Harnessing Quantitative Creativity to Explore the Uncanny Valley of Synthetic Faces

Aditi Ramaswamy
aditi.ramaswamy@kcl.ac.uk
and **Hana Chockler**
hana.chockler@kcl.ac.uk

Abstract

AI-generated, or synthetic, faces have proliferated across the Internet in recent years, and have become increasingly difficult to distinguish from real facial images. Even state-of-the-art synthetic image detection models have not been highly accurate, achieving F1 scores below 0.9, and generally do not follow human-interpretable logic or produce explanations that can be understood by the average Internet user. Humans also generally perform poorly at determining whether a given face is synthetic or real, and often make their decisions based on gut feeling (the "uncanny valley" effect) rather than explainable logic. However, an approach predicated on a task-oriented definition of computational creativity could provide insight into differentiating factors between synthetic and real faces that can explain the origin of the uncanny valley effect and be understood more easily by laypeople. To study this, we propose a preliminary experiment that we aim to expand into a full paper.

Introduction

Facial images synthesized by generative AI models, henceforth known as synthetic facial images, have grown increasingly common over the last few years, and have experienced a drastic spike in quality (Dunn et al. 2026). The faces produced by large models such as FLUX (Labs et al. 2025) are of such realism that at first glance they harbor no visible features that signal their AI origin.

We have found that state-of-the-art synthetic facial image detection models tend to lag behind the newest architectures, with accuracy rates dropping on more lifelike images and those produced by architectures beyond the standard Generative Adversarial Networks (GANs). To justify this claim, we ran a series of open-access synthetic image detection models on 7997 synthetic and 7000 real images taken from the following models: StarGAN (Choi et al. 2017), Taming Transformer (Esser, Rombach, and Ommer 2020), Dual Condition Diffusion (Kim et al. 2023), and FLUX (Labs et al. 2025). The real images come from the Flickr Faces High Quality (FFHQ) dataset (Karras, Laine, and Aila 2018) and Kaggle (Dhote 2024). We chose this set of images because it encompasses a large number of data points from multiple commonly used architectures, some of which have been traditionally more difficult for detection models to correctly classify (Ojha, Li, and Lee 2023). The detection models we chose—UniFD (Ojha, Li, and Lee 2023), EFFORT (Yan et al. 2025), HIFI (Guo et al. 2023), and RECCE (Cao et al. 2022). Each of these models uses a combination of frequency-domain and spatial-domain features to perform their classifications. Table 1 shows the F1 scores for each of the models on the dataset of 14997 images, with the last column showing what a random coin-toss model would score on the same dataset. As shown here, none of the models score above a 0.9, with two of them performing not much better than completely random. Additionally, all of these models had fairly low accuracy rates for datasets of highly realistic synthetic images, such as those produced by FLUX. Even UniFD and EFFORT, the higher-performing models that draw from frequency-domain information, do not do well with certain architectures, and degrade significantly in quality when simple frequency-domain distortions are applied to synthetic images.

Figure 1: On the left is F036, an image of a real individual taken from the CUHK Face Sketch Database, and on the right, an image generated using FLUX with F036 as a prompt, with strength = 0.2 and guidance = 15.

UniFD	EFFORT	HIFI	RECCE	RANDOM
0.8743	0.8135	0.6255	0.5947	0.5161

Table 1: F1 Info for UniFD, EFFORT, RECCE, and HIFI on 7000 REAL images and 7997 SYNTHETIC images

From these results, we can gather that current approaches

generally do not do particularly well at detecting high-quality synthetic images, which in turn means that a new approach is needed. Utilizing human knowledge and intuition has the potential to not only bridge this gap, but to do so in a way that allows for more easily human-interpretable detection results. This is why we are proposing an experiment that will combine previously described measures of creativity in generative AI models with the observed correlation between manual face detection and the "uncanny valley effect": we believe that quantifying aspects of uncanniness in synthetic faces could lead to more accurate and interpretable synthetic face detection results.

Background and Prior Literature

The Uncanny Valley

The term "uncanny valley effect" originally referred to the discomfort a human viewer might feel upon encountering a robot that is extremely lifelike, but still inherently recognizable as a non-human entity (Mori and Macdorman 2017). In terms of artificially generated or manipulated faces, it has been associated with a mix of hyper-realistic features and inhuman traits, as well as the concept of "category uncertainty", or the idea that a given face cannot easily be classified as real or digitally created at first glance (MacDorman and Chattopadhyay 2016). Neurological research has shown that even hyper-realistic AI-generated faces spark different brain activity in observers than real photographs of human faces, although those observers are unable to tell the two categories apart visually at an accuracy level higher than near-random, and cannot articulate any tangible cause for perceived uncanniness (Moshel, Carlson, and Grootswagers 2022). However, short of EEG imaging, which is inefficient and expensive to conduct on a large scale, few techniques have been proposed that can successfully quantify uncanniness.

Uncanniness has previously been associated with both human and AI creativity, with greater uncanniness associated with a higher reported measure of perceived creativity within generative AI output (Hwang and Jeong 2026). Hwang & Jeong's study uses human responses on an arbitrarily set numerical scale to rank perceived creativity or a given output, and predetermines ground truth creativity as part of the experimental design. Similarly, a study by Kaate et al. measuring the potentially beneficial impact of deepfaked images in representations of user personas discusses perceived uncanniness as a fundamental aspect of deepfakes (Kaate et al. 2023). This study also uses subjective measures, even acknowledging the drawbacks of this tactic and incorporating a quantitative behavioral metric as well; however, this metric also is not granular enough to specifically filter for deepfakes that are viewed as particularly uncanny by users.

In our own paper, we ask: can this potential relationship between high uncanniness and high creativity in generative AI outputs be quantified in a manner that does not depend on subjective and unstable factors, but instead relies on a mathematical definition for computational creativity? And building off this, could we potentially use these results to differentiate between real facial photographs and high-quality synthetic facial images?

Task-Oriented Creativity

To measure uncanniness, we draw from the definition of Task-Oriented Creativity (TOC) as described in (Ramaswamy, Chockler, and Navaratnarajah 2025), which sets out formulae to measure three components of creativity as derived from prior computational creativity research. As stated in this paper, the definitions are predicated on the following terms: M is the generative model, A_O is the artifact set of a given output image O_M (generated by M), and A_I is the artifact set of the input prompt I, which can be an image or a text.

Definition 1 (Satisfaction of Prompt Requirements) *A generated image O_M is said to be* satisfying requirements *of I if $A_I \subseteq A_O$. In other words, all artifacts in I are present in O_M.*

Definition 2 (Cohesion) Cohesiveness $C(O_M)$ *of an output image O_M is defined wrt the input I. Given the set A_I of the artifacts in I and the set A_O of the artifacts in O_M, let P be the set of all pairs (a, b), where $a \in A_O \setminus A_I$ and $b \in A_I$. Furthermore, let $M_N = \max_{(c,d)\in A_O\setminus A_I}((\hat{c}) \cdot (\hat{d}))$. Then, cohesiveness of O_M is defined as*

$$C(O_M) = \frac{\sum_{a:(a,b)\in P} \max_{b:(a,b)\in P}((\hat{a}) \cdot (\hat{b})) + M_N}{|P| + 1}.$$

Definition 3 (Novelty) Novelty $D(O_M)$ *of an output image O_M is defined as the proportion of new artifacts in the image, if O_M contains all artifacts of I. In other words, $D(O_M)$ is $\frac{|A_O \setminus A_I|}{|A_O|}$ if $A_I \subseteq A_O$ and is 0 otherwise.*

Revisited Definitions

The original (Ramaswamy, Chockler, and Navaratnarajah 2025) paper dealt with simple sets of macro-artefacts like "apple pie", meaning that additions and subtractions to these sets would be easily to detect both manually and with an object detection model. However, this does not necessarily apply to facial images. If our aim is to differentiate real photographs from highly realistic synthetic images, predicating our results on the mere presence of large facial landmarks like "nose" would be pointless, as even low-quality synthetic facial images tend to contain facial landmarks. Instead, we must use more fine-grained measures tailored to pick up subtle differences in a practical setting.

Definition 1 is general enough to hold for a synthetic facial image study, although we must be specific about the set of artefacts to include so we can eliminate very low-quality facial images and focus on higher-quality facial images, as seen in modified Definition 4:

Definition 4 (Satisfaction of Facial Requirements) *Given a facial image F and the entire set of 68 facial landmarks F_L, F is said to be* satisfying requirements *if $F_L \subset F$. (Jabberi et al. 2023).*

The simplest way to measure this in a large experimental setting is to use DeepFace or another facial extraction library—if the chosen model finds a human face, it automatically meets the threshold. In a user study setting, an image

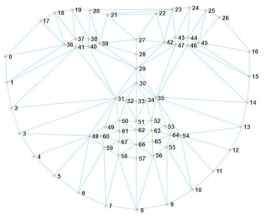

Figure 2: The 68 facial landmarks that are needed for efficient facial detection (Jabberi et al. 2023).

would pass this requirement if it is determined to be a facial image by human observers.

The (Ramaswamy, Chockler, and Navaratnarajah 2025) definition of cohesion relies on the presence of observable artefacts that can be detected by a model and have a quantifiable semantic relationship, as measured through the cosine similarity various artefact labels. However, in terms of generative AI, an alternate approach that is far more granular would be to use a spatial autocorrelation algorithm to measure the relationships between components of an image on a pixel-by-pixel basis. As further support for this, we noted that minor discrepancies have been observed between synthetic and real facial images within the spatial domain, and have been used by numerous synthetic detection models with varying degrees of success (Wang et al. 2020; Guo et al. 2023; Cao et al. 2022). These discrepancies are not visible to the naked eye, but involve hidden pixel patches, clusters, and disjointed areas that remain as products of the generative process, particularly for Generative Adversarial Network (GAN)-produced images.

The two primary spatial autocorrelation algorithms are Moran's I and Geary's C, but the former has been shown to perform better for multivariate data (Lin 2023). For a given image X of size N pixels, with a corresponding weight matrix w, Moran's I is calculated according to Equation 1, where $\bar{X}$ represents the mean of the image. A high Moran's I score indicates a high level of spatial autocorrelation within an image, while a low score indicates more disjointed areas interspersed throughout an image.

$$I(X) = \frac{N}{\sum_{i=0,j=0}^{N} w_{ij}} \times \frac{\sum_{i=0}^{N}\sum_{j=0}^{N} w_{ij}(X-\bar{X})}{(X-\bar{X})^2} \quad (1)$$

Hence, we redefine cohesion as follows in Definition 5:

Definition 5 (Cohesion) *Given a facial image F, cohesion can be measured by applying $I(F)$.*

Novelty according to Definition 3 is relatively easy to measure in a setting where multiple objects are present—have any been added? However, when dealing with high-quality facial images, we must take a different approach, since large artefacts being added, such as an extra nose, would automatically render this a low-quality image that can easily be classified. Prior research suggests that synthetic images most likely to fool humans with a "hyper-real" effect fall closer within the average area of what has been called the "face-space", or the measure by which humans intuitively analyze the uniqueness of feature configurations (Miller et al. 2023). AI-generated faces are also often deemed as more "attractive" than real faces due to their perceived average-ness, and this trait in turn is more likely to dupe observers into classifying them as real photographs (Makowski et al. 2025). Therefore, we suggest that novelty of a given image X be determined by measuring facial symmetry and distance to the golden ratio using a set of key facial landmarks within X, following a proposal for measuring attractiveness by (Schmid, Marx, and Samal 2008). The higher this distance is, and the more asymmetrical the face is, the higher the novelty factor.

(Schmid, Marx, and Samal 2008) reference a series "neo-classical canons" for facial proportions, which we can use to measure the distance between a given facial image F and a hypothetical perfectly average, symmetric face. We can use a simplified version to describe this average face F, using the 68 facial landmarks shown in Figure 2:

1. nose width (W_N, the difference between Landmarks 31 and 35) = inter-ocular width (W_O, the difference between Landmarks 41 and 46)
2. mouth width (W_M, the difference between Landmarks 48 and 54) = $1.5 \times W_N$
3. face width (W_F, the difference between Landmarks 0 and 16) = $4 \times W_N$
4. nose length (L_N, the difference between Landmarks 27 and 33) = lower face length (L_{LF},, the difference between Landmarks 8 and 51)
5. symmetry: the distance between Landmarks 0 and 30 (W_1) = the distance between Landmarks 15 and 30 (W_2)

We can calculate the results of applying these criteria to the actual landmarks in F_L: for example, how far is $|W_1 - W_2|$ from 0? Then we can average these results and use that as the novelty factor, as seen in modified Definition 6:

Definition 6 (Novelty) *Given a facial image F, novelty is defined as*

$$\frac{|W_N - W_O| + |W_M - 1.5W_N| + |W_F - 4W_N| + |W_1 - W_2| + |L_N - L_{LF}|}{5}$$

The closer this value is to 0, the less "novel" (more symmetrical and proportionate) F is.

Methodology

We hypothesize that, in conjunction with (Hwang and Jeong 2026)'s discussion of perceived uncanniness and creativity, synthetic images that humans correctly classify as AI-generated will score lower on cohesion and higher on novelty, even if they satisfy the facial requirement set out in Definition 4 and there is nothing overtly "wrong" with them. Conversely, we hypothesize that synthetic images classified incorrectly as real by humans will have low novelty, meaning that they fall closer to the average face. To test this, we plan to generate 10-step chains for a small dataset of facial images,

and perform a user study wherein individuals will be asked to classify a sampling of images from each chain step as either synthetic or real. We will then revisit the reconfigured TOC measures, applying them to the datasets of correctly and incorrectly classified images in order to determine whether the uncanniness factor can explain why humans may misclassify some synthetic images but not others.

As a preliminary study, we have chosen to use a set of 188 images from the Chinese University of Hong Kong (CUHK) student database, as found on Kaggle (Zhang, Wang, and Tang 2011), along with images we have produced ourselves by using the CUHK dataset as seed image input to the FLUX.1-dev model. We have chosen this dataset primarily because it features front-facing, clear facial images with blank backgrounds, allowing our chosen generative model to focus on the faces rather than background objects.

In a full-length paper, we would have implemented the 10-step chain-style setup described in (Ramaswamy, Chockler, and Navaratnarajah 2025) on all of these images to calculate the TOC between each, but due to limited space and computational resources we have only been able to implement them on 16/188 images thus far.

To check whether the images satisfy the facial requirement, we ran them through DeepFace's facial extraction algorithm, and confirmed that all of the real and synthetic images did in fact have clear, detectable faces. This is also true from a visual assessment of the chains, an example of which can be seen in Figure 3.

Figure 3: An example chain seeded with one of the CUHK dataset images. The real seed is the leftmost image in the top row.

We also ran the PySAL (Rey and Anselin 2010) implementation of Moran's I on the images from each chain step, and compared them to each other using the Brunner-Munzel test (Brunner and Munzel 2000), a more sensitive variation of the commonly used nonparametric two-category Mann Whitney U test. We found that, although the Moran's I value for each step did not differ from the next at a statistically significant level (with the threshold for p set at the standard 0.05), the seed images did display a slight but statistically significant decrease in spatial autocorrelation than the images from the last step in the chain. The median values yielded by this test are shown in Table 2, with only even steps displayed for brevity.

We also used the dlib facial landmark detection model (King 2009; Sagonas et al. 2013) to implement the novelty measure described in Definition 6. In Table 3 we show the median novelty scores derived from alternating steps of the chains (for brevity):

00	02	04	06	08	10
0.9774	0.9798	0.9821	0.9823	0.981	0.9816

Table 2: Median Moran's I values for alternating steps in the chain in Figure 3, from seed (00) to step 10.

00	02	04	06	08	10
28	34.2	36.3	36.55	36.35	35.9

Table 3: Median novelty score values for alternating steps in the chain in Figure 3, from seed (00) to step 10.

We also used the Brunner-Munzel test to compare these values, and found that the consecutive values for comparisons between steps 00 (the real seed), 01, 02, and 03 differed significantly, while the novelty values for rest of the steps did not differ significantly from each other. We noted an increase in novelty as the chain progressed, with the faces generated from synthetic images more distant from the real seed straying farther from the proportions described by (Schmid, Marx, and Samal 2008).

Further studies we plan to execute include occluding parts of correctly classified images and determining whether patches of low cohesion exist, determining facial symmetry thresholds between correctly and incorrectly classified images to measure disparities in novelty, and experimenting with distorting ratios subtly to flip classifications. The latter approach is informed by prior efforts to deliberately generate recognizably fake images through the "amplification" of image features such as color saturation and facial landmark positions (Broad, Fol Leymarie, and Grierson 2020). Instead of the obviously fake images produced by this procedure, we could modify it to perform subtle enhancements in the areas pointed out by Broad, Leymarie, & Grierson to create images that look realistic but might register as subtly uncanny with viewers. We also noted a shift in many of the chains that were seeded with male students, where each subsequent image appeared more aligned with stereotypical femininity—likely due to inherent gender biases within FLUX (Guo et al. 2026). Measuring differences between chains seeded with male and female images could yield interesting insights as well. However, before embarking on these, we must first perform a user study to come up with adequate human classification data.

Conclusions and Future Research

The "uncanny valley" is a contested term that may refer to a large number of phenomena in the general area of facial pattern recognition. Attempting to quantify it through the lens of creativity is a large and ill-defined task that we hope to shed some clarity on through our proposed experiment, which we hope to run and turn into a full paper over the coming months. Although the connection between the uncanny valley and creativity in the context of AI-generated imagery may seem tenuous at first glance, we believe that this is a promising method of combining human intuition with mathematical principles to address the failures of state-of-the-art detection models.

References

Broad, T.; Fol Leymarie, F.; and Grierson, M. 2020. Amplifying the uncanny.

Brunner, E., and Munzel, U. 2000. The nonparametric behrens-fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal* 42(1):17–25.

Cao, J.; Ma, C.; Yao, T.; Chen, S.; Ding, S.; and Yang, X. 2022. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4113–4122.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2017. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint arXiv:1711.09020*.

Dhote, K. 2024. Human faces dataset.

Dunn, J. D.; White, D.; Sutherland, C. A. M.; Miller, E. J.; Steward, B. A.; and Dawel, A. 2026. Too good to be true: Synthetic ai faces are more average than real faces and super-recognizers know it. *British Journal of Psychology* n/a(n/a).

Esser, P.; Rombach, R.; and Ommer, B. 2020. Taming transformers for high-resolution image synthesis.

Guo, X.; Liu, X.; Ren, Z.; Grosz, S.; Masi, I.; and Liu, X. 2023. Hierarchical fine-grained image forgery detection and localization. In *CVPR*.

Guo, W. M.; Qian, Q.; Hasan, K.; and Du, S. 2026. Position: Universal aesthetic alignment narrows artistic expression.

Hwang, Y., and Jeong, S.-H. 2026. Effects of creativity in ai advertising: The mediating role of perceived uncanniness. *Cyberpsychology, Behavior, and Social Networking* 29(2):109–113. PMID: 41460705.

Jabberi, M.; Wali, A.; Chaudhuri, B. B.; and Alimi, A. M. 2023. 68 landmarks are efficient for 3d face alignment: what about more? 3d face alignment method applied to face recognition. *Multimedia Tools Appl.* 82(27):41435–41469.

Kaate, I.; Salminen, J.; Santos, J.; Jung, S.-G.; Olkkonen, R.; and Jansen, B. 2023. The realness of fakes: Primary evidence of the effect of deepfake personas on user perceptions in a design task. *International Journal of Human-Computer Studies* 178:103096.

Karras, T.; Laine, S.; and Aila, T. 2018. A style-based generator architecture for generative adversarial networks. *CoRR* abs/1812.04948.

Kim, M.; Liu, F.; Jain, A.; and Liu, X. 2023. Dcface: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023*, 12715–12725.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10:1755–1758.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space.

Lin, J. 2023. Comparison of moran's i and geary's c in multivariate spatial pattern analysis. *Geographical Analysis* 55(4):685–702.

MacDorman, K. F., and Chattopadhyay, D. 2016. Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition* 146:190–205.

Makowski, D.; Te, A. S.; Neves, A.; Kirk, S.; Liang, N. Z.; Mavros, P.; and Chen, S. A. 2025. Too beautiful to be fake: Attractive faces are less likely to be judged as artificially generated. *Acta Psychologica* 252:104670.

Miller, E. J.; Steward, B. A.; Witkower, Z.; Sutherland, C. A.; Krumhuber, E. G.; and Dawel, A. 2023. Ai hyperrealism: Why ai faces are perceived as more real than human ones. *Psychological Science* 34(12):1390–1403.

Mori, M., and Macdorman, K. F. 2017. The uncanny valley: The original essay by masahiro mori. In *IEEE Spectrum*.

Moshel, M.; Carlson, T.; and Grootswagers, T. 2022. Are you for real? decoding realistic ai-generated faces from neural activity. *Vision Research* 199:108079.

Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards universal fake image detectors that generalize across generative models. In *CVPR*.

Ramaswamy, A.; Chockler, H.; and Navaratnarajah, M. 2025. Defining and quantifying creative behavior in popular image generators. In Oliveira, H. G.; Spendlove, B.; Gervás, P.; and Ventura, D., eds., *Proceedings of the 16th International Conference on Computational Creativity*, 362–367. Campinas, Brazil: Association for Computational Creativity.

Rey, S. J., and Anselin, L. 2010. *PySAL: A Python Library of Spatial Analytical Methods*. Berlin, Heidelberg: Springer Berlin Heidelberg. 175–193.

Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; and Pantic, M. 2013. 300 faces in-the-wild challenge: the first facial landmark localization challenge. In *International Conference on Computer Vision Workshops (ICCV Workshops)*.

Schmid, K.; Marx, D.; and Samal, A. 2008. Computation of a face attractiveness index based on neoclassical canons, symmetry, and golden ratios. *Pattern Recognition* 41(8):2710–2717.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*.

Yan, Z.; Wang, J.; Jin, P.; Zhang, K.-Y.; Liu, C.; Chen, S.; Yao, T.; Ding, S.; Wu, B.; and Yuan, L. 2025. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Forty-second International Conference on Machine Learning*.

Zhang, W.; Wang, X.; and Tang, X. 2011. Coupled information-theoretic encoding for face photo-sketch recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 513–520. IEEE.