Seeing Faces in Things: A Model and Dataset for Pareidolia

Mark Hamilton[1,2], Simon Stent[3], Vasha DuTell[1], Anne Harrington[1], Jennifer Corbett[1], Ruth Rosenholtz[4], and William T. Freeman[1]

[1] MIT, [2] Microsoft, [3] Toyota Research Institute, [4] NVIDIA

Abstract. The human visual system is well-tuned to detect faces of all shapes and sizes. While this brings obvious survival advantages, such as a better chance of spotting unknown predators in the bush, it also leads to spurious face detections. "Face pareidolia" describes the perception of face-like structure among otherwise random stimuli: seeing faces in coffee stains or clouds in the sky. In this paper, we study face pareidolia from a computer vision perspective. We present an image dataset of "Faces in Things", consisting of five thousand web images with human-annotated pareidolic faces. Using this dataset, we examine the extent to which a state-of-the-art human face detector exhibits pareidolia, and find a significant behavioral gap between humans and machines. We find that the evolutionary need for humans to detect animal faces, as well as human faces, may explain some of this gap. Finally, we propose a simple statistical model of pareidolia in images. Through studies on human subjects and our pareidolic face detectors we confirm a key prediction of our model regarding what image conditions are most likely to induce pareidolia. Dataset and Website: https://aka.ms/faces-in-things

Keywords: Pareidolia · Face Detection · Human Psychophysics

1 Introduction

Hamlet: Do you see yonder cloud that's almost in the shape of a camel?
Polonius: By the Mass and 'tis, like a camel indeed.
Hamlet: Methinks it is a weasel.
Polonius: It is back'd like a weasel.
Hamlet: Or like a whale.
Polonius: Very like a whale.

— *Hamlet, Act III, Scene ii*, William Shakespeare

Pareidolia is a type of visual "apophenia", which refers to the perception of patterns in random data. This occurs frequently in human perception as we look at clouds, mountain skylines, and burnt toast. Pareidolia is even described in an exchange in Hamlet [45]. When it was first described, pareidolia was seen as an early symptom of psychosis [7, 46]. Today we know pareidolia is common among healthy humans [47] and infants [23]. It is also not confined to humans:

Fig. 1: You print out an exciting new computer vision paper to review, but as you sit down at your desk to start reading you knock over your coffee cup. At first, you are annoyed, but then, you laugh! The sight of the stain induces "pareidolia" in your brain: rather than an unsightly blemish, you see a happy face. In this paper we explore the phenomenon of face pareidolia: Why don't we see faces all the time? Why do we see them at all when they are clearly so different from human faces? Can a better understanding of face pareidolia help computer vision–based face detection?

rhesus macaques, for example, have been shown to spend more time fixating on pareidolic than non-pareidolic images, in a manner similar to humans [51].

As an intriguing phenomenon of our visual system, pareidolia presents many opportunities in the study of the visual perception of both humans and machines. It offers a controlled setting in which to study object detection: we can present random signals to the visual system and study what detections arise. Do computer vision detectors exhibit similar misidentifications, and if not, why not? Why don't humans see pareidolic effects everywhere, in any textured region?

To help answer these questions, we introduce an annotated dataset of five thousand pareidolic face images, called "Faces in Things". With this dataset, we examine whether modern computer vision face detection systems, trained to robustly detect human faces, exhibit pareidolia. We show that a state-of-the-art neural network trained on the popular WIDER FACE detection benchmark [62] fails to detect pareidolic faces well, even when detection thresholds are relaxed. By fine-tuning the same model on the Faces in Things training data we create a simple and strong baseline for the task of pareidolic face detection, which shows that significantly higher machine pareidolic performance is within reach.

Next, we explore how we might bridge this gap to supervised—or ultimately, human—performance, without access to pareidolic training data? Could pareidolia appear in a face detector in a more natural way? The Faces in Things dataset provides a clean testbed to explore these questions in machines. We test a variety of different interventions ranging from image augmentation techniques to additional sources of training data. We find one possible mechanism that accounts for roughly half of the performance gap: when models are fine-tuned to detect *animal* faces, pareidolic face detection is significantly improved. This suggests that face pareidolia may arise in part from a more general, evolutionary need to detect diverse faces in the natural environment.

Finally, we consider why pareidolic faces are not all around us, and why certain textures seem to cause the effect more often. We propose two simple mathematical models, a simple Gaussian process model, and a second deep feature-based model, that capture important features of pareidolia. In particular, we show how these simple models both predict a "Goldilocks" zone, where

conditions are ideal to induce pareidolia. We confirm the existence of this zone with experiments on both human subjects and face detection models.

Through the contributions of our open-source dataset, models, and experimental findings, we bring the study of the intriguing phenomenon of face pareidolia to the computer vision community.

2 Related Work

Face detection. One of the most famous early examples of face detection was the Viola-Jones face detector [54, 55]. This detector used binary Haar features through simple-to-compute integral images and achieved greater precision and efficiency than early neural-network-based detectors [42, 43] and other feature-based methods [39, 61]. Following the deep learning breakthroughs of the 2010s, methods transitioned from hand-crafted features to learned features, and convolutional neural networks (CNNs) achieved close to human levels of performance on ever-larger datasets [8, 27–29, 33, 38, 50, 65]. For a broader survey of face detection methods, we refer the interested reader to [38, 64]. In our work, we use the recent RetinaFace model [8] as a strong face detection baseline.

Neuroscience of face pareidolia. The face is a highly unique stimulus for the human visual system [26, 52]: we find faces easy to spot and difficult to ignore. Face detection can occur in both noise and highly degraded images [5]. Prior work shows that face detection occurs in a dedicated brain region, the Fusiform Face Area [34]. But exactly what constitutes a face for the visual cortex and what are the mechanisms underlying pareidolia? A recent study into the temporal dynamics of neuro-imaging data during pareidolic face viewing showed results consistent with "a broadly-tuned face detection mechanism that privileges sensitivity over selectivity" [58]. Pareidolic faces do more than give the impression of the presence of faces: [48] show that they can trigger an additional face-specific attentional process, consuming more time and processing power than similar non-pareidolic stimuli, and even enhancing the detection of face-pareidolic objects [49]. Analyses in [30] revealed a network of neurons in the brain specialized to detect face pareidolia. Their results suggested that face processing has a strong top-down component whereby sensory input with even the slightest suggestion of a face can result in the interpretation of a face. Such top-down information might be supportive of some form of inverse rendering as a cognitive mechanism to explain the remarkable robustness of human perception of faces in degraded viewing conditions [11]. While our dataset allows the study of several types of face detection models, we focus our study on feed-forward neural networks which are known to yield close to human performance on challenging "in-the-wild" datasets [62].

Face pareidolia in computer vision. Face detection and face recognition have been core topics in computer vision for many decades, but the study of face pareidolia—and its deep relationship with visual object representation learning—has been relatively overlooked. Face pareidolia has some similarities with the problem of cross-modal recognition or cross-depiction: recognizing the

Fig. 2: Examples of face pareidolia from our "Faces in Things" dataset. Faces in Things consists of five thousand images annotated with bounding boxes (shown here), and facial attributes such as perceived emotion, gender, and intentionality.

same objects across different modalities irrespective of how the object is visually depicted. This has been explored particularly in the context of detecting faces, people and objects across modalities such as photography, different art movements, cartoons and sketches [4, 15, 36, 60]. The importance of capturing spatial relationships for robust cross-modal detection has also been highlighted [4]. The work of Castrejon *et al.* [6] showed how, when learning cross-modal scene representations with neural networks, units would emerge in the shared representation that tended to activate on consistent concepts, independently of the modality. This tendency was used by Abbas & Chalup [1], who found that mid-level units learned during human face detection could generalize to detect semantically similar facial key-points in pareidolic images, showing promise for pareidolia to emerge. However, the authors only evaluated the method qualitatively over a small test set of ten images. One route to a larger dataset may be through pareidolic image generation, which shows promise but does not yet produce convincingly natural images [12]. Curating a larger dataset, "Totally-Looks-Like" [41] explored the perceptual judgment of image similarity between humans and CNNs, using images which had been paired by humans as visually similar but semantically disparate. They found that visual representations extracted from CNNs such as ResNet [18] perform poorly in terms of reproducing the matching selected by humans. Though this dataset is of similar size to ours (6k samples), it is not specifically tailed towards face pareidolia and offers no bounding box or key-point annotations. Other datasets such as COCO-Periph [17] have been used to show that object detection behavior in CNNs and transformers diverge from human perception in peripheral vision.

In summary, there has yet to be a computational model of how or why pareidolia might arise proposed in the literature. Moreover, despite the abundance of face detection datasets [38], there is no large-scale dataset to directly support the study of face pareidolia. A large-scale pareidolia dataset would help the community explore the mechanisms underlying pareidolia, which may in turn help us to understand and harness human visual attention (which is drawn towards face-like objects), reduce pareidolic false positives in face detectors, help design-

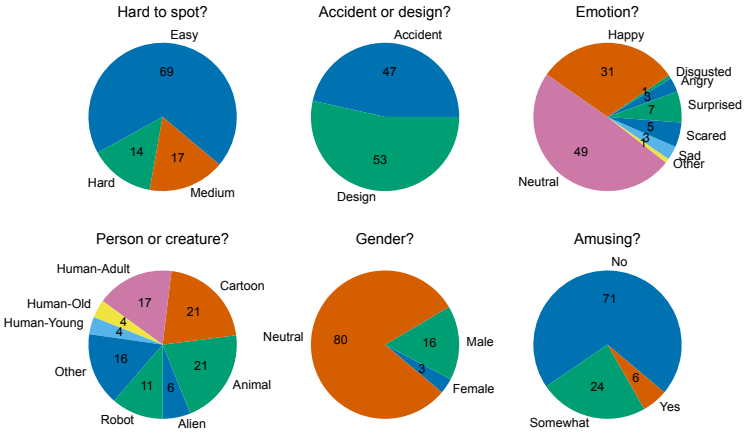

Fig. 3: Attributes of the Faces in Things Dataset. We find that 31% of faces are considered challenging to spot; faces are largely (31%) judged as happy; approximately half (47%) are judged as accidental rather than by design; animals and humans are seen in roughly equal numbers; and we observe a slight bias (16% vs 3%) towards male over female faces, similar to biases observed in prior studies [56, 57].

ers avoid or create pareidolia, improve pareidolic animation, and create systems that better understand how humans perceive the world.

3 Faces in Things Dataset

To address this gap, we begin by sampling candidate pareidolic images from the LAION-5B dataset [44]. This dataset consists of 5.85 billion CLIP-filtered image-text pairs, of which 40% of captions contain English. We use CLIP retrieval [2] to build a raw image set based on text queries including "pareidolia", "faces in things", "accidental faces", and "[object] looks like a face". We download images, check for duplicates, then downsample to 512×512 pixels while preserving the aspect ratio with white-space padding. We used the VGG Image Annotation tool [10] to manually annotate images, removing samples that contain the faces of actual humans or animals. Some examples of annotated images are shown in Fig. 2. Our annotations include the bounding boxes of pareidolic faces and basic facial attributes as summarized in Fig. 3. Though beyond the scope of the current paper, we note that these attributes could be useful for other future studies. We divide the dataset at random into training (70%) and testing (30%) sets. We refer to this as the 'Pareidolic' dataset.

4 Experiments

Datasets. We use the following additional datasets. Fig. 4 shows the average faces within our dataset (Pareidolic) and the WIDER FACE (Human), and AnimalWeb (Animal) datasets.

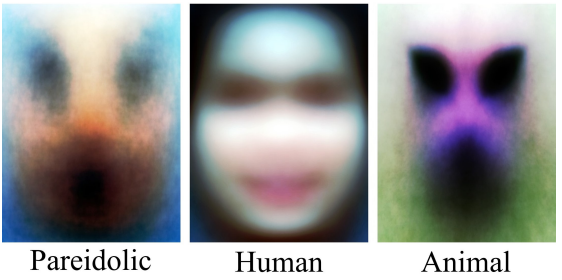

Fig. 4: The Appearance of an Average Pareidolic Face. Per-channel histogram-equalized average images for registered pareidolic faces (our Faces in Things dataset), human faces (samples from the WIDER FACE dataset [62]), and animal faces (AnimalWeb [24]). The average pareidolic face, while less distinct than human or animal, has surprisingly clear eye, nose, and mouth features, and vertical symmetry.

WIDER FACE [62] is a popular face detection benchmark dataset with 32,203 images and 393,703 faces. It contains a high degree of variability in scale, pose, makeup, lighting, emotion, and occlusion, organized across 61 event classes. We use the provided 40%/10%/50% splits for training, validation, and testing. We refer to this as the 'Human' dataset.

AnimalWeb [24] is a collection of 22,451 faces from 334 diverse species and 21 animal orders across biological taxonomy. These faces are captured 'in-the-wild' and are consistently annotated with 9 landmarks on key facial features. We convert these landmarks to bounding boxes, by finding the tightest box that captures the points and expanding this box's width and height by 15%. We refer to this as the 'Animal' dataset.

WIDER FACE Corruptions. To measure whether pareidolia could arise from common data augmentations we corrupt the WIDER FACE images using the level 3 strength of the corruptions used in both the COCO-C [35] and ImageNet-C [19] datasets. We also include a Sobel filtering corruption [13] which has been shown to reduce a model's dependence on texture information [22].

Models and Training. We use RetinaFace [8] which achieves state-of-the-art performance on WIDER FACE easy and medium subsets and is the third-best face detector on the hard subset, missing the top model by less than a percentage point of Average Precision (AP). We perform experiments using both their MobileNet [21] and ResNet50 [18] backbones and use the Pytorch_Retinaface [3] repository to ensure the same experimental conditions, dataset characteristics, and preprocessing. We use pre-trained models provided by this repository and fine-tune them for 10 epochs with the AdamW optimizer [31] using a learning rate of 10^{-4} and a weight decay of 5×10^{-4}. We verify that fine-tuning using this strategy on the original WIDER FACE training dataset does not hurt model performance. When fine-tuning on Faces in Things (Pareidolia), AnimalWeb (Animal), WIDER FACE (Human), and Corruption datasets we randomly replace images in the original WIDER FACE stream of training data with data

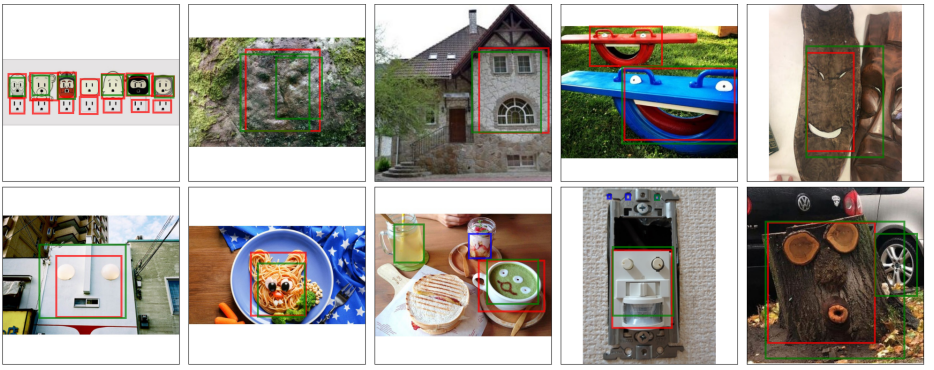

Fig. 5: Qualitative Analysis of Transfer Experiments. On a sample of held-out test images, we visualize the confident ($p > 10\%$) detections of our ground truth (red), our model fine-tuned on human faces (blue), and our model fine-tuned on animal faces (green). It is evident from these and Table 1 that fine-tuning on animal faces significantly boosts the model's ability to detect pareidolic faces.

from the target dataset 90% of the time. This allows the network to learn the new task without catastrophic forgetting [25]. These changes to the optimizer, learning rate, number of epochs, and stream of training data are the only changes we make to the training paradigm of [8]. Figures in this work use the MobileNet architecture of RetinaNet unless specified otherwise. AP evaluation computations share the same setting and parameters as [8].

4.1 Does a SOTA Face Detector Exhibit Pareidolia?

We measure the Average Precision (AP) of the MobileNet and ResNet50 RetinaNet architectures on the Faces in Things dataset. The first row of Table 1 shows results for existing pre-trained models, and the second row shows those for models fine-tuned on the original WIDER FACE training data. These act as control groups to ensure our transfer learning procedure does not interfere with our measurement of the effects of other interventions. Though these models exhibit pareidolia to a small extent, they fall far short of a model fine-tuned to detect pareidolic faces. Fig. 5 also depicts some of these predictions with blue boxes. On the whole, the models trained only on human faces are largely silent across the Faces in Things dataset.

4.2 How Might Pareidolia Emerge?

The WIDER FACE dataset is known for its diversity of lighting, pose, makeup, emotion, and scale of faces. This fact, coupled with the results of Section 4.1 begs the question: What else do models need to experience pareidolia as humans do? The Faces in Things dataset provides a clean and robust setting to explore

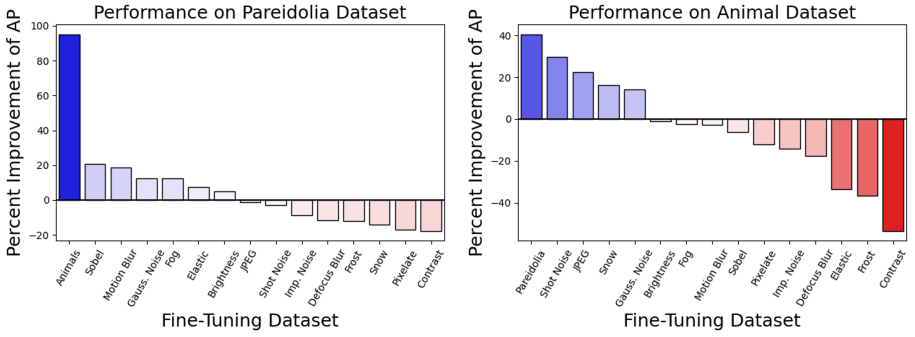

Fig. 6: Measuring the effect of several training interventions on pareidolic face detection The left plot shows that fine-tuning RetinaNet on animals improves pareidolic face detection more than any other intervention. Conversely, the right plot shows that pareidolic fine-tuning improves animal face detection performance.

the development of pareidolia in algorithms. Unlike in humans, where it is impossible to causally intervene on their facial training data, we can easily modify an algorithm's training data. This makes it possible to explore whether one can induce pareidolia in algorithms using specific stimuli.

To this end, we investigate whether a variety of training data interventions can induce pareidolia in algorithms. In particular, we measure the effect of adding several data augmentations from the COCO-C [35] and ImageNet-C [19] datasets and explore a Sobel filtering augmentation which reduces models' dependence on texture. Additionally, we also measure the effect of adding animal faces to the training data. Animal faces show a far greater breadth of variation in coloration, structure, and appearance than human faces. Recognizing animal faces provides many evolutionary advantages including gaze detection during hunting and avoiding onlooking predators. The generality required to detect this wide space of faces could yield a greater number of "false positives" that lead to the sensation of pareidolia. Indeed, some recent studies provide some corroborating evidence for this hypothesis. Firstly, Rhesus Monkeys exhibit pareidolia [51] showing this effect does not only occur in humans. Secondly, the experience of pareidolia is a rapid cognitive process and not a "late re-interpretation" of input signals [16, 58] which the authors conclude is evidence that pareidolia could be linked to the need to quickly react to predators.

We plot the change to pareidolic face detection performance as a function of each training intervention in the left panel of Figure 6. Of the different corruption interventions, we find that Sobel filtering, motion blur, Gaussian noise, and fog tend to slightly improve pareidolic face detection while most other corruptions do not improve pareidolic face detection performance. Most strikingly, the addition of animal faces to the training data roughly doubles the algorithms' ability to detect pareidolic faces compared to the control group, closing around half of the gap between a human-trained model and a pareidolia trained model.

Finetuning	AP	
	MobileNet	ResNet50
None	7.9%	2.8%
Human (Control)	9.8%	3.6%
Animal	16.7%	15.4%
Pareidolia	33.9%	27.1%
Animal + Pareidolia	**36.4%**	**31.7%**

Table 1: Effect of Fine Tuning on Pareidolic Face Detection. Our results show that WIDER FACE-trained RetinaFace models do not detect many pareidolic images. Fine-tuning these models on animal faces approximately doubles pareidolic face detection rates. Interestingly, adding animal faces alongside pareidolic faces (30%/70% split respectively) can improve performance over fine-tuning on pareidolic faces alone.

Reciprocally, the right-hand plot of Figure 6 shows that fine-tuning on pareidolic images yields the greatest improvement in animal face detection. We further explore this phenomenon in Table 1, where we show that this effect occurs across both MobileNet and ResNet50 architectures. Finally, we also show that adding a small number of animal faces (30% animal 70% pareidolic) can improve pareidolic face detection performance over pareidolic images alone.

To understand this effect better, Fig. 7 visualizes the inner representations of this model across the three datasets (Human = WIDER FACE, Animal = AnimalWeb, Pareidolic = Faces in Things). Specifically, we extract multi-scale features from the animal and pareidolia fine-tuned RetinaNet shared feature layer before the application of the classification and regression heads. We average pool these features across the bounding box for each face and visualize them with t-SNE [20]. This figure shows that RetinaNet's representations of animal and pareidolic faces tend to cluster together and are distinct from its representations of human faces. This lends evidence to the relative similarity of pareidolic and animal faces compared to human faces. We also reiterate that we filtered the Faces in Things dataset to avoid images of real animals.

5 Modeling Pareidolia

Though many prior works have measured pareidolia, there has yet to be a simple mathematical model that describes the high-level structure of this phenomenon. In this section we provide two simple formal models of pareidolia and show that they both exhibit a testable prediction: the existence of a peak in pareidolic face detection as a function of an image's complexity. Section 5.4 presents experimental evidence of this "pareidolic peak" in both humans and machines.

5.1 Gaussian Model of Pareidolia

A model of pareidolia needs to describe two processes: (1) the random process that generates candidate images, and (2) the face detection process which determines when an image is pareidolic. We begin with a simple Gaussian model for

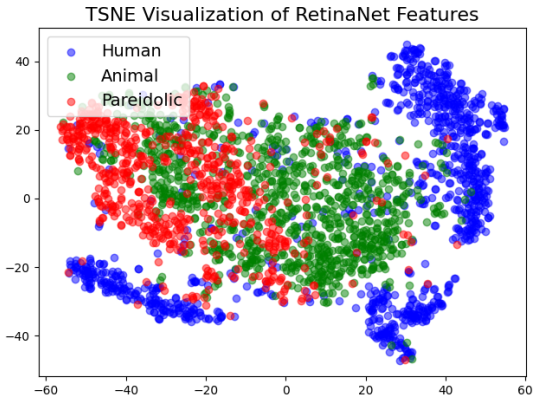

Fig. 7: Visualizing RetinaNet Representations across Datasets. Animal+Pareidolia fine-tuned RetinaNet representations tend to group animal and pareidolic faces together. This lends evidence to the hypothesis that the perception of animal and pareidolic faces are linked. (To highlight the commonality of pareidolic animal detection we note the similarity of these points to a frog.)

each. We model the image generation process as a sum of independent normal modes, each contributing a zero-mean Gaussian of a specified variance, multiplied by the mode image y_i. For example, as in [9, 53], these modes could be the principal components of a mean-subtracted image dataset. In this setting the generated image, y, is a weighted sum of the normal modes:

$$\boldsymbol{y} = \sum_i n_i \boldsymbol{y}_i \quad \text{where,} \quad n_i \sim N(0, \sigma_i) \tag{1}$$

To model our face detection process we capture the intuition of matching an image to a template image and note that this can be generalized to distributions of template images. In particular, the target pareidolic image is represented as a vector, $\boldsymbol{a}$, of statistically independent target coefficients, a_i, for each mode. The probability that this mode contributes towards the face detection, $P(a_i)$, is the probability of detecting the pareidolic value, a_i, at the ith mode. We assume a Gaussian detection process: $P(a_i|y_i) \sim N(a_i, \gamma_i)$. Because each mode's coefficient is a zero-mean Gaussian distribution, $P(y_i) \sim N(0, \sigma)$, we have:

$$P(a_i) = \int_{y_i} P(a_i, y_i) \mathrm{d}y_i \tag{2}$$

$$= \int_{y_i} P(a_i|y_i) P(y_i) \mathrm{d}y_i \tag{3}$$

$$= \frac{1}{2\pi\gamma_i\sigma_i} \int_{y_i} e^{-\frac{(y_i - a_i)^2}{2\gamma_i^2}} e^{-\frac{y_i^2}{2\sigma^2}} \mathrm{d}y_i \tag{4}$$

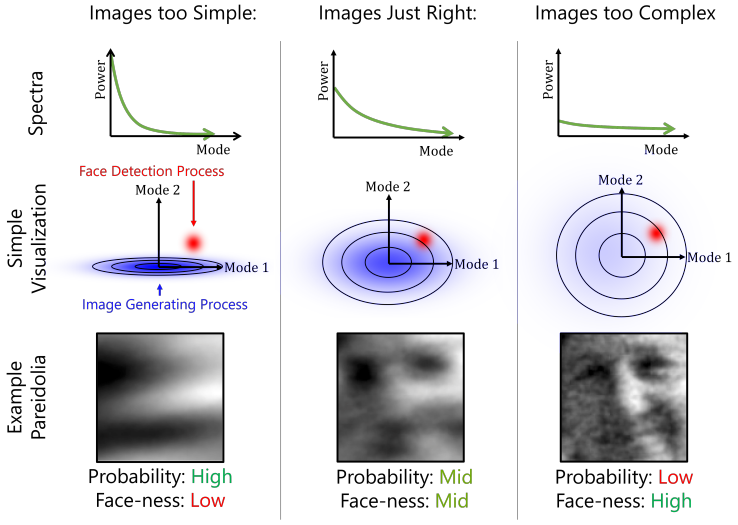

Fig. 8: Illustration of the proposed Gaussian model for pareidolia with three example generating distributions. To make pareidolia likely, the generating distribution needs a proper distribution of spatial frequencies. A process with too few spatial frequencies (left) is likely to only generate weak face-like details ("face-ness": low). In contrast, with too many frequencies (right), faces can be modeled with exquisite detail ("face-ness": high), but the likelihood of drawing any particular desired combination become vanishingly small. The most likely pareidolic images form when the generating distribution has the right spectrum (middle), enabling reasonable faces to emerge with reasonable likelihood. In other words, this model predicts that pareidolic faces will match the low frequencies of faces but differ in the higher frequency details.

Note that σ_i^2 is the variance of the random process generating the pareidolia, while γ_i^2 is the variance of the likelihood term — how far a mode is allowed to vary from the target mode value before it stops looking like the target image, $\boldsymbol{a}$.

We can complete the square in Eq. 4 to write the product of Gaussians in $P(a_i)$ as a single Gaussian. Integrating that Gaussian over all possible observations y_i gives the probability of finding the pareidolic value a_i from mode i:

$$P(a_i) = \frac{1}{\sqrt{2\pi(\gamma_i^2 + \sigma_i^2)}} e^{-\frac{a_i^2}{2(\sigma_i^2+\gamma_i^2)}} \tag{5}$$

5.2 Predicting Peak Pareidolia

For a given mode's detection variance, γ_i^2, and target mode coefficient, a_i, Eq. 4 allows us to find the optimal mode variance to generate pareidolia, i.e., to maximize $P(a_i)$. Unfortunately, we seldom have the flexibility to design a random process one mode at a time. But we may have the option to select between image generation processes that have different numbers of modes, M. Since each mode

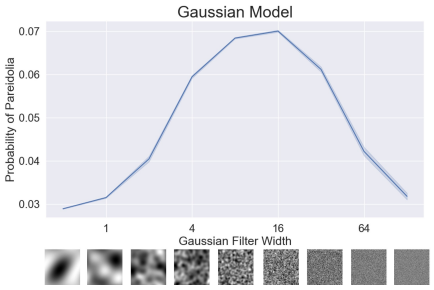

Fig. 9: Probability of pareidolia (Eq. 6) in the Gaussian model ($\sigma = 10$) across images with different spatial frequency distributions (Sec. 5.4). This assumes spatial frequencies are uncorrelated and thus underestimates the probability of pareidolia, however peak pareidolia is still present.

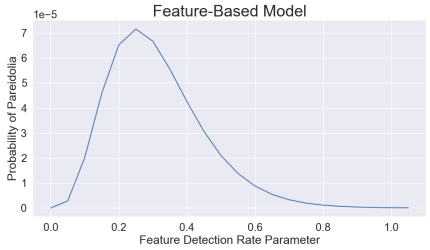

Fig. 10: Probability of pareidolia under the feature-based example of Eq. (8) as a function of the rate of feature detection, $\lambda_i = \lambda$ for all i, within the random images. Note the low probability of pareidolia for both feature-free ($\lambda \to 0$) and feature-rich ($\lambda \gg 0$) random images.

is independent, the probability of a pareidolic detection of the target object template is the product of detecting the target coefficient for each of the M modes:

$$P(a) = \prod_{i}^{M} P(a_i) \tag{6}$$

We plot some predictions of our Gaussian model in Fig. 9 for a target template with a $\frac{1}{f}$ power spectrum (standard deviation of each mode inversely proportional to mode number) on noise images of varying complexity. We note the existence of a peak in pareidolic detection probability for random image generation processes with a mid-range number of spatial modes, as measured by the width of Gaussian that modulates power in Fourier space. Too few modes in the random generation process, and no image will ever have enough complexity to render the target well. Too many modes and pareidolia becomes unlikely because so many modes need to match the desired target values. Each added mode multiplies the pareidolia probability by another small factor. In between, there is what we call *peak pareidolia*. As the detection value, σ, becomes more stringent (smaller) the peak pareidolia value occurs at a larger number of modes and becomes less probable. We illustrate this effect in Figure 8

5.3 Higher-Level Feature Model of Pareidolia

The Gaussian model for pareidolia above lays out important aspects of pareidolia, but relies on a naive model of object detection, the squared distance from a template image. We assume that a more realistic model of human perception would incorporate higher-level features and introduce a still simple, yet more realistic, feature-based model.

We assume that the detection of an object requires particular features to be detected in certain spatial regions, e.g. an eye in the top left and right, a nose in the center, and a mouth in the bottom. Such an approach has been used in computer vision object detection algorithms, e.g. [14, 59, 63]. Any given object template has some number of regions, R_i, indexed by i, within which a given feature, F_i, must be detected. The other features, $F_{j \neq i}$, should not be detected in region i. For a given random image where we hope to detect pareidolia, we assume that feature existence is a spatial Poisson process. In this process, the probability of n feature instances for any given feature i over some area B_i is

$$P(n_i) = \frac{(\lambda_i |B_i|)^{n_i}}{n_i!} e^{-\lambda_i |B_i|} \tag{7}$$

To detect a pareidolic instance of the object template, we must detect one feature of the correct type, F_i, in each region i of the face template, and zero features of the wrong type, $F_{j \neq i}$ in each region i. Assuming independence of the feature detections, and for simplicity setting all the feature detection rates to be the same, $\lambda_i = \lambda$, and all the template areas to be the same, $B_i = 1$, we have for the probability, $P(O)$, of pareidolic detection of object O:

$$P(O) = \prod_{i}^{M} \lambda^1 e^{-\lambda(M-1)} \tag{8}$$

For the case of $M = 4$, a simple detection model for two eyes, a nose, and a mouth, we have $P(o) = \lambda^4 e^{-16\lambda}$, which is plotted in Fig. 10. In this feature-based object detection model, we also find the existence of "peak pareidolia". Again, it is governed by a parameter describing the complexity of the random image, in this case a Poisson process rate parameter, λ, that governs the probability of a feature detection per unit area. For too low a rate, the model doesn't generate enough features to satisfy the object template, for too high a rate, the probability of seeing only the right features in just the right places becomes very small. In between is the most probable rate for pareidolia.

5.4 Measuring the Pareidolic Peak in Humans and Machines

Both mathematical models of Section 5 predict the existence of a peak of pareidolic face detection as a function of image complexity. We show the existence of this pareidolic peak in both humans and machines. In particular, we perform a psychophysics experiment where human subjects view noise images of varying complexity and report how many pareidolic faces they saw in each image, from zero to nine. Campbell [5] demonstrated that a 12x12 array of random, binary squares is sufficient to evoke human and animal faces. We generate noise images of varying complexity by randomly sampling Fourier coefficients and modulating these coefficients with a zero mean σ^2 variance Gaussian in Fourier space. We show some samples of these images on the x-axis of Fig. 11. Intuitively, the Gaussian envelope in frequency space filters out most frequencies higher than σ

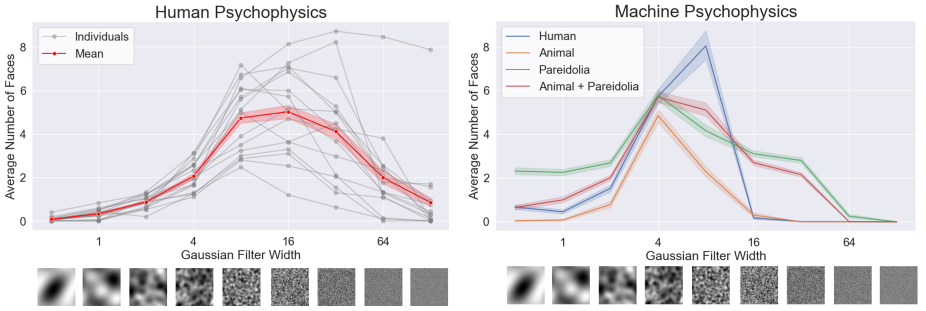

Fig. 11: Measuring Peak Pareidolia. Left: Subjects were asked how many faces they see in each noise image. We plot the average number of faces detected as a function of noise frequency (examples on x-axis), the mean over all subjects and its 95% confidence interval in red. Right: average number of faces detected by our fine-tuned models. This reveals the "peak pareidolia" effect predicted in Section 5 across humans and machines.

after applying an inverse Fourier transform. We detail our image stimuli creation method in the Supplement.

We find that humans exhibit the model-expected peak pareidolia, with a maximum number of faces detected at a frequency filter width of 16 (Fig. 11, left). The existence of a pareidolic peak at or near this filter width is consistent among *all* subjects even for those that reported fewer faces overall. Although response time did decrease slightly at higher frequencies, it did not fall off completely at the highest frequency levels, indicating that fewer reported faces were not the result of subjects "giving up" on the task. We provide additional details and analysis of this experiment in the Supplement.

Finally, we evaluate our fine-tuned models from Section 4 on the same images to test whether machines also exhibit peak pareidolia (Fig. 11, right). In particular, we showed the models $5,000$ sampled noise images of varying frequency levels and counted the number of face detections they make with confidence $> 10\%$. We find the same characteristic "pareidolic peak" where models detect the most faces in medium-complexity images.

6 Conclusion

We have taken initial steps towards the mathematical modelling of pareidolia and build a richly annotated dataset of images for face pareidolia. We showed through experiments on modern face detectors that detecting animal faces may partly explain the emergence of pareidolia in a complex vision system. The Faces in Things dataset can help the community address other questions about how and why pareidolic behavior emerges, a hallmark of humans' robust recognition system. We hope that our findings and dataset will spark further study of pareidolia and its potential use to improve computer vision systems.

Acknowledgments

We would like to thank Karen Hamilton for her hundreds of hours of annotations for the Faces in Things dataset. We also thank Abhishek Dutta who created the VGG Image Annotation tool, for his kind generosity to support our project.

We would like to thank the Microsoft Research Grand Central Resources team for their gracious help performing the experiments in this work. Special thanks to Oleg Losinets and Lifeng Li for their consistent, gracious, and timely help, debugging, and expertise. Without them, none of the experiments could have been run.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 2021323067. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their employers, or the National Science Foundation.

This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/) and the CSAIL MEnTorEd Opportunities in Research (METEOR) Fellowship.

Research was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-192-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

The authors acknowledge the MIT SuperCloud [40] and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported within this paper.

References

1. Abbas, A., Chalup, S.: From Face Recognition to Facial Pareidolia: Analysing Hidden Neuron Activations in CNNs for Cross-Depiction Recognition. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (Jul 2019). `https://doi.org/10.1109/IJCNN.2019.8852013`, iSSN: 2161-4393
2. Beaumont, R.: Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. `https://github.com/rom1504/clip-retrieval` (2022)
3. biubug6: Retinaface in pytorch (Nov 2021), `https://github.com/biubug6/Pytorch_Retinaface`
4. Cai, H., Wu, Q., Corradi, T., Hall, P.: The Cross-Depiction Problem: Computer Vision Algorithms for Recognising Objects in Artwork and in Photographs. arXiv:1505.00110 [cs] (May 2015), `http://arxiv.org/abs/1505.00110`, arXiv: 1505.00110
5. Campbell, F.: How much of the information falling on the retina reaches the visual cortex and how much is stored in the visual memory. Pattern recognition mechanisms **54**, 83–94 (1983)

6. Castrejon, L., Aytar, Y., Vondrick, C., Pirsiavash, H., Torralba, A.: Learning Aligned Cross-Modal Representations from Weakly Aligned Data. In: CVPR (Jun 2016)
7. Conrad, K.: Die beginnende Schizophrenie. Versuch einer Gestaltanalyse des Wahns (1958)
8. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: RetinaFace: Single-stage Dense Face Localisation in the Wild. CVPR (2020)
9. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley (2012)
10. Dutta, A., Zisserman, A.: The VIA annotation software for images, audio and video. In: Proceedings of the 27th ACM International Conference on Multimedia. MM '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3343031.3350535, https://doi.org/10.1145/3343031.3350535
11. Egger, B., Siegel, M.H., Arora, R., Soltani, A.A., Yildirim, I., Tenenbaum, J.: Inverse rendering best explains face perception under extreme illuminations. In: CogSci (2020)
12. Endo, Y., Asanuma, R., Shimojo, S., Akashi, T.: Systematic face pareidolia generation method using cycle-consistent adversarial networks. IEEJ Transactions on Electrical and Electronic Engineering **19**(4), 535–541 (2024)
13. Farid, H., Simoncelli, E.P.: Optimally rotation-equivariant directional derivative kernels. In: International Conference on Computer Analysis of Images and Patterns. pp. 207–214. Springer (1997)
14. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE TPAMI **32**(9) (2010)
15. Ginosar, S., Haas, D., Brown, T., Malik, J.: Detecting people in cubist art. In: ECCV Workshops (2014)
16. Hadjikhani, N., Kveraga, K., Naik, P., Ahlfors, S.P.: Early (n170) activation of face-specific cortex by face-like objects. Neuroreport **20**(4), 403 (2009)
17. Harrington, A., DuTell, V., Hamilton, M., Tewari, A., Stent, S., Freeman, W.T., Rosenholtz, R.: Coco-periph: Bridging the gap between human and machine perception in the periphery. In: The Twelfth International Conference on Learning Representations
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
19. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations (2019)
20. Hinton, G.E., Roweis, S.: Stochastic neighbor embedding. Advances in neural information processing systems **15** (2002)
21. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
22. Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information clustering for unsupervised image classification and segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9865–9874 (2019)
23. Kato, M., Mugitani, R.: Pareidolia in Infants. PLoS ONE **10**(2) (Feb 2015). https://doi.org/10.1371/journal.pone.0118539, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4331561/
24. Khan, M.H., McDonagh, J., Khan, S., Shahabuddin, M., Arora, A., Khan, F.S., Shao, L., Tzimiropoulos, G.: AnimalWeb: A large-scale hierarchical dataset of annotated animal faces. In: CVPR (2020)

25. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
26. Leopold, D.A., Rhodes, G.: A comparative view of face perception. Journal of Comparative Psychology **124**(3), 233 (2010)
27. Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G.: A convolutional neural network cascade for face detection. In: CVPR (2015)
28. Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., Wang, C., Li, J., Huang, F.: DSFD: dual shot face detector. In: CVPR (2019)
29. Liao, S., Jain, A.K., Li, S.Z.: A Fast and Accurate Unconstrained Face Detector. PAMI **38**(2), 211–223 (Feb 2016). `https://doi.org/10.1109/TPAMI.2015.2448075`
30. Liu, J., Li, J., Feng, L., Li, L., Tian, J., Lee, K.: Seeing Jesus in toast: Neural and behavioral correlates of face pareidolia. Cortex **53**, 60–77 (Apr 2014). `https://doi.org/10.1016/j.cortex.2014.01.013`, `http://www.sciencedirect.com/science/article/pii/S0010945214000288`
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
32. Mannos, J., Sakrison, D.: The effects of a visual fidelity criterion of the encoding of images. IEEE transactions on Information Theory **20**(4), 525–536 (1974)
33. Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L.: Face detection without bells and whistles. In: ECCV (2014)
34. Mcgugin, R., Gatenby, C., Gore, J., Gauthier, I.: High-resolution imaging of expertise reveals reliable object selectivity in the FFA related to perceptual performance. Proceedings of the National Academy of Sciences of the United States of America **109**, 17063–8 (Oct 2012). `https://doi.org/10.1073/pnas.1116333109`
35. Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A.S., Bethge, M., Brendel, W.: Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint arXiv:1907.07484 (2019)
36. Mishra, A., Nandan Rai, S., Mishra, A., Jawahar, C.V.: Iiit-cfw: A benchmark database of cartoon faces in the wild. In: VASE ECCV Workshops (2016)
37. Proverbio, A.M., Galli, J.: Women are better at seeing faces where there are none: an erp study of face pareidolia. Social cognitive and affective neuroscience **11**(9), 1501–1512 (2016)
38. Ranjan, R., Sankar, S., Bansal, A., Bodla, N., Chen, J.C., Patel, V., Castillo, C., Chellappa, R.: Deep Learning for Understanding Faces: Machines May Be Just as Good, or Better, than Humans. IEEE Signal Processing Magazine **35**, 66–83 (Jan 2018). `https://doi.org/10.1109/MSP.2017.2764116`
39. Rein-Lien Hsu, Abdel-Mottaleb, M., Jain, A.: Face detection in color images. PAMI **24**(5), 696–706 (May 2002). `https://doi.org/10.1109/34.1000242`, `http://ieeexplore.ieee.org/document/1000242/`
40. Reuther, A., Kepner, J., Byun, C., Samsi, S., Arcand, W., Bestor, D., Bergeron, B., Gadepally, V., Houle, M., Hubbell, M., Jones, M., Klein, A., Milechin, L., Mullen, J., Prout, A., Rosa, A., Yee, C., Michaleas, P.: Interactive supercomputing on 40,000 cores for machine learning and data analysis. In: 2018 IEEE High Performance extreme Computing Conference (HPEC). pp. 1–6. IEEE (2018)
41. Rosenfeld, A., Solbach, M.D., Tsotsos, J.K.: Totally looks like-how humans compare, compared to machines. In: ACCV (2018)

42. Rowley, H.A., Baluja, S., Kanade, T.: Human Face Detection in Visual Scenes. In: Touretzky, D.S., Mozer, M.C., Hasselmo, M.E. (eds.) NeurIPS (1996), http://papers.nips.cc/paper/1168-human-face-detection-in-visual-scenes.pdf
43. Rowley, H.A., Baluja, S., Kanade, T.: Neural network-based face detection. PAMI **20**(1), 23–38 (1998)
44. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS Datasets and Benchmarks Track (2022), https://openreview.net/forum?id=M3Y74vmsMcY
45. Shakespeare, W.: The Tragedy of Hamlet, Prince of Denmark. The Folio Society (1954)
46. Sims, A.: Symptoms in the mind: An introduction to descriptive psychopathology. Bailliere Tindall Publishers (1988)
47. Summerfield, C., Egner, T., Mangels, J., Hirsch, J.: Mistaking a house for a face: neural correlates of misperception in healthy humans. Cerebral cortex **16**(4), 500–508 (2006)
48. Takahashi, K., Watanabe, K.: Gaze Cueing by Pareidolia Faces. i-Perception **4**(8), 490–492 (Dec 2013). https://doi.org/10.1068/i0617sas, https://doi.org/10.1068/i0617sas
49. Takahashi, K., Watanabe, K.: Seeing Objects as Faces Enhances Object Detection. i-Perception **6**(5) (2015). https://doi.org/10.1177/2041669515606007, https://doi.org/10.1177/2041669515606007
50. Tang, X., Du, D.K., He, Z., Liu, J.: PyramidBox: A Context-Assisted Single Shot Face Detector. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV (2018)
51. Taubert, J., Wardle, S.G., Flessert, M., Leopold, D.A., Ungerleider, L.G.: Face Pareidolia in the Rhesus Monkey. Current Biology **27**(16), 2505–2509 (Aug 2017). https://doi.org/10.1016/j.cub.2017.06.075, http://www.sciencedirect.com/science/article/pii/S0960982217308126
52. Tsao, D.Y., Livingstone, M.S.: Mechanisms of face perception. Annu. Rev. Neurosci. **31**, 411–437 (2008)
53. Turk, M., Pentland, A.: Eigenfaces for recognition. J. of Cognitive Neuroscience **3**(1) (1991)
54. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR (2001)
55. Viola, P., Jones, M.J.: Robust real-time face detection. IJCV **57**(2), 137–154 (2004)
56. Wardle, S.G., Ewing, L., Malcolm, G.L., Paranjape, S., Baker, C.I.: Children perceive illusory faces in objects as male more often than female. Cognition **235** (2023). https://doi.org/https://doi.org/10.1016/j.cognition.2023.105398, https://www.sciencedirect.com/science/article/pii/S001002772300032X
57. Wardle, S.G., Paranjape, S., Taubert, J., Baker, C.I.: Illusory faces are more likely to be perceived as male than female. Proceedings of the National Academy of Sciences **119**(5) (2022). https://doi.org/10.1073/pnas.2117413119, https://www.pnas.org/doi/abs/10.1073/pnas.2117413119
58. Wardle, S.G., Taubert, J., Teichmann, L., Baker, C.I.: Rapid and dynamic processing of face pareidolia in the human brain. Nature communications **11**(1), 4518 (2020)
59. Weber, M., Welling, M., Perona, P.: Unsupervised learning of models for recognition. In: ECCV (2000)

60. Westlake, N., Cai, H., Hall, P.: Detecting People in Artwork with CNNs. In: ECCV Workshops (2016)
61. Yang, M.H., Kriegman, D., Ahuja, N.: Detecting faces in images: a survey. PAMI **24**(1), 34–58 (Jan 2002). https://doi.org/10.1109/34.982883
62. Yang, S., Luo, P., Loy, C.C., Tang, X.: WIDER FACE: A Face Detection Benchmark. In: CVPR (2016)
63. Yuille, A.L.: Deformable templates for face recognition. Journal of Cognitive Neuroscience **3**(1), 59–70 (1991)
64. Zafeiriou, S., Zhang, C., Zhang, Z.: A survey on face detection in the wild: Past, present and future. Computer Vision and Image Understanding **138**, 1–24 (Sep 2015). https://doi.org/10.1016/j.cviu.2015.03.015, https://linkinghub.elsevier.com/retrieve/pii/S1077314215000727
65. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: S3fd: Single shot scale-invariant face detector. In: ICCV (2017)

A Appendix

A.1 Additional Information on Frequency-Dependent Noise Generation

To generate noise of different frequencies for our experiments we leveraged the fact that low frequency information is localized close to the origin in the Fourier transform of an image. To this end, we can generate a random noise images by randomly sampling images in the Fourier space, filtering them, and transforming back to image space. Specifically, we modulate a random Fourier spectra by a Gaussian centered at 0 with a variable width. The width controls the frequency of the noise created. Larger width images let more frequencies pass through in Fourier space, and the resulting image has higher frequency patterns.

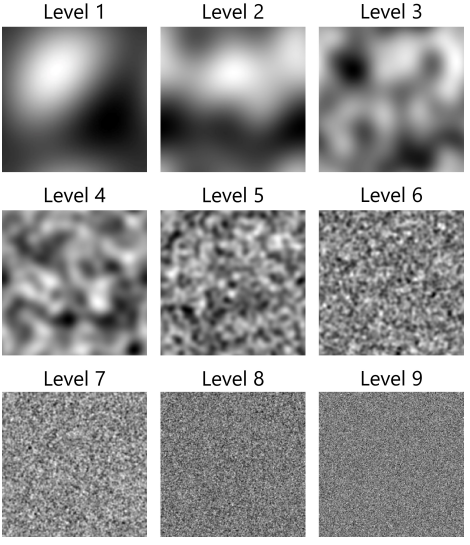

Fig. 12: Different frequency noise levels used in human experiments from Fig. 11.

For the noise levels shown in Fig. 12 we use a width of $2^{level-2}$. We include the following pseudo-code to be precise:

```
import numpy as np
from scipy.fftpack import fft2, fftshift, ifft2
def generate_noise(width, size):
    modes = np.randn(size, size)
    dft = fft2(modes)
    gauss = fftshift(np.exp((-xx ** 2 - yy ** 2) / (2 * width
     ** 2)))
    return np.real(ifft2(dft * gauss))
```

A.2 Human Psychophysics Experiment Setup

In the main experiment, 14 subjects (6 female, 8 male) were shown noise images filtered as described in A.1 at resolution 1024x1024. At each of 9 filter widths, 10 random noise pulls were created, making 90 unique images. Each unique image was repeated 3 times, for a total of and 270 presentations per subject. The images were shown in random order, which was different for each subject. Subjects were split into two groups of 7 subjects, with the two groups view different sets of 90 unique images. Experiments were performed in PsychoPy, and the experimental code both for generating and displaying experimental stimlui is available at https://github.com/vdutell/pareidolicNoise.

Subjects were seated in a dimly-lit room in front of a laptop with screen resolution 2500x1664, with the 1024x1024 image subtending the entire screen vertically, with grey padding on the horizontal edges (except for 2 additional subjects in the control experiment described below, where image subtended half screen height). Subjects were instructed to sit at a comfortable distance from the screen (approximately 30 inches), and were asked to count the number of faces seen in each noise image, and report the number from [0-9] on the laptop keyboard, reporting 9 if they saw 9 or more. There were no time-outs, no response feedback, and subjects were instructed to self-pace. The experiment took approximately one hour to complete. Subjects were told that there were no "correct" answers, and instructed to count any face, animal or human as long as they "felt they saw some kind of face". Subjects were allowed to take breaks as needed.

All participants provided informed consent prior to participation, in compliance with the Common Rule (45 CFR 46), and this study was assessed as exempt from review by the Institutional Review Board, pursuant to 45 CFR 46.101(b)(2). Participants took between 45-90 minutes complete the study and were paid $20 for their participation. Subjects were not excluded for having corrective lenses, but confirmed to be able to see the screen clearly. One subject reported having vision issues beyond using corrective lenses (possible prosopagnosia, see below).

For the analysis, trials were removed where subjects responded in less than 100 milliseconds (likely to be a mistake), as well as trials where the subject took more than 2 minutes to respond (likely to be taking a break). No subjects were removed due to outlying or erroneous data. Subjects' time to completion varied from approximately 45-90 minutes.

We analyze the effect of different psychophysical conditions with mixed effects ANOVA for all 16 subjects using image seed group, gender, image field of view (FOV) as between-subject factors, and Gaussian filter width as a within-subject factor. Uncorrelated p-values are reported for the two ANOVA analyses (image seed and gender). FOV ANOVA is omitted due to unbalanced sample sizes (14 and 2). No significant differences were found in responses for the two subject groups that were shown image sets from two different random seeds ($p > 0.2$, Fig. 13, Left). This indicates that peak pareidolia is not an artifact of a 'lucky draw' from the random image set generated.

In addition to the peak in number of faces reported, we also found that subject's response time (RT) mirrored a similar curve for low to medium filter widths. However, for the largest filter widths shown, the RT curve did not fall off as steeply as the pareidolia curve (Fig. 13, Right). This indicates that subjects took time to look for faces in high frequency data and did not simply 'give up' on the task.

A.3 Additional Human Psychophysics Results

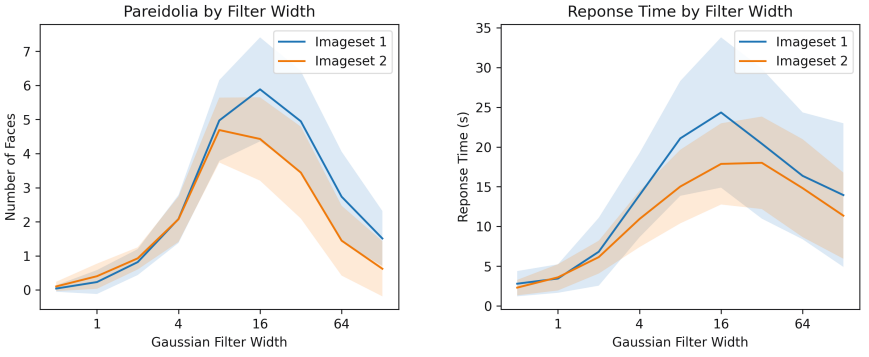

Fig. 13: Left: Pareidolia reported for two subject groups with different random seed images shows no significant difference between groups. Right: Response time (RT) mirrors number of faces for small to medium filter widths, but does not fall off as sharply. Plots report ±1 standard deviation.

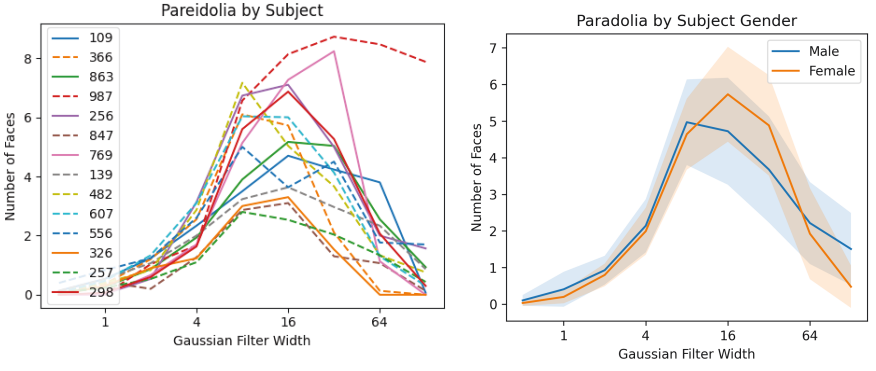

Fig. 14: Left: Pareidolia by subject. Right: Female and Male subjects both demonstrate peak pareidolia, and do not show significant differences. Plot reports ±1 standard deviation.

Furthermore, no significant differences were found between male and female subjects ($p > 0.8$, Fig 14, Right). We note that previous work found females have stronger pareidolic-like neural responses [37]. Interestingly, one subject self-reported having difficulty recognizing face identity (possible undiagnosed prosopagnosia), yet still demonstrated peak pareidolia, though reported fewer faces than most other subjects (Fig. 13, Left, Subject ID 326).

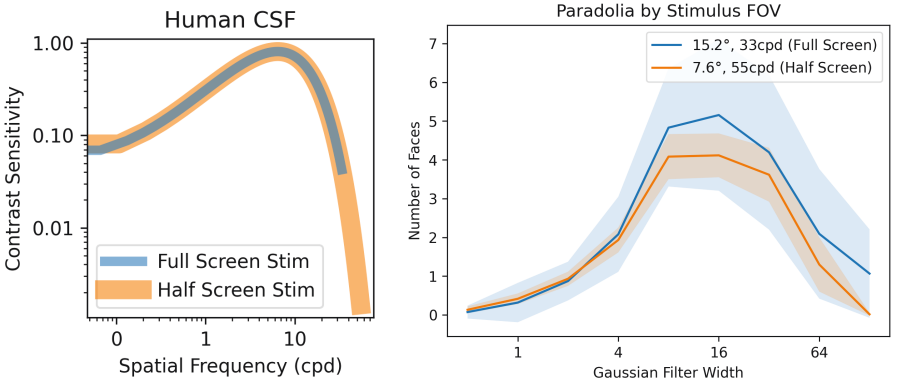

Fig. 15: Left: Human Contrast Sensitivity Function for 1024x1024 image shown at full screen height Field of View (full FOV, blue), and at half size (half FOV, orange). Right: Peak pareidola measured in humans for stimuli viewed at full and half FOV. Full Screen data for 14 subjects, Half screen data for two subjects. Plot reports ± 1 standard deviation.

Despite finding similar peak pareidolia between humans and and machines, one unaccounted difference between their visual systems is in the Contrast Sensitivity Function (CSF). The CSF describes the attenuation of spatial frequency sensitivity for humans at frequencies below and above around 10 cycles per degree (cpd). We calculate the CSF for the range of our experiment assuming the viewing distance of 30 inches, and the scotopic-mesiopic viewing conditions of our setup using the CSF equation from [32]. We plot this for the full screen viewing experiment in Figure 15, Left, blue line.

Because our model posits that high frequencies disrupt pareidolia by making mode alignment unlikely, we explored whether the human CSF, which reduces sensitivity to high frequencies, might explain the results. To determine if the CSF had a measurable effect on the human pareidolia results in our experimental setup, we ran two additional subjects in the same experiment, but with images presented at half the screen width, and therefore with half Field of View (FOV). This increased the Nyquist frequency from 33cpd to 55cpd (screen monitor limit). The Human CSF for this range of frequencies is shown in Figure 15, Left, orange line. We find that for the two subjects tested, results trend to

similarly to the full FOV condition (Figure 15, Right) demonstrating that peak pareidolia is not solely caused by the human CSF.

A.4 Additional Annotation Details

Each image in the Faces in Things dataset is annotated by the following series of questions:

1. Is there a face?
   (Yes / No / Several)
2. If Yes / Several: draw bounding box over face(s)
3. Is the face difficult to spot?
   (Easy / Medium / Hard)
4. Was the face generated by accident, or by design?
   (Accident / Design)
5. What emotion does the face show?
   (Neutral / Happy / Sad / Surprised / Angry / Disgusted / Scared / Other)
6. What does the face most resemble?
   (Human-Baby, Human-Child, Human-Adult, Human-Older, Alien, Animal, Cartoon, Robot, Other)
7. What gender do you think the face is?
   (Neutral / Female / Male)
8. Is this example of pareidolia amusing?
   (No / Somewhat / Yes)
9. How common is this type of face pareidolia?
   (Uncommon / Somewhat / Common)

For consistency, a single annotator was tasked with annotating raw data until a set of five thousand face-containing images had been collected. Data was then manually checked by the authors to correct errors, confirm the reasonableness of the annotations and to flag any faces that were considered unsafe for viewing. Duplicates were automatically detected and removed by thresholding the similarity between or DINOv2 class tokens at 0.85. We note that the subjectivity of this task and that the annotations represent a biased view of the dataset from a single annotator's perspective. While it would be interesting to annotate the same data with multiple annotators to build a distribution of answers and model this subjectivity, it was outside the scope of our current project and we leave it as a direction for future work.

A.5 Additional Average Face Renderings

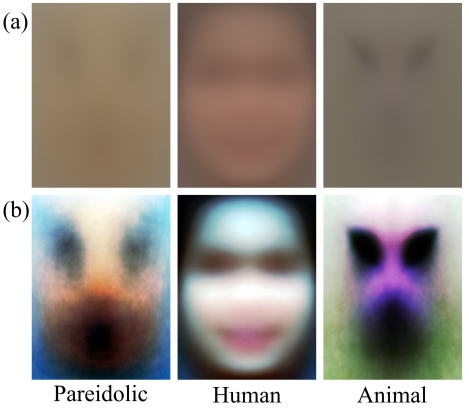

Fig. 16: The Appearance of an Average Pareidolic Face. Shown here are the (a) raw average and (b) per-channel histogram-equalized average images for registered pareidolic faces (our Faces in Things dataset), human faces (samples from the WIDER FACE dataset [62]), and animal faces (AnimalWeb [24]). The average pareidolic face, while less distinct than human or animal, has surprisingly clear eye, nose, and mouth features, and vertical symmetry.

A.6 Average Faces across Different Conditions

Out of curiosity, we plot in Fig. 17 the histogram-equalized averages for faces in our dataset that have been classified as Happy (31% of the data) or otherwise (Neutral / Sad / Surprised / Angry / Disgusted / Scared / Other).

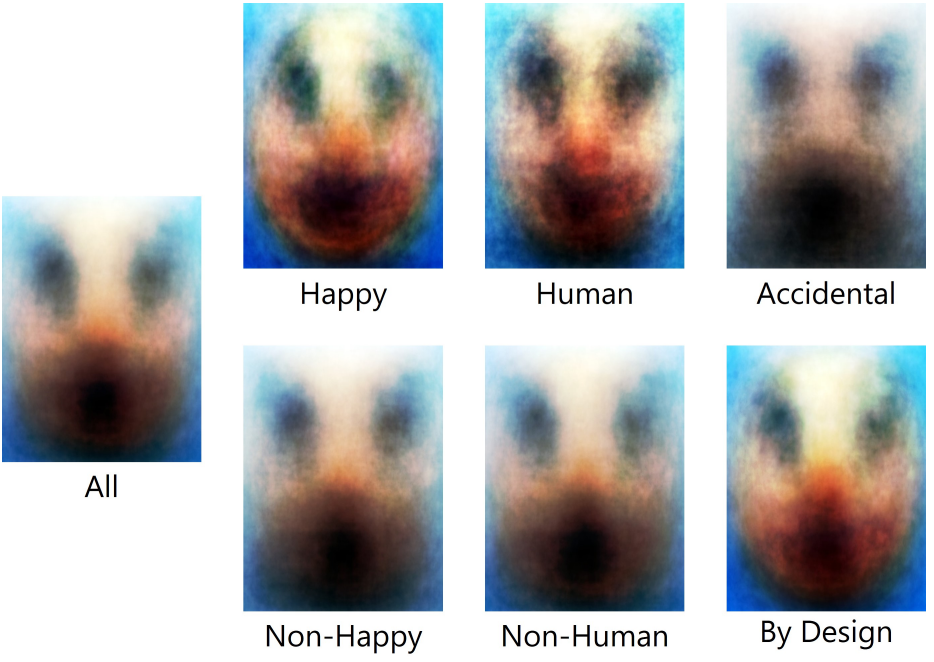

Fig. 17: Average faces from the Faces in Things dataset that fit certain label criteria.

A.7 Why focus on Pareidolia?

Many computer vision researchers are inspired by the human visual system and its ability to robustly recognize patterns in the world. Face pareidolia is fascinating because it is a human visual representation phenomenon that is not well understood. Our dataset, models, and experimental analyses shed light on how and why it might arise. These contributions may help the community to: better understand and harness human visual attention (which is drawn towards face-like objects), reduce pareidolic false positives in face detectors, help designers avoid or create pareidolia, improve pareidolic animation, and create systems that understand how we perceive the world.

A.8 Analyzing the Viola-Jones face detector.

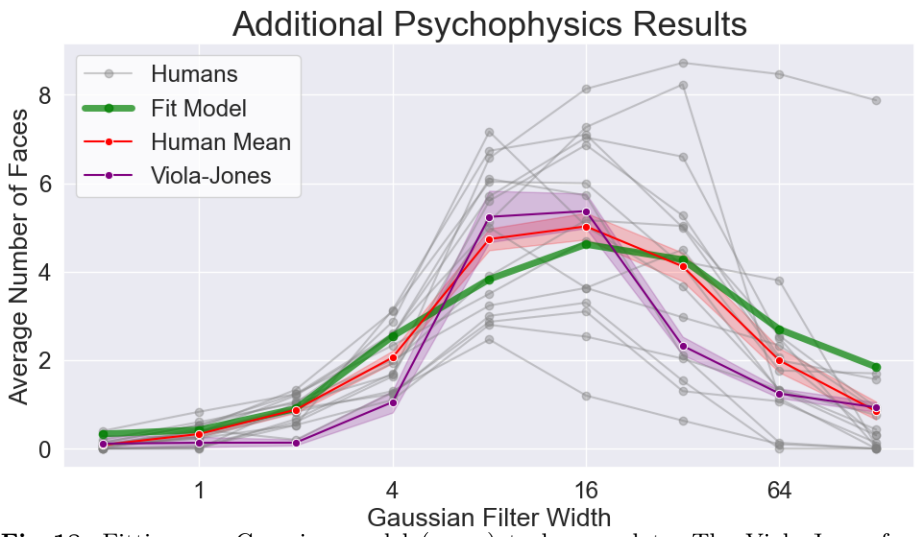

Fig. 18: Fitting our Gaussian model (green) to human data. The Viola-Jones face detector also shows peak pareidolia (purple).

We perform a set of additional experiments with the Viola-Jones face detector and find that it also displays a "pareidolic peak" as seen in in Figure 18.

A.9 Fitting the Gaussian model of Pareidolia to human experiments.

We fit our Gaussian model to human data in Figure 18. The fit parameters ($\sigma = 6$) are similar to the flot of Figure 9 plots ($\sigma = 10$).

A.10 Simultaneous Classification and Detection.

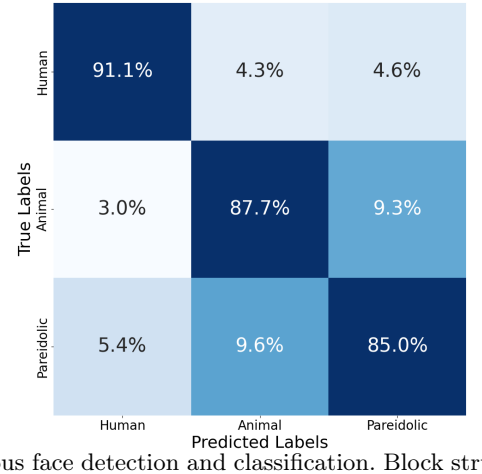

Fig. 19: Simultaneous face detection and classification. Block structure shows similarity between animal and pareidolic faces.

We show a simultaneous classification analysis in Figure 19. We thank the reviewer as this further quantifies our findings that models are more likely to confuse animal and pareidolic faces with each other than with human faces.