# Modeling the Viewpoint Theory of Mental Rotation

**Yeon Tae Chung** YEONTAECHUNG@GATECH.EDU
**Vijay Marupudi** VIJAYMARUPUDI@GATECH.EDU
**Siddhartha Vemuri** SVEMURI8@GATECH.EDU
**Athanassios Economou** THANOS@GATECH.EDU
**Sashank Varma** VARMA@GATECH.EDU
Georgia Institute of Technology, Atlanta, GA 30332 USA

## Abstract

Multiple theories have been proposed to explain mental rotation. The analog theory was inspired by the seminal experiment that introduced the Shepard–Metzler (SM) task, which found a linear trend between the angle between two objects and the time to judge them as the same or different (Shepard & Metzler, 1971). This finding was taken as evidence of objects being mentally represented and spatially transformed (e.g., rotated). Subsequent theories have questioned whether mental rotation is such a kinematic process. The viewpoint theory, in particular, explains how response time (RT) is linearly related to angular disparity in terms of the *similarity* between viewpoints. Here, we evaluated two models that instantiate this theory in an experiment on the SM task and a variant of it. One model is an algorithm from image processing and operates at the raw level of images, computing the normalized mean-shifted cross-correlation (NMSCC). The other model is based on a computational architecture adapted from computer vision and operates at the structured level of vector spaces: the cosine similarity of latent vectors in AlexNet, a convolutional neural network (CNN). Neither model demonstrated the expected performance profile consistent with the human data. In response, we introduced a process account as an alternative way to interpret the results of the experiments under the analog theory.

## 1. Introduction

People can infer two objects are the same by the congruence of their shapes. While congruence can be determined by simply moving objects into alignment, it is unclear how people can do this with objects in images without physically interacting with them. In the 1970s, Shepard and his collaborators famously proposed that people do so by spatially transforming mental images of objects in their mind (Shepard & Cooper, 1982). They describe mental rotation as one such mental transformation in which mental images of objects are rotated in the mind. To evaluate this proposal, they designed what has come to be called the Shepard–Metzler (SM) task, which asks people to judge whether objects like those depicted in Figure 1 are the same or different without physical interaction. A striking finding was that response time (RT) increased linearly with the angular disparity between the orientations of compared objects (Shepard & Metzler, 1971). They interpreted this as evidence of mental rotation and mental imagery more generally.

Given that cognitive processing in mental transformation seems to follow motor principles of motion, they proposed an analog theory of mental imagery (Shepard, 1978). This theory posits

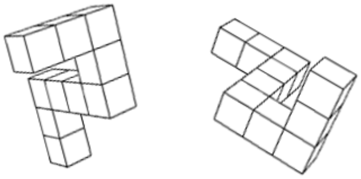

*Figure 1*. Sample pair of congruent Shepard–Metzler (SM) figures used in the original SM task.

kinematics of the mind as the mental counterpart to equivalent laws of physics (Shepard, 1994). Mental rotation, then, is a continuous process of rotating objects in the mind akin to manually reorienting physical objects (Shepard & Metzler, 1971). A logical next step is to instantiate this theory as a computational model to test whether the claims it makes about representation and process are sufficient for reproducing the human data. There have been many such models over the years (Funt, 1983; Hamrick & Griffiths, 2014). One goal of this study is to look towards existing computational methods from image processing and computer vision for new sources.

The analog theory was almost immediately challenged in the context of mental imagery. Famously, Pylyshyn's (1973) symbolic theory of mental imagery proposes mental images are but propositional representations capturing abstract associations between the constituent parts of objects. Mental rotation is, then, merely logical inference over such representations. In short, this theory proposes mental imagery is an epiphenomenon. This debate on the phenomenology of representation in the mind between proponents of imagistic cognition, often led by Kosslyn and his collaborators (Kosslyn & Pomerantz, 1977; Kosslyn et al., 2002), and proponents of symbolic cognition, often Pylyshyn standing alone (Pylyshyn, 1973; 2002), raged for decades. It was finally resolved with the maturation of functional neuroimaging, which brought the ability to directly observe which areas of the brain are active during mental imagery. Early studies showed that these were areas associated with vision and visuospatial reasoning, not those associated with language processing or logical reasoning (Kosslyn et al., 1995). Subsequent studies found evidence of a linear function: activation in the intraparietal sulcus (IPS), a part of the visual system associated with spatial transformations, increases with the angular disparity between orientations of compared objects (Just et al., 2001; Kosslyn et al., 2001). With these findings, cognitive scientists largely side with the analog theory.

Still, subsequent theories have been proposed that also find support. Here, we consider the viewpoint theory, which shares with the analog theory the assumption that people can form mental images (Edelman, 1995). It differs from the analog theory in proposing that similarity, not rotation, is the mechanism for comparison. Specifically, it claims that for objects at different orientations, the greater the angular disparity between them, the less similar their viewpoints are, and thus the longer it takes to judge them as the same or different. Put succinctly, mental imagery is real, but mental rotation is an epiphenomenon according to this theory. We instantiated the theory in two ways by adapting a method of computation from image processing and another from computer vision: the normalized mean-shifted cross-correlation (NMSCC) (Bradski, 2000) and AlexNet (Krizhevsky et al., 2012), respectively. We then evaluated our two models as cognitive science theories in an experiment on the SM task and a variation of it focusing on rotational symmetry. Finally, we framed the experimental results of our two models in the context of both the viewpoint and analog theories of mental rotation.

## 2. Literature Review

### 2.1 The Shepard–Metzler Task

It takes people time proportional to the angular disparity between the orientations of two objects, like those depicted in Figure 1, to determine whether they are the same or different (Shepard & Metzler, 1971). In the SM task, this angular disparity is the primary independent variable and the time taken to make this judgement (i.e., RT) is the primary dependent variable. Objects are either congruent or mirrored; this is the other independent variable. Accuracy of judgements is another dependent variable. Because people are typically highly accurate on this task, researchers generally focus on RT.

The angular disparity between orientations of objects is defined as the shortest angle between them. This is important because people somehow know the optimal direction of rotation. For example, for a 2D shape in the picture plane, people have the choice of rotating it clockwise or counterclockwise, and they choose to take the shortest path. Additionally, knowing the shortest path of rotation beforehand only minimally affects people's performance (Cooper & Shepard, 1973; Shepard & Metzler, 1971). These are major points of curiosity that we return to in the General Discussion.

There are also several other aspects of the SM task that are important to consider. One is that even when the objects are mirrored, RT is still a linear function of the angular disparity—albeit with slower RTs and at a slower rate of RT over angular disparity (Parsons, 1987). This is a puzzling finding because the angular disparity between orientations of mirrored objects is technically ill-defined (Shepard & Metzler, 1971). Different theories of mental rotation explain this finding in different ways.

Another aspect to consider is that the objects need to be chiral (i.e., have "handedness") so that a mirrored object can be distinguished from the object it mirrors. Chirality is necessary to ensure that geometric transformations that account for chirality like rotation are required to determine congruence. Otherwise, people can shortcut mental rotation and purely rely on discriminative features of objects to make their judgements.

Lastly, the SM task originally involved showing two objects simultaneously so that they do not overlap. Subsequent studies explored sequential presentation of objects centered at the same point (Cooper & Shepard, 1973; Metzler & Shepard, 1974). The same linear trend between angular disparity and RT was found in this case. This detail is important because both viewpoint models can be understood to be performing this sequential version of the task, processing images of objects separately before comparing them. It is of no consequence for the purposes of replication that the experiments of the current investigation implement the sequential version.

### 2.2 The Viewpoint Theory

Like the analog theory, the viewpoint theory of mental rotation is an imagistic account of mental imagery that proposes people can form mental images of objects. It differs, however, in proposing that mental rotation is *not* a spatial transformation but rather arises from the similarity between the viewpoints of objects from which they are viewed (Edelman, 1995).

Supporting the viewpoint theory, Edelman and Bülthoff (1992) found that when participants are shown a limited set of viewpoints of unfamiliar 3D objects and then tested on novel viewpoints of those objects, both RTs and error rates increased with angular disparity from the orientations of known viewpoints. They speculated that the linear trend, originally taken as evidence for a spatial process, is mediated by similarity: as angular disparity increases, similarity decreases, requiring more processing time and increasing error rates. Their explanation of mental rotation is taken as the viewpoint theory. For this theory to be a viable account of mental rotation, there must be a similarity measure with scores that linearly decrease with increasing angular disparity. To fill this role, many similarity measures have been tested with varying degrees of success (Edelman & Weinshall, 1991; Niall, 2020; 2023; Stewart et al., 2022). We extend this prior work by testing both a simple algorithm from image processing, NMSCC (Bradski, 2000), and a computational architecture from computer vision, the convolutional neural network (CNN) AlexNet (Krizhevsky et al., 2012).

## 3. Research Questions

We investigated four research questions concerning the viewpoint theory by forming hypotheses to test in two computational experiments.

First is the question of whether the viewpoint theory can successfully account for the major finding in mental rotation: the linear trend between angular disparity and RT in both the congruent (Shepard & Metzler, 1971) and mirrored (Parsons, 1987) cases of the SM task. In both cases, we hypothesized the similarity between object viewpoints will decrease linearly with increasing angular disparity between the orientations of the depicted objects.

Second is the question of whether mirrored objects can be differentiated from congruent ones via similarity alone. RT is found to be faster in the congruent than the mirrored cases (Parsons, 1987). Thus, our second hypothesis was that similarity in the mirrored case will be lower than in the congruent case for the same angular disparity.

The first two hypotheses can be summarized in the predicted results shown in Figure 2. In accordance with the first hypothesis, similarity score changes linearly with angular disparity in both the congruent case and the mirrored case. The two linear trends in the intervals 0–180° and 180–360° reflect the surprisingly optimal choice of rotation direction by people during mental rotation because angular disparity is a maximum of 180°. Similarity scores of mirror-image comparisons are also always lower than those of congruent comparisons in accordance with the second hypothesis.

Third is the question of whether the linear trend is sensitive to the rotational symmetry of the viewpoint. It is quantified as the number of orientations across 0–360° where the viewpoint is the same. For example, viewpoints with onefold rotational symmetry have a single orientation of self-alignment at a rotation angle of 0°. This is the case for the SM figures shown in Figure 1. The plots of Figure 2 reflect our predictions for those SM figures in that there is one oscillation in similarity score (i.e., a change in slope at 180°). We hypothesized that there will be *n* number of these oscillations, corresponding to the *n*-fold rotational symmetry of the viewpoint.

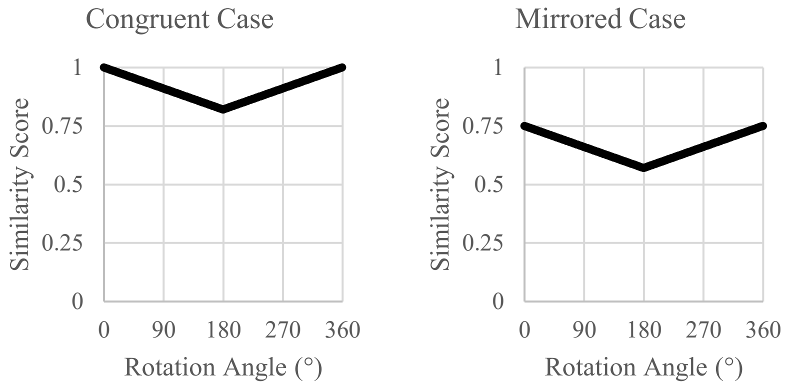

*Figure 2*. Idealized replication of the original SM task supporting the first two hypotheses.

The fourth is the question of whether simple image-based similarity measures from image processing sufficiently explain human performance, or whether the more complex computational architectures of computer vision models, which extract structured information (e.g., vector spaces of neural networks), are necessary. We hypothesized that the AlexNet model, drawing on more information, will perform better than the NMSCC model.

## 4. Models

### 4.1 NMSCC

The simple similarity measure we chose from image processing is NMSCC. Specifically, we used the implementation of the OpenCV library (Bradski, 2000). In this implementation, the similarity between two images of the same size is a scalar value in the interval [−1, 1], with 1 indicating maximal similarity. This is because the measure reduces to the Pearson correlation coefficient when comparing images of the same size. The statistical measure can be trivially reused as a similarity measure in the context of image processing because digital images can be represented as a collection of color channels (e.g., red, green, and blue). Pairings of color channels suitable for correlation analysis can be made as each position of a pixel in one image can correspond to the same position in another of the same size. In this context, the coefficient can be used to measure how color intensities of one image correspond to those at the same positions in another. In this way, a similarity score can be computed. We use this similarity measure directly as the NMSCC model.

### 4.2 AlexNet

The structured similarity measure we used was derived from the AlexNet computational architecture (Krizhevsky et al., 2012) from computer vision. As a CNN, it has multiple connected layers of units. The weights of these connections support the computation of informationally rich internal representations of input images. We used these representations in our model to compute a structured similarity measure.

AlexNet was trained on images from the ImageNet dataset (Deng et al., 2009). This dataset consists of more than 14 million images, each associated with one of 1000 classification labels (e.g., "goldfish"). We use the pretrained version of AlexNet provided by the TorchVision library (Paszke et al., 2019). This version closely follows the original design outlined by Krizhevsky et al. (2012).

The first layer of AlexNet is the input layer receiving the color intensities of every pixel in an image input. The last layer is the output layer that outputs the probability of membership in each of the 1000 categories of ImageNet. The layers in between are where internal representations of input images are computed. We chose the first of the final three "fully connected" layers of the CNN to extract internal representations of images. We made this choice because these layers compute a holistic representation of input images.

The activations of the units in the first fully connected layer form a vector. We use the cosine similarity of two of these vectors, one for each compared image, as a similarity measure. This measure computes the cosine of the angle between two vectors, the result of which is in the same interval $[-1, 1]$ as the NMSCC model with the same interpretation of 1 indicating maximal similarity.

## 5. Experiment 1

In this experiment, we tested whether the viewpoint theory can account for the original finding of mental rotation by Shepard and Metzler (1971) in the congruent case and by Parsons (1987) in the mirrored case. Rather than measuring RT like these studies, we examined whether similarity score follows a linear trend with angular disparity. With respect to our first hypothesis, we expected the linear trend to occur in two intervals: 0–180° with a negative slope and 180–360° with a positive slope. Thus, we expected the overall function of a complete rotation to be bilinear in the interval 0–360°. With respect to our second hypothesis, we expected similarity to be lower for mirrored objects. Thus, we expected the model findings to approximate the idealized predictions of Figure 2. With respect to our fourth hypothesis, we expected the structured representations of AlexNet to perform better than the statistical measure of NMSCC.

### 5.1 Method

#### *5.1.1 Stimuli*

We used viewpoints of the same object shown in the two SM figures of Figure 1. The object is drawn in perspective and oriented in the picture plane such that all 10 cubes are unambiguously visible. Instead of using the same line drawings as shown in Figure 1, however, we used shaded renderings of the object, as shown in Figure 3. They contain more visual cues of 3D structure with shading. This is especially relevant for AlexNet due to it being trained on images from ImageNet, which are primarily of the physical world with light and shadows.

Notably, adjacent cubes are not flush with one another in the shaded rendering to maintain the quality of individual cubes being visible. This allows the shadows cast by each cube to outline its

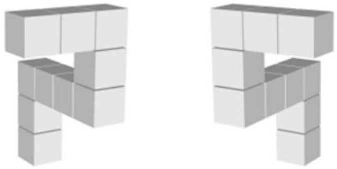

*Figure 3*. A shaded rendering of the same SM figure, used in the current study, and its mirror image reflected across the vertical axis.

own boundary. Additionally, to control the distribution of shading, the direction of lighting is not changed relative to the orientations of the object.

### *5.1.2 Procedure*

To generate the stimuli for the congruent case, the left viewpoint in Figure 3 was taken as the base image of comparison. It was rotated clockwise in the picture plane by increments of 1° in the interval 0–360°. Images from each rotation were paired with the base image, which represented the 0° rotation, as stimuli for the congruent case. The same process was used to generate the mirrored stimuli, only with the rotated images being paired with the right viewpoint in Figure 3.

Rotation in the picture plane was chosen because 2D shapes in viewpoints remain unchanged across rotations in this plane. When images are rotated in depth, parts of images transformed further away will be scaled down in resolution and be less visible. We controlled depth to rule out this confound.

Under the viewpoint theory, we expected to replicate the results of the original experiment by Shepard and Metzler (1971) via a linear trend between similarity and angular disparity. As shown in Figure 2, this means the models should produce a bilinear function in the interval 0–360°. We therefore fit a bilinear function with the hinge point at 180° to both models.

## 5.2 Results

### *5.2.1 Accounting for Shepard and Metzler's (1971) Results*

The first research question asks whether either viewpoint model demonstrates the linear trend between angular disparity and RT found by Shepard and Metzler (1971). We hypothesized that they do by demonstrating a linear trend with similarity.

For congruent stimuli, the NMSCC model's results showed a moderate linear trend between similarity and angular disparity. This is indicated by an $R^2$ of .50 when fitting a bilinear function; see Table 1 for the relevant statistics and the left plot of Figure 4 for the similarity function itself. The sign of the estimated slope was negative in the interval 0–180°, −.47, and positive in the interval 180–360°, .47, matching the bilinear function observed in human data (Cooper & Shepard, 1973) under the assumption similarity is inversely related to RT. Thus, the NMSCC model's replication of Shepard and Metzler's (1971) results was partial.

In contrast, the AlexNet model's results in the congruent case show a weaker bilinear fit. The

*Table 1*. Statistics of similarity score as a function of rotation angle for the two models and the two cases of stimuli in the SM task.

|  | Stimuli | Bilinear Fit ($R^2$) | Estimated Slopes |  | Similarity Scores |  |
|---|---|---|---|---|---|---|
|  |  |  | 0–180° | 180–360° | Low | High |
| NMSCC | Congruent | .50 | −.47 | .47 | .56 | 1 |
|  | Mirrored | .44 | −.12 | .21 | .50 | .77 |
| AlexNet | Congruent | .06 | −.13 | .12 | .66 | 1 |
|  | Mirrored | .03 | −.1 | .09 | .67 | 1 |

$R^2$ is only .06; see Table 1. Still, the estimated slopes, −.13 in the interval 0–180° and .12 in the interval 180–360°, also match the empirical finding with humans. Statistics aside, it is rather difficult to see a bilinear function here; compare the AlexNet model's similarity function in the left plot of Figure 5 with the idealized results in the left plot of Figure 2.

With respect to the fourth hypothesis in the case of congruent stimuli, the NMSCC model better captured the bilinear function observed in human data (Cooper & Shepard, 1973) than the AlexNet model.

### *5.2.2 Accounting for Parsons' (1987) Results*

Our second hypothesis is that similarity scores for mirrored objects will always be lower than the similarity scores for congruent objects as this was observed by Parsons (1987) for human RT data; see Figure 2. Put simply, we predicted that the intervals of similarity scores will not overlap.

For the NMSCC model, the range of similarity scores for the congruent stimuli and the mirrored stimuli overlapped significantly. Table 1 provides the relevant statistics: the interval is [.56, 1] in the congruent case and [.5, .77] in the mirrored case. This is also observable across the left and right plots of Figure 4. This is inconsistent with our second hypothesis, and it is problematic because people are slower in mirror-image comparisons than congruent comparisons

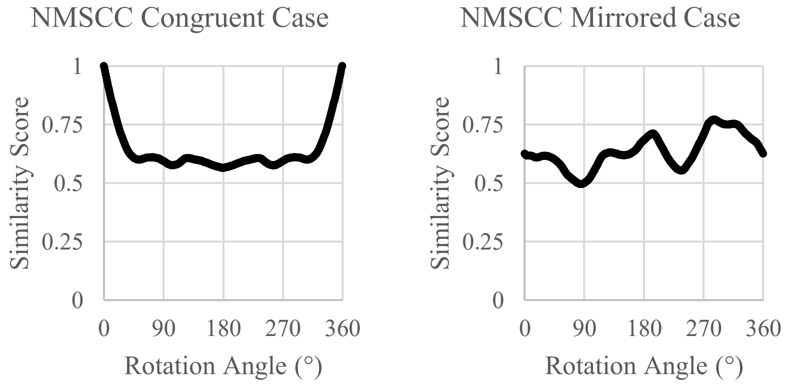

*Figure 4*. Similarity score as a function of rotation angle produced by the NMSCC model using congruent stimuli, on the left, and mirrored stimuli, on the right.

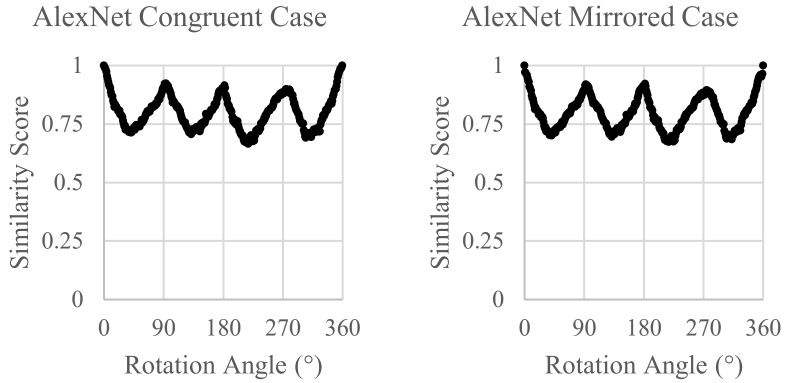

*Figure 5*. Similarity score as a function of rotation angle produced by the AlexNet model using congruent stimuli, on the left, and mirrored stimuli, on the right.

(Parsons, 1987). The NMSCC model does not match this aspect of human performance.

The AlexNet model produced the same pattern of results: the range of similarity scores for congruent stimuli and mirror-image stimuli overlapped significantly. It produced similarity scores in the interval [.66, 1] and [.67, 1] for congruent stimuli and mirrored stimuli, respectively; see Table 1 and Figure 5.

With respect to the fourth hypothesis, neither model produced lower similarity scores for the mirrored stimuli. Consequently, there is no basis to choose between them as viewpoint models.

### 5.3 Discussion

The results of Experiment 1 showed that neither instantiation of the viewpoint theory is sufficient as a viewpoint model of mental rotation.

The NMSCC model shows a moderate bilinear trend between similarity score and rotation angle, aligning with the Shepard and Metzler (1971) results. However, the intervals of similarity scores it produced overlapped significantly across congruent stimuli and mirrored stimuli, which is inconsistent with the finding that people are slower to compare congruent objects versus mirrored objects (Parsons, 1978).

Surprisingly, the AlexNet model is even less aligned with the human data. There is no evidence of a bilinear trend between similarity score and rotation angle, while there is a significant overlap in the intervals of its produced similarity scores between congruent stimuli and mirrored stimuli. This latter finding may be explained by the training procedure of AlexNet. Recall it was a pretrained model from TorchVision, trained on ImageNet. To improve the performance of the model for object classification, the dataset was augmented by adding mirror images of existing images (Krizhevsky et al., 2012). This augmentation has the positive effect for image classification of improving model generalization across mirror images. A negative consequence may be its inability to discern mirrored stimuli from congruent stimuli.

Although neither model fully aligned with human data, the NMSCC model's results are more consistent with the first and second hypotheses than AlexNet model's results. Thus, the NMSCC model may be a more psychologically plausible implementation of the viewpoint theory than the AlexNet model. This is contrary to our fourth hypothesis.

Interestingly, the AlexNet model produced similarity scores that largely oscillated in 4 cycles with rotation angle for both congruent and mirrored stimuli; see Figure 5. This could have been due to the model extracting a representative shape with fourfold rotational symmetry from the viewpoints of Figure 3 and their rotations. For example, the many quadrilaterals in SM figures are roughly such shapes. We followed up on this finding in Experiment 2.

## 6. Experiment 2

In this experiment, we used simple 2D shapes with varying folds of rotational symmetry to test the sensitivity of the viewpoint models to rotational symmetry. This was motivated by the results of Experiment 1 and is relevant to the third research question that asks whether *n*-fold rotational symmetry of figures corresponds to the *n* cycles of similarity score oscillations. Given that SM figures are comprised of quadrilaterals, is the periodicity observed for the AlexNet model in Experiment 1 due to the roughly fourfold rotational symmetry of these shapes? If so, then this would indicate that AlexNet formed an internal representation of a shape with fourfold rotational symmetry for the SM figures in Experiment 1.

### 6.1 Method

#### *6.1.1 Stimuli*

To test the third hypothesis, we used regular polygons; see Figure 6. It is notable that there is no regular polygon with twofold rotational symmetry. However, the vesica piscis shown in Figure 7 is a comparable shape that fits the pattern. It can be seen as "regular" in the sense it has equal-degree angles and equal-length sides. Furthermore, its two sides are mirrored curves, giving it bilateral symmetry. We therefore used it as our exemplar for twofold rotational symmetry.

We evaluated folds of rotational symmetries 2–6 as there is evidence of human sensitivity to sixfold rotational symmetry from grid cell firing patterns of the brain (Doeller et al., 2010). The perceptual limits of symmetry detection remain unclear at present, however. Therefore, we

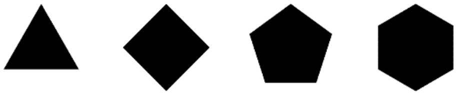

*Figure 6*. Regular polygons with folds of rotational symmetry 3–6.

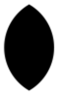

*Figure 7*. The vesica piscis, a lens shape with twofold rotational symmetry.

additionally evaluated regular polygons with 7–12 folds of rotational symmetry.

### *6.1.2 Procedure*

This experiment followed the same procedure as Experiment 1. For each of the 11 regular shapes, similarity scores were computed across rotation angles 0–360° for each of the two viewpoint models. We fit a sinusoidal function to these similarity scores using least squares to estimate the frequency parameter of the function, as well as its other 3 parameters. This is the critical parameter as it tells us the number of cycles of oscillations in similarity scores. Under our third hypothesis, we expected frequency to match the fold of rotational symmetry of each regular shape.

## 6.2 Results

Table 2 shows the fit of the sinusoidal function to the similarity scores over the interval of rotation angles for the two viewpoint models. The average $R^2$ values are high across rotational symmetry of folds 2–12: .93 for the NMSCC model and .88 for the AlexNet model. These findings are consistent with our third hypothesis—that the models exhibit cycles in similarity scores with rotation angle.

*Table 2*. Average sinusoidal fits of similarity score as a function of rotation angle across 2–12 folds of rotational symmetry per model.

| | Average Sinusoidal Fit ($R^2$) | Standard Error |
|---|---|---|
| NMSCC | .93 | < .01 |
| AlexNet | .88 | .02 |

We further evaluated our hypothesis that the number of cycles of oscillations in similarity

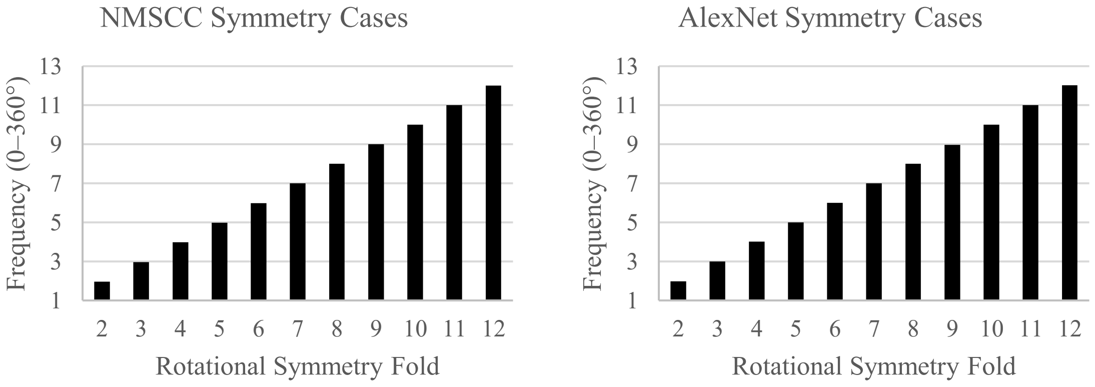

*Figure 8*. Frequency over fold of rotational symmetry produced by the NMSCC model, on the left, and the AlexNet model, on the right, using regular shapes.

score will match the fold of rotational symmetry of compared shapes by examining the frequency parameter of the model fits. The results are shown in Figure 8. Both models demonstrate a one-to-one correspondence between the frequency parameter and rotational symmetry fold of regular shapes, supporting our third hypothesis for simple shapes.

### 6.3 Discussion

The third research question asks whether the frequency of the models' similarity scores observed in Experiment 1 was driven by the rotational symmetry of the stimuli. As SM figures are 2D projections (i.e., viewpoints) of cubes viewed as a collection of quadrilaterals, the rotational symmetry of those quadrilaterals may explain the periodicity observed for the AlexNet model in Experiment 1; see Figure 4. To evaluate this possibility, Experiment 2 used simple (i.e., regular) 2D shapes of varying rotational symmetry. Both models demonstrated a strong correspondence between fold of rotational symmetry and frequency of oscillations in similarity score for these shapes. For the AlexNet model, this suggests AlexNet processed the SM figures in terms of their local fourfold rotational symmetry rather than their global onefold rotational symmetry.

This finding makes sense given how computer vision models like AlexNet are trained to identify discriminative features relevant for image classification. It could be the case AlexNet preferentially detects local patterns (e.g., the local fourfold rotational symmetry of 2D shapes in SM figures) without consideration for global spatial arrangement.

## 7. General Discussion

Imagistic theories of mental imagery can be distinguished into two classes. The classic analog theory proposes that people can spatially transform mental images; hence, the original term "mental rotation" (Shepard, 1978). The later viewpoint theory proposes that the similarity between images of objects at different viewpoints accounts for the linear trend between angular disparity and RT (Edelman, 1995; 1998; Morales & Firestone, 2022). We evaluated two computational models instantiating the viewpoint theory: the NMSCC model implements image-based processing whereas the AlexNet model implements structured information processing. We tested four hypotheses addressing different aspects of this theory.

The first research question asks whether the models can replicate the signature phenomenon of mental rotation: a linear trend between angular disparity and RT for both congruent and mirrored objects (Shepard & Metzler, 1971; Parsons, 1987). The central claim of the viewpoint theory is that longer RTs correspond to lower similarities between object viewpoints. The NMSCC model showed moderate support for this hypothesis, particularly for congruent stimuli. The bilinear similarity function exhibited by this model mirrors the pattern observed in human RTs during mental rotation; see Figure 4. By contrast, the AlexNet model showed a substantially weaker linear trend between angular disparity and similarity, contradicting our prediction that it would be the model that better captures the original results of Shepard and Metzler (1971); see Figure 5.

The second research question asks whether the models can distinguish between congruent stimuli and mirrored ones through similarity. The viewpoint theory predicts similarity scores for congruent objects will be consistently higher than those for mirrored objects. This prediction stems from similarity concerns alone, and if true, would match the human data showing RTs for mirrored objects are consistently slower than for congruent objects (Parsons, 1987). Neither model produced the expected pattern. Both the NMSCC model and the AlexNet model demonstrated substantial overlap between similarity scores for congruent stimuli and for mirrored stimuli; see Table 1. Thus, neither model adequately captures human performance.

The third research question asks whether the linear trend between similarity and angular disparity is also sensitive to the rotational symmetry of 2D shapes. SM figures are composed of cubes, where each face is a quadrilateral when projected onto the viewing plane. Although distorted by orientation and perspective, these quadrilaterals have roughly fourfold rotational symmetry (i.e., they are approximately congruent for every rotation of 90°). The similarity function for AlexNet contained four peaks and valleys, suggesting that it was attending to the component square faces of cubes rather than the overall shape. This possibility found support in Experiment 2, which used simple 2D shapes (e.g., regular polygons).

The fourth research question asks whether simple image-based similarity measures suffice for implementing the viewpoint theory, or whether the more structured forms of processing in the AlexNet computer vision model are necessary. Counter to our expectations, the simpler NMSCC model captured the bilinear pattern seen in human data better than the more complex AlexNet model. This suggests that low-level image-based processing may be a more psychologically plausible model of mental rotation. Still, neither instantiation of the viewpoint theory offers a satisfying account for the human data. Both computational models must be further developed if they are to serve as viable cognitive models.

### 7.1 Inferring the Optimal Direction of Rotation: Towards a Process Account

An enduring puzzle of mental rotation is how people consistently follow the optimal path of rotation when comparing objects even without advance knowledge of what it is (Cooper & Shepard, 1973; Shepard & Metzler, 1971). This is particularly interesting for mirrored objects, where the concept of a "shortest path" is technically ill-defined because no rotation can bring mirror images into alignment (Shepard & Metzler, 1971). Yet, people still respond in the SM task as a linear function of angular disparity for mirrored objects (Parsons, 1987).

The discrepancy between our models' performance and human behavior can be organized in terms of Marr's (1982) tri-level analysis of computational models in cognitive science. The NMSCC model operates at the lowest level of implementation, directly processing raw visual data. The AlexNet model operates at a higher level of algorithms and representations, extracting and transforming visual data into latent vector spaces of visual information instead.

We can restate the puzzle of optimality as a claim at the highest, computational level of Marr's (1982) analysis: mental rotation is *optimal* in always following the minimal (i.e., the shortest) path. This raises the question of whether either model considered here behaves consistently with this proposed optimality at the highest level. This is actually a difficult question to answer because neither model is a process account—neither specifies what is happening moment-by-moment during mental rotation. This is the reason we adopted the *linking hypothesis* that model similarity scores map to human RTs in our previous analyses.

That said, a viewpoint model can be straightforwardly adapted into a process model for mental rotation in the picture plane:

(1) Form the current image (i.e., the mental image to be rotated) and the target image (i.e., the mental image to be matched).
(2) Rotate the current image by a small increment in each of the two directions in the picture plane, clockwise or counterclockwise.
(3) Compute the similarity between each rotated image and the target image.
(4) Choose the rotated image with the greater similarity score to the target image as the new current image; doing so implicitly determines the path of rotation.
(5) Repeat steps 2–4 until the similarity score of the current and target images peaks. If the peak value is close to 1, conclude the two objects are congruent. Otherwise, conclude that they are mirrored.

In this process model, similarity serves both as the means for comparing objects and as the heuristic guiding the rotation process. Whereas the two models instantiating the viewpoint theory explain the linear trend between RT and angular disparity via the linking hypothesis that similarity shares this trend with them, the process model explains the trend as a product of an iterative process *involving* similarity. The process model follows the shortest path of rotation without advance knowledge of what this might be for both congruent stimuli and mirror-image stimuli, addressing the two puzzles of mental rotation (Cooper & Shepard, 1973; Shepard & Metzler, 1971).

There are two notable precedents for such a process model in the literature. The earliest is Funt's (1983) parallel-process model, which computes rotations in small increments. More similar to the present proposal is Hamrick and Griffiths' (2014) threshold model, hill climbing model, and Bayesian quadrature model, which all iteratively sample similarity to small increments of clockwise and counterclockwise rotations in the picture plane before committing to one in successive iterations. These models were originally designed with the analog theory in mind, and, indeed, the process model is an instantiation of the analog theory.

The process model's explanatory power critically depends on the goodness of the similarity measure because the shortest path of rotation must fall out of differences in similarity alone at any orientation. Still, it need not be "perfect." For the process model to align with the findings of Shepard and Metzler (1971), only a strictly monotonic trend between angular disparity and similarity is necessary. Put differently, the trend need not be perfectly linear; what matters is that optimality is incrementally approached.

With this sketch of a process model in hand, we can revisit the results of Experiment 1 and reassess the adequacy of the NMSCC model and the AlexNet model. Figure 4 shows that the NMSCC model does not implement a similarity function suitable for determining the optimal path of rotation. There is a large interval spanning approximately 90–270° where similarity scores plateau in the congruent case. In this region of relatively constant similarity scores, steps 3 and 4 of the process model would be unable to determine the optimal direction for the next rotation. Moreover, for mirrored stimuli, there are multiple local maxima that could trap the process model in a suboptimal path of rotation. Figure 5 shows that the similarity function of the AlexNet model also lacks the desired performance profile. The four cycles of oscillating similarity scores would make it an unreliable guide for incremental rotation towards alignment.

### 7.2 Limitations

One limitation of our study is that we only evaluated two computational models as instantiations of the viewpoint theory. Future work should evaluate other algorithms from image processing and other computational architectures from computer vision. These may include newer CNNs than AlexNet and also models that take the form of vision transformer (ViT) architectures, which are newer still.

Another limitation is that our experiments focused exclusively on rotation in the picture plane to control potential confounds from depth-related scaling effects. However, people can perform mental rotation in depth planes too (Shepard & Metzler, 1971). Future research should extend our investigations to rotations in depth planes to provide a more complete coverage.

## References

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*, *25*(11), 120–125.

Cooper, L. A., & Shepard, R. N. (1973). Chronometric studies of the rotation of mental images. In W. G. Chase (Ed.), *Visual information processing: Proceedings of the Eighth Annual Carnegie Symposium on Cognition* (pp. 75–176). Academic Press.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, *463*(7281), 657–661.

Edelman, S., & Weinshall, D. (1991). A self-organizing multiple-view representation of 3D objects. *Biological Cybernetics*, *64*(3), 209–219.

Edelman, S., & Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, *32*(12), 2385–2400.

Edelman, S. (1995). Representation of similarity in three-dimensional object discrimination. *Neural Computation*, *7*(2), 408–423.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, *21*(4), 449–467.

Funt, B. V. (1983). A parallel-process model of mental rotation. *Cognitive Science*, *7*(1), 67–93.

Hamrick, J. B., & Griffiths, T. L. (2014). What to simulate? Inferring the right direction for mental rotation. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 577–582.

Just, M. A., Carpenter, P. A., Maguire, M., Diwadkar, V. A., & McMains, S. (2001). Mental rotation of objects retrieved from memory: A functional MRI study of spatial processing. *Journal of Experimental Psychology: General*, *130*(3), 493–504.

Kosslyn, S. M., & Pomerantz, J. R. (1977). Imagery, propositions, and the form of internal representations. *Cognitive Psychology*, *9*(1), 52–76.

Kosslyn, S. M., & Sussman, A. L. (1995). Roles of imagery in perception: Or, there is no such thing as immaculate perception. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 1035–1041). MIT Press.

Kosslyn, S. M., Thompson, W. L., Wraga, M., & Alpert, N. M. (2001). Imagining rotation by endogenous versus exogenous forces: Distinct neural mechanisms. *NeuroReport*, *12*(11), 2519–2525.

Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2002). Mental imagery doesn't work like that. *Behavioral and Brain Sciences*, *25*(2), 199–208.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (Vol. 25). Curran Associates. (pp. 1097–1105).

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Metzler, J., & Shepard, R. N. (1974). Transformational studies of the internal representation of three-dimensional objects. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola Symposium* (pp. 231–247). Erlbaum.

Morales, J., & Firestone, C. (2022). Empirical evidence for perspectival similarity. *Psychological Review*, *129*(1), 135–144.

Niall, K. K. (2020). "Mental rotation" in depth as the superficial correlation of pictures. *Methods in Psychology*, *2*, Article 100019.

Niall, K. K. (2023). Mental rotation in depth as the optical difference of pictures. *Attention, Perception, & Psychophysics*, *85*(2), 368–387.

Parsons, L. M. (1987). Visual discrimination of abstract mirror-reflected three-dimensional objects at many orientations. *Perception & Psychophysics*, *42*(1), 49–59.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates. (pp. 8024–8035).

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin*, *80*(1), 1–24.

Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioral and Brain Sciences*, *25*(2), 157–182.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Shepard, R. N. (1978). The mental image. *American Psychologist*, *33*(2), 125–137.

Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations*. MIT Press.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review*, *1*(1), 2–28.

Stewart, E. E., Hartmann, M., Morgenstern, Y., Storrs, K. R., Maiello, G., & Fleming, R. W. (2022). Mental object rotation based on two-dimensional visual representations. *Current Biology*, *32*(21), 4526–4534.