Regression Residual Reasoning with Pseudo-labeled Contrastive Learning for Uncovering Multiple Complex Compositional Relations

Chengtai Li[1,2] , **Yuting He**[1,3] , *__Jianfeng Ren__[1,4] , **Ruibin Bai**[1,4] , **Yitian Zhao**[2] , **Heng Yu**[1] and **Xudong Jiang** [5]

[1]Digital Port Technologies Lab, School of Computer Science, University of Nottingham Ningbo China
[2]Cixi Institute of Biomedical Engineering, Ningbo Institute of Industrial Technology, Chinese Academy of Sciences
[3]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[4]Nottingham Ningbo China Beacons of Excellence Research and Innovation Institute, University of Nottingham Ningbo China
[5]School of Electrical & Electronic Engineering, Nanyang Technological University
{chengtai.li, yuting.he, jianfeng.ren, ruibin.bai, heng.yu}@nottingham.edu.cn, yitian.zhao@nimte.ac.cn, exdjiang@ntu.edu.sg

Abstract

Abstract Visual Reasoning (AVR) has been widely studied in literature. Our study reveals that AVR models tend to rely on appearance matching rather than a genuine understanding of underlying rules. We hence develop a challenging benchmark, Multiple Complex Compositional Reasoning (MC2R), composed of diverse compositional rules on attributes with intentionally increased variations. It aims to identify two outliers from five given images, in contrast to single-answer questions in previous AVR tasks. To solve MC2R tasks, a Regression Residual Reasoning with Pseudo-labeled Contrastive Learning (R3PCL) is proposed, which first transforms the original problem by selecting three images following the same rule, and iteratively regresses one normal image by using the other two, allowing the model to gradually comprehend the underlying rules. The proposed PCL leverages a set of min-max operations to generate more reliable pseudo labels, and exploits contrastive learning with data augmentation on pseudo-labeled images to boost the discrimination and generalization of features. Experimental results on two AVR datasets show that the proposed R3PCL significantly outperforms state-of-the-art models.

1 Introduction

Visual recognition tasks [Li *et al.*, 2022; Jiang *et al.*, 2022; Ding *et al.*, 2022; Zhang *et al.*, 2023; Zhang *et al.*, 2024] focus on categorizing images into classes, while humans can not only understand the appearance but also reason the underlying rules for real-world objects [Song *et al.*, 2023; Małkiński and Mańdziuk, 2023]. Abstract Visual Reasoning (AVR) has emerged recently to research the induction, summarization, and application of visual rules, which discovers the underlying rules based on contextual information, *e.g.*, Raven's Progressive Matrices (RPM) [Zhang *et al.*, 2019], Odd-one-out [Zerroug *et al.*, 2022], Abstraction and Reasoning Corpus [Chollet, 2019], etc. Most AVR methods [He *et al.*, 2023; Xu *et al.*, 2023; Yang *et al.*, 2023] contain two main building blocks: a visual perception module to understand the image scenes and an analogical reasoning module to conduct logical reasoning based on the perceived visual attributes.

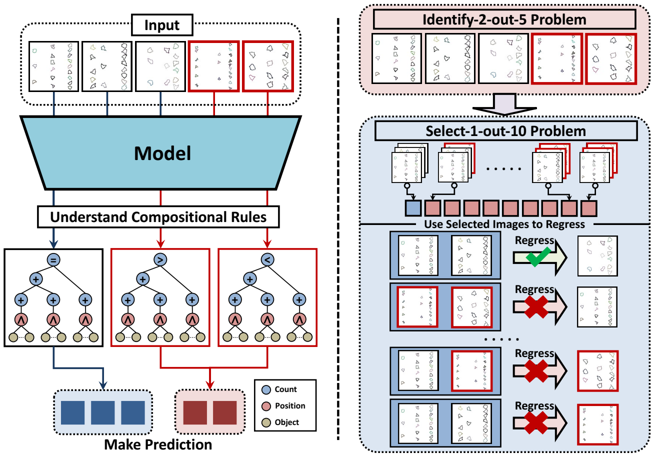

Figure 1: **Left:** A sample of proposed MC^{2}R task, which identifies two outliers among five images by analyzing the subtle differences in underlying compositional rules. **Right:** The proposed R^{3}M, which transforms the identify-2-out-5 problem into a select-1-out-10 problem, and solves it using iterative regression residual reasoning.

AVR tasks were initially constructed using simple regular shapes such as polygons and lines to assess human's abstract visual reasoning [Zhang *et al.*, 2019]. The underlying rules are intentionally made simple so that humans can discover the rules from limited images, *e.g.*, a question panel of 3 × 3

*Corresponding author.

images in RAVEN [Zhang *et al.*, 2019]. In real applications, the patterns we humans comprehend often arise from intricate compositions of multiple rules on complex scenes. Recently, Compositional Visual Relation (CVR) [Zerroug *et al.*, 2022] has been developed to simulate diverse compositional relations in the real world, in which 103 compositional rules are much more comprehensive and complex than rules in other datasets [Zhang *et al.*, 2019]. Furthermore, instead of regular shapes, the CVR dataset is composed of irregular curved shapes, posing additional challenges for visual perception.

The emergence of CVR datasets partially solves the problems in AVR tasks, but there are still some challenges. 1) As most existing datasets are composed of simple rules or regular shapes, the model may complete the tasks via surface matching of image appearance, instead of uncovering the underlying rules. 2) The irregular curved shapes in CVR [Zerroug *et al.*, 2022] and the regular shapes in RAVEN [Zhang *et al.*, 2019] and PGM [Barrett *et al.*, 2018] have limited variations in attributes, which not only leads to the problem of appearance matching but also artificially simplifies the tasks. 3) The problem formulation of AVR tasks is relatively simple, *e.g.*, the 3×3 question panel in RAVEN [Zhang *et al.*, 2019] and one outlier out of four images in CVR [Zerroug *et al.*, 2022].

To tackle these problems, a new AVR task of **M**ultiple **C**omplex **C**ompositional **R**easoning (**MC^{2}R**) is proposed. As shown in Figure 1, given three images following a compositional rule, and two outliers following slightly different compositional rules, the target is to identify the two outliers. The proposed MC^{2}R addresses the challenges of existing AVR tasks in three aspects. 1) Outliers look similar to normal images as the underlying rules only differ slightly, which avoids the problem of potential shortcuts in appearance matching, and enforces AVR models to focus on the subtle rule differences. 2) The attribute variations across images are intentionally increased, to avoid possible appearance matching and greatly challenge identifying outliers. 3) Compared to identifying one outlier out of four images in CVR [Zerroug *et al.*, 2022], the proposed MC^{2}R task is much more challenging, requiring identifying two outliers from five images. As shown in experiments, the MC^{2}R task greatly challenges existing AVR models and pushes the frontier of AVR research.

To effectively detect the subtle rule differences, **R**egression **R**esidual **R**easoning with **P**seudo-labeled **C**ontrastive **L**earning (**R^{3}PCL**) is proposed, containing a **R**egression **R**esidual **R**easoning **M**odule (**R^{3}M**) for logic reasoning and a backbone of ResNet-50 with **P**seudo-labeled **C**ontrastive **L**earning (**PCL**) for visual perception. Due to the tiny differences in compositional rules, the limited number of images per question panel, and the large attribute variations, it is difficult to directly identify the outliers. As shown in Figure 1, we transform the original problem of identify-2-out-5 into the problem of determining the only selection of three normal images from $C_5^3 = 10$ possible selections. The key motivation here is that by selecting the three images that follow the same compositional rule, we can regress any one image with minimal error by using the other two, while the regression error for other selections containing outliers will be larger. Inspired by this, a **R**egression **R**esidual **R**easoning **B**lock (**R^{3}B**) is proposed, where one image is regressed using the other two through a series of regression networks, and the regression error is minimal when three images follow the same compositional rule. Then, a set of R^{3}Bs are cascaded to form the proposed R^{3}M. As a result, the proposed method could better capture the tiny differences in compositional rules and more accurately identify the outliers.

Lastly, the increased attribute variations across images and the tiny rule differences pose great challenges to visual perception. To tackle the distribution discrepancy between training and test data due to large sample variations, the proposed PCL incorporates a data augmentation of image rotation of 90°, 180°, and 270° to improve the generalizability of the extracted features. More importantly, the proposed PCL incorporates a contrastive learning strategy to better distinguish normal images and outliers. After pseudo-labeling the five images, the more representative outlier is selected through a set of min-max operations, and the most representative pair of normal images is identified as the most similar pair. The proposed PCL then simultaneously maximizes the similarity between the pair of normal images and minimizes the similarity between the outlier and the normal images. After training converges, the proposed PCL improves the model in handling subtle rule differences between normal images and outliers.

Our contributions can be summarized as follows. 1) To avoid possible shortcuts in appearance matching, an MC^{2}R task with a benchmark dataset is proposed to challenge existing AVR models in comprehending complex visual compositional relations. 2) To better detect the tiny rule differences, the proposed model transforms the MC^{2}R task into a select-1-out-10 problem, and designs a Regression Residual Reasoning module to determine the three images that follow the same compositional rule. 3) To boost the generalizability, the proposed visual perception module incorporates a data augmentation mechanism into the contrastive learning framework, which contrasts the pseudo-labeled normal images and outliers to enhance the discriminant power of the model. 4) Extensive experimental results show that the proposed MC^{2}R dataset is much more challenging than existing benchmark datasets, and the proposed method significantly outperforms state-of-the-art solutions on two benchmark datasets.

2 Related Work

Abstract Visual Reasoning. Abstract Visual Reasoning explores the capacity to draw analogies regarding abstract visual relations across diverse scenes. It is often used to evaluate human's intelligence in applying relevant knowledge and past experiences in unfamiliar contexts [Małkiński and Mańdziuk, 2023]. AVR contains several tasks: Raven's Progressive Matrices (RPM) [Zhang *et al.*, 2019; Barrett *et al.*, 2018], Odd-one-out [Zerroug *et al.*, 2022], Abstraction and Reasoning Corpus (ARC) [Chollet, 2019], etc. RAVEN [Zhang *et al.*, 2019] and PGM [Barrett *et al.*, 2018] are typical RPM benchmarks, where the question panel involves a 3×3 grid with 8 contextual images and a missing one, and the task is to identify the missing image from 8 candidates. The rules in RPMs are often too simple and fail to reflect the complexity of compositions. Odd-one-out is better suited for creating complex compositional rules [Zerroug *et al.*, 2022], where the task is to identify the outlier from four given images. ARC derives the masked block based on the example grid-like color blocks [Chollet, 2019]. As many AVR tasks were initially designed to evaluate human's abstract reasoning, they are intentionally composed of simple patterns following simple rules.

Compositionality assists humans in gaining a deeper understanding of the essence of objects [Hofstadter, 2001]. It has been explored in diverse fields, including logical reasoning [Lake and Baroni, 2018], visual reasoning [Johnson *et al.*, 2017; Thrush *et al.*, 2022; Zerroug *et al.*, 2022], and mathematics [Saxton *et al.*, 2018]. Typically, the CVR dataset [Zerroug *et al.*, 2022] has been created to evaluate the reasoning of compositionality, which applies compositional relations to a set of attributes and generates a wide range of tasks that adhere to compositional rules. In this paper, the proposed MC^{2}R task addresses the limitations of existing AVR tasks, typically CVR tasks, and challenges existing solution models.

Many methods have been developed to solve AVR tasks, which extract visual attributes using a perception module and exploit their relations using a reasoning module [Barrett *et al.*, 2018; Wu *et al.*, 2020; He *et al.*, 2023; Yang *et al.*, 2023]. As early AVR tasks are composed of simple regular shapes, shallow networks such as ResNet-18 [Hu *et al.*, 2021] and ResNet-50 [Barrett *et al.*, 2018; Zhang *et al.*, 2019] are often utilized for visual perception. Yang *et al.* [2023] argued that deep neural networks are not suitable for AVR problems where the objects are relatively small, so they adopted four residual blocks for visual perception. He *et al.* [2023] argued that local attributes such as `shape`, `size` and `color` can be well detected by shallow networks, while global attributes such as `number` and `position` may be better detected using deep networks, so they developed Hierarchical ConViT to perceive the visual attributes at different receptive fields.

To deduce the relations between attributes, early models [Barrett *et al.*, 2018] often utilize multi-layer perceptrons or ResNet blocks. Recently, several specialized reasoning blocks have been developed. Barrett [Barrett *et al.*, 2018] first encoded context-context relations and context-multiple-choice relations based on extracted visual features, and iteratively applied Relation Networks to deduce inter-panel relations. Wu *et al.* [2020] introduced the Scattering Compositional Learner (SCL) for feature extraction through object networks and attribute networks, and designed a similar SCL for abstract reasoning. Yang *et al.* [2023] developed PredRNet to deduce abstract rules by directly mapping from the eight context images to the missing image. Very recently, He *et al.* [2023] developed an attention-based relational reasoner to discern the underlying relations. Despite these recent advancements, there may still exist shortcuts for appearance matching due to the nature of existing AVR tasks, which hinders understanding of underlying rules.

Contrastive Learning. Contrastive learning maintains representation consistency across various data views while enhancing discrimination between different data points [Son, 2022]. Related research mainly focuses on two directions. 1) The design of training strategies for enhancing the generalization capabilities of a backbone [Chen *et al.*, 2020; Chen *et al.*, 2021]. SimCLR [Chen *et al.*, 2020] preserves the similarity between a sample and its augmented versions, and ensures their dissimilarity from other samples, with a learnable nonlinear transformation to enhance the feature representation. Chen *et al.* [2021] developed the widely adopted MoCo (Momentum Contrast), utilizing a queue to augment negative samples and a momentum encoder to maintain the feature consistency. 2) The formulation of the loss function to enhance the discrimination among samples [Chuang *et al.*, 2022]. Specifically, the InfoNCE loss [Oord *et al.*, 2018] follows the structure of BCE loss but incorporates a hyperparameter to address the challenge of identifying dissimilar negative samples. Robinson *et al.* [2020] devised an adjustable sample distribution to highlight challenging negative samples. Chuang *et al.* [2022] presented RINCE to enforce the symmetry property in the loss function, thereby enhancing the robustness of the derived features. The proposed PCL exploits pseudo labels and contrastive learning with data augmentation to extract generalized high-level abstract features that better distinguish normal images and outliers.

3 Proposed Methodology

3.1 Problem Formulation

As AVR tasks are often constructed using simple rules on regular shapes [Zhang *et al.*, 2019], the solution may be derived from appearance matching instead of uncovering the underlying rules. In addition, most existing problem formulations are single-answer questions, *e.g.*, identifying one outlier from four images in Odd-one-out [Zerroug *et al.*, 2022] or selecting the most suitable option to complete a 3×3 question panel in RPMs. Furthermore, image variations are often limited due to simple regular shapes in AVR datasets. All these simplify the tasks and artificially boost the performance of AVR models.

To address these issues, a more challenging AVR task of **M**ultiple **C**omplex **C**ompositional **R**easoning (**MC^{2}R**) is proposed. As shown in Figure 1, given a question panel of five images, $\{\boldsymbol{X}_i \in \mathbb{R}^{H \times W \times C}\}_{i=1}^{5}$, three are generated from the same compositional rule formed by applying specific relations on random object attributes in a tree-like manner and two are generated as outliers by applying slightly different compositional rules on the same attributes, the task is to identify the two outliers, where H, W, and C are the height, width, and number of channels of images, respectively.

The MC^{2}R tasks well address the issues of existing AVR datasets. Firstly, each question panel is constructed using compositional rules rather than simple rules, and the differences in compositional rules between outliers and normal images are intentionally kept minimal. Secondly, instead of regular shapes, curved contours are utilized to increase variations in image appearance. Compared to the CVR dataset [Zerroug *et al.*, 2022], the attribute variations are intentionally randomly sampled from a much larger range, bringing significant appearance differences between images generated from the same rule. Lastly, as the proposed MC^{2}R identifies two outliers from five given images, this multi-answer question is much more challenging than the single-answer questions in PGM [Barrett *et al.*, 2018], RAVEN [Zhang *et al.*, 2019], and CVR [Zerroug *et al.*, 2022]. All these measures prevent the

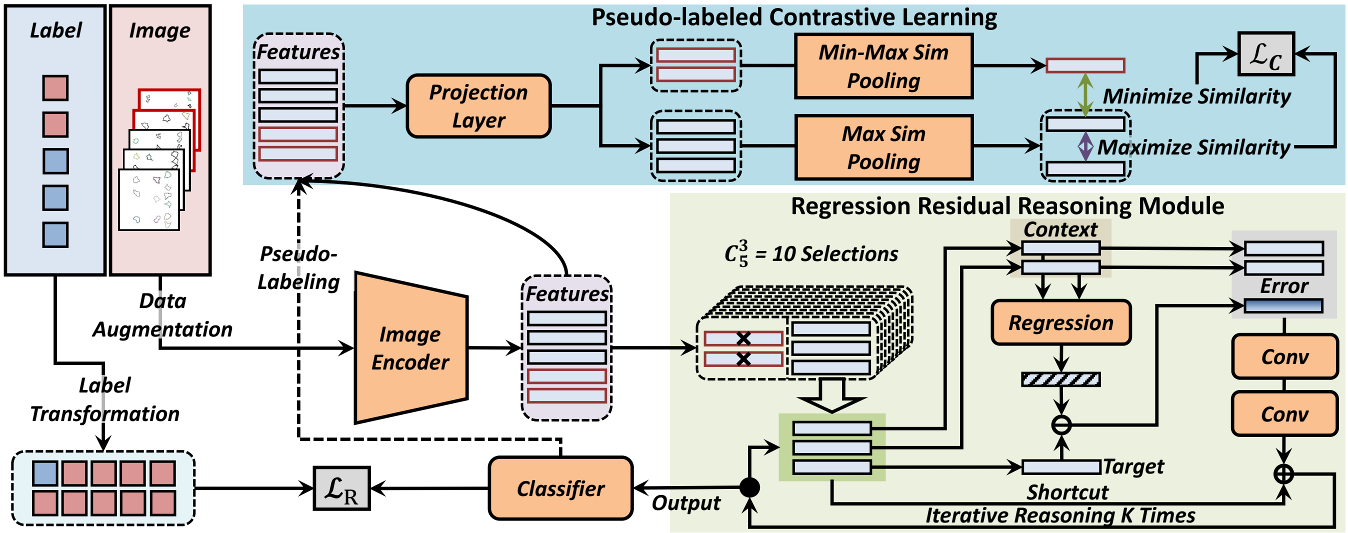

Figure 2: Overview of proposed R^{3}PCL, containing two main modules. 1) Image encoder with Pseudo-labeled Contrastive Learning (PCL), to select two normal images and one outlier with high confidence using pseudo-labels, thereby extracting discriminant and generalized features through contrastive learning. 2) Regression Residual Reasoning Module (R^{3}M), containing a set of Regression Residual Reasoning Blocks (R^{3}Bs) to abstract rules, in which one image is regressed using the other two, and the error is minimized if three images follow the same rule.

model from completing tasks through appearance matching and greatly increase the difficulty of the MC^{2}R task.

Given the input $\boldsymbol{X}_i$ with the label $\boldsymbol{y}_i = 1$ if $\boldsymbol{X}_i$ is an outlier and 0 otherwise, the training process can be represented as,

$$\hat{\boldsymbol{y}} = \mathcal{F}_\Theta(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4, \boldsymbol{X}_5; \Theta), \tag{1}$$

where $\hat{\boldsymbol{y}}$ is the one-hot encoding of regressed labels, Θ denotes the model parameters, and $\mathcal{F}_\Theta$ is the mapping function.

3.2 Overview of Proposed R^{3}PCL

To solve MC^{2}R tasks, **R**egression **R**esidual **R**easoning with **P**seudo-labeled **C**ontrastive **L**earning (**R^{3}PCL**) is proposed, which contains two main modules. 1) The visual perception module consisting of an image encoder with a pre-trained ResNet-50 [Chen *et al.*, 2021] as the backbone, and a **P**seudo-labeled **C**ontrastive **L**earning (**PCL**) block to boost the feature representations. In contrast to unsupervised learning in most contrastive learning models [Khosla *et al.*, 2020], the proposed PCL pseudo-labels normal images and outliers, and subsequently designs a set of min-max similarity pooling operations to select the more representative outlier and a max-pooling operation to select the pair of most similar normal images. Then, to enhance the feature representation, contrastive learning with data augmentation is designed to minimize the similarity between the outlier and normal images and maximize the similarity between normal images. 2) **R**egression **R**esidual **R**easoning **M**odule (**R^{3}M**), which converts the identify-2-out-5 formulation into the select-1-out-10 formulation. Then, a set of **R**egression **R**esidual **R**easoning **B**locks (**R^{3}Bs**) are designed to identify the triplet of three normal images, where one image can be regressed by the other two with minimal error. The proposed R^{3}M detects subtle differences in underlying composition rules by minimizing the regression error among three normal images, and hence more accurately identifies them and, equivalently, the two outliers.

3.3 Pseudo-labeled Contrastive Learning

The proposed **PCL** promotes similarity among images adhering to the same compositional rule while discouraging similarity among images following different rules. To achieve this, we leverage the initial prediction results of R^{3}M to pseudo-label normal images and outliers. Denote $\{\boldsymbol{N}_i\}_{i=1}^3$ and $\{\boldsymbol{O}_j\}_{j=1}^2$ as the features for pseudo-labeled normal images and outliers respectively. These pseudo-labels may contain errors and hence directly employing them introduces noise. To tackle this problem, similarity pooling operations are designed to improve the quality of pseudo-labels, where the cosine similarity, $\mathcal{S}(\boldsymbol{F}_i, \boldsymbol{F}_j) = \frac{\boldsymbol{F}_i \times \boldsymbol{F}_j}{|\boldsymbol{F}_i| \times |\boldsymbol{F}_j|}$, is used. It can be observed that normal images following the same compositional rule exhibit higher similarity than images following different rules. We hence select the pair of normal images that exhibit the largest feature similarity,

$$\mathcal{F}_{max}(\mathcal{S}(\boldsymbol{N}_1, \boldsymbol{N}_2), \mathcal{S}(\boldsymbol{N}_1, \boldsymbol{N}_3), \mathcal{S}(\boldsymbol{N}_2, \boldsymbol{N}_3)), \tag{2}$$

where $\mathcal{F}_{max}(\cdot)$ is the max-pooling operation. Denote $\boldsymbol{N}_1^c$ and $\boldsymbol{N}_2^c$ as the two normal images with the highest confidence.

Next, we select one outlier with higher confidence from two pseudo-labeled outliers. Intuitively, outliers are least similar to the normal images. We hence employ min-max-pooling to identify the outlier. Specifically, we first derive the nearest neighbor for each pseudo-labeled outlier by the max-pooling operation, and then choose the one with the least similarity as the outlier by the min-pooling operation $\mathcal{F}_{min}(\cdot)$,

$$\mathcal{F}_{min}\left(\left\{\mathcal{F}_{max}\left(\{\mathcal{S}(\boldsymbol{N}_i, \boldsymbol{O}_j)\}_{i=1}^3\right)\right\}_{j=1}^2\right). \tag{3}$$

Denote $\boldsymbol{O}^c$ as the outlier. The PCL loss $\mathcal{L}_C$ is defined as,

$$\mathcal{L}_C = -\log \frac{\exp(\mathcal{S}(\boldsymbol{N}_1^c, \boldsymbol{N}_2^c))}{\exp(\mathcal{S}(\boldsymbol{N}_1^c, \boldsymbol{N}_2^c)) + \sum_{i=1}^2 \exp(\mathcal{S}(\boldsymbol{N}_i^c, \boldsymbol{O}^c))}. \tag{4}$$

3.4 Regression Residual Reasoning Module

Challenges of MC^{2}R. There are many challenges in solving the MC^{2}R tasks. 1) For each question panel, there are only three normal images and two outliers, so it is inherently challenging to uncover the rules from such limited images. In addition, there are two correct outliers for each question panel, while most existing AVR tasks contain single-answer questions. 2) The compositional rules for normal images and outliers only differ slightly, so their appearance differences are challenging to observe, while the attribute variations across images are intentionally increased to avoid appearance matching. 3) The MC^{2}R tasks contain as many as 103 diverse compositional rules, compared to several simple rules in many AVR tasks. Furthermore, they are composed of levels of compositions, in contrast to existing datasets without rule compositions [Barrett *et al.*, 2018; Zhang *et al.*, 2019]. To tackle the challenges, the proposed R^{3}M transforms the identify-2-out-5 problem into a select-1-out-10 problem, and then designs a set of R^{3}Bs to identify the three normal images, and hence identify the two outliers.

Problem Transformation. It is hard to directly identify the two outliers from five given images. Hence, the multi-answer task is transformed into a single-answer problem of select-1-out-10. Specifically, given five images, there are $C_5^3 = 10$ possible ways to select three images, and only one consists of three normal images, *i.e.*, $\{\boldsymbol{F}_i^s\}_{i=1}^3 = \mathcal{F}_{se}(\{\boldsymbol{F}_j\}_{j=1}^5)$, where $\mathcal{F}_{se}(\cdot)$ denotes the function to select three images $\{\boldsymbol{F}_i^s\}_{i=1}^3$ from five given images $\{\boldsymbol{F}_j\}_{j=1}^5$. If three normal images are selected, one image can be well regressed by the other two with minimal error, while the error will be large if there is one or two outliers in these three images. By identifying the image triplet with the smallest regression error $\mathcal{E}(\{\boldsymbol{F}_i^s\}_{i=1}^3)$, the three normal images will be accurately identified. Other formulations such as selecting more than three images will include outliers and selecting fewer than three may result in insufficient information for accurate regression.

Regression Residual Reasoning Block. To capture the subtle differences in diverse compositional rules between normal images and outliers using limited images, a set of Regression Residual Reasoning Blocks is designed for abstract reasoning. As shown in Figure 2, given three images, two of them are treated as the context images to regress the remaining one,

$$\hat{\boldsymbol{F}}_3^s = \mathcal{F}_r(\boldsymbol{F}_1^s, \boldsymbol{F}_2^s), \tag{5}$$

where $\mathcal{F}_r$ is the regression function. We then calculate the regression error $\tilde{\boldsymbol{F}}_3^s = \boldsymbol{F}_3^s - \hat{\boldsymbol{F}}_3^s$ between the regression $\hat{\boldsymbol{F}}_3^s$ and the target, concatenate $\tilde{\boldsymbol{F}}_3^s$ with the two context features to form $\hat{\mathbf{Y}} = [\boldsymbol{F}_1^s, \boldsymbol{F}_2^s, \tilde{\boldsymbol{F}}_3^s]$, and feed $\hat{\mathbf{Y}}$ into two convolutional layers to enhance the interactions within these features, $\tilde{\mathbf{Y}} = \mathcal{F}_C(\hat{\mathbf{Y}})$, where $\mathcal{F}_C$ represents the operations of two convolutional layers. Moreover, we design a residual structure to shortcut the original features $\mathbf{Y}^0 = [\boldsymbol{F}_1^s, \boldsymbol{F}_2^s, \boldsymbol{F}_3^s]$ as,

$$\mathbf{Y}^1 = \mathbf{Y}^0 + \tilde{\mathbf{Y}}. \tag{6}$$

Iterative Regression Residual Reasoning. There are multiple compositions in each rule and a single R^{3}B is insufficient for adequate reasoning. The proposed iterative regression residual reasoning hence hierarchically combines K R^{3}Bs,

$$\mathbf{Y}^{k+1} = \mathcal{F}_{R^3B}^k(\mathbf{Y}^k), 0 \leq k < K, \tag{7}$$

where $\mathcal{F}_{R^3B}^k$ denotes the reasoning block in the k-th iteration. This simulates the iterative reasoning of humans. When faced with an unfamiliar rule, humans engage in a continuous process of guessing, validating, and ultimately confirming the rule, corresponding to the iterative reasoning process in R^{3}M.

Loss Function for R^{3}M. We convert the label of each question panel into a one-hot vector $\boldsymbol{y} = \{y_i\}_{i=1}^{10}$, where $y_i = 1$ if three normal images are selected, and 0 otherwise. Following previous AVR models [Benny *et al.*, 2021], two fully connected layers are designed as the classifier $\mathcal{F}_\phi(\cdot)$, to determine which one among 10 selections consists of three normal images. The Binary Cross Entropy Loss is used in this paper,

$$\mathcal{L}_R = -\sum_{i=1}^{10} \sigma(\boldsymbol{y}_i) \log \sigma(\mathcal{F}_\phi(\mathbf{Y}^K)), \tag{8}$$

where $\mathcal{L}_R$ denotes the regression loss, and $\sigma(\cdot)$ denotes the sigmoid function. The final loss for our end-to-end model is,

$$\mathcal{L} = \mathcal{L}_R + \lambda \mathcal{L}_C, \tag{9}$$

where λ is the weighting factor to balance the two losses.

4 Construction of MC^{2}R Dataset

The MC^{2}R dataset contains 9 elementary tasks involving one relation for each task, and 94 compositional tasks combining two or more relations for each task, *i.e.*, 20 using single relations, 65 using a pair of relations, and 9 using more than two relations. For each task, 10,000, 500 and 1,000 samples are generated for training, validation and testing, respectively.

Each image consists of three elements. 1) **Object**, a closed irregular contour with 9 fundamental attributes: `shape`, `position`, `color`, `size`, `rotation` angle, and `flip` state of an object, `inside` representing that one object contains another, `contact` that two objects are contacting, and `count` representing the number of objects. 2) **Relation**, defined on attribute as: *logical operators* such as and $\wedge$, or $\vee$, not $\neg$; *relational operators* such as greater than $>$, less than $<$, equal to $=$; *arithmetic operators* such as plus $+$, minus $-$. 3) **Rule**. As shown in Figure 1, objects are treated as bottom-level nodes, relations are treated as intermediate and upper-level nodes, and a rule is composed by a hierarchical composition of nodes in a tree-like manner.

Each MC^{2}R task is created with three images following the same compositional rule, and two outliers following slightly different rules, *e.g.*, by changing a relation node in the rule from $=$ to $>$. The rule differences are intentionally minimized to challenge AVR models and avoid appearance matching. The variations of attributes that are not used in the compositional rules are also intentionally increased, *e.g.*, if the compositional rule includes `rotation`, the `shape` and `color` of objects, which are unrelated to the compositional rule, are kept as diverse as possible. Compared to single-answer questions in CVR, it is much more challenging to reason the multiple answers on the MC^{2}R dataset, on top

Model	20	50	100	200	500	1000	AUC	10000
WReN [2018]	26.8/10.6	27.6/10.5	28.5/11.0	30.1/11.0	36.4/11.6	42.3/12.6	32.0/11.2	64.5/37.6
SCL [2020]	25.8/10.1	25.8/10.0	28.3/10.1	34.1/10.2	43.2/10.5	46.2/12.5	33.9/10.6	56.9/30.4
PredRNet [2023]	26.9/11.4	30.0/12.6	31.5/13.1	36.2/15.6	45.9/23.3	54.5/41.2	37.5/19.5	92.2/59.3
ResNet-50 [2016]	27.5/10.6	28.2/10.5	29.9/10.7	33.9/13.1	52.1/17.1	59.2/28.1	38.4/15.0	93.7/75.3
ViT-small [2020]	27.3/10.1	27.8/10.1	28.0/10.4	28.1/10.5	29.9/11.1	31.4/11.7	28.7/10.7	58.7/14.1
IN-ResNet-50 [2016]	32.0/12.5	35.1/14.5	39.0/16.4	43.8/19.9	57.7/27.7	69.5/41.2	46.2/22.0	-/-
IN-ViT-small [2020]	27.9/11.4	28.2/11.9	28.6/13.4	30.0/14.9	35.6/17.6	47.2/20.7	32.9/15.0	-/-
CLIP-ResNet-50 [2021]	28.7/14.4	32.0/15.9	40.8/18.3	46.9/20.2	59.7/32.3	74.4/38.1	47.1/23.2	-/-
CLIP-ViT-base [2021]	31.1/11.6	37.4/11.5	43.9/17.6	56.0/20.4	68.9/29.0	78.8/35.9	52.7/21.0	-/-
SSL-ResNet-50 [2021]	44.3/19.4	50.3/22.7	55.3/24.9	59.5/27.4	68.9/37.7	79.2/44.9	59.6/29.5	93.1/82.3
SSL-ViT-small [2021]	39.3/15.9	39.5/16.6	40.8/17.1	44.1/17.6	53.3/18.7	60.7/21.7	46.3/17.9	81.6/38.9
Proposed R^{3}PCL	**44.8/19.8**	**52.1/27.8**	**57.6/35.1**	**63.0/45.5**	**76.4/59.6**	**88.9/72.9**	**67.1/43.4**	**95.7/87.9**

Table 1: Comparison with state-of-the-art models on the CVR [Zerroug *et al.*, 2022] (**Left**) and MC^{2}R (**Right**) datasets. Following the same settings as in [Zerroug *et al.*, 2022], the results are reported for using different numbers of training samples per task on both datasets.

of the increased attribute variations. Compared to simple rules on regular shapes in PGM [Barrett *et al.*, 2018] and RAVEN [Zhang *et al.*, 2019], the proposed MC^{2}R is constructed using much more complex and diverse compositional rules on irregular contours, and hence more challenging.

5 Experimental Results

5.1 Experimental Settings

The proposed R^{3}PCL is compared with three state-of-the-art models in solving RPMs but adapted for solving CVR and MC^{2}R problems, *i.e.*, **WReN** [Barrett *et al.*, 2018] utilizing a Relation Network to infer inter-feature relations for RPM problems; **SCL** [Wu *et al.*, 2020] containing an object network for encoding objects, an attribute network for encoding attributes, and a relation network for unveiling inherent relations; **PredRNet** [Yang *et al.*, 2023] comprising an image encoder with four residual blocks to extract visual features and a predictive reasoning module to capture the relations between contextual images and candidate answers.

R^{3}PCL is also compared with eight state-of-the-art models for solving CVR problems, *i.e.*, **ResNet-50** [He *et al.*, 2016] and **ViT-small** [Dosovitskiy *et al.*, 2020] are utilized as the backbone image encoder, and incorporated with two fully connected layers as the reasoning module, same as in [Zerroug *et al.*, 2022]. **IN-ResNet-50** [He *et al.*, 2016] and **In-ViT-small** [Dosovitskiy *et al.*, 2020] pre-trained on ImageNet [Deng *et al.*, 2009], **CLIP-ResNet-50** and **CLIP-ViT-base** employing the visual encoder of CLIP [Radford *et al.*, 2021], and **SSL-ResNet-50** and **SSL-ViT-small** utilizing MoCo-v3 [Chen *et al.*, 2021] to pre-train the backbone with 1.03 million samples are chosen for comparisons, and the remaining configurations are consistent with ResNet-50 [He *et al.*, 2016] and ViT-small [Dosovitskiy *et al.*, 2020].

The input image dimensions are $128 \times 128 \times 3$ pixels. The number of R^{3}Bs is set to $K = 3$ and the weighting factor λ is set to 1, as indicated by the ablation study. The Adam optimizer is employed with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. The batch size is set to 64. The maximum number of epochs is set to 100, and training stops if there is no substantial performance gain in 10 epochs.

5.2 Comparisons on CVR Dataset

Extensive comparison experiments are conducted on the CVR dataset, following the same settings as in [Zerroug *et al.*, 2022], by reporting the reasoning accuracy for 20, 50, 100, 200, 500, 1,000, and 10,000 training samples per task. As shown in Table 1, the proposed R^{3}PCL significantly and consistently outperforms all the compared models under all the settings. Specifically, compared to the second-best methods, SSL-ResNet-50 [Chen *et al.*, 2021], the performance gains using 500 and 1,000, and 10,000 training samples are 7.5%, 9.7% and 2.6%, respectively, showing the superior ability of R^{3}PCL in solving compositional reasoning problems.

Figure 3 presents the accuracy of SSL-ResNet-50 and R^{3}PCL in reasoning pairwise compositional rules using 1,000 training samples per task on the CVR dataset. Our method consistently outperforms SSL-ResNet-50 for all pairs of elementary rules except those where both methods perform perfectly. In particular, the proposed R^{3}PCL presents almost perfect results on the diagonal representing the accuracy on elementary rules and their compositions, while SSL-ResNet-50 performs poorly, especially rules related to `rotation` and `count` that require high-level image semantics.

5.3 Comparisons on MC^{2}R Dataset

Extensive experiments are conducted on the MC^{2}R dataset, following the same settings as in [Zerroug *et al.*, 2022]. From Table 1, the following can be observed. 1) The proposed R^{3}PCL outperforms all the compared methods for all the settings, demonstrating its ability to distinguish the subtle differences in compositional rules despite the challenges of the MC^{2}R dataset. 2) When the number of training samples per task is less than 200, many methods have very low accuracy, even approaching a random guessing rate of 10%, while our method achieves a higher accuracy of 19.8%, 27.8%, 35.1% and 45.5% using 20, 50, 100, and 200 training samples, respectively. This not only indicates that our MC^{2}R dataset is

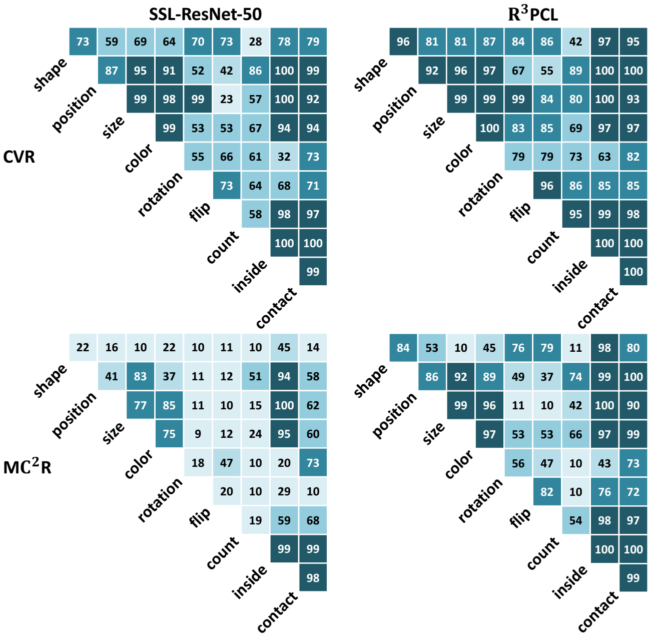

Figure 3: Comparison with previous best-performing SSL-ResNet-50 [Chen *et al.*, 2021] for individual tasks using 1,000 training samples per task on the CVR [Zerroug *et al.*, 2022] and MC^{2}R datasets.

more challenging, but also demonstrates the stronger reasoning abilities of our R^{3}PCL. 3) Compared to the second-best method, SSL-ResNet-50, the performance gain of R^{3}PCL grows as the number of training samples increases, as more samples assist the model to better capture visual information and compositional rules. Even when there are 10,000 training samples per task and the benefits of contrastive learning with data augmentation are reduced, the proposed R^{3}PCL still yields a large performance gain of 5.6%.

Overall, all the evaluated methods perform poorer on the MC^{2}R dataset than on the CVR dataset, showing that the MC^{2}R is more challenging. Even so, the proposed R^{3}PCL significantly outperforms all the compared methods on both datasets, demonstrating its superior ability in reasoning compositional rules.

Figure 3 shows the reasoning accuracy of SSL-ResNet-50 and R^{3}PCL for individual tasks on the MC^{2}R dataset. The proposed method significantly outperforms SSL-ResNet-50 for all pairs of elementary rules. While SSL-ResNet-50 struggles with rules related to `rotation` and `flip`, the proposed R^{3}PCL performs much better on these tasks, demonstrating its effectiveness in capturing high-level image semantics and uncovering the subtle rule differences.

5.4 Ablation Studies

We ablate the two major components of R^{3}PCL on the MC^{2}R dataset. SSL-ResNet-50 [Chen *et al.*, 2021] serves as the baseline, and its perception module and reasoning module are replaced by PCL and R^{3}M respectively. As shown in Table 2, the proposed PCL achieves a performance gain of 2.3%, 4.0%, and 12.9% for 200, 500, and 1,000 training samples per task respectively, indicating that it could better distinguish normal images and outliers through contrastive learning. The performance gains brought by R^{3}M are 11.7%, 13.5%, and 14.0% respectively, proving that R^{3}M could better capture the subtle rule differences and identify outliers. By adopting PCL and R^{3}M simultaneously, the accuracy is further boosted to 45.5%, 59.6%, and 72.9%, respectively. The ablation results demonstrate the effectiveness of the two proposed modules.

PCL	R^{3}M	200	500	1000
✗	✗	27.4	37.7	44.9
✔	✗	29.7	41.7	57.8
✗	✔	39.1	51.2	58.9
✔	✔	**45.5**	**59.6**	**72.9**

Table 2: Ablation study of major components of proposed R^{3}PCL.

Param.	K				λ			
	1	2	3	4	0.25	0.5	1	2
Acc (%)	71.2	71.5	**72.9**	70.4	71.6	71.1	**72.9**	72.3

Table 3: Ablation study of key parameters K and λ.

Our model is ablated using different numbers of R^{3}Bs with 1,000 training samples per task on the MC^{2}R dataset. As shown in Table 3, when K is increased from 1 to 3, there is a noticeable performance improvement from 71.2% to 72.9%, indicating that iterative reasoning better deduces complex compositional rules. The performance at $K = 4$ drops to 70.4%, due to possible over-fitting by using too many R^{3}Bs.

Lastly, we evaluate the weighting factor λ in Equation (9). As shown in Table 3, for $\lambda = 0.25$ or 0.5, the weight of $\mathcal{L}_C$ is too small, causing the perception module to insufficiently extract discriminative and generalized features. When $\lambda = 2$, the accuracy drops slightly compared to $\lambda = 1$ due to overemphasis on contrastive learning. Hence, λ is set to 1 by default.

6 Conclusion

It is challenging to reason over compositional rules on the CVR dataset, but the single-answer question formulation and limited attribute variations potentially allow models to complete tasks through appearance matching. To address this problem, an MC^{2}R benchmark is developed to challenge compositional reasoning, which has increased attribute variations, smaller rule differences between normal images and outliers, and multiple correct answers. The proposed R^{3}PCL well tackles the challenges of MC^{2}R tasks by enhancing feature representations through Pseudo-labeled Contrastive Learning and strategically transforming the original problem into a select-1-out-10 problem, and effectively solves the task through iterative regression residual reasoning. The experimental results on the CVR and MC^{2}R datasets demonstrate that the proposed R^{3}PCL significantly and consistently outperforms all the compared models under various settings. Code is available here.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 72071116, and in part by the Ningbo Science and Technology Bureau under Grants 2022Z173, 2022Z217, 2023Z138 and 2023Z237.

Contribution Statement

The contributions of Chengtai Li and Yuting He to this paper were equal.

References

[Barrett *et al.*, 2018] David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning (ICML)*, pages 511–520. PMLR, 2018.

[Benny *et al.*, 2021] Yaniv Benny, Niv Pekar, and Lior Wolf. Scale-localized abstract reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12557–12565, 2021.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020.

[Chen *et al.*, 2021] X Chen, S Xie, and K He. An empirical study of training self-supervised vision transformer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9620–9629, 2021.

[Chollet, 2019] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

[Chuang *et al.*, 2022] Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16670–16681, 2022.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009.

[Ding *et al.*, 2022] Ruxin Ding, Jianfeng Ren, Heng Yu, and Jiawei Li. Dynamic texture recognition using PDV hashing and dictionary learning on multi-scale volume local binary pattern. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1840–1844. IEEE, 2022.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[He *et al.*, 2023] Wentao He, Jialu Zhang, Jianfeng Ren, Ruibin Bai, and Xudong Jiang. Hierarchical convit with attention-based relational reasoner for visual analogical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 37, pages 22–30, 2023.

[Hofstadter, 2001] Douglas R Hofstadter. Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*, pages 499–538, 2001.

[Hu *et al.*, 2021] Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 1567–1574, 2021.

[Jiang *et al.*, 2022] Zhanghao Jiang, Ke Xu, Heshan Du, Huan Jin, Zheng Lu, and Qian Zhang. Occlusion-invariant representation alignment for entity re-identification. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3266–3270. IEEE, 2022.

[Johnson *et al.*, 2017] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997. IEEE, 2017.

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems (NeurIPS)*, 33:18661–18673, 2020.

[Lake and Baroni, 2018] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning (ICML)*, pages 2873–2882. PMLR, 2018.

[Li *et al.*, 2022] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4804–4814, 2022.

[Małkiński and Mańdziuk, 2023] Mikołaj Małkiński and Jacek Mańdziuk. A review of emerging research directions in abstract visual reasoning. *Information Fusion*, 91:713–736, 2023.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack

Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

[Robinson *et al.*, 2020] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations (ICLR)*, 2020.

[Saxton *et al.*, 2018] David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. Analysing mathematical reasoning abilities of neural models. In *International Conference on Learning Representations (ICLR)*, 2018.

[Son, 2022] Jeany Son. Contrastive learning for space-time correspondence via self-cycle consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14679–14688, 2022.

[Song *et al.*, 2023] Xingke Song, Jiahuan Jin, Chenglin Yao, Shihe Wang, Jianfeng Ren, and Ruibin Bai. Siamese-discriminant deep reinforcement learning for solving jigsaw puzzles with large eroded gaps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2303–2311, 2023.

[Thrush *et al.*, 2022] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248, 2022.

[Wu *et al.*, 2020] Yuhuai Wu, Honghua Dong, Roger Grosse, and Jimmy Ba. The scattering compositional learner: Discovering objects, attributes, relationships in analogical reasoning. *arXiv preprint arXiv:2007.04212*, 2020.

[Xu *et al.*, 2023] Jingyi Xu, Tushar Vaidya, Yufei Wu, Saket Chandra, Zhangsheng Lai, and Kai Fong Ernest Chong. Abstract visual reasoning: An algebraic approach for solving raven's progressive matrices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6715–6724, 2023.

[Yang *et al.*, 2023] Lingxiao Yang, Hongzhi You, Zonglei Zhen, Dahui Wang, Xiaohong Wan, Xiaohua Xie, and Ru-Yuan Zhang. Neural prediction errors enable analogical visual reasoning in human standard intelligence tests. In *International Conference on Machine Learning (ICML)*. PMLR, 2023.

[Zerroug *et al.*, 2022] Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A benchmark for compositional visual reasoning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:29776–29788, 2022.

[Zhang *et al.*, 2019] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5317–5327, 2019.

[Zhang *et al.*, 2023] Jialu Zhang, Jianfeng Ren, Qian Zhang, Jiang Liu, and Xudong Jiang. Spatial context-aware object-attentional network for multi-label image classification. *IEEE Transactions on Image Processing*, 2023.

[Zhang *et al.*, 2024] Jialu Zhang, Xiaoying Yang, Wentao He, Jianfeng Ren, Qian Zhang, Yitian Zhao, Ruibin Bai, Xiangjian He, and Jiang Liu. Scale optimization using evolutionary reinforcement learning for object detection on drone imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 410–418, 2024.