REVISITING MODELS OF VISUAL CATEGORIZATION USING DEEP LEARNING MODELS: THE GENERATION OF NATURALISTIC VISUAL STIMULI THROUGH GANs

BY

ISABELLA LONGORIA-VALENZUELA

COGNITIVE NEUROSCIENCE HONORS THESIS

APRIL 2022

ADVISOR: PROFESSOR THOMAS SERRE

Submitted in partial fulfillment of requirements for the degree of Bachelor of Arts with Honors in the Department of Cognitive, Linguistic, and Psychological Science

ACKNOWLEDGMENTS

First and foremost, I would like to thank Lakshmi Govindarajan for his selfless help, encouragement, and training – I could not have completed this thesis without you. You have been an invaluable mentor to me since my days writing "for" loops in CLPS0950. Your wisdom both as a neuroscientist and a human being is appreciated more than you will ever know. Thank you for always magically having the solution to the toughest experimental problems and fixing them with your delightful sarcasm.

A huge thank you to my wonderful thesis advisor, Professor Thomas Serre, for placing me under Lakshmi's guidance and always providing me with support, feedback, and resources throughout the thesis process. You have been extremely patient and kind with me as I discover my interests and learn about computational cognitive neuroscience, and for that, I will always be grateful. Thank you both for believing in me and introducing me to so many amazing people and opportunities in this field.

I would also like to thank Lore Goetschalcx for her guidance throughout this process as someone also from a psychology background. Thank you for introducing me to the exciting computational framework of GANs and for making me feel more comfortable in this new computationally based field. In particular, thank you for walking me through the AMT setup process with such patience, compassion, and flexibility.

Thank you to all the other members of the Serre Lab for their gracious feedback related to my code, experimental paradigm, and visual stimuli.

Thank you to my friends for always bringing a smile to my face throughout this long thesis process.

Last but certainly not least, thank you to my amazing family for their unconditional support throughout my journey in science and for tolerating my blabbing about neuroscience since the age of 17. I would not be here today without your love.

TABLE OF CONTENTS

ABSTRACT

Our daily interaction with the visual world raises questions about how humans identify and categorize visual stimuli. Previous research on the Perceptual Magnet Effect (PME), which argues that a perceptual warping occurs at an ambiguous category boundary between category X and category Y, has focused on replicating this perceptual warping in other perceptual domains and using naturalistic visual stimuli. In the present study, GAN-generated stimuli were evaluated against human similarity judgments in a pairwise comparison similarity task. Specifically, the presented study used visual stimuli from the popular CIFAR-10 and FF-HQ datasets. By comparing average human similarity ratings during each interpolation step from category X to category Y, GAN-generated stimuli were measured on their ability to replicate the warping seen in the PME. The results suggest that the CIFAR-10 and FF-HQ visual stimuli show hints of perceptual warping within-class category and across-instance category boundaries, respectively, but additional experimental fine-tuning is needed to strengthen the results. These findings provide an important first step in using GAN-generated stimuli to replicate psychophysics experiments analyzing perceptual phenomena like the PME. In addition to improving the sample size and scaling up high-resolution visual stimuli, future work should aim to investigate the possible application of GAN-generated stimuli with modified latent spaces to other domains like memory and attention.

INTRODUCTION

Studying visual recognition allows us to understand how humans perceive and encode visual experiences in the world around them. One key aspect of visual recognition, categorization, highlights how humans group visual stimuli into distinct concepts. In psychological research, categorization has been eagerly studied to understand how both human and artificial brains identify everyday objects. Early psychological theories of categorization emphasized two possible explanations: A rule-based model of categorization, where people learn rules or definitions for categories (Hull et al., 1920) (Bruner et al., 1956), and a similarity-based model, where categories are defined in terms of a resemblance to other objects (Rosch et al., 1970).

These two rule-based and similarity-based theories, later modified to the commonly known prototype and exemplar models of categorization, have encouraged the conception of numerous mathematical models to support each theory. Creating such models allows researchers to construct experiments to distinguish between the two theories using visual stimuli specifically crafted to shift category membership depending on the model being examined.

One phenomenon that often guides the evaluation of exemplar and prototype theories is the Perceptual Magnet Effect (PME). The PME underscores how categories and category boundaries influence perception, and previous literature has been focused on its influence on speech sounds. Kuhl et al. (1991) first coined the term after finding that poor phonetic category discrimination occurs near the phonetic category prototypes. More broadly, the Perceptual Magnet Effect can be defined as a phenomenon where category prototypes pull – like a magnet – neighboring stimuli closer to them.

Some have suggested that categorical perception outside of speech may be conceptually different from the PME (Kuhl and Iverson, 2000), but most prior work has demonstrated that the PME can be qualitatively compared to other categorical effects. For example, Feldman and Griffiths (2007) related perceptual warping seen in speech perception to visual stimulus production. A common visualization of the Perceptual Magnet Effect can be seen in ***Figure 1***.

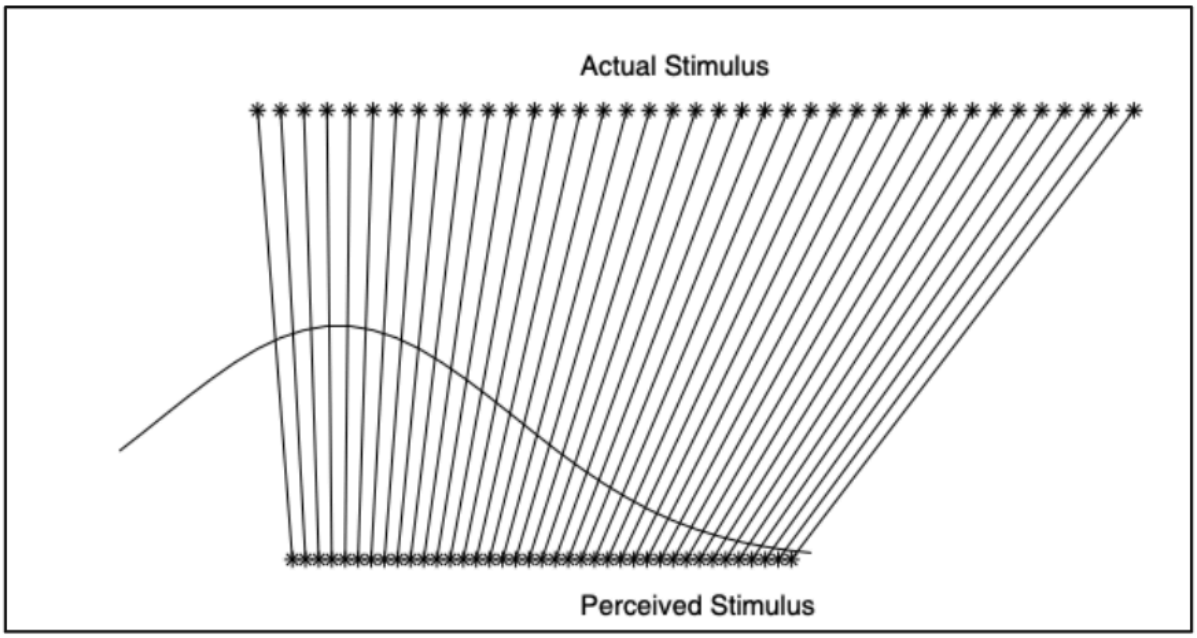


Figure 1 **- The Perceptual Magnet Effect**: Griffith and Feldman's (2007) visualization of the Perceptual Magnet Effect. Each * under "actual stimuli" represents a physical stimulus within a single category, while each * above "perceived stimuli" represents the perceived stimuli within that same category, The curve represents the relationship between the actual stimuli and the perceived stimuli for that category. As illustrated above, category prototypes pull neighboring stimuli closer to them, causing the discrepancy between actual stimulus distance and perceived stimulus difference to have a non-linear effect that is more prominent at category boundaries.

Specifically, Feldman and Griffiths (2007) compared category boundaries in speech to category boundaries in visual stimuli using stimuli from a Huttenlocher et al. (2000) study. In Huttenlocher et al. (2000), subjects were given the category structure and, consequently, used that structure to categorize visual stimuli when faced with uncertainty in a memory encoding

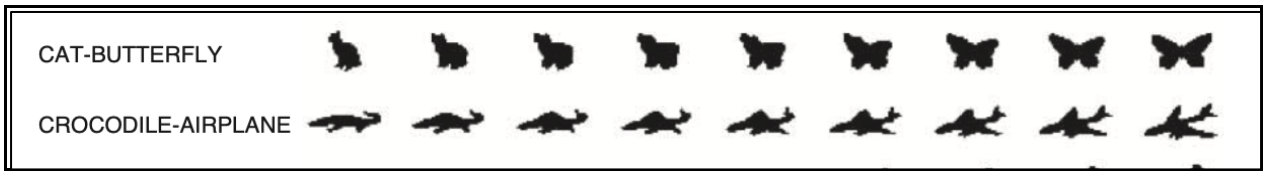
task. Importantly, in these other-domain categorical effects, perceptual space also shrinks near category centers, producing what can be called a "perceptual warping."

A study by Guenther and Gjaja (1996) furthered Kuhl et al. (1992)'s work by illustrating *how* the PME emerges from the uneven distribution of speech stimuli. The researchers suggest that perceptual warping occurs from the specific distribution of speech sounds and not from category labels or "exemplars." Specifically, on a neuronal level, the neural firing preferences in speech sound categorization take on a Gaussian distribution. Furthermore, Guenther and Gjaja (1996) found that the central sounds possessed a stronger neural representation than the peripheral category speech sounds. As a result, a speech sound halfway between the category periphery and the category center will perceptually appear closer to the center, or the "exemplar" category sound, rather than the peripheral category sound.

Griffiths et al. (2009) unified previous literature on the PME through a Bayesian computational model that deployed for both perceptual and psychophysical category mapping strategies to assess the perceptual warping. Their model highlights potential reasons for the visual categorical variability seen in the PME across different domains, specifically in speech sounds, colors, and faces. The researchers suggest that normalized Gaussian noise, perceptual warping, or other category latent space shifts may influence category boundaries. Within perception, latent space describes the underlying psychological space of perceived features and properties that guide similarity judgments between two visual stimuli (Suchow et al. 2018). Griffiths et al. (2009) also point out that these categorical judgements may be based on learned categories, specifically by implicit categories formed by the specific distribution of exemplar stimuli. Their study accounts for both one-category and multiple-category cases, allowing their computational framework to be generalizable to the many categorical possibilities a listener

encounters during natural speech perception. Similarly, the visual stimuli used in the present study require the participants to identify stimuli from two categories. For example, participants must distinguish between an image of a dog and a cat or between an airplane and a boat without prior knowledge of what categories were to be presented, much like everyday visual classification.

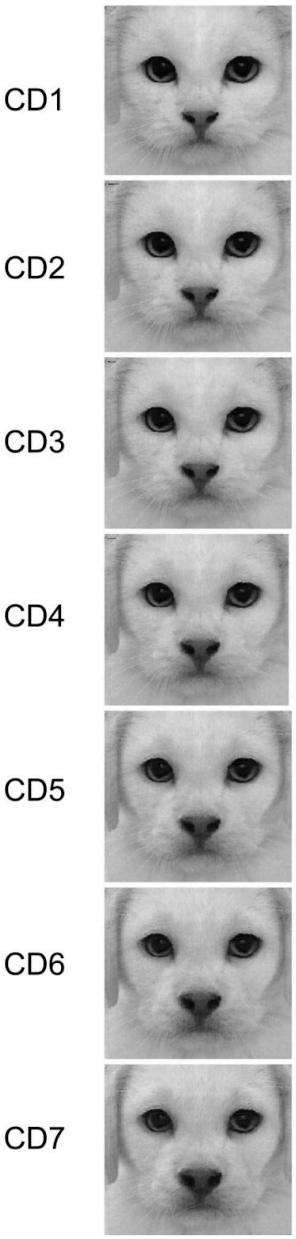
Previous research has used simple visual stimuli to study the PME and its influence on visual perception and categorization. The ability to physically manipulate the stimuli properties allows researchers to characterize changes in perceptual and cognitive ability. For example, Hartendorp et al. (2010) used silhouette stimuli to evaluate the categorization of morphed objects during a free-naming experiment. The simplistic, contextless nature of stimuli, illustrated in ***Figure 2***, allowed the researchers to conclude that categorical perception depends on the wholeness of the visual stimuli structure during the perceptual morphing from one stimulus to another. Notably, in experiments like Hartendorp et al. (2010) and others, the "toy," algorithmically generated stimuli do not provide an ecologically valid representation of the everyday perception of our visual world.



(above) ***Figure 2*** **- Example of algorithmically generated, "toy" stimuli used in previous perceptual categorization experiments**: Example interpolations generated in the Hartendorp et al. (2010) study. The stimuli morphed from a silhouette of a cat to a silhouette of a butterfly, and from a crocodile to an airplane, respectively.

Using handcrafted naturalistic stimuli, Hampton, Estes, and Simmons (2005) conducted a series of experiments evaluating how human participants categorized borderline perceptual stimuli in the context of other similar stimuli or within and across a category boundary. The first few experiments in the study used pairs of hues between the colors purple and blue, and their

final experiments used perpetually similar black and white photographs of a dog and cat, as shown in the handcrafted vertical interpolation of ***Figure 3***. Although Hampton, Estes, and Simmons (2005) were able to make valuable contributions as to how stimuli are categorized based on their visual and perceptual contrast to a category relevant stimulus, the small dataset and limited modalities tested impact the overall generalizability of their findings. Importantly, the stimuli used were handcrafted naturalistic stimuli, which require extensive effort and time to generate. Thus, it remains unclear how these results would generalize to additional stimuli classes.



(left) ***Figure 3*** **- Example of handcrafted naturalistic stimuli used in previous perceptual categorization experiments**: Handcrafted interpolation generated in a study by Hampton, Estes, and Simmons' (2005). The stimuli morphs from a photo of a cat (frame CD1) to a photo of a dog (frame CD7). The stimuli were cropped to only show the face and were gray scaled to allow for minimal contextual and background interference.

Constructing automated but naturalistic stimuli in psychophysics, categorical perception, and PME experiments continues to be an important next step in the field. Generating these images will enhance the ecological validity of the stimuli itself and will provide a more effective naturalistic setting for identifying category boundaries – one that more closely resembles our everyday visual context. Related subfields of scene recognition and face perception have successfully used naturalistic stimuli to study visual perception (Fei-Fei et al., 2005) (Torralba et

al., 2003). These experiments tend to create stimuli in an isolated computational and neuroscience-based modeling framework. Therefore, to bridge findings in both disciplines, this study grounds itself in a computational modeling framework from both computational neuroscience and psychology literature.

To generate automated, naturalistic stimuli for visual recognition tasks, researchers have been forced to consider the perceptual and mathematical features that underlie categorical perception in humans. One way to construct human-like computational models of categorization is by evaluating how closely model-perceived perceptual similarity aligns with human-perceived similarity. Most research in this field, like the popular ImageNet Large Scale Visual Recognition Challenge (ILSVRC), seek to establish a benchmark for predicting class categories, rather than evaluating the perceptual progression between categories. Roads and Love (2021) expanded the ILSVRC validation set to create ImageNet Human Similarity Judgment (ImageNet-HSJ), creating a more general benchmark of human perception and reasoning. Using the ImageNet-HSJ allowed the researchers to assess how well human similarity judgements of visual stimuli (i.e., respondents provided a similarity rating while viewing a beer bottle and a soda can versus a beer bottle and a cigarette) aligned with popular computational models of visual perception. Their findings point to the experimental usefulness of psychological embedding spaces, which are extracted from human judgements, that can then be used to infer categorical similarity between visual stimuli.

Another recent study by Peterson et al. (2018) combined cognitive science and machine learning methods to introduce a method of estimating the structure of human categories. In their proof of concept, the researchers used a dcGAN (deep convolutional generative adversarial network) and a biGAN (bidirectional generative adversarial network) to mathematically evaluate

category boundaries and classification on images from the Asian Faces and ILSVRC12 datasets. Afterwards, their model-generated samples were compared to human classification ratings on images from all ten categories present in the prior datasets. Peterson et al. (2018)'s findings importantly demonstrated that generative models allowed for the visualization of multi-modal category templates and for a better approximation of human classification ratings. Furthermore, the results from Peterson et al. (2018) highlight the possibility that diverse datasets and categories can be effectively classified by a generative model.

The current study hopes to expand upon the current categorical perception literature by specifically improving the visual stimuli themselves as an important first step in validating and refining automated, naturalistic stimuli against human similarity ratings. To do so, generative adversarial networks (GANs) were chosen to generate the photorealistic stimuli. Prior literature in artificial intelligence (AI) has pointed to GANs as a valuable computational method for constructing images that balance both experimental control and ecological validity (Goetschalckx et al., 2021) Most relevant to the current study is a GAN's powerful ability to generate a continuous space of stimuli while maintaining latent space control. Unlike handcrafted, naturalistic stimuli, which take a long time to generate, GANs can be scaled up to generate thousands of controlled, high-quality images in a much shorter period. Thus, GANs present a promising computational method for generating sequences of automated, naturalistic stimuli between two image categories that can subsequently be measured against human similarity judgements.

In the present study, GAN-generated visual stimuli were evaluated against human similarity judgements in an attempt to replicate the perceptual warping seen in the Perceptual Magnet Effect. Like the Peterson et al. (2018) study, this study trained the well-established

CIFAR-10 dataset on a conditional dcGAN and a StyleGAN. Additionally, the FF-HQ dataset and the silhouette dataset from Hartendorp et al. (2010) assess perceptual differences in visual stimuli quality and context. All experimental stimuli were then interpolated to create 16-step visual interpolations from stimuli class A to stimuli class B. To compare the three datasets of generated visual stimuli to human perceptual similarity judgements, an Amazon Mechanical Turk experiment was run to collect similarity ratings from pairwise presentations of the generated images. If the GAN-generated stimuli effectively capture the categorical boundary effects seen in the PME, human similarity ratings should exhibit a perceptual warping in its transformation from a visual stimulus in class A to a visual stimulus in class B.

METHODS

Participants:

Participants were recruited online via Amazon Mechanical Turk (AMT) to complete the human similarity judgment task for the CIFAR-10, FF-HQ, and the Hartendorp Silhouette datasets. No participants were excluded from the final analysis, as exclusion criteria were pre-set in AMT. To ensure high quality results, all participants had to be in the United States, had to have been accepted as a participant in 95% of their previous experiments, and had to have participated in over 1,000 experiments via the AMT platform. The task was conducted entirely online via AMT and all participants were compensated for their participation with a monetary reward per image comparison submitted.

For the CIFAR-10 dataset, each comparison of two perceptually sequential images were evaluated by nine unique participants, making for a total of 4,275 similarity ratings for the batch of 475 pairwise comparison. In the FF-HQ dataset, each comparison of two perceptually sequential images were also evaluated by nine unique participants, resulting in a total of 675

similarity ratings for the batch of 75 pairwise comparisons. Finally, on the Hartendorp Silhouette dataset, each comparison of two perceptually sequential images were evaluated by nine unique participants, making for a total of 288 similarity ratings for the batch of 32 pairwise comparisons.

Human Similarity Judgment Task:

To measure perceptual similarity between the generated stimuli participants completed a perceptual task that asked them to rate the similarity between two of the images on a scale from 1-9. The template was modified from an AMT Item Equity template to put the two generated images side by side on the screen. Each trial consisted of pairwise comparison between two generated images from randomized positions in two different classes. Alongside the two generated images, instructions asked the participants to click on the score (1-9), 1 being the extremely different and 9 being extremely similar, based on how similar they thought the two images were. The scale was labeled as follows: 1 - extremely different, 2 - very different, 3 - mostly different, 4 - somewhat different, 5 - neither different nor similar, 6 - somewhat similar, 7 - mostly similar, 8 - very similar, and 9 - extremely similar. Each participant was expected to spend around 25 minutes evaluating the collection of pairwise stimuli but was given up to 1 hour to do so. An example trial of the similarity judgment task presented to participants is shown in ***Figure 4***.

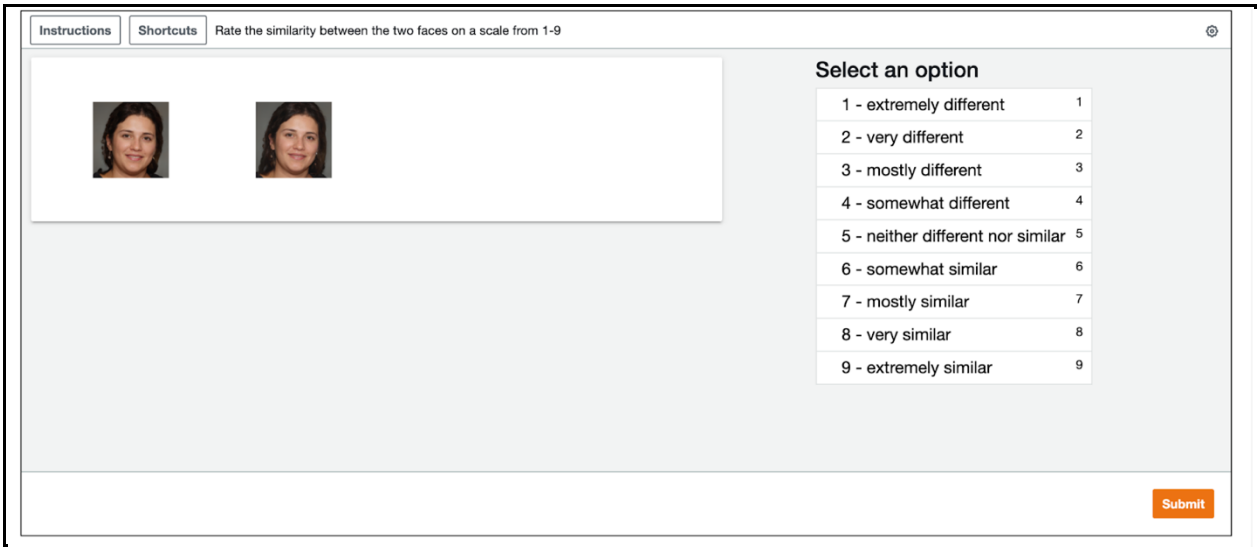


Figure 4 **- Example Trial of the Human Similarity Judgment Task on the FF-HQ dataset on Amazon Mechanical Turk**: Visualization of the similarity judgment task that participants completed. Participants were instructed to look at the two stimuli on the left and select a box on the right based on how similar they perceived the two images to be. For the FF-HQ dataset, participants were asked to evaluate the similarity between the two "faces" rather than "images." After clicking on an option bar, the participant had to select a submit button before moving to the next trial. Here, the participant was viewing pairwise stimuli from an interpolation between a young man and a young woman with long hair.

CIFAR-10 Dataset:

The first set of visual stimuli used in the human similarity judgment task was generated using a conditional StyleGAN trained on the CIFAR-10 dataset. The CIFAR-10 dataset consists of 60,000 32x32 color images from 10 different image classes. There are 6,000 images in each of the 10 classes. The 10 classes used during training were airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.

FF-HQ Dataset:

To provide quality control on the contextual variability of the CIFAR-10 dataset, the second set of visual stimuli evaluated in the human similarity judgment task were generated using a conditional StyleGAN trained on the Flick-Faces-HQ (FF-HQ) dataset. The FF-HQ dataset consists of 70,000 high quality faces taken from Flickr in a 1024 x 1024 resolution. In the present study, a subset of the face stimuli were chosen for interpolation.

The Hartendorp Silhouette Dataset:

Finally, to compare machine-generated, naturalistic stimuli to algorithmically generated, "toy" stimuli, a final set of images used in the human similarity judgment task were taken from the morphed silhouette images used in Hartendorp et al. (2010). The entire dataset consisted of 15 morphed interpolations made using the Sqirlz-Morph software. The morphs were constructed between two living objects, two non-living objects, and between living and non-living objects. Additionally, the series were created between different basic-level categories. In the present study, a subset of the morphs (cat to butterfly, crocodile to airplane, gun to rabbit, and truck to peacock) were selected to be shown and evaluated by participants during the human similarity judgment task.

Neural Network Architecture:

To determine which specific GAN architecture would produce the best naturalistic images, two GAN architectures were tested and compared for image quality: a conditional dcGAN and a conditional StyleGAN. In a basic GAN architecture, two computational models work against each other during training. In the first model, the generator learns to generate samples of a given image distribution. The second model, the discriminator, learns to identify whether the samples generated by the generator model are real or fake (Goetschalckx et al. 2021).

The first network model, a dcGAN, is a direct extension of a GAN architecture and uses convolutional and convolutional-transpose layers in both the generator and the discriminator. These transposed convolutional layers perform up-sampling of the 2D image size initially fed into the network (Inkawhich, *DCGAN tutorial*). The network for this study was modified for the

CIFAR-10 dataset from an official PyTorch implementation of a DcGAN, which can be found here: https://github.com/pytorch/examples/blob/main/dcgan/main.py.

After initially establishing baseline image results, the DcGAN model was retrained to be class conditional in order to interpolate between image classes. To do so, the DcGAN architecture was modified via this PyTorch implementation of a cDCGAN (conditional dcGAN) by GitHub user "togheppi": https://github.com/togheppi/cDCGAN. Later, additional modifications were made to increase learning, improve image quality, and accommodate the CIFAR-10 image dimensions. Initially, the network was trained on 50 epochs, but was later changed to 100 and then to 200 to further enhance network learning. The cDCGAN was also re-trained to increase the image pixel size from 32 to 64.

The conditional StyleGAN neural network is an alternative architecture for generative adversarial networks (GANs). Unlike a traditional GAN, the latent code in a StyleGAN is not fed through an input layer but rather through a learned constant, an intermediate latent space that then controls the generator model through adaptive instance normalization at each convolutional layer. The computational model was modified to accommodate a CIFAR-10 dataset from this PyTorch implementation of a StyleGAN2: https://github.com/NVlabs/stylegan2-ada-pytorch. For the FF-HQ dataset (https://github.com/NVlabs/ffhq-dataset), a StyleGAN3 was modified from an official PyTorch implementation found here: https://github.com/huangzh13/StyleGAN.pytorch.

Class Conditional Network Training:

After training both the cDCGAN and the StyleGAN, images generated from the generator of each model were compared. Conclusively, the StyleGAN was found to produce images with greater clarity and quality. This comparison is in line with the current literature on StyleGANs, as StyleGANs are known for being able to produce a high image quality that is often

compromised in other GAN architectures when training on the CIFAR-10 dataset (Karras et al., 2020). Additionally, the StyleGAN architecture improves traditional distribution quality metrics, which in turn produces better interpolation properties (Karras et al., 2020). Training the StyleGAN resulted in the generation of 23,040 randomized images from the CIFAR-10 dataset. In the FF-HQ dataset, 80 face stimuli were selected for interpolation after training. Notably, the Hartendorp Silhouette stimuli were algorithmically generated, so they were not trained on a neural network.

Interpolation:

To interpolate each of the generated images along class boundaries in the CIFAR-10 dataset, a python script accounting for the class label, random seed generation, and the number of steps in each interpolation series (16) was created. Another Python script generated both individual images representing each position within the series and a complete 16-step interpolation series image. Sequential stimuli positions (i.e., Position 3 and Position 4, Position 5, and Position 6) in all interpolations were then paired to be used as the pair-wise stimuli presented in the AMT online experiment.

Specifically, for the CIFAR-10 stimuli, interpolation between class labels was completed from Class A to Class B, and from Class B to Class A for all 10 class labels. A total of 1,666 pairs of sequential images were generated for all interpolations in the complete CIFAR-10 set. For the FF-HQ dataset, five 16-step interpolations were produced. A subset of pairwise comparisons taken from the CIFAR-10 interpolations, consisting of 475 comparisons, was shown to the AMT participants during the human similarity judgment task. In the FF-HQ dataset, five interpolations, resulting in a total of 75 comparisons, were shown to AMT participants.

Finally, the Hartendorp dataset consisted of 9-step interpolations that were split into pairwise comparisons, resulting in the presentation of 32 comparisons during the human similarity judgment task. An example of an interpolation series and the generated stimuli from each of the three experimental datasets can be found below in ***Figure 5a*** (Hartendorp)***, Figure 5b*** (CIFAR-10)**,** and ***Figure 5d*** (FF-HQ)**.**

RESULTS

Figure 5 **- Example Interpolations (9-step) from Different Visual Stimuli Datasets**:
a) Algorithmically generated, "toy" dataset: an example interpolation between a gun and a rabbit in the Hartendorp Silhouette dataset. **b)** Automated, naturalistic dataset: CIFAR-10 example interpolation from a dog to a cat. **c)** Another example of an automated, naturalistic dataset: an interpolation between a fox and a wolf in the Animal Faces Dataset, which was not used in the current experiment. **d)** A high-resolution automated, naturalistic dataset (context-controlled): FF-HQ dataset example interpolation between a young man and woman with long hair. Each interpolation from a-d also increased in resolution, with CIFAR-10, the Animal Faces, and the FF-HQ datasets having a pixel resolution of 64x64, 512x512, and 1024x1024, respectively.

Hartendorp Silhouette Dataset:

First, to evaluate the feasibility of the Perceptual Magnet Effect (PME) against human similarity judgments on Amazon Mechanical Turk, the Hartendorp Silhouette dataset was used as a positive control. An example interpolation from this dataset can be seen in ***Figure 5a.*** As illustrated in ***Figure 6***, average similarity ratings for the Hartendorp dataset reveal a perceptual warping effect, as indicated by the lower average similarity rating in positions 4-7. Thus, these results demonstrate that the PME can be replicated using the current experimental paradigm on algorithmically generated, "toy" stimuli like the Hartendorp dataset.

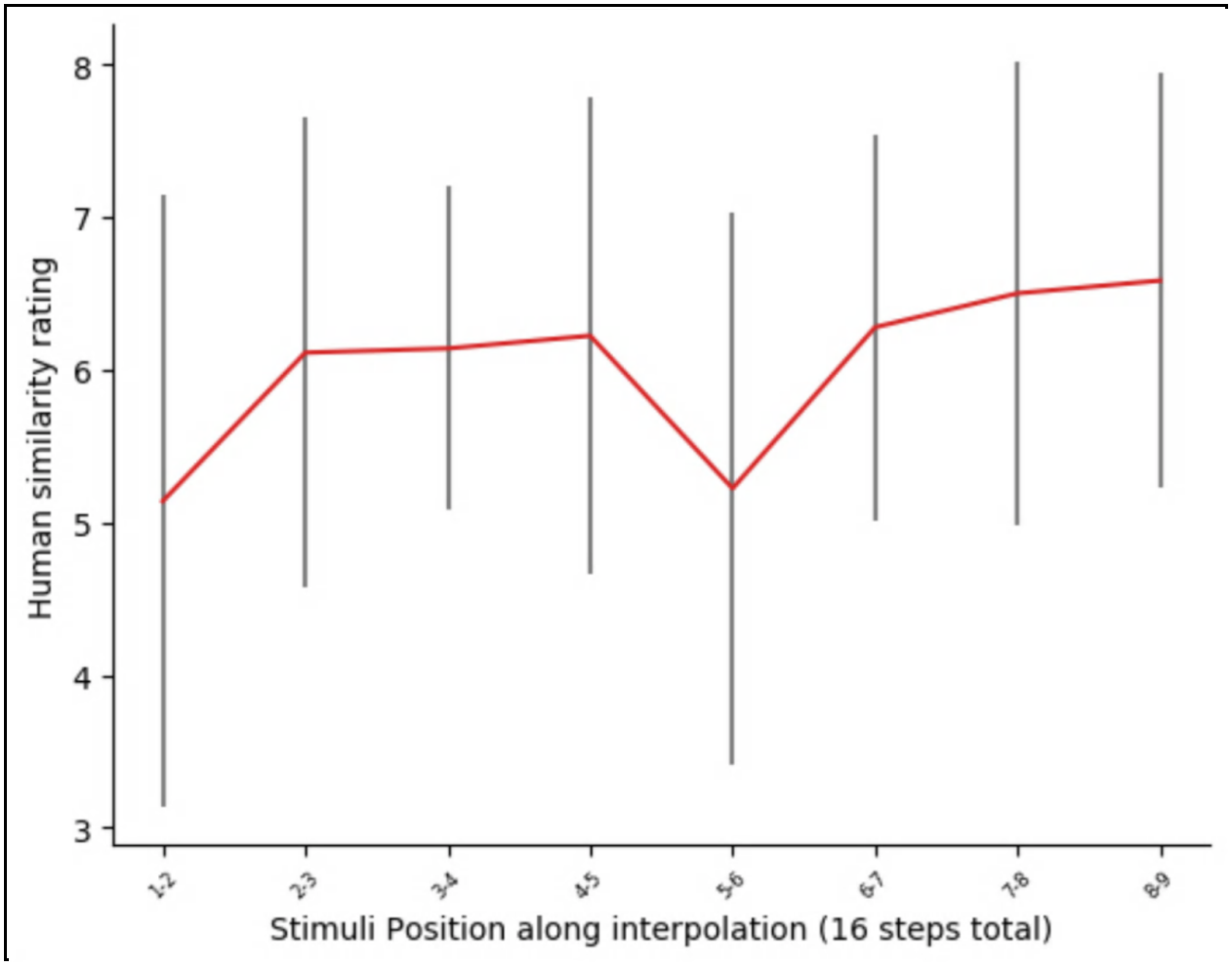


Figure 6 **- Average Human Similarity Rating on the Hartendorp Silhouette Dataset:** Average human similarity ratings on the human similarity judgment task for the Hartendorp Silhouette dataset along each of the 9 interpolation steps. The drop in average similarity rating between stimuli positions 4-7 illustrate

where participants perceive the largest perceptual difference between the two stimuli. Error bars represent one standard deviation from the average human similarity rating at that interpolation position.

CIFAR-10 Dataset:

To replicate the encouraging findings seen in the Hartendorp dataset on automated, naturalistic stimuli, the experimental paradigm was repeated using the CIFAR-10 and FF-HQ datasets. To generate visual stimuli that could be interpolated between multiple classes (class A to class B, class B to class C, class C to class B), the CIFAR-10 dataset was trained on a conditional StyleGAN. An example of the resulting visual stimuli from CIFAR-10 training can be seen via an example interpolation in ***Figure 5b*.** Notably, the human similarity judgment task on the CIFAR-10 dataset did not reveal a strong PME across all 10 CIFAR-10 classes ("within-instance").

However, analyzing the average human similarity rating within a class did indicate a perceptual curvature like that in the PME. For example, As seen in ***Figure 7***, average human similarity ratings were the highest near category prototypes (stimuli interpolation positions 1 and 16) and were the lowest near category boundaries (stimuli interpolation positions 4 through 7) in Class 0 (airplanes). This trend indicates that participants perceptually evaluated stimuli near the category boundary as being more dissimilar, even though the physical latent space between each interpolation was the same. Thus, a noticeable "perceptual warping" occurred within a category in the CIFAR-10 dataset. ***Figure 8*** illustrates similar warping effects for another class, Class 9 (trucks), with the lowest average similarity rating occurring at stimuli interpolation positions 8 through 11.

The lack of a strong PME in the CIFAR-10 is likely due to several confounding factors. For example, the low resolution of the generated stimuli (64 x 64) could have forced AMT participants to use other metrics, like texture, color, or context to quantify stimulus similarity.

a)

b)

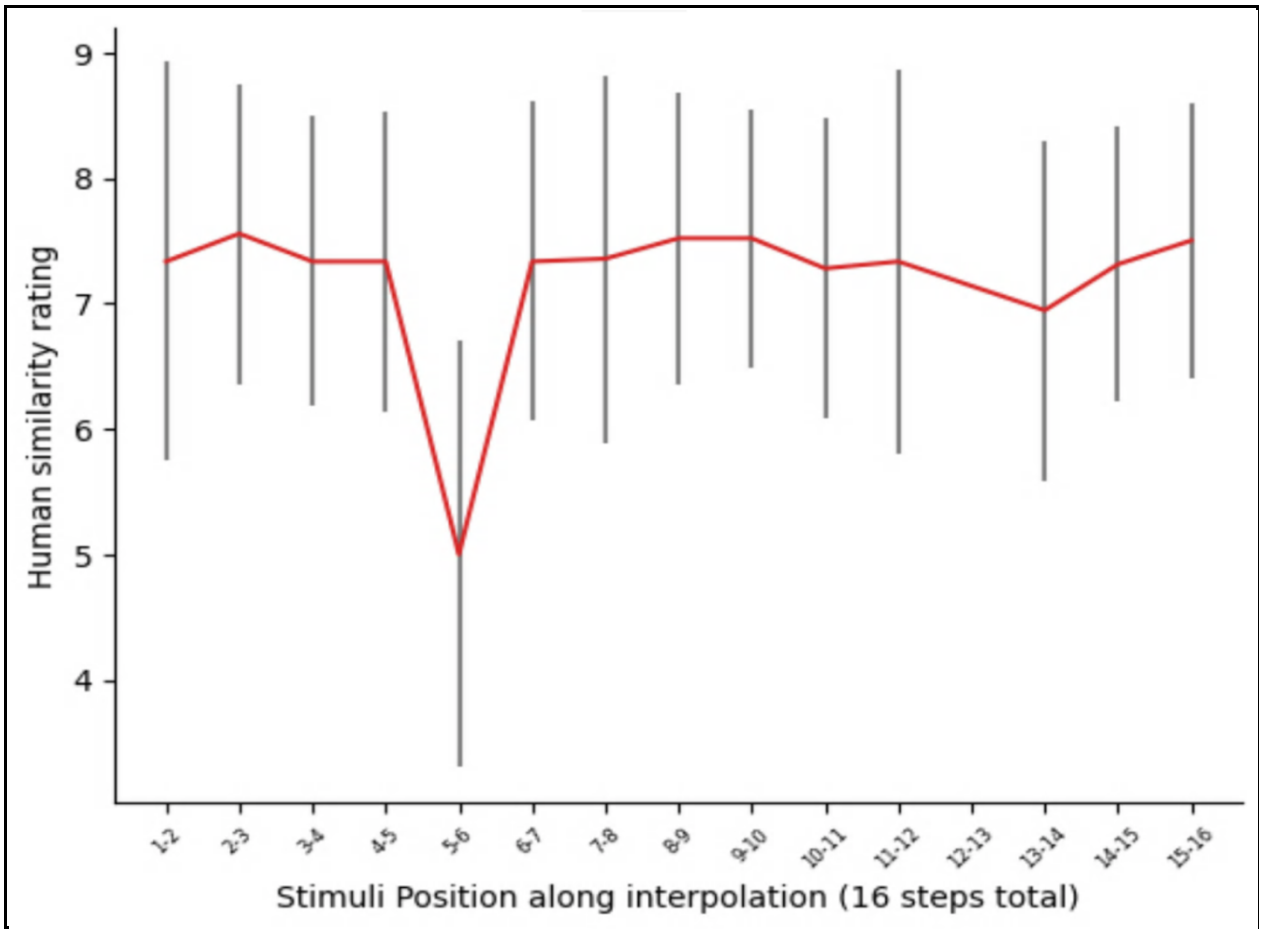
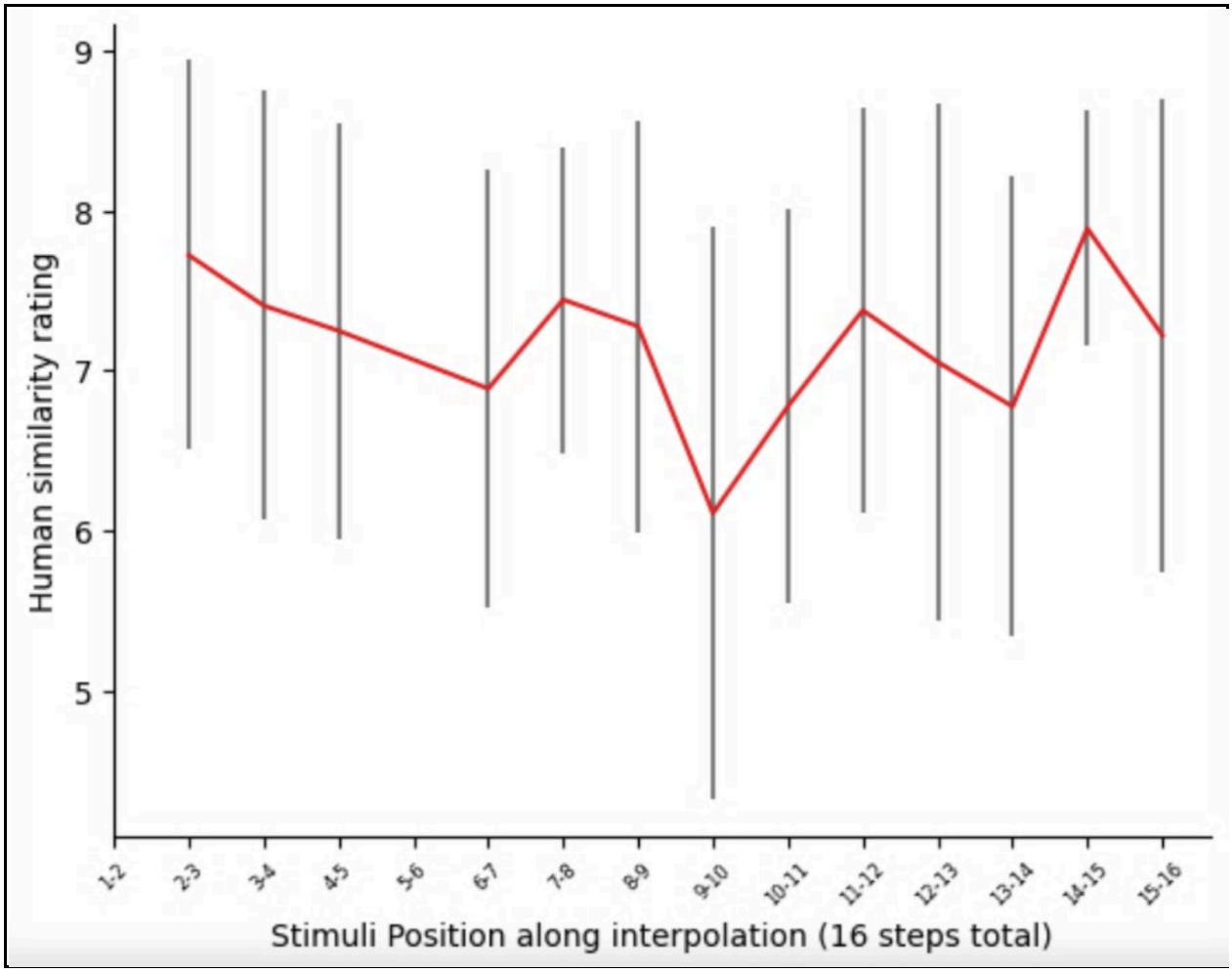


Figure 7 **- Within-class Interpolation on the CIFAR-10 Dataset Evaluated Against Human Similarity Ratings: a)** An example interpolation between Class 0 (airplanes) and Class 4 (deer). **b)** Average human similarity ratings on the human similarity judgment task between Class 0 (airplanes) and any of the other 9 classes for each of the 16 interpolation steps. The red box in ***Figure 7a*** illustrates where humans are most likely to perceive the largest perceptual difference between two stimuli within the interpolation, even though each position along the interpolation changes by equal steps. Within this

interpolation, the largest perceptual difference also happens to be close to the category boundary between the airplane and the deer. Error bars represent one standard deviation from the average human similarity rating at that interpolation position.

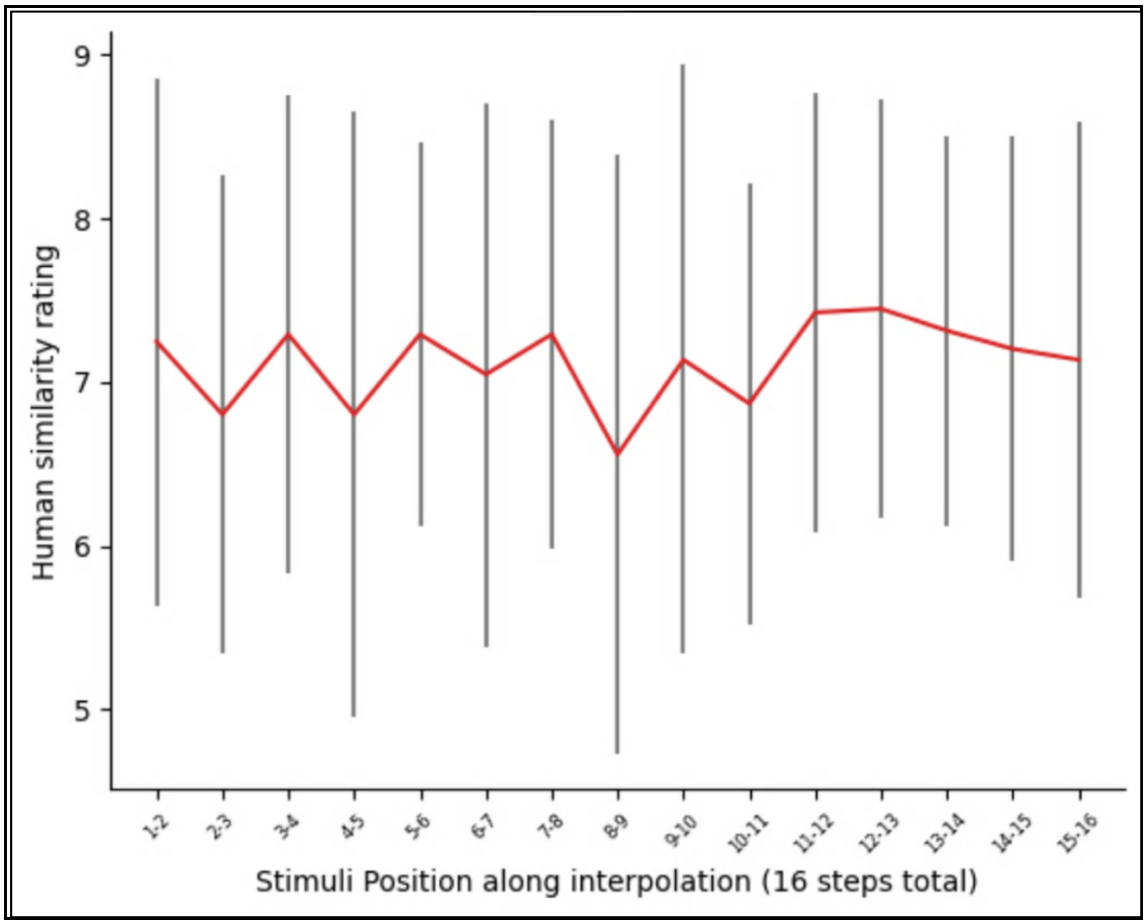


(above) ***Figure 8*** **- Within-class Interpolation on the CIFAR-10 Dataset Evaluated Against Human Similarity Ratings:** Average Human Similarity ratings on the human similarity judgment task for Class 9 (trucks) and any of the other 9 classes for each of the 16 interpolation steps. The drop in average similarity rating between stimuli positions 8-11illustrate where participants perceive the largest perceptual difference between the two stimuli. Conversely, the "ends" of the interpolation (position 1 and 16) have higher average similarity ratings, thus indicating where humans perceive the smallest perceptual difference between the two stimuli. Error bars represent one standard deviation from the average human similarity rating at that interpolation position.

FF-HQ Dataset:

To untangle one of these confounding factors, context, another automatic, naturalistic dataset, FF-HQ, was evaluated by AMT participants. Unlike the CIFAR-10 dataset, the FF-HQ is

context controlled, meaning that during interpolation, the background context stays constant. Additionally, the dataset stimuli had a higher resolution of 1024x1024. An example of a generated interpolation from the FF-HQ dataset can be observed in ***Figure 5d***. As shown in ***Figure 9***, average human similarity ratings exhibited hints of perceptual warping, with a lower average human similarity rating occurring in stimuli interpolation positions 7 through 10. Due to the time it takes to generate the high-quality resolution seen in the FF-HQ dataset, a smaller number of stimuli and interpolations were generated in comparison to the CIFAR-10 dataset. Thus, the weaker PME effect seen in the FF-HQ dataset could be attributed to a smaller number of interpolations and the perceptual curve may be strengthened with additional interpolations or more participants. Additionally, the smaller number of interpolations did not permit within-class analysis, as was done on the CIFAR-10 classes.



(above) ***Figure 9*- Average Human Similarity Rating on the FF-HQ Dataset:** Average human similarity ratings on the human similarity judgment task for the FF-HQ dataset along each of the 16 interpolation steps. The average human similarity reveals a slight perceptual difference between stimuli interpolation positions 7-10. Error bars represent one standard deviation from the average human similarity rating at that interpolation position.

DISCUSSION

The present study demonstrates promising results for using automatic, naturalistic visual stimuli to evaluate perceptual similarity. Average human similarity ratings on the Hartendorp Silhouette dataset revealed that the human similarity judgment task on AMT was a viable experimental paradigm to test perceptual warping in visual stimuli. Though the across-instance average similarity ratings on the CIFAR-10 dataset did not demonstrate significant perceptual warping,

within-class analysis on the CIFAR-10 dataset indicated lower human similarity ratings at category boundaries, and thus producing a noticeable warping effect. Similarly, the FF-HQ dataset illustrated hints of perceptual warping at category boundaries.

Future work should carefully consider sample size, experimental protocol, and inherent noise from AMT when evaluating human similarity judgements using GAN-generated stimuli. Perceptual warping effects for each of the three experimental stimuli datasets might have been more robust with a larger sample size of both participants and interpolations tested. Every pairwise comparison in the three datasets was evaluated by 9 unique participants, but more participants should be included to strengthen observed trends. Though the Hartendorp Silhouette dataset did illustrate perceptual warping, the warping and the category boundary for the interpolations may become pronounced with the addition of the 11 other interpolations included in the original Hartendorp et al. (2010) study. Due to timing constraints and pre-set Amazon Mechanical Turk experimental paradigm configurations, larger datasets like CIFAR-10 and FF-

HQ could only be evaluated in groups of 475 comparisons at time. Ideally, a revised experimental setup would allow AMT participants to complete trials for all comparisons within the CIFAR-10 dataset in one sitting. Particularly for the FF-HQ dataset, more face stimuli should be evaluated during the human similarity judgment task to standardize for the natural variability of faces in the dataset.

Presentation of the visual stimuli may have also affected the accessibility and accuracy of the human similarity judgment task. Due to the low resolution of certain datasets, particularly the 64x64 resolution of the CIFAR-10 dataset, the visual stimuli may have looked unclear to some of the participants during the task. Furthermore, the standard configurations of the AMT task setup display each image at its current pixel size, making each visual stimuli physically smaller and potentially less discernible, as can be seen in ***Figure 4.*** In terms of the experimental protocol, participants were only given simple instructions to complete the task. The instructions were as follows: "Determine the similarity between the two images [or in the case of FF-HQ: faces], on a scale from 1-9, with 1 being extremely different and 9 being extremely similar." Because participants may have different perceptions of what they deem "similar" or "different," future iterations of this paradigm should provide a reference pairwise comparison for how to calibrate similarity ratings.

Additionally, presentation quality and participant response quality can only be minimally controlled on Amazon Mechanical Turk. In an ideal experimental setup, the experimental paradigm would be hard-coded, rather than modified from a premade Amazon template, to best fit the current study's goals. Increasing the number of participants is also essential due to the variability in participant quality on AMT. Though the present study was configured with the intention of excluding participants who did not have a reliable response record on the platform,

there is still a possibility that participants were randomly clicking on similarity rating responses during the task or were disengaged during the task's completion.

Although the evaluation of human similarity judgements on the FF-HQ datasets allowed more contextual control, other perceptual cues should be taken into consideration during stimuli generation and experimental testing. For example, in the positive control condition using the Hartendorp Silhouette dataset, participants were not able to use other perceptual cues, like texture and color, to evaluate stimuli similarity. Therefore, although the Hartendorp dataset was effective in illustrating the potential of the experimental paradigm on "toy" stimuli, the average human similarity ratings may not encapsulate all components of human perceptual similarity judgment or demonstrate as much of a pronounced effect as it would in other domains.

In addition to improving the experimental paradigm and the visual stimuli, future research should focus on manipulating the latent space between each stimulus position to regularize the perceptual space. A useful starting point would be taking the human similarity judgments (HSJ) collected in the aforementioned study by Roads and Love (2021) and applying it to the GAN latent space used in this present study. The Roads and Love (2021) study directly applied the generated HSJ data to a psychological embedding (latent space) trained on the human similarity judgements. Because the GAN latent space reflects the step size inherent to the neural network and not the psychological and perceptual latent space that humans observe, embedding the HSJ could regularize the GAN latent space. Re-running the human similarity judgment task on this modified latent space could highlight the potential of GAN-generated stimuli in mimicking human-like models of visual categorization and perception.

Conversely, GAN-generated stimuli can also be "perceptually straightened," rather than regularized, through latent space manipulations. A study by Hénaff et al. (2019) proposed that

humans internally transform perceptual judgments needed to perceive continuous stimuli. This "temporal straightening" hypothesis provides a methodology for examining the perceptual curvature of an internal trajectory from human perceptual and similarity judgments. Hénaff et al. (2019) investigated this hypothesis on natural videos and found that humans internally "straighten" perceptual latent space when the visual stimuli are highly curved, as would be the case with the PME. Thus, extending these findings to GAN-generated stimuli could reveal meaningful differences between our "internal" and "external" visual perception of categories.

Other future work could use these high-quality generated stimuli to study other cognitive domains like memory and attention. For example, Goetschalckx et al. (2019) manipulated GAN-generated images to investigate memorability, aesthetics, and emotional valence. By navigating the latent space along a desired quality, the researchers were able to visualize what properties make a visual stimulus memorable. Their results demonstrated that the latent space manipulations did correlate with differential human memory performance. Unlike the present study, Goetschalckx et al. (2019) explored these image properties by training images on a class conditional BigGAN rather than a StyleGAN. Furthermore, the visual stimuli originated from pretrained ImageNet images. Therefore, future extensions in this field should also consider comparing different types of conditional GANs and high-quality image datasets to generate optimal experimental results. Furthermore, their model framework, called GANalyze, also alters perceptual attributes like brightness, color, and object size in their assessment of memorability. Importantly, their successful manipulations highlight the potential of using GANalyze to investigate how these attributes could influence human attention and other aspects of vision.

A similar latent space manipulation was explored by Suchow et al. (2018) in their study of human face perception. Rather than manipulating memorability, as Goetschalckx et al. (2019)

did, this study controlled perceptual latent space relating to human identity and appearance using a variational autoencoder. The researchers were able to generate high-quality, photorealistic human portraits using a smooth and navigable psychological latent space. Applying the knowledge from Suchow e al. (2018) study to the current GAN-generated stimuli, particularly the FF-HQ dataset, could help in the development of GAN-stimuli that can be manipulated to generate a face, place, or object that is usually only accessible in a person's mind.

In conclusion, the current study stresses the potential utility of GAN-generated images to investigate how humans define categories in everyday visual stimuli. It will be worth evaluating different (and perhaps newer) GANs, training on larger datasets, recruiting more participants, and improving relevant experimental paradigms. Many questions in this line of work remain unanswered. What properties within these GAN-generated stimuli contribute to their perceived similarity? To what extent does the "type" (faces, landscapes, animals, etc.) of visual stimuli impact category boundary effects? And finally, what are the limits of computer-generated stimuli, and to what extent can we use them to evaluate fundamental aspects of human cognition?

REFERENCE

Bruner, J. S., Goodnow, J. J. & Austin, G. A. A Study of Thinking (Wiley, New York, 1956).

Fei-Fei, L. & Perona, P. A Bayesian hierarchical model for learning natural scene categories. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) 2, 524–531 (IEEE, 2005).

Feldman, N. H., & Griffiths, T. L. (2007). A Rational Account of the Perceptual Magnet Effect.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782. https://doi.org/10.1037/a0017196

Goetschalckx, L., Andonian, A., & Wagemans, J. (2021). Generative adversarial networks unlock new methods for cognitive science. *Trends in Cognitive Sciences*, *25*(9), 788-801. https://doi.org/10.1016/j.tics.2021.06.006

Goetschalckx, L., Andonian, A., Oliva, A., & Isola, P. (2019). GANalyze: Toward visual definitions of cognitive image properties. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv.2019.00584

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. Journal of the Acoustical Society of America, 100(2), 1111-1121.

Hampton, J. A., Estes, Z., & Simmons, C. L. (2005). Comparison and contrast in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1459–1476. https://doi.org/10.1037/0278-7393.31.6.1459

Hartendorp, M. O., Van der Stigchel, S., Burnett, H. G., Jellema, T., Eilers, P. H., & Postma, A. (2010). Categorical perception of morphed objects using a free-naming experiment. *Visual Cognition*, *18*(9), 1320–1347. https://doi.org/10.1080/13506285.2010.482774

Hénaff, O. J., Goris, R. L., & Simoncelli, E. P. (2019). Perceptual straightening of natural videos. *Nature Neuroscience*, *22*(6), 984–991. https://doi.org/10.1038/s41593-019-03774

Hull, C. L. Quantitative aspects of evolution of concepts: An experimental study. Vol. 28 (Psychological Review Company, 1920).

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? Journal of Experimental Psychology, 129(2), 220-241.

Inkawhich, N. (n.d.). *DCGAN tutorial*. DCGAN Tutorial - PyTorch Tutorials 1.11.0+cu102 documentation. Retrieved April 22, 2022, from https://pytorch.org/tutorials/beginner/dcgan_faces_tutorial.html

Peterson, J., Aghi, K., Suchow, J., Ku, A., & Griffiths, T. (2018). Sampling from object and scene representations using deep feature spaces. *Journal of Vision*, *18*(10), 403. https://doi.org/10.1167/18.10.403

Roads, B. D., & Love, B. C. (2021). Enriching ImageNet with human similarity judgments and psychological embeddings. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr46437.2021.00355

Rosch, E. H. Natural categories. Cogn. Psychol. 4, 328–350 (1973).

Rosch, E. & Mervis, C. B. Family resemblances: studies in the internal structure of categories. Cogn. Psychol. 7, 573–605 (1975).

Suchow, J. W., Peterson, J. C., & Griffiths, T. L. (2018). Learning a face space for experiments on human identity. https://doi.org/10.48550/arXiv.1805.07653

Torralba, A., Murphy, K. P., Freeman, W. T. & Rubin, M. A. Context-based Vision System for Place and Object Recognition (IEEE, 2003).