Search for Extraordinary Experiences Division

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
[2] Electronic Arts, SEED, Vancouver, BC, Canada
[3] Electronic Arts, SEED, Montreal, QC, Canada

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

Motivation

- Modern video games require **scale**
  - require 10,000+ generated **diverse, believable**, character heads

Motivation

- Modern video games require scale
  - require 10,000+ generated diverse, believable, character heads
- Character heads are challenging
  - Manual curation or quality validation (QV) of thousands of heads is not practical

Motivation

- Modern video games require scale
  - require 10,000+ generated diverse, believable, character heads
- Character heads are challenging
  - Manual curation or quality validation (QV) of thousands of heads is not practical
- **Can we automate QV?**
  - Reduce workload while maintaining consistency and quality

Motivation

- Modern video games require scale
  - require 10,000+ generated diverse, believable, character heads
- Character heads are challenging
  - Manual creation or quality validation (QV) of thousands of heads is not practical
- **Can we automate QV?**
  - Reduce workload while maintaining consistency and quality
- Need to define “quality” or “acceptability” of heads (QV criteria)
  - Aesthetic acceptability is context-dependent

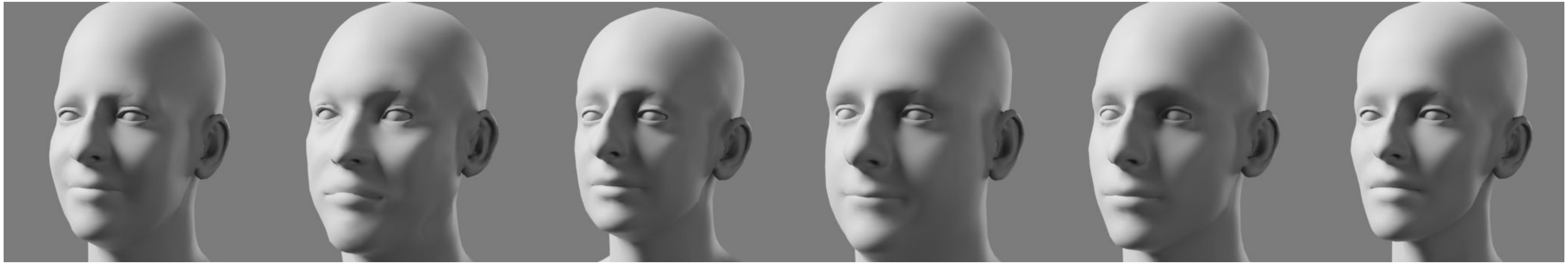
What “acceptability” is not:

🚫 No traditional Facial Beauty Prediction (FBP)

🚫 Not About “Uncanny Valley”: we enforce “consistent style” not realism

✅ **Do you “like” these heads or not?**

[*The heads below do not represent any product or art style, see Experiment Setup*]



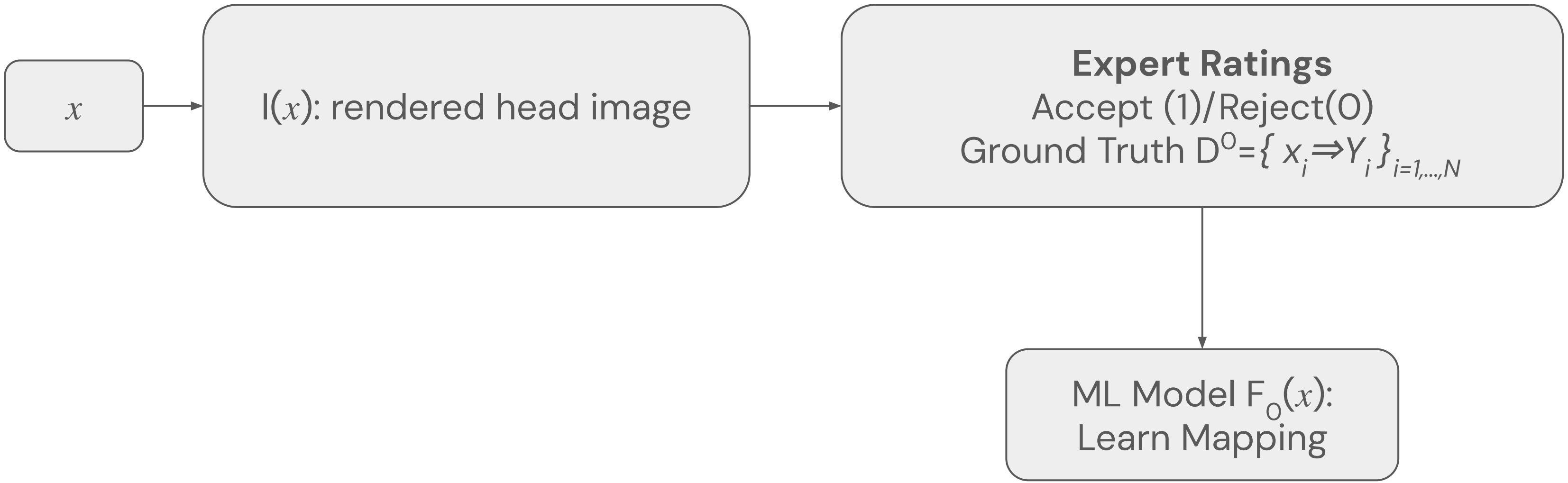
Why automate?

- **Scale**: large number of heads to evaluate (many thousands)
- **High dimensionality** of the parametric space for the head model (>600)
- **Art directors** time is limited and can't validate each head

Need a Scalable Solution

Proposed Solution. Step 1.

Goal is to build an ML model that mimics the art director accept/reject ratings



Result: the model F_0 approximating the art director

Proposed solution. Step 2.

Learn from Proxy Experts via Ensemble Modeling

1. Crowdsourced rating
   - Internal non-experts (proxy experts or “crowd”, e.g., engineers, sales, …) rate the same heads
2. Filter for reliability
   - Exclude proxy experts with low correlation with art direction.
   - The remaining experts $j=1,\dots,k$ correspond to k datasets $D^k=\{ x_i \Rightarrow Y_i \}^k_{i=1,\dots,N}$
3. Train & Ensemble
   - *Train k* ML models predicting preferences of the “crowd” members $F_j(x)$, $j=1,\dots,k$.
   - *Combine via* ensemble models F_j to predict expert ratings (ground truth): $E(x)=E(F_1(x),\dots,F_k(x))$
4. ***Result***
   - **The ensemble *E* approximates F_0 (the art director)**

Proposed solution. Step 2.

Learn from Proxy Experts via Ensemble Modeling

1. Crowdsourced rating
   - Internal non-experts (proxy experts or “crowd”, e.g., engineers, sales, ...) rate the same heads
2. Filter for reliability
   - Exclude proxy experts with low correlation with art direction.
   - The remaining experts $j=1,\dots,k$ correspond to k datasets $D^k=\{ x_i \Rightarrow Y_i \}^k_{i=1,\dots,N}$
3. Train & Ensemble
   - *Train k* ML models predicting preferences of the “crowd” members $F_j(x)$, $j=1,\dots,k$.
   - *Combine via* ensemble models F_j to predict expert ratings (ground truth): $E(x)=E(F_1(x),\dots,F_k(x))$
4. ***Result***
   - ***The ensemble E* approximates F_0 (the art director)** **← scalable, low-cost QV**

Where is the gain?

- High Fidelity to Expert Judgment
  - On the original heads, the proxy ensemble *E* approximates the expert quite well (low FPR comparable to F_0).
- Accuracy Filtering of Acceptable/Rejectable Heads
  - Using crowd ensemble *E* and/or expert model F_0 on the remaining heads produces the required ratings “accept” or “reject”
- Generalization to New Heads
  - A new dataset of heads generated without changing the art direction or generation pipeline can be rated in a similar manner via E(x) and/or F_0
- Retrain and Reuse:
  - With notable changes in art direction or generation pipeline, we can retrain crowd models $F_j(x)$, $j=1,\dots,k$ and feed them into the *E* ensemble to operate until the art direction provides new rating “ground truth”. After that, we retrain the ensemble *E* only

Experiment setup

- Head Generation
  - FLAME model with only 60 parameters selected for randomization
  - 200 heads with ¾ portraits rendered in Blender
  - Grey scale, no texture, no scalp or facial hair
  - Natural setup: gender, age, and ethnicity agnostic setup

- Rating Protocol
  - Expert: single art director provides ground thrush
  - Crowd: 7 proxy experts; no training, minimal instructions: “Like the image?”. 2 raters removed for low expert correlation

- Modeling & Evaluation (expert and ensemble models)
  - Logistic Regression (baseline, poor performance), XGBoost, Random Forest, SVC
  - Also tested Weighted Bayesian Votes
  - Repeated 64 times random train-test splits to average results

Experiment results

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	$0.76_{\pm 0.06}$	$0.77_{\pm 0.09}$	$0.72_{\pm 0.09}$	$0.21_{\pm 0.10}$

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

Low FPR achieved with SVC and Random Forest

Examples of rated images

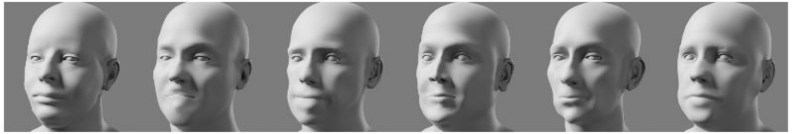


Fig. 2: Six least voted heads out of 200 with average score 0.

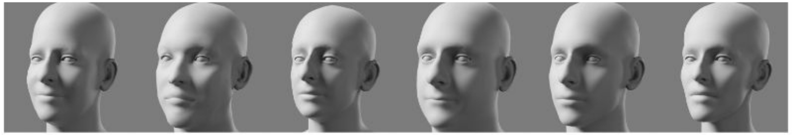


Fig. 3: Six top voted heads out of 200 with average score ≈ 1.

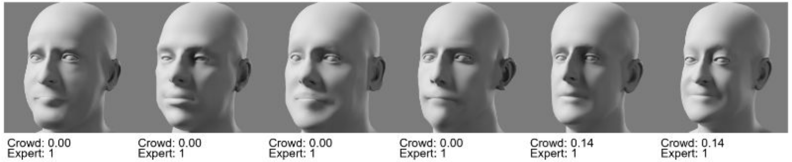


Fig. 4: Approved by expert, disliked by the crowd.

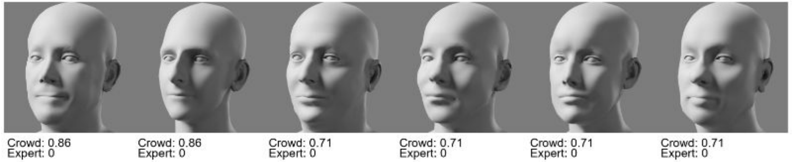


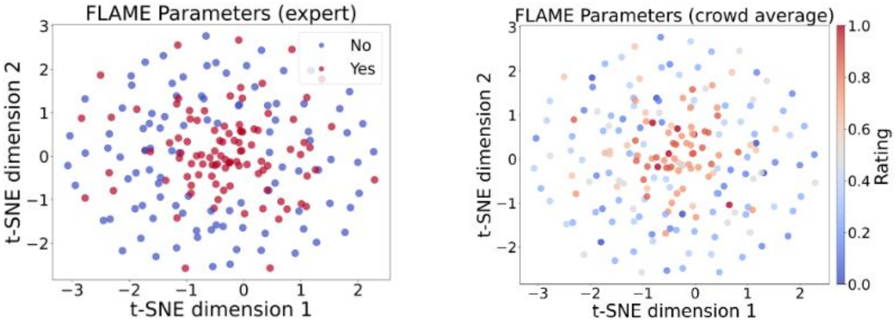
Fig. 5: Disapproved by expert, favored by the crowd.

- Ratings of many images agree between the expert and the selected crowd
- Extreme shapes are universally rejected
- Some shapes represent “interesting” but not necessarily “beautiful” heads
- That suggests that “averganess” criteria doesn’t apply directly to acceptability.

Takeaway: Acceptability ≠ Beauty

“Averageness” bias (central to facial beauty prediction) doesn’t fully explain what gets accepted

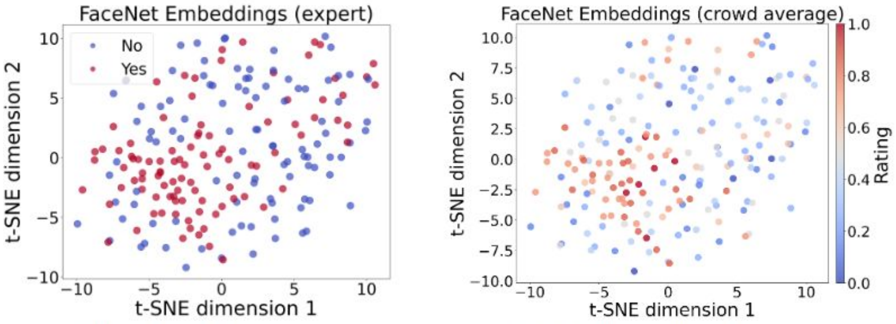
tSNE shape of the acceptable parameters region



(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.

- tSne Embedding Observations
  - Acceptable heads cluster in a dense central region
  - Rejected heads form a surrounding "donut" or ring shape (explains Logistic Regression failure)

- Future work
  - Can we uncover structure (beyond single mode Gaussian) in the "acceptable" heads shapes?

Conclusion

Context matters

- custom approach to rating "acceptability" or "likeability" of human heads

Learning from Art Direction

- We explored possibility of capturing subjective preferences of art direction with ML models

Scalable Evaluation with Proxy Experts

- In-house crowdsourcing can reduce time demand on the art direction by training proxy models and ensembling them

Q&A

Q: How do you generate thousands of head shapes?

A: One possible approach is described in our **IEEE Face and Gesture 2022 paper "Practical Parametric Synthesis of Realistic Pseudo-Random Face Shapes," Igor Borovikov, Karine Levonyan, and Mihai Anghelescu**.

The idea is to train mapping from a latent representation to the space of authoring parameters.

Artists use authoring parameters to define the shape of a head by moving sliders in a visual editor. The problem is that sliders do not enforce any correlations between parameters. Naive randomization of authoring parameters may result in "unnaturally" looking heads.

A latent space trained from real human faces, like in FaceNet, captures the relationships between features. Mapping the latent vectors to the authoring space would produce a distribution of heads with properly correlated features, allowing for the generation of a large number of natural-looking heads. As a bonus, we can control the variety of the heads by drawing samples from the latent space sufficiently far from each other.

Please refer to the paper for additional details.

Thank you for your attention!

Please feel free to reach to the authors with comments and questions.

seed.ea.com