations. We refer to this control as the *random label* model.

The third control was to create $\mathbf{W}$ by sampling from a multivariate normal distribution (512 dimensions) rather than finding it through SVR. This control condition, called *random vector*, was selected to determine a lower bound for how much an image can be transformed to match a target label by simply moving in a random direction within the sample space.

5.2. Evaluation

Experimental data were anonymized shortly after collection. Due to this and the timelines of the experiments, it

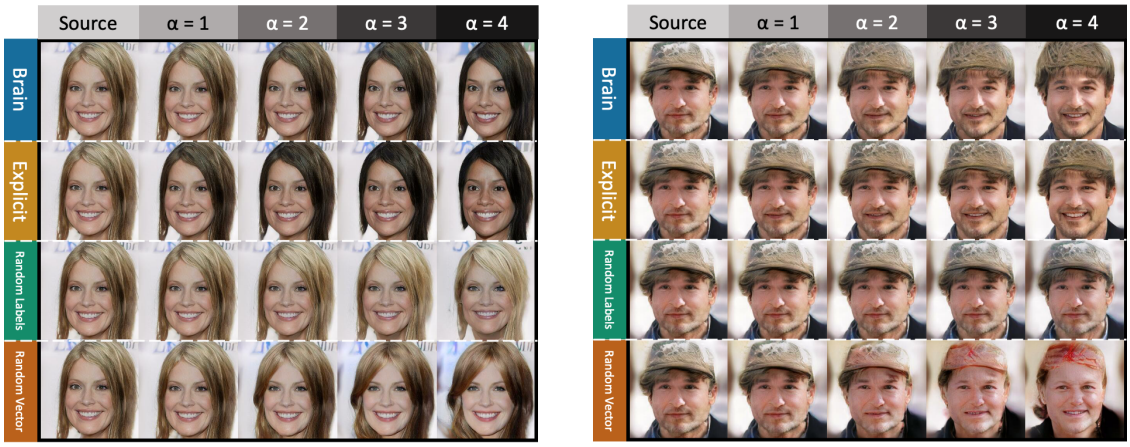


Figure 3: Sample results of the transformations performed by the Brain model and three controls for tasks *dark hair* (left) and *smile* (right), given the same source image where the semantic feature of interest is not salient. Recall that *explicit* refers to transformations made using a model trained from manual labels, while *brain* refers to transformations made using a model trained from classified brain responses.

was not feasible to have the original participants evaluate the results of the transformations. To account for this, two independent evaluators who had not participated in the neurophysiological experiment were recruited to assess the results of the transformations. One evaluator performed the primary evaluation while the second was used to calculate the kappa statistic for inter-rater reliability.

The generated outputs were evaluated in a randomized blind user study completed by these evaluators making use of a graphical user interface. Given a source image, four edited images were produced by the Brain model and three control conditions, for a total of four image sets per source image. Evaluators were presented with one image set at a time in random order (thus the same source image was not evaluated multiple times in a row). Each image set consisted of the source image and four images transformed using a multiple ($\alpha = 1, 2, 3, 4$) of $\mathbf{W}$. The ordering of these images was also randomized.

Evaluators were asked for a numeric rating in a five-level Likert scale the degree to which each image matched a target visual feature, with 0 indicating *no match* with the label and 4 indicating *complete match*. Additionally, two other metrics were collected for each image: realism, and identity preservation. The evaluator was asked to provide a binary rating for each image as to whether or not it appeared realistic, and another binary rating if it appeared to depict the same person as in the other images.

15 randomly selected source images were transformed by the Brain model and 3 control conditions for each of the 8 semantic saliency tasks. In total, the primary evaluator annotated 9,600 images, while the second evaluator evaluated 512 of the same generated images to estimate the reliability of the annotation process.

6. Results

In this section, we provide the evaluation results of the semantic editing experiment. We also give a brief overview of the neurophysiological findings.

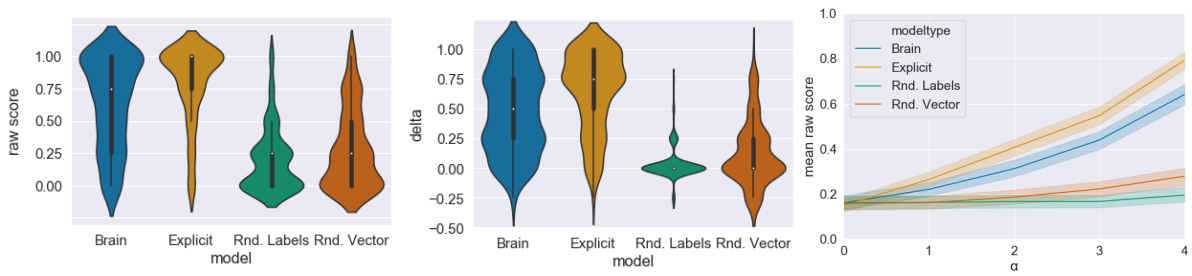
6.1. Neurophysiological Experiment

Shown in Figure 2 is an ERP plot for average evoked responses at the Fz electrode to smiling images for two conditional tasks: task "smiling" and task "not smiling". The P300 effect is clearly shown and confirms that although participants were seeing the same images, their neurophysiological responses to the images depended on the task. Thus, a smiling face during task "smiling" will produce a large positivity at the Fz electrode as it matches the task. Smiling faces during the "not smiling" task, however, do not produce this positivity, as they do not match the target description.

For the classification of brain responses across all participants and all semantic saliency tasks, the mean F1 score was 0.67 (Min=0.54, Max=0.87, SD=0.12), which is consistent with the performance typically expected of BCIs using similar equipment, data preprocessing, and classification techniques [6].

6.2. Semantic Editing Experiment

The results of the semantic editing experiment show that the Brain model performed similarly to the Explicit model. The Brain and Explicit models consistently produced images where the salience of semantic features was appropriately changed without significantly altering other visual features while the random controls did not. Sample results can be found in Figure 3 and Figure 5. Both the brain model



(a) (b) (c)

Figure 4: Violin plots of semantic editing performance for the brain model and controls. In Figure 4a, the rating value for the final step across all tasks is shown for each model. In Figure 4b, the difference between rating values for the final step and the source image across all tasks are shown for each model. Both brain and explicit models performed significantly better than the random controls. Comparing the brain and explicit model results, we see from Figure 4c slight differences in performance between the brain and explicit models as α increases.

Task	Raw Ratings				Deltas				Realism				Identity Preservation			
	Brain	Explicit	R1	R2	Brain	Explicit	R1	R2	Brain	Explicit	R1	R2	Brain	Explicit	R1	R2
Blond	0.74	0.92	0.16	0.25	0.65	0.81	0.04	0.14	0.90	0.90	0.93	0.63	0.97	0.97	1.00	0.85
Female	0.68	0.84	0.13	0.18	0.62	0.79	0.06	0.13	0.80	0.87	0.93	0.80	0.98	0.91	1.00	0.94
Young	0.80	0.86	0.47	0.50	0.30	0.43	0.04	0.10	0.73	0.70	0.83	0.70	0.98	0.98	0.99	0.88
Smiling	0.54	0.73	0.12	0.93	0,23	0.64	0.03	0.13	0.87	0.97	0.87	0.63	0.99	1.00	1.00	0.83
Dark-haired	0.65	0.93	0.06	0.10	0.65	0.93	0.04	0.08	0.80	0.87	0.80	0.67	0.95	0.95	1.00	0.85
Male	0.77	0.90	0.22	0.32	0.62	0.76	0.05	0.16	0.73	0.80	0.90	0.73	0.99	0.98	1.00	0.90
Old	0.26	0.34	0.09	0.18	0.16	0.27	0.01	0.05	0.67	0.47	0.63	0.67	0.93	0.92	0.99	0.82
Not smiling	0.69	0.83	0.33	0.48	0.36	0.53	0.01	0.18	0.83	0.80	0.80	0.73	0.99	0.99	1.00	0.86
Mean	*0.64*	*0.79*	*0.19*	*0.27*	*0.48*	*0.65*	*0.04*	*0.12*	*0.79*	*0.80*	*0.84*	*0.70*	*0.97*	*0.96*	*1.00*	*0.87*

Table 1: Results for all measures, with non-target starting image, shown for the four models - brain, explicit, random labels (R1), and random vector (R2). All measures are between 0 and 1, with 0 indicating the worst performance and 1 indicating the best possible performance. Across all tasks, both models performed significantly better than random controls ($p < 0.001$) (Bonferroni corrected) based on the deltas. For *realism* and *identity preservation*, the performance of R1 serves as an upper bound as the target images produced by the R1 baseline were more or less indistinguishable from the source image.

and explicit control model performed better than the random controls across all metrics: final scores, deltas, identity preservation, and realism. However, comparing the distribution of both deltas and raw scores for I_f, the Brain model has a wider distribution, as shown in Figure 4. Additionally, each step of the Brain model introduces a smaller change than the Explicit model. The full results of the evaluation are shown in Table 1.

Comparing the Brain model and the Explicit model, across all tasks, differences in performance between the Explicit model and the Brain model were statistically significant (two-sided t-test, Bonferroni corrected $p < 0.05$). However, comparing between tasks, differences in performance were only significant for task dark-haired (Bonferroni corrected $p < 0.01$).

For identity preservation and realism, both the Brain and Explicit models performed similarly, with no significant differences found between the two. The R1 model, which did not change the source images much, maintained the identity of the source images well. The Brain and Explicit models produced images that were significantly more realistic than the random vector model, although the effect size was small (Bonferroni corrected $p < 0.01$). For the preservation of identity, no significant differences were found between the Brain and Explicit models. Both models performed significantly better than the random vector model (Bonferroni corrected $p < 0.0001$). Between the two evaluators, Cohen's kappa was 0.88 for the performance of semantic editing, 1.00 for preserving identity, and 0.99 for synthetic image realism. Therefore, all evaluated metrics present a high inter-rater agreement.

7. Discussion and Conclusions

In this work, we sought answers for the following research questions:

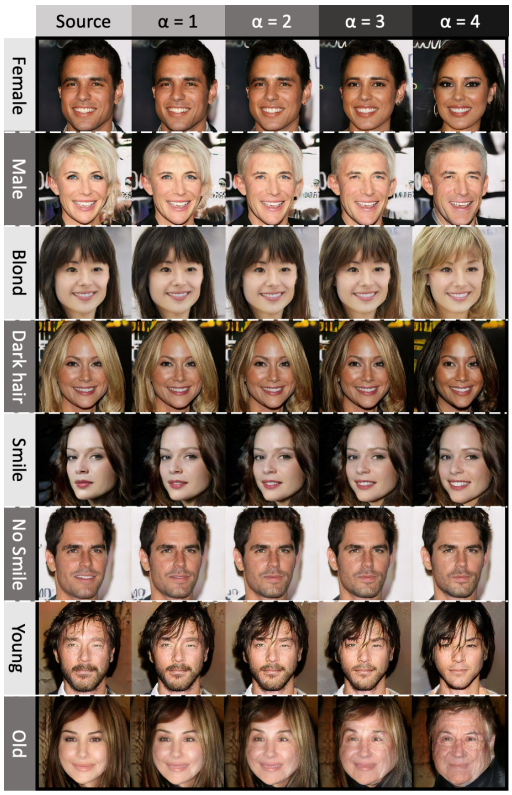


Figure 5: Sample results of the transformations performed by the Brain model for each task. The name of the task indicates the target output description. For example, for task *young*, the goal is to produce an image that looks *younger*. Similarly, for task *dark hair*, the goal is to produce an image with *darker* hair.

RQ1: Can brain responses be used as supervision signals for semantic image editing? We have shown that brain responses can be used to detect the saliency of semantic features of interest within images, and these data are of sufficient quality to supervise the semantic editing of images.

RQ2: How does brain-supervised semantic editing perform compared with editing informed by manual labels? While the explicit model performed better than the brain-based model, the differences in performance are small enough to warrant further investigation in brain-based methods.

While the implementation described here involves learning the latent space over a pre-trained GAN, we believe that brain-supervision may generalize to many other methods for GAN control and beyond. The same P300 relevance effect could be utilized to rapidly explore and select latent space manipulations and features produced by unsupervised transformation techniques. It may also be possible to incorporate brain signals as auxiliary information to be used directly in training representations.

The fundamental limitation of our approach — the accuracy of the semantic saliency estimation from EEG responses — stems not from limitations of the paradigm itself but the quality of currently available sensor technology. EEG remains a relatively noisy and spatially inexact technique to capture brain activity. Thus, the typical range of 0.65 — 0.80 classification accuracy achieved in binary classification problems using EEG signals is usually outmatched by explicit labeling techniques. However, this performance gap is less significant when taking into consideration how rapidly brain responses can be recorded (2 stimuli presentations per second) relative to manual techniques. With improved sensor technology and/or better imaging techniques, it is not unreasonable to expect that brain-supervised methods will surpass their manual alternatives sometime in the near future.

As brain-imaging sensor technology continues to improve and become more affordable, the prospect that BCIs will become a common interaction paradigm becomes increasingly likely. It is therefore worthwhile to begin setting the foundation for how information from such sensor technology could be integrated into existing and future image processing methods. This would not be to simply augment or complement the performance of existing models but to fundamentally change how these models are supervised and controlled.

Here we have demonstrated, for the first time, that semantic image editing can be conducted using responses from the brain. More broadly, this work presents a novel paradigm: the incorporation of physiological feedback in supervised model training and control. This paradigm extends beyond the current supervision signals used by the computer vision and machine learning research communities by more efficiently utilizing peoples' natural abilities to detect semantic features and semantic saliency. Furthermore, using measurements directly from the brain allows for the implicit identification of semantic dimensions that may be otherwise difficult to quantify using traditional manual labeling techniques. This entails a vision in which computer vision systems can learn semantic saliency important for their users, or even semantic image representations, directly from human brain responses to visual information.

Acknowledgments

This research was partially funded by the Academy of Finland. Computing resources were provided by the Finnish Grid and Cloud Infrastructure (urn:nbn:fi:research-infras-2016072533). We thank Michiel Spapé for his contributions to the neurophysiological experimentation and advice.

References

[1] Lena M Andreessen, Peter Gerjets, Detmar Meurers, and Thorsten O Zander. Toward neuroadaptive support technologies for improving digital reading: a passive bci-based assessment of mental workload imposed by text difficulty and presentation speed during reading. *User Modeling and User-Adapted Interaction*, 31(1):75–104, 2021.

[2] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE international conference on image processing (ICIP)*, pages 2089–2093. IEEE, 2017.

[3] Aayush Bansal, Yaser Sheikh, and Deva Ramanan. Shapes and context: In-the-wild image synthesis & manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2317–2326, 2019.

[4] Apratim Bhattacharyya, Shweta Mahajan, Mario Fritz, Bernt Schiele, and Stefan Roth. Normalizing flows with multi-scale autoregressive priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2020.

[5] Benjamin Blankertz, Guido Dornhege, Matthias Krauledat, Michael Schröder, John Williamson, Roderick Murray-Smith, and Klaus-Robert Müller. The berlin brain-computer interface presents the novel mental typewriter hex-o-spell. 2006.

[6] Benjamin Blankertz, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Müller. Single-trial analysis and classification of erp components — a tutorial. *NeuroImage*, 56(2):814 – 825, 2011. Multivariate Decoding and Brain Reading.

[7] Anton Cherepkov, Andrey Voynov, and Artem Babenko. Navigating the gan parameter space for semantic image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3671–3680, 2021.

[8] Keith M Davis III, Lauri Kangassalo, Michiel Spapé, and Tuukka Ruotsalo. Brainsourcing: Crowdsourcing recognition tasks via collaborative brain-computer interfacing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[9] Keith M Davis III, Michiel Spapé, and Tuukka Ruotsalo. Collaborative filtering with preferences inferred from brain signals. In *Proceedings of the Web Conference 2021*, pages 602–611, 2021.

[10] Carlos de la Torre-Ortiz, Michiel M Spapé, Lauri Kangassalo, and Tuukka Ruotsalo. Brain relevance feedback for interactive image generation. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 1060–1070, 2020.

[11] Garoe Dorta, Sara Vicente, Neill DF Campbell, and Ivor JA Simpson. The gan that warped: Semantic attribute editing with unpaired data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5365, 2020.

[12] Harris Drucker, Chris JC Burges, Linda Kaufman, Alex Smola, Vladimir Vapnik, et al. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997.

[13] Manuel JA Eugster, Tuukka Ruotsalo, Michiel M Spapé, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. Natural brain-information interfaces: Recommending information by relevance inferred from human brain signals. *Scientific reports*, 6(1):1–10, 2016.

[14] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5744–5753, 2019.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[16] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020.

[17] James E Hoffman, Robert F Simons, and Michael R Houck. Event-related potentials during controlled and automatic target detection. *Psychophysiology*, 20(6):625–632, 1983.

[18] Lauri Kangassalo, Michiel Spapé, and Tuukka Ruotsalo. Neuroadaptive modelling for generating images matching perceptual categories. *Scientific reports*, 10(1):1–10, 2020.

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[20] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1809–1817, 2017.

[21] Avisek Lahiri, Arnav Kumar Jain, Sanskar Agrawal, Pabitra Mitra, and Prabir Kumar Biswas. Prior guided gan based semantic inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13696–13705, 2020.

[22] Ren Li, Jared S Johansen, Hamad Ahmed, Thomas V Ilyevsky, Ronnie B Wilbur, Hari M Bharadwaj, and Jeffrey Mark Siskind. The perils and pitfalls of block design for eeg classification experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):316–333, 2020.

[23] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3809–3817, 2019.

[24] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *Advances in neural information processing systems*, 29:469–477, 2016.

[25] Steven J Luck. *An introduction to the event-related potential technique*. MIT press, 2014.

[26] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. *arXiv preprint arXiv:1805.11145*, 2018.

[27] Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. Sesame: semantic editing of scenes by adding, manipulating or erasing objects. In *European Conference on Computer Vision*, pages 394–411. Springer, 2020.

[28] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[29] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14104–14113, 2020.

[30] Zuzana Pinkosova, William J McGeown, and Yashar Moshfeghi. The cortical activity of graded relevance. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–308, 2020.

[31] Baptiste Rozière, Morgane Riviere, Olivier Teytaud, Jérémy Rapin, Yann LeCun, and Camille Couprie. Inspirational adversarial image generation. *IEEE Transactions on Image Processing*, 30:4036–4045, 2021.

[32] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020.

[33] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6809–6817, 2017.

[34] Michiel Spape, Keith Davis, Lauri Kangassalo, Niklas Ravaja, Zania Sovijarvi-Spape, and Tuukka Ruotsalo. Brain-computer interface for generating personally attractive images. *IEEE Transactions on Affective Computing*, 2021.

[35] Nancy K Squires, Kenneth C Squires, and Steven A Hillyard. Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and clinical neurophysiology*, 38(4):387–401, 1975.

[36] Praveen Tirupattur, Yogesh Singh Rawat, Concetto Spampinato, and Mubarak Shah. Thoughtviz: Visualizing human thoughts using generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 950–958, 2018.

[37] Jacques J Vidal. Real-time detection of brain events in eeg. *Proceedings of the IEEE*, 65(5):633–641, 1977.

[38] Miao Wang, Guo-Ye Yang, Ruilong Li, Run-Ze Liang, Song-Hai Zhang, Peter M Hall, and Shi-Min Hu. Example-guided style-consistent image synthesis from semantic labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1495–1504, 2019.

[39] Yi Wei, Zhe Gan, Wenbo Li, Siwei Lyu, Ming-Ching Chang, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. Maggan: High-resolution face attribute editing with mask-guided generative adversarial network. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[40] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 168–184, 2018.

[41] Guoxing Yang, Nanyi Fei, Mingyu Ding, Guangzhen Liu, Zhiwu Lu, and Tao Xiang. L2m-gan: Learning to manipulate latent space semantics for facial attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2951–2960, 2021.

[42] Huiting Yang, Liangyu Chai, Qiang Wen, Shuang Zhao, Zixun Sun, and Shengfeng He. Discovering interpretable latent space directions of gans beyond binary attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12177–12185, 2021.

[43] Shuai Yang, Zhangyang Wang, Zhaowen Wang, Ning Xu, Jiaying Liu, and Zongming Guo. Controllable artistic text style transfer via shape-matching gan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4442–4451, 2019.

[44] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. Attention-aware multi-stroke style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1467–1475, 2019.

[45] Thorsten O Zander, Laurens R Krol, Niels P Birbaumer, and Klaus Gramann. Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity. *Proceedings of the National Academy of Sciences*, 113(52):14898–14903, 2016.