Cognition-Inspired Autonomy in Puzzle Rearrangement

Qiaosi Lei
Beijing No.101 High School
24joss.lei@beijing101id.com

Rui Wang
Beijing No.101 High School
25rachel.wang@beijing101id.com

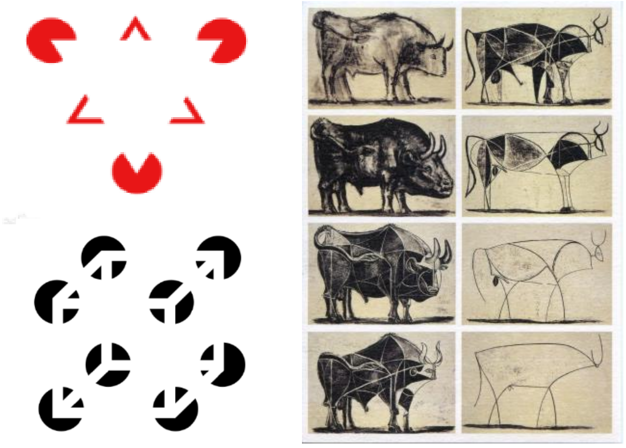
Abstract

While generative AI for content generation gets popular, the next step is to explore more intricate cognitive abilities with autonomous mechanisms to support cognitive reasoning. The tangram, known as a Chinese puzzle, is widely applied in training and evaluating human cognitive abilities. We utilize tangram as the experiment platform to investigate and mimic how human image and reason with the abstract shape consist of constrained polygons. Comparing to the tasks of existing pixel-oriented generative AI models for open boundary abstract-to-concrete generation, the autonomous arrangement of tangram must face the challenges of concrete-to-abstract and abstract-to-abstract mapping with strict geometric constraints on tans in the form of various polygons and sophisticated spatial relationship among them for forming the target shape. To deal with the challenges, firstly we design cognitive reasoning tasks with gradually increasing complexity for human cognition abilities including completeness *and closure, and abstract representation. Then we investigate the human cognitive process for the above tasks including perception, memory, reasoning and imagination, and exploit the automatic generation method based on Denoising Score Matching centered models with incrementally optimized structure and training policy. To support above tasks, we take great efforts to generate and customize task-specific datasets. The qualitative and quantitative experiment results demonstrate that our method could achieve the goals of the designed cognitive reasoning tasks with effectiveness and efficiency, underscoring its potential for advancing the cognitive capability of generative AI in extended fields.*

Keywords—generative AI, cognitive reasoning, tangram arrangement, denoising score matching.

I. Introduction

Generative AI has gradually become a popular topic in the field of artificial intelligence. It utilizes deep learning algorithms to analyze patterns from vast amounts of data in order to generate high-quality content. At present, several representative systems, such as DALL-E, Midjourney, and Stable Diffusion [1], have achieved high-quality image generation based on textual descriptions. These systems have been successfully applied in various fields, including games and media. From the perspective of human cognition, above system takes the first step of abstract-to concrete mapping. It is the next step to explore more intricate cognitive abilities of humans and develop an autonomous mechanism that supports cognitive reasoning. Cognition psychology explores states and processes that involve the occurrence, transformation, and storage of information-bearing structures in the mind or brain [31]. For example, two of the most important concepts in Gestalt Theory are abstract representation, and completeness and closure [32] with detail depicted below.



Completeness and Closure

Abstract Representation

Figure 1. Examples of human cognition abilities.

- **Completeness and Closure**: According to Gestalt principles of organization, incomplete forms are perceived by people with a tendency as complete, synthesizing the missing units to create complete forms [32].
- **Abstract Representation**: Mapping the concepts from one domain to another, such as metaphors in

cognitive semantics. In cognitive psychology theory, it is a a mental representation of a stimulus in an abstract or essential form that is not tied to any one of its variable surface forms [20].

Tangram, known as a Chinese puzzle, is a typical puzzle game widely applied in training and evaluating human cognitive abilities, such as spatial thinking, creativity, and problem-solving skills. Tangram is a collection of seven polygons, called tans: five isosceles right triangles, a square and a parallelogram. These pieces are arranged, using Euclidean isometries, to form dissections of prescribed or unknown polygons as illustrated in Figure 2. As a representation and abstraction of multidimensional space in two-dimensional space, it supports various complex arrangement methods based on mathematical rules and can even be utilized to solve or prove mathematical problems [2]

Cognitive psychology research has explored the use of tangram as a cognitive task for humans and has revealed several factors that can affect performance, including working memory [25], mental rotation [26], visual perception [27], strategy use [28], and feedback [29], etc. Tangram is known to be effective in training geometry reasoning and spatial ability [21], which have long been viewed as measures of practical and mechanical abilities that are useful in predicting success in technical occupations [22].

Currently, tangram is being used in some artificial intelligence research projects, such as visual feature extraction and low-resolution visual tasks based on pre-trained models. These studies primarily concentrate on the perceptual aspects of solving tangram arrangements and do not involve more complex cognitive tasks. In the literature, we found no prior work with the primary objective of automatically generating tangram arrangements. Autonomous generation and selection of tangram patterns, as well as the use of cognitive reasoning methods to find optimal or creative solutions, is of great significance to explore the application of generative AI in a broader range of cognition related fields.

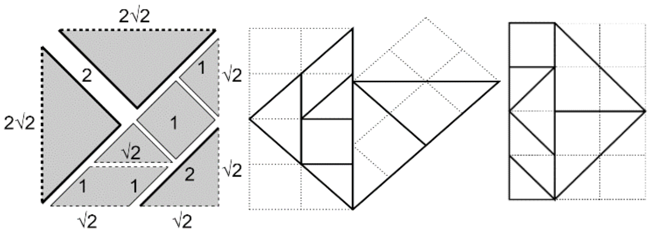


Figure 2. Tangram arranged with restricted spatial constraints.

We utilize tangram as the experimental platform to investigate the autonomous mechanism and cognitive reasoning ability of artificial intelligence. Comparing to the tasks of existing generative AI models, autonomous arrangement of tangram must face to following challenges

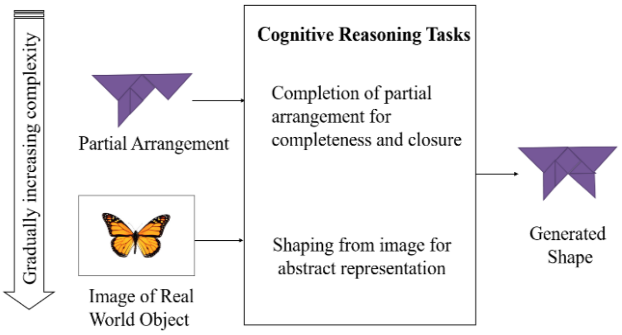

Gradually increasing complexity
Partial Arrangement
Image of Real World Object
Cognitive Reasoning Tasks
Completion of partial arrangement for completeness and closure
Shaping from image for abstract representation
Generated Shape

Figure 3. Cognitive reasoning tasks for tangram arrangement.

- Intricate cognitive process for concrete-to-abstract and abstract-to-abstract mapping, in contrast to abstract-to-concrete image generation by current generative AI models. The arrangement shape is a highly abstract representation of real-world objects, and there exists more complicated cognitive reasoning and mapping relationship.
- Strict constraints on tans in form of various polygons, as the basic element of generation with diverse shape, edge, angle, rotational variance of geometry, etc. , in contrast to the simple and rotational invariant pixel in image generated by current generative AI models.
- Strict constraints on dynamic spatial relationship among tans including the relative and absolute position, inadmissibility of gap, edge and angle matching among multiple tans, etc., in contrast to the fixed relationship among pixels in the image.

Combining with psychology concepts such as image [33], inference [34], memory [35], and imagination [36], we applied Gestalt Theory to the rearrangement model. We design cognitive reasoning tasks with gradually increasing complexity for the human cognition abilities following the gradually thinking process of human, trying to make the model to build up thinking patterns which are similar to that of human being with detail described blow and illustrated in Figure 3.

Completion of Partial Arrangement: As the initial task for abstract-to-abstract reasoning, it aims to explore the ability of automatic generation model in achieving the completeness and closure principles of Gestalt theory. Based on partial arrangement of the tangram, it will complete the entire arrangement for target object according to the characteristics of the remaining blocks. These principles state that an impression presents the most perfect form when it is consistent with its environment, and that any gap or inconsistence should be minimized as much as possible.

Shaping from Image: In this intricate task for concrete-to-abstract reasoning, the final arrangement is generated from real-world image containing real world object. This task is used to explore the ability of an automatic generation model for abstract representation and imagination.

Focusing on the challenges of autonomous arrangement of restricted abstract graphics, we firstly investigate the cognitive process of human on above tasks where human brain reflects the properties of objective objects and the relationships between objects through perception, memory, reasoning and imagination. Based on the cognitive process, we devise the Denoising Score Matching [3] centers automatic generation technology for arranging tangram, and incrementally optimize the models for the series of tasks with preprocessing, segmentation and various training policy. To support the experiments of above tasks, we take great effort to create and customize task specific datasets by collecting images from internet and integrate with selected subset of KILOGRAM dataset [2] with more than 1000 labeled instances of shapes and relevant image. Based on the datasets, we conducted a series of qualitative and quantitative experiments with specific evaluation metrics. On the basis of strict spatial constraints, this project examines how the AI system combines autonomous generation and cognitive reasoning methods to efficiently solve the problem of abstract tangram to support cognitive tasks including abstract expression, conceptual structure, and pattern continuity.

Our contributions are summarized as follows:

- We design cognitive reasoning tasks inspired by Gestalt principles with increasing cognition complexity for artificial intelligence to develop its ability of automatically cognitive reasoning.
- We collect and work out the dedicated datasets for validating the effectiveness and efficiency of the models by conducting a series of qualitative and quantitative experiments.
- We devise the Denoising Score Matching based automatic generation technology for arranging tangram, and incrementally optimize the models for the series of tasks.

II. Related Works

Generative AI: DALL-E, Midjourney, and Stable Diffusion [1] are representative examples of the current field of artificial intelligence image generation. The above systems are based on multimodal models that support both language and image. DALL-E2 is based on Transformer and VAE models, Midjourney is based on the Transformer and CLIP models, and Stable Diffusion incorporates a hybrid model of UNet and Transformer into the Diffusion process. The design goal of these systems is to enhance the accuracy of the generated image during the generation process, using language prompts without any additional constraints, and focusing on pixel-based image generation. Currently, they are unable to support the organization and creation of puzzle-specific tasks using discrete modules as units with strict spatial constraints and abstract semantic representation.

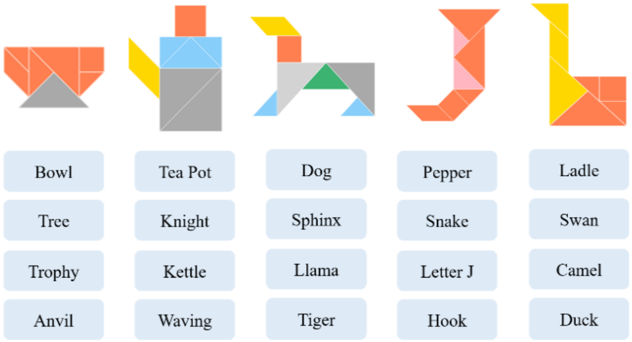


Figure 4. The examples of KILOGRAM: tangrams with multiple annotations representing the concepts imaged from the shape of tangrams.

Tangram: Tangram pictures are abstract pictures which may be used as stimuli, which were introduced by Clark and Wilkes-Gibbs [24], in various fields of experimental psychology and are classically used in the literature on cognitive processes, such as visual perception, language, and memory [23]. As a representation and abstraction of multidimensional space in two-dimensional space, tangram supports various complex arrangement methods based on mathematical rules and can even be utilized to solve or prove mathematical problems [2]. The tangram problem is being applied in some research work in the field of artificial intelligence. For example, [4] supported abstract visual research work using the KILOGRAME tangram dataset, [5] we explored the use of the tangram dataset to train pre-training models for completing low-resolution visual tasks. However, the work mentioned above does not include the automatic generation of the arrangement patterns.

Object Rearrangement: In the broader field of object rearrangement, numerous works closely related to robot planning (such as [6], [7], [8], [9]) have studied methods for object rearrangement. In the typical scene of rearrangement process, the goal of these systems is to obtain the initial state of the scene and transform it into the target state specified by the user. The related technology has reference value for this project. However, the input and output in such tasks are organized as objects of equal levels of abstraction, and the resulting arrangements have no specific semantic limitations. Therefore, the above method cannot be directly applied to tasks with abstract cognitive task requirements and strict spatial constraints for arrangement of tangram.

III. Dataset Generation

3.1 Dataset Generation for Completion of Partial Arrangement

For the cognitive reasoning tasks with gradually increasing complexity, as KILOGRAM dataset is not designed for such purpose originally and unable to meet requirements, we take great efforts to build the task specific datasets by developing the program to generate a part of it automatically as well as manually collecting images from internet and integrate with selected subset of KILOGRAM dataset.

KILOGRAM is a large collection of annotated tangrams and includes 1016 shapes of curated and digitized tangrams [4]. It is suited for generative tasks as well as instruction-following tasks. But randomly generated reference games may include ambiguities that make the models impossible to solving when KILOGRAM is used, for the same shape may have different ways of the rearrangement of the tangrams. We build our datasets based on KILOGRAM, and also used some of the existing annotations as references. The examples of KILOGRAM dataset are shown in Figure 4.

For the task of completion of partial arrangement, we generate our own dataset by keeping more than 4 tans of the tangram unchanged as the hint to ground truth with specific semantics, while disarranging others and follow a specific distribution in the arrangement space, as illustrated in Figure 5. We generate the dataset with all the 1016 instances in TANGRAM as demonstrated in Figure 5 mainly in two steps, extracting and reshaping the data from KILOGRAM, and then disarranging each tan.

A tangram is represented by a list of vertices, representing each tan's shape respectively. The shape of a tan is defined by a set of vertices, $V = \{(x_1, y_1), (x_2, y_2), \ldots\}$. We convert it to a vector that contains the information of both pose and position in the arrangement space.

$$T_i = \{t_1, t_2, \ldots t_7\}$$
$$t_{ij} = \{x_{ij}, y_{ij}, \theta_{ij}\}$$

where T_i is the i^{th} tangram of the dataset that consists of seven tans as polygons; x_{ij} are the coordinates of a tant_{ij} in the coordinate system of the arrangement space; θ_{ij} is the rotational angle of a tan, t_{ij}, indicating the initial pose.

The position of each tan of the^{th}i tangram, $[x_{ij}, y_{ij}]$, is denoted as the coordinates of its geometric centroid [16] calculated from the vertices of its shape.

$$C_{i,j,x} = \frac{1}{6A_{i,j}} \sum_{i=0}^{N-1} (x_{i,j,k} + x_{i,j,k+1})(x_{i,j,k} y_{i,j,k+1} - x_{i,j,k+1} y_{i,j,k})$$

$$C_{i,j,y} = \frac{1}{6A_{i,j}} \sum_{i=0}^{N-1} (y_{i,j,k} + y_{i,j,k+1})(x_{i,j,k} y_{i,j,k+1} - x_{i,j,k+1} y_{i,j,k})$$

where $C_{i,j,x}$ and $C_{i,j,y}$ are the coordinate of the geometric centroid of the j^{th} tan of the i^{th} tangram. A vertex is denoted by $(x_{i,j,k}, y_{i,j,k})$, which means the k^{th} vertex of that tan. N is the number of vertices of the polygon of specific tan. The last vertex $(x_{i,j,N}, y_{i,j,N})$ is equal to the first vertex $(x_{i,j,0}, y_{i,j,0})$. Notion $A_{i,j}$ is the area of the tan [15] and is defined as follows.

$$A_{i,j} = \frac{1}{2} \sum_{i=0}^{N-1} (x_{i,j,k} y_{i,j,k+1} - x_{i,j,k+1} y_{i,j,k})$$

The rotational angle, $\theta_{i,j}$, of each tan is defined as the angle different from corresponding template shape as the standard pose with the geometric centroid as the center of rotation, and the vertices coincide respectively. Each template shape states the status when that tan's rotation angle is equal to zero. Note the set of vertices of a tan's template by V. A tan in the dataset is considered to be transformed from the template V, thus is denoted by V'. So $(x_{ijk}, y_{ijk}) \in V$ and $(x'_{ijk}, y'_{ijk}) \in V'$.

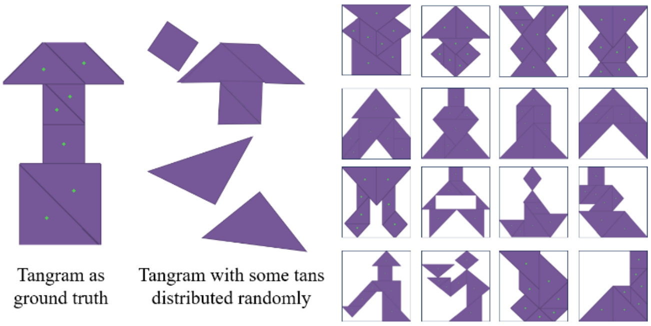


Figure 5. Example of the data set for completion of partial arrangement

$$R_\theta \begin{bmatrix} x_{ijk} \\ y_{ijk} \end{bmatrix} + \begin{bmatrix} d_x \\ d_y \end{bmatrix} = \begin{bmatrix} x'_{ijk} \\ y'_{ijk} \end{bmatrix}$$

$$R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

where R_θ is a rotation matrix in respect of θ. Since $[x_{ijk}, y_{ijk}]^\top$ and $[x'_{ijk}, y'_{ijk}]^\top$ are known, and $[d_x, d_y]^\top$ is simple to obtain. Then, we can solve $\begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$ from it.

$$\begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} = A^{-1} \begin{bmatrix} x'_{ijk} - d_x \\ y'_{ijk} - d_y \end{bmatrix} = \begin{bmatrix} \frac{x_{ijk}(x'_{ijk} - d_x)}{x_{ijk}^2 + y_{ijk}^2} + \frac{y_{ijk}(y'_{ijk} - y_{ijk})}{x_{ijk}^2 + y_{ijk}^2} \\ \frac{-y_{ijk}(x'_{ijk} - d_x)}{x_{ijk}^2 + y_{ijk}^2} + \frac{x_{ijk}(y'_{ijk} - d_y)}{x_{ijk}^2 + y_{ijk}^2} \end{bmatrix}$$

We can get a numerical solution from this. By using anti-trigonometric function, we can obtain $\theta' = arccos(\cos\theta)$.

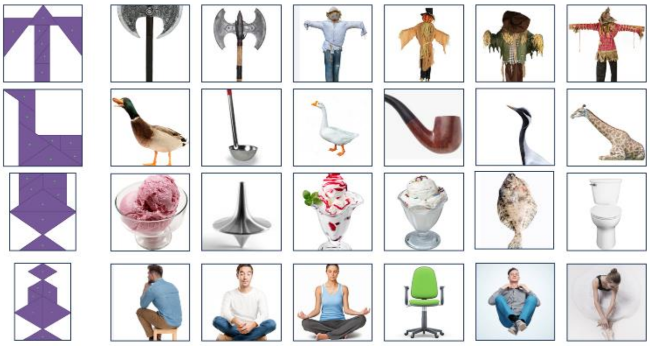


Figure 6. Examples of dataset for task of shaping from image: the first row is tangrams, and the next 6 rows are corresponding real-world images.

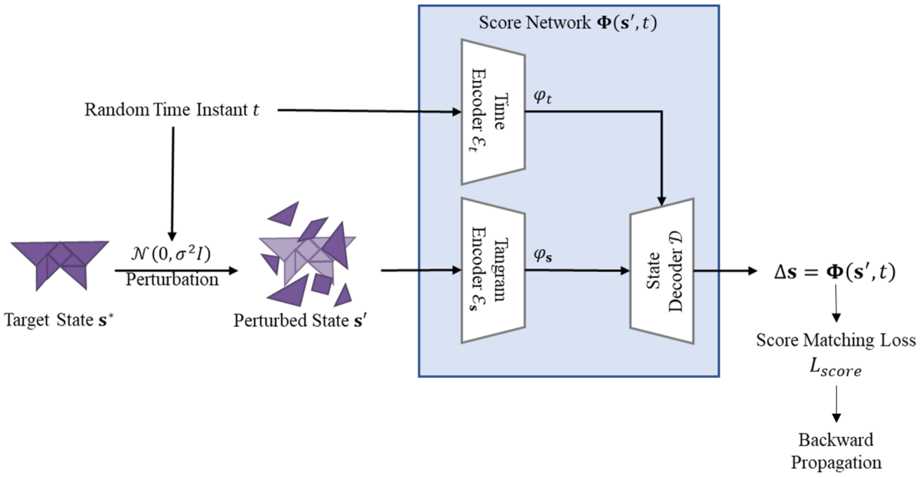

Score Network $\mathbf{\Phi}(\mathbf{s}', t)$
Random Time Instant t
Time Encoder $\mathcal{E}_t$
φ_t
$\mathcal{N}(0, \sigma^2 I)$
Perturbation
Target State $\mathbf{s}^*$
Perturbed State $\mathbf{s}'$
Tangram Encoder $\mathcal{E}_s$
φ_s
State Decoder $\mathcal{D}$
$\Delta\mathbf{s} = \mathbf{\Phi}(\mathbf{s}', t)$
Score Matching Loss
L_{score}
Backward Propagation

Figure 7. Model structure for completion of partial arrangement.

The rotation angle $\theta \in [-\pi, \pi]$, but the range of *arccosine* is $[0, \pi]$. We can do this to obtain θ.

$$\theta = \begin{cases} -\theta', & \sin\theta < 0 \\ \theta', & \sin\theta \geq 0 \end{cases}$$

3.2 Dataset Generation for Shaping from Image

For the task of shaping from image, we build the model to generate tangram arrangements based on images of real-world objects to explore the concrete-to-abstract ability for abstract representation and imagination with examples shown in Figure 6. For KILOGRAM was only designed for abstract-to-abstract mapping between textual annotation and tangram, we manually create a new dataset contains 12 tangrams with 632 images and annotations. Among the 632 images, 126 images are selected for test. The dataset includes different types of real-world objects, such as human gestures, animals, insects, buildings, letters, symbols, and other common objects.

Each tangram has 40 to 50 images, corresponding to the number of annotations it has. An annotation is a phrase that describes what the tangram looks like respectively. (e.g., the image of annotation “giraffe” should be an image with a giraffe in it.

The dataset is created in following 3 steps in general. First, we used the annotations of each tangram in KILOGRAM to search the corresponding real-world images online. In order to minimize interference during model training, we then selected 1316 proper images from more than 9740 images, with relative clear background and right pose as the corresponding tangram. In the 1316 images, we selected some objects of the similar shapes and refer to the annotations in KILOGRAM and carried out a relatively in-depth and perfected data preparation of 12 tangrams. In the final steps, we manually process the selected dataset for training and testing, transforming them to unified format and size, and tagging for automated indexing, etc. For example, we find some annotations of tangrams in KILOGRAM only looks right when the real-world image rotates at certain angle, so we adjust the rotational angle to make the object on the image fit the annotations. We proved the dataset can be used for training and testing properly, and the detail of the result is shown in Sec 5.2.

IV. Tasks and Methods

Follow the cognitive process, we design the technologies for the tasks with components for steps in cognitive process including perception, memory, reasoning and imagination, based on the generated task specific datasets.

4.1 Completion of Partial Arrangement

The goal of this task is to complete a partially arranged tangram. We adhere to the concept of completeness and closure, as mentioned in the Introduction. The input is from the partial-arranged tangram we developed. According to Gestalt Principle, completeness refers to the property of a system where all necessary elements or operations exist, ensuring that every statement is either provable or disprovable within the system [38]. We investigate the process that human complete this task. In most of the cases, the rest tans are placed onto the board one by one, to search for the most reasonable shape while trying. We project this idea to make the model to complete the task following the similar approach.

We chose the generation model based on Denoise Score Matching [3] as the foundational model for this project. It achieved good results in the field of image generation, challenging the dominance of some previous mainstream generation models, such as Generative Adversarial Network (GAN), Variational Autoencoder (VAE) in the field of generation.

The flow of the outline is shown in Figure 7. The main idea of the algorithm is to add noise to the observed data. By

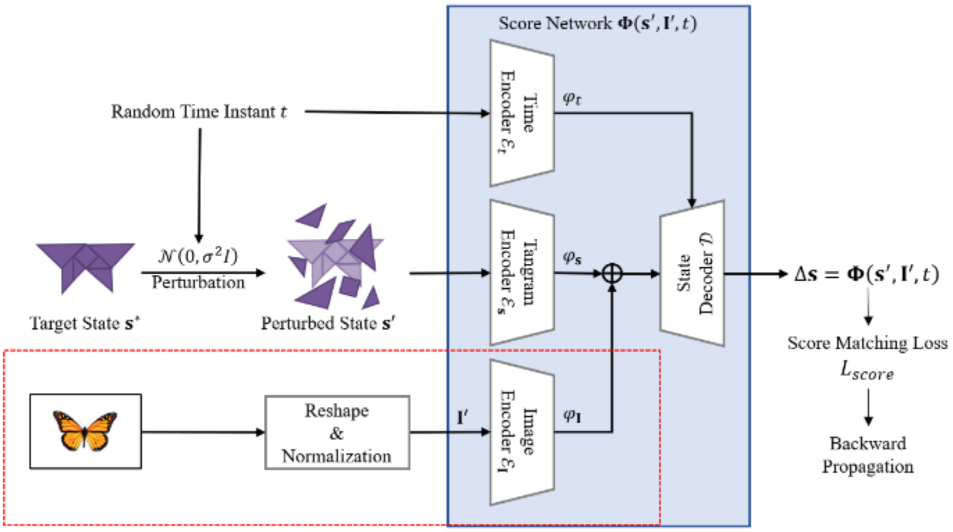

Score Network $\mathbf{\Phi}(\mathbf{s}', \mathbf{I}', t)$
Random Time Instant t
Time Encoder $\mathcal{E}_t$
φ_t
$\mathcal{N}(0, \sigma^2 I)$
Perturbation
Target State $\mathbf{s}^*$
Perturbed State $\mathbf{s}'$
Tangram Encoder $\mathcal{E}_s$
φ_s
State Decoder $\mathcal{D}$
$\Delta\mathbf{s} = \mathbf{\Phi}(\mathbf{s}', \mathbf{I}', t)$
Score Matching Loss
L_{score}
Backward Propagation
Reshape & Normalization
$\mathbf{I}'$
Image Encoder $\mathcal{E}_I$
φ_I

Figure 8. Introduce real-world object into score network with image encoder for shaping from image.

learning a scoring function on a dataset, samples that conform to the training set are obtained by sampling after noise disturbance.

The state of a tangram was previously defined as a set T. For the sake of convenience for the following definitions, we define a vector form of the state of a tangram denoted bys_i.

$$\mathbf{s}_i = [x_{i,1}, y_{i,1}, \theta_{i,1}, \dots, x_{i,7}, y_{i,7}, \theta_{i,7}]^{\mathrm{T}}$$

where $\mathbf{s}_i$ is 21-dimensional vector that describes a specific sample of tangram, such that $\mathbf{s}_i \in S$. The distribution of the entire dataset is described as $p_{target}(\boldsymbol{s})$, such that $\mathbf{s}_i \sim p_{target}(\mathbf{s})$. We assume that the model starts to rearrange the tangram randomly placed following Gaussian distribution. Thus, during the training process, we add perturbation to the target data $\mathbf{s}_i$ to get $\boldsymbol{s}_i^{'}$.

$$\boldsymbol{s}_i^{'} = \mathbf{s}_i + \sigma^2 \epsilon, \epsilon \sim \mathcal{N}(0, I)$$

The gradient of the log-likelihood$\nabla_s \, log \, p_{target}(\mathbf{s})$, in this case, means the best direction to move and rotate tans toward the target arrangement. Such mechanism can avoid us from dealing with the intractable normalization constant [17].

$$\mathbf{\Phi}(s) = \nabla_{\mathbf{s}} \, log \, p_{target}(\mathbf{s})$$

The model to estimate the gradient of the log-likelihood $\nabla_{\mathbf{s}} \, log \, p_{target}(\mathbf{s})$ is denoted by $\mathbf{\Phi}$, the score network. And, From the perturbed state of tans $\boldsymbol{s}^{'}_i$, we are having a distribution $q(\boldsymbol{s}^{'}|\boldsymbol{s}) = \mathcal{N}(\boldsymbol{s}; \boldsymbol{s}, \sigma^2 I)$. Therefore, the goal of the task is described as the following.

$$\mathbb{E}_{q(\mathbf{s}^{'}|\mathbf{s}), p_{target}(\mathbf{s})}[\|\mathbf{\Phi}(\mathbf{s}^{'}) - \nabla_{\mathbf{s}^{'}} \log q(\mathbf{s}^{'}|\mathbf{s})\|_2^2]$$

$$= \mathbb{E}_{q(\boldsymbol{s}^{'}|\boldsymbol{s}), p_{target}(s)}[\|\mathbf{\Phi}(\boldsymbol{s}^{'}) - \frac{\boldsymbol{s}^{'} - \boldsymbol{s}}{\sigma^2}\|_2^2]$$

To train the score network, we apply the method of Denoising Score-Matching (DSM) [18]. The perturbation coefficient σ^2 is related to time $t \in [0,1]$. By training with a series of time instants, the model can learn the gradient of likelihood in any extent of perturbation.

By referring to the certain tans that indicate the possible rearrangement for the whole tangram shape, our model shows its capability on finding the trend of the arrangement. For example, when the model receives a partial arrangement of the tangram shapes that resembles a butterfly, it identifies the wings of the butterfly and then determines that the most likely solution for rearranging the tangram pieces is to form a butterfly shape. Our model also applies the principles of completeness and closure to other partial tangram rearrangement problems, producing rearrangement solutions corresponding to different types of finalized rearrangement trends.

Specifying the model, it is, in this case, a function related to time, denote by $\mathbf{\Phi}(\mathbf{s}, t)$. Structurally, encoders, denoted by $\mathcal{E}$, collects information or extract features from the input; the decoder, denoted by $\mathcal{D}$, generate the score based on the outputs of encoders.

There are two encoders: $\mathcal{E}_s$, encodes the pose of the input tangram, and $\mathcal{E}_t$, encodes the time. Superficially saying, since the intensity of noise is correlated with the time in both training process and inference process, the time encoder can prompt the model the extent of disorder of the current arrangement. Input each encoder with the corresponding information, we will obtain $\varphi_s = \mathcal{E}_s(\mathbf{s})$ and $\varphi_t = \mathcal{E}_t(t)$,

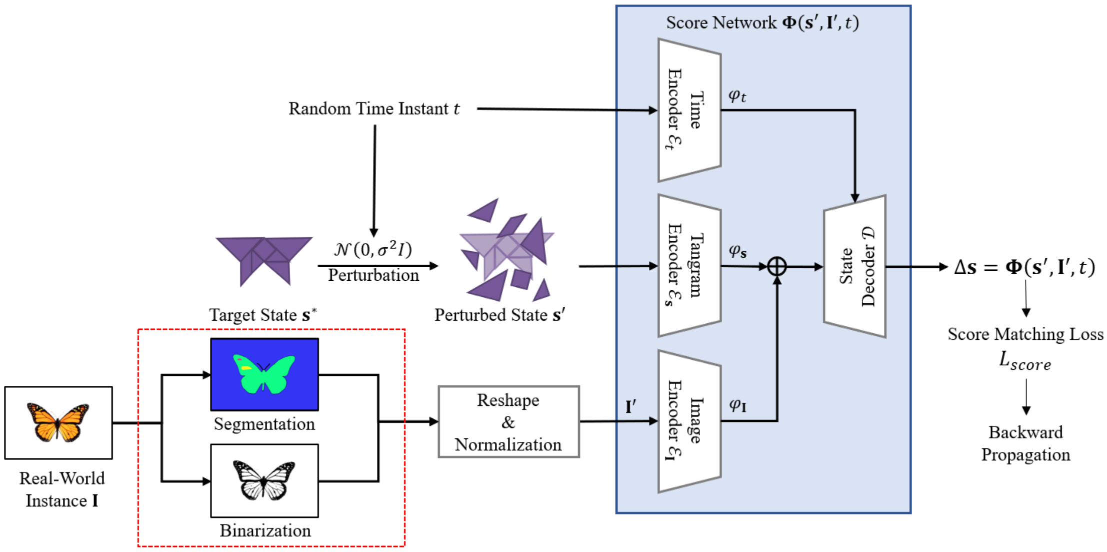

Score Network $\mathbf{\Phi}(\mathbf{s}', \mathbf{I}', t)$
Random Time Instant t
Time Encoder $\mathcal{E}_t$
φ_t
$\mathcal{N}(0, \sigma^2 I)$
Perturbation
Target State $\mathbf{s}^*$
Perturbed State $\mathbf{s}'$
Tangram Encoder $\mathcal{E}_s$
φ_s
State Decoder $\mathcal{D}$
$\Delta\mathbf{s} = \mathbf{\Phi}(\mathbf{s}', \mathbf{I}', t)$
Score Matching Loss
L_{score}
Backward Propagation
Segmentation
Binarization
Real-World Instance $\mathbf{I}$
Reshape & Normalization
$\mathbf{I}'$
Image Encoder $\mathcal{E}_I$
φ_I

Figure 9. Enable the model to concentrate on the real-world object via semantic segmentation and Otsu's binarization.

called the tangram pose embedding and the time embedding respectively.

The score, which is approximately equivalent to$\Delta\mathbf{s}$, is equal to the output of the decoder, where$\Delta\mathbf{s} = \mathcal{D}\,(\varphi_s, t)$. The time t is correlated closely with the embedding ω in the decoder. Specifically, the decoder is implemented with multiple neural layers, and the time embedding $\mathcal{E}_t$ is concatenated to the output of the last layer and, the concatenated vector, becomes the input of the next layer.

4.2 Shaping from Image

In this task with increased complexity, we provide a real-world image which is similar to the tangram, and then ask the model to work out a corresponding arrangement of the tangram. This task is built on the basics of the task of completion of partial arrangement, while adding different task specific component to achieve incrementally improved performance, including introducing image embedding into score network, concentrating on the real-world object via semantic segmentation and binarization.

4.2.1 Introducing Image Embedding into Score Network

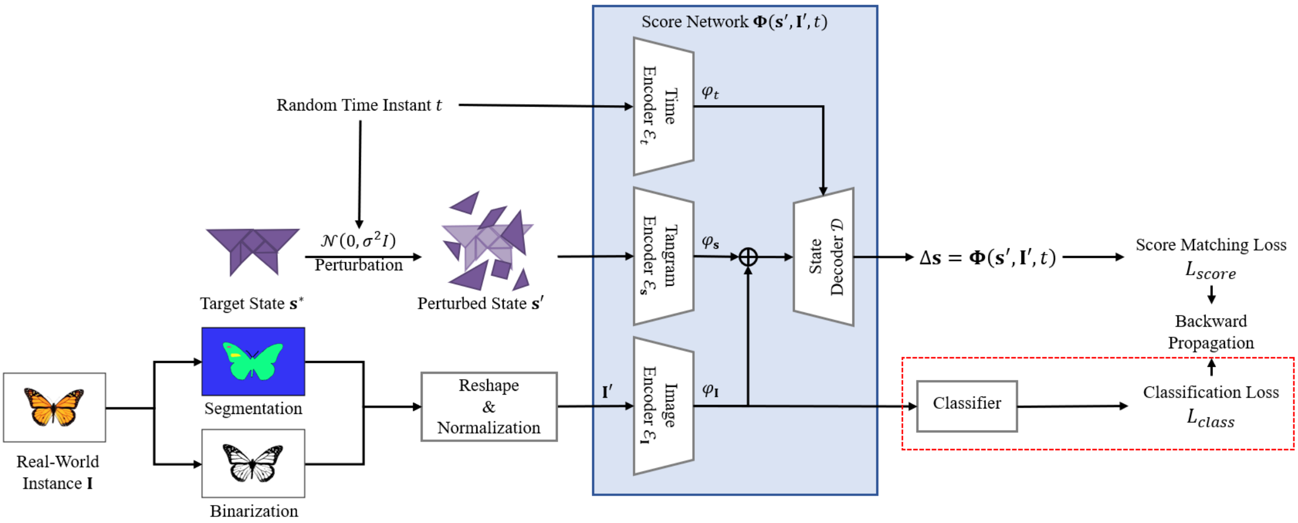
Rearranging tangrams is just the first step to exploring the cognitive ability for abstract-to-abstract mapping. In this intricate task, we would further explore the imagination and abstract representation ability for concrete-to-abstract mapping to understand real-world object in an image and correspondingly generate a tangram with the similar shape.
In the cognitive process of human beings, perception plays a key role for the above task. So, we try to integrate the model of perception into the score network. The model abstracts the shape feature of the real-world object and applies to unfamiliar objects with similar shape. The model also demonstrates the ability of inference, which is a cognitive process that derives general laws from specific instances or draws new conclusions from existing principles [39].

In the dataset, each image includes only one major semantically meaningful object as the concrete visualization of the target tangram. To achieve the goal, the network should be able to understand how the input image is shaped with a specific underlying semantic structure. Therefore, we add a convolutional network-based image encoder for introducing image embedding into the original score network as shown in Figure 8.

Formally, let an image from the dataset be represented by a multidimensional vector$\mathbf{I}$. Specifically, an image will be resized to the resolution of 224×224, with three channels, while preprocessing. Normalization will be perform after resizing, so the value of each pixel in a channel ranges from 0 to 1, which for $\forall \mathbf{e} \in \{Independent\ Base\ Vectors\}$, $0 \leq \mathbf{I}^{\mathrm{T}}\mathbf{e} \leq 1$. Preprocessed image (reshaped and normalized) is denoted by $\mathbf{I}^{'}$.

The image encoder, $\mathcal{E}_I$, is implemented by a set of convolutional layers. Each layer is consisted of the following sequential structure: C-C-P-A. Here with kernel size of 3*3 and stride of 1, notion C represents a two-dimensional convolution layer; With kernel size of 2*2 and stride of 1, notion P represents a pooling layer; analogously, notion A represents an activation layer, which we are using ReLU here.

The output of the image encoder is$\varphi_I = \mathcal{E}_I(\mathbf{I}^{'})$. With previously defined φ_s and φ_t, the output, $\Delta\mathbf{s}$, is equal to $\mathcal{D}([\varphi_s; \varphi_I], \varphi_t)$. Notion $[\varphi_s; \varphi_I]$ is equal to the concatenation of the tangram pose embedding and the image embedding.



Score Network $\Phi(\mathbf{s}', \mathbf{I}', t)$
Random Time Instant t
Time Encoder $\mathcal{E}_t$
φ_t
$\mathcal{N}(0, \sigma^2 I)$
Perturbation
Target State $\mathbf{s}^*$
Perturbed State $\mathbf{s}'$
Tangram Encoder $\mathcal{E}_s$
$\varphi_{\mathbf{s}}$
State Decoder $\mathcal{D}$
$\Delta\mathbf{s} = \Phi(\mathbf{s}', \mathbf{I}', t)$
Score Matching Loss
L_{score}
Backward Propagation
Real-World Instance $\mathbf{I}$
Segmentation
Binarization
Reshape & Normalization
$\mathbf{I}'$
Image Encoder $\mathcal{E}_I$
$\varphi_{\mathbf{I}}$
Classifier
Classification Loss
L_{class}

Figure 10. Learn the shape-aware representation via composite loss for focusing more on shape instead of semantics.

In each iteration of training, the model is fed with a batch of data in which single slice of datum consists of: (1) a 21-dimentional tangram state vector s and (2) a reshaped and normalized image. The training and sampling process of this task is similar to that of the completion of partial arrangement described in section 4.1.

We are not using pretrained models like VGG nor ResNet which had been trained on either ImageNet[29] or COCO[30] in our task. As mentioned previously, our goal is to enable the tangram arranging model to imitate the shape of the object in an image. Clearly, information of structure and shape are crucial for the model to perform this task. On the other hand, classification labels provided by ImageNet and COCO are for semantics and can't meet the requirement of our tasks.

4.2.2 Concentrating on The Real-World Object via Semantic Segmentation and Otsu's Binarization

Arranging tangrams to mimic the shape of the corresponding the real-world object in the input image is theoretically an abstraction of its shape consist of one or more parts. Nevertheless, real-world images include extra information that is not essential to abstract the shapes of the target objects, including diverse colors, background, light effects. Such information may introduce interference into the abstracted features and furthermore the image embedding as input to score network during both training and inference process.

Due to above reason, and also for the purpose of improving robustness of the model, we try two methods separately to achieve that goal to avoid disturbance of the additional information and enable the model concentrating more on the geometric shape, as shown in Figure 9.

Based on the assumption that the parts of target object are connected seamlessly as the most prominent area in the image, we utilize binarization to process the image for either removing the background objects or isolating them from the target object and neglecting them in further steps. To achieve the best and similar performance on as much varied images as possible, we used Otsu's binarization, which could choose an optimal global threshold to split greyscale pixels evenly into two parts. In this way, the foreground objects of each image could be extracted from the background, and most unrelated details are hidden—except for the approximate shape of the object in the image.

Another method we try is semantic segmentation that can provide extra shape information, the parts of the object in an image. We utilize Segment Anything [19] service with pretrained model which can perform segmentation on any images and does not mask parts with fixed labels. So, images in the dataset will not be constrained on how the segmentation model work, and we do not need to label the segments.

Segment Anything generate a list of masks. We cast each masked part into a unique color for purposes of training and visualization. The casted colors are evenly distributed in the HSL color space and normalized before sending to the model. The training and sampling processes are similar to the task of completion of partial arrangement described in 4.1.

4.2.3 Learning Shape-Aware Representation

To improve the performance of the model, we add a shape classification model in parallel to the score network for further adjustment of gradient during the training process and making the model concentrate more on the shape of the real-world object in the input image, as shown in Figure 10. This is done by adding full-connection layers for the classification of shape behind the convolution layers for feature extraction. The objective of classification is to infer which tangram in the dataset is the most likely the abstraction of input image.

Each tangram in the dataset is labeled sequentially with an index (i.e., subscript i as previously used). And, since a set of images with shapes that are similar to a tangram is paired to that tangram, those images are labeled with the same index as the tangram. In our dataset, images in each group vary semantically, and we also eliminated the differences in color and brightness with the method discussed in 4.2.2. Thus, the most impactful factor to the classifier is the shape of object in the image.

Let input x_i be an image with shapes similar to the^{th}i tangram T_i. Then, the ground truth$y_i = 1$, while $\forall k \neq i$, $y_k = 0$. Denote the classifier by$c(x)$. Let$z = c(x_i)$, z_k is the predicted probability of whetherx_i shapes is similar toT_i. The loss function will thus be defined as follows.

$$L_{class} = -\sum_{k=1}^{N} y_k \ln z_k$$

However, while the convolution layers are for both shape classification and image feature extraction as a part of the score network, the gradient that propagated from the score will also be considered. We denote the image encoder by $\mathcal{E}_I(\mathbf{I})$ in sction 4.2.1, and the rest part of the score network by$\hat{\mathbf{\Phi}}(\mathbf{s}, t)$, and the classifier by $c(\varphi_I)$. Since $\varphi_I = \mathcal{E}_I(\mathbf{I})$, $L_{score} = Loss_{score}(\hat{\mathbf{\Phi}}(\varphi_I))$ and $L_{class} = Loss_{class}(c(\varphi_I))$. The gradient, of the weight w_f, for the last layer of $\mathcal{E}_I(\mathbf{I})$ is calculated as follows.

$$\frac{\partial L_{score}}{\partial w_f} + \frac{\partial L_{class}}{\partial w_f} = \frac{\partial \mathcal{E}_I(x)}{\partial w_f}\left(L_{score}^{'} \cdot \hat{\mathbf{\Phi}}^{'} + L_{class}^{'} \cdot c^{'}\right)$$

The gradient of bias b_f of the last layer of $\mathcal{E}_I(x)$is similar to the gradient ofw_f.

V. EXPERIMENT

5.1 Evaluation Methods

Based on the task specific datasets generated, we conduct a series of experiments to evaluate the experimental results separately from both quantitative and qualitative perspectives, including ablation study for the models of shaping from image.

5.1.1 Quantitative Evaluation Metrics

For the similarity between the generated tangram and ground truth are reflected and impacted by the displacement and rotation of all the tans, we define Displacement Error (*DE*), Rotation Error (*RE*) and Comprehensive Error (*CE*) as evaluation metrics for quantitative measurement of the similarity.

For the convenience of defining these errors, we define vectors of displacement, rotation, and pose, the comprehensive indicator including both displacement and rotation， independently for each tan of a tangram. For the displacement, define $\mathbf{d}_{ij} = [x_{ij} \quad y_{ij}]^{\mathsf{T}}$. The rotation, precisely, is indirectly indicated to avoid circumstances which tans' rotation is visually similar (e.g., 179° and $-179°$, visually 2° error, but their L2 distance is358°). So, we define $\mathbf{r}_{ij} = [\cos\theta_{ij} \quad \sin\theta_{ij}]^{\mathsf{T}}$, named as rotation indicator. For the pose, define $\mathbf{p}_{ij} = [x_{ij} \quad y_{ij} \quad \cos\theta_{ij} \quad \sin\theta_{ij}]^{\mathsf{T}}$.

Then, define Displacement Error (*DE*) for a batch of tangrams as the average of the sum of the L2 distances of displacements between each generated tan and the ground truth over N test samples.

$$DE = \frac{1}{7N}\sum_{i=1}^{N}\sum_{j=1}^{7}\|\boldsymbol{d}_{ij}^{GT} - \boldsymbol{d}_{ij}\|_2$$

Define Rotation Error (*DE*) for a batch of tangrams to as the average of the sum of the L2 distances of rotation indicators between each generated tan and the ground truth over N test samples.

$$RE = \frac{1}{7N}\sum_{i=1}^{N}\sum_{j=1}^{7}\|\boldsymbol{r}_{ij}^{GT} - \boldsymbol{r}_{ij}\|_2$$

Define Comprehensive Error (*CE*) for a batch of tangrams as the average of the sum of the L2 distances of poses between each generated tan and the ground truth over N test samples.

$$CE = \frac{1}{7N}\sum_{i=1}^{N}\sum_{j=1}^{7}\|\boldsymbol{p}_{ij}^{GT} - \boldsymbol{p}_{ij}\|_2$$

For two tasks, we apply above evaluation metrics to evaluate the efficiency of the final models. For the models for the task of shaping from image, we also conduct the ablation experiments, removing components from the augmented models, keeping the rest of the model, data pre-processing methods, and training hyperparameters unchanged, to verify the effectiveness of the incrementally optimized mechanisms. This comprehensive assessment approach enables a meticulous appraisal of the effectiveness and accuracy of the experimental procedures, shedding light on the intricate dynamics governing the arrangement of tans in the space.

5.1.3 Baseline Models

For the purpose of comparison, we design the straightforward models as the baseline models with structure shown in Figure 11 and 12. The baseline models directly learns what state to

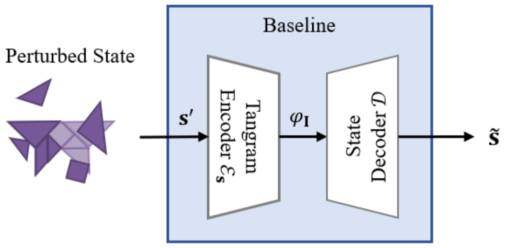


Figure 11 Baseline model for completion of partial arrangement.

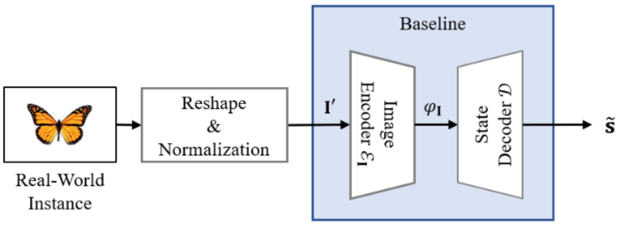


Figure 12. Baseline model for shaping from image.

arrange the tangram. Formally, we define the output of the baseline model as the following.

$$\hat{\mathbf{s}} = [x_{i,1}, y_{i,1}, \cos\theta_{i,1}, \sin\theta_{i,1}, \ldots, x_{i,7}, y_{i,7}, \cos\theta_{i,7}, \sin\theta_{i,7}]$$

here s' is a 28-dimensional vector. By calculating $\theta \in [-\pi, \pi]$ from the numerical values of$\cos\theta$ and $\sin\theta$, the output of model $\hat{\mathbf{s}}$ will be converted to the previously defined 21-dimensional vector, the state of the tangram,$\mathbf{s}$.

For the task of shaping from image, we use the real-world images as the input of the baseline model. The images are firstly passed through convolution layers that have the same structure as what we used for our model. The extracted feature will be sent into a set of full-connection layers to compute the output $\hat{\mathbf{s}}$. The size of parameters and the training method of the baseline model are as same as the proposed model.

5.3. Qualitative Evaluation

We further evaluate the semantics of the generated tangram via observation by human tester. From the temporal perspective, we also visualize the process of generation. It is taken as a supplementary to overcome the limitations of the corresponding evaluation in *DE, RE* and *CE*.

5.2 Experiment Results

5.2.1 Evaluation Results for the Task of Completion of Partial Arrangement

For the task of completion of partial arrangement, we conduct the comparison experiment between our model and the baseline model based on the generated dataset. The experiment result in shown in Table 1. Our model, Score Network, demonstrates the corresponding ability for abstract-to-abstract mapping with performance significantly better than that of the baseline model, proving the effectiveness and efficiency.

Table 1. Result of Comparison Experiment for Completion of Partial Arrangement

Model	***CE***	***DE***	***RE***
Baseline	0.13341	0.70170	0.72067
Score Network	0.01199	0.03828	0.04361

5.2.2 Ablation Study for the Task of Shaping from Image

An ablation experiment was meticulously orchestrated on the foundational score network, involving the systematic manipulation of various features. These manipulations encompassed the selective removal and integration of diverse elements, including image encoder, semantic segmentation and binarization, and composite loss. The DE RE and CE metrics for the various algorithms are shown in Table 2. Where the proposed approach shown in Figure 9, with all above features integrated, achieves overall better results compared to the baseline model.

- *DE*: The proposed model achieves the lowest error of 0.0170. It means the model moved all tans on mostly correct positions.
- *RE*: The proposed model achieves the lowest error of 0.0261. This means it turns tans to the mostly correct orientation.
- *CE*: The proposed model achieves the lowest error of 0.0261 in total, which means its arrangement is mostly correct and accurate.

Table 2. Result of Ablation Study for Shaping from Image.

Model	***CE***	***DE***	***RE***
Baseline	0.4958	0.1085	0.4674
Score Network with Image Encoder	0.0542	0.0222	0.0494
Score Network with Image Encoder and Binarization	0.0490	0.0199	0.0388
Score Network with Image Encoder and Semantic Segmentation	0.0380	0.0173	0.0289
Score Network with Image Encoder, Semantic Segmentation and Composite Loss	**0.0261**	**0.0170**	**0.0199**

It shows that the proposed approach can effectively and efficiently improve the accuracy of shaping from image task, with strong ability of concrete-to-abstract mapping demonstrated. As well, removing the image encoder, semantic segmentation and binarization, and composite loss mechanisms negatively affects all 3 metrics. This indicates that the structures we added plays a key role for improving the accuracy.

5.2.3 Qualitative Evaluation Results of Shaping from Image

This task requires a higher level of cognitive reasoning capability. The model must learn the features of various real-world objects and memorize their characteristics. We visualize the process of tangram generation with intermediate and final results as illustrated in Figure 13, where the changes of pose indicated by both the geometric center of each tan and

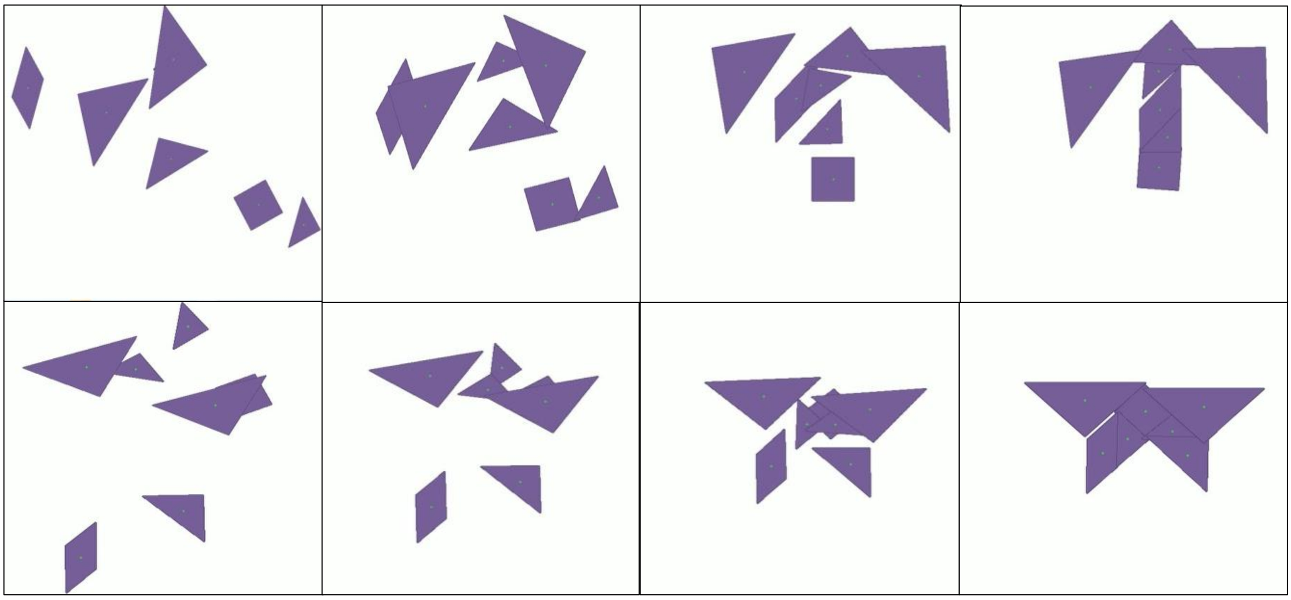


Figure 13. Visualization of the tangram generation process with randomly distributed tans in the initial state approaching the target shape.

the rotational angle could be observed. In the initial state, the tans are distributed randomly in the arrangement space. During the arrangement process, the pose of the tans keeps changing simultaneously until the final state is achieved with minimized difference from the tangrams as ground truth.

We observe multiple images of a different real-world objects and corresponding tangram as the ground truth, comparing with the arrangement generated by the baseline and proposed models, with the example shown in Figure 14. For the proposed model, the pose, gaps between tans, and the overall shape of the tangram generated closely approximate the ground truth with minimal error. The semantics were clear and easy to understand, and can be easily associated to the real-world objects. Still using the example of the butterfly-shaped tangram. We provide the model with various types of butterflies for training, allowing it to learn the patterns and features of butterfly shapes. For testing the model, we provide image with objects different from butterflies but share similar shape features, such as a table with a long board and an outstretched short table leg. The model then gives an arrangement of tangrams that is similar to the butterfly-shaped tangram, as it identifies the similarities between the shape of a butterfly and a table. This phenomenon indicates that the model abstracts the shape feature of the butterfly and applies the learned shape to unfamiliar objects. The model also demonstrates the ability of cognitive reasoning that derives general laws from specific instances or draws new conclusions from existing principles. The observation proved that the proposed model could achieve the best result with minimal difference from the ground truth.

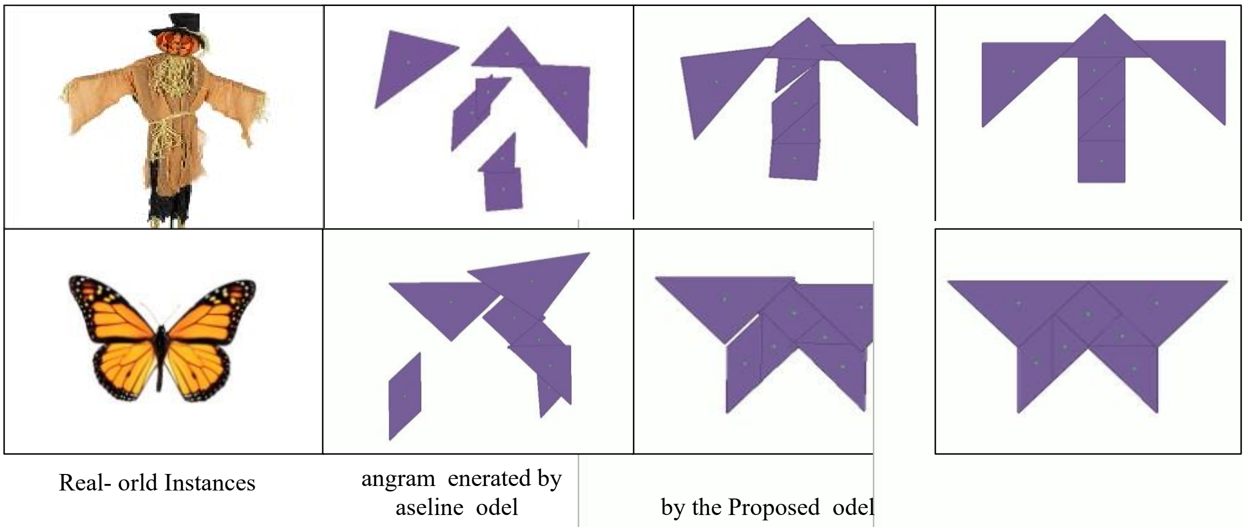


Figure 14. Comparison among images with real-world objects, tangrams generated by the baseline model, our model and as ground truth

VI. Conclusion

In this project, we utilize tangram as the experiment platform to investigate the autonomous mechanism and cognitive reasoning ability with rare prior work found. Comparing to the tasks of existing generative AI models, we summarize the challenges to autonomous arrangement of tangram including concrete-to-abstract and abstract-to-abstract mapping, strict constraints on tans in form of various polygons and spatial relationship among them. To meet the challenges, firstly we design cognitive reasoning tasks with gradually increasing complexity for the human cognition abilities including completeness and closure, and abstract representation. Then we investigate the human cognitive process for above tasks including perception, memory, reasoning and imagination, and explore the automatic generation method based on Denoising Score Matching centered models with incrementally optimized model structure and training policy. For training and testing of the above cognitive reasoning tasks, we make great efforts to create and customize task specific datasets by collecting images from internet and integrate with selected subset of KILOGRAM dataset. We conducted both the qualitative and quantitative experiments to evaluate both effectiveness and efficiency of the method. The experiment results demonstrate that our method could achieves the goals of the designed cognitive reasoning tasks, with good performance on automatic generating traits of tangram either from shapes or real-world images, underscoring its potential for advancing cognitive capability of generative AI in extended fields.

VII. Acknowledgements

We would like to express our sincere thanks to Dr. Fangwei Zhong from Peking University and Dr. Yu Chen Zhou from Talent Institute Beijing No.101 High School as the supervisors of our project.

VIII. References

[1] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, Hong Lin. AI-Generated Content (AIGC): A Survey. arXiv:2304.06632

[2] Sarah Sophie Pohl, Christian Richter. The complete characterization of tangram pentagons. arXiv:2006.09698.

[3] Song Y, Ermon S. Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems, 2019, 32.

[4] Ji A, Kojima N, Rush N, et al. Abstract Visual Reasoning with Tangram Shapes. arXiv:2211.16492 [cs.CL]. 2022.

[5] Yizhou Zhao, Liang Qiu, Pan Lu, Feng Shi1, Tian Han, Song-Chun Zhu. Learning from the Tangram to Solve Mini Visual Tasks. arXiv:2112.06113 [cs.CV]. 2021. https://arxiv.org/pdf/2112.06113

[6] Wei Q A, Ding S, Park J J, et al. Lego-net: Learning regular rearrangements of objects in rooms. arXiv:2301.09629.

[7] Dhruv Batra, Angel X Chang, Sonia Chernova, Andrew J Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, et al. Rearrangement: A challenge for embodied ai. arXiv:2011.01975.

[8] Ohad Ben-Shahar and Ehud Rivlin. Practical pushing planning for rearrangement tasks. IEEE Transactions on Robotics and Automation, 14(4):549–565, 1998.

[9] Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, Tingfa Xu. LayoutGAN: Generating Graphic Layouts with Wireframe Discriminators. arXiv:1901.06767.

[10] Ramesh, Aditya, et al. Zero-shot text-to-image generation. International Conference on Machine Learning. PMLR, 2021.

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752

[12] Robert J. Sternberg; Karin Sternberg. Cognitive Psychology (6th ed.). Belmont, CA: Cengage Learning. 2012. ISBN 978-1-133-31391-5.

[13] Wu, M., Zhong, F., Xia, Y., & Dong, H. (2022). TarGF: Learning Target Gradient Field to Rearrange Objects without Explicit Goal Specification. arXiv.2209.00853

[14] Nicholas FitzGerald, Yoav Artzi, and Luke Zettlemoyer. 2013. Learning distributions over logical forms for referring expression generation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1914–1925.

[15] Wikipedia contributors. (2023). Centroid. Wikipedia. https://en.wikipedia.org/wiki/Centroid

[16] Paul Bourke. (1988). Calculating The Area And Centroid Of A Polygon.

[17] Song, Y. (2021, January 9). How to train your Energy-Based Models.

[18] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In International Conference on Learning Representations, 2021.

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, Ross Girshick. Segment Anything. arXiv:2304.02643

[20] Abstract representation. (2012). In Springer eBooks (p. 30). https://doi.org/10.1007/978-1-4419-1428-6_3014

[21] Renavitasari, I. R. D., & Supianto, A. A. (2018). Educational Game For Training Spatial Ability Using Tangram Puzzle. IEEE. https://doi.org/10.1109/siet.2018.8693164

[22] Dennis, I. (2013). Human Abilities. https://doi.org/10.4324/9780203774007

[23] Fasquel, A., Brunellière, A., & Knutsen, D. (2022). A modified procedure for naming 332 pictures and collecting norms: Using tangram pictures in psycholinguistic studies. Behavior Research Methods. https://doi.org/10.3758/s13428-022-01871-y

[24] Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. Cognition, 22(1):1–39.

[25] Ayaz, H., Shewokis, P. A., Izzetoglu, M., Cakir, M., & Onaral, B. (2012). Tangram solved? Prefrontal cortex activation analysis during geometric problem solving. IEEE. https://doi.org/10.1109/embc.2012.6347022

[26] Kmetova, M., & Lehocká, Z. N. (2021). Using Tangram as a Manipulative Tool for Transition between 2D and 3D Perception in Geometry. Mathematics, 9(18), 2185. https://doi.org/10.3390/math9182185

[27] Rizki Diaz, I., Supianto, A. & Tolle, H. (2018). Log Data Analysis of Player Behavior in Tangram Puzzle Learning Game. International Association of Online Engineering. https://www.learntechlib.org/p/207197/

[28] Huang, Y. (2019). Tangram: Bridging Immutable and Mutable Abstractions for Distributed Data Analytics. USENIX. https://www.usenix.org/conference/atc19/presentation/huang

[29] Krizhevsky, A. (2012). ImageNet Classification with Deep Convolutional Neural Networks. https://proceedings.neurips.cc/ paper/2012.html

[30] Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in context. In Lecture Notes in Computer Science (pp. 740–755). https://doi.org/10.1007/978-3-319-10602-1_48

[31] Mental Representation (Stanford Encyclopedia of Philosophy). (2020, January 21). https://plato.stanford.edu/entries/mental-representation/

[32] Meyers, R. A. (2009b). Encyclopedia of Complexity and Systems Science. In Springer eBooks. https://doi.org/10.1007/978-0-387-30440-3

[33] Encyclopedia of Psychology and Religion. (2014). In Springer eBooks. https://doi.org/10.1007/978-1-4614-6086-2

[34] Blanchard, E., Frasson, C., & Lajoie, S. P. (2012). Encyclopedia of the Sciences of Learning. In Springer eBooks. https://doi.org/10.1007/978-1-4419-1428-6

[35] Smirnov, A. A. (1973). Problems of the psychology of memory. In Springer eBooks. https://doi.org/10.1007/978-1-4684-1968-9

[36] Jalobeanu, D., & Wolfe, C. T. (2022). Encyclopedia of Early Modern Philosophy and the Sciences. In *Springer eBooks*. https://doi.org/10.1007/978-3-319-31069-5

[38] Completeness - Psychology Dictionary of Arguments. (n.d.). https://philosophy-science-humanities-controversies.com/listview-list-psychology.php?concept=Completeness

[39] Inference. (2008). In Springer eBooks (p. 1947). https://doi.org/10.1007/978-3-540-29678-2_2421

APPENDIX

Source of the selected topic and research background

The topic of this project comes from the combination of our interests and the current hotspot of AI research. It is further exploration and expansion of the research in autonomy and cognitive reasoning of generative AI.

The work and contribution of each team member(s)

The project is completed by Qiaosi Lei and Rui Wang under the guidance of the supervisors. Qiaosi Lei is responsible for the part of models of cognitive process. Rui Wang is responsible for dataset generation and model evaluation.

The relationship between the supervisors and the student, the role supervisors played in the process of writing thesis, and whether the tutoring is paid

Dr. Fangwei Zhong is the off-campus supervisor of the Branch AI Laboratory of Center on Frontiers of Computing Studies Peking University, Talent Institute, Beijing No. 101 High School. Dr.Yuchen Zhou is the supervisor of the Branch AI Laboratory of Center on Frontiers of Computing Studies Peking University, Talent Institute, Beijing No. 101 High School. We are project team members of the laboratory.

Research completed with the assistance of others

The project is completed independently under the guidance of the supervisors.

Team Profile

Qiaosi Lei: International Humanities Experimental Class, grade 12, Beijing No. 101 High School. He won following science related awards in middle/high school:

- Second Prize, Yau High School Science Award – Computer 2022 - China division
- Bronze Award of Rhino-Bird Science Talent Development Program 2022 organized by Tencent Company and Tsinghua University.
- Silver Award in the USAD SEED North China Reginal Competition

He learned deep learning, linear algebra and calculus etc. via self-learning and completed following AI related projects:

- Construction and Application of Converged Virtual and Reality Environment for Embodied AI: Conducted research to establish an intelligent agent training environment and devised improved techniques encompassing three-dimensional reconstruction using ORB-SLAM3, articulated part detection, two-dimensional semantic segmentation, and three-dimensional projection.

- Computer Vision - Object Detection：Chinese Academy of Sciences Intelligent Science Innovation Talent Development Program 2020， acquired knowledge in machine learning and deep learning, trained a VGG16 model and achieved a precision rate of 91% on the dataset provided by the mentor.

Rui Wang: International Accelerated Class, grade 11, Beijing No. 101 High School. She won following science related awards in middle/high school:

- First Prize, CTB Global Youth Innovation Challenge 2023 Global Finals hosted in Harvard University.
- Second Prize in the Australian Mathematics Competition.

She also participated in a research project on mechanical arm rigid body poses and co-authored the paper "Measurement and Analysis of Rigid Body Poses in Three-Dimensional Space." And she is proficient in PyTorch and deep learning models, with development capabilities in Python, Java, C++, and MATLAB.

Supervisor Profile

Fangwei Zhong: Ph.D., a Boya Postdoctoral Researcher at Peking University, working with Prof. Song-Chun Zhu. Before that, he received Ph.D in Computer Science from EECS, Peking University, supervised by Prof. Yizhou Wang, and received B.Sc in Communication Engineering from Beijing Jiaotong University. His current research interests are robot learning, multi-agent learning, and computer vision, particularly in building embodied agents with physical and social common sense.

Yuchen Zhou: Ph.D., certificated research fellow, supervisor of the Branch AI Laboratory of Center on Frontiers of Computing Studies Peking University, Talent Institute, Beijing No. 101 High School. He is a senior member of the ACM and IEEE, and former member of Technical Committee, Embedded System Society, China Computer Federation. With 20 years of technical innovation experience in IBM, he served as senior research manager of AI perception in IBM Research China, a member of IBM Academy of Science and Technology, IBM Master Inventor, chair of technical committee and patent review committee of the center, etc. He won 3 outstanding technical achievement awards, published 1 book, participated and contributed to 2 international standards, obtained around 50 international patents and published more than 30 papers.