Usable and Fast Interactive Mental Face Reconstruction

Florian Strohm
florian.strohm@vis.uni-stuttgart.de
University of Stuttgart
Stuttgart, Baden-Württemberg
Germany

Mihai Bâce
mihai.bace@vis.uni-stuttgart.de
University of Stuttgart
Stuttgart, Baden-Württemberg
Germany

Andreas Bulling
andreas.bulling@vis.uni-stuttgart.de
University of Stuttgart
Stuttgart, Baden-Württemberg
Germany

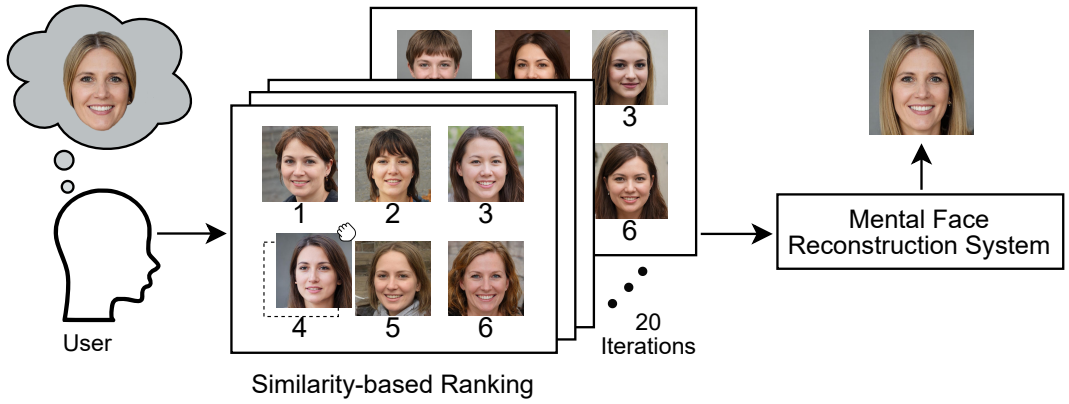


Figure 1: We propose an interactive system that is able to reconstruct face images residing only in a user's mind. Over multiple iterations, the system shows different face images to the user, which they have to rank according to the perceived similarity with their mental image. This feedback is used to extract relevant image features across iterations and combine them to visually reconstruct the user's mental image.

ABSTRACT

We introduce an end-to-end interactive system for mental face reconstruction – the challenging task of visually reconstructing a face image a person only has in their mind. In contrast to existing methods that suffer from low usability and high mental load, our approach only requires the user to rank images over multiple iterations according to the perceived similarity with their mental image. Based on these rankings, our mental face reconstruction system extracts image features in each iteration, combines them into a joint feature vector, and then uses a generative model to visually reconstruct the mental image. To avoid the need for collecting large amounts of human training data, we further propose a computational user model that can simulate human ranking behaviour using data from an online crowd-sourcing study (N=215). Results from a 12-participant user study show that our method can reconstruct mental images that are visually similar to existing approaches but has significantly higher usability, lower perceived workload, and is 40% faster. In addition, results from a third 22-participant lineup study in which we validated our reconstructions on a face ranking task show a identification rate of 55.3%, which is in line with prior work. These results represent an important step towards new interactive intelligent systems that can robustly and effortlessly reconstruct a user's mental image.

CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**; • **Human-centered computing** → *User models.*

KEYWORDS

mental image reconstruction, faces, user modelling, deep learning

ACM Reference Format:
Florian Strohm, Mihai Bâce, and Andreas Bulling. 2023. Usable and Fast Interactive Mental Face Reconstruction. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), October 29-November 1, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3586183.3606795

UIST '23, October 29-November 1, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0132-0/23/10…$15.00
https://doi.org/10.1145/3586183.3606795

1 INTRODUCTION

Visually reconstructing mental images, i.e., images of objects, faces, or concepts one only has in their mind, has held a long-time fascination and has thus attracted significant research interests. Mental image reconstruction is promising for a range of applications, such as for the creation of personalised game avatars or the reconstruction of a criminal's face from witnesses' memory. At the same time this

task is highly challenging: Mental images are encoded in the complex neural dynamics of the brain that are still only poorly understood [28]. Early work has therefore explored brain sensing modalities, such as electroencephalogram (EEG) [3, 10, 37, 47] or functional magnetic resonance imaging (fMRI) [1, 9, 16, 25, 36, 38, 42] for this task. More recently, to avoid the need for special-purpose sensing equipment that is either invasive (EEG) or expensive and not practical for everyday use (fMRI), other works have started to explore *passive* monitoring of human gaze for mental image reconstruction. While gaze is known to be a strong mediator of cognitive processes, including memory [4], prior works have only achieved limited reconstruction quality so far [34, 35, 40, 41].

In this work, we study a practically more feasible approach in which human and AI *collaborate* to jointly reconstruct a user's mental image based on *active* user feedback. Mental image reconstruction in the form of facial composite generation as studied in this work is particularly challenging as small changes to facial features can make the face be perceived as being completely different [27]. This task can be broadly grouped into constructive, holistic, and hybrid methods. Constructive composite systems maintain a large catalogue of images for each facial feature that users have to select from, such as the eyes, nose, and mouth [6, 12, 23, 24]. The key limitation of this approach is that humans struggle to correctly identify specific facial features in isolation but rather perceive and recall faces holistically [13]. Holistic composite creation tools such as EvoFIT [14] try to address this limitation by allowing users to iteratively compose entire faces through an evolutionary algorithm. However, controlling and guiding holistic image generation is more challenging compared to changing a specific feature based on a catalogue. To overcome the limitations of constructive and holistic approaches, hybrid methods (e.g. CG-GAN [46]) allow users to iteratively combine faces through interactive refinement and additionally provide constructive control over specific facial features. Despite significant advances, existing hybrid composite generation tools are limited in terms of usability as they require manual edits which are slow and tedious. Moreover, previous methods rely on random search within the high-dimensional appearance space around faces that have been selected by users for further refinement.

We address these limitations by proposing a novel interactive mental face reconstruction system (MFRS) that is specifically geared towards maximising the information gained from user feedback and that does not rely on random exploration. Our method only requires users to iteratively rank sets of face images proposed by our system with respect to the subjective visual similarity to their mental image. Using these images and the corresponding user-provided ranking as input, our system extracts facial appearance information about the mental image over multiple iterations. To integrate this information across iterations, we propose an end-to-end, data-driven model that predicts a feature vector encoding of facial features that are likely part of the mental image. To visually decode the image, our MFRS finally uses a state-of-the-art generative face model [21]. Collecting a large amount of human ranking feedback to train our method is impractical. We instead propose a computational user model to simulate human ranking behaviour. The user model consists of a pre-trained face identification network [11] that we fine-tuned on a face similarity task using human labels that we crowd-sourced from 215 users using Amazon Mechanical Turk (AMT). This approach allows us to generate arbitrary amounts of human rankings to train our reconstruction system.

The contributions of our work are three-fold: First, we propose an end-to-end trainable method for mental face reconstruction that learns to integrate explicit user rankings and facial image information across multiple iterations. Second, we present a computational user model for ranking that we trained on a novel crowd-sourced dataset of human face ranking information. This, in turn, allows us to synthesise arbitrary amounts of data to train our method. [1] Third, we report evaluations of our method on data collected from 12 users that show that our method is significantly more usable and faster than existing methods without sacrificing image reconstruction quality. Finally, a study with 22 independent participants shows that the human recognition rate of our system is on par with the current state of the art.

2 RELATED WORK

Our work is related to prior work on mental face reconstruction with 1) implicit and 2) explicit feedback.

2.1 Mental Face Reconstruction with Implicit Feedback

Related work on face reconstruction using implicit feedback has focused on EEG, fMRI and gaze. The first work for fMRI-based face reconstruction was conducted by Cowen et al. [8], where they linearly mapped fMRI responses to principal components of the face manifold called Eigenfaces, which allowed the reconstruction of faces by linearly combining them. Nestor et al. [30] inferred a face feature space from fMRI responses directly instead of precomputing Eigenfaces from images only. They trained an SVM to discriminate the face identities based on the fMRI response and subsequently calculated template faces for each axis in the discovered space using the training data. To reconstruct faces, they mapped an fMRI response into this feature space and accordingly interpolated the template faces. Later, Nemrodov et al. [29] adopted this approach to successfully reconstruct faces from EEG responses instead. VanRullen and Reddy [42] proposed an approach using a VAE-GAN architecture to encode faces into a low-dimensional latent space. A simple regression model maps fMRI responses onto the latent image space, from which the face can be reconstructed using the VAE-GAN decoder. Unlike previous work, they were also able to reconstruct images from fMRI responses elicited when participants only thought of a face, without actually seeing it. Recently, Dado et al. [9] collected fMRI responses for synthetic faces generated from a pre-trained PGGAN [19]. Similar to VanRullen and Reddy [42], they then trained a linear regression model to map fMRI responses to the latent space of this GAN, allowing the reconstruction of faces from fMRI responses.

While the discussed methods for face reconstruction from fMRI and EEG achieve promising results, their usage is highly invasive and expensive. Additionally, a main drawback of such methods is that they are hard to generalise to new participants, as these models are trained on specific subjects.

[1] Project code and collected data is available at https://perceptualui.org/publications/strohm23_uist/

Recently, Strohm et al. [40] proposed a method for gaze-based mental face reconstruction. The subjects observed a set of faces, and the so evoked gaze behaviour was used to predict which features of a face are relevant. They combined the extracted features from multiple faces and generated the mental image using a pre-trained decoder. While their method is less invasive and expensive than fMRI and EEG-based methods, it requires prior knowledge about the target. They later proposed an iterative system [41], removing the need for prior knowledge but they still could only operate in a controlled environment and with high-precision gaze data. Instead, our system operates in the less constrained domain of real faces and only requires simple user feedback.

2.2 Mental Face Reconstruction with Explicit Feedback

Given the current limitations of methods using implicit feedback, methods based on explicit feedback are still dominating. Earlier systems were solely based on the constructive face creation paradigm [6, 12, 23, 24], which allow witnesses to separately choose facial features like eyes and nose from a large template catalogue. However, the performance of such systems is limited by insights into human face perception, which show that humans perceive faces holistically and struggle to correctly recognise isolated facial features [13]. Therefore, Frowd et al. [14] proposed the EvoFIT system, that allows the holistic interpolation of faces in the Eigenface space. Users are shown a collection of faces and have to select those that they perceive to be similar to their mental image in some aspect over multiple iterations. Based on their proposed evolutionary algorithm, the EvoFIT system then generates a new collection of faces based on the Eigenfaces of the selected faces. A similar approach has been proposed by Gibson et al. [15], with the additional functionality of controlling age and adding details to the face such as wrinkles. Later, Bontrager et al. [2] proposed deep interactive evolution which deploys the evolution algorithm proposed by Frowd et al. [14] in the latent space of a pre-trained GAN instead of the Eigenface space, improving image quality. Xu et al. [45] further improved reconstruction results utilising a GAN conditioned on facial landmarks and performed a search in the landmark space. They iteratively trained an online classifier based on user relevance feedback to find the optimal landmarks. Zaltron et al. [46] proposed CG-GAN which extended the work by Bontrager et al. [2] with additional constructive functionalities. Using binary face labels such as *glasses* and *beard*, they discovered axes in the latent space of a pre-trained GAN and allowed users to change faces along these axes. Finally, Chiu et al. [5] proposed a method that allows the user to explore randomly sampled one-dimensional sub-spaces of a pre-trained GAN. Users iteratively select the best point using sliders until they cannot improve the result further.

In contrast to prior work, we propose a new method which learns to extract information from user feedback by training a system end-to-end, replacing the naive evolutionary algorithm or random exploration techniques.

3 INTERACTIVE MENTAL FACE RECONSTRUCTION

The goal of mental face reconstruction is to generate an image f_{rec} of a face that only resides in a user's mind f_{m}, such that f_{rec} matches the identity of f_{m}. To generate images we make use of a generative model G, StyleGAN2 [21], which was pre-trained on the FFHQ dataset [20] and widely used for face generation tasks. Given a vector z of size 512 encoding latent image features, the generator G generates a single image of a human face. Therefore, the goal of our proposed mental face reconstruction system (MFRS) is to find a vector z_{rec}, such that:

$$G(z_{\text{rec}}) = f_{\text{rec}} \overset{id}{=} f_{\text{m}}.$$

To gain information about z_{rec} our system shows a set of n predefined auxiliary face images $\mathcal{F}_{\text{aux}} = \{f_1, ..., f_n\}$ to a user. Subsequently, the user's task is to rank these auxiliary images according to the perceived similarity with their mental image f_{m}. A reconstruction network utilises this user ranking to extract information about f_{m}. Gaining enough information about f_{m} would likely require a large number of auxiliary images n to be shown to the user. As the number of possible rankings grows with $n!$, the ranking task quickly becomes infeasible for users. To address this, our approach is based on an iterative design paradigm, where the model collects evidence about f_{m} over multiple iterations. This allows us to keep the number of auxiliary images n shown per iteration low. Consequently, we set the number of auxiliary faces to show in each iteration to $n = 6$ for our MFRS, which has also been used in related work and proven not to overwhelm users [32, 40]. Furthermore, we set the number of iterations used in our final MFRS to 20, which we determined through experimental testing. This allows us to collect enough information about the mental image while preventing user fatigue and keeping the system execution time low. The architecture of our proposed iterative MFRS is shown in Figure 2.

The auxiliary face images $\mathcal{F}^i_{\text{aux}} = \{f^i_1, ..., f^i_6\}$ for each iteration i are generated once with StyleGAN2 [21] by decoding randomly sampled latent vectors. Therefore, for each set of auxiliary images $\mathcal{F}^i_{\text{aux}}$ we also have the corresponding latent vectors $Z^i_{\text{aux}} = \{z^i_1, ..., z^i_6\}, i = 1 \dots 20$. Similarly to Zaltron et al. [46], we divide the input space to our MFRS into four different categories based on sex (female/male) and age (young/old). Since there are 20 iterations with six faces each, our system requires $20 * 6 = 120$ different auxiliary images for each category. We used the InsightFace[2] toolbox to determine the sex and age (young $\leq$ age 40, following [26]) of each generated face to assign them one of the four categories.

3.1 Reconstruction Network

The goal of the reconstruction network is to predict a latent vector z_{rec} which, when decoded with the generator G, results in a face image that resembles the mental image f_m as closely as possible. Input to the reconstruction network are all sets of auxiliary latent vectors Z^i_{aux}, where the vectors of each set are ordered based on the user ranking. A sequence of recurrent and dense layers extracts information from each iteration separately and is then the input

[2] https://insightface.ai/projects

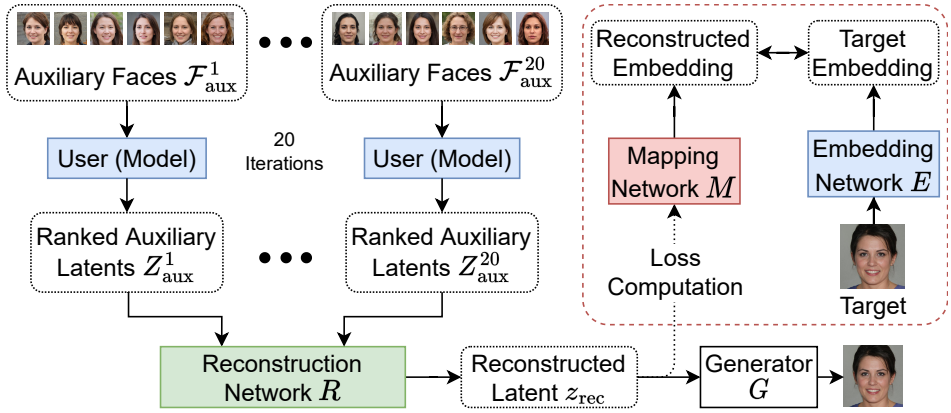


Figure 2: Our system shows users sets of auxiliary faces over multiple iterations, which the users have to rank according to the similarity with their mental image. A reconstruction network predicts the latent vector of the mental image according to the ranked auxiliary latents. This vector can be decoded with a pre-trained generator to reconstruct the image. To optimise the reconstruction network, a pre-trained mapping network maps the predicted latent vector into a meaningful embedding space, where it is compared against the true embedding vector.

Figure 3: Three example images generated with a state-of-the-art StyleGAN2 [21] generator. The mean absolute difference between the corresponding latent vectors z_A and z_B is higher compared to the difference between z_B and z_C, although images A and B are visually more similar. However, when using face embeddings extracted with ArcFace [11], the difference between the face embeddings e_A and e_B is smaller compared to e_B and e_C.

to another final sequence of recurrent and dense layers in order to combine the information across iterations and predict z_{rec}.

Training the reconstruction network to directly optimise for z_{rec} is unlikely to ensure that $z_{\text{rec}} \overset{id}{=} z_{\text{m}}$, since two similar latent vectors are not necessarily decoded by G into faces that are perceived to be similar by humans. For example, Figure 3 shows three faces that were generated by a state-of-the-art StyleGAN2 [21] generator. Face A and B are visually more similar to each other than face B and C. However, the mean absolute difference between z_A and z_B is higher compared to z_B and z_C. One reason for this is that the latent space encodes image features that are not relevant to the similarity of the generated faces such as background, pose and lighting. Additionally, facial features which are perceived to be similar by humans do not necessarily have to be close in the latent space. To address this issue, we instead aim to optimise a more meaningful embedding vector that allows us to calculate a perceived similarity between faces. Models such as ArcFace [11] map faces into an embedding space which is highly relevant for identity recognition as they are trained specifically for this task. Contrary to the latent space z, using such face embeddings $e_{A,B,C}$ for faces in Figure 3 results in a mean absolute difference between e_A and e_B that is smaller compared to e_B and e_C. Therefore, we define the loss function to train the reconstruction network as follows:

$$\mathcal{L} = -\frac{E(G(z_{\text{rec}})) \cdot E(G(z_{\text{M}}))}{\|E(G(z_{\text{rec}}))\| \, \|E(G(z_{\text{M}}))\|} \tag{1}$$

where G is a pre-trained generator network that maps from latent to image space, and E is a pre-trained face embedding network that maps from image to embedding space.

3.2 Mapping Network

Training with the loss function defined in Equation 1 is slow and unstable as gradients have to be propagated through the generator and embedding networks before actually optimising the reconstruction network. Therefore, instead of actually calculating $E(G(z))$, we build a compact mapping network which approximates this mapping: $M(z) \approx E(G(z))$. Given a pre-trained generator and embedding networks, arbitrary amounts of training tuples (z, e) can be generated to learn M by sampling random latent vectors z and calculating $E(G(z)) = e$. The network can then be trained by minimising the mean squared error between the true and predicted embedding:

$$\mathcal{L}_{\text{mapping}} = (e_{\text{pred}} - e_{\text{true}})^2 \tag{2}$$

Using this mapping network allows us to efficiently calculate the loss when training the reconstruction network by calculating $M(z)$ instead of $E(G(z))$ reducing the loss function in Equation 1 to

$$\mathcal{L}_{\text{rec}} = -\frac{M(z_{\text{rec}}) \cdot M(z_{\text{M}})}{\|M(z_{\text{rec}})\| \, \|M(z_{\text{M}})\|}. \tag{3}$$

Algorithm 1: The following algorithm defines our proposed user model. First, the auxiliary and target faces are mapped into an embedding space. Then, the cosine similarities between each of the auxiliary face embeddings and the target face embedding are calculated. Finally, the auxiliary face latents are ranked according to these similarities, i.e. the most similar face ranks first while the least similar face ranks last.

Input: Auxiliary faces of current iteration $\mathcal{F}^i_{\text{aux}}$, corresponding latents Z^i_{aux}, target face f_{m}, and a face embedding network E.
Output: Ranked auxiliary latents $(z_{R1}, z_{R2}, ..., z_{Rn}), z_R \in Z^i_{\text{aux}}$
similarities ← [ ]
for f in $\mathcal{F}_{\text{aux}}$ **do**
    similarity ← `cosineSimilarity`$(E(f), E(f_{\text{m}}))$
    similarities.append(similarity)
end for
rankedIndices ← `argSort`$(-\text{similarities})$
rankedLatents ← Z^i_{aux}[rankedIndices]
return rankedLatents

3.3 Computational User Model for Face Similarity Ranking

Training the mental face reconstruction network end-to-end requires a dataset that is very expensive to collect. For one training sample, a user has to memorise a generated face such that z_{m} is known and subsequently rank 6 auxiliary images 20 times according to their similarity with the memorised face. Therefore, we propose a novel user model that simulates human behaviour for the ranking task, which enables us to synthesise arbitrary amounts of training data. At the core of the user model stands a face embedding network that extracts embedding vectors that allow the computation of the similarity between two faces. Using this network, we define the user model in Algorithm 1: Given a set of auxiliary faces $\mathcal{F}^i_{\text{aux}}$, their latent vectors Z^i_{aux}, a target face f_m, and an embedding network E, the cosine similarities between target and auxiliary face embeddings are calculated. Subsequently, the auxiliary latent vectors are sorted according to these similarities, where the most similar face is ranked first followed by the other faces in order of decreasing similarity.

Existing models to extract face embeddings are trained on the task of face identification [11, 31, 43, 44]. While models trained on this task extract meaningful embeddings, comparing and ranking faces is not explicitly learned. Sadovnik et al. [32] analysed this and showed that measuring identity is not necessarily measuring similarity, which results in rankings dissimilar from human judgement. Therefore, it is beneficial to fine-tune an existing model for face identity on a small face similarity dataset based on human feedback to align the user model and human behaviour. For fine-tuning we use a small dataset consisting of triplets (f_a, f_p, f_n), where f_a is some reference face (anchor) and f_p a face that is more similar to f_a (positive pair) than f_n is to f_a (negative pair) based on human judgement. Using this dataset, the embedding network can be fine-tuned with a triplet margin loss objective defined as:

$$\mathcal{L}_{\text{c}} = max((f_a - f_p)^2 - (f_a - f_n)^2 + m, 0) \tag{4}$$

where m defines the required margin between the positive and negative pair to achieve a loss of zero. This requires the network to adjust the embeddings such that faces that are perceived to be similar by humans also have to be similar in the embedding space.

4 DATA COLLECTION

We collected data from humans ranking our 20 sets of six auxiliary faces based on the similarity to a target face. This provides us with a validation set to optimise hyper-parameters of our MFRS on real human data. Additionally, triplets can be generated from this data to fine-tune the face embedding network as discussed in Subsection 3.3.

4.1 Procedure

We implemented the data collection experiment as a website and collected data via Amazon's Mechanical Turk (AMT) platform. Participants had to complete 23 trials which consisted of a memorisation and a ranking step. During memorisation, participants had to observe a face f_m randomly generated by StyleGAN2 [21] until they had memorised it. The target face was fixed for all 23 trials, and participants could refresh their memory during the other memorisation steps. Afterwards, they were shown each set of auxiliary faces $\mathcal{F}^i_{\text{aux}}$ of the corresponding sex/age category of the target and were instructed to rank the six images according to the perceived similarity with the memorised face. In addition to the 20 iterations showing each auxiliary set, participants had to complete three additional test trials in between, which were used to ensure data quality. During test trials, we replaced one of the auxiliary faces with the target face. If participants properly engaged in the data collection study, we expected them to rank the target face as most similar.

4.2 Dataset Statistics

We cleaned the dataset by strictly removing data from participants that failed to pass all three test trials. From 317 participants, 62 failed all three test trials, 33 failed two, and seven failed one, resulting in a total of 215 participants in our dataset. Rejecting about a third of the participants based on standard attention checks is common for data collected via AMT [33].

Our collected dataset includes 15 target images for which we collected data from two different participants, allowing us to calculate the rank correlation between participants. Figure 4 shows the agreement between two human raters for each rank. Each cell (i, j) shows the probability that two independent human raters assign ranks i and j to the same face. While the chance of humans assigning the same rank is at most 34%, chances for strong disagreements are low. The average Kendall rank correlation coefficient between participants was 0.267 ($p<0.05$), suggesting that humans tend to rank faces similarly, but that there is also considerable variability in rankings.

After removing the 15 duplicates, we randomly held out 50 out of 200 remaining participants as a validation set to measure the MFRS performance on real human data during training. The data of the remaining 150 participants were split into 140 used to generate the triplets to fine-tune the ArcFace [11] embedding network and 10 for validation. For each iteration a participant completed, we can generate $\binom{6}{2} = 15$ different (f_a, f_p, f_n) triplets, where $f_a = f_m$ and (f_p, f_n) are all 15 possible tuples of the six auxiliary images. The image ranked higher in a tuple is defined as the negative example

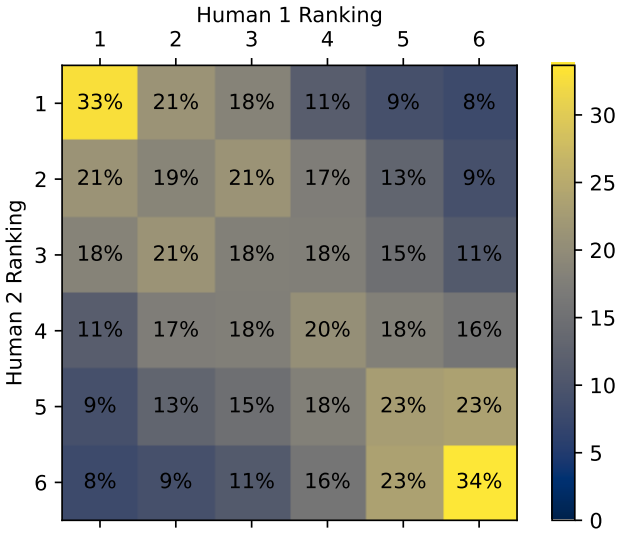


Figure 4: Face ranking agreement between humans. Each cell (i, j) shows the probability that two independent human raters assign ranks i and j to the same face. We observe a positive correlation between the human rankings with an average Kendall's Tau of 0.267. Furthermore, humans tend to agree more for the most and least similar face, while the rankings are more noisy for middle ranks.

f_p while the lower ranked image is defined as the positive example f_n. This yields a total of $15 * 20 = 300$ triplets for fine-tuning per participant, resulting in a total of 45000 triplets for training and 3000 for validating.

5 EXPERIMENTS

5.1 Implementation Details & Model Training

Embedding Network. For the embedding network used in our computational user model and MFRS loss calculation, we use the state-of-the-art face recognition network ArcFace [11] which was trained on the IBUG-500K dataset (11.96 million images with 493K identities). The network consists of a ResNet50 [17] feature extractor with frozen weights extracting 2048 $4x4$ feature maps. The feature maps are flattened and input to an output model consisting of a Batch-Normalization [18], Dropout (40% drop rate) [39], fully-connected (512 neurons) and a Batch-Normalization layer. The weights are initialised with the pre-trained ArcFace model. The ResNet50 feature extractor is frozen while the output model is fine-tuned using the collected triplets described in the Section 4. The network was trained with the contrastive loss with a margin of 0.1 (Equation 4) and Adam [22] optimiser with default parameters and a batch size of 32.

To evaluate the effectiveness of the fine-tuning step, we compared the performance of the user model as defined by Algorithm 1 before (a) and after (b) fine-tuning the embedding network (see Figure 5). Each cell (i, j) in the two matrices indicates the probability that humans assign rank i and the user model rank j to the same face. We notice that the probabilities along the main and adjacent diagonals increased overall, indicating higher agreement between user ranking and the fine-tuned user model. Before fine-tuning, the user model achieves an average Kendall's Tau of 0.205 ($p<0.05$) on the validation set, while after fine-tuning it achieved a value of 0.297 ($p<0.05$), resulting in an improvement of 45%. Compared with the ranking agreement between humans in Figure 4, our computational user model appears to provide rankings very similar to humans. Since we trained our MFRS only on rankings predicted by the user model, these results are important as it allows our MFRS to generalise to actual human feedback.

Mapping Network. The mapping network takes a 512-dimensional vector as an input, which is first input to a normalisation layer identical to the normalisation in StyleGAN2 [21], ensuring that it can be correctly interpreted and decoded by StyleGAN2 after training. Following the normalisation are five blocks each consisting of a fully-connected layer, LeakyReLU activation function with a slope of 0.2, and a Batch-Normalization [18] layer. Each fully-connected layer has 1024 neurons except for the last one which maps onto a 512-dimensional embedding. To train the mapping network, we sampled one million random latent vectors from a normal distribution and generated corresponding face images with StyleGAN2. We used our fine-tuned embedding network to generate the corresponding face embeddings. The model was trained to minimise the mean squared error with the Adam optimiser [22], batch size of 32 and exponential learning rate decay (initial learning rate of 0.005, decay rate of 0.9 and 80000 decay steps).

Reconstruction Network. The reconstruction network takes 20 times six ranked latent vectors of size 512 as an input, which correspond to the six ranked auxiliary faces for each of the 20 iterations. Each set of ranked auxiliary latents is then input to a separate feature extraction module specific to each iteration. Each of these modules consists of a recurrent layer with 1024 gated recurrent units (GRUs) [7] to capture the user ranking information followed by three blocks consisting of a fully-connected layer with 512 neurons, LeakyReLU activation with slope 0.2 and a Batch-Normalization layer. The output of each iteration module is used in a final prediction module consisting of a recurrent layer with 2048 GRUs followed by four blocks each consisting of a fully-connected layer with 1024 neurons, a LeakyReLU activation with a slope of 0.2 and a Batch-Normalization layer. A final fully-connected output layer with 512 neurons predicts the 512-dimensional latent vector z_{rec} of the target image.

To train the model, we generated one million target images by randomly sampling latent vectors from a normal distribution and decoding them with StyleGAN2 [21]. Together with the fixed sets of auxiliary images and the user model, we simulated the human ranking of the auxiliary images for each of the generated target images. The loss function used was the cosine similarity between the embeddings using the mapping network to map the predicted latent vector by the reconstruction network into the embedding space as defined in Equation 3. We trained the model with the Adam optimiser [22] with default parameters except for a learning rate of 0.0001 and batch size of 32 for 100K steps. As the validation set, we use the 50 left-out participants of the data collection study and keep the model achieving the lowest validation loss. Due to the small

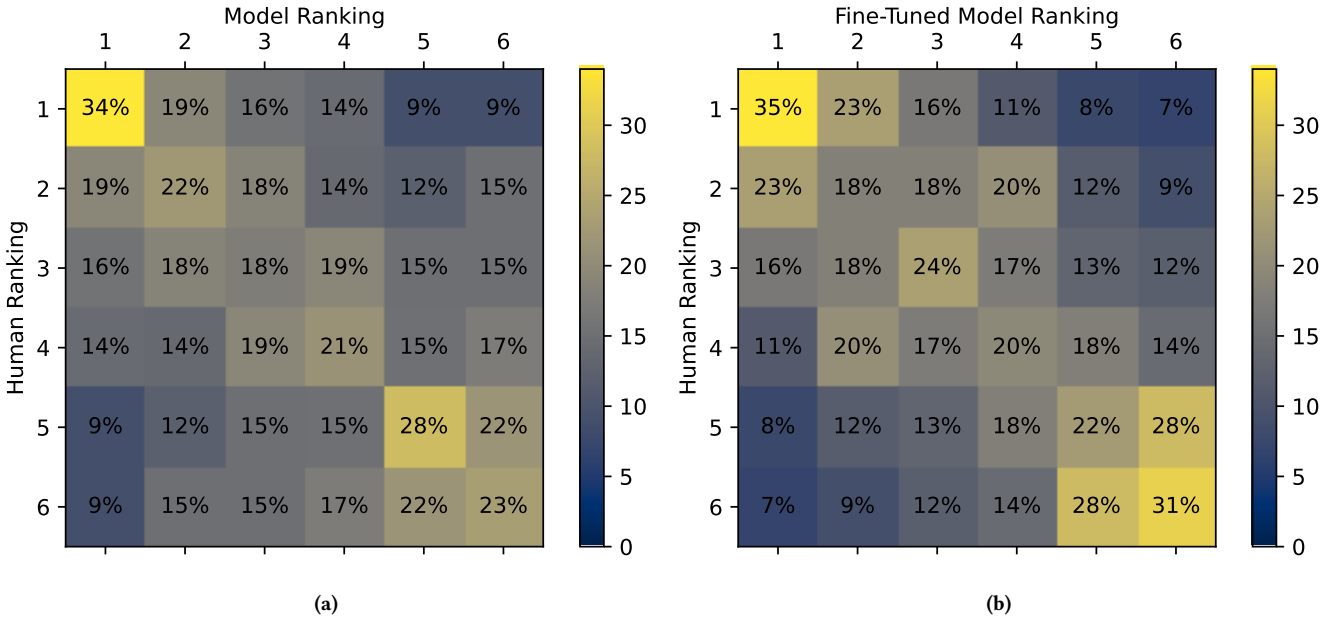


Figure 5: Face ranking agreement between humans and our user model. Each cell (i, j) shows the probability that the human raters assign rank i and the user model rank j to the same face. Figure (a) shows the agreement for the user model before and (b) after fine-tuning. Comparing figures (a) and (b), we observe that more probability mass is distributed across the main diagonal, showing higher agreement between human and model ranking.

validation set of real user data, the validation loss fluctuates during training. Therefore, we train a total of six reconstruction networks and use them as an ensemble in the user study. After participants finished the 20 iterations using our system, we showed them the six reconstructions and asked them to select the best image.

5.2 User Study

For evaluation of our MFRS, we compare the performance against the current state-of-the-art method CG-GAN [46]. We conducted a user study with 12 participants (four female) aged between 20 and 35 (Mean=25.4; SD=5.2). Participants were recruited through the university mailing list and financially compensated for their effort. Our study design was approved by the university's ethics commission.

5.3 Procedure

After giving their informed consent, participants had to complete two reconstruction experiments, one using our MFRS and one using CG-GAN. For each experiment, an explanation of the corresponding reconstruction system was given and participants could familiarise themselves with it. Once they were confident in using the system, a random target face was shown which they had to memorise. Since the generator of CG-GAN was pre-trained on the celebA-HD dataset [19] while our StyleGAN2 was pre-trained on the FFHQ dataset [20], we selected an equal number of random target faces from both datasets for a fair comparison. No time limit was given for the memorisation step, but participants were not able to see the image again after starting the experiment. The same target face was used for both experiments allowing us to compare the reconstructions for each participant. The order of the two experiments was counterbalanced between participants to average out possible memorisation effects. Following memorisation, participants used the first system without a time limit to reconstruct their mental image. After they were done, the reconstructed mental image of the system was presented. To compare the two systems, we collected five evaluation metrics:

- **Mental rating**: Participants were asked to rate the similarity between the image that they had memorised and the produced reconstruction on a seven-point Likert scale. The target mental image was not shown for this rating – users had to provide a rating based only on their memory of the target image.
- **Visual rating**: Participants were asked to visually rate the similarity between the target and the reconstructed image on a seven-point Likert scale. In contrast to the *Mental rating*, the two images were now shown side by side.
- **System Usability Scale (SUS)**: A questionnaire for assessing the system's usability. It produces a single usability score ranging from 0 to 100, with higher scores corresponding to more usability.
- **NASA Task Load Index (NASA-TLX)**: A multidimensional questionnaire for assessing the users' perceived workload. The final metric is a combination of the results from each subscale: mental demand, physical demand, temporal demand, overall performance, effort, and frustration level. The workload score is a value between 0 and 100 (lower is better).

Method	Mental Rating ↑	Visual Rating ↑	SUS ↑	NASA-TLX ↓	Time (mins) ↓
Ours	$4.0^{*} \pm .8$	$\mathbf{4.1 \pm .9}$	$\mathbf{85^{*} \pm 13}$	$\mathbf{27^{*} \pm 18}$	$\mathbf{10^{*} \pm 4.1}$
CG-GAN	$\mathbf{4.8^{*} \pm .8}$	$3.9 \pm .9$	$59^{*} \pm 13$	$43^{*} \pm 11$	$17^{*} \pm 5.6$

Table 1: Comparison of our system with CG-GAN [46] based on the user study results. The best result in each column is highlighted in bold. Our method outperforms CG-GAN in every metric except for the mental rating. * indicates statistically significant differences at the $p < 0.05$ level. ↑ and ↓ indicate if a higher or lower value is better.

- **Task completion time**: The amount of time participants needed to obtain the reconstruction. In the CG-GAN condition, participants could spend as much time as they wanted, until they were happy with the reconstruction result.

The same procedure was then repeated for the second experiment. Participants had to take a three-minute break between the experiments and could take additional breaks before and after using the systems. The duration of the study was about one hour, depending on how long participants needed to reconstruct their mental face with each system.

5.4 Results

Figure 6 shows example reconstructions from the conducted user study. The top row shows the target image that participants had to memorise. The second row shows reconstructions that resulted from our MFRS and the last row results from CG-GAN. Additional reconstructions of our MFRS are shown in Figure 11 and Figure 12 in the appendix. Quantitative results of the user study are shown in Table 1. For each metric, we computed whether the difference between our method and CG-GAN was significant using a paired t-test or Wilcoxon signed-rank test, depending on whether the compared samples stemmed from a normally distributed population, which we identified through a Shapiro-Wilk test. For a Bonferroni–Holm corrected p-value of < 0.05 we assumed that the difference was significant, indicated with an asterisk (*) in Table 1.

In terms of *mental rating*, study participants rated the reconstructions produced by CG-GAN higher than our method, 4.8 (SD=.80) vs. 4.0 (SD=.82). The *visual rating* of our system was higher than CG-GAN's, 4.1 (SD=.95) vs. 3.9 (SD=.95). However, the differences between the two conditions were not statistically significant. We also compared the change in scores from the *mental rating* to the *visual rating*, which reflects the users' perception of the reconstruction before and after seeing the target and the reconstructed image side by side. The increase in visual rating (4.1) for our method was not significantly different from the mental rating (4.0). However, the sharp decrease in rating for CG-GAN from a 4.8 mental rating to a 3.9 visual rating was statistically significant. Furthermore, Table 1 reports the average System Usability Scale (SUS) score, NASA Task-Load-Index (NASA-TLX) and task completion time. For all three metrics, our method significantly outperformed CG-GAN. The average SUS score increased from 59.0 (SD=12.7) to 84.6 (SD=12.9) with our system, a significant improvement in usability. Similarly, the perceived user workload is drastically reduced from a NASA-TLX rating of 43.4 (SD=10.9) to only 27.2 (SD=18.4). Finally, the average time it took participants to reconstruct their mental image was reduced by 40%, from about 17 mins (SD=5.6 mins) to about 10.0 mins (SD=4.1 mins).

5.5 Lineup Study

In addition to the metrics collected during the main study as reported in Table 1, we conducted another evaluation study to measure the identification rate of the produced reconstructions in a lineup study. Depending on the task, e.g. in forensics, it might not be necessary to produce perfectly accurate reconstructions as long as the portrayed person can be identified. To calculate the identification rate we need to create a lineup of potential candidate images containing the true target and similar looking faces. The task of participants is to rank this lineup according to the similarity with a given reconstruction. Similar to Zaltron et al. [46] we then define the identification rate ID based on the ranking as:

$$ID = \frac{\#\text{Rank 1}}{\#\text{Votes}} \times 100. \tag{5}$$

Creating a reasonable lineup is crucial for this metric and can bias the outcome. If the lineup contains faces that are dissimilar in appearance to the target, identification is trivial and the identification rate is inflated. To create the lineup Zaltron et al. [46] added noise to the latent vector of the generated target faces, which allowed them to create faces similar to the target. Since our targets are real faces, it is not straightforward to generate variations of the faces. Instead, we searched the three nearest neighbours in the respective datasets the target faces were selected from, FFHQ or CelebA-HQ. To find the nearest neighbours we selected faces with the most similar embedding vector using the ArcFace [11] embedding space, excluding image variations of the same person. This results in 12 lineups consisting of four candidate faces, each paired with the reconstructions from CG-GAN and our MFRS. Example lineups used for this study are shown in Figure 7. Although this study enables us to compare the identification rate of our system with CG-GAN, it is crucial to acknowledge that our lineups always include the target face. This aspect cannot be guaranteed in real-world scenarios and may lead to an inflation of the identification rate for both methods.

We recruited 22 colleagues and friends as independent raters for an online evaluation study without financial compensation. Each participant was randomly assigned to one of two groups and had to complete 12 trials in random order. For each trial they had to rank a lineup according to the similarity with the reconstruction from either our MFRS or CG-GAN, depending on the group. The first group had to rank a subset of six lineups based on our MFRS reconstruction and another six based on the CG-GAN reconstruction, while this assignment was flipped for the second group. This ensured that each participant saw each lineup only once to avoid unwanted side effects, while still evaluating all reconstructions between groups. Based on the results of this study we achieve a identification rate IR of 55.3% while CG-GAN achieves a identification rate of 56.1%. The target face was ranked within the top three using MFRS reconstructions in 94.0% of cases and with CG-GAN in 95.0% of cases. A p-value of 0.54 for a Wilcoxon signed-rank test

Figure 6: Example reconstructions from the user study. Each column shows the results for one participant. The first row shows the target face the participant had to memorise, the second row shows the reconstruction with our MFRS and the third row shows the reconstruction with CG-GAN.

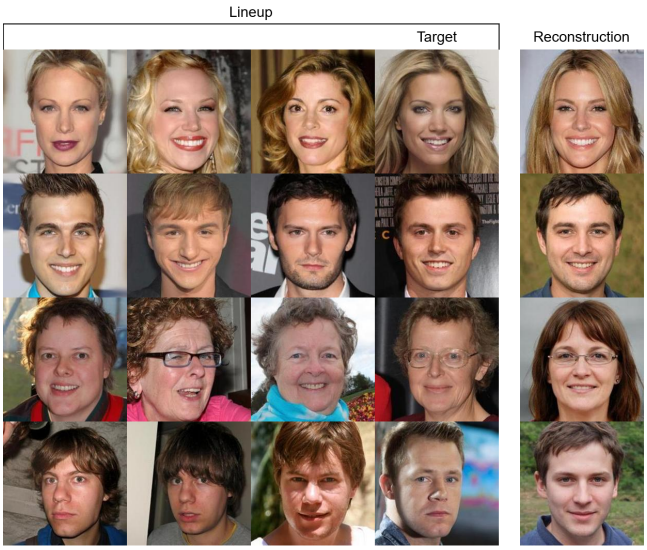


Figure 7: Example lineups used in our lineup study. Participants had to rank the lineup according to the similarity with the reconstruction.

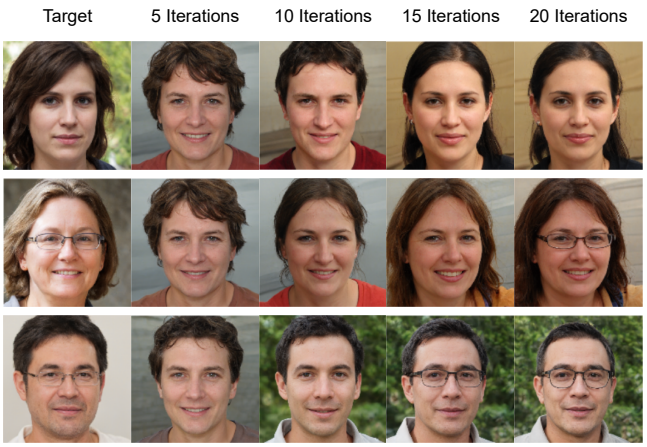


Figure 8: Reconstructions of our proposed system that show an increasing similarity with the target face with an increasing number of iterations. The first column shows three example target faces from the validation set. The following columns show the reconstruction for 5, 10, 15 and 20 iterations.

indicates that there is no significant difference between our MFRS and CG-GAN in terms of the identification rate.

5.6 Ablated Models

In addition to the user study, we conducted several ablation studies to evaluate different system components and hyper-parameters. One such hyper-parameter is the number of iterations participants had to complete. To evaluate this, we modified the MFRS to not only predict one latent vector z_{rec} based on all 20 iterations but to predict 20 vectors $z^i_{\text{rec}}, i = 1, 2, \ldots, 20$ using only the information from iterations 1 to i. For this, we trained the reconstruction network similarly as before except that we averaged the loss over all 20 predictions:

$$\mathcal{L}_{\text{rec}} = -\frac{1}{20}\sum_{i=1}^{20}\frac{M(z^i_{\text{rec}})\cdot M(z_{\text{M}})}{\left\|M(z^i_{\text{rec}})\right\|\left\|M(z_{\text{M}})\right\|}. \tag{6}$$

This allowed us to visualise the MFRS reconstruction after every iteration. Figure 8 shows three target faces from the validation set and reconstructions of our MFRS after 5, 10, 15 and 20 iterations.

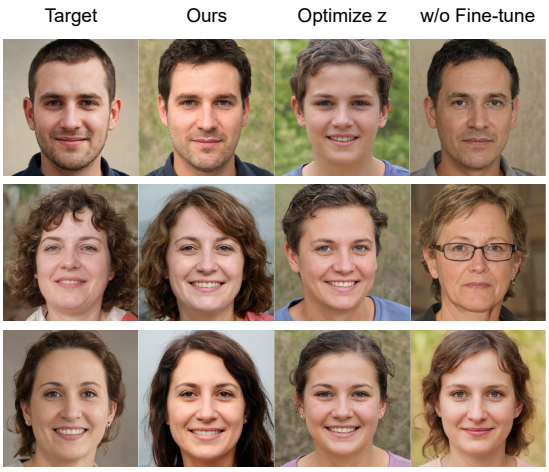


Figure 9: Results of our conducted ablation experiments. The first column shows three example targets from the validation set and the second column the respective reconstruction with our proposed system. The last two columns show ablated versions where we train the reconstruction network to optimise the latent vector z of an image directly instead of the embedding (column 3) or train with the non fine-tuned user model (column 4).

As expected, we observe that the similarity between target and reconstruction increases the more user ranking information the MFRS receives.

Another important component of our system is the mapping network, which allowed us to compute the loss in the embedding space instead of the latent space of the generator. To evaluate the importance of this step, we trained the reconstruction network such that it optimises the mean squared error between the target z_m and reconstructed z_{rec} latent vector directly instead of the similarity in the embedding space. Figure 9 shows three example target faces from the validation set in column one, the reconstructions with our method in column two and reconstructions without the mapping network in the third column. The results indicate that our method produces visually more similar mental face reconstructions with the mapping network.

Finally, we evaluated the importance of the fine-tuning step of the embedding network used in the user model. For this, Figure 9 additionally shows qualitative results of our method without fine-tuning the user model in the last column. Results appear inferior when compared to our method with fine-tuning.

6 DISCUSSION

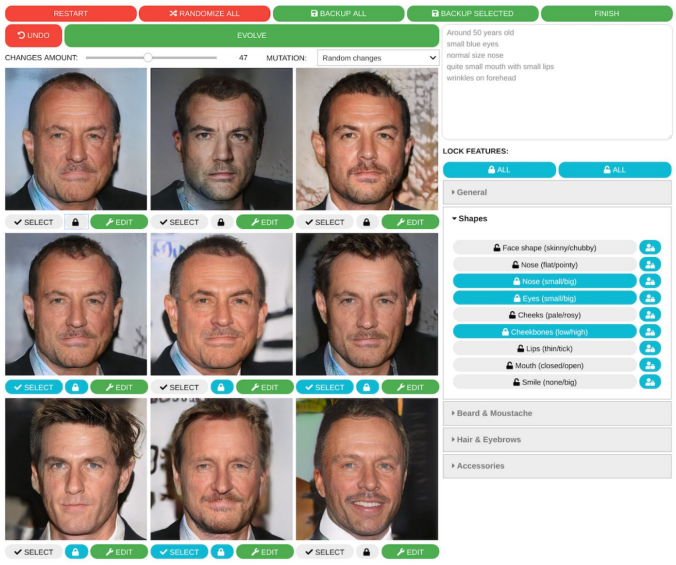
In the sections that follow, we discuss our system's performance and the difficulty in evaluating mental image reconstruction systems.

6.1 Comparison with CG-GAN

Results from our user study (see Table 1) demonstrated significant performance improvements in system usability, workload, and task completion time without sacrificing the reconstruction quality when compared to the current state of the art [46]. To assess the quality of the reconstructions, we introduced two metrics: the *mental rating* and *visual rating*. The *mental rating* captures the users' perception of similarity to the target image at the end of an experiment based on what the users remember. In contrast, the *visual rating* compares the two images side by side, i.e. the target image and the reconstruction, thereby allowing users to more accurately judge the reconstruction quality. While CG-GAN outperformed our method on the mental rating (4.8 vs. 4.0), our method's performance on the visual rating was comparable to CG-GAN's (4.1 vs. 3.9). These results show a sharp decrease in performance for CG-GAN from a mental rating of 4.8 to a visual rating of 3.9, which suggests a difference in what participants remembered, compared to the actual target image they had to memorise. One reason for this discrepancy might be that CG-GAN always visualises the current state of the face being edited. Because of this, we hypothesised that the participants' mental images changed during the lengthy editing process with CG-GAN. This interesting finding was validated through post-study interviews with the participants. Three of the twelve participants mentioned that they were surprised to see the actual target after using CG-GAN, as they expected it to look different. In addition, they reported that it was difficult to memorise the target while actively manipulating facial features. While further work is necessary to validate this hypothesis, it is likely that a participant's mental image changes and moves closer to the image that they try to generate using, for example, CG-GAN. The impact of such a "mental shift" depends on the application. While it may be minimal for generating faces of digital avatars or characters, it is severe for facial composite generation in criminal investigations. However, it remains unclear why and to which degree this mental image shift occurs. We suspect that the degree of familiarity with the target image plays an important role. In our user study, familiarity was low as participants were exposed to the target face for the first time. While the scenario of unfamiliar target faces is more in line with the task of composite generation from a witness's memory, a comparison of the systems with familiar faces will be interesting for future work.

For many tasks, especially in forensics, the objective is not to maximise the perceived similarity of the reconstruction to the target, but rather the recognition rate of the target given a reconstruction. Through an additional user study, we estimated the recognition rate of our MFRS and CG-GAN to be 55.3% and 56.1%, respectively, with no significant difference between the systems. This finding reinforces that the significant usability improvements of our MFRS do not come at the expense of reconstruction quality. Furthermore, it confirms the overall effectiveness of our system in producing reasonable reconstructions. Considering that the CelebA-HQ [19] and FFHQ [20] datasets comprise 30K and 70K images, respectively, and we select the nearest neighbours for the lineups, a recognition rate of 55.3% is substantial. In 94% of the cases, the target face was not assigned the last rank, indicating that our reconstruction was closer to the target than to at least one of the three nearest neighbours. Moreover, the comparable identification rates observed between CG-GAN and our MFRS suggest that the higher mental rating of CG-GAN does not necessarily result in a higher identification rate. This further underscores the possibility of an unintended side-effect associated with CG-GAN, as previously discussed.

(a)



(b)

Figure 10: Figure (a) shows the user interface of our system. Users are presented with six faces that they can rank via drag and drop. Figure (b) shows the main interface of CG-GAN. Users are presented with 9 faces that they can manipulate via randomising, mutating or manual editing.

While our method is similar to CG-GAN in terms of reconstruction quality, our face reconstruction system significantly outperforms it on three usability metrics. Our method achieved a significantly higher SUS score, a lower task load index, and a much faster task completion time. We believe the main reason for the high SUS scores is our user-friendly system design: splitting the reconstruction objective into multiple small ranking tasks is easy to explain, understand, and execute for participants. This also allows for a much simpler user interface: Figure 10 shows the main interface of our proposed system at the top and the main interface of CG-GAN at the bottom. Visually comparing these two interfaces shows that the interface of our system is much simpler, which is likely another contributing factor to the significantly higher usability score.

While our system reconstructs the mental face based on user rankings, participants have to actively reconstruct their mental image with CG-GAN. Finding a good initial face with CG-GAN heavily relies on randomness without much user control. Similarly, optimising the reconstructions through the evolutionary algorithm is based on random exploration but also requires the users to select appropriate faces. Manual editing of faces adds a significant amount of complexity and participants reported frustration because certain face controls were missing or the existing controls were heavily entangled with multiple features, often preventing them from implementing their intended changes. The increased complexity and higher user frustration are also reflected in the significantly higher task load index of 43.4 compared to 27.2 with our system (see Table 1).

6.2 Analysis of the Reconstruction System

We conducted several ablation experiments to study the importance of each component of our method. Our results (Figure 8), including comparisons to CG-GAN (Figure 6, Table 1), suggest that performing 20 iterations is a good trade-off between reconstruction quality and task completion time. Even though the change between 15 and 20 iterations led to minor changes in reconstruction quality (Figure 8), we decided to go for the best mental face reconstruction, which was still 40% faster than CG-GAN. Our proposed loss function in Equation 6 allowed us to not only visualise the reconstructions at every iteration but also has interesting usability implications. While participants had to complete all 20 iterations in our evaluation study, the modified optimisation objective allows for stopping the reconstruction process at any iteration. Users could see the reconstruction result after a few iterations, deciding when to stop the system. However, how and when to visualise intermediate reconstructions is an open question: visualising the reconstruction after every iteration could cause a "mental shift", as discussed in Subsection 6.1. Another possibility is to define an automatic stop criterion through the system, e.g. stop if the difference between two iterations is below a threshold α: $M(z^i_{\text{rec}}) - M(z^{i+1}_{\text{rec}}) < \alpha$.

Another important factor is the selection of auxiliary faces that are presented to the participants. While we chose the auxiliary faces for each iteration at random, it might be possible to carefully select these to increase overall system performance. While we do not expect this to influence the reconstruction quality, it might reduce the number of iterations needed to extract the same amount of information. We conducted a preliminary experiment in which the objective was to learn the auxiliary faces. Instead of passing the ranked auxiliary latent vectors into the reconstruction network, we provided a constant value as input which was followed by a fully connected layer to predict the auxiliary latent vectors for a given iteration. Since the user model is fully differentiable, it could be integrated into the reconstruction network to rank the learned faces during training. However, this approach has two main drawbacks. First, we would not be able to collect real human ranking data for the auxiliary faces to fine-tune the user model as those are not known before training the reconstruction network. Second, the

learned auxiliary faces would often be extremely different in appearance. While this maximises the information the reconstruction network can extract from the faces in each iteration, it would make it more difficult for humans to rank and compare, resulting in noisy feedback.

6.3 Challenges in Evaluating Mental Image Reconstruction Systems

Looking beyond the discrepancy between CG-GAN's mental and visual rating, both methods achieved an average visual rating of around four out of seven. While this indicates, on the one hand, that reconstructions were similar in quality, it also highlights, on the other hand, a significant gap to the actual mental image. Visually comparing the target and reconstructed faces in Figure 6, Figure 11 and Figure 12 leaves the same impression. Given the immense variability in face appearance, the reconstructed faces of both systems look similar compared to the targets overall. However, participants never thought that they looked like the same person, which raises the question of how much better such reconstruction systems can become. As we are reconstructing high-fidelity images from human memory, this task is inherently noisy. Consequently, the evaluation of such systems remains an open challenge. While the mental rating considered the users' mental image, their mental encoding may itself be influenced by the system as discussed in Subsection 6.1. The visual rating is also limited because participants often focused on features unrelated to the identity such as facial expression and hair. Therefore, future systems might benefit from disentangling these features by showing neutral faces without hair and accessories, which is also done similarly in classical systems like EvoFit [14]. After reconstructing the inner facial features and the head shape, these features could then be identified and added post-hoc. This could potentially improve reconstruction performance and allow for a visual rating more focused on the identity of individuals.

7 CONCLUSION

In this work, we presented an end-to-end trainable, interactive system for mental face reconstruction. In stark contrast to prior works that required users to explicitly reconstruct mental images using tedious and time-consuming tools, our system only requires users to rank images of faces according to the similarity to their mental images over multiple iterations. Our reconstruction system combines the image features from each iteration into a single vector that is visually decoded into an image with a state-of-the-art generative model. Through quantitative and qualitative evaluation in a 12 and 22-participant user study, we demonstrated superior performance in terms of system usability, cognitive load, and usage time without sacrificing reconstruction quality or identification rate – our method's reconstruction performance was comparable to prior work. As such, our work presented a new interactive intelligent system that can be used to quickly and effortlessly reconstruct a user's mental image and has yielded interesting insights that could help to further improve such systems in the future.

ACKNOWLEDGMENTS

Florian Strohm and Andreas Bulling were funded by the European Research Council (ERC) under the grant agreement 801708. Mihai Bâce was funded by a Swiss National Science Foundation (SNSF) Postdoc.Mobility Fellowship (grant number 214434).

REFERENCES

[1] Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. 2019. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI. In *Advances in Neural Information Processing Systems*. 6517–6527.

[2] Philip Bontrager, Wending Lin, Julian Togelius, and Sebastian Risi. 2018. Deep interactive evolution. In *International Conference on Computational Intelligence in Music, Sound, Art and Design*. 267–282.

[3] Andrea Bruera and Massimo Poesio. 2022. Exploring the representations of individual entities in the brain combining EEG and distributional semantics. *Frontiers in artificial intelligence* (2022), 25.

[4] Andreas Bulling and Daniel Roggen. 2011. Recognition of Visual Memory Recall Processes Using Eye Movement Analysis. In *Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. 455–464. https://doi.org/10.1145/2030112.2030172

[5] Chia-Hsing Chiu, Yuki Koyama, Yu-Chi Lai, Takeo Igarashi, and Yonghao Yue. 2020. Human-in-the-loop differential subspace search in high-dimensional latent space. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 85–1.

[6] Donald F Christie and Hadyn D Ellis. 1981. Photofit constructions versus verbal descriptions of faces. *Journal of Applied Psychology* 66, 3 (1981), 358.

[7] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning*.

[8] Alan S Cowen, Marvin M Chun, and Brice A Kuhl. 2014. Neural portraits of perception: reconstructing face images from evoked brain activity. *Neuroimage* 94 (2014), 12–22.

[9] Thirza Dado, Yağmur Güçlütürk, Luca Ambrogioni, Gabriëlle Ras, Sander Bosch, Marcel van Gerven, and Umut Güçlü. 2022. Hyperrealistic neural decoding for reconstructing faces from fMRI activations via the GAN latent space. *Scientific reports* 12, 1 (2022), 1–9.

[10] Hiroto Date, Keisuke Kawasaki, Isao Hasegawa, and Takayuki Okatani. 2019. Deep learning for natural image reconstruction from electrocorticography signals. In *2019 IEEE International Conference on Bioinformatics and Biomedicine*. 2331–2336.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4690–4699.

[12] Hadyn D Ellis, Graham M Davies, and John W Shepherd. 1978. A critical examination of the Photofit system for recalling faces. *Ergonomics* 21, 4 (1978), 297–307.

[13] Martha J Farah, Kevin D Wilson, Maxwell Drain, and James N Tanaka. 1998. What is" special" about face perception? *Psychological Review* 105, 3 (1998), 482.

[14] Charlie D Frowd, Peter JB Hancock, and Derek Carson. 2004. EvoFIT: A holistic, evolutionary facial imaging technique for creating composites. *ACM Transactions on applied perception* 1, 1 (2004), 19–39.

[15] Stuart J Gibson, Chris J Solomon, Matthew IS Maylin, and Clifford Clark. 2009. New methodology in facial composite construction: From theory to practice. *International Journal of Electronic Security and Digital Forensics* 2, 2 (2009), 156–168.

[16] Yağmur Güçlütürk, Umut Güçlü, Katja Seeliger, Sander Bosch, Rob van Lier, and Marcel A van Gerven. 2017. Reconstructing perceived faces from brain activations with deep adversarial neural decoding. *Advances in Neural Information Processing Systems* 30 (2017).

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[18] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*. 448–456.

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

[20] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4401–4410.

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8110–8119.

[22] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[23] Christine E Koehn and Ronald P Fisher. 1997. Constructing facial composites with the Mac-a-Mug Pro system. *Psychology, Crime and Law* 3, 3 (1997), 209–218.

[24] Kenneth R. Laughery and Richard H. Fowler. 1980. Sketch artist and Identi-kit procedures for recalling faces. *Journal of Applied Psychology* 65, 3 (June 1980), 307–316.

[25] Yunfeng Lin, Jiangbei Li, and Hanjing Wang. 2019. DCNN-GAN: Reconstructing Realistic Image from fMRI. In *2019 16th International Conference on Machine Vision Applications*. 1–6.

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*. 3730–3738.

[27] Andrew J Logan, Gael E Gordon, and Gunter Loffler. 2017. Contributions of individual face features to face discrimination. *Vision research* 137 (2017), 29–39.

[28] Thomas Naselaris, Cheryl A Olman, Dustin E Stansbury, Kamil Ugurbil, and Jack L Gallant. 2015. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *Neuroimage* 105 (2015), 215–228.

[29] Dan Nemrodov, Matthias Niemeier, Ashutosh Patel, and Adrian Nestor. 2018. The neural dynamics of facial identity processing: insights from EEG-based pattern analysis and image reconstruction. *Eneuro* 5, 1 (2018).

[30] Adrian Nestor, David C Plaut, and Marlene Behrmann. 2016. Feature-based face representations and image reconstruction from behavioral and neural data. *Proceedings of the National Academy of Sciences* 113, 2 (2016), 416–421.

[31] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. In *Proceedings of the British Machine Vision Conference*.

[32] Amir Sadovnik, Wassim Gharbi, Thanh Vu, and Andrew Gallagher. 2018. Finding your lookalike: Measuring face similarity rather than face identity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2345–2353.

[33] Antonios Saravanos, Stavros Zervoudakis, Dongnanzi Zheng, Neil Stott, Bohdan Hawryluk, and Donatella Delfino. 2021. The hidden cost of using Amazon Mechanical Turk for research. In *HCI International 2021-Late Breaking Papers: Design and User Experience: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings 23*. Springer, 147–164.

[34] Hosnieh Sattar, Andreas Bulling, and Mario Fritz. 2017. Predicting the category and attributes of visual search targets using deep gaze pooling. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2740–2748.

[35] Hosnieh Sattar, Mario Fritz, and Andreas Bulling. 2020. Deep gaze pooling: Inferring and visually decoding search intents from human gaze fixations. *Neurocomputing* 387 (2020), 369–382.

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

[37] Sophia M Shatek, Tijl Grootswagers, Amanda K Robinson, and Thomas A Carlson. 2019. Decoding images in the mind's eye: The temporal dynamics of visual imagery. *Vision* 3, 4 (2019), 53.

[38] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. 2019. Deep image reconstruction from human brain activity. *PLoS computational biology* 15, 1 (2019), e1006633.

[39] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[40] Florian Strohm, Ekta Sood, Sven Mayer, Philipp Müller, Mihai Bâce, and Andreas Bulling. 2021. Neural Photofit: Gaze-based Mental Image Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 245–254.

[41] Florian Strohm, Ekta Sood, Dominike Thomas, Mihai Bâce, and Andreas Bulling. 2022. Facial Composite Generation with Iterative Human Feedback. In *Proceedings of Machine Learning Research*.

[42] Rufin VanRullen and Leila Reddy. 2019. Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications biology* 2, 1 (2019), 1–10.

[43] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. 2018. Additive margin softmax for face verification. *IEEE Signal Processing Letters* 25, 7 (2018), 926–930.

[44] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5265–5274.

[45] Caie Xu, Ying Tang, Masahiro Toyoura, Jiayi Xu, and Xiaoyang Mao. 2019. Generating Users' Desired Face Image Using the Conditional Generative Adversarial Network and Relevance Feedback. *IEEE Access* 7 (2019), 181458–181468.

[46] Nicola Zaltron, Luisa Zurlo, and Sebastian Risi. 2020. Cg-gan: An interactive evolutionary gan-based approach for facial composite generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 2544–2551.

[47] Xiao Zheng, Wanzhong Chen, Mingyang Li, Tao Zhang, Yang You, and Yun Jiang. 2020. Decoding human brain activity with deep learning. *Biomedical Signal Processing and Control* 56 (2020), 101730.

Figure 11: Additional reconstructions of our method on the validation set using real human ranking. Each image pair shows the target face on the left and our reconstruction on the right.

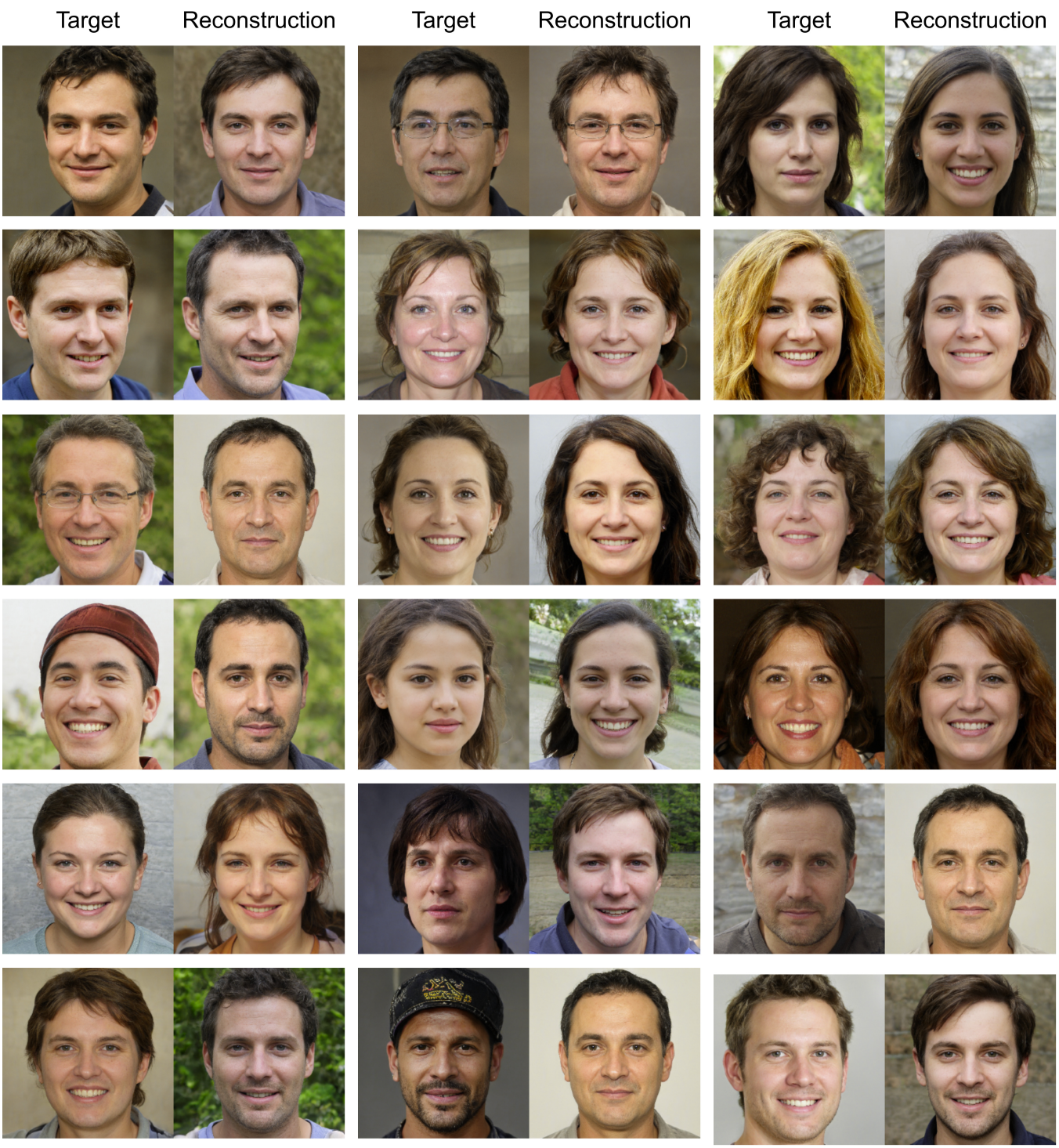


Figure 12: Additional reconstructions of our method on the validation set using real human ranking. Each image pair shows the target face on the left and our reconstruction on the right.