# Make Use of Mooney Images to Distinguish between Machines and Humans

**Jingmeng Li (jmli17@fudan.edu.cn)**
Department of Computer Science, Fudan University, 2005 Songhu Road
Yangpu District, Shanghai, China

**Hui Wei* (weihui@fudan.edu.cn)**
Department of Computer Science, Fudan University, 2005 Songhu Road
Yangpu District, Shanghai, China

## Abstract

Completely automated public Turing test to tell humans apart (CAPTCHA) aims to exploit the ability gaps between machines and humans to distinguish between them. However, the rapid development of artificial intelligence technology in the past decade has significantly narrowed the gap in some tasks based on natural images (*e.g.*, object detection and recognition). Mooney images (MIs) are important research materials in the field of cognitive science. Compared to natural images, we perceive MIs relying more on the iteration between feedforward and feedback processes. In this paper, we explored an intriguing question: **Can MIs be used to distinguish between machines and humans?** Before this study, we first proposed a framework HiMI that generated the high-quality MIs from natural images and also allowed flexible adjustment of the perceived difficulty. Next, we designed two MI-based Turing test tasks related to foreground-background segregation and object recognition, respectively. We compared the performance of human subjects and the deep neural networks on these two tasks. The experimental results indicate the significant gaps between the deep neural networks and humans, providing evidence for the potential of MIs in the design of CAPTCHA schemes. We hope that HiMI will contribute to more research related to MIs in the fields of cognitive science and computer science.

**Keywords:** Turing test; CAPTCHA; Mooney image; closed-loop information processing; object detection; figure-ground segregation; deep neural networks.

## Introduction

The Turing test, proposed by computer scientist Alan Mathison Turing, aims to judge whether a machine attains a level of intellectual capacity akin to that of humans (French, 2000). Researchers later proposed the schemes of completely automated public Turing tests to tell humans apart (CAPTCHA) to ensure the security of websites and online applications (Xu, Liu, & Li, 2020). According to the format of data, these CAPTCHA schemes can be broadly categorized into three types: text, image, and audio. Text-based and image-based CAPTCHA schemes are the most widely applied and typically rely on visual tasks of target detection and recognition (Alqahtani & Alsulaiman, 2020; Shi et al., 2020). Specifically, users are required to identify English letters, Chinese characters, or target objects within them.

With the rapid improvement in the performance of recognition algorithms, the security of many CAPTCHAs is under severe threat (Zhao et al., 2018; Searles et al., 2023). Algorithms can be roughly divided into two categories based on their development process: traditional methods and those based on deep neural networks. Traditional methods are inspired by perceptual processing theories (*e.g.,* Gestalt theory (Wagemans, Elder, et al., 2012; Wagemans, Feldman, et al., 2012)) and object recognition theories (*e.g.,* Object template theory, Recognition by components theory (Biederman, 1987)). Designers of text-based CAPTCHAs typically employ some techniques to enhance recognition difficulty (Bursztein, Martin, & Mitchell, 2011; Wang et al., 2023). For instance, introducing noise into the CAPTCHA, increasing overlap and intersection between texts, and inducing a certain degree of deformation. Moreover, the complexity of content and interference noise in natural images result in incomplete feature extraction, thereby affecting the performance of traditional methods. Deep neural networks have made significant strides in recognition accuracy compared to traditional methods. With the massive training data and powerful computation, the performance of these methods even surpasses that of humans (Alqahtani & Alsulaiman, 2020).

In daily life, we can effortlessly recognize pedestrians and vehicles on the road, words in a book, or food on a desktop. The efficiency of the visual system leads us to overlook the process from the initial visual signals to the emergence of perception. Research indicates that approximately half of the cerebral cortex in primates is involved in visual perception (Felleman & Van Essen, 1991; DiCarlo, Zoccolan, & Rust, 2012). Therefore, this process involves high computational complexity. Fig. 1 (A) displays the closed-loop information processing process for object recognition in the human visual system. It comprises two iterative components: feedforward/bottom-up and feedback/top-down (Theeuwes, 2010). During the bottom-up process, the visual system integrates low-level visual signals through perceptual organization to obtain higher-level visual features and extract crucial cognitive cues (*e.g.*, shape, texture). Our brain combines these cues with rich object knowledge, generating visual expectations. During the feedback process, the visual system actively adjusts perceptual organization by integrating missing information, filtering noise, and reducing redundancy to meet these expectations.

The Mooney image (MI) are a type of stylized image that consists of discrete speckles with irregular shapes and sizes, colored only in black and white (Mooney, 1957; Mitra et al., 2009; Hegdé, Thompson, & Kersten, 2007). It is obtained by binarizing natural images and setting a threshold to pre-



In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* 

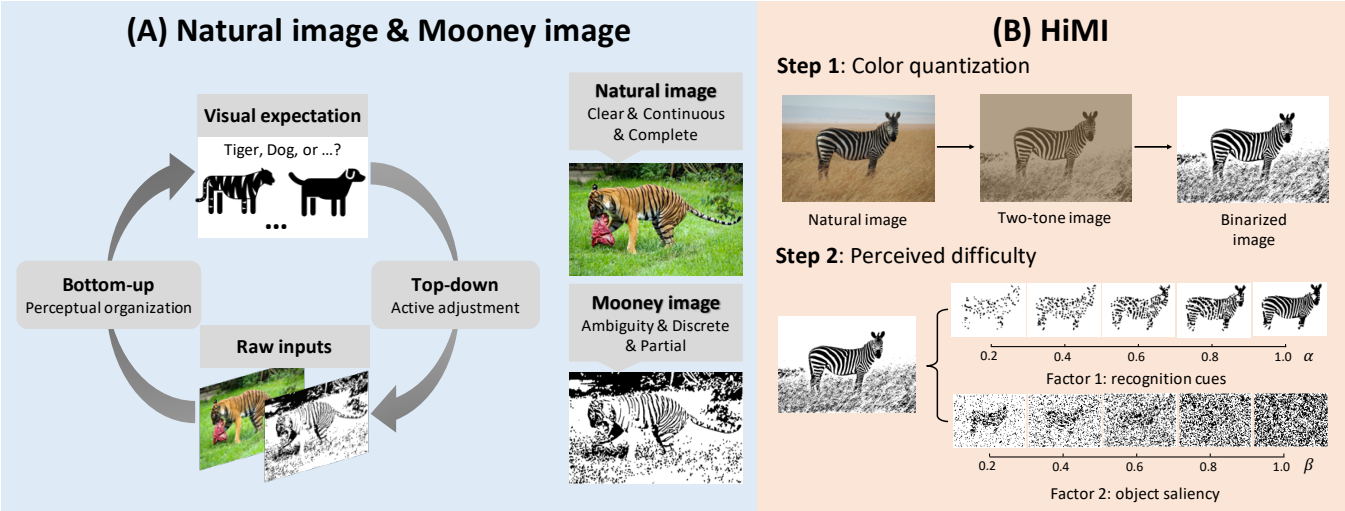


Figure 1: (A) Comparison between the natural and MIs. Visually, natural images contain rich cues, complete and continuous object information, while in MIs, there are fewer cues, the content is incomplete and discrete. According to the closed-loop information processing process of human vision system, perceiving the target object from MIs relies more on the iteration between the bottom-up process and the top-down process. (B) The overview of HiMI. It consists of two steps, color quantization and perceived difficulty control. Color quantization aims to reduce the color space to 1-bit while preserving as much of the natural image content as possible. We consider two factors, object saliency and recognition cues, to control the perceived difficulty of MIs.

serve the highlighted areas. Fig. 1 (A) displays an example of the MI. When some speckles are appropriately organized together, we can perceive a tiger. From the visual comparison between the natural image and the MI, we can observe that the MI contains fewer visual cues, and its content is partial and discrete. These characteristics make it an important research material in cognitive science. The discrete speckles increase the difficulty of perceptual organization in the bottom-up process, enabling researchers to more clearly observe and record the occurrence of perceptual organization (Andrews & Schluppeck, 2004; Mooney, 1957; Grützner et al., 2010). In addition, MIs miss some visual content and cues, increasing its ambiguity. Therefore, it is also used to explore related research on utilizing prior-knowledge to visual disambiguation (Hegdé & Kersten, 2010).

While deep neural networks improves the performance of machines in certain visual tasks, they lack the biological mechanisms, which may result in differences in visual representations and consequently lead to gaps in behavior and capabilities. Neurobiological studies found that long-range horizontal connections between neurons in the same hierarchical visual cortex contribute to the integration of visual signals in the bottom-up process (Das & Gilbert, 1995). Houtkamp et al. proposed the pathfinder challenge to investigate the principle of good continuation in Gestalt theory (Houtkamp & Roelfsema, 2010). Inspired by this, Kim et al. trained a deep neural network using the pathfinder challenge dataset and evaluated the performance of multiple deep models (Linsley, Kim, Veerabadran, Windolf, & Serre, 2018; Kim, Linsley, Thakkar, & Serre, 2019). The experimental results demonstrate the gap in solving this challenge between deep neural networks and human. Furthermore, the interpretability of deep neural networks is poor. Specifically, we cannot discern the basis for the final decisions made by deep neural networks. For instance, research by Geirhos et al. revealed instances where deep neural networks incorrectly classified the cat with the elephant texture as an elephant (Geirhos et al., 2018).

In this paper, we explored an intriguing question: **Can MIs be employed in CAPTCHA schemes to distinguish machines from humans?** Before this study, we first proposed a universal framework HiMI that can use the natural image to generate high-quality MIs. HiMI consists of two steps: color quantization and perceived difficulty control. In the first step, we reduce the color space of the image to 1-bit while preserving as much of the natural image content as possible. Next, we consider two factors object saliency and recognition cues, and set the corresponding parameters to control the perceived difficulty of MIs. We demonstrated the effectiveness of these two parameters through user experiments. After using HiMI to generate diversified MIs, we designed two MI-based Turing test tasks related to the figure-ground segregation and object recognition respectively. We compared the performance of human subjects and deep neural networks on solving these two tasks. The experimental results show a significant gap between deep neural networks and humans in solving these two MI-based perceptual tasks, which provides evidence that the MI can be used to distinguish between machines and humans.

## HiMI: High-quality Mooney Images

We need to generate high-quality MIs for the following MI-based Turing test. The existing MI generation methods typically set a threshold for natural images and retain the content of the highlighted regions. However, this processing has some issues. Firstly, there is inconsistency in brightness between different images. If we set a uniform threshold for all images, inappropriate threshold values will result in excessive loss of image content. In addition, the brightness within a single image is uneven. If the background region has higher brightness than the foreground region, an inappropriate threshold will result in the final MI missing the foreground object to be recognized. The design goal of the image-based CAPTCHA is to minimize the impact on human recognition performance while enlarge the impact on machines as much as possible. Therefore, we need to design a generation method that can flexibly adjust the perceived difficulty of MI.

### Two Factors for Perceived Difficulty

**Object saliency.** Before recognizing the target object, the visual system needs to perform figure-ground segregation. Research indicates that visual attention (including endogenous spatial attention and exogenous spatial attention) influences the occurrence of figure-ground segregation (Vecera & Farah, 1994; Kimchi, Yeshurun, Spehar, & Pirkner, 2016; Vecera, 2000). Vecera et al. further demonstrated that exogenous spatial attention influenced the role of bottom-up Gestalt cues in figure-ground segregation (Vecera, Flevaris, & Filapek, 2004). The pattern (*e.g.*, texture, color) differences between the foreground and background are closely related to exogenous spatial attention. An example is that some organisms have evolved camouflage colors to self-protect by reducing predators' exogenous spatial attention. The greater pattern difference between the foreground and background, the faster occurrence of figure-ground segregation.

**Recognition cues.** The amount of object recognition cues in MIs can affect the accuracy. In object recognition, we use object features such as color, texture, and shape. Traditional recognition theories emphasize that shape is more important in object recognition. Psychological-behavioral experiments have shown that surface information (color, texture) speeds up recognition but does not significantly improve recognition accuracy (Gegenfurtner & Rieger, 2000). Biederman's recognition-by-components asserts that surface information only plays a role in low-level vision and provides cues for the organization and integration of visual signals while object recognition tasks rely on shape (Biederman, 1987). However, this view cannot explain discrimination between horses and zebras. If we only provide subjects with the shape of a zebra, they will likely mistake the zebra for a horse. The "shape + surface" computational framework for object recognition suggests that surface and shape information play a joint role in high-level visual processing, and that the role of surface information depends heavily on differences in structural properties between the objects in question (Tanaka, Weiskopf, & Williams, 2001).

According to "shape + surface" theory, the process of perceiving the tiger in the MI shown in Fig. 1 (A) can be described as follows. The visual system first obtains shape information, such as edges or contour segments, based on the physical features of visual signals in the bottom-up process. Then, it reorganizes the shape and surface information in the top-down process to discover more holistic combinations and form a cognitive hypothesis of a tiger based on a prior knowledge. Parsing the process backward gives us the following insight. Under the condition that a prior knowledge is available, the visual system reorganizes signals and collects cues in an iterative way. The results of cue collection affect the speed and accuracy of visual expectation which in turn affects the speed and accuracy of object recognition in the MI.

### Generation Process

As shown in Fig. 1 (B), we use a zebra image to introduce HiMI. Compared to the 24-bit color space of natural images, MIs are two-tone (1-bit color space). Therefore, we need to perform color quantization on natural images. We use Color-CNN proposed by Hou et al. to reduce the color space while preserving image content as much as possible (Hou, Zheng, & Gould, 2020). Next, we control the perceived difficulty based on the results of color quantization (binarized image). We set two parameters, $\alpha$ and $\beta$, to control the amount of recognition cues and object saliency, respectively. The first parameter, $\alpha$, is used to adjust the proportion of recognition cues. For example, $\alpha = 0.2$ means 20% of the surface information will be randomly selected to be rendered by speckles. After the parameter $\alpha$ is set, the density of speckle in the object region can be calculated. Then, the second parameter $\beta$ controls the density of noise speckles around the object. For example, $\beta = 0.2$ means the speckle-density of the surroundings is 20% of the object region.

### Natural Image Datasets

Two public image datasets (Animal 2K dataset (Li, Zhang, Maybank, & Tao, 2022) and PASCAL VOC2012 (Everingham & Winn, 2012)) are used in our study. Animal 2K is created by Li et al. for natural image matting studies in computer vision, and it includes 2000 images in 20 animal categories. Most images in the Animal 2K contain only one animal. In addition, the images in the dataset are high resolution, which makes data processing easier. PASCAL VOC2012 is a classical dataset for multiple computer vision tasks such as image classification, object detection, and image segmentation. It contains 20 classes of objects with a total of 11,530 images.

### Effectiveness of Two Parameters

The purpose of this experiment is to test the effectiveness of two parameters in HiMI. Subjects are presented with MIs that reflected only changes in one parameter $\alpha$ or $\beta$ to reduce the influence of other factors on the experimental results.

**Participants.** A total of 100 students participated in this experiment (mean age = 22.8 years; 50 female). The participants come from the School of Computer Science, School of Psychology, School of Life Sciences, and School of Mathematics. None of the subjects had visual cognitive impairment. The 100 subjects were divided equally into ten groups: $G_1$, ..., $G_{10}$. The ratio of male to female in each group was kept the same as the overall ratio.

**Stimuli.** We randomly select one image from each category in Animal 2K and PASCAL VOC2012 datasets. Totally fourty images were then used to generate MIs. We employed the concept of controlling variables to generate MIs for testing these two factors. When generating MIs for validating the factor recognition cues (α), we set the parameter β to 0 to minimize the impact of object saliency and sequentially adjusted the parameter α to 0.2, 0.4, 0.6, 0.8, and 1. When generating MIs for validating the factor object saliency (β), we set the parameter α to 1 to reduce the impact of object recognition cues and sequentially adjusted the parameter β to 0.2, 0.4, 0.6, 0.8, and 1. Each natural image had ten corresponding MIs. In the experiment, ten MIs corresponding to a natural image were presented to subjects in the ten groups, respectively.

**Procedure.** Subjects sat 34 cm in front of a 24-inch monitor with a resolution of $1920 \times 1080$. The screen sequentially displayed 40 MIs, each containing only one target object. The subjects' task was to observe the MI and identify the target object, and then verbally state the object's category. Afterward, the subjects clicked the "Next" button to play the next MI and repeated the process.

**Results.** The time taken from the presentation of a MI on the screen (start of observation) to the subject stating the category of the object in the image (end of observation) is referred to as the reaction time (RT). A smaller RT value means that the subject recognize the object from the MI more quickly and with less difficulty. For each MI, a group of RT values was obtained from 10 subjects, and the average of these 10 values was the subjects' RT for recognizing the target object in this MI. After obtaining the RTs of 40 MIs under a certain parameter setting (e.g., $\alpha = 0.2$ or $\beta = 0.2$), we calculated the mean of these 40 values as the RT for the current parameter setting. Fig. 2 presents the statistical results of subjects' RTs when observing MIs under different settings of the recognition cues (α) and the object saliency (β). The experimental results demonstrate that both two parameters α and β can significantly influence the speed of object recognition and thus control the perceived difficulty of the generated MIs.

In the experiment, the subjects' recognition results for each MI were recorded. A correct recognition was recorded as 1; otherwise, 0. Each stimuli had 10 values of 0 or 1, and the mean value was the recognition accuracy that reflects the difficulty of recognizing the target objects from the MI. Fig. 3 illustrates the impact of the two parameters α and β on the recognition accuracy. When fixing the parameter β at 0 and sequentially adjusting the parameter α, the recognition accuracy gradually improves. However, when fixing the parameter α at 1 and adjusting the parameter β, there is no significant change in the accuracy.

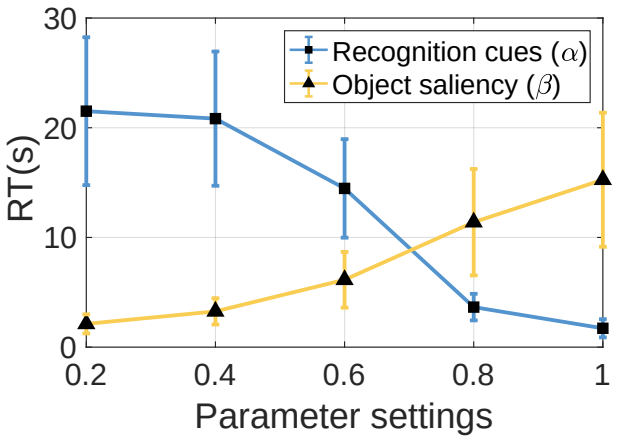


Figure 2: Statistical results of RTs with 90% confidence intervals.The results demonstrate that both recognition cues (α) and object saliency (β) have a significant impact on the speed of object recognition.

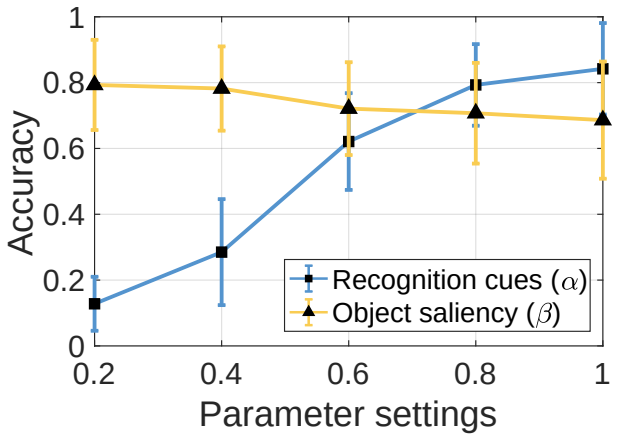


Figure 3: Statistical results of recognition accuracy with 90% confidence intervals. We can observe that the factor recognition cues (α) has a significant impact on the accuracy, while object saliency (β) only has a slight impact on it.

## MI-based Turing Test

In this section, we will explore an intriguing question: Can Mooney images be used in CAPTCHA to distinguish between humans and machines? Inspired by the famous Google reCAPTCHA v2, we design two tasks (shown in Fig. 4): **(1) select all MIs that contain the target category of objects from a set of MIs**; **(2) select the one located in the area of object from four red dots**.

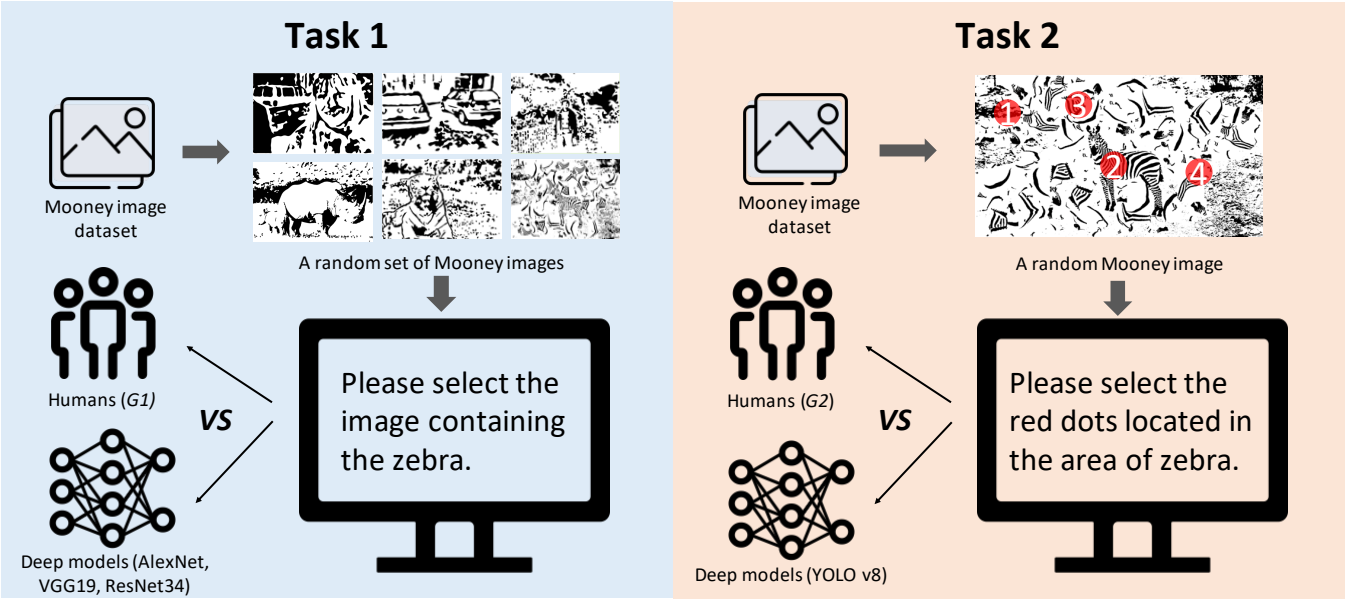


Figure 4: Explanation for two MI-based Turing test tasks. Task 1 is related to object recognition. We randomly select six MIs from $\text{Animal}_{\text{MI}}$ and $\text{PASCAL}_{\text{MI}}$, including $N$ ($1 \leq N \leq 3$) images that contain the objects of the target category. The human subjects and the deep neural networks need to find all the images that contain the target objects. Task 2 is related to the figure-ground segregation. We randomly select one from the six MIs of Task 1, and combine the object segmentation groundtruth provided by dataset to generate four red dots. The human subjects and deep vision model need to find the one located within the area of the target object.

## Participants and Deep Models

We invited the same 100 subjects (average age = 22.8 years; 50 female) to participate in this experiment. The subjects were divided equally into two groups: $G1$ (average age = 22.5 years) and $G2$ (average age = 23.1 years). The proportion of male and female subjects in the two groups was the same as the overall proportion. We ensured that these 100 subjects had the required prior knowledge before the experiment. YOLO, a classical one-stage detection model, has performed impressive results on the task of object deteciton, and it is also used in the Google reCAPTCHA v2 solver. In Task 1, we use AlexNet (Krizhevsky, Sutskever, & Hinton, 2017), VGG19 (Simonyan & Zisserman, 2014) and Resnet34 (He, Zhang, Ren, & Sun, 2016) as the task networks. In Task 2, we test the ability of YOLO v8 (Redmon, Divvala, Girshick, & Farhadi, 2016) to detect objects based on MIs.

## Stimuli

We generate MIs based on the Animal 2K and the PASCAL VOC2012 datasets. We adjust the generated results using two parameters $\alpha$ and $\beta$. Here, we use $\text{HiMI}_{(\alpha,\beta)}$ to denote the generated results the with different parameter settings. In Table 2, we take some representive examples to explain the meanings of the symbols related to these two parameters in this experiment. We firstly generate two corresponding Mooney-style datasets, denoted as $\text{Animal}_{\text{MI}}$ and $\text{PASCAL}_{\text{MI}}$, respectively. For each natural image in the training set, we use $\text{HiMI}_{(1,\{0,0.5,1\})}$ to generate its MIs with different perceived difficulty. For each natural image in the test set, we use $\text{HiMI}_{(1,[0,1])}$ to generate its MIs.

Table 1: Meanings of symbols related to parameters $\alpha$ and $\beta$ in this experiment.

| Parameters | Symbols | Meaning |
|---|---|---|
| $\alpha, \beta$ | 1 | set parameter to 1 |
| | {0,0.5,1} | set parameter to 0, 0.5, 1 in order |
| | [0,1] | randomly set it to a value within [0, 1] |

## Procedure

**Model training.** We firstly use the officially available pre-trained models, and then fine-tune them on Animal 2K, $\text{Animal}_{\text{MI}}$, PASCAL VOC2012, and $\text{PASCAL}_{\text{MI}}$. For training, we use a batch size of 16 and train these model for 300 epochs with an initial learning rate of 0.01.

**Task 1.** We randomly select six MIs from the test set, including $N$ ($1 \leq N \leq 3$) images that contain the target category of objects. We conducted 50 task experiments for each setting of $N$. Subjects in $G1$ are instructed to select all images containing the target category of objects from this set of MIs, and the task for the deep neural networks is to classify these six MIs. The human approach to solving this task relies on two strategies. First, subjects can accurately identify $N$ MIs

that contains the target category of objects. Second, subjects can successfully identify the other $(6-N)$ MIs, and then arrive at the correct result through the process of elimination. Therefore, we established two rules to evaluate the deep neural networks' results. First, if the deep neural network can correctly classify $N$ target MIs, the task is considered successfully solved. Second, if the deep neural network can classify the remaining $(6-N)$ MIs, it is also considered successful.

**Task 2.** We randomly select one from the set of six MIs, and generate four red dots based on the groundtruth of object segmentation provided by the dataset. Note that one of the red dots must be located in the area of the target object. Subjects in $G2$ are required to select a red dot from these four red dots. The task for the deep vision models is to segment the object from the MI, and the result will be successful if the segmented object region only includes a red dot.

## Results

Fig. 5 shows the success rates of Task 1. When $N = 1$, human subjects achieved a success rate of 85.4%, while the success rates of the three deep neural networks (AlexNet, VGG19, and ResNet34) were 23.9%, 20.7%, and 33.4%, respectively. As the number of MIs containing the target object increased in the set of six MIs, both human subjects and the three deep neural networks presented a decrease in success rates. However, human subjects maintained a significantly higher success rate compared to these three deep neural networks. In Table 2, we present the comparison of the success rates in completing Task 2 between human subjects and the deep vision model YOLO v8. Human subjects achieved a success rate of 87.2% in solving Task 2, while YOLO v8 only achieved a success rate of 33.5%. The above experimental results demonstrate that human subjects exhibit stronger abilities than deep neural networks in solving these two MI-based vision tasks.

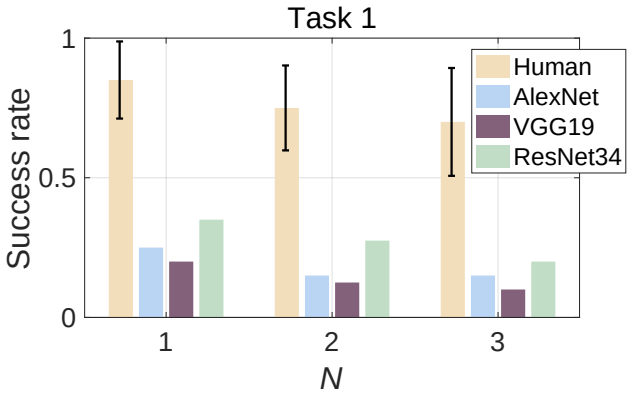


Figure 5: Statistical results of success rate on Task 1. Compared to deep neural networks, humans have a significant advantage in the success rate of completing Task 1.

Table 2: Statistical results of success rate on Task 2.

| | Human | YOLO v8 |
|---|---|---|
| Success rate | 87.2% | 33.5% |

## Conclusion & Discussion

While the MI is an important research material in the field of cognitive science due to their distinctive characteristics, there is still a lack of an efficient generation scheme. Therefore, in this paper, we proposed HiMI for generating high-quality MIs from natural images. It involves two steps: color quantization and perceived difficulty control. The first step aims to reduce the color space of natural images to 1-bit while preserving the semantic content of the images. In the second step, we set two parameters to adjust object saliency and recognition cues, enabling users to change the perceived difficulty of MIs according to their needs. We verified the effectiveness of the two parameters through controlled experiments. We initially fixed the recognition cues and adjusted object saliency, finding a noticeable impact on subjects' recognition speed with a relatively minor effect on accuracy. Conversely, when we fixed object saliency and altered recognition cues, we observed a significant impact on both subjects' recognition speed and accuracy. Additionally, what cognitive research can we use HiMI for? (1) By two factors, object saliency and recognition cues, we can respectively adjust subjects' exogenous spatial attention and endogenous spatial attention during MI observation. Although some research has identified the impact of these two spatial attentions on behaviors like object detection, the mechanisms underlying how they affect low-level visual representations and how their interaction occurs remain unclear (Fernández, Li, & Carrasco, 2019). (2) Visual disambiguation refers to the ability to interpret ambiguous information in a reasonable manner, which is important in an ever-changing external environment. We currently understand the significant role of prior object-knowledge in disambiguation, but the neural circuits behind disambiguation induced by prior knowledge remain unclear (Hegdé & Kersten, 2010).

We are in an era of rapid development in artificial intelligence technology. With the support of data and computational resources, the deep neural networks can achieve performance in certain visual tasks based on natural images that matches or even surpasses humans. This suggests that these tasks no longer effectively reflect the differences in abilities between machines and humans. In this study, we attempted to utilize visual tasks based on MIs to distinguish between machines and humans. Task 1 involves selecting the one containing the target object from a set of six MIs. Task 2 requires selecting one of the four dots in a single MI located within the target object area. Experimental results indicate a significant performance gap between the deep neural networks and humans.

## Acknowledgement

This work was supported by the NSFC Project (Grant number: 61771146).

## References

Alqahtani, F. H., & Alsulaiman, F. A. (2020). Is image-based captcha secure against attacks based on machine learning? an experimental study. *Computers & Security*, *88*, 101635.

Andrews, T. J., & Schluppeck, D. (2004). Neural responses to mooney images reveal a modular representation of faces in human visual cortex. *Neuroimage*, *21*(1), 91–98.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, *94*(2), 115.

Bursztein, E., Martin, M., & Mitchell, J. (2011). Text-based captcha strengths and weaknesses. In *Proceedings of the 18th acm conference on computer and communications security* (pp. 125–138).

Das, A., & Gilbert, C. D. (1995). Long-range horizontal connections and their role in cortical reorganization revealed by optical recording of cat primary visual cortex. *Nature*, *375*(6534), 780–784.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Everingham, M., & Winn, J. (2012). The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep*, *2007*(1-45), 5.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, *1*(1), 1–47.

Fernández, A., Li, H.-H., & Carrasco, M. (2019). How exogenous spatial attention affects visual representation. *Journal of Vision*, *19*(11), 4–4.

French, R. M. (2000). The turing test: the first 50 years. *Trends in cognitive sciences*, *4*(3), 115–122.

Gegenfurtner, K. R., & Rieger, J. (2000). Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology*, *10*(13), 805–808.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Grützner, C., Uhlhaas, P. J., Genc, E., Kohler, A., Singer, W., & Wibral, M. (2010). Neuroelectromagnetic correlates of perceptual closure processes. *Journal of Neuroscience*, *30*(24), 8342–8352.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hegdé, J., & Kersten, D. (2010). A link between visual disambiguation and visual memory. *Journal of Neuroscience*, *30*(45), 15124–15133.

Hegdé, J., Thompson, S., & Kersten, D. (2007). Identifying faces in two-tone ('mooney') images: A psychophysical and fmri study. *Journal of Vision*, *7*(9), 624–624.

Hou, Y., Zheng, L., & Gould, S. (2020). Learning to structure an image with few colors. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 10116–10125).

Houtkamp, R., & Roelfsema, P. R. (2010). Parallel and serial grouping of image elements in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(6), 1443.

Kim, J., Linsley, D., Thakkar, K., & Serre, T. (2019). Disentangling neural mechanisms for perceptual grouping. *arXiv preprint arXiv:1906.01558*.

Kimchi, R., Yeshurun, Y., Spehar, B., & Pirkner, Y. (2016). Perceptual organization, visual attention, and objecthood. *Vision Research*, *126*, 34–51.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

Li, J., Zhang, J., Maybank, S. J., & Tao, D. (2022). Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision*, *130*(2), 246–266.

Linsley, D., Kim, J., Veerabadran, V., Windolf, C., & Serre, T. (2018). Learning long-range spatial dependencies with horizontal gated recurrent units. *Advances in neural information processing systems*, *31*.

Mitra, N. J., Chu, H.-K., Lee, T.-Y., Wolf, L., Yeshurun, H., & Cohen-Or, D. (2009). Emerging images. *ACM transactions on graphics (TOG)*, *28*(5), 1–8.

Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *11*(4), 219.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 779–788).

Searles, A., Nakatsuka, Y., Ozturk, E., Paverd, A., Tsudik, G., & Enkoji, A. (2023). An empirical study & evaluation of modern captchas. In *32nd usenix security symposium (usenix security 23)* (pp. 3081–3097).

Shi, C., Ji, S., Liu, Q., Liu, C., Chen, Y., He, Y., ... Wang, T. (2020). Text captcha is dead? a large scale deployment and empirical study. In *Proceedings of the 2020 acm sigsac conference on computer and communications security* (pp. 1391–1406).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tanaka, J., Weiskopf, D., & Williams, P. (2001). The role of color in high-level vision. *Trends in cognitive sciences*,

*5*(5), 211–215.

Theeuwes, J. (2010). Top–down and bottom–up control of visual selection. *Acta psychologica*, *135*(2), 77–99.

Vecera, S. P. (2000). Toward a biased competition account of object-based segregation and attention. *Brain and Mind*, *1*, 353–384.

Vecera, S. P., & Farah, M. J. (1994). Does visual attention select objects or locations? *Journal of Experimental Psychology: General*, *123*(2), 146.

Vecera, S. P., Flevaris, A. V., & Filapek, J. C. (2004). Exogenous spatial attention influences figure-ground assignment. *Psychological Science*, *15*(1), 20–26.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & Von der Heydt, R. (2012). A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, *138*(6), 1172.

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., Van der Helm, P. A., & Van Leeuwen, C. (2012). A century of gestalt psychology in visual perception: Ii. conceptual and theoretical foundations. *Psychological bulletin*, *138*(6), 1218.

Wang, P., Gao, H., Guo, X., Xiao, C., Qi, F., & Yan, Z. (2023). An experimental investigation of text-based captcha attacks and their robustness. *ACM Computing Surveys*, *55*(9), 1–38.

Xu, X., Liu, L., & Li, B. (2020). A survey of captcha technologies to distinguish between human and computer. *Neurocomputing*, *408*, 292–307.

Zhao, B., Weng, H., Ji, S., Chen, J., Wang, T., He, Q., & Beyah, R. (2018). Towards evaluating the security of real-world deployed image captchas. In *Proceedings of the 11th acm workshop on artificial intelligence and security* (pp. 85–96).