Geometric Structures of Visual Ambiguity:

How CLIP Represents Optical Illusions in Embedding Space

Seema Rida, Osheen Tikku, Keshav Tiwari

University of California San Diego

COGS 118B: Unsupervised Machine Learning

Professor Virginia De Sa

Introduction

Sensation and perception are distinct yet interconnected processes in the brain. Sensation involves gathering information from the world through sensory receptors, while perception is the way in which the brain interprets those sensory signals (Goldstein 2014). Perception varies across individuals and can be affected by several factors like our expectations, context, and prior experience. Optical illusions provide a compelling perspective of this distinction: although sensory input remains constant, perceptual interpretation can actually differ (Freedheim & Weiner, n.d.). Classic examples like the Checker Shadow illusion (Adelson, 2000) or the Müller-Lyer illusion (Howe & Purves, 2005) show how the visual system relies on heuristics that don't always directly correspond to the physical stimulus.

Foundation models such as CLIP integrate visual and linguistic information and have proven to replicate human-like response patterns under certain tasks (Bommasani et al., 2021). Ngo et al. (2023) demonstrated that CLIP can actually be "fooled" by several classic optical illusions. However, existing research has focused on low-level geometric and lightness illusions, and representations of ambiguous figures in CLIP remain less explored.

To address this limitation, we focus on the rabbit-duck illusion that first appeared in the German magazine "Fliegende Blatter" in 1892, and later analyzed by psychologist Joseph Jastrow in 1899. Ambiguity in figures like the rabbit-duck illusion provide insights on how perception alternates while sensory input remains constant. The illusion offers a clean binary perceptual switch which makes it an ideal test case for investigating our central question: How are optical illusions represented in CLIP's embedding space, and can CLIP's interpretations be systematically biased using mechanisms analogous to those that influence human perception? We hypothetize that CLIP embeddings of the illusions images will fall somewhere in the middle of the embeddings of unambiguous images of rabbits and ducks. Our goal is to evaluate whether CLIP is able to capture the structure of perceptual ambiguity.

Related Work

Human Perception in Foundation Models

We've seen recent advances in machine learning showing how large-scale vision models can superficially mimic human perceptual and cognitive processes. Muttenthaler et al. (2023) revealed a fundamental misalignment that explains that vision foundation models successfully encode local semantic similarities, but fail to capture multilevel semantic structure that exists within human visual information. For example, our brains naturally organize ideas hierarchically. Standard training objectives seem to fail at presenting different relationships and rather focus on differentiating between similar items without the context of that broader conceptual hierarchy that humans experience.

Optical Illusions in Deep Learning

Ngo et al. (2023) systematically test if CLIP can be fooled by optical illusions. They tested 11 different illusions with CLIP using the illusion image as well as text prompts. Their work addresses the low-level perceptual effects like shape and light perception. Sun & Dekel (2021) take a different approach with ImageNet-trained models looking at the Scintillating grid illusion. They discover that these models show nonmonotonic responses, which mirror human perception of the illusion. Both studies focus on illusions that involve misperception of physical properties rather than perceptual ambiguity.

Limited work examines how vision-language models represent images that can be interpreted in multiple equally valid ways. Previous work on optical illusions in deep learning has focused on misperception of physical properties (e.g. Muller-Lyer illusion where two lines are actually the same length) and illusory perception of nonexistent features (e.g. Kanizsa triangle where you perceive a nonexistent triangle). Ambiguous figures are categorized differently; when you look at the rabbit-duck illusion, your brain isn't making a mistake or seeing something that isn't there, it's simply choosing between two interpretations that are legitimately present in the image. Sometimes you see a rabbit, sometimes a duck, and sometimes can spontaneously switch between the two i.e. bistable perception. Since both interpretations are correct, a ground truth does not exist. Instead, the ambiguity itself is the phenomenon we become interested in exploring.

Methods

Dataset

We first compiled a dataset of grayscale rabbit and duck images so we could establish the endpoints of CLIP's embedding space. Using an existing dataset that contained 60 images of rabbits and ducks, we manually picked out photos that had mainly an individual rabbit/duck as its subject. More specifically, we were looking for images that had clear rabbit ears and duck beaks since the illusion we are testing is based on which feature the human brain perceives the left side of the image as. We made sure to remove any images that weren't clear or contained external artifacts such as humans and/or obstacles. From this we ended up with 38 images per animal category (76 total).

Model and Classification

We used the CLIP model implementation provided by the course (Srinivas-R, 2025). Images were preprocessed to 224x224 pixels via CLIP's standard feature extractor. To test the model's classifications, we computed the cosine similarity where the model classified the image based on which category ("duck" or "rabbit") had a greater similarity number. We tested the initial CLIP model with some sanity checks on image-text and image-image similarity. Once we verified this, we moved on to testing our base cases which included 5 total cases: (1) a real rabbit image, (2) a real duck image, (3) a drawing of a rabbit, (4) a drawing of a duck, and (5) the base rabbit- duck illusion with no modifications.

Perceptual Bias Manipulations

To test for perceptual bias, we want to check if CLIP's interpretation can be shifted through different human-like biasing mechanisms. We came up with three conditions to replicate these biasing methods: (1) spatial realignment, (2) feature cropping, (3) image rotation, and (4) re-coloring. For spatial realignment, we cropped the illusion to proportions of ¾ and ½ on its left and right side as well ½ on the top and bottom. For feature cropping, we cropped the original illusion image to focus on the eye, ears/beak, and mouth. To replicate a human tilting their head when viewing an image, we rotated the image by ± 30, 45, 60, 90, and 180 degrees. In order to test whether color perturbations make a difference to CLIP's interpretation of the duck-rabbit illusion, we systematically applied three different colors (red, blue, green) to seven distinct spatial regions of the base illusion image. Additionally, we loosely colored two images with realistic colors (green and yellow for duck, brown for rabbit) to check if semantically appropriate colors would bias classification more than the arbitrary RGB colorizations in our experiment.

Under each of these conditions, we also collected self reports from each member of the group to see what they viewed when looking at the modified illusion. This helped give us a baseline for how human cognition viewed the illusion so we could compare it to the CLIP classifications.

Results

Sanity Check

Before we tested the CLIP classification model against the test images, we wanted to make sure that its category embeddings had validity. When comparing an image of a duck to the duck text, its cosine similarity was 0.2639. When an image of a rabbit was compared to the rabbit text, its cosine similarity was higher with a score of 0.3098. Confused by the higher rabbit image-text embedding association, we experimented with different photos of ducks and rabbits and found that the cosine similarity score varied based on which image was used for the sanity check. That being said, the scores seemed to range between 0.25 and 0.30 and seemed stable enough to continue with the rest of our study. We also ran a validity check for image to image comparison. As seen in *Figure 1,* the highest cosine similarity score was between two photos of the same animal (Rabbit1 vs Rabbit2 and Duck1 vs Duck2) while the cosine similarity score between a duck and rabbit image were lower. This was the desired result of this sanity check since photos of the same animal should be closer in the embedding space than photos of different animals.

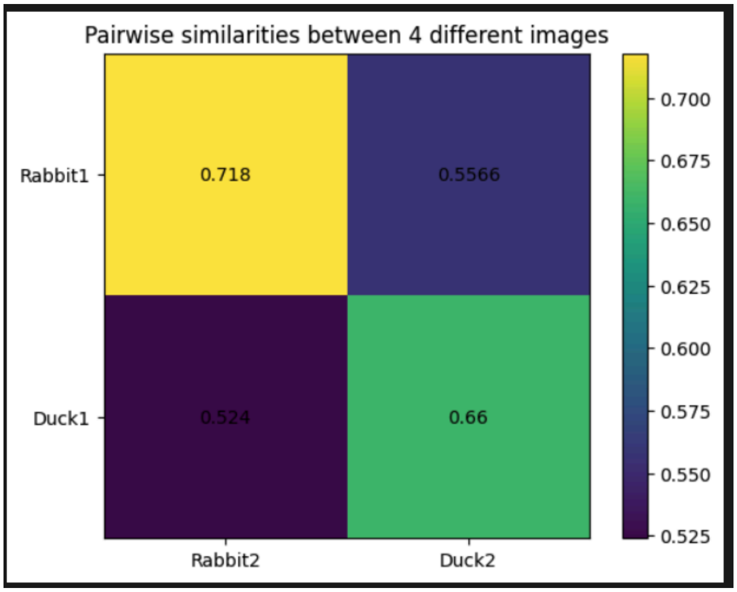

Figure 1. Pairwise similarities between two rabbit and duck images

Bases Cases

Under our base test cases, the CLIP model performed as expected. As seen in *Figure 2*, the real and drawn rabbits were classified as rabbits with a high cosine similarity score (~0.30). Similarly, the real and drawn ducks were classified as ducks with high cosine similarity scores as well (~0.29). This confirms that within the CLIP embedding space, the duck and rabbit clusters were well defined and identified correctly when presented with obvious instances of each category. When presented with the rabbit-duck illusion, we expected to see it fall somewhere in the middle of the embeddings of canonical examples from each category (e.g. between the “duck” cluster and “rabbit” cluster). This was represented in our initial testing as the illusion had a cosine similarity of 0.2822 to the rabbit cluster and 0.2607 to the duck cluster. While these similarities were almost identical, we still see the model with a slight preference for the rabbit cluster.

Test Case	Rabbit Similarity	Duck Similarity	Classification
Real Rabbit	0.3035	0.1964	Rabbit
Drawing Rabbit	0.3093	0.1628	Rabbit
Real Duck	0.1528	0.2927	Duck
Drawing Duck	0.1283	0.2920	Duck
Illusion	0.2822	0.2607	Rabbit

Figure 2. Cosine similarities & resulting classification for CLIP embeddings of base test images

Spatial Realignment

Our first test condition was presenting CLIP with versions of the illusion with its center realigned. As seen in *Figure 3*, when observing the right section of the illusion, CLIP classified the image as rabbit with the rabbit cosine similarities (0.2849 for the right half; 0.2976 for the right three-quarters) exceeding the duck cosine similarities (0.2155 and 0.2532 respectively). Similarly, when presented with the left half and three-quarters, the illusion was classified as duck with rabbit cosine similarities of 0.1635 and 0.2218 as opposed to the higher duck cosine similarities of 0.1835 and 0.2769. When realigned vertically, there was once again a preference for duck with rabbit cosine similarities of 0.1741 and 0.1093 and duck cosine similarities of 0.2109 and 0.1541 for the upper and lower half respectively. For every case, except for the lower 1/2, we found that our self report matched with the classification by CLIP. We would like to note, however, that the lower half realignment resulted in no identifying features (e.g. eyes, mouth, ears/beak) being represented so the classification may not carry as much meaning as opposed to the other realignments.

Test Case	Rabbit Similarity	Duck Similarity	Classification	Self Report
Right 1/2	0.2849	0.2155	Rabbit	Rabbit
Left 1/2	0.1635	0.1835	Duck	Duck
Right 3/4	0.2976	0.2532	Rabbit	Rabbit
Left 3/4	0.2218	0.2769	Duck	Duck
Upper 1/2	0.1741	0.2109	Duck	Duck
Lower 1/2	0.1093	0.1541	Duck	Inconclusive

Figure 3. Cosine similarities & resulting classification for CLIP embeddings of realigned illusion images

Feature Cropping

To test whether focusing on specific anatomical features would bias CLIP, we cropped the rabbit-duck illusion so that we capture and isolate three features: the eye, ear/beak region, and the mouth. CLIP classified all three images as duck, with corresponding cosine similarities represented in *Figure 4.* The ear/beak image, which is the main main feature of the illusion, showed the closest competition between the two categories (rabbit: 0.1591, duck: 0.1726). The eye and mouth images show stronger preference for duck (eye: 0.1887 duck vs. 0.1586 rabbit; mouth: 0.1927 duck vs. 0.1617 rabbit), but our self report found those images inconclusive.

Test Case	Rabbit Similarity	Duck Similarity	Classification	Self Report
Eye	0.1586	0.1887	Duck	Inconclusive
Ear/Beak	0.1591	0.1726	Duck	Duck (Beak)
Mouth	0.1617	0.1927	Duck	Inconclusive

Figure 4. Cosine similarities & resulting classification for CLIP embeddings of cropped illusion features

Image Rotation

We rotated the image by ± 30, 45, 60, 90, and 180 degrees. These rotated test cases produced consistent and significant bias in CLIP's classification of the illusion. As shown in *Figure 5*, CLIP classified all rotated versions as rabbit, and negative angles (counterclockwise rotation) in particular produced the strongest rabbit similarity by 12.7% to 20.1% compared to baseline. Positive angles (clockwise rotation) and the 180° case produced weaker rabbit classification with values ranging from 0.2737 to 0.2944 . Our self report matched the counterclockwise rotation test cases classifications but then diverged from CLIP's classification for clockwise rotation test cases.

Test Case	Rabbit Similarity	Duck Similarity	Classification	Self Report
-30°	0.3352	0.2085	Rabbit	Rabbit
-45°	0.3389	0.2084	Rabbit	Rabbit
-60°	0.3320	0.1981	Rabbit	Rabbit
-90°	0.3181	0.1788	Rabbit	Rabbit
30°	0.2886	0.2832	Rabbit	Duck
45°	0.2944	0.2849	Rabbit	Duck
60°	0.2800	0.2698	Rabbit	Duck
90°	0.2847	0.2433	Rabbit	Duck
180°	0.2737	0.2510	Rabbit	Duck

Figure 5. Cosine similarities and resulting classification for CLIP embeddings of rotated illusion images

Illusion Colorization

We originally hypothesized that adding color to distinct regions of the image could bias CLIP's attention to particular features, potentially shifting its classification toward either duck or rabbit. *Figure 6* shows the similarity scores received for each color-region combination along with scores for the realistic colorizations for duck and rabbit. Each color group largely produced mixed results. For the red group, the left half produced the strongest rabbit bias (+0.0465), whereas the left 3/4th showed the strongest duck bias (-0.0235). As far as the full tint is concerned, all three colors showed almost equal similarities, suggesting an indifference towards full colorizations from CLIP. This finding is supported by the averages across regions for each color suggesting no difference in CLIP classification either.

Color	Region	Rabbit Similarity	Duck Similarity	Classification
Real Rabbit	**Full**	**0.2625**	**0.2496**	**Rabbit**
Real Duck	**Full**	**0.2078**	**0.2686**	**Duck**
Red	Full	0.2532	0.2539	Duck
	Top 1/2	0.2544	0.2446	Rabbit
	Bottom 1/2	0.2571	0.2344	Rabbit
	Left 1/2	0.2751	0.2286	Rabbit
	Right 1/2	0.2329	0.2441	Duck
	Left 3/4	0.2307	0.2542	Duck
	Right 3/4	0.2301	0.2517	Duck
	Average	**0.2476**	**0.2445**	**Rabbit**
Green	Full	0.2578	0.2634	Duck
	Top 1/2	0.2365	0.2549	Duck
	Bottom 1/2	0.2588	0.2463	Rabbit
	Left 1/2	0.283	0.2262	Rabbit
	Right 1/2	0.2284	0.2437	Duck
	Left 3/4	0.2325	0.2612	Duck
	Right 3/4	0.2272	0.2651	Duck

	Average	**0.2463**	**0.2515**	**Duck**
Blue	Full	0.2588	0.2705	Duck
	Top 1/2	0.2413	0.2668	Duck
	Bottom 1/2	0.248	0.2367	Rabbit
	Left 1/2	0.2698	0.2416	Rabbit
	Right 1/2	0.2439	0.2438	Rabbit
	Left 3/4	0.224	0.2681	Duck
	Right 3/4	0.2277	0.2681	Duck
	Average	**0.24479**	**0.25651**	**Duck**

Figure 6. Cosine similarities & resulting classification for CLIP embeddings of recolored illusion images across regions

	Red	**Green**	**Blue**	**Real Rabbit**	**Real Duck**
Full	-0.0007	-0.0056	-0.0117	0.0129	**-0.0608**
Top Half	0.0098	-0.0184	-0.0255		
Bottom Half	0.0227	0.0125	0.0113		
Left Half	0.0465	**0.0568**	0.0282		
Right Half	-0.0112	-0.0153	0.0001		
Left 3/4	-0.0235	-0.0287	-0.0441		
Right 3/4	-0.0216	-0.0379	-0.0404		

Figure 7: Heatmap for CLIP score differences (rabbit sim - duck sim) across region x color. The orange coloring indicates rabbit-biased while the blue coloring indicates duck-biased.

To visualize bias across region and color, *Figure 7* shows the difference in similarity scores between rabbit and duck classifications in a heatmap. The figure reveals that coloring the left half of the image drew a strong bias towards rabbit (+0.0568), which would imply that the colorization acted as a blocker to the duck part of the image, to bias CLIP towards the right hand side. However, coloring either of the 3/4ths, left or right, revealed a strong bias towards duck, which quickly disproves that theory. The largest bias of all was seen in the real duck colorization where CLIP was significantly biased towards duck (-0.0608), suggesting some level of semantic correspondence to color for classification. While the real rabbit colorization only produced a modest rabbit bias (+0.0129), this still suggests that ecologically valid color schemes engage CLIP's semantic understanding more effectively than arbitrary color overlays. Crucially, the RGB colors themselves did not create a significant bias, albeit blue colorizations had slightly higher average differences. This suggests that spatial location matters more than color choice. Most significantly, even the largest color-induced shifts represented only a minor change in similarity scores from the baseline, whereas rotation produced shifts of a much higher magnitude. This further suggests that CLIP's visual processing prioritizes geometric and spatial factors over the color information in classification tasks.

Discussion

Our results show that CLIP interprets the rabbit-duck illusion in partial alignment with human perception, when systematically biased with our 4 test cases. In the spatial realignment condition, our results demonstrate that CLIP is affected by spatial realignment in the same way that human perception is. Changing the center affects how the model views the illusion. When the amount that a certain feature is in focus (such as the ear/beak), it influences the overall illusion's similarity to the rabbit and duck categories that the model was trained on. This tells us that CLIP has the ability to perform feature recognition in instances where it's given the appropriate spatial context. The feature cropping condition results suggest a great limitation. Isolated features presented without their surrounding context, cause CLIP to classify all images as duck with low similarity scores. This reveals a feature in CLIP, demonstrating that CLIP relies on global context more than local features for classification. The image rotation condition revealed fundamental divergences between CLIP and natural human perception. This condition was consistent in classifying the images as rabbit across all angles tested. This shows a limitation in CLIP's ability to semantically process images when orientation is manipulated.

Our findings support prior related work on optical illusions and AI. Ngo et al. (2023) showed that CLIP can be fooled by low-level perceptual illusions that involve lightness and geometric perturbations. Our results show that CLIP is able to encode the perceptual ambiguity of optical illusions in geometrically structured ways. This is proven by the illusion's intermediate embedding position (rabbit sim: 0.2822, duck sim: 0.2607) that mirrors the structural ambiguity that fools human perception. Furthermore, our work supports Muttenthaler et al. (2023), which suggests that vision foundation models fail to capture the multilevel semantic structure which exists in human visual information. When rotated, our human self-report was able to recognize angles at which the shape looked more like a duck/rabbit, however, CLIP rigidly classified all rotations as rabbits. This suggests that while CLIP can encode local semantic similarities (making distinctions between unrotated ducks and rabbits), it can not replicate the context-dependent reinterpretation of the illusion, seen in multi-level semantic structures in human perception.

The illusion colorization approach, while failing in causing significant perturbations in similarity scores, provides critical insight into CLIP's visual processing. The discussion revolves around the crucial insight that color manipulations produced minimal effects, relative to geometric manipulations. Our first explanation of the failure of color-based variance is that CLIP was trained on 400 million diverse image-text pairs, consisting of deep variation in lighting, color grading etc. Thus, in order to conduct effective classification, CLIP must be robust to color variations. The intuition behind this is that a duck, for example, pictured at sunset (red-tinted) or underwater (blue-tinted) is still a duck. Thus, our RGB overlays must have been neglected by CLIP's ability to see through color manipulations, suggesting a prioritization of geometric features like shape and structure. Another reason for this prioritization of shape over color could be that both ducks and rabbits exhibit significant color variation within their species, and consequently CLIP's training data, with rabbits ranging from white to brown and ducks ranging from grey to green. This implies that structural properties, such as long ears and bills, are more robust to reliably distinguish rabbits from ducks. While the real duck colorization produced almost tripled the effects of arbitrary colorizations, the difference from baseline scores caused by weak color-semantic associations (2% change through real duck color) pale in comparison to that caused by geometric manipulation (20.1% change through rotation).

One limitation of our dataset was that we could've used more photos when training our model on duck versus rabbit. In the grand scheme of things, 37 images for each category isn't very much and it's possible that the cosine similarities and classifications on the illusions would have been different if CLIP had been trained on more images. By including more standardized images that had the same resolution, background noise, centering of subjects etc, we could have reduced the influence of outliers and established better representations in CLIP's embedding space. Another limitation was that we didn't look at the geometric structure of the embedding space and only focused on cosine similarity scores. While cosine similarity helps determine image-text closeness, we potentially missed out on details like dimensionality and density. Implementing PCA visualizations may have revealed critical geometric insights about the illusion's exact spatial location relative to duck and rabbit centroids, whether it occupied a distinct region, the position of the decision boundary and the potential existence of clusters within manipulation groups (colored, cropped, rotated). Creating higher dimensional visualizations would have provided a stronger spatial structure to our understanding of CLIP's visual processing.

Lastly, our self report methodology had limited statistical validity, given our small team sample size of three. For more reliable human benchmarking, we would require a significantly larger number of participants. Moreover, the sequential presentation of images during self-reporting tasks caused priming behaviors suggesting that participants continued to lean towards the same interpretation across trials. This may have been influenced by the natural order of reading/viewing left-to-right. We could have randomized the order of ducks and potentially introduced confidence ratings rather than binary classification to better model the self-reporting against the continuous similarity scores (or their differences). Other measurements such as reaction time to classification may have been better indicators of perceptual ambiguity too.

Once we have overcome the limitations to a reasonable degree, our next steps would be to extend this experiment by conducting CLIP classification of different illusion types, or testing the same illusion with different unsupervised learning methods. We could test CLIP on multiple different illusion types such as ambiguous figures like Necker's Cube, color illusions like the viral 2015 dress, geometric illusions like the Müller-Lyer arrows, and illusory contours like Kanizsa's triangle. These distinct illusions could enable us to compute a perception ambiguity score/profile for CLIP relative to human perception. Our initial hypothesis, consistent with our findings, would be that CLIP handles geometric illusions better than the color-based ones, exhibiting ambiguity scores that match human perception. A systematic experiment with sufficient statistical validity in human self-reporting could aid in benchmarking vision language models on ambiguous visual perception tasks.

Alternatively, we could also apply other learning methods covered in class like K-means clustering and Gaussian Mixture Models to analyze CLIP's embedding space structure. Applying K-means with k=2, to our embeddings for all images, including the baseline illusion and its manipulations, we could observe whether the algorithm separates ducks from rabbits. Additionally, we could fit a GMM with 2 components to get probabilistic cluster assignments and even use PCA for the visualizations of the clusters and decision boundaries currently missing from our analysis. Our hypothesis would be that two distinct clusters would form out of K-means and our baseline illusion would be at the decision boundary. The manipulated images would potentially move gradually towards their respectively classified clusters, and the GMM would assign roughly equal probabilities to the classification of the baseline illusion. This extension could help us evaluate the effectiveness of clustering methods relative to our current approach of similarity-scoring.

Finally, we hope this project brings attention to the importance of evaluating foundation models not just on accuracy, but on how they represent perceptual ambiguity. Our findings reveal how CLIP's interpretation of ambiguous figures differs from human perception in important ways. Understanding these gaps is a step toward developing vision-language models that not only perform well on benchmarks, but also process visual information in ways that are scientifically and cognitively grounded in human perceptual mechanisms.

References

- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
- Freedheim, D. K., & Weiner, I. B. (2003). Handbook of psychology: Volume 1. History of psychology. Wiley.
- Muttenthaler, L., Doerig, A., Linhardt, L., Vandermeulen, R. A., & Kornblith, S. (2023). Improving neural network representations using human similarity judgments. arXiv preprint arXiv:2306.04507.
- Ngo, J., Csail, M., Sankaranarayanan, S., & Isola, P. (2023). Is CLIP fooled by optical illusions? Retrieved from https://jerryngo.com/assets/pdf/clip_illusion_preprint.pdf
- Srinivas-R. (2025). CLIP_THINGS.ipynb [Computer software]. GitHub. https://github.com/Srinivas-R/COGS118B_FA25_Project/blob/main/CLIP_THINGS.ipynb