2025 19th International Conference on Automatic Face and Gesture Recognition (FG)

Towards ML-based Assessment of Synthetic Characters Heads

Igor Borovikov[1], Karine Levonyan[1], Panda Elliott[2], and Etienne Danvoye[3]
[1] Electronic Arts, SEED, Redwood City, CA, USA, {iborovikov, karine}@ea.com
[2] Electronic Arts, SEED, Vancouver, BC, Canada, pelliott@ea.com
[3] Electronic Arts, SEED, Montreal, QC, Canada, edanvoye@ea.com

***Abstract*— Virtual environments present ever-growing requirements for their population of synthetic characters. In many applications, various character heads must provide a balanced representation of age, gender, and ethnicity. With a character count well above 10,000, manually checking and verifying the target metrics is impractical. This paper outlines a possible pipeline for generating parametric avatar heads. The main focus is the final stage, where generated character heads are evaluated for aesthetic quality metrics. The proposed quality assurance (QA) approach uses ML models trained on sparse data obtained from human evaluation. The QA ML models' training data collection leverages in-house crowdsourcing and aims to match the assessment initially provided by the expert art direction. We illustrate the approach with heads generated using FLAME.**

I. INTRODUCTION

Realistic 3D avatars are an active area of research and development in gaming and virtual reality (VR). A detailed, realistic head is vital for delivering likable, believable virtual characters. Head shape modeling is the key part of such exploration. This paper focuses on the perceptual quality of 3D parametric avatar head shapes produced in large numbers using a mixture of various methods, including generative. We use Machine Learning (ML) to predict the artistic evaluation of such heads to automate a complete production pipeline. Artistic evaluation is part of the Quality Assurance (QA) workflow. A positive aesthetic rating of the produced heads increases user acceptance of the generated avatars. A more formal definition of positive aesthetic rating follows in later sections. The section on previous work indicates that aesthetic evaluation was primarily concerned with beauty rather than other utilitarian metrics. The paper intends to make a practical contribution to that area of research.

To put our work in the proper context, we begin with a brief overview of parametric and morphable models, followed by a quick mention of AI-driven generative techniques. The introductory discussion (with aesthetic aspects in the back of the mind) helps to determine the motivation behind our decisions for the in-house avatar generation pipeline before defining metrics and diving into the quality assessment. An important note: In our work, heads are considered static and neutral, i.e., we exclude facial expressions, animation rigging, general facial animation, and speech in particular. Scalp and facial hair play a significant role in the perception of the human face, but we take that aspect into account tangentially, deferring their selection to the art direction.

A. Overview of Head Shape Representation

Numerous practical systems for human head modeling utilize established approaches like 3D Morphable Models (3DMM) and parametric models. With a general 3D modeling tool at the lowest level, a morphable modeling system allows direct vertex manipulation, including texture coordinates, applying various blendshapes, and detailed animation rigging. At such a detailed level, a system offers exceptional control over minute details and extends beyond human shapes, but it heavily relies on 3D artists' mastery and custom assets. A far cry from mass production, sharing such one-off models between products or in academia is neither feasible nor practical. The aesthetic qualities of such models are handled individually.

Quantitative observations of actual head shapes lead to standardization and a more structural approach. A 3D Morphable Mesh (3DMM) system captures facial structures via a linear combination of basic shapes (aka in the industry "blendshapes") [9]. The blendshapes can be derived from the statistical analysis of real-world scans. Structured modeling leads to higher-level parametric representations, e.g., the Basel Face Model (BFM) [24], [20], and FLAME [29]. The neutral head in such systems represents the shared base shape, and the individual features come from blendshapes combined with weights (parameters). These parametric models enable avatar creators to generate diverse facial structures while maintaining compatibility with standardized rendering and production pipelines. The range of weights for blendshapes is usually controlled by the technical validity of the resulting mesh, i.e., by the necessity to avoid creases, folds, and self-intersections. From an aesthetic point of view, a narrow range of weights keeps generated heads in reasonable agreement with typical artists' expectations but may negatively affect the variety of possible heads.

The 3DMM remains an active area of research, with later contributions addressing the limitations of the original approach. Introduction of non-linearity [52] improves representation accuracy. A combination of 3DMMs leverages data from different sources [43]. More complete models cover the entire head, including the face, cranium, ears, eyes, teeth, and tongue [42]. Focusing on concrete regions like ear [16] extends the approach even further. A relatively recent review of the 3DMMs is in [17]. In addition to compact and accurate shape representation, parametric models in commercial software enable end-user avatar customization. Systems such as Ready Player Me [45] and Epic Games' MetaHuman

979-8-3315-5341-8/25/$31.00

Creator [19] allow users to modify head proportions, facial features, and skin details interactively. The data formats for these systems allow their use in popular game engines. Also, many video games featuring proprietary systems for avatar customization use custom parametric representation and blendshapes. In commercial applications, design decisions play a critical role. The design usually favors pleasing avatar heads attractive to a broad multicultural audience. That dictates additional limits on top of technical correctness for the controls. It also enforces necessary correlations between parameters, e.g., by introducing curated templates available for blending [19].

Generative methods are rapidly gaining popularity due to exciting results. The most direct way of using the generative approach in the industry is to reconstruct 3D heads from images generated by commercial systems like Adobe Firefly [1] or other readily available counterparts. The generative systems can process text inputs to produce human portraits that closely meet the specifications. The transition to a 3D head model requires monocular reconstruction, which is well-studied; a recent review is available in [8]. The paper [31] demonstrates a monocular reconstruction technique in application to video game avatars. Other public and proprietary neural models produce human heads using generative approaches. Direct sculpting from 2D generative images is also possible [34]. A complete 3D generative approach for human heads can provide a shortcut by eliminating the 2D stage [62]. During the generation phase, including keywords for aesthetic aspects (e.g., "attractive") may steer the generation in the desired direction. However, that requires inventive prompt engineering to keep the imagery sufficiently diverse. Building a complete automated pipeline with such an approach may be cumbersome.

Methods such as Neural Head Avatars are another active research area (see a review in [28]). Such models implicitly capture 3D facial morphologies and often surpass the expressiveness of traditional parametric or 3DMMs. Training neural parametric models requires extensive (and expensive) datasets of human head scans to distill them in a compact yet expressive form of embedding. Publicly available models like FaceNet [46] may provide a practical replacement for less accessible ones. Embeddings enable the mass production of heads by mapping latent vectors to the interpretable parameters [10]. This method may scale up in production but lacks efficient pipeline controls to meet required metrics such as gender, age, and ethnicity with sufficient variety in each category. The aesthetic aspects also remain difficult to control.

Currently, the head generation of human avatars draws artistic or aesthetic evaluation from humans in the loop (character artists, game designers). It also relies on implicit (embeddings) or explicit (parameter ranges) constraints. In interactive applications, generation systems' are geared towards attractiveness. That makes the problem of aesthetic evaluation less critical when the generation process objective is to produce a limited number of heads. However, the bias towards "attractive" faces may result in a loss of visual variety. With variety and mass production becoming a priority, the built-in attractiveness may need additional support from an automated ML-based evaluation.

II. Previous Work

Various motivations trigger the exploration of attractiveness, beauty prediction, or similar aggregate attributes of human faces. Facial Beauty Prediction (FBP) is a commonly accepted name for this research area. The subject is widely studied, but due to either a strictly academic or commercial approach, the datasets and models mentioned in this section are proprietary and unavailable for commercial use in applications to virtual worlds.

We skip works based on measured anthropometric features as input (primarily applicable in plastic and restorative surgery, e.g., [22]) and focus on visual inputs.

Various definitions of facial beauty revolve around key concepts like the golden ratio, facial symmetry, and the averageness hypothesis with support from early datasets and models. These are discussed in the still-relevant [60] and [59]. The methods continue development by utilizing general popular characteristics like the golden ratio, e.g., [39], [26], [54], [25], [7] to the application of classic Computer Vision (CV) [21] and more recent ML and AI-based approaches (discussed next), all the way to the recent generative text-to-image techniques [6].

Early works establish the applicability of ML, e.g., [18]. They utilize visual input with promising results, showing that the subject lends itself to ML and can approximate human perception even from a limited-size dataset of images. Instead of full image input, smaller dimensionality inputs, like facial landmarks, are one of the popular features. Landmarks and similar geometric features appear helpful for beauty definition and prediction using various ML models in, e.g., [27], [61], [14], [51], [15]. However, the landmarks can be explicitly ignored with still good results [21].

Training some of the more powerful models for facial attractiveness became possible with larger datasets like SCUT-FBP [57] and SCUT-FBP5500 [30]. However, their data is limited to Caucasian and Asian faces. Conditioning on ethnicity is frequent in the field of facial beauty. Nigerian [26], Turkish [38], Gujarati [49], North Indian [35], and Bengali [2] (and many more) faces are the subjects of exploring the golden ratio. To balance the narrow focus, the racial fairness approach in [37] highlights another facet of beauty exploration.

The rise of online dating motivates research leveraging large-scale crowdsourced ratings, as seen in studies like "Hot-Or-Not" [5]. The potential dating angle introduces certain cultural biases both in the input imagery and in the scores. Similarly, the social media angle focuses on altering or selecting images to enhance the beauty aspect [56], [33]. The large-scale celebrity datasets CelebA and CelebA-HQ [32] offer the "Attractive" attribute but are not available for commercial applications.

Authors in [58] propose leveraging attention mechanisms for FBP through transfer learning by tuning the pre-trained

SCUT-FBP500 dataset to their target dataset. Recent advancements include the work of [11], who demonstrate that vision transformers outperform traditional CNN-based neural classifiers for facial beauty assessment. A recent introduction of Anchor-Net [4] relies on an ensemble-based approach utilizing semi-supervised learning. This method predicts beauty scores by coupling ResNet predictions with the relative distances between faces in the embedding space, effectively capturing perceived beauty differences.

Since relevant data and models are not widely available for commercial applications, the industry is left to collect its in-house datasets and train models tailored to their specific needs.

III. Contribution

This work expands beyond previous studies by exploring virtual worlds as an application domain, including synthetic heads created with generative methods and considering practical beauty standards. Our analysis also relies on a novel scoring approach utilizing a panel of proxy experts approximating the principal expert.

Most published work on facial beauty concerns real humans, while the presented work explores the acceptability of synthetic heads as the main subject. The subject is related to the "uncanny valley" problem of synthetic avatars (e.g., [13], [47], [48], [36], and more) but is not the same since we are exploring heads generated from the same pipeline and sharing the same level of details, feature completeness, and distinct artistic style. To that end, we propose a more practical standard for beauty, limiting it to the level of "acceptable," i.e., a level that allows releasing it with a product without triggering negative user feedback. That leads to a simple binary classification and preserves "interesting" heads that will not necessarily win a beauty contest but are acceptable and engaging. In that sense, the mainstream of the FBP and "uncanny valley" research only applies on a conceptual level.

A proposed name for our approach is "proxy crowdsourcing". We tap into in-house crowdsourcing to approximate and augment the evaluations provided by the art direction. The art directors' time is considered more valuable than the time provided by in-house volunteers willing to give ratings to a small set of synthetic heads as a fun distraction from their primary duties. Since the respondents work for the same company, their perception of acceptability is aligned to a degree and introduces useful bias into our experiments.

IV. In-house Head Generation Pipeline Overview

While we present results obtained with open source parametric model FLAME [29], they can translate to the in-house pipeline, which we outline in this section. The in-house head shape parametric model conceptually follows a typical pattern of having a handcrafted "neutral head" or "base shape" representing an average of all heads. A collection of blendshapes (over 600) covers all major areas of a complete head, allows for asymmetries, and comprises the head shape model. Initially, the blendshapes were handcrafted from

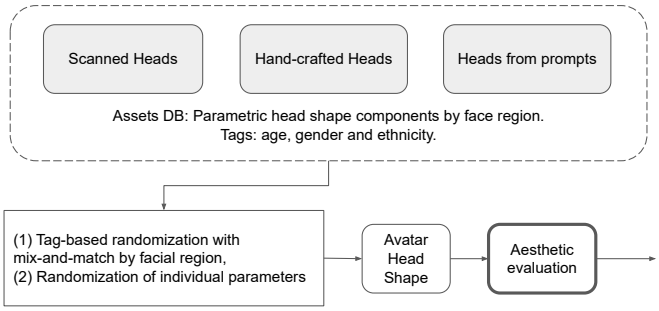


Fig. 1: Head Shapes Generation Pipeline.

anatomical data. Next, registration and tweaks of blendshapes to the scanned heads ensure their completeness, i.e., the parametric model can represent scanned shapes with the required accuracy. The resulting structure of the in-house parametric head model is similar to the FLAME model at a conceptual level.

Figure 1 shows a schematic representation of our pipeline. It generates head shapes using various data sources. The available head shape assets feed into a library of components organized by facial regions and tags specifying age, gender, and ethnicity. The first generation step involves mixing and matching parts by regions with randomized weights while respecting tag consistency.

Since the source assets are limited, tag-based randomization produces an insufficient variety of shapes. To address that, we introduce the second stage, where we randomize individual parameters to create wider variations of the produced heads. A more extensive range of randomization can push head shapes into the region of "unacceptable," i.e., not following beauty standards by a notable margin. Mild randomization produces more likable heads in line with the "averageness hypothesis" [60]. Still, the variety of the produced heads is too low for the target application, and "interesting" heads are missing. In our early experiments, such heads mostly disappear when parameters clamp under approximately one standard deviation of the initial batch. The range of produced parameters depends on the initial randomization range, age, gender, and ethnicity. The described trade-off between variety and acceptability necessitates a customized approach to rating the randomized heads as "acceptable."

The final artifact of our pipeline is a set of parameters that define the shape and discrete elements of an avatar of a young age. A younger age allows us to ignore the impact of facial texture lacking wrinkles. Also, this paper does not consider makeup, facial hair, and scalp hair. The last box in Figure 1 is the focus of this paper and ensures that the produced heads are aesthetically pleasing without limiting randomization to an unnecessarily narrow range. It also prevents extreme or undesirable shapes from being offered to users in production.

V. Panel of Proxy Experts

We aim to answer the question in this section: Can we approximate professional artistic judgment with a crowd-

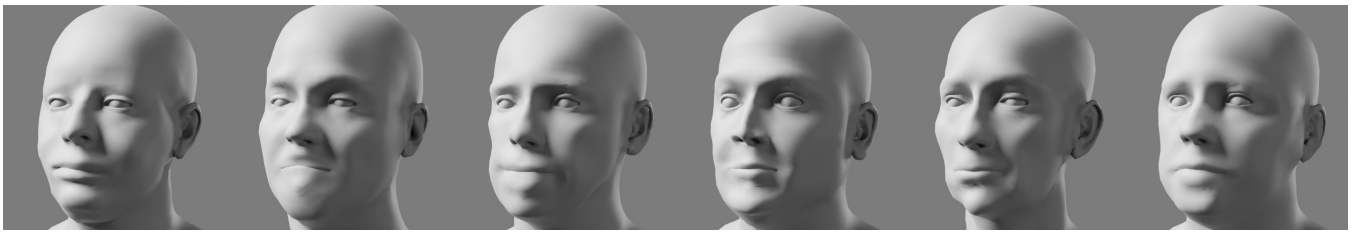


Fig. 2: Six least voted heads out of 200 with average score 0.

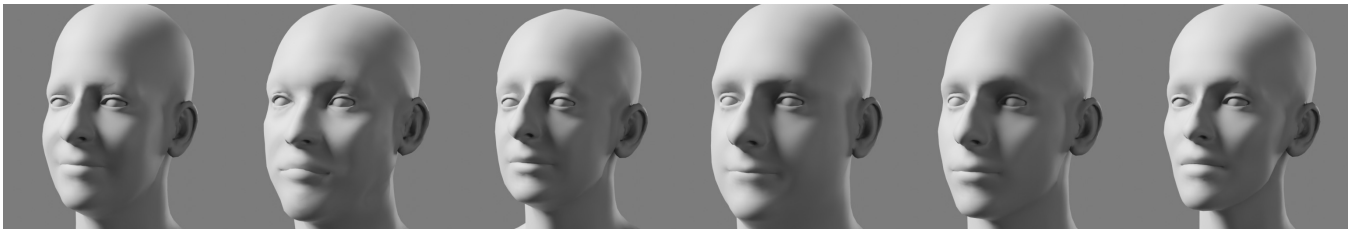


Fig. 3: Six top voted heads out of 200 with average score ≈ 1.

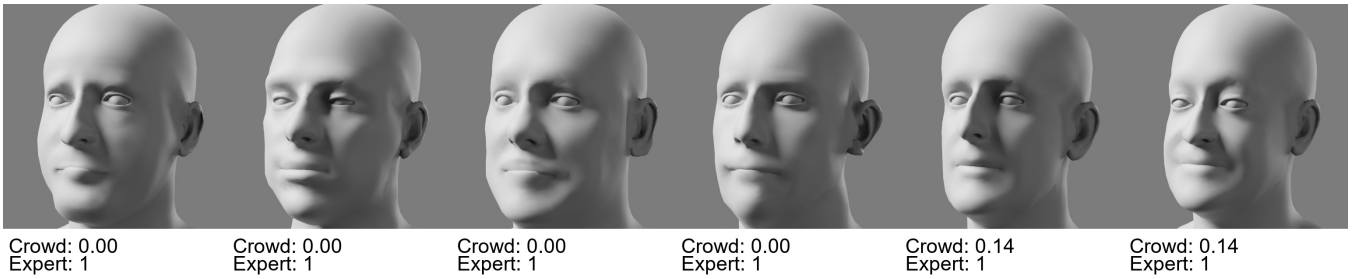


Fig. 4: Approved by expert, disliked by the crowd.

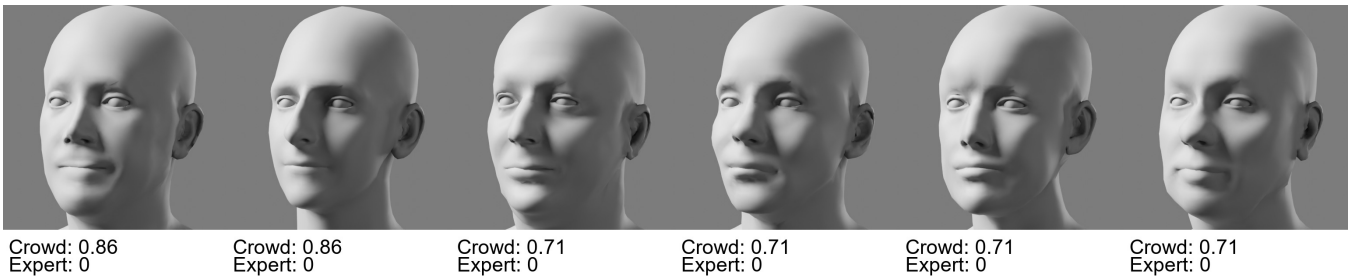


Fig. 5: Disapproved by expert, favored by the crowd.

sourced one from non-artist respondents? The motivation behind this comes from the scarcity of art direction time, while other disciplines (e.g., customer support) may be more available and willing to provide their judgment. Since their judgment is likely noisy and biased, we aim to build a robust ML model leveraging crowdsourced ratings and using art ratings as ground truth.

We start with defining the terms. In our models, the "**principal expert**" (or "**expert**" for brevity) is a single individual, usually an art director or other professional artist. The ratings from the expert comprise **ground truth**. The ground truth is what we want to predict with our models. The non-expert **respondents** in our experiments are the "**crowd**" or the "**proxy experts**." We will use these terms interchangeably. The crowd forms the "panel of proxy experts".

Next, we introduce notations. The training batch B of generated heads consists of N images $I_{i=1,\dots,N}$ rendered using identical settings from 3D models defined by known generation parameters $x_{q=1,\dots,m}$, i.e., an image is a deterministic function of generation parameters $I = f(x)$. These parameters contain floating point values and discrete choices in our production pipeline. In the FLAME experiment described here, these parameters are limited to floating point numbers defining weights for principal components of the

TABLE I: Correlations of crowd-sourced raters with the expert.

Rater	Mean	Pearson Correlation	P-value
crowd4	0.245	-0.030	0.680
fair coin x1000	0.500	0	0.490
crowd3	0.280	0.160	0.020
crowd7	0.560	0.230	0
crowd5	0.795	0.260	0
crowd1	0.500	0.350	0
crowd6	0.325	0.440	0
crowd2	0.470	0.450	0
expert (ground truth)	0.495	1	0

parametric head model.

The art director (the expert) provides "likability" binary ratings Y for the images: $Y = 0$ for "reject" and $Y = 1$ for "accept." The art director uses their subjective judgment function $F_0(I)$ to produce these values: $r = F_0(I(x))$, or $r = F_0(x)$ for brevity. Ratings comprise our first training dataset $D = (x_i, Y_i)_{i=1,\ldots,N}$ and places the problem into the classic supervised ML context. This dataset is, by definition, the ground truth.

Our first objective is to fit a model to the function $F_0(x)$. We can apply a variety of methods to that end. However, the scarcity of data points (200 in most of our experiments) and relatively high dimensionality of the parametric space (300 for the FLAME neutral head shape) suggest simpler, less powerful models to avoid overfitting. Candidates could be Logistic Regression, Random Forest, SVC, XGBoost, or other models. In this work, we rely on their implementations provided by scikit-learn [40]. Model selection may be considered a designer choice with a preference for better explainability. We use the default split of D into the training and test data to avoid overfitting the single available dataset. We conduct repeated experiments using different random seeds for the split, similar to how bagging treats the data [40], [12].

With art preferences approximated by a fit model of $F_0(x)$, we introduce k proxy experts (the regular independent respondents), indexed as $j = 1, \ldots, k$ with index 0 reserved for the principal expert. Using the same batch B of N images, we collect ratings from $j = 1, \ldots, k$ independent respondents: $D(j) = (x_i, Y_i(j))_{i=1,\ldots,N}$. Next, similarly to the expert, we train individual models $F_j(x)$ for proxy experts. These $F_j(x), j \neq 0$ approximate the preferences of the individual crowd members.

The final step is ensembling $F_{j=1,\ldots,K}$ into a single model: $E(x) = E(F_1(x) \ldots, F_k(x))$ (where we may also include x explicitly as a feature) and fit E to the ground truth on the same batch B. As with individual models, we prefer a smaller model for E with a simple structure allowing us to add or remove individual proxy experts at a low cost of re-training. Note that including $F_0(x)$ as a feature in the ensemble doesn't result in a performance better than that of the principal expert itself. However, our experiments show that the fit ensemble E may sufficiently approximate $F_0(x)$. That is unsurprising since $F_0(x)$ is the only model trained on ground truth, while proxy models use predictions from the corresponding crowd members. Such models will unlikely give a positive score to the images from Figure 4 and will likely generate False Positives for the images from Figure 5. Eliminating such disagreements from the dataset (e.g., by voting and applying a threshold) may be desirable and will likely improve the performance of the ensemble of the crowd models. We also leave this for future exploration.

Bayesian Weighted Votes [44] is an approach related to our field and showing effectiveness. It predicts ground truth from noisy votes. Adding variational aspect as in [50] may be beneficial, but we leave it to future exploration. The outline of Bayesian weighted votes is straightforward.

We aim to predict whether a head is **Good** ($Y = 1$) or **Bad** ($Y = 0$) using k voters. Each voter j produces a prediction $D \in \{0, 1\}$, where 1 indicates "Good". The voters are imperfect and characterized by their correlation $Q_j \in (0, 1)$ with the expert.

The likelihoods for each voter D_j conditioned on the true label $Y \in \{0, 1\}$ are:

$$
\begin{aligned}
P(D_j = 1 \mid Y = 1) &= Q_j \\
P(D_j = 0 \mid Y = 1) &= 1 - Q_j \\
P(D_j = 1 \mid Y = 0) &= 1 - Q_j \\
P(D_j = 0 \mid Y = 0) &= Q_j
\end{aligned}
$$

For each voter score D_j, the log-likelihood ratio is:

$$
\log\left(\frac{P(D_j \mid Y = 1)}{P(D_j \mid Y = 0)}\right) = (2D_j - 1)\log\left(\frac{Q_j}{1 - Q_j}\right)
$$

This captures whether the voter supports class 1 or class 0 and how strongly, based on its reliability.

Using Bayes' theorem, the posterior log-odds of $Y = 1$ versus $Y = 0$ given all voters outputs is:

$$
\log\frac{P(Y = 1 \mid D)}{P(Y = 0 \mid D)} = \log\left(\frac{\rho}{1 - \rho}\right) + \sum_{j=1}^{k}(2D_j - 1)\log\left(\frac{Q_j}{1 - Q_j}\right)
$$

where $\rho = P(Y = 1)$ is the prior probability of a head being good.

Decision Rule: We classify the head as "Good" (i.e., $\hat{Y} = 1$) if the posterior odds favor class 1:

$$
\hat{Y} = \begin{cases} 1 & \text{if } \log\dfrac{P(Y = 1 \mid D)}{P(Y = 0 \mid D)} > 0 \\ 0 & \text{otherwise} \end{cases}
$$

Favoring Pearson correlation with proper normalization instead of agreement measure produces results, shown in Table II. The last row of the table corresponds to Bayesian

TABLE II: The table shows the performance of individual, Bayesian Weighted Votes, and ensemble models in the panel of proxy experts. Individual respondents' models aim to predict corresponding crowd votes. The same type of classifier is used for "Ensemble" rows (third row in each cell). We omit combinations of different classifiers for brevity (e.g., SVC for individual models and Random Forest for ensembling). The table shows the mean and standard deviation of 64 experiments with different train-test splits. SVC and Random Forest achieve the best FPR with reasonable accuracy.

Classifier	Ensembling	Accuracy	Precision	Recall	FPR
Logistic Regression	Mean of crowd models	$0.56_{\pm 0.06}$	$0.55_{\pm 0.08}$	$0.73_{\pm 0.11}$	$0.61_{\pm 0.11}$
	Bayesian Weighted Votes	$0.73_{\pm 0.02}$	$0.73_{\pm 0.03}$	$0.73_{\pm 0.02}$	$0.27_{\pm 0.04}$
	Ensemble	$0.59_{\pm 0.08}$	$0.56_{\pm 0.09}$	$0.77_{\pm 0.10}$	$0.57_{\pm 0.13}$
	Expert	$0.68_{\pm 0.08}$	$0.64_{\pm 0.09}$	$0.80_{\pm 0.10}$	$0.43_{\pm 0.12}$
Random Forest	Mean of crowd models	$0.70_{\pm 0.06}$	$0.82_{\pm 0.11}$	$0.52_{\pm 0.10}$	$0.12_{\pm 0.07}$
	Bayesian Weighted Votes	$0.76_{\pm 0.02}$	$0.81_{\pm 0.03}$	$0.68_{\pm 0.03}$	$0.15_{\pm 0.03}$
	Ensemble	$0.79_{\pm 0.06}$	$0.96_{\pm 0.05}$	$0.60_{\pm 0.11}$	$0.02_{\pm 0.03}$
	Expert	$0.79_{\pm 0.06}$	$0.97_{\pm 0.05}$	$0.60_{\pm 0.11}$	$0.02_{\pm 0.03}$
XGBoost	Mean of crowd models	$0.63_{\pm 0.07}$	$0.66_{\pm 0.11}$	$0.59_{\pm 0.10}$	$0.31_{\pm 0.11}$
	Bayesian Weighted Votes	$0.75_{\pm 0.02}$	$0.79_{\pm 0.03}$	$0.69_{\pm 0.03}$	$0.18_{\pm 0.03}$
	Ensemble	$0.73_{\pm 0.06}$	$0.80_{\pm 0.12}$	$0.61_{\pm 0.12}$	$0.15_{\pm 0.10}$
	Expert	$0.76_{\pm 0.06}$	$0.81_{\pm 0.09}$	$0.66_{\pm 0.11}$	$0.15_{\pm 0.08}$
Support Vectors	Mean of crowd models	$0.70_{\pm 0.06}$	$0.84_{\pm 0.10}$	$0.51_{\pm 0.09}$	$0.10_{\pm 0.06}$
	Bayesian Weighted Votes	$0.79_{\pm 0.01}$	$0.83_{\pm 0.02}$	$0.74_{\pm 0.02}$	$0.15_{\pm 0.02}$
	Ensemble	$0.79_{\pm 0.06}$	$0.97_{\pm 0.04}$	$0.60_{\pm 0.11}$	$0.02_{\pm 0.03}$
	Expert	$0.79_{\pm 0.06}$	$0.97_{\pm 0.04}$	$0.60_{\pm 0.11}$	$0.02_{\pm 0.03}$
Weighted Bayesian Voting	Ratings as votes	$0.76_{\pm 0.06}$	$0.77_{\pm 0.09}$	$0.72_{\pm 0.09}$	$0.21_{\pm 0.10}$

weighted votes ensembling directly from the crowd votes. Like ground truth, these votes are unavailable outside of the training-test dataset. Hence, we replace them with trained models with results presented in the rest of Table II.

With models $F_0, F_1, \ldots, F_k$ and the ensemble E trained, we apply the prediction from E in the production pipeline as the final step to decide if a particular head can be shipped with the product. Finally, before the deployment in production, we may apply bagging [12] to the models trained in our experiments with different subsampling of the training data. Another way to use the trained models is to filter out rejected heads with the proxy experts ensemble and then pass the remaining heads to the expert model for the next pass. Finally, the human expert may review a small sample to ensure a low rate of False Positives. Such a workflow is reminiscent of the general idea of boosting, where weak learners can ensemble sequentially to improve performance.

The shift in facial feature distribution between the batches of images may invalidate the models. A change of art direction may also degrade the ensemble's performance. Either of these events may require a new round of scoring and re-training.

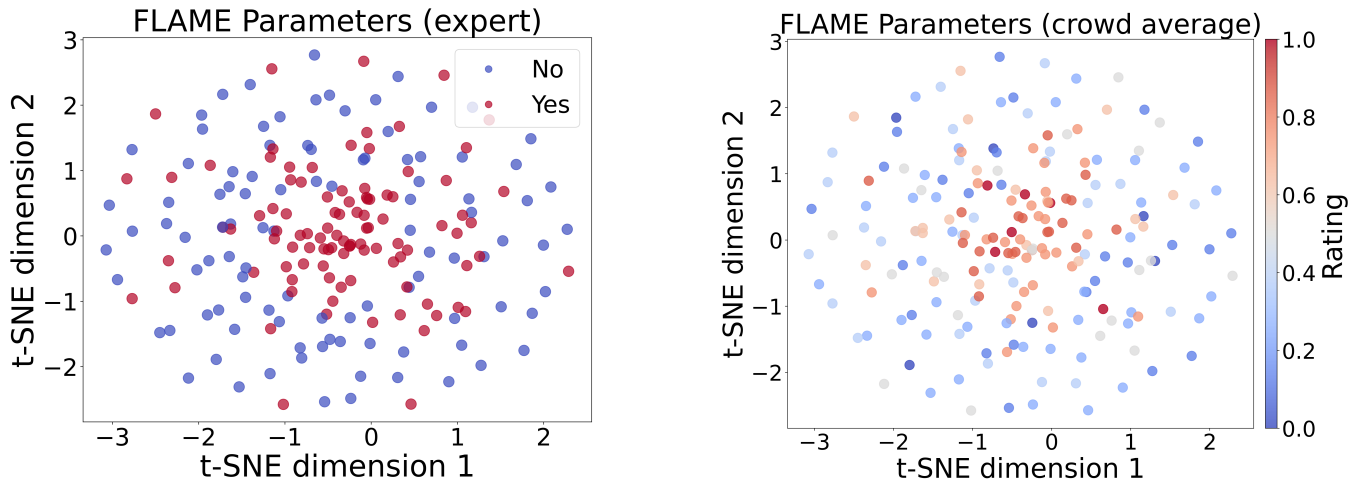
To summarize this section: We aim to capture the art direction's preferences with crowdsourced input and use it instead of or together with the expert model.

VI. Experiment Setup

In our experiment, an artist rates 200 images generated with FLAME by randomizing shape parameters within a relatively large range of values and with asymmetry constraints partially tuned down. For each portrait, the rater has to answer "yes" or "no", indicating whether the head is "likable". Our experiment's principal expert providing ground truth is a professionally trained, well-recognized artist. The heads dataset is balanced, and the mean rating from the art expert is 0.495.

Fig. 6: Web-based Rating tool: the instructions are minimal. The placeholders "Generic" and "Any Gender" allow more specific questionary for use production.

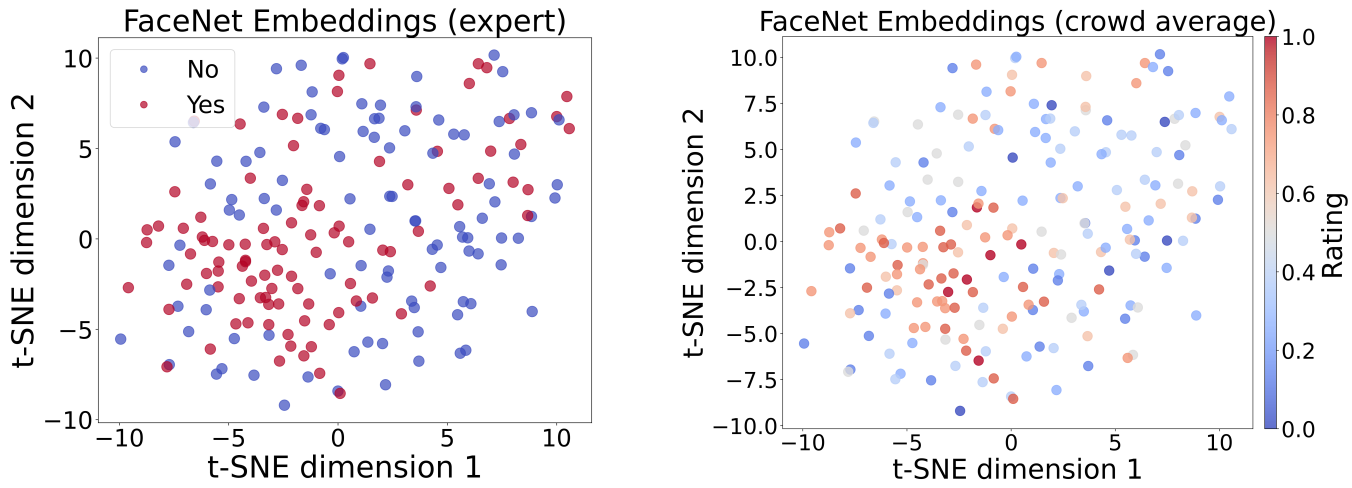
Next, we invite seven proxy experts who rate the same



(a) t-SNE visualization of the expert ratings

(b) t-SNE visualization of the crowd average ratings

Fig. 7: Visualizations of generation parameters embeddings showing (a) expert ratings and (b) crowd average ratings.



(a) t-SNE visualization of the expert ratings

(b) t-SNE visualization of the crowd average ratings

Fig. 8: Visualizations of FaceNet embeddings showing (a) expert ratings and (b) crowd average ratings.

heads. The proxy experts' backgrounds include engineering and administration, with no professionally trained artists in the "crowd." The correlations Table I shows their agreement with the expert ratings.

The proxy experts do not receive detailed instructions before providing scores, see Figure 6 showing a minimalist interface. Moreover, they are asked not to "overthink" and use their intuitive judgment instead of analyzing image details. This way, we aim to capture the subjectivity of the crowd's judgment. Figure 4 and Figure 5 illustrate the difference in the evaluation of the heads where the scores of the expert and the crowd are in disagreement. Note that the order of image presentation may affect the ratings: if there is a long stretch of potential rejections, the subjective quality standard may drop and introduce undesirable bias. We randomly shuffle images for each new respondent for our experimentation to mitigate such effects. The effect is likely measurable, and we plan to explore it in future work.

Since we have a balanced dataset, we use a fair coin toss sampled a thousand times as the baseline; see row 2 in Table I. Interestingly, one responder (crowd4, see the first row) shows a slightly negative correlation, voting 0 for most heads, hence a larger P-value. Such imbalanced ratings affect training classifiers: all predicted values are zeros in most data splits for training and tests, invalidating the predictions of several classifiers. The crowd member crowd3 is similarly problematic. For this reason, we exclude both crowd3 and crowd4. A low correlation with the expert and relatively high P-value may be an early indication for excluding proxy experts from further exploration as their ratings were poor predictors. However, we can exploit a strong negative correlation. Our small-scale experiment doesn't encompass such respondents.

For illustration, we include only the most and least-liked heads from the entire dataset of images: Figure 3 and Figure 2 correspondingly. These extremes illustrate how the heads

look like in the rest of the dataset. The less-liked heads show more ragged features, while the top-rated ones appear smoother with well-correlated features.

Table II summarizes the experiment's results. Since the dataset is small, we use multiple independent random splits for training and test data following the standard procedure implemented with default parameters in scikit-learn. While that doesn't guarantee statistical significance, it suggests that the results can be similar in more extensive experiments.

The observations are common across the model selection. The Random Forest and SVC models of the expert $F_0(x)$ achieved performance that exceeds the random guess baseline in the data's 80-20 percent split. But Logistic Regression did not generalize well. A random guess with the probability matching ground truth distribution would result in a False Positive Rate (FPR) close to 0.5. Approximating the ground truth probability (0.495, see Table I) of ones as 0.5, we get:

$$FPR = \frac{FP}{FP + TN} = \frac{0.25}{0.25 + 0.25} = 0.5$$

Mean performance of the individual models $\overline{F_i(x)}, i = 1, \dots\ k$ trained for the crowd is not as good as for the model $F_0(x)$ of the expert. Ensembling individual models of the crowd $E(F_1(x)) \dots, F_k(x))$ improves performance. For Random Forest and Support Vectors Classifiers, the metrics are on par with the model of the expert, supporting our initial guess that crowdsourcing may offer an alternative to the expert.

An insight into the low performance of Logistic Regression comes from Figure 7. We produce clusters in the generation parameters space with t-SNE [23], [53] as implemented in [40]. Highly-rated heads form a distribution with a single mode, similar to Gaussian. The rejected heads form a ring around the accepted heads. Linear decision boundaries for such data can not deliver good predictions. As expected, non-linear classifiers show much better performance in this case. The t-SNE diagrams support the "averageness hypothesis" but do not trivialize the problem of rating heads.

One aspect of the non-triviality consideration comes from the technical art choice of the neutral head, described by all zeroes in the generation parameters space. The zero point of the PCA space in FLAME is one generic candidate. However, art may prefer a different neutral head to better cater to the concrete application and to make the best use of the in-house data. The center of the accepted heads cluster may not necessarily coincide with such a zero.

The second aspect is that we are not solving the classic FBP problem. The objective of keeping "interesting heads" may introduce additional structure to the distribution of accepted heads. Such a structure may be directly interpretable with techniques applied to embeddings [41]. To that end, various embedding may provide additional insights. FaceNet embeddings capture facial features directly in a non-linear space and may indicate the "interesting heads" structure independent of the design zero point; see Figure 8. Moving to a higher dimension of FaceNet embedding (512) may appear to be overkill compared to FLAME parameters (300 for the static head). However, t-SNE visualization may be helpful in future experiments with larger parametric models.

One of our objectives is a low FPR since preventing the approval of "unlikable" faces is more critical than mistakenly rejecting "likable" ones, aligning with our experimental design to prioritize user satisfaction and accuracy. From that perspective, the lowest FPR comes from the expert model and ensemble trained with Random Forest and Support Vectors, which is superior to ensembling with Bayesian Weighted Votes. With low FPR, we can also envision a hybrid workflow using the crowd model to eliminate the extremely bad heads and then send the crowd-accepted heads to the art director for final approval. This would mean that they would have a smaller batch to review, saving their (valuable) time.

VII. Conclusion and Future Work

We plan to run broader scale rating experiments to obtain additional data points and uncover structures beyond the simple "good" Gaussian centered at zero. Such experiments may also help train a model for the crowd, potentially replacing the expert without asking them to rate images unless art direction requires changes. The t-SNE diagrams and the results table show promising agreement between the accepted and rejected heads with the expert rating. Clustering the majority using voting may be sufficient. More sophisticated methods can emerge from a signal-processing approach, where we aim to recover noisy signals (expert ratings) from biased noise observations by the crowd.

Synthetic faces of high quality with properly correlated features are valuable as training data for various face reconstruction models. Applying our technique to filter randomly generated heads intended for such purpose can improve the quality of the produced data. One positive outcome is a reduced distribution shift in applications to authentic faces in the wild. The "acceptability filter" will substantially reduce the number of produced heads. Still, we can address that with established methods (reviewed in [55]) and with more recent techniques geared towards face reconstruction [3]. We plan to explore this subject more deeply in the future research.

In conclusion, with subjective evaluation geared towards a concrete product, relying on widely accepted models of likability and mainstream FBP is hard. Restrictive or unclear licenses of the public datasets make the field challenging to commercial applications when relying on publicly available data and models. With the simple approach proposed in this paper, we expect to address the issue by training models from the ratings provided by the art director and the proxy experts employed by crowdsourcing. The proposed approach carries the bias introduced by the proxy experts from the same company or industry. The positive impact of such a bias may allow us to uncover the structure of the manifold of "interesting heads" that do not follow the metrics of classic FBP. That would require larger-scale experimentation, which we plan for future work.

REFERENCES

[1] Adobe. Adobe firefly - generative ai for creativity, 2024. Accessed: 2024-03-05.

[2] V. R. Ahuja, A. Ahuja, and N. Thosar. Evaluation and comparison of facial appearance using the golden ratio: An anthropometric study in preschool and school-going children of santhal tribe in west bengal. *Cureus*, 16, 2024.

[3] A. Atzori, F. Boutros, N. Damer, G. Fenu, and M. Marras. If it's not enough, make it so: Reducing authentic data demand in face recognition through synthetic faces. *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10, 2024.

[4] J. Bae, S.-J. Buu, and S. Lee. Anchor-net: Distance-based self-supervised learning model for facial beauty prediction. *IEEE Access*, 12:61375–61387, 2024.

[5] M. W. Baxter and D. Walker. Are you "hot or not"? *Atlantic Economic Journal*, 36:367–368, 2008.

[6] I. C. Bernal, J. Andre, M. Patel, and M. I. Newman. Beauty re-defined: A comparative analysis of artificial intelligence-generated ideals and traditional standards. *Cureus*, 16, 2024.

[7] D. Bhatnagar and T. Gupta. Golden ratio and facial beauty with computer vision. *Journal of Analysis and Computations*, 2023.

[8] Z. Bin, L. Yong, X. Li, L. Dan, L. Zihao, and X. Sun. A review of research on 3d face reconstruction methods. *Proceedings of the 2024 9th International Conference on Intelligent Information Technology*, 2024.

[9] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 1999.

[10] I. Borovikov, K. Levonyan, and M. Anghelescu. Practical parametric synthesis of realistic pseudo-random face shapes. *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–7, 2023.

[11] D. E. Boukhari. Facial beauty prediction based on vision transformer. *International Journal of Electrical and Electronic Engineering & Telecommunications*, 2024.

[12] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[13] T. J. Burleigh, J. R. Schoenherr, and G. L. Lacroix. Does the uncanny valley exist? an empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers in Human Behavior*, 29(3):759–771, 2013.

[14] H. Chen, W. Li, X. Gao, and B. Xiao. Novel multi-feature fusion facial aesthetic analysis framework. *IEEE Transactions on Big Data*, 9:1302–1320, 2023.

[15] Y. Chen, H. Mao, and L. Jin. A novel method for evaluating facial attractiveness. *2010 International Conference on Audio, Language and Image Processing*, pages 1382–1386, 2010.

[16] H. Dai, N. E. Pears, and W. Smith. A data-augmented 3d morphable model of the ear. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 404–408, 2018.

[17] B. Egger, W. A. P. Smith, A. K. Tewari, S. Wuhrer, M. Zollhöfer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, C. Theobalt, V. Blanz, and T. Vetter. 3d morphable face models - past, present and future. *ArXiv*, abs/1909.01815, 2019.

[18] Y. Eisenthal, G. Dror, and E. Ruppin. Facial attractiveness: Beauty and the machine. *Neural Computation*, 18:119–142, 2006.

[19] Epic Games. Metahuman creator - high-fidelity digital humans, 2024. Accessed: 2024-03-05.

[20] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Lüthi, and T. Vetter. Morphable face models - an open framework. *Proceedings of the 13th IEEE International Conference on Automatic Face Gesture Recognition (FG)*, pages 75–82, 2018. Basel Face Model 2017 (BFM17).

[21] D. Gray, K. Yu, W. Xu, and Y. Gong. Predicting facial beauty without landmarks. In *European Conference on Computer Vision*, 2010.

[22] H. Harrar, S. Myers, and A. Ghanem. Art or science? an evidence-based approach to human facial beauty a quantitative analysis towards an informed clinical aesthetic practice. *Aesthetic Plastic Surgery*, 42:137 – 146, 2018.

[23] G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. In *Neural Information Processing Systems*, 2002.

[24] IEEE. *A 3D Face Model for Pose and Illumination Invariant Face Recognition*, Genova, Italy, 2009.

[25] E. Imre and A. Yılmaz. The effect of the golden ratio in facial anatomy on beauty perception. *Anatomy*, 2024.

[26] K. A. Iteire, F. Chukwudebe, V. O. Ukwenya, F. Johnson, R. Uwejigho, and F. Enemali. Conceptualization of facial beauty among female students in a southwestern nigerian university using the golden ratio model. *Nigerian Journal of Experimental and Clinical Biosciences*, 10:81 – 89, 2022.

[27] T. J. Iyer, R. K, R. Nersisson, Z. Zhuang, A. N. J. Raj, and I. Refayee. Machine learning-based facial beauty prediction and analysis of frontal facial images using facial landmarks and traditional image descriptors. *Computational Intelligence and Neuroscience*, 2021, 2021.

[28] M. jung Sun, D. Yang, D. Kou, Y. Jiang, W. W. Shan, Z. Yan, and L. Zhang. Human 3d avatar modeling with implicit neural representation: A brief survey. *2022 14th International Conference on Signal Processing Systems (ICSPS)*, pages 818–827, 2022.

[29] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017.

[30] L. Liang, L. Lin, L. Jin, D. Xie, and M. Li. Scut-fbp5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1598–1603, 2018.

[31] J. Lin, Y. Yuan, and Z. Zou. Meingame: Create a game character face from a single portrait. *ArXiv*, abs/2102.02371, 2021.

[32] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[33] R. Loucas, B. Sauter, M. Loucas, S. Leitsch, O. Haroon, A. Macek, S. Graul, A. Kobler, and T. Holzbach. Is there an "ideal instagram face" for caucasian female influencers? a cross-sectional observational study of facial proportions in 100 top beauty influencers. *Aesthetic Surgery Journal. Open Forum*, 6, 2024.

[34] Y. Men, B. Lei, Y. Yao, M. Cui, Z. Lian, and X. Xie. En3d: An enhanced generative model for sculpting 3d humans from 2d synthetic data. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9981–9991, 2024.

[35] S. Mittal, G. Aneja, A. Mittal, P. H. Teja, M. Gagain, and A. Verma. Comparison of facial attractiveness with golden proportion anthropometrically in young north indian females. *International Dental Journal of Student's Research*, 2024.

[36] S. M. Moon and J. K. Min. A study on the visual realism of digital humans and the uncanny valley phenomenon. *Journal of The Korean Society of Illustration Research*, 2024.

[37] E. Nguyen, S. E. Akwafuo, D. Bein, and B. Ojeme. Racially inclusive approach to facial beauty modeling using machine learning. *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4467–4473, 2024.

[38] Ö. Özgür, L. C. Mazlum, and B. S. Kızılok. Beauty beyond the golden ratio: A study of perception regarding facial proportions and symmetry in the turkish population. *Gazi Medical Journal*, 2025.

[39] P. M. Pallett, S. W. Link, and K. Lee. New "golden" ratios for facial beauty. *Vision Research*, 50:149–154, 2010.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[41] R. Plesh, J. Kriaj, K. Bahmani, M. Banavar, V. truc, and S. Schuckers. Discovering interpretable feature directions in the embedding space of face recognition models. *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2024.

[42] S. Ploumpis, E. Ververas, E. O. Sullivan, S. Moschoglou, H. Wang, N. E. Pears, W. Smith, B. Gecer, and S. Zafeiriou. Towards a complete 3d morphable model of the human head. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:4142–4160, 2019.

[43] S. Ploumpis, H. Wang, N. E. Pears, W. Smith, and S. Zafeiriou. Combining 3d morphable models: A large scale face-and-head model. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10926–10935, 2019.

[44] V. C. Raykar, S. Yu, L. H. Zhao, G. Hermosillo, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *J. Mach. Learn. Res.*, 11:1297–1322, 2010.

[45] Ready Player Me. Ready player me documentation, 2024. Accessed: 2024-03-05.

[46] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[47] M. Seymour, K. Riemer, and J. Kay. Interactive realistic digital avatars - revisiting the uncanny valley. In *Hawaii International Conference*

on System Sciences, 2017.
[48] M. Seymour, L. I. Yuan, A. R. Dennis, and K. Riemer. Have we crossed the uncanny valley? understanding affinity, trustworthiness, and preference for realistic digital humans in immersive environments. *J. Assoc. Inf. Syst.*, 22:9, 2021.
[49] K. Shah and M. Patel. Assessment of facial golden proportions in gujarati population – a retrospective study. *The Journal of Dental Panacea*, 2023.
[50] E. Simpson, S. J. Roberts, and C. J. Lintott. Bayesian combination of multiple , imperfect classifiers. 2011.
[51] N. Sultan. A study on an automatic system for analyzing the facial beauty of young women. 2014.
[52] L. Tran and X. Liu. Nonlinear 3d face morphable model. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7346–7355, 2018.
[53] L. van der Maaten and G. E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
[54] D. Vincent, J. Nazar, G. V. Abraham, and R. M. Philip. Beauty beyond numbers: The golden ratio and facial aesthetics. *The Journal of Dental Panacea*, 2024.
[55] P. K. Vinodkumar, D. Karabulut, E. Avots, C. Ozcinar, and G. Anbarjafari. Deep learning for 3d reconstruction, augmentation, and registration: A review paper. *Entropy*, 26, 2024.
[56] J. Wang, Y. Gong, and D. Gray. Female facial beauty attribute recognition and editing. In *Human-Centered Social Media Analytics*, 2014.
[57] D. Xie, L. Liang, L. Jin, J. Xu, and M. Li. Scut-fbp: A benchmark dataset for facial beauty perception. *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1821–1826, 2015.
[58] C.-T. Yang, Y.-C. Wang, L.-J. Lo, W.-C. Chiang, S.-K. Kuang, and H.-H. Lin. Implementation of an attention mechanism model for facial beauty assessment using transfer learning. *Diagnostics*, 13, 2023.
[59] D. Zhang. Facial beauty analysis system by computer models. 2017.
[60] D. D. Zhang, F. Chen, and Y. Xu. Computer models for facial beauty analysis. In *Cambridge International Law Journal*, 2016.
[61] J. Zhao, F. Deng, J. Jia, C. Wu, H. Li, Y. Shi, and S. Zhang. A new face feature point matrix based on geometric features and illumination models for facial attraction analysis. *Discrete & Continuous Dynamical Systems - S*, 2019.
[62] Y. Zhuang, Y. He, J. Zhang, Y. Wang, J. Zhu, Y. Yao, S. Zhu, X. Cao, and H. Zhu. Towards native generative model for 3d head avatar, 2024.