CVPR
#2762

Do computer vision foundation models learn the low-level characteristics of the human visual system?

Anonymous CVPR submission

Paper ID 2762

Abstract

Computer vision foundation models, such as DINO or OpenCLIP, are trained in a self-supervised manner on large image datasets. Analogously, substantial evidence suggests that the human visual system (HVS) is influenced by the statistical distribution of colors and patterns in the natural world, characteristics also present in the training data of foundation models. The question we address in this paper is whether foundation models trained on natural images mimic some of the low-level characteristics of the human visual system, such as contrast detection, contrast masking, and contrast constancy. Specifically, we designed a protocol comprising nine test types to evaluate the image encoders of 45 foundation and generative models. Our results indicate that some foundation models (e.g., DINO, DINOv2, and OpenCLIP), share some of the characteristics of human vision, but other models show little resemblance. Foundation models tend to show smaller sensitivity to low contrast and rather irregular responses to contrast across frequencies. The foundation models show the best agreement with human data in terms of contrast masking. Our findings suggest that human vision and computer vision may take both similar and different paths when learning to interpret images of the real world. Overall, while differences remain, foundation models trained on vision tasks start to align with low-level human vision, with DINOv2 showing the closest resemblance.

1. Introduction

Computer vision foundation models, such as DINO [8] or OpenCLIP [22, 31], show exceptional ability to generalize to different tasks and are becoming cornerstones of many computer vision methods. They owe their exceptional performance to self-supervised training on very large image datasets. The human visual system also owes much of its capability to being able to perceive the world, over many years from infancy to childhood [6]. A question arises: if the neural network and the visual system are trained by being exposed to a large number of images of the world, will they share their low-level vision characteristics? If they do, we will know that those low-level characteristics arise naturally and likely reflect the statistics of real-world scenes. If they do not, it means that human low-level vision characteristics are specific to the optical/biological limitations of human vision rather than natural image statistics. Our analysis is meant to shed some light on how the vision, either biological or computational, may develop from observing samples of the world, taking either the same or different routes to accomplish their respective tasks.

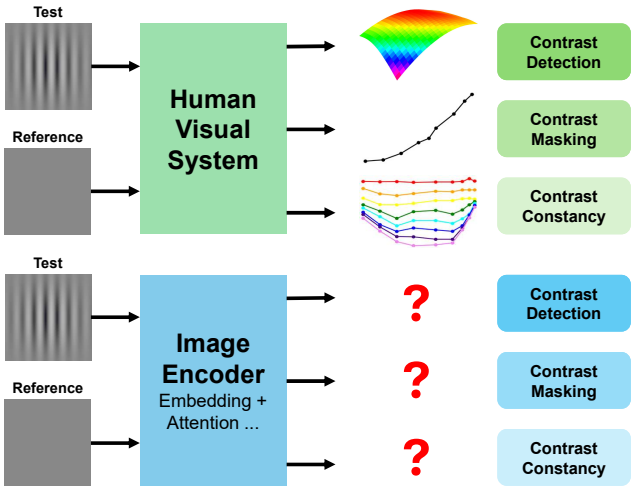

Figure 1. To determine whether image encoders of foundation models exhibit a similar low-level characteristic as human vision, we test them on psychophysical stimuli for which human data is available. We want to test the alignment of contrast encoding between human and computational vision models.

In particular, we are interested in the characteristics that are well understood and measured in human vision science using psychophysical methods: contrast detection [4], contrast masking [25] and contrast constancy [17]. Contrast detection and contrast masking quantify the ability of the visual system to detect small contrast patterns, either on uni-

form backgrounds (contrast detection) or on backgrounds with patterns (contrast masking). Contrast detection and masking capture the "bottlenecks" of the visual system — the characteristic that may prevent us from detecting patterns that are too dark or too small. Similarly, cameras used for computer vision are limited by the MTF of the lens, sensor resolution, photon and sensor noise, and we can expect that computer vision methods may need to deal with similar limitations.

Contrast constancy is the term used in vision science to describe the invariance of the visual system to spatial frequency [17] and partially luminance [24, 30]. Georgeson and Sullivan [17] showed that the perceived magnitude of the contrast that is well above the detection threshold (supra-threshold) appears to us the same regardless of spatial frequency. This is a very important characteristic as it allows us to see contrast (and therefore objects) the same regardless of the viewing distance; otherwise, the frequencies would change with the viewing distance and hence the contrast appearance. A partial constancy (invariance) is also observed across luminance [24, 30], though there is a significant deviation from constancy at lower luminance levels, once the visual system needs to rely on the rod vision. The invariance is also an important feature of many computer vision methods. For example in SIFT features [27] have been designed to be invariant to the changes in contrast, brightness, scale, and rotation. In our experiments, we used the supra-threshold contrast matching test to assess whether the models exhibit the characteristic of contrast constancy.

Numerous works on adversarial attacks demonstrated that the classification performance of deep learning models can be greatly degraded by visually inconsequential changes [19]. At the same time, human vision does not suffer from such adversarial vulnerability [41]. This is one of the most salient arguments put forward to state that deep architectures are different from human vision. Here, we propose a different methodology to study this question. We consider the deep neural network to be a black box and compare its responses to well-understood and measured characteristics of the human visual system. In particular, we want to check whether the foundation vision models share the same "bottlenecks" and invariance properties as the visual system. To achieve this, we test foundation models on basic vision stimuli, such as Gabor patches and band-limited noise, and compare the response of those models with the psychophysical data collected from human observers.

In summary, our contributions are as follows:

- We developed a protocol to evaluate the similarity between machine vision models and the human visual system. This protocol includes contrast detection, contrast masking, and contrast constancy, subdivided into nine distinct test types that collectively capture the low-level fundamental characteristics of human vision.
- We tested the image encoders of 45 foundation and generative models. The results reveal similarities between certain foundation models (*e.g.*, DINOv2 and OpenCLIP) and human vision, particularly in the contrast masking test. However, differences persist across other tests.

2. Related Work

Since the advent of deep learning, machine vision models based on DNNs and foundation models [12, 20, 23, 29] have successfully handled numerous advanced visual tasks. However, researchers have observed that machine vision operates differently from human vision. [15] revealed that standard convolutional neural networks (CNNs) trained on ImageNet are strongly biased toward texture recognition rather than shape, which contrasts with human visual patterns. [41] provided further evidence that deep neural networks (DNNs) differ significantly from the HVS, demonstrating poor robustness in object classification under 3D viewpoint changes and image distortions, and showing vulnerability to adversarial examples, which are rarely problematic for humans. [5] pointed out that DNNs performing well in benchmark tests share little overlap with biological vision mechanisms and fail to account for many findings in psychological studies of human vision. This highlights a clear distinction between machine and human vision, leading to the rise of interest in domain adaptation [7] and making networks robust to adversarial attacks [43].

But there is also evidence that the gap between neural network-based machine vision models and human vision is gradually narrowing. [37] compared vision transformers (ViT) [12] and CNNs, finding that ViT not only achieves superior task accuracy but also exhibits weaker inductive biases, with error patterns more consistent with human errors. [16] discovered that the long-standing robustness gap between humans and CNNs in handling distortions is shrinking. [9, 18] also demonstrated that foundation models like DINO [8] and CLIP [32] can generate more accurate and robust metrics for low-level perceptual similarity.

Most of the aforementioned studies focus on high-level task performance (*e.g.*, accuracy, consistency, ...), which may not reveal whether computation models suffer from the same bottlenecks and rely on the same invariances as human vision. To that end, [2, 26] have attempted to reveal CSF characteristics within pretrained architectures by training a head with a contrast discrimination classifier. The problem with this approach is that it introduces a bias by relying on a classifier trained to compare contrast. Such studies also make an incorrect assumption that CSF explains both near-threshold and super-threshold vision, while contrast constancy results (see Section 4.3) show that this is not the case. In contrast, we examine networks' low-level characteristics without additional task-specific training, considering both near-threshold and supra-threshold vision.

3. Testing framework

We first explain the tested models, testing methods, result visualization, and the strategy we used to summarize and quantify our results.

3.1. Tested models and testing methodology

The objective is to evaluate the responses of machine vision foundation models to stimuli commonly used in human vision research and to compare these responses with psychophysical human data. We tested a representative set of 45 models, encompassing the most influential large vision foundation models, including the variants of DINO [8], DINOv2 [11, 29], OpenCLIP [22, 31], SAM [23], SAM-2 [33], and MAE [21], as well as the encoder used for the latent space of generative model Stable Diffusion (SD-VAE, [34]). Additionally, we report the responses of ColorVideoVDP [28], which is an image and video quality metric that explicitly models low-level human vision and acts as a reference for a low-level human vision model. All models and their variants are listed in Figure 6.

To test these models, we need to be able to compare pairs of images and assess the "perceived" difference between them. For example, for a pattern detection task, a pair of images could be a Gabor patch (test) and a uniform field (reference), as shown on the left of Figure 2. As such patterns are calibrated in physical light units in vision science, we generate these patterns as luminance maps, scaled in physical units of cd/m^2. These luminance maps are then mapped from the linear space to the sRGB color space using a display model (display peak luminance of 400 cd/m^2) and fed into the image encoder of the foundation model for feature extraction. Note that the sRGB space is almost universally used to represent training datasets and is expected input for the tested encoders. To ensure that models can operate on small contrast values, we modified them to accept floating-point values (instead of 8-bit integers) as input. This was necessary as the quantization artifacts in 8-bit images are often larger than the detection thresholds of human vision.

To investigate whether the distances in the feature space reflect the perceptual detection thresholds and invariances, we experimented with a series of distance measures, including L_1 and L_2. We found the cosine similarity expressed as a relative angle (S_{ac}) yielded results most consistent with the psychophysical data. S_{ac} is defined as:

$$S_{\text{ac}} = \frac{1}{\pi} \arccos\left(\frac{F_{\text{T}} \cdot F_{\text{R}}}{\|F_{\text{T}}\| \, \|F_{\text{R}}\|}\right), \tag{1}$$

where F_{T} and F_{R} are the test and reference feature vectors (feature maps reshaped into one dimension), and $\cdot$ denotes the dot product. $S_{\text{ac}} = 0$ indicates two input images are equivalent, while $S_{\text{ac}} = 1$ indicates large differences.

To compare the encoder responses with psychophysical data, we need to be able to map the image sampling frequency into the spatial frequency on the retina. For that, we select the effective resolution of 60 pixels-per-degree, which is typical for modern monitors. We note, however, that the choice of this parameter is arbitrary and the model similarity scores can be shifted by a small multiplier along the spatial frequency axis. The luminance maps are generated at the resolution of 224×224 pixels, corresponding to the size of 3.7×3.7 visual degrees.

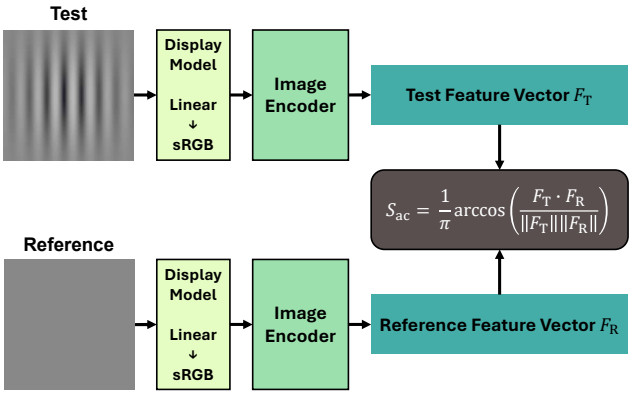

Figure 2. Pipeline for computing S_{ac}. Generate test and reference images in the linear luminance space (cd/m^2), transform them to the sRGB color space using a display model, and input each into the image encoder. Reshape the output features into one-dimensional vectors F_{T} and F_{R}, then compute S_{ac}.

For ColorVideoVDP, we do not use sRGB encoding as the metric can directly work on physical units. We directly use its quality score instead of S_{ac}. Note that the primary focus of this study is on foundation models, with the ColorVideoVDP metric used solely for comparison purposes.

3.2. Model alignment score

Most of our results will be represented as contour plots of model responses, providing qualitative interpretation. Here, we explain our measure of model alignment, which provides quantitative scores.

As an example, we take the leftmost contour plot in row (a) of Figure 5. Each point on the contour plot corresponds to S_{ac} between the test image as shown in Figure 3 and a uniform field of the same mean luminance. The dashed line represents human contrast detection data, predicted with castleCSF [3]. A well-aligned model should show one of the contour lines that follows the dashed castleCSF line. We cannot directly use the S_{ac} values along the dashed line as the measure of alignment because some models result in $S_{\text{ac}} = 0$ for most points near the detection threshold (they detect no difference). Therefore, instead, we rely on the measure of change in S_{ac} in the neighborhood of the dashed line.

For a well-aligned model, the perceived differences in the neighborhood of the detection threshold (dashed line)

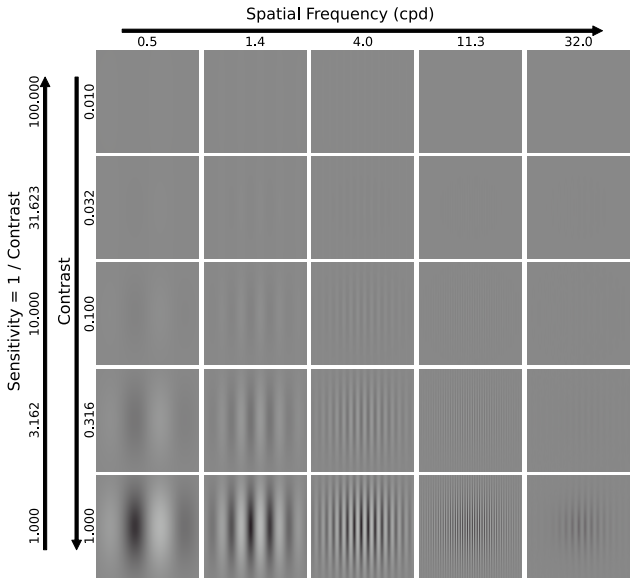

Figure 3. Gabors with different spatial frequencies (x-axis) and contrast (y-axis) used as the test images in the contrast detection tests (Section 4.1). "cpd" denotes cycles per degree. Note that the high-frequency patterns can be rendered with aliasing artifacts on the screen or in print — those were not present in our tests.

should increase as the contrast increases (the sensitivity decreases), Furthermore, the values should be similar along the dashed line. The measure these two properties, we sample the S_{ac} values for the points that are shifted in contrast (vertical direction) from the dashed line by a multiplier m, where $0.5 \leq m \leq 2$ (note the logarithmic scale in the contour plot in Figure 5). We collect such data for multiple frequencies (or other dimensions) along the dashed line and calculate the Spearman rank order correlation between the multipliers m and the S_{ac} values. If the properties mentioned above are preserved, the correlation coefficient value r_s should be close to 1.

The above strategy is used for all contrast detection and contrast masking experiments. For the contrast matching experiment, we use the root mean squared error (RMSE) between the model and human matching data, expressed as the logarithm of contrast.

4. Experiments

4.1. Contrast detection and CSF

We begin by testing the foundation models' ability to detect low-contrast (near-threshold) patterns (*e.g.*, Gabor patches, band-limited noise) and compare their performance with the human data. As the reference human data, we rely on the caslteCSF [3], which is the recent contrast sensitivity function, modeling contrast detection of both achromatic and chromatic patterns.

Spatial Frequency Contrast sensitivity of the human eye is typically associated with the variation across the spatial frequency. The visual system exhibits a band-pass characteristic, with a peak sensitivity between 2 and 4 cycles per degree (cpd), depending on the luminance and other parameters of the stimulus. The lower sensitivity of the visual system at lower frequencies is associated with the mechanism of lateral inhibition [4], which helps to reduce the influence of (low-frequency) illumination on the perceived images. Such invariance to illumination is a desirable property if the goal is to recognize objects regardless of illumination conditions. The drop in sensitivity at high frequencies is associated with the limitations of eye optics (achromatic contrast) and cone density (chromatic contrast). Here we want to test whether the computer-vision foundation models pick up similar traits when trained on natural images.

To test encoder responses across frequencies, we generated a 2D array of image pairs, in which the reference image had uniform luminance, and the test image contained a Gabor patch, as shown in Figure 3 (refer to Supplementary for the visualization of other stimuli). We generate such Gabors for achromatic (see Figure 4-a) and chromatic modulation (see Figure 4-c,d). We also tested band-limited noise (see Figure 4-b). The test patterns had fixed size but varying spatial frequency (x-axis) and contrast (y-axis). Because most of the contrast detection data are plotted as the function of sensitivity, we follow this convention and plot the sensitivity, which is the inverse of the contrast. This corresponds to the reversal of the axis on the logarithmic plots, which we use in our analysis. The other parameters of the stimuli are listed in Table 1. Although we tested multiple variants of each foundation model, here we show only the variant with the highest complexity and best performance on its original high-level tasks. The contour plots for other variants can be found in the Supplementary.

The contour plots of foundation model responses are shown together with a contrast sensitivity function (castleCSF [3]) in rows (a)–(d) of Figure 5. If the foundation models had the same contrast detection characteristics as the human eye, we would expect the smallest difference (the lowest S_{ac}) contour line to follow the black dashed curve of castleCSF. This is not the case for any of the tested foundation models. Some models, and in particular DINOv2 and SD-VAE, show overall band-pass characteristics, in particular for noise and achromatic Gabors. SD-VAE has a drop of sensitivity at lower frequencies much larger than that of the visual system. The responses to chromatic patterns (rows (c) and (d)) tend to be less regular than to achromatic patterns and lack the shift toward lower frequencies, which is observed in the human data. Both OpenCLIP and SAM-2 show a very inconsistent response across the spatial frequencies.

From the data, we can conclude that foundation models

CVPR
#2762

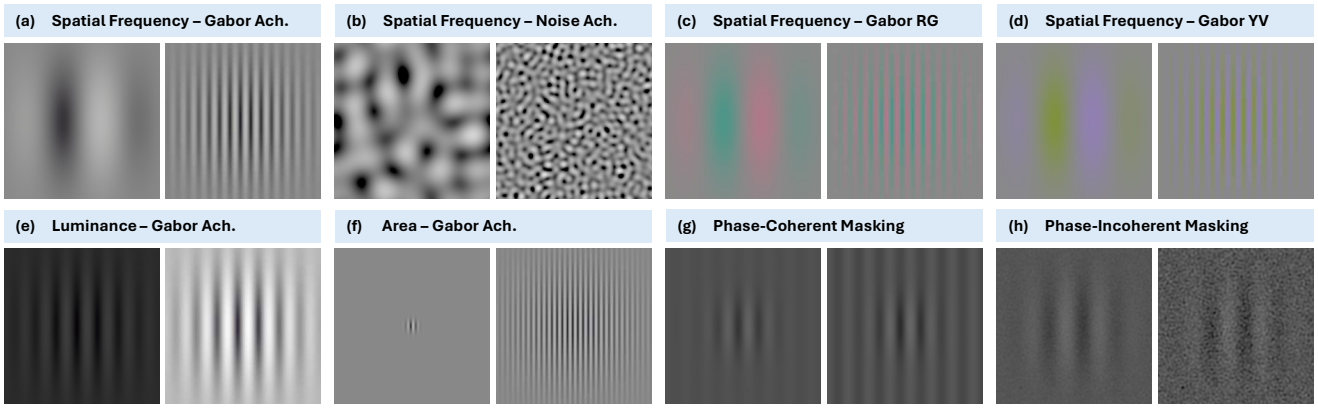

Figure 4. Examples of test images for contrast detection and contrast masking. The detailed explanation of the stimuli can be found in the Supplementary. "Ach." denotes achromatic.

Table 1. Key parameters for all our tests. Note that "Radius" does not apply when the pattern is not a Gabor.

Test	Spatial Frequency (cpd)	Luminance (cd/m^2)	Radius (degree)	Contrast
Spatial Frequency - Gabor Achromatic	0.5 - 32	100	1	0.001 - 1
Spatial Frequency - Noise Achromatic	0.5 - 32	100	-	0.001 - 1
Spatial Frequency - Gabor RG	0.5 - 32	100	1	0.001 - 0.2
Spatial Frequency - Gabor YV	0.5 - 32	100	1	0.001 - 0.2
Luminance - Gabor Achromatic	2	0.1 - 200	1	0.001 - 1
Area - Gabor Achromatic	8	100	0.1 - 1	0.001 - 1
Phase-Coherent Masking	2 (mask) / 2 (test)	32	0.5 (test)	0.005 - 0.5 (mask) / 0.01 - 0.5 (test)
Phase-Incoherent Masking	0 - 12 (mask) / 1.2 (test)	37	0.8 (test)	0.005 - 0.5 (mask) / 0.01 - 0.5 (test)
Contrast Matching	5 (reference) / 0.25 - 25 (test)	10	-	0.005 - 0.629 (reference)

do not follow the sensitivity pattern of the visual system, but some may show a band-pass characteristic that is associated with the CSF. Many models show lower sensitivity at lower frequencies, which may indicate that those models obtained some invariance to (low-frequency) illumination through training.

Luminance The sensitivity of the human eye increases with the luminance. In dim light, human contrast sensitivity increases proportionally to the square root of retinal illuminance, following the DeVries-Rose law. Conversely, in bright light, sensitivity follows Weber's law, remaining independent of illuminance [36]. In this experiment, we want to test whether the computer vision models lose sensitivity at lower luminance levels. Such a loss could be justified by camera noise, which increases in terms of contrast as the light intensity decreases [1].

To produce contour plots, we generated a 2D array of image pairs in a similar manner as for spatial frequency variations in the section above, but instead of varying spatial frequency, we varied luminance, as shown in Figure 4-e. As a reminder, we did not pass the absolute luminance values directly to each model but instead converted them to display-encoded (gamma-corrected) sRGB space — the colour space used for training datasets (see Figure 2). Such an encoding partially compensates for perceptual non-uniformity of luminance.

The results, shown in row (e) of Figure 5, indicate a systematic drop in sensitivity with luminance for all tested models. The drop in sensitivity of those models is faster than that observed in the human data, in particular for OpenCLIP.

We can conclude that the models trained on sufficiently large datasets mimic the sensitivity of the visual system to luminance but with a faster drop in sensitivity. This, however, may be the result of using sRGB representation for training datasets, which were presumably mostly well-exposed and contained relatively little information in darker image regions.

Area Stimulus area (size) significantly affects sensitivity, as larger stimuli activate more retinal cells. Sensitivity increases with area up to the saturation point, denoted as the critical area [35]. The response to stimuli of different sizes tests the model's ability to pool information across the visual field.

CVPR
#2762

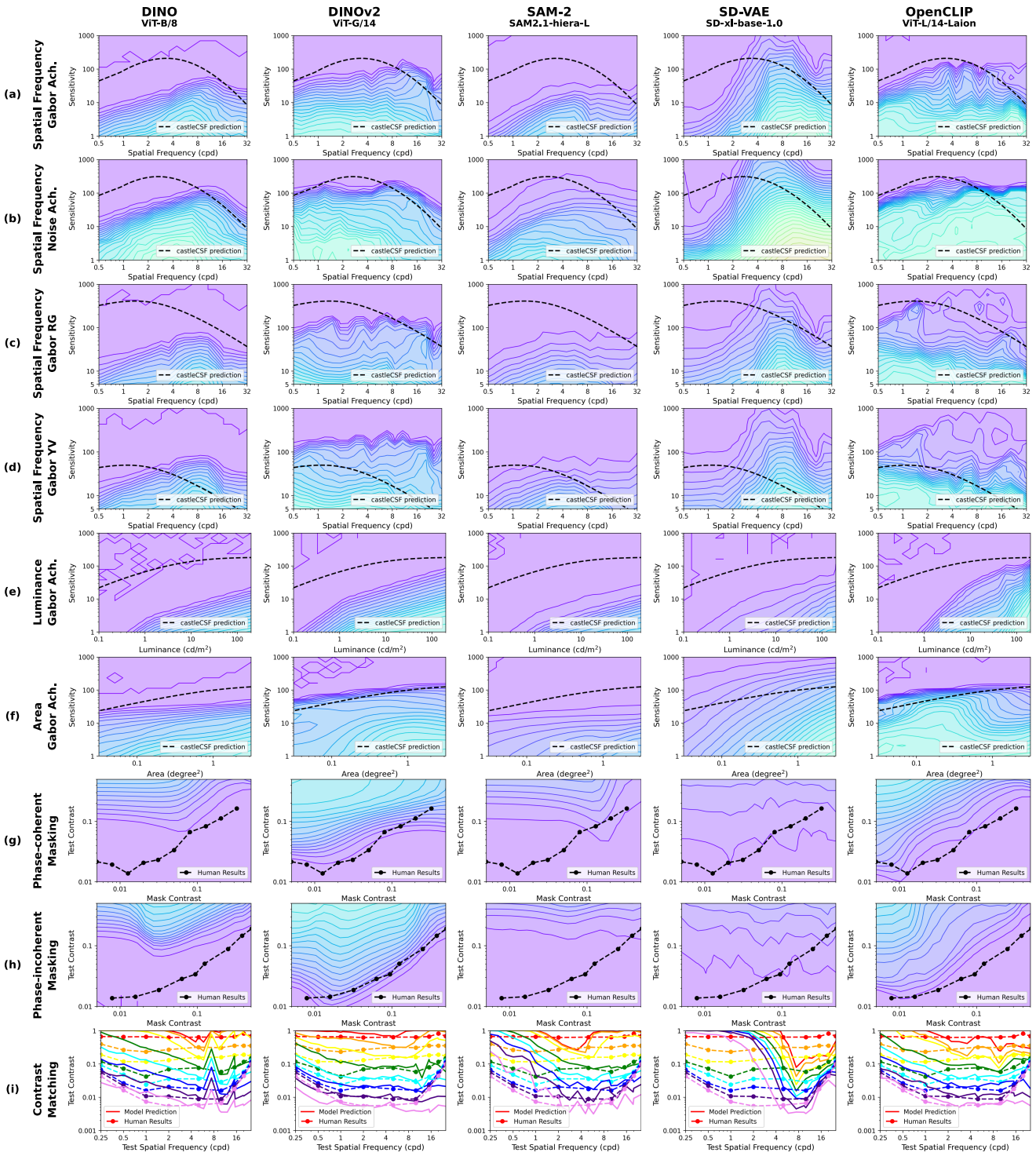

Figure 5. Selected representative experimental results. Each row represents a test, and each column corresponds to a model, selected as the best-performing in their original tasks. (a)-(f): contour plots of contrast detection S_{ac}, with the ground truth castleCSF [3]. (g),(h): contour plots of contrast masking S_{ac}, with the ground truth from [13] and [14], respectively. (i): results from the contrast matching experiment, where different colors represent different C_r values. The dashed lines are human results [17], while the solid lines are model-predicted results from Equation 2. Results for all models are in the supplementary material.

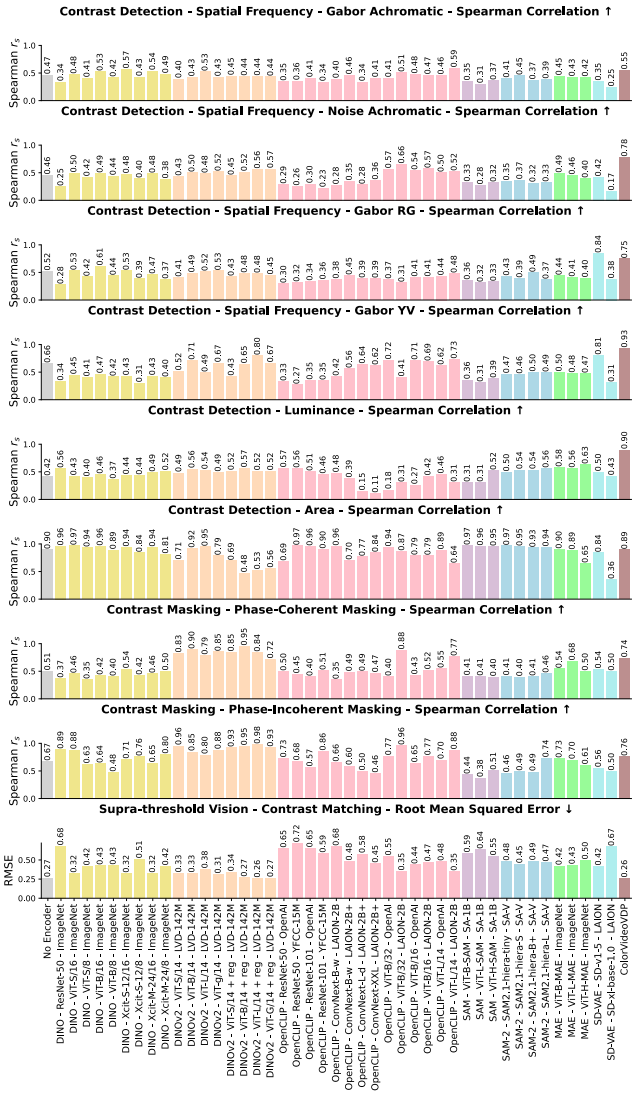

Figure 6. The quantified similarity error between all 45 models and HVS under 9 different tests. For the Contrast Detection and Contrast Masking tasks, Spearman Correlation was used as the metric, with higher values (closer to 1) indicating greater similarity to human vision. For the Supra-threshold Contrast Matching task, RMSE was used as the metric, with lower values (closer to 0) indicating better similarity.

In this experiment, we vary the size of the Gaussian envelope limiting stimulus size to observe its effect on model responses. We follow the same procedure as in the two previous sections and generate pairs with a uniform field and a Gabor patch (see Figure 4-f). The parameters of the Gabor patch are listed in Table 1.

The results, shown in row (f) of Figure 5, indicate that most models show summation across the area (the S_{ac} values increase with the area) and the rate of the increase varies across the models; DINO and SAM-2 have a smaller increase, SD-VAR has a higher increase of S_{ac} than the human data, and only DINOv2 roughly matches human performance. OpenCLIP shows no consistent patterns. We can conclude that many models show spatial pooling characteristic that shares the trend observed in human data, though the actual slope of the increase is typically different.

4.2. Contrast masking

Contrast masking explains the decreased visibility of a signal (test) due to the presence of a supra-threshold background (mask). The masking function defines the relationship between the threshold test contrast required for signal detection and the mask contrast. Put simply, a pattern is more difficult to detect in the presence of another pattern of similar spatial frequency and orientation. A typical masking characteristic of the visual system is shown in rows (g) and (h) of Figure 5. It consists of a relatively shallow segment at low mask contrast (near the detection threshold), with the slope increasing for high mask contrast. The shape of the curve is influenced by the specific properties of the mask and test signals [10, 40]. We will consider the case in which the masker is a sinusoidal grating of the same frequency as the test pattern (phase-coherent masking, row (g)) and when the masker is noise (phase-incoherent masking, row (h)) [10, 13, 14], as shown in Figure 4-g,h.

First, we consider the fundamental form of *phase-coherent masking* in which the masker image (reference in Figure 2) is a sinusoidal grating and the test image is the masker plus a Gabor patch of the same frequency and phase as the masker (see Figure 4g). Contrast masking data, shown as the dashed black curve in row (g) of Figure 5, shows the smallest contrast of the test Gabor that is detectable in the presence of the masker of a given contrast. As the contrast of the masker increases, the smallest detectable contrast of the test also needs to increase. However, such an increase starts only for a masker that has the contrast sufficiently high to be detected. If the contrast of the masker is near the detection threshold, we can observe a dipper effect — the contrast detection is facilitated by a masker [13, 39, 42]. Such an effect can be only observed in phase coherent masking.

As an example of *phase-incoherent masking*, we will consider a masker with band-limited noise and a test with a Gabor patch — see Figure 4h. The human detection thresholds for such masking patterns are similar to those for phase-coherent masking, except that the dipper effect disappears [38].

The differences predicted by DINOv2 and OpenCLIP are surprisingly well-aligned with the human contrast masking data — their responses roughly match the slopes of the human data. The alignment is stronger for the phase-incoherent masking. This is particularly notable for OpenCLIP, which did not show any consistent trends for contrast

detection. Other models do not show strong alignment with the human data.

Overall, computational models are better aligned with human data for contrast masking than for contrast discrimination. One possible explanation is that the signals that induce contrast masking are plentiful in natural images, but contrast detection stimuli, which involve barely noticeable patterns on uniform backgrounds, are rare. Therefore, computational models are more likely to pick up the characteristic that is well represented in the training datasets.

4.3. Supra-threshold contrast matching

While contrast detection and contrast masking explain the just detectable (near-threshold) contrast, most of the vision tasks, such as detection or recognition, involve well-visible (supra-threshold) contrast. Supra-threshold human vision has been studied in contrast-matching experiments in which the magnitude of one contrast is visually matched to the magnitude of another contrast of a different frequency [17] or luminance [24, 30]. One of the most significant findings of those studies is *contrast constancy* [17] — the ability of the visual system to match physical contrast across frequencies and luminance levels. The results of the seminal study of Georgeson and Sullivan [17] on matching contrast across frequencies are shown as dashed lines in row (i) of Figure 5. At small contrast, the dashed lines show a band-pass shape that follows the contrast sensitivity function. However, as the contrast is increased, the lines become flat showing little influence of frequency on contrast perception. This is an important property that lets us see objects to have the same appearance regardless of the viewing distance. Such a scale invariance is also important for neural networks that are tasked to detect or recognize objects regardless of their size.

We followed the experimental setup from [17], where the reference was a 5 cpd, 10 cd/m^2 sinusoidal grating, presented at eight distinct contrast levels c_r. The test stimulus had the same luminance but a different spatial frequency ρ_t. In [17], observers adjusted the test stimulus contrast c_t until its apparent contrast matched that of the reference (contrast matching). In our experiments, we match contrast encodings of sinusoidal gratings: the ($S_\mathrm{ac}()$, eq. (1)), between a feature vector of a sinusoidal grating, $F(\rho, c)$ of frequency ρ and contrast c, and a uniform field, $U = F(\rho_\mathrm{t}, 0)$. We find the test contrast c_t that minimizes the expression:

$$\underset{c_\mathrm{t}}{\operatorname{argmin}} \left(\frac{S_\mathrm{ac}(F(\rho_\mathrm{r}, c_\mathrm{r}), U)}{S_\mathrm{ac}(F(\rho_\mathrm{r}, 1), U)} - \frac{S_\mathrm{ac}(F(\rho_\mathrm{t}, c_\mathrm{t}), U)}{S_\mathrm{ac}(F(\rho_\mathrm{t}, 1), U)} \right)^2, \quad (2)$$

where the reference frequency $\rho_\mathrm{r} = 5$ cpd. The denominators in the expression are used to normalize contrast across frequencies. We experimented with other contrast encodings, including a direct comparison of feature vectors, but the formula above resulted in the best contrast constancy properties across the models.

The matching contrast predictions for foundation models are visualized as continuous lines in row (i) of Figure 5. The plots show that only DINOv2 and OpenCLIP roughly follow the dashed contrast constancy lines. Both models show less attenuation (more constancy) at the highest spatial frequencies, which could be advantageous when the model needs to work with small-scale features. Both models show attenuation of low frequencies (below 1 cpd), suggesting worse contrast constancy in that frequency range. Other models, including DINO and SAM-2, suffer from large instability across frequencies, or very heavy attenuation of low frequencies in the case of SD-VAE. To conclude, we can observe only partial contrast constancy for selected models.

4.4. Model alignment scores

As the analysis of all 45 variants of the models is infeasible in the scope of this paper, we prepared quantitative results according to the method explained in Section 3.2 and summarized them in Figure 6. Those let us make three observations:

First, ColorVideoVDP, which is a visual metric that models human low-level vision, is better aligned with the human data in almost all contrast detection tasks and in terms of contrast constancy, as expected. However, certain variants of OpenCLIP and DINOv2 can match or surpass the ColorVideoVDP alignment in terms of contrast masking. Second, the alignment scores of different foundation model variants (*e.g.*, OpenCLIP) show significant variations and alignment scores appear unrelated to the complexity of the variants or their performance on higher-level tasks. Finally, DINOv2 variants, which have been trained to solve vision tasks, show the greatest alignment with the human data among all foundation models.

5. Conclusions

If we believe that the goal of both biological and computational low-level vision is to efficiently encode visual information, we can expect that computational models trained on large natural image datasets will share similarities with human vision. In this work, we find that selected computational models, e.g., variants of DINOv2 and OpenCLIP, show surprisingly high alignment with supra-threshold human contrast masking and contrast matching data, but little alignment with the near-threshold contrast detection. This means that computation models do not have the same "bottlenecks" as human vision, but, through training, they attain invariance and efficient contrast coding that resembles that of the visual system. We hope that our testing protocol with basic psychophysical stimuli will provide a useful tool for examining future computational models of vision.

References

[1] Cecilia Aguerrebere, Julie Delon, Yann Gousseau, and Pablo Musé. *Study of the digital camera acquisition process and statistical modeling of the sensor raw data*. 2013. Citation Key: Aguerrebere2013. 5

[2] Arash Akbarinia, Yaniv Morgenstern, and Karl R Gegenfurtner. Contrast sensitivity function in deep networks. *Neural Networks*, 164:228–244, 2023. 2

[3] Maliha Ashraf, Rafał K Mantiuk, Alexandre Chapiro, and Sophie Wuerger. castlecsf—a contrast sensitivity function of color, area, spatiotemporal frequency, luminance and eccentricity. *Journal of Vision*, 24(4):5–5, 2024. 3, 4, 6

[4] Peter GJ Barten. *Contrast sensitivity of the human eye and its effects on image quality*. SPIE press, 1999. 1, 4

[5] Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E Hummel, Rachel F Heaton, et al. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46: e385, 2023. 2

[6] Oliver Braddick and Janette Atkinson. Development of human visual function. *Vision Research*, 51(13):1588–1609, 2011. 1

[7] Yancheng Cai, Bo Zhang, Baopu Li, Tao Chen, Hongliang Yan, Jingdong Zhang, and Jiahao Xu. Rethinking cross-domain pedestrian detection: a background-focused distribution alignment framework for instance-free one-stage detectors. *IEEE transactions on image processing*, 32:4935–4950, 2023. 2

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1, 2, 3

[9] Francesco Croce, Christian Schlarmann, Naman Deep Singh, and Matthias Hein. Adversarially robust clip models induce better (robust) perceptual metrics. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. 2

[10] Scott J Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III*, pages 2–15. SPIE, 1992. 7

[11] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023. 3

[12] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[13] John M Foley. Human luminance pattern-vision mechanisms: masking experiments require a new model. *JOSA A*, 11(6):1710–1719, 1994. 6, 7

[14] Karl R Gegenfurtner and Daniel C Kiper. Contrast detection in luminance and chromatic noise. *JOSA A*, 9(11):1880–1888, 1992. 6, 7

[15] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2

[16] Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021. 2

[17] MA Georgeson and GD Sullivan. Contrast constancy: deblurring in human vision by spatial frequency channels. *The Journal of physiology*, 252(3):627–656, 1975. 1, 2, 6, 8

[18] Abhijay Ghildyal, Nabajeet Barman, and Saman Zadtootaghaj. Foundation models boost low-level perceptual similarity metrics. *arXiv preprint arXiv:2409.07650*, 2024. 2

[19] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2014. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[22] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 1, 3

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3

[24] J.J. Kulikowski. Effective contrast constancy and linearity of contrast sensation. *Vision Research*, 16(12):1419–1431, 1976. Citation Key: Kulikowski1976a. 2, 8

[25] G E Legge and J M Foley. Contrast masking in human vision. *Journal of the Optical Society of America*, 70(12):1458–71, 1980. Citation Key: Legge1980. 1

[26] Qiang Li, Alex Gomez-Villa, Marcelo Bertalmío, and Jesús Malo. Contrast sensitivity functions in autoencoders. *Journal of Vision*, 22(6):8–8, 2022. 2

[27] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 2

[28] Rafal K Mantiuk, Param Hanji, Maliha Ashraf, Yuta Asano, and Alexandre Chapiro. Colorvideovdp: A visual difference predictor for image, video and display distortions. *arXiv preprint arXiv:2401.11485*, 2024. 3

[29] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-

Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 3

[30] Eli Peli. Suprathreshold contrast perception across differences in mean luminance: effects of stimulus size, dichoptic presentation, and length of adaptation. *Journal of the Optical Society of America A*, 12(5):817, 1995. Citation Key: Peli1995. 2, 8

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 3

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[33] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[35] Jyrki Rovamo, Olavi Luntinen, and Risto Näsänen. Modelling the dependence of contrast sensitivity on grating area and spatial frequency. *Vision Research*, 33(18):2773–2788, 1993. 5

[36] Jyrki Rovamo, Juvi Mustonen, and Risto Näsänen. Neural modulation transfer function of the human visual system at various eccentricities. *Vision Research*, 35(6):767–774, 1995. 5

[37] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021. 2

[38] A Van Meeteren and JM Valeton. Effects of pictorial noise interfering with visual detection. *JOSA A*, 5(3):438–444, 1988. 7

[39] Andrew B Watson. Efficiency of a model human image code. *JOSA A*, 4(12):2401–2417, 1987. 7

[40] Andrew B Watson and Joshua A Solomon. Model of visual contrast gain control and pattern masking. *JOSA A*, 14(9): 2379–2391, 1997. 7

[41] Felix A Wichmann and Robert Geirhos. Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, 9(1):501–524, 2023. 2

[42] Hugh R Wilson and James R Bergen. A four mechanism model for threshold spatial vision. *Vision research*, 19(1): 19–32, 1979. 7

[43] Fei Yin, Yong Zhang, Baoyuan Wu, Yan Feng, Jingyi Zhang, Yanbo Fan, and Yujiu Yang. Generalizable black-box adversarial attack with meta learning. *IEEE transactions on pattern analysis and machine intelligence*, 46(3):1804–1818, 2023. 2