

# On Human-like Biases in Deep Neural Networks for the Perception of Slant from Texture

YUANHAO WANG, Brown University Department of Computer Science, , USA  
QIAN ZHANG, Brown University Department of Computer Science, , USA  
CELINE AUBUCHON, Brown University Department of Cognitive, Linguistic, and Psychological Sciences, , USA  
JOVAN KEMP, Brown University Department of Cognitive, Linguistic, and Psychological Sciences, , USA  
FULVIO DOMINI, Brown University Department of Cognitive, Linguistic, and Psychological Sciences, , USA  
JAMES TOMPKIN, Brown University Department of Computer Science, , USA

Depth estimation is fundamental to 3D perception. This study investigates whether unsupervised deep neural networks (DNN) might display the same biases as humans when predicting the sign of curvature and depth of surfaces under different viewing conditions (field of view) and surface parameters (slant and texture irregularity). To this end, we trained both unsupervised and supervised DNN models on the renderings of slanted surfaces with random Polka dot patterns and analyzed their internal latent representations. The results demonstrated that the unsupervised models replicate human biases well across all experiments, while supervised DNN models do not exhibit similar biases. The latent spaces of the unsupervised models can be linearly separated into axes representing field of view and optical slant. For supervised models, this ability varies substantially with model architecture and the kind of supervision (continuous slant vs. sign of slant). These findings suggest that unsupervised DNN models can share similar predictions to the human visual system, leading to potential use in generating and analyzing hypotheses about depth and slant perception for future human testing.

CCS Concepts: • **Applied computing**;

Additional Key Words and Phrases: Perception, Slant, Texture

## ACM Reference Format:

Yuanhao Wang, Qian Zhang, Celine Aubuchon, Jovan Kemp, Fulvio Domini, and James Tompkin. 2023. On Human-like Biases in Deep Neural Networks for the Perception of Slant from Texture. 1, 1 (May 2023), 9 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

## 1 INTRODUCTION

Deep neural networks have achieved success in a wide range of applications, such as image and speech recognition, natural language processing, and game playing. Since deep neural networks were originally inspired by the structure and function of the human brain, comparing deep neural networks to the human cognitive system has been an area of interest for researchers in the field of artificial intelligence and cognitive science.

The human visual system is a complex network of biological structures with remarkable but imperfect capabilities. Some works have tried to evaluate convolutional neural networks (CNN) as explanatory models for human vision using simulated psychophysical studies. A recent study by Storrs et al. [2021] considered the ambiguity between the perception of glossiness and surface curvature,

where low-gloss high-curvature surfaces look the same as high-gloss low-curvature surfaces. They found that unsupervised neural networks made similar predictions to the human visual system in gloss perception: unsupervised networks were able to reproduce specific patterns of success and failure in distinguishing high and low gloss images commonly made by humans. Supervised networks trained to predict high or low gloss did not share this property.

Does the finding that unsupervised networks can exhibit human-like biases hold in visual tasks other than gloss perception? We try to answer this question in the context of depth perception, specifically in the task of estimating the perceived slant of textured surfaces. We consider simple Polka dot textures. Even with no other depth cues like disparity or shading, human beings can estimate surface slant because the texture deforms under perspective. Such a setting lets us investigate three additional factors influencing the estimation of slant from texture: the field of view (FOV), the sign of the surface curvature (concave or convex), and texture pattern regularity.

With regard to depth perception, it is well-documented in vision science that humans are prone to bias [Campagnoli et al. 2022; Domini and Caudek 2003; Johnston 1991; Langer and Siciliano 2015; Liu and Todd 2004; Todd et al. 2007; Watt et al. 2005]. In judging slant from texture, previous research by Todd et al. [2005] found evidence for four biases. First, the perceived sign of curvature of a surface became ambiguous when the FOV was small. Second, an increase in FOV produced a corresponding increase in the magnitude of the perceptual gain (i.e., the judged slant divided by the ground truth). Third, humans perceive more depth from convex surfaces than from concave surfaces. Finally, there is a greater perceptual gain when the surface texture pattern is more regular.

To test whether unsupervised deep neural networks share these biases, we followed the psychophysical experiment settings in Todd et al. [2005], and generated synthetic renderings depicting surfaces with concave or convex dihedral angles, varying physical slant, random Polka dot textures, and a varying FOV. Then, we trained unsupervised generative models capable of reconstructing input images of such surfaces to learn the statistical regularities in the training data. From analyzing the learned network latent spaces, our study reveals that unsupervised models can reproduce human-like biases. They exhibit a higher error in judging the sign of curvature when FOV is smaller, perceive greater slant when FOV is increased, perceive more slant in convex surfaces than concave surfaces, and when texture regularity level is increased. These results are consistent with the findings in human depth perception research, and

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).  
© 2023 Copyright held by the owner/author(s).  
XXXX-XXXX/2023/5-ART  
<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

suggest a similarity in the predictions made between the unsupervised deep neural networks and the human visual system (to say nothing of the mechanisms for those predictions). Across four different neural network architectures, we find some variation but overall similar trends in bias.

For the evaluated deep neural network models trained with supervision of the signed continuous-valued surface slant, we discover that models do not exhibit bias on test stimuli. When considering the latent space separation of curvature sign, physical slant angle, and FOV, we find more significant differences across architectures than in unsupervised models, with some architectures leading to good factor separation and others not. Further, we also train a set of models with weaker supervision only of the sign of the surface—concave or convex. This mirrors the ‘binary’ high/low gloss choice in the work of Storrs et al. [2021]. With weaker supervision, models still do not exhibit bias on test stimuli, and the models’ latent factors are less well separated for all architectures. This suggests that latent space visualizations must be read carefully and not independently—supervised models with appropriate architectures and labels can still factor physically-meaningful variables even if their predictions are unbiased, and that models trained on impoverished labels (continuous slant vs. sign of slant) may lead to false assumptions about the ability of an architecture and training routine to factor variation.

## 2 BACKGROUND

### 2.1 Human perceptual biases in depth estimation

Understanding the mechanism of human depth perception is essential in 3D vision research. Humans integrate many sources of information to estimate depth, including binocular disparity, texture, shading, defocus, and motion, ultimately forming a three-dimensional percept of an object. A body of research has emerged to study human perceptual biases in depth judgment, and these biases can be categorized based on their associated visual cues.

For disparity, despite retinal and extra-retinal cues often providing sufficient information to achieve veridical perception, the visual system still produces errors. Johnston et al. [1991] showed that the veridicality of human perception depends on the distance, with objects appearing elongated at a close viewing distance and flattened at a far distance. Ambiguities also exist within focus/de-focus cues: the sign of depth is ambiguous [Watt et al. 2005], and increasing blur gradient away from fixation point increases perceived slant [Langer and Siciliano 2015]. Liu et al. [2004] reported that participants exhibited biases in shape from shading as they misperceived convex surfaces as deeper. In addition, adding more sources of information from shading (i.e., specular highlights and cast shadows) increased perceived depth for convex surfaces. When using motion as a cue, the perceived depth depends on the deformation component of the optic flow field. This information is ambiguous and can lead to biases in depth perception [Domini and Caudek 2003]. Although increasing the number of available cues can potentially disambiguate depth information and lead to veridical perception, it has been shown that adding cues increases perceived depth without necessarily making it more accurate [Campagnoli et al. 2022].

The cue that is important to our study is texture. Studies have found that texture regularity, the field of view, and the sign of surface

curvature can lead to perceptual biases in judging slant from texture. For instance, humans tend to perceive more slant from regular textures or textures with discrete elements. Compressing elements along one direction also increases perceived slant ([Todd et al. 2005, 2007]). Convex surfaces appear to elicit greater slant responses than concave surfaces, and large fields of view produce greater amounts of perceived slant than small fields of view. Additionally, humans are more prone to making errors in judging the signs of surface curvature when the FOV is small.

### 2.2 Unsupervised models may predict human perception

Many of the key ideas in machine learning took inspiration from the biological findings in the human brain. Most notably, neural networks mimic the design of interconnected biological neurons that send electrical signals to each other in a brain, and the convolutional neural network (CNN) was inspired by the hierarchical structure of the ventral visual pathway. Naturally, evaluating deep neural networks (DNN) as a model of the visual system has been a research area of interest. Many studies have found that DNNs trained for object recognition are good at predicting the representations of images in high-level ventral visual areas of the human and nonhuman primate brain [Kubilius et al. 2019; Lindsay 2021; Ponce et al. 2019; Schrimpf et al. 2018; Xu and Vaziri-Pashkam 2021].

The work most relevant to our methodology is that of Storrs et al. [2021]. The authors investigated the connection between intermediate representations in unsupervised models and the patterns of ‘success’ and ‘failure’ in human perception of gloss. They trained a variational auto-encoder (VAE) on a synthetic dataset consisting of renderings of bumpy surfaces with either high or low specular reflectance and found disentanglement of distal scene properties in the model’s latent space. Then, they trained a linear support vector machine classifier to generate quantitative gloss predictions. The authors found that, surprisingly, the latent codes of the unsupervised generative model could be used to predict human bias of gloss perception better than supervised networks or other control models. In our work, we try to test if unsupervised generative networks could also predict human bias related to slant estimation as well.

## 3 METHOD

To test whether unsupervised models exhibit human-like perceptual biases for slant estimation from texture cues, we first rendered a dataset of stimuli images. Next, we trained unsupervised generative deep neural network models to implicitly learn the statistical distributions of the data, along with supervised equivalent models for comparison. Then, we analyze the internal latent representations of each model to a) evaluate how stimuli are laid out within it, and to b) assess whether simple distance measures from linear classifiers on the latent space can produce biased outcomes.

### 3.1 Synthetic data

We produced an *in silico* replica of the real-world human psychophysical experimental setup of Todd et al. [2005] (Fig. 1a). We used a surface with concave or convex dihedral angles that were bilaterally symmetrical about the vertical axis; whether the surface is concave

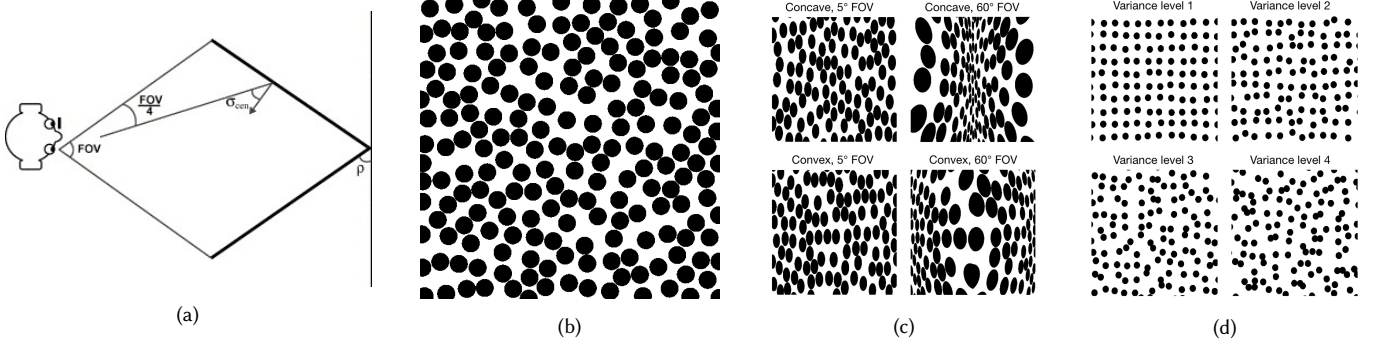


Fig. 1. **Reproducing slant psychophysical experiments *in silico*.** (a) A schematic top-view representation of the physical scene geometry used to depict the stimuli in the human psychophysical study (reproduced from Todd et al. [2005]). (b) Random dot pattern projected onto a flat surface for reference. (c) Examples of our synthetic reproduction of the stimuli with a consistent optical slant of  $60^\circ$  and different dihedral angles and field of view (FOV). (d) Examples of grid textures with varying regularities. Images have variance levels from 1 to 4.

or convex is referred to as the sign of curvature. The virtual perspective camera was positioned at the center front of the surface, and its distance was adjusted to capture the entire width of the surface. Note that, in this setting, the surface slant ( $\rho$  in Fig. 1a) is directly correlated with the relative depth of the scene. For consistency and clarity, we choose to use the term slant for the rest of the paper. For a human being to successfully estimate the physical slant, they must be able to separate the effects of varying view angles and slants on the deformations of the textures as no other cues exist.

We generated stimuli with control over two parameters: the field of view (FOV), and the optical slant at the center of each surface ( $\sigma_{cen}$ ). FOV ranged between  $5^\circ$  and  $60^\circ$ , and the values of  $\sigma_{cen}$  ranged between  $25^\circ$  and  $60^\circ$ . The ranges of optical slants for the concave and convex surfaces were matched (maximum value  $\sigma_{max} = \sigma_{cen} + FOV/4$ , minimum value  $\sigma_{min} = \sigma_{cen} - FOV/4$ ). However, the physical slants ( $\rho$ ) defined as  $\rho = \sigma_{cen} + FOV/4$  for concave surfaces, and  $\rho = \sigma_{cen} - FOV/4$  for convex surfaces had mismatched range.

For each combination of FOV,  $\sigma_{cen}$ , and curvature sign, we generated 10 random black and white Polka dot textures (Fig. 1b). The dots were uniformly distributed with no overlaps and had the same size. We mapped each pattern onto the surface and rendered the scene using a perspective camera. The dataset consisted of 2000 images each of  $256 \times 256$  pixels, and all images were generated using Python. Figure 1c shows stimuli of different convexity and FOV.

To further examine the impact of surface texture regularity on perceived slant from texture, we also generated stimuli with different Polka dot regularity (Fig. 1d). We began with a grid of uniform dots. Then, we shifted the center of each dot by  $\alpha$ , where  $\alpha \in R^2$ ,  $\alpha \sim Uniform(-b, b)^2$ . This allowed us to manipulate the regularity of the texture by adjusting  $b$ . In our experiments, we used 5 levels of variances/irregularities, ranging from level 0 (perfect grid) to level 4 (most irregular). As before, we mapped the surface dot patterns by the dihedral angles. We used FOV values in the range between  $5^\circ$  and  $60^\circ$ , and  $\sigma_{cen}$  values between  $25^\circ$  and  $60^\circ$ . This dataset comprised 10,000 images each of  $256 \times 256$  pixels.

### 3.2 Unsupervised generative model

Unsupervised generative models are a class of neural networks trained to reproduce high-dimensional inputs. When trained on a large number of data points sampled from a distribution, their low-dimensional latent vectors are forced to encode the distribution as efficiently as possible. We trained generative models to reconstruct input 2D images. Our models are all auto-encoders (Fig. 2a): an encoder compresses the input image to a low-dimensional latent space (often called a *bottleneck*), and a decoder restores the original input from the latent space. We evaluated several variants to investigate whether the study's findings were architecture-independent, with the primary architecture being the common U-Net [Ronneberger et al. 2015]. The model architectures are:

- (1) **VGG-based auto-encoder (VGG-AE)**: An auto-encoder that uses the VGG16 architecture [Simonyan and Zisserman 2014]. The encoder uses max pooling to downsample, and the decoder uses bilinear upsampling.
- (2) **Variational auto-encoder (VAE)**: We use the VAE proposed by Kingma et al. [2013]: Instead of passing the latent vector directly to the decoder, we add to the latent vector random Gaussian noise with learned distribution parameters (Fig. 2b). Storrs et al. [2021] used PixelVAE, a variant to address blurry samples, but the principle is the same.
- (3) **U-Net** [Ronneberger et al. 2015]: This auto-encoder adds residual connections between equivalent-spatial-sized layers of the encoder and the decoder networks. This lets high-resolution information pass directly from the encoder to the decoder, bypassing the bottleneck.
- (4) **U-Net-**: This model removes the residual connections between the encoder and the decoder, so that all the information passed to the decoder is contained in the latent vector.

### 3.3 Supervised model

We conducted experiments with supervised DNN models using two different architectures, namely the ResNet ([He et al. 2016]) with 18 layers and a U-Net-based model that utilized the encoder of the U-Net. We augmented both architectures with an additional dense layer preceding the final layer, and treated its output space



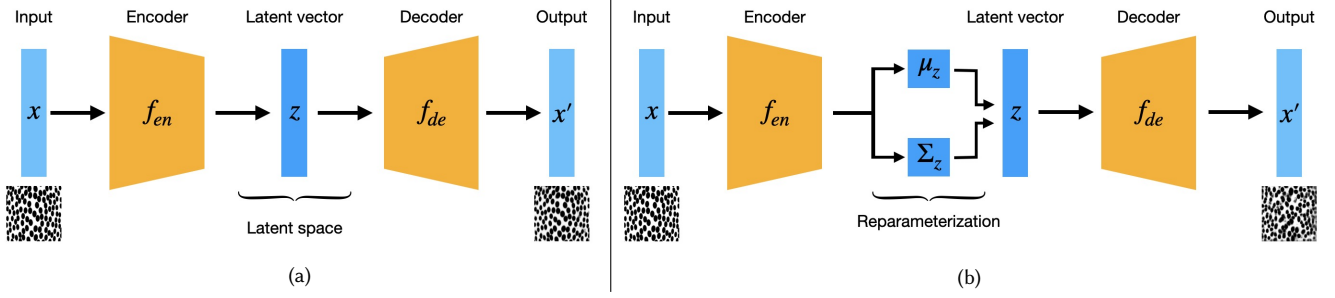


Fig. 2. Schematic illustration of neural network models: (a) auto-encoder and (b) variational auto-encoder.

as the latent space similar to that of the unsupervised model. The models were trained using both stronger supervision, with ground truth physical slant labels, and weaker supervision, with ground truth sign of curvature labels. This allowed us to not only compare the behaviors of the supervised and unsupervised models, but also examine the impact of different architectures and training objectives on the outcomes of the supervised models.

### 3.4 Network training and losses

The unsupervised objective was to reconstruct the input images. We penalize an L2 reconstruction loss, i.e., the sum of the squared difference between each pair of matching pixels in the reconstructed and input images. For supervised models, we have two settings. 1) We ask the network to predict the signed physical slant of the surface, as humans do in Todd et al. [2005]; and 2) We ask the network just to predict the sign of the physical slant of the surface; this mirrors the ‘binary’ high/low gloss choice in the work of Storrs et al. [2021].

Data are split into training and testing images with a 80/20% split. We train each model for 100 epochs using the Adam optimizer, with a learning rate of  $5 \times 10^{-4}$ ,  $\beta_1 = 0.5$ , and  $\beta_2 = 0.999$ . As model training can show variability, we train an ensemble of 10 instances of each model and average their outputs. After training, we fed unseen test images to the trained encoder to extract the latent vectors.

### 3.5 Methods of analysis

*Sign of curvature prediction.* Unsupervised generative models cannot make predictions given stimulus. However, it is possible to define a classifier upon the latent space. This assumes that a generative model can arrange stimuli in the latent space according to their statistical properties. For example, stimuli with the same sign of curvature may form clusters. Ideally, the physical properties form simple continuous arrangements. This would allow boundaries to be drawn using a linear classifier such as a Support Vector Machine (SVM); it being linear allows only simple arrangements in the latent space to lead to meaningful interpretation. If stimuli are misplaced in the latent space, such as a concave stimulus being within convex stimuli, we can interpret this as an ‘error’ in judging convexity.

*Magnitude of perceived slant.* How far the latent code of a stimulus lies from the decision boundary can potentially be used as a measure of perceived slant. We compute the Euclidean distance of each latent code to the decision hyperplane, which we term the ‘latent distance’. We suppose that the latent distance is positively correlated with the magnitude of the perceived slant, and that stimuli with latent vectors

lying on the decision boundary may be considered by the model as flat. A larger latent distance indicates that the model perceives the surface as more slanted, so that the model is more confident at predicting its sign of curvature. Although the numerical value of the latent distance does not have physical meaning, we will use it to compare the perceived slants of different stimuli within the test set.

## 4 EXPERIMENTAL FINDINGS

### 4.1 Physical factors are disentangled in unsupervised model latent spaces

We are interested in the extent to which the unsupervised model can learn to disentangle our physical factors of interest in its latent representation—sign of curvature, field of view, optical slant, and texture regularity. First, we trained models on the dataset without texture irregularity discrepancies. After compressing the latent vectors to 2D using principal component analysis (PCA) and t-distributed stochastic neighbor embedding (tSNE), we observed grouping of concave and convex surfaces when choosing a suitable latent dimension (Fig. 3a). However, entanglement would occur if the latent dimension was too small or too large. Furthermore, FOV and optical slant values varied smoothly within each cluster along two linearly separable axes, suggesting that both variables were well-disentangled in the latent space of the model.

Next, we trained the unsupervised model on the dataset with varying levels of texture irregularities and visualized the latent spaces of all four models in Fig. 3b. Notably, the U-Net model was unable to disentangle texture irregularity levels in its latent space. However, the same model could learn to disentangle texture irregularity if the residual connections between the encoder and the decoder were removed. In fact, all the unsupervised models without residual connections tested (U-Net-, VGG-AE and VAE) exhibited discernible clusters based on the irregularity levels in the latent space. Images with less regular textures tended to concentrate in the middle region of the latent space, while images with more regular textures were more dispersed, indicating that the models were more capable of distinguishing input images with more regular textures.

We conducted a quantitative analysis on latent space clustering effects. Representational similarity analysis [Nili et al. 2014] showed that, for unsupervised models, pairs of images with the same convexity were represented by more similar vectors in the latent space than pairs of images with different convexity (Fig. 3c; T-test comparing average Euclidean distances between same-convexity versus

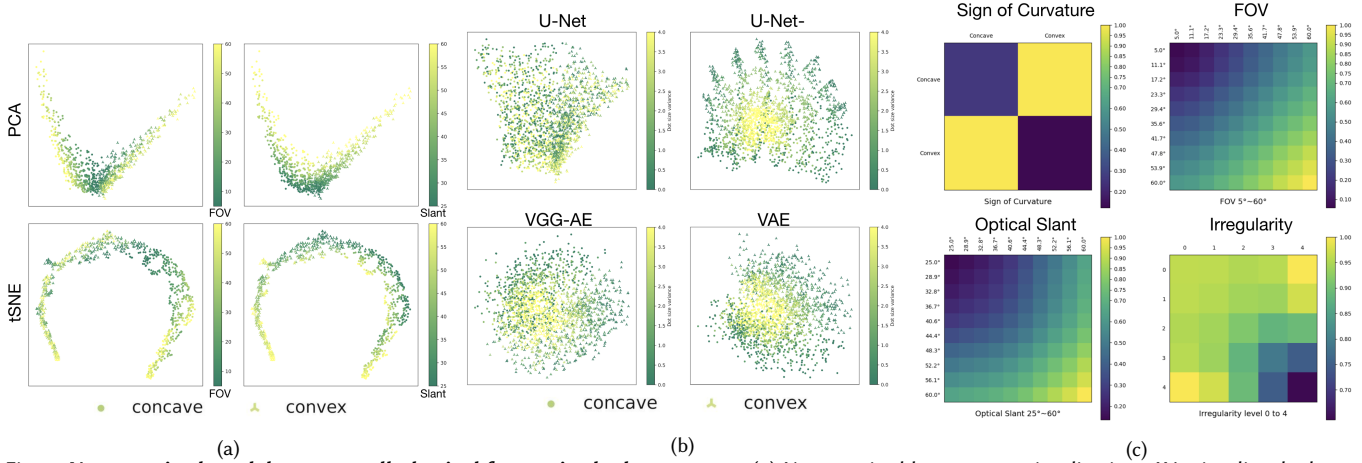


Fig. 3. **Unsupervised model captures all physical factors in the latent space.** (a) Unsupervised latent space visualizations. We visualize the latent space of the U-Net using PCA (first two principal components) and tSNE for dimensionality reduction (rows). Data points are color-coded by the FOV and optical slant values (columns). The unsupervised latent space disentangles surface convexity, FOV, and optical slant successfully. (b) Unsupervised latent space visualizations when trained on the dataset with varying texture irregularities. We visualize the latent spaces of all four unsupervised models, with data points color-coded by the texture irregularity level. All three alternative models disentangle texture irregularity levels, while the U-Net fails to do so. (c) Representational dissimilarity matrices showing mean distances between all pairs of latent vectors, grouped by the sign of curvature, FOV, optical slant values, and texture irregularity levels. Distances are defined as  $1 - corr$ , normalized to the range from 0 to 1, where  $corr$  is the Pearson correlation coefficient. Notably, the bottom right matrix uses U-Net while the rest uses the default U-Net. With the exception of U-Net with texture irregularity, the unsupervised models exhibit strong clustering by all factors.

Table 1. Correlations between FOV, optical slant and the first two principal components of the latent vectors.

	Optical slant	FOV	FOV (concave)	FOV (convex)
1st P.C.	0.819	0.078	-0.190	0.239
2nd P.C.	0.480	0.252	0.551	-0.342

different-convexity image pairs:  $t = -200.20$ ;  $P < 0.001$ ; Cohen's  $d = -0.71$ ; 95% confidence interval (CI) of difference,  $-21.50 - -21.09$ . Regarding FOV and optical slant, it was observed that images with similar FOV or optical slant values tended to have smaller latent distances; however, the dominant trend was that the latent representations became more diverse as FOV or optical slant increased (strong correlations between the FOV/optical slant and the averaged representational dissimilarity:  $r = 0.979$  for FOV and  $r = 0.956$  for optical slant). Moreover, the latent representations were more dissimilar when textures were more regular, which corroborated our prior observations from examining the latent spaces.

To quantify the 'smoothness' of latent space variations, we computed the strength of correlation between FOV, optical slant and the first two principal components of the latent vectors. There was strong correlation between the optical slant and the first principal component ( $R = 0.819$ ), and moderately strong correlation between FOV and the second principal component when divided into concave and convex groups (Tab. 1; concave:  $R = 0.551$ , convex:  $R = -0.342$ ). These results suggest that the unsupervised latent space smoothly captures FOV and optical slant variations.

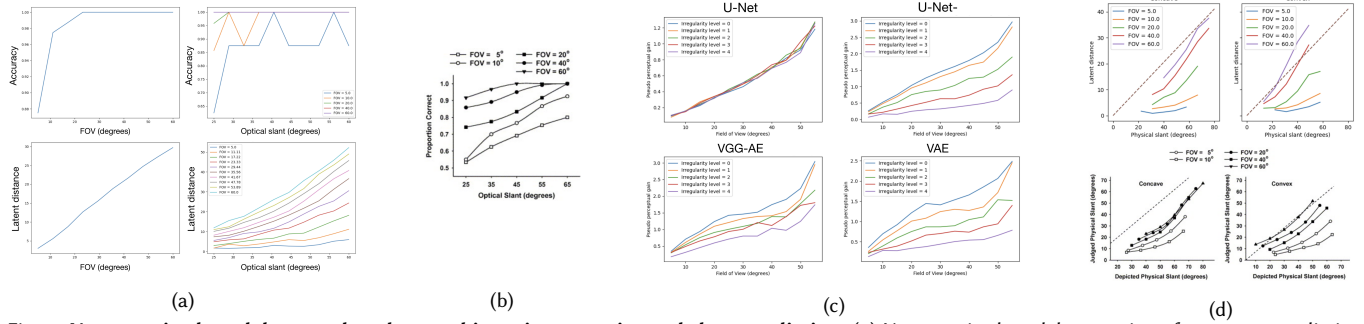
## 4.2 The sign of curvature is ambiguous when the field of view is small

The linear SVM lets us determine the judged sign of curvature for each test instance to calculate a classification accuracy. The overall accuracy across the 10 instances of the unsupervised model is 96.4% ( $\pm 0.91\%$ ). The model can predict the sign of curvature perfectly when the FOV is greater than  $25^\circ$ , but shows more errors when the FOV is smaller (Fig. 4a). The accuracy also increases with the optical slant. We found moderately strong correlation between FOV and the mean classification accuracy ( $r = 0.674 \pm 0.028$ ), and stronger correlation between the optical slant and the mean classification accuracy ( $r = 0.813 \pm 0.027$ ). These findings are consistent with those from Todd et al. [2005] (Fig. 4b). Like humans, unsupervised models were more likely to misjudge the sign of curvature when the FOV was small. All supervised models were able to make classifications with 100% accuracy. Hence, no such bias could be inferred.

## 4.3 Unsupervised models perceive more slant with greater FOV, optical slant and convex surfaces

We use latent distance as a proxy for the magnitude of the perceived slant (Sec. 3.5). Experiments showed that the latent distance had a strong linear correlation with the FOV ( $r = 0.998 \pm 0.001$ ; Fig. 4a (lower left)), meaning the model perceived more slant as the FOV increased. Again, this is in line with the human perceptual bias (Sec. 2.1). Furthermore, the averaged latent distance was also positively correlated with the optical slant for each FOV value ( $r = 0.980 \pm 0.002$ ; Fig. 4a lower right).

For the human bias that perceives more slant from convex surfaces than concave surfaces, we observed a similar bias: the unsupervised model exhibited a systematic bias towards perceiving more



**Fig. 4. Unsupervised models reproduce human biases in convexity and slant prediction.** (a) Unsupervised model mean sign of curvature prediction accuracy and mean latent distance (rows) as functions of FOV and optical slant per FOV (columns). The mean prediction accuracy is 100% when the FOV is above  $25^\circ$  and increases with FOV and optical slant when the FOV is below  $25^\circ$ . The mean latent distance increases with both FOV and optical slant. (b) Results from human psychophysical study ([Todd et al. 2005]). Humans accuracy in judging the sign of curvature also increases with both FOV and optical slant. (c) The pseudo perceptual gain as a function of the FOV per irregularity level for each tested model. With the exception of the U-Net, all other models produce greater pseudo perceptual gains with more regular textures. (d) Top row: unsupervised model's latent distance as a function of the physical slant in concave and convex cases; bottom row: human psychophysical study results from [Todd et al. 2005]. Both the unsupervised model and the human obtain larger perceptual gains from convex surfaces than from concave surfaces.

slant from convex surfaces than concave surfaces. The predicted latent distance in the convex cases was generally larger than that in the concave cases (Fig. 4d), suggesting that the model perceived more slant from convex surfaces (cf. equivalent figure from Todd et al. [2005] reproduced beneath). Our qualitative observation is supported by the T-test results on the mean difference of the latent distance between the convex and concave groups, where the physical slant was between  $26.25^\circ$  and  $58.75^\circ$  ( $t = 30.32, p < 0.001$ ).

To better understand the effect of each variable, we computed generalized linear models [Nelder and Wedderburn 1972] separately on the set of concave and convex instances, using physical slant and FOV as independent variables and latent distance as the dependent variable. In both groups, physical slant and FOV are positively correlated with the latent distance with high pseudo R values (Tab. 2). Three-way analysis of variance (ANOVA) showed that FOV, convexity, and physical slant have statistically significant impacts on the latent distance (FOV:  $F_{9,721} = 12.6, p < 0.001$ ; convexity:  $F_{1,721} = 41.7, p < 0.001$ ; physical slant:  $F_{6,721} = 39.0, p < 0.001$ ).

We conducted correlation analyses to identify the particular attributes of texture upon which the model's slant judgments were based. We considered several variables including the length, width, area, and vertical density of the Polka dots. For each variable, we calculated the minimum, maximum, median values, and the range of values. Subsequently, we computed the correlation coefficients between each of the above-mentioned measures and the latent distance (Tab. 3). Results show strong correlations between the latent distance and multiple texture statistics, suggesting that the unsupervised model may use these attributes to represent the data.

#### 4.4 Unsupervised latent spaces disentangle texture regularity

[Todd et al. 2005] demonstrated that more regular textures led to a greater perceptual gain, defined as the human judged slant divided by the ground truth slant (Fig. 4d). Does a similar effect of texture regularity on the slant perception exists in the context of unsupervised learning? We trained both supervised and unsupervised

**Table 2. Generalized linear model results.** The independent variables are physical slant (P.S.) and FOV, and the dependent variable is latent distance.

Convexity	Variable	Coeff.	S.t.d. err.	P-value	Pseudo R
Concave	P.S	0.043	0.007	<0.010	0.844
	FOV	0.125	0.010	<0.001	
Convex	P.S	0.056	0.005	<0.001	0.740
	FOV	0.075	0.005	<0.001	

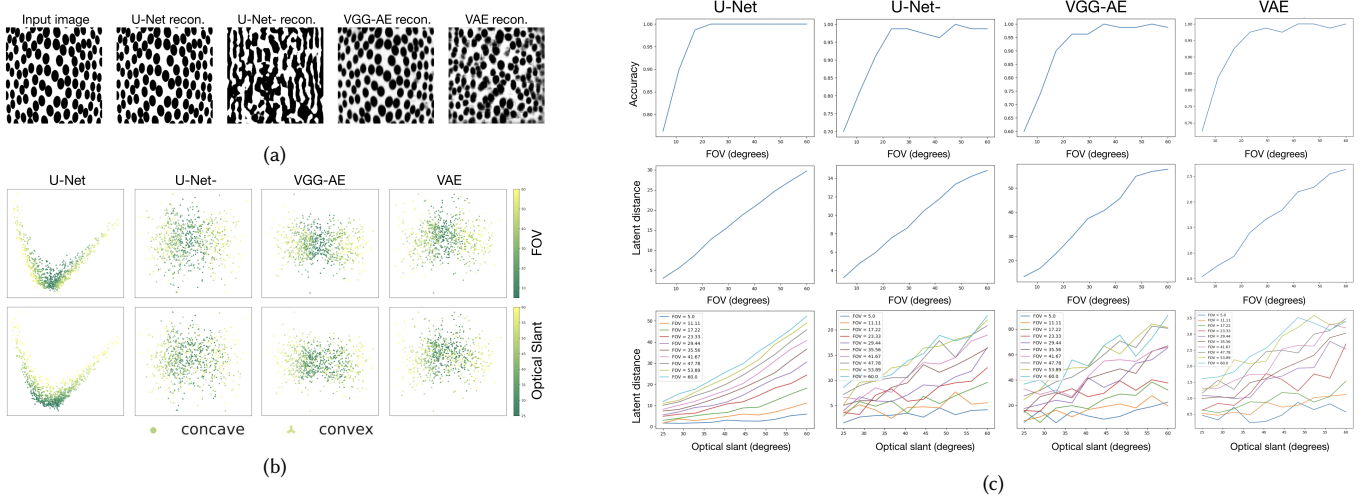
**Table 3. Correlation coefficients (R) between texture attributes and the model's judged slant (latent distance).**

	Length	Width	Area	Spatial Density
Minimum value	0.924	0.852	0.835	-0.858
Median value	0.911	0.739	0.843	0.939
Maximum value	0.852	0.353	0.609	0.945
Range	-0.904	-0.908	-0.819	0.950

models using a dataset consisting of synthetic renderings of polka dot patterns with varying degrees of irregularities. Subsequently, we analyzed the models' responses to subsets of the dataset that corresponded to each level of irregularity.

In Sec. 4.1, we showed that unsupervised models that lack residual connections between the encoder and decoder were able to disentangle texture irregularity levels in the latent spaces. Further investigations have revealed that these models also exhibit a perceptual bias comparable to that of humans. To facilitate the comparison, we define pseudo perceptual gain as the normalized latent distance divided by the normalized physical slant. The pseudo perceptual gain is greater at lower irregularity levels for U-Net-, VGG-AE and VAE, while no such disparity was observed in U-Net (Fig. 4c). Without bypassing the latent bottleneck, unsupervised models produce





**Fig. 5. Effects of unsupervised model architecture.** (a) The reconstruction results of the unsupervised models. Significant differences in reconstruction quality across different models. (b) Latent space visualizations for the unsupervised models. Data points were color-coded by FOV and optical slant values (rows) for all 4 unsupervised models (columns). Sign of curvature, FOV and optical slant are disentangled by all models. (c) Latent space trend comparisons. Each row depicts a plot of a particular parameter against another, with the first row showing the FOV vs. sign of curvature prediction accuracy, the second row the FOV vs. the latent distance, and the third row the optical slant vs. the latent distance per FOV. In each column, the plots are arranged from left to right to correspond to the models U-Net, U-Net-, VGG-AE, and VAE. The general trends are consistent across all models.

greater perceptual gains from more regular textures. Despite differences in the experimental settings, the prediction patterns of humans and unsupervised models are related.

#### 4.5 The effects of model architecture

In the previous section, the experimental results differed significantly for models with different architectural designs (with or without residual connections). To evaluate the generalizability of our findings to a broader class of unsupervised models, we replicate the previous experiments using four architectures (Sec. 3.2).

We observed a significant variation in reconstruction quality across models. U-Net achieved good quality, VGG-AE and VAE suffered from blurriness and artifacts, and U-Net- failed to reconstruct finer details (Fig. 5a). The superior performance of the U-Net can be attributed to the residual connections that allowed direct passage of information to the decoder, allowing the bottleneck to ignore high-frequency information. Without residual connections, U-Net- lacked detailed stimuli reconstruction capabilities.

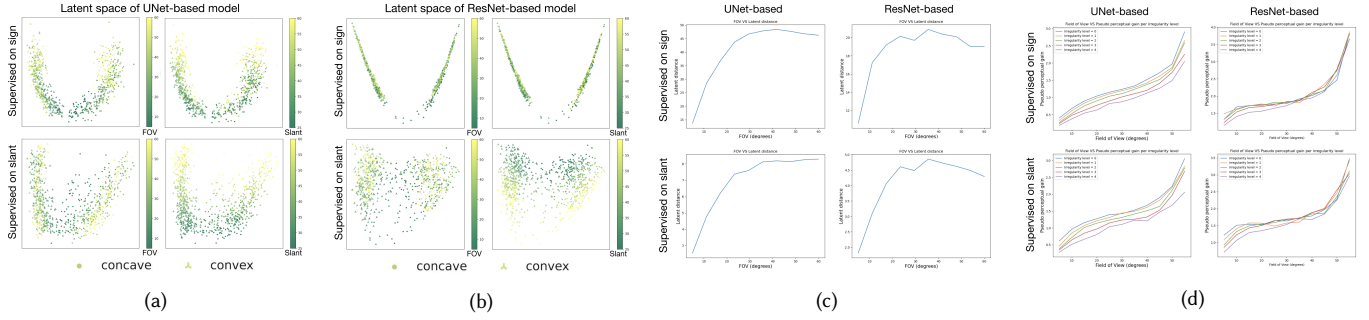
Despite differences in reconstruction, all models were capable of disentangling the sign of curvature in their latent spaces. The latent spaces of U-Net-, VGG-AE and VAE models formed distinct clusters for concave and convex surfaces, albeit without the "V" shape in the PCA plots observed in the case of the U-Net model (Fig. 5b). Additionally, the three alternative models also learned to disentangle FOV and optical slant in their latent spaces. Comparable to the baseline U-Net model, images with lower FOV or optical slant values tended to be located closer to the classification boundary in the latent spaces. Subsequent analysis revealed that differences in same-convexity and cross-convexity representational similarity were statistically significant, but their clustering effects were weaker than those observed in the U-Net model (U-Net-:  $t = -187.57$ ,  $P <$

0.001, Cohen's  $d = -0.67$ ; VGG-AE:  $t = -30.34$ ,  $P < 0.001$ , Cohen's  $d = -0.15$ ; VAE:  $t = -80.81$ ,  $P < 0.001$ , Cohen's  $d = -0.29$ ).

Further analyses indicated that the alternative unsupervised models display comparable trends to the U-Net in predicting the sign of curvature and the perceived slants, despite underlying latent spaces being less well-structured. They all achieved high overall accuracy in surface convexity prediction: 84.0% ( $\pm 1.2\%$ ) for U-Net-, 98.0% ( $\pm 0.1\%$ ) for VGG-AE, and 94.7% ( $\pm 0.2\%$ ) for VAE. Fig. 5c shows that all three unsupervised models exhibit high accuracy when the field of view (FOV) is large, and relatively lower accuracy when the FOV is small. Additionally, the average latent distance increases linearly with FOV and optical slant in all models. Although the optical slant vs. latent distance curves are less smooth for the U-Net-, VGG-AE and VAE due to the sub-optimal behavior of their learned latent spaces, the general trends are still observable. Moreover, T-tests show that the alternative unsupervised models are systematically biased towards perceiving more slant from convex surfaces (U-Net-:  $t = -30.46$ , VGG-AE:  $t = -35.5$ , VAE:  $t = -33.7$ ;  $p < 0.001$  in all cases). In conclusion, all the perceptual biases observed in the U-Net hold true for the alternative models.

#### 4.6 Supervised models produce unbiased outcomes, but the latent behaviors depend on the training setting

Supervised models with the objective to predict the provided labels are unbiased when successfully trained. In our experiments, we found that when supervised with sign of curvature labels, our models were able to make class predictions with 100% accuracy. Additionally, when trained with ground truth physical slant labels, the predicted slants did not have a statistically significant mean difference from the ground truth slants ( $p = 0.688$ ), nor was there a



**Fig. 6. Effects of supervised model architecture and training objective.** (a) U-Net-based supervised model latent space visualizations. We trained the model using the sign of curvature labels (top row) and physical slant values (bottom row). Data points are color-coded by the FOV and optical slant values (see legend). The latent space of the U-Net-based supervised model captures FOV and slant regardless of the training objective, and the clustering effect is more substantial with stronger supervision (slant values). (b) For ResNet18, when trained with curvature sign labels, there is a far separation of concave and convex stimuli in the latent space, but FOV and optical slant are entangled; when trained physical slant labels, optical slant is well disentangled and FOV is also better separated. (c) Plots of the latent distance against the FOV for both training objective (rows) and model architectures (columns). U-Net-based supervised model has smoother curves, indicating a more structured latent space. (d) The pseudo perceptual gain as a function of the FOV per irregularity level for each supervised model (columns) and training objective (rows). Under both training settings, the supervised U-Net disentangles texture irregularity level and the ResNet fails to do so.

statistically significant mean difference in predicted slants between concave and convex images ( $p = 0.394$ ).

We examine the supervised latent spaces under different architectures and training settings. Figures 6a and 6b depict the latent space visualizations for the U-Net-based and ResNet-based models, respectively. They were trained using either sign of curvature labels or ground truth physical slant labels. Results indicate that both models, when supervised by the sign of curvature, exhibited a significant separation of concave and convex clusters in their respective latent spaces (U-Net:  $t = -1036.07$ ,  $P < 0.001$ , Cohen's  $d = -3.66$ ; ResNet:  $t = -1461.17$ ,  $P < 0.001$ , Cohen's  $d = -5.17$ ). However, in the ResNet-based model, the FOV and optical slant appeared to be fully entangled (FOV:  $t = -18.07$ ,  $P < 0.001$ , Cohen's  $d = -0.09$ ; Slant:  $t = -4.27$ ,  $P < 0.001$ , Cohen's  $d = -0.02$ ), whereas the U-Net disentangled them (FOV:  $t = -90.96$ ,  $P < 0.001$ , Cohen's  $d = -0.46$ ; Slant:  $t = -58.76$ ,  $P < 0.001$ , Cohen's  $d = -0.29$ ).

On the other hand, model latent spaces were better structured when supervised by ground truth physical slant labels. This enabled both models to accurately cluster the optical slant (U-Net:  $t = -258.48$ ,  $P < 0.001$ , Cohen's  $d = -1.27$ ; ResNet:  $t = -86.29$ ,  $P < 0.001$ , Cohen's  $d = -0.43$ ), a variable directly related to physical slant, and to achieve better results at separating out FOV in the case of ResNet ( $t = -40.77$ ,  $P < 0.001$ , Cohen's  $d = -0.20$ ). The FOV vs. latent distance plots (Fig. 6c) indicate that the latent distance first increases with FOV and then becomes flat in all cases, but the curves are smoother for the U-Net, indicating better latent space disentanglement. When trained on textures with different irregularities, the U-Net can disentangle texture irregularity under both training objectives, and the pseudo perceptual gain is greater at lower texture irregularity levels (Fig. 6d). However, the ResNet fails at identifying texture irregularities in its latent space.

Across model architecture differences, supervised models show stronger clustering with their supervised properties and weaker clustering with other properties, and their latent spaces can better capture trained and untrained factors with stronger supervision.

## 5 DISCUSSION

How the human visual system learns to process and integrate visual depth cues to form a 3D percept remain unresolved in vision research. Notably, the human visual system exhibits various biases in depth perception, resulting in systematic deviations of the human predictions of depth from the ground truth under specific conditions.

Unsupervised DNNs can learn statistical distributions from high-dimensional inputs and compactly store information in latent representations. We have demonstrated that unsupervised DNNs are capable of replicating human biases in a range of tasks related to judging slant from texture. Specifically, unsupervised models made more errors in determining the sign of surface curvature when the FOV is smaller (Sec. 4.2), and perceived more slant when (a) FOV is greater (Sec. 4.3), (b) surfaces are convex rather than concave (Sec. 4.3), (c) the surface texture patterns are more regular (Sec. 4.4). In comparison, no bias was observed when models were supervised by the ground truth. Physical properties that were not trained on were also entangled in some supervised model latent spaces, whereas all factors were well-disentangled in the unsupervised latent spaces.

We found that modifying common design choices in DNN architectures (depth, width, convolution kernel size, activation functions) did not substantially affect the main conclusions of our study, despite impacting image reconstruction quality. One exception was the residual connections between the encoder and the decoder used in the U-Net. We empirically found that adding these additional paths that bypass the latent space bottleneck resulted in more structured latent spaces. Despite preserving other perceptual biases, this modification impaired model ability to disentangle texture irregularity. Further investigation is required to validate and explain this effect.

These results also have implications for the study of 3D human perception. Todd et al. [2007] show that the patterns of biases discussed above can emerge if observers rely solely on image-level changes, such as measuring the scaling difference in the projected texture elements. Unsupervised shape from texture deep learning models can give a signal that is highly related to the shape of the



surface, even though it does not produce the veridical estimate, and with an internal representation that is also highly correlated to geometric scaling measures of texture (Tab. 3). Beyond predictive capability, that neural networks find a good optimum under their training mechanism, and that they are found to represent the texture information similarly when feedback on the veridical structure is impossible, may suggest further study into a potential deeper connection to the human mechanism of slant perception.

Finally, the psychophysical study by Todd et al. [2005] showed the impact of different texture types—plaids, regular and irregular contours and blobs—on human perception of depth. Future work could study the effect of texture types in the context of DNN models.

## REFERENCES

- Carlo Campagnoli, Bethany Hung, and Fulvio Domini. 2022. Explicit and implicit depth-cue integration: evidence of systematic biases with real objects. *Vision Research* 190 (2022), 107961.
- Fulvio Domini and Corrado Caudek. 2003. 3-D structure perceived from dynamic information: A new theory. *Trends in Cognitive Sciences* 7, 10 (2003), 444–449.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- Elizabeth B Johnston. 1991. Systematic distortions of shape from stereopsis. *Vision research* 31, 7-8 (1991), 1351–1360.
- Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
- Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. 2019. Brain-like object recognition with high-performing shallow recurrent ANNs. *Advances in neural information processing systems* 32 (2019).
- Michael S Langer and Ryan A Siciliano. 2015. Are blur and disparity complementary cues to depth? *Vision Research* 107 (2015), 15–21.
- Grace W. Lindsay. 2021. Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience* 33, 10 (sep 2021), 2017–2031. [https://doi.org/10.1162/jocn\\_a\\_01544](https://doi.org/10.1162/jocn_a_01544)
- Baoxia Liu and James T Todd. 2004. Perceptual biases in the interpretation of 3D shape from shading. *Vision research* 44, 18 (2004), 2135–2145.
- John Ashworth Nelder and Robert WM Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* 135, 3 (1972), 370–384.
- Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. 2014. A toolbox for representational similarity analysis. *PLoS computational biology* 10, 4 (2014), e1003553.
- Carlos R Ponce, Will Xiao, Peter F Schade, Till S Hartmann, Gabriel Kreiman, and Margaret S Livingstone. 2019. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell* 177, 4 (2019), 999–1009.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18. Springer, 234–241.
- Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. 2018. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv* (2018), 407007.
- Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- Katherine R Storrs, Barton L Anderson, and Roland W Fleming. 2021. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour* 5, 10 (2021), 1402–1417.
- James T Todd, Lore Thaler, and Tjeerd MH Dijkstra. 2005. The effects of field of view on the perception of 3D slant from texture. *Vision Research* 45, 12 (2005), 1501–1517.
- James T Todd, Lore Thaler, Tjeerd MH Dijkstra, Jan J Koenderink, and Astrid ML Kappers. 2007. The effects of viewing angle, camera angle, and sign of surface curvature on the perception of three-dimensional shape from texture. *Journal of vision* 7, 12 (2007), 9–9.
- Simon J Watt, Kurt Akeley, Marc O Ernst, and Martin S Banks. 2005. Focus cues affect perceived depth. *Journal of vision* 5, 10 (2005), 7–7.
- Yaoda Xu and Maryam Vaziri-Pashkam. 2021. Publisher Correction: Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature communications* 12 (2021).