Visual Expertise Explains Image Inversion Effects

Martha Gahl
Computer Science and Engineering
University of California San Diego
La Jolla, CA 93093
mgahl@ucsd.edu

Shubham Kulkarni
Computer Science and Engineering
University of California San Diego
La Jolla, CA 93093
skulkarn@ucsd.edu

Nikhil Pathak
Computer Science and Engineering
University of California San Diego
La Jolla, CA 93093
npathak@ucsd.edu

Alex Russell
Computer Science and Engineering
University of California San Diego
La Jolla, CA 93093
awrussell@ucsd.edu

Garrison W. Cottrell
Computer Science and Engineering
University of California San Diego
La Jolla, CA 93093
gcottrell@ucsd.edu

Editors: Marco Fumero, Emanuele Rodolà, Clementine Domine, Francesco Locatello, Gintare Karolina Dziugaite, Mathilde Caron

Abstract

We present an anatomically-inspired neurocomputational model, including a foveated retina and the log-polar mapping from the visual field to the primary visual cortex, that recreates image inversion effects long seen in psychophysical studies. We show that visual expertise, the ability to discriminate between subordinate-level categories, changes the performance of the model on inverted images. We first explore face discrimination, which, in humans, relies on configural information. The log-polar transform disrupts configural information in an inverted image and leaves featural information relatively unaffected. We suggest this is responsible for the degradation of performance with inverted faces. We then recreate the effect with other subordinate-level category discriminators and show that the inversion effect arises as a result of visual expertise, where configural information becomes relevant as more identities are learned at the subordinate-level. Our model matches the classic result: faces suffer more from inversion than mono-oriented objects, which are more disrupted than non-mono-oriented objects when objects are only familiar at a basic-level, and simultaneously shows that expert-level discrimination of other subordinate-level categories respond similarly to inversion as face experts.

1 Introduction

Since 1969, researchers have been studying the effects of inverting images (Yin, 1969). Some researchers have focused on defining the bounds of inversion effects: what the measurable effect is

Proceedings of the I edition of the Workshop on Unifying Representations in Neural Models (UniReps 2023).

for what types of images (Farah et al., 1995; Yin, 1969; Jacques et al., 2007; Rezlescu et al., 2017). Others looked to explain how inversion effects arise: what part of the brain was active during inversion tasks or what level of experience a participant had with the stimuli in the experiment (Gauthier et al., 2000; Gauthier & Bukach, 2007; Gauthier et al., 2014; Kanwisher et al., 1997, 1998; Richler et al., 2011; Wang et al., 2014). We look to understand the signal processing in the visual system that leads to observed inversion effects to help inform more human-like computational vision systems.

In Yin (1969), participants studied a set of images (training phase), and then they were shown pairs of images and asked to select the image that was in the study set (testing phase). Trials with upright images and trials with inverted images were compared to determine the inversion effect. Using images of faces resulted in a strong and significant inversion effect; performance was much worse for inverted faces. Images of houses, a mono-oriented category, had a lesser, but still significant effect. Images of airplanes had an insignificant inversion effect. We draw two conclusions from this work: the effects of inversion on performance are greater when images of faces are used as the stimuli than other objects (Yin, 1969). The second conclusion is that not all objects produce the same inversion effects. Mono-oriented objects, which are objects that are typically seen in only one orientation such as the houses in Yin's 1969 work, do show an inversion effect, though it is smaller than that of faces (Yin, 1969). In our work we ask how these differing inversion effects could arise out of the same system, and how and why the signal processing differs depending on the stimulus observed.

Since Yin's 1969 paper on inversion, a great deal of research has focused on explaining inversion effects. Why is it that different stimuli - faces, objects, mono-oriented objects - produce different inversion effects? One explanation, supported by brain imaging and behavioral studies, is that visual expertise changes the way we process visual stimuli.

Visual expertise is defined with respect to Rosch's basic-level categories (Rosch et al., 1976). In a category hierarchy, the basic-level is the level at which objects are most commonly labeled, such as "chair", "tree", or "car". Basic-level categories define broad categories of objects that share properties such as general appearance, function, and common parts (Rosch et al., 1976). Visual expertise is defined as having proficiency in differentiating subordinate-level sub-classes of basic-level categories. For example, subordinates of the basic-level category "car" could include "Toyota Corolla", "Honda Civic", or "Hyundai Sonata". Most people are face experts in this sense. Identity is a subordinate-level judgment of "face" because faces share the same features (eyes, nose, mouth, ears, etc.) in the same general configuration.

Faces are processed holistically, which means it is not just the features of the face that are used for classification, but the *configuration* of the features. Instead of just using what the eyes or nose look like to recognize a person, we use information such as the distance between the eyes, or the distance from the nose to the mouth. When such stimuli are inverted, the configuration is disrupted, and we are left with only featural processing (Gauthier et al., 2000, 1999, 2003). The research into expertise suggests that experts in other domains, such as cars or birds, also use configural information when viewing basic-level categories in which the participants are experts (Gauthier et al., 2000).

We conduct experiments to ask if there is a way to characterize inversion effects, and thus the visual processing of the available configural information, in different stimuli in terms of levels of expertise. In doing so, we explore the changes in visual signal processing that occur between novice level and expert level. To do this, we build an anatomically inspired network that incorporates foveation (high resolution central vision and low resolution peripheral vision) and the log-polar mapping of the visual field on to the primary visual cortex which is seen in primate vision (Polimeni et al., 2006). In order to more accurately compare the signal processing in the visual system to the signal processing in a CNN, we incorporate approximations of the transformations to visual stimuli that occur in the brain into our model.

Using this model, we test the inversion effects of different types of stimuli across increasing expertise in order to gain an insight into how and why visual processing changes based on the visual stimulus. Our model is consistent with the view that expertise plays a significant role in the way we process visual inputs, and leads to the inversion effects seen in previous work. We show that visual expertise with a stimulus category and the level of discrimination of the task are responsible for the inversion effects seen.

2 Methods

2.1 Data Transformations

Cropping We augment our data by performing random cropping on the images. We take four random crops of each image, with each crop including approximately 65% of the pixels in the image. The crops cannot extend past the edge of the image, so none of the crops include any padding.

Rotation We randomly rotated our training images to be between -15° and 15°, because scenes may be viewed with some small amount of rotation from the tilting of the viewer's own head. To study the effects of inversion in the network, our validation images are shown at 0° and 180°.

Foveation We foveate each crop using the algorithm described in Jiang et al. (2015). The foveation leaves the center of the crop (the point of fixation) at a high resolution and transforms the periphery to a lower resolution. The further a pixel is from the center of fixation, the greater the degree of blurring. This mimics the foveation of the retina (Jiang et al., 2015). A visualization is provided in Figure 1.

Log-polar transform We further preprocessed our images to create an anatomically-realistic mapping of the visual field onto the visual cortex. Previous work has shown the validity of using the log-polar transformation as a 2D approximation of the mapping of the visual field onto the visual cortex in primates (Polimeni et al., 2006) and in computational visual models (Gahl et al., 2020; Remmelzwaal et al., 2020; Su & Wen, 2022). We incorporate the log-polar mapping in our network so as to more closely approximate the signal processing in human vision. For each case of rotation, we first rotated the image, then foveated the image, and finally took the log-polar transformation of the image. After the images have undergone a log-polar transformation, the changes in degrees of rotation appear as changes in vertical translation. This vertical shift causes pixels to "fall off" one edge of the primary visual cortex in the brain and wrap around to the opposite side. Instead of appearing as a simple translation up and down, inversion in images that have undergone a log-polar transformation results in a fundamental rearranging of image components. A visualization of the log-polar transform is provided in Figure 1. With the log-polar transform, cropping has a large effect as it changes the fixation point with each crop. To see this, note the difference between two fixations in Figure 1.

2.2 Data

To test the effects of expertise in visual processing, we use four different datasets. To model experts, the first three datasets are images of faces, cars, and dogs, generally mono-oriented objects, with targets at the subordinate-level. To model novices, who mainly know basic-level labels, the fourth dataset contains 128 categories with basic-level targets. Of the 128 categories, 124 are ImageNet categories that contain objects that are seen in a variety of orientations in natural scenes. The remaining four categories are mono-oriented at a basic-level instead of a subordinate-level: faces, cars, houses, and dogs. We randomly chose our sets with an 80/10/10 train/test/validation split.

Faces We curated a face dataset with 128 separate identities and approximately 200 example images per identity. The images for each identity portray the person in a variety of contexts, with differing backgrounds, lighting conditions, orientations, and facial expressions. Example images from the dataset are shown in Figure 2A.

Subordinate-level mono-oriented objects: Cars Cars are an appropriate choice for mono-oriented objects because they are almost exclusively seen upright in natural scenes. "Car" is also a basic-level category with a number of sub-classes. Car expertise is associated with discrimination at the level of model, e.g., 2010 Toyota Camry. The dataset used is the Comprehensive Cars dataset (Yang et al., 2015). These images include cars of a variety of makes, models, and years. The dataset contains images across 163 car makes and 1,716 car models. We select the 128 largest car model categories with approximately 200 example images per car model. The cars are of varying orientations, lighting conditions, and backgrounds. Examples of cars from the Comprehensive Cars dataset are shown in Figure 2B.

Subordinate-level mono-oriented objects: Dogs In the same way that models of cars are subordinates of the category "car", dog breeds are subordinates of the category "dog". We use 117 dog breeds included in ImageNet (Deng et al., 2009). Like all of ImageNet, these images are highly

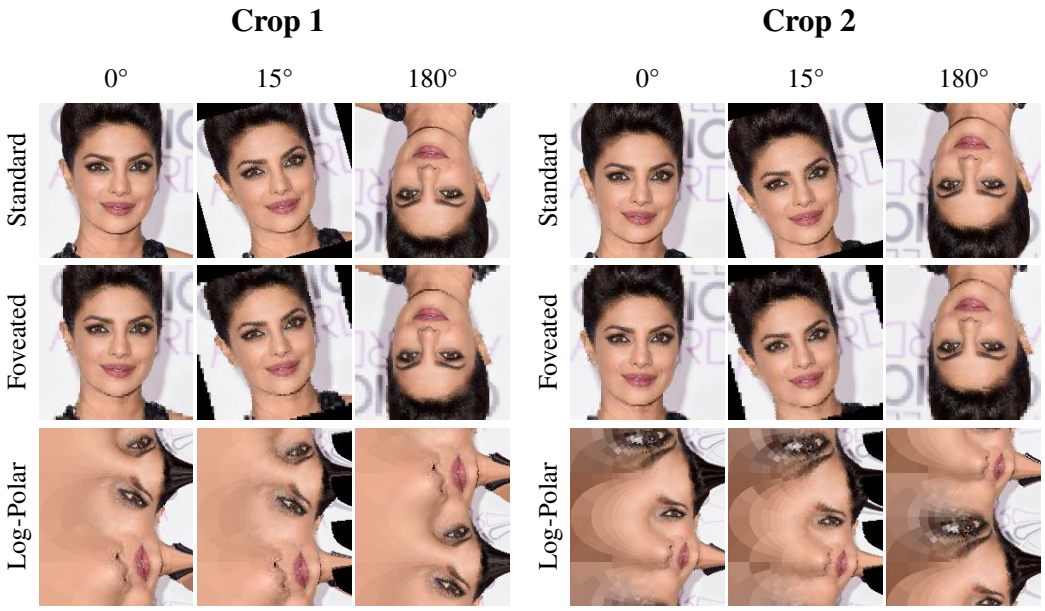

Figure 1: A visualization of the log-polar transformation for two crops of the same image. The top row shows standard images, or a depiction of the visual field. The bottom row shows images that have been transformed, using the log-polar mapping. For each crop, we show both image types at three amounts of rotation: 0°, 15°, and 180°.

variable in pose, context, and scale. Because of this we use all available images from ImageNet in these 117 categories. Examples of dogs in the dataset are shown in Figure 2C.

Basic-level categories We use a subset of 124 categories from ImageNet (Deng et al., 2009) which were chosen specifically because they are naturally viewed in multiple orientations, such as "ladle", "screwdriver", or "dumbbell" (Figure 2D). We avoided categories such as "clock" or "candle" which, although can be oriented multiple ways, are naturally seen in a limited number of orientations. The labels for the stimuli in this category are at the basic-level, instead of at the subordinate-level as in the previous three datasets (Rosch et al., 1976). Due to the within-category variance, we used all examples in ImageNet of these categories in order to achieve acceptable performance. We also included four categories (bringing the total number of categories to 128) that are mono-oriented. They are: "face", "car", "dog", and "house". Using these mono-oriented objects in our experiment with basic-level categorization allows us to compare how performance changes between experts and novices as general visual expertise increases. We aggregate data from each of the subordinates of "face", "car", and "dog" to get a varied sample of images for these categories. We matched the number of images approximately to the number of images the ImageNet categories included.

2.3 Model

We perform all experiments on two networks: a standard CNN and LPNet. LPNet is a CNN that is trained from scratch with foveated, log-polar transformed images. The standard CNN does not include foveation or the log-polar transform. We use ResNet-50 (He et al., 2016) as the architecture for both networks. Adding transformations to an existing network, instead of designing a new architecture, creates potential for the transformations to be added to any architecture and used for a variety of visual tasks. Similarly to widespread adoption of neural networks, inspired by the brain's architecture, the transformations of LPNet may lead to improvements in visual processing on unrelated tasks.

3 Experiment I: Faces and Objects

Based on Yin (1969), we first explore the difference in the effect of inversion based on whether a participant is viewing faces (objects of expertise) or objects for which the subject is a novice, or which are only known at the basic-level.

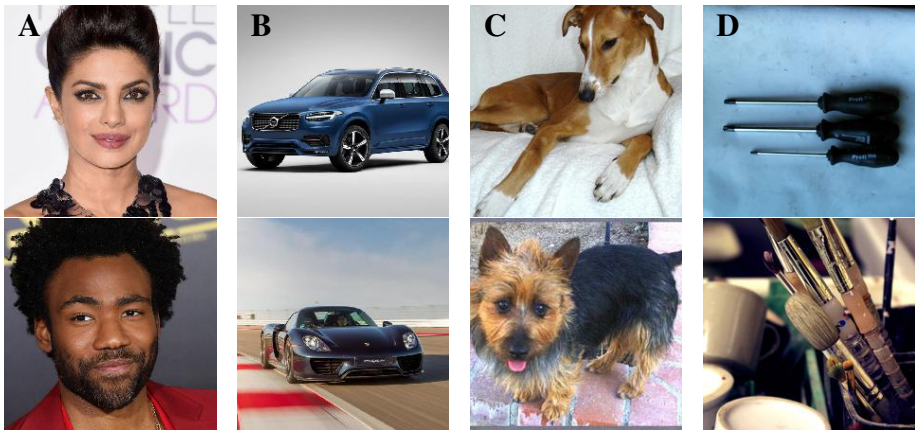

Figure 2: Example images used for recognition experiments: (A) a lab-gathered face dataset, (B) Comprehensive Cars dataset, (C) ImageNet (dog categories only), and (D) ImageNet (dog categories excluded).

3.1 Experimental Setup

When using subordinate-level visual stimuli, such as individual face identities, the network has to learn to discriminate between very similar stimuli. For example, faces share features and the same overall organization of features. Just as with humans, being able to discriminate very few subordinates demonstrates a low level of expertise with a particular basic-level category of visual stimuli. Being able to discriminate visually between many subordinates demonstrates high visual expertise with that category. Our network learned to perform categorization tasks using 4, 8, 16, 32, 64, or 128 categories of either faces or non mono-oriented objects.

Over training, we increase the number of classifications the network makes in order to mimic the increase of visual expertise over time. The network is trained for 40 epochs with 4 category outputs (identities for faces, object categories for the "novice" network). After 40 epochs, 4 new categories are added. There is a drop in accuracy because unseen categories are added, but during the next 40 epochs the network learns to discriminate the additional 4 categories. This continues for a total of 240 epochs, across 4, 8, 16, 32, 64, and 128 category outputs. Each phase of training is essentially performing pretraining of the network for the next phase of training by learning features that are helpful for discrimination.

During training, cropped images are rotated randomly between -15° and 15°. Testing images are presented in two conditions: upright (0°) and inverted (180°). The upright condition provides a baseline so that we can measure the relative amount the accuracy decreases after inversion. We report the effect of inversion as the percent of upright accuracy lost due to inversion. We calculate this as $\boldsymbol{Acc_{lost}} = (Acc_{up} - Acc_{inv})/Acc_{up}$. Using this measure, we can determine the effect inversion had on the network's recognition capabilities.

We include four configurations of networks and data: (1) CNN with faces (2) CNN with objects (3) LPNet with faces and (4) LPNet with objects. All experiments use the Adam optimizer and a mini batch size of 48.

3.2 Results

We ran all experiments five times and averaged the results. Figure 3 shows the accuracy throughout training. The first row is for standard CNNs and the second row is LPNet. The effect of the stimuli and the category level of the stimuli, either basic-level or subordinate-level, is clear. For the face stimuli, even in early phases of training, there is a performance gap between the upright validation accuracy and the inverted validation accuracy. As the network learns to differentiate more identities, this performance gap continues to increase, indicating an increasing inversion effect. When using objects as the training stimuli, the difference in performance between the two validation conditions is

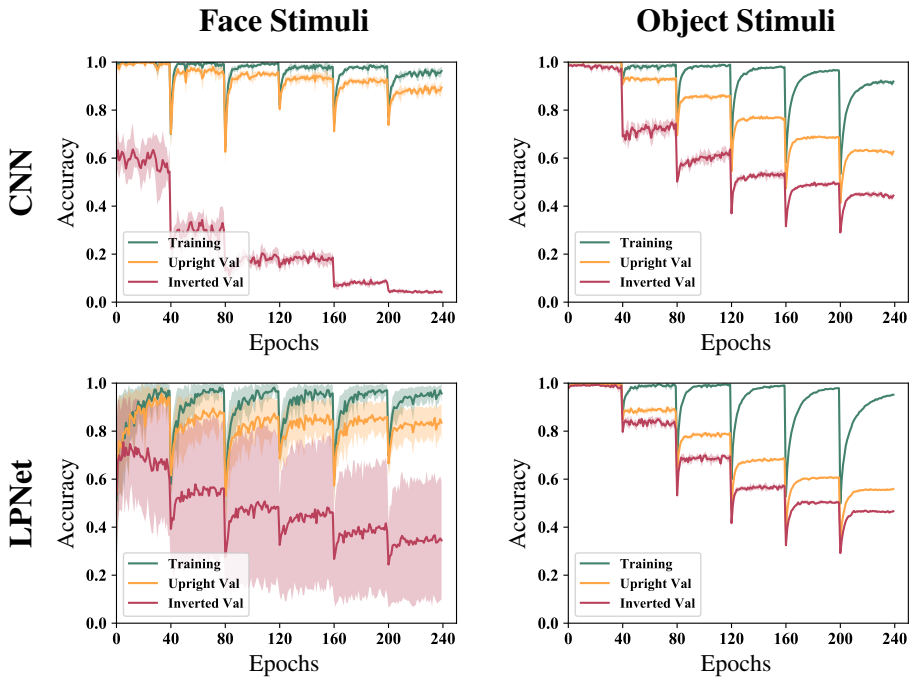

Figure 3: Accuracy throughout training on standard CNNs and LPNet for face and object stimuli. The green line is training accuracy, the orange line is accuracy on validation images at 0°, and the magenta line is accuracy on validation images rotated 180°. Shaded regions are +/-1 standard deviation.

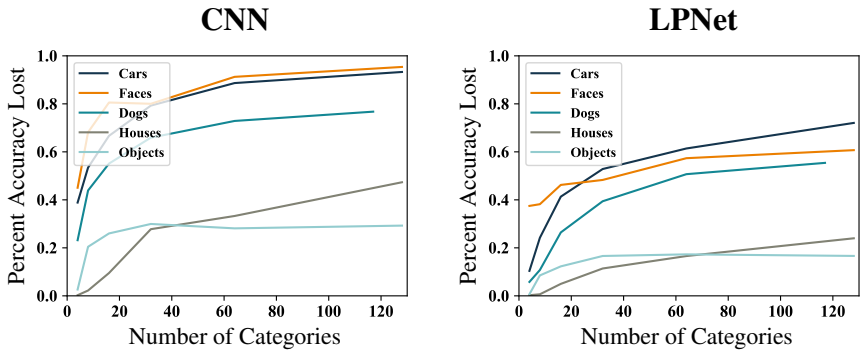

Figure 4: Percent accuracy lost, or the effect of image inversion, across all phases of training on standard CNNs and LPNet. Shown for faces, car models, and dog breeds (all expert networks), and houses and objects from the basic-level categorizer.

overall much smaller, with no apparent gap during the first phase of training. In addition, the size of the gap changes less throughout training.

The results on the CNN and LPNet have similar trends in that both show performance gaps between the two validation conditions which increase as the number of identities increases. This gap is larger for the CNN, in part because the inverted condition is more difficult for it. With inverted faces, the CNN fails nearly completely with an accuracy of approximately 4%. This is because the log-polar transform provides some rotation invariance that standard CNNs do not have. In the log-polar plane, changes in image rotation become vertical shifts as is seen in Figure 1. When using blur pooling, as our networks do, CNNs are shift invariant Zhang (2019). Therefore, LPNet is invariant to small amounts of rotation, similarly to human visual capabilities.

While these raw data give an insight to the effect of visual expertise on network performance, the effect of image inversion, particularly relative to other experiments, is less clear. Therefore, we plot the percent accuracy lost, which is accuracy change after inversion relative to the upright validation accuracy. Plotting the change relative to the upright validation accuracy allows us to directly measure

how much inverting an image will disrupt performance in any given experiment. We can also compare percent accuracy lost across experiments with different numbers of categories, which inherently have different difficulties, and across different datasets, for which differing complexity of images cannot be controlled. Figure 4 shows the percent accuracy lost first for a CNN and then for LPNet. A higher value indicates a larger percentage of the accuracy is lost due to inversion, or a larger inversion effect.

In Figure 4, both the CNN and LPNet models clearly show that image inversion significantly impacts face stimuli categorization and has a more limited impact on object stimuli categorization. This mirrors the first conclusion of Yin 1969. One of the classes we include in the object dataset is "house" in order to recreate Yin's testing of house stimuli. We separately plot the percent accuracy lost for only the "house" class in Figure 4. The percent accuracy lost for houses is much lower than that of faces, but slightly higher than objects. This mirrors the second conclusion of Yin 1969.

Both CNN and LPNet models show similar trends in percent accuracy lost. Each data type has an approximately logarithmic increase across increasing expertise, and each plateaus at a different level. We attribute these trends to increasing visual expertise. Figure 4 also shows how the percent accuracy lost overall on a standard CNN is higher than the corresponding values on LPNet. The curves are shifted up for the CNN because there is such a significant performance decrease on inverted images; the CNN is losing a larger proportion of the upright validation accuracy. The shift of the curves can be attributed to the log-polar transform, and the more human-like response of LPNet to inversion.

LPNet is rotation invariant, but the invariance has limits. When an image in the log-polar plane shifts vertically, some pixels wrap around to the opposite edge of the representation. This is because rotation is periodic. With small amounts of rotation, few pixels wrap around, the majority of features are undisturbed, and the configural information within the representation is mostly retained. This can be seen in Figure 1. This is why humans, and LPNet, are invariant to small amounts of rotation. Figure 1 also shows the log-polar representation for an inverted image. The more the image is rotated, the more the log-polar representation shifts, and the more pixels wrap around to the opposite edge of the representation. This is most severe with an inverted image. With an inverted image, the log-polar representation shows a rearrangement of features and a loss of configural information. It is this loss of configural information that causes a rotation invariant system to be unable to recognize inverted images at the same frequency as upright images. The CNN sees a significant performance decrease on inverted images, not because of loss of configural information (as the image does not undergo the same featural rearrangement that the log-polar representations do), but because it is not rotation invariant. This makes LPNet a more realistic model of human vision.

4 Experiment II: Expertise Effects with Car Models and Dog Breeds

In the previous experiment we showed that the inversion effect for faces was significant and the inversion effect for objects using a network with ability at the basic-level was not, recreating Yin (1969) results in a computational model. Here we explore whether other categories of stimuli with many subordinate-level labels, such as cars and dogs, experience the same inversion effect as that of faces. In Yin (1969), no other stimuli had comparable inversion effects to faces. Using increasing expertise, and an increasing reliance on configural information, we challenge this finding.

4.1 Experimental Setup

To study expertise in other subordinate-level categories, we trained two different sets of networks: one to distinguish the sub-classes of car models and one to distinguish the sub-classes of dog breeds. We again run experiments with both standard CNNs and LPNet.

Aside from the data chosen, this experiment follows the same setup and procedure as the previous experiment. We increase the number of identities being differentiated during each phase of training. For car models, our phases of training include 4, 8, 16, 32, 64, and 128 category outputs. The mean number of examples per car model is 151 images. For dogs, our phases of training include 4, 8, 16, 32, 64, and 117 category outputs. This is limited by the number of dog breeds available in ImageNet. These images inherently have more variation both between individuals and in pose (dogs sitting, standing, jumping, etc.) as they are not rigid objects like cars, so we used all available images from the dog categories of ImageNet, which averages to approximately 1500 images per dog breed.

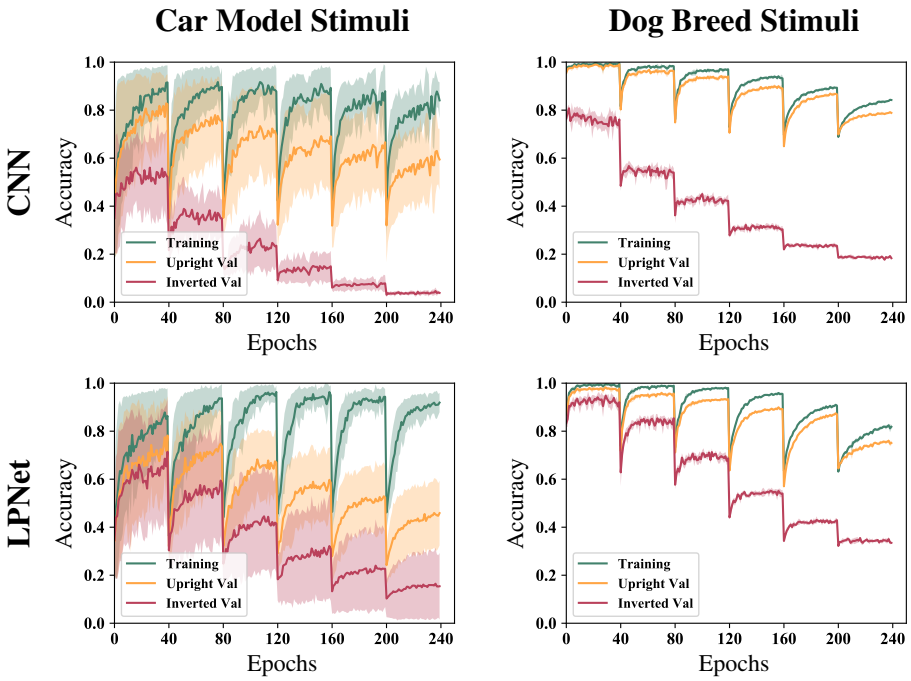

Figure 5: Accuracy throughout training on standard CNNs and LPNet for car model stimuli and dog breed stimuli.

4.2 Results

The results for these experiments are shown in Figure 5. The first row of plots shows results on a standard CNN and the second row shows results on LPNet. For all plots, the validation accuracy gap is smaller at the beginning of training, and increases as more categories are added. Like the previous experiment, it is much more difficult for the CNN to distinguish inverted images than it is for LPNet, because of the rotation invariance provided by LPNet.

Comparing LPNet results from both experiments, the car and dog upright validation accuracies (Figure 5) and the face upright validation accuracies (Figure 3) vary widely. In addition, the net number of percentage points lost because of inversion in these experiments varies. This is because of differences in data that are hard to control for, like more complicated backgrounds for cars and dogs, similar contexts for faces, and the amount of available training data. Despite these differences, when looking at the percent accuracy lost in Figure 4, it is clear that each of the expert LPNets experienced a very similar inversion effect. This is because percent accuracy lost measures how much of the upright validation accuracy was lost due to inversion, not the net number of percentage points lost.

Discriminating between different subordinate categories may become harder or easier depending on the stimulus class itself or the context of the images. However, the percent accuracy lost shows that image inversion affects LPNet trained as an expert on car models or dog breeds in almost the same way it affects LPNet trained on faces.

We see that LPNet trained on faces and LPNet trained on cars have the most variance across trials of all experiments. Some of this can be attributed to variation in the data itself, although the variance is much smaller for the corresponding CNN experiments. The face experiment and the car experiment both use an average number of examples per category between 150 and 200. The object experiment and the dog experiment however each average more than 1,000 images per category. This suggests that LPNet requires more examples during training for generalization.

4.3 Comparison with Human Data

We present our results alongside results from Yin's 1969 paper in Figure 6 Yin (1969). In Yin's experiments, human participants are shown 40 novel faces, and asked to study them. They are then shown 24 pairs of images, and asked which of the two images in the pair appeared in the study set.

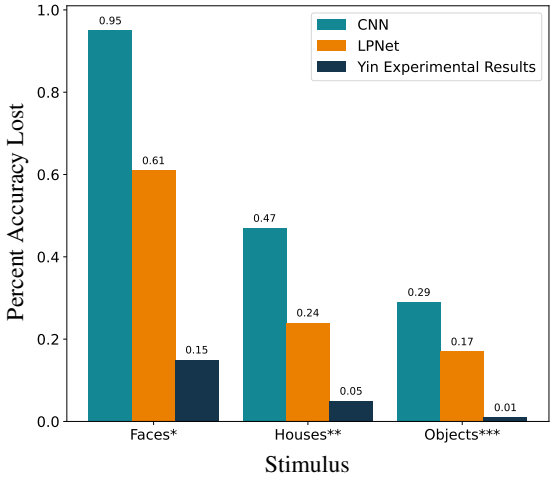

Figure 6: Percent accuracy lost for CNN, LPNet, and Yin's experimental results Yin (1969). We perform t-tests comparing CNN and LPNet values for each stimulus type. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Yin reports mean error for each stimulus category and each condition: upright or inverted Yin (1969). We calculate percent accuracy lost for Yin's human results by subtracting the mean error from the 24 test images to get a mean accuracy. We then use the mean accuracy for the upright and inverted conditions in our formula for percent accuracy lost.

In Figure 6, all three sets of results, LPNet, CNN, and human Yin (1969), show the same trend: percent accuracy lost for experts in subordinate-level categories is the highest, followed by mono-oriented objects at a basic level, and finally objects at a basic level (airplanes in Yin (1969)). There is a large difference in the magnitudes of the percent accuracy lost, which we attribute to differences in task. For the human task, participants are making binary decisions about which image in a pair was in a study set Yin (1969). Chance for this experiment is 50%. The computational models are making 128-way classifications, where chance is less than 1%. Even taking task differences into account, the CNN shows a significantly larger image inversion effect than LPNet as indicated in Figure 6. LPNet more closely matches human data and is therefore more biologically realistic.

5 Conclusion

We explored image inversion effects and the impact of expertise on visual signal processing. In experiment I, by using images of faces and objects, we showed that LPNet, a convolutional neural network that includes a foveated retina and the log-polar mapping from the visual field to the primary visual cortex, can reproduce experimental results of image inversion. We showed that there is a larger effect on performance from inverting images of faces than images of objects. We then explored the result that images of mono-oriented objects have an inversion effect between that of faces and objects by showing performance of a house-novice network.

In experiment II, we demonstrated that expert networks on mono-oriented objects, discriminating subordinate-level categories, do not show an effect between that of faces and objects as suggested in Yin (1969). Rather, we showed that LPNet trained to distinguish car models or dog breeds showed similar inversion effects as faces. This suggests that inversion effects are not dependent on the stimulus, but rather on the level of visual expertise with the stimulus and on the categorization level of the stimulus (expert or novice viewing subordinate-level or basic-level categories).

People are face experts, and we process face stimuli holistically, using the configural information of the face (Gauthier et al., 2000, 1999, 2003). This holistic processing is disrupted during inversion, which produces the much studied face-inversion effect. By recreating the same inversion effects with expert networks discriminating subordinate-level categories of "dogs" and "cars", we show that this inversion effect is dependent not on the stimulus being a face, but on the level of expertise the network has with a stimulus class.

Image inversion effects are dependent on the level of expertise because of an expert's reliance on configural information to perform fine-grain discrimination. LPNet is a more realistic model of human vision because of the log-polar transform, which provides rotation invariance. This is seen with inverted images, when CNNs nearly completely fail to perform differentiation while LPNet is still able to. We have shown that, with the log-polar transform, the inversion of an image causes a rearrangement of features. We suggest this disrupts the configural information used for discrimination.

When LPNet is not an expert, the loss of configural information has a minimal impact on discrimination ability, and it is able to perform the task with only minor performance changes. We believe this is explained by novices relying on featural information for recognition. As LPNet does more fine-grain discrimination between categories, it relies more heavily on the configural information held within the image.

This suggests that expertise with visual stimuli changes the signal processing occurring and that, for any network performing a visual task, the expertise and discrimination level will affect the information the network uses and its ultimate performance. Our results support the hypothesis that the source of the inversion effect seen with increasing visual expertise, including face expertise, is disruption of configural information. This work is limited by our ability to interpret LPNet. Based on biological results, we suggest it uses configural data, but direct evidence is unavailable.

References

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

Farah, M. J., Tanaka, J. W., and Drain, H. M. What causes the face inversion effect? *Journal of Experimental Psychology: Human Perception and Performance*, 21:628–634, 1995.

Gahl, M., Yuan, M., Sugumar, A., and Cottrell, G. The face inversion effect and the anatomical mapping from the visual field to the primary visual cortex. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 2020.

Gauthier, I. and Bukach, C. Should we reject the expertise hypothesis? *Cognition*, 103:322–330, 2007.

Gauthier, I., Tarr, M., Anderson, A., Skudlarski, P., and Gore, J. Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature neuroscience*, 2:568–73, 07 1999.

Gauthier, I., Gore, P. S. J. C., and Anderson, A. W. Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3:191–197, 2000.

Gauthier, I., Curran, T., Curby, K. M., and Collins, D. Perceptual interference supports a non-modular account of face processing. *Nature Neuroscience*, 6:428–432, 2003.

Gauthier, I., McGugin, R. W., Richler, J. J., Herzmann, G., Speegle, M., and Gulick, A. E. V. Experience moderates overlap between object and facerecognition, suggesting a common ability. *Journal of Vision*, 14:1–12, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Jacques, C., d'Arripe, O., and Rossion, B. The time course of the inversion effect during individual face discrimination. *Journal of Vision*, 7, 6 2007.

Jiang, M., Huang, S., Duan, J., and Zhao, Q. Salicon: Saliency in context. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1072–1080, 2015.

Kanwisher, N., McDermott, J., and Chun, M. M. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302–4311, 1997.

Kanwisher, N., Tong, F., and Nakayama, K. The effect of face inversion on the human fusiform face area. *Cognition*, 68, 1998.

Polimeni, J., Balasubramanian, M., and Schwartz, E. Multi-area visuotopic map complexes in macaque striate and extra-striate cortex. *Vision Research*, 2006.

Remmelzwaal, L. A., Mishra, A. K., and Ellis, G. F. R. Human eye inspired log-polar pre-processing for neural networks. *2020 International SAUPEC/RobMech/PRASA Conference*, pp. 1–6, 2020.

Rezlescu, C., Susilo, T., Wilmer, J. B., and Caramazza, A. The inversion, part-whole, and composite effects reflect distinct perceptual mechanisms with varied relationships to face recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 43:1961–1973, 2017.

Richler, J. J., Cheung, O. S., and Gauthier, I. Holistic processing predicts face recognition. *Psychological Science*, 22:464–471, 2011.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, 1976.

Su, B. and Wen, J.-R. Log-polar space convolution layers. *Proceedings of the 36th Conference on Neural Information Processing Systems*, 2022.

Wang, P., Gauthier, I., and Cottrell, G. Experience matters: Modeling the relationship between face and object recognition. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 2014.

Yang, L., Luo, P., Loy, C. C., and Tang, X. A large-scale car dataset for fine-grained categorization and verification. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3973–3981, 2015.

Yin, R. K. Looking at upside-down faces. *Journal of Experimental Psychology*, 81:141–145, 1969.

Zhang, R. Making convolutional networks shift-invariant again. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.