RuleMatch: Matching Abstract Rules for Semi-supervised Learning of Human Standard Intelligence Tests

Yunlong Xu[1*], **Lingxiao Yang**[2*], **Hongzhi You**[3], **Zonglei Zhen**[4]
Da-Hui Wang[4], **Xiaohong Wan**[4], **Xiaohua Xie**[2], **Ru-Yuan Zhang**[1#]
[1]Shanghai Jiao Tong University, China
[2]Sun Yat-sen University, China
[3]University of Electronic Science and Technology of China, China
[4]Beijing Normal University, China
yxu103@u.rochester.edu, yanglx9@mail.sysu.edu.cn, hongzhi-you@uestc.edu.cn, {zhenzonglei, wangdh, xhwan}@bnu.edu.cn, xiexiaoh6@mail.sysu.edu.cn, ruyuanzhang@sjtu.edu.cn

Abstract

Raven's Progressive Matrices (RPM), one of the standard intelligence tests in human psychology, has recently emerged as a powerful tool for studying abstract visual reasoning (AVR) abilities in machines. Although existing computational models for RPM problems achieve good performance, they require a large number of labeled training examples for supervised learning. In contrast, humans can efficiently solve unlabeled RPM problems after learning from only a few example questions. Here, we develop a semi-supervised learning (SSL) method, called RuleMatch, to train deep models with a small number of labeled RPM questions along with other unlabeled questions. Moreover, instead of using pixel-level augmentation in object perception tasks, we exploit the nature of RPM problems and augment the data at the level of abstract rules. Specifically, we disrupt the possible rules contained among context images in an RPM question and force the two augmented variants of the same unlabeled sample to obey the same abstract rule and predict a common pseudo label for training. Extensive experiments show that the proposed RuleMatch achieves state-of-the-art performance on two popular RAVEN datasets. Our work makes an important stride in aligning abstract analogical visual reasoning abilities in machines and humans. Our Code is at https://github.com/ZjjConan/AVR-RuleMatch.

1 Introduction

Deep Neural Networks (DNNs) [Krizhevsky *et al.*, 2012; Simonyan and Zisserman, 2014; He *et al.*, 2016; Vaswani *et al.*, 2017] are currently the *de facto* methods for many tasks in both academic research and industrial applications [Conneau *et al.*, 2016; Bahdanau *et al.*, 2014; Kim *et al.*, 2016; Ren *et al.*, 2015; Karpathy *et al.*, 2014]. However, most existing DNNs face two fundamental challenges. First, current DNNs heavily rely on a large number of labeled examples for training to achieve promising results. Second, current DNNs achieve extraordinary performance mostly in low-level perceptual tasks, such as computer vision or speech recognition, but are believed to still lag far behind humans in high-level analogical reasoning abilities. This is in stark contrast to human cognition, where even a baby can quickly learn object relationships and make inferences with only a few supervised samples. The ability to efficiently learn and generalize from a few labeled examples reveals remarkable abstract analogical reasoning abilities in humans.

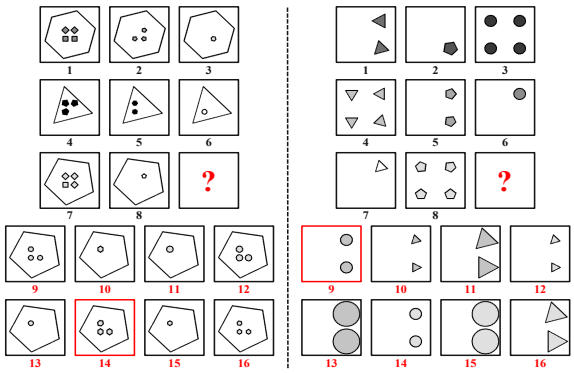

Figure 1: Two RPM questions. In each RPM, eight context images (*i.e.*, $1 \sim 8$) are provided to form a problem matrix. The goal is to select the correct answer from eight answer images (*i.e.*, $9 \sim 16$) to fill in the missing one (denoted using **?**) in that problem matrix, making three rows or three columns form similar abstract rules.

In human psychology, the ability of abstract analogical reasoning is typically assessed by the standard Intelligence Quotient (IQ) test -Raven's Progressive Matrices (RPMs) [Raven and Court, 1938]. Figure 1 shows two example RPM questions from the RAVEN [Zhang *et al.*, 2019a] dataset. The goal is to select the correct answer from 8 choice images to fill in the missing 9-th panel (denoted by **?**) of the context images, so that three rows or three columns to form a coherent abstract rule. An observer must first recognize visual attributes of objects in the context images and infer abstract rules embedded within the context images, and reason about the correct answer. Importantly, RPM is purely vision-based

and, unlike tuning tests, does not depend on language skills. RPM problems are therefore widely used as a standard intelligence test across regions and populations.

In cognitive science, AVR has been traditionally solved by neuro-symbolic approaches and several related models have been proposed [Lovett and Forbus, 2017; Lovett *et al.*, 2010]. However, traditional RPM problems used in human psychology are hand-crafted by experts, and the limited amount hinders the development of deep learning models. Recently, several large-scale RPM datasets have been established [Zhang *et al.*, 2019a; Benny *et al.*, 2021; Hu *et al.*, 2021a; Barrett *et al.*, 2018], which have greatly accelerated this line of research in deep learning. For example, the RAVEN dataset contains 42,000 training questions, 14,000 validation questions, and 14,000 test questions [Zhang *et al.*, 2019a]. Based on these large-scale RPM datasets, several deep learning models for AVR have been proposed and have shown impressive performance [Zhang *et al.*, 2019b; Zhuo and Kankanhalli, 2022; Hu *et al.*, 2021a; Wang *et al.*, 2020; Jahrens and Martinetz, 2020; Benny *et al.*, 2021; Zheng *et al.*, 2019; Spratley *et al.*, 2020; Zhang *et al.*, 2021; Wu *et al.*, 2020]. However, these models are trained in a fully supervised fashion. In contrast, in everyday life, humans learn primarily based on a few labeled examples and a large number of unlabeled examples. For example, human babies learn the concept of "dog" by receiving "dog" instructions from their parents only a few times and subsequently naturally associating unlabeled dog examples without explicit instruction. Similarly, a standard RAVEN test provides only a few example questions and instructions. An observer must combine knowledge of the labeled examples with unlabeled questions encountered in the test to complete the test. There is evidence that humans have remarkable abilities to integrate labeled and unlabeled examples [Gibson *et al.*, 2013]. This type of learning in humans is defined as semi-supervised learning (SSL), where only a few labeled examples and other unlabeled examples can be used for training. To our best knowledge, there has been no thorough investigation of SSL in RPM problems. Investigating SSL in abstract visual reasoning tasks is particularly valuable for building robust machines equipped with human-like cognitive functions.

In this work, we develop an SSL method that uses only a small number of labeled training examples in addition to other unlabeled examples to learn abstract relations in RPM problems. As shown in multiple previous works [Sohn *et al.*, 2020; Berthelot *et al.*, 2019], data augmentations play an important role in SSL methods. However, conventional data augmentations in single image classification typically perturb low-level visual features (e.g., flipping, adding noise), which may be suboptimal for the rule-based relationship reasoning across images in AVR. Therefore, we propose to perturb abstract rules embedded in an RPM question (see 8 context images in Figure 1) to generate augmentation samples. Then, we calculate consistency between different augmented samples to obtain pseudo labels. Such pseudo labels along with these samples will be used to improve the model in the traditional supervised fashion. Our SSL method outperforms several conventional SSL methods and achieves state-of-the-art performance on two popular RAVEN datasets.

2 Related Work

2.1 Machine Learning Models for RPMs

To create human-like intelligent machines, one important step is to systematically quantify and compare intelligence in both humans and machines. The recent surge of research interest in RPMs in machine learning is because RPMs can serve as a common testbed for both human and machine intelligence.

The first obstacle in this line of research is the lack of appropriate large-scale RPM datasets for model training. Several works [Barrett *et al.*, 2018; Zhang *et al.*, 2019a; Hu *et al.*, 2021a; Benny *et al.*, 2021] have demonstrated the feasibility of automatically generating large-scale RPM datasets. Based on these datasets, several machine learning models [Barrett *et al.*, 2018; Spratley *et al.*, 2020; Zhang *et al.*, 2019b; Zhuo and Kankanhalli, 2022; Zheng *et al.*, 2019; Hu *et al.*, 2021a; Benny *et al.*, 2021; Jahrens and Martinetz, 2020; Zhang *et al.*, 2021; Wang *et al.*, 2020] have been proposed with impressive robustness and accuracy. For example, DCNet improves the ability to reason about relationships reasoning by computing contrastive effects between candidate answers [Zhuo and Kankanhalli, 2022]. MRNet discovers different types of abstract relations based on visual features at different spatial scales [Benny *et al.*, 2021]. Although existing studies have shown impressive performance on one or two benchmarks, their methods are still trained in a completely supervised fashion, relying on a large number of labeled samples. This is known to deviate significantly from RPM reasoning in humans. We therefore mainly focus on SSL, which is more reasonable in real-life learning regimes.

2.2 Human's Semi-supervised Human Learning

Imagine a mother pointing to a dog and saying to her baby, "This is a dog". The baby can easily associate the concept of the dog with the appearance of this example. However, the baby sees the majority of dog examples in everyday life without such explicit instruction. Arguably, human concept learning in everyday life is largely in a SSL fashion rather than supervised learning as used in most current DNNs. However, existing studies on human SSL are rare compared to supervised and unsupervised learning. Zhu *et al.* [2007] provided perhaps the first piece of direct evidence of SSL in humans. The experimenters let human observers learn a few labeled artificial objects and then asked them to classify several unlabeled objects. The finding that the distribution of the unlabeled objects significantly biased human classification strongly suggests that humans learn by integrating both labeled and unlabeled examples. This notion was further extended by similar results in social categorical learning [Kalish *et al.*, 2011], and the test-item effects [Zhu *et al.*, 2010] during human category learning.

Computational models of human SSL generally fall into two categories. One approach is to augment supervised or unsupervised cognitive models. For example, Gibson *et al.* [2013] proposed that exemplar-based and prototype-based supervised learning models can be extended in an SSL fashion. The other approach is to formalize SSL in an online learning fashion. It has been shown that labeled examples can be used to form the initial categorical distributions and

these distributions can be updated by unlabeled examples via non-parametric Bayesian belief updating [Gibson *et al.*, 2013]. These neuro-symbolic models are successful in explaining human SSL behavior. Unlike successful applications of DNNs in supervised [Yamins *et al.*, 2014] and unsupervised learning [Zhuang *et al.*, 2021] in primates, DNNs of semi-supervised learning in humans are rare.

2.3 Semi-supervised Machine Learning

SSL is an active area of research in machine learning community. The goal of SSL is to train models with a limited set of labeled data, and then improve that trained models by exploring weakly-labeled or unlabeled data [Zhu, 2005; Zhu and Goldberg, 2009]. In the deep learning era, this problem setting has been used in many related works, including general image classification [Tarvainen and Valpola, 2017; Laine and Aila, 2016; Laine and Aila, 2016; Sohn *et al.*, 2020; Berthelot *et al.*, 2019], object detection [Jeong *et al.*, 2019; Misra *et al.*, 2015; Tang *et al.*, 2021; Xu *et al.*, 2021; Kaul *et al.*, 2022], node classification [Welling and Kipf, 2016; Yang *et al.*, 2016], and image generation [Bodla *et al.*, 2018; Katsumata *et al.*, 2022]. Existing SSL methods generally fall into two categories. The first approach is to develop different methods to regularize model adaption. For example, Temporal Ensembling [Laine and Aila, 2016] uses an exponential moving average of label predictions to penalize inconsistent model outputs, and Mean Teacher Model [Tarvainen and Valpola, 2017] improves on this method by averaging model weights. The second approach is to exploit data augmentation to generate new samples and utilize these generated samples to propose new loss functions. For example, Hu *et al.* [2021b] proposes a standard supervised loss function on labeled data, an unsupervised loss on unlabeled samples belonging to the same categories, and a pair loss on unlabeled samples belonging to different categories. Our work inherits the advantages of model regularization and data augmentation and proposes a new method for generating consistency reference and augmentation samples according to abstract rules in RPMs.

3 Methods

In this work, we proposea a new semi-supervised learning framework - RuleMatch (Figure 3), which utilizes rule-based relationships to exploit unlabeled samples to improve the deep models for the human intelligence tests.

3.1 Problem Setup

We first formulate the semi-supervised setting in RPM problems. Suppose N_l RPM questions with groundtruth labels $\mathcal{X} = \{([\boldsymbol{X}_i^c; \boldsymbol{X}_i^a], y_i) \mid i = 1, \cdots, N_l\}$ are provided, where $\boldsymbol{X}_i^c \in R^{k^c \times H \times W}$ are the collections of context images of the i-th RPM question. Similarly, $\boldsymbol{X}_i^a \in R^{k^a \times H \times W}$ are the answer images. y_i is the groundtruth label and k is the number of panels. In RAVEN-like tasks, each RPM question contains $k^c = 8$ and $k^a = 8$ images, and y_i ranges from $[1, 8]$ to indicate the correct answer, *i.e.,* allowing three rows or columns to constitute the same abstract rule. Besides $\boldsymbol{X}$, N_u unlabeled RPM questions are also provided $\mathcal{U} = \{([\boldsymbol{U}_i^c; \boldsymbol{U}_i^a], \emptyset) \mid i = 1, \cdots, N_u\}$, where $\boldsymbol{U}$ represents collections of images with unknown label $\emptyset$. Our goal in this paper is to combine the union set of $\mathcal{X} \cup \mathcal{U}$ to improve the robustness and accuracy of current popular reasoning networks. This is also the central goal of semi-supervised learning in traditional computer vision tasks [Berthelot *et al.*, 2019; Sohn *et al.*, 2020; Zhu, 2005; Misra *et al.*, 2015].

3.2 Background: FixMatch

Our algorithm is inspired by the recently proposed semi-supervised learning algorithm – FixMatch [Sohn *et al.*, 2020]. FixMatch attempts to optimize a linear combination of two cross-entropy loss functions in a stochastic manner. In detail, let $\boldsymbol{X}_i = [\boldsymbol{X}_i^c; \boldsymbol{X}_i^a]$ and $\boldsymbol{U}_i = [\boldsymbol{U}_i^c; \boldsymbol{U}_i^a]$ be all panels of the i-th labeled and unlabeled RPM question respectively, FixMatch assumes the loss function as follows:

$$\begin{aligned} l =& \frac{1}{B_l} \sum_{b=1}^{B_l} H(y_b, p_m(y|\alpha(\boldsymbol{X}_b))) \\ &+ \frac{\lambda_u}{B_u} \sum_{u=1}^{B_u} \mathbf{1}(\max(q_u) \geq \tau) H(\hat{q}_u, p_m(y|\mathcal{A}(\boldsymbol{U}_u))), \end{aligned} \tag{1}$$

where B_l and B_u are the batch sizes for labeled and unlabeled examples respectively. $p_m(y|\alpha(\boldsymbol{X}_b))$ is the predicted score distribution produced by a model with labeled input. $q_u = p_m(y|\alpha(\boldsymbol{U}_u))$ is the predicted score distribution by the same model for the unlabeled example $\boldsymbol{U}_u$. $\hat{q}_u = \arg\max(q_u)$ and τ is the threshold. $H(p, q)$ is the cross-entropy loss between two probability distribution p and q. λ_u is a balancing term.

FixMatch estimates a pseudo label of each unlabeled example over a *weakly-augmented* input $(\alpha(\cdot))$, and then uses the estimated label and a *strongly-augmented* version $(\mathcal{A}(\cdot))$ of the input for loss optimization. This two-type-augmentation introduces a form of consistency regularization that is crucial to FixMatch. Another important factor is the confidence threshold τ for selecting high-confidence predicted examples to reduce the negative impacts of noisy examples. In the following sections, we will analyze the reason why FixMatch is suboptimal for RPM problems.

3.3 Our Algorithm: RuleMatch

FixMatch [Sohn *et al.*, 2020] was originally proposed for image recognition (i.e., natural or digit image categorization), and the confidence threshold τ serves as an important factor for selecting useful unlabeled examples. This is feasible in image recognition because natural images usually have some similar statistical patterns across different categories, such as colors, textures, and some mid-level semantic parts. The deep networks trained with enough examples have good generalization abilities to recognize such categories, and they can even be transferred to discriminate novel objects [Donahue *et al.*, 2014; Sharif Razavian *et al.*, 2014] to some extent.

However, for RPM problems, a model should not only recognize objects that appeared in each panel but also extract abstract relationships between these objects. Finally, the model must infer the rules with given contexts and allow the selected answer together with the context images to follow a consistent rule. It is worth noting that: (1) the same objects appearing in

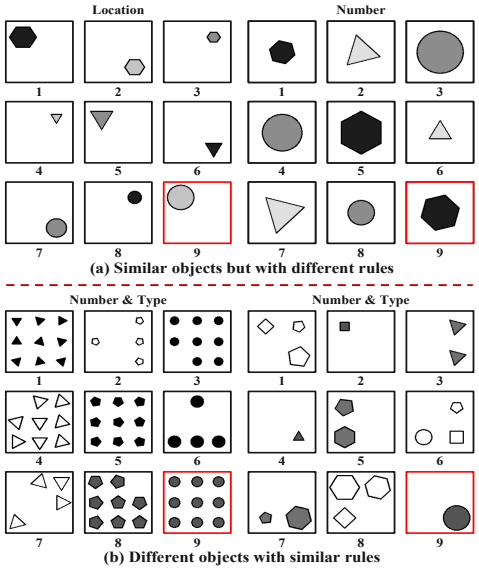

Figure 2: An illustration of the difficulty of RPM problems. (a) Two RPM questions contain similar objects in their panel images, but they have totally different rules as shown in the top of each RPM question. (b) Two RPM questions have similar rules, but the second RPM question contain much more diverse objects. It clearly shows that merely recognizing objects does not guarantee understanding of abstract rules, resulting in suboptimal unlabeled data selection.

the panels may not follow the same rules (see Figure 2 (a)); (2) different objects may also form similar abstract rules (see Figure 2 (b)); (3) small variations in objects may yield completely different rules. Due to these intrinsic characteristics of RPM problems, we argue that it is hard to select appropriate unlabeled samples using only a simple confidence threshold.

Based on the above analyses, we propose a simple yet effective algorithm – RuleMatch, as shown in Figure 3. Our RuleMatch follows the main component of FixMatch by utilizing different types of augmentation on unlabeled data, but differs in the way that how pseudo labels are estimated. Instead of a simple confidence threshold, we propose a rule-based agreement framework to select useful information from a large number of unlabeled data. After that, the selected unlabeled data will be combined with other labeled data to optimize a cross-entropy loss. To train robust rule-wise networks, we propose three types of rule-level augmentations.

Rule Consistency Loss

As shown in Figure 3. To generate pseudo labels from unlabeled data, we use our *RuleAug* (depict in the next paragraph) to produce two views of the same unlabeled RPM question: $\boldsymbol{U}_{u1} = RuleAug(\boldsymbol{U}_u)$ and $\boldsymbol{U}_{u2} = RuleAug(\boldsymbol{U}_u)$. As such, the model is used to calculate the predicted scores of both questions, denoted $q_{u1} = p_m(y|\boldsymbol{U}_{u1})$ and $q_{u2} = p_m(y|\boldsymbol{U}_{u2})$. Given the two predictions, we modified the second term in Eq (1) and formulate our rule consistency loss as:

$$l_u = \frac{\lambda_u}{B_u} \sum_{u=1}^{B_u} \mathbf{1}(\hat{q}_{u1} == \hat{q}_{u2}) H(\hat{q}_{u1}, p_m(y|\textit{RuleAug}(\boldsymbol{U}_u))), \tag{2}$$

where $\hat{q}_{u1} = \arg\max(q_{u1})$ and $\hat{q}_{u2} = \arg\max(q_{u2})$. Eq (2) shows that unlabeled data are included for model training if and only if the two views of rules are consistent. This design has two advantages over Eq (1). First, the consistency loss focuses on rule-wise recognition rather than visual attributes or categories of objects, and should be more appropriate for existing RPM problems. Second, Eq (2) removes the hyper-parameter τ, reducing the effort for parameter tuning. Moreover, we only employ the *RuleAug* on unlabeled data. This manipulation increases the robustness of the learned models to rule disruptions and produces more consistent predictions between two augmented views of the same question.

Rule Augmentation

As shown in Figure 1 & Figure 2, a panel in each RPM contains objects. These objects further form different attributes (*e.g.*, "number" in Figure 2 (b)). Finally, a row of three panel images forms some specific rules, which will be used for learning and reasoning. In summary, rules in RPM questions contain different levels of information. Based on these observations, we propose a rule-wise augmentation technique to enhance model robustness against rule disruption. Our technique contains three strategies to disrupt rules, including *object-level*, *panel-level*, and *row-level*.

Our first proposed strategy disrupts rule in object-level, denoted as *MaskObj*. Specifically, let us divide each context panel image into a 2×2 grid and use $[tl, tr, bl, br]$ to denote top-left, top-right, bottom-left, and bottom-right parts of each image in i-th RPM question. *MaskObj* can be formulated as:

$$\begin{aligned} MaskObj(\boldsymbol{U}_i^t): \quad & \boldsymbol{U}_i^t[r] = 0 \\ s.t. \quad & r = RS([tl, tr, bl, br]) \\ & t = RS([1, ..., 8]) \end{aligned} \tag{3}$$

where RS stands for the random selection of an element from the list contained in $[\cdot]$. *MaskObj* masks out 1/4 portion of an image, systematically disrupting the objects contained in the t-th context image. Such masking breaks the abstract rule embedded in this RPM question, and produces an augmented version of the original RPM question.

Our second and third augmentation strategies share a similar philosophy by breaking rules in single row or column, but in different ways. We formulate them as:

$$MaskPanel(\boldsymbol{U}_i^t): \boldsymbol{U}_i^t = 0 \tag{4}$$

$$MaskRow(\boldsymbol{U}_i): \boldsymbol{U}_i^{row} = 0, \tag{5}$$

where t is similar to Eq (3), and row is the random selection of a single row for the i-th RPM question, which can be formulated as $row = RS([row_1, row_2])$. Here, $row_1 = [1, 2, 3]$ or $row_2 = [4, 5, 6]$. The second augmentation method masks one of eight context images to break a rule. The third augmentation masks the entire first or the second row. The ultimate goal of rule-based masking is to systematically disrupt the rule in an RPM question and enforce a model to learn the correct answer from unmasked portion of the question. This approach allows the model to learn robust features by pushing it to obtain highly consistent predicted answers from two variants of augmented questions, and this process can in turn select high-quality unlabeled data for model training.

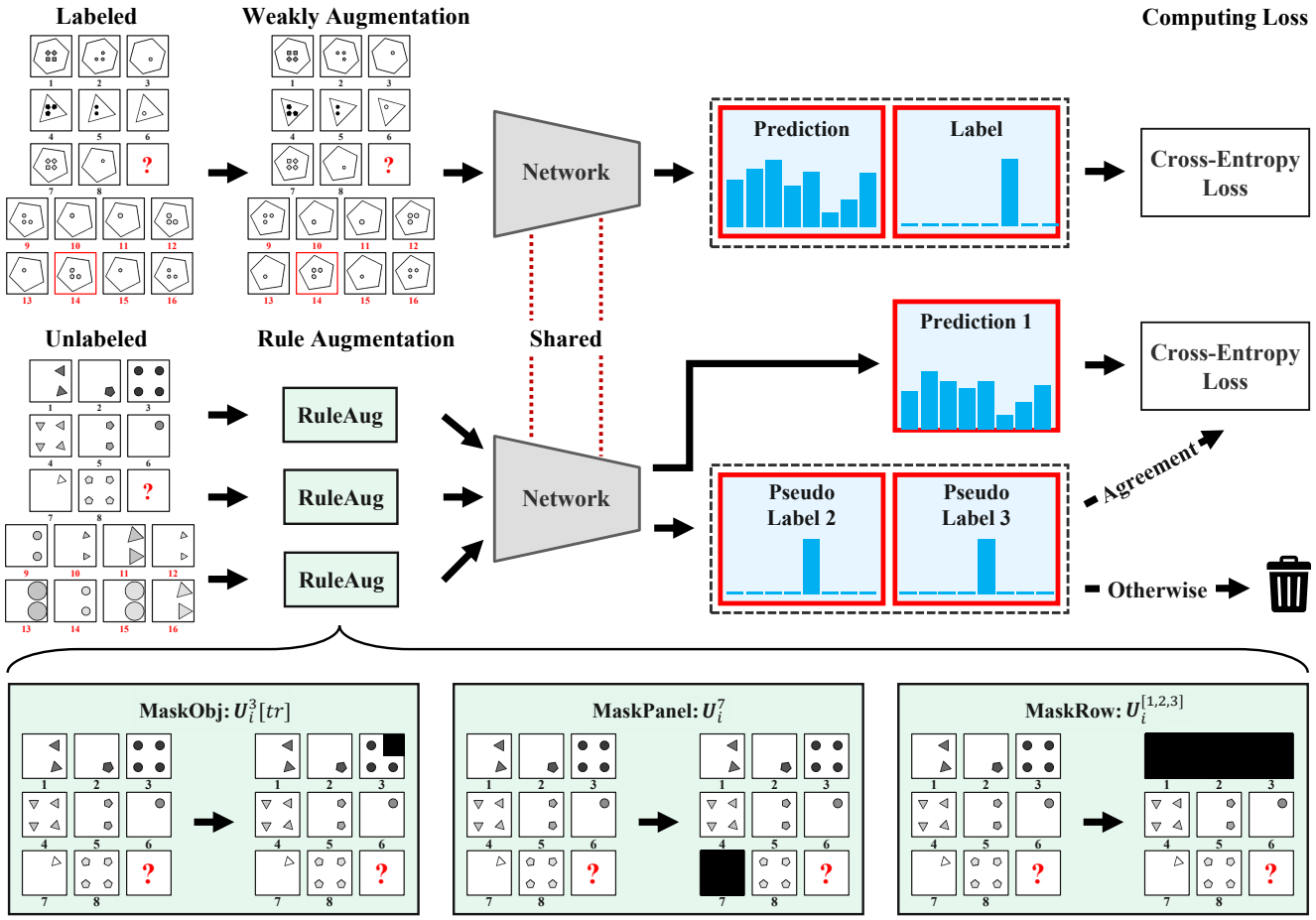

Figure 3: An overview of the proposed RuleMatch algorithm. Our RuleMatch uses the proposed rule-based augmentation to augment unlabeled data in three different views, and then employs the network to estimate pseudo labels from two of them. If the estimated labels are consistent across the two views, this unlabeled example and its pseudo label are included for network training, along with the weakly-augmented labeled data using supervised cross-entropy loss.

Based on our three strategies proposed above, we then formulate our rule-based augmentation method as $RuleAug = RS([MaskObj, MaskPanel, MaskRow])$. It is worth noticing that our augmentation is only applied to context images, which contain complete rules for each RPM question.

4 Experiments

In this section, we report results on two popular RAVEN benchmark datasets, I-RAVEN [Hu *et al.*, 2021a] and RAVEN-FAIR [Benny *et al.*, 2021], to verify our method.

4.1 Datasets & Implementations

RAVEN [Zhang *et al.*, 2019a] mimics RPM problems in human psychology and is one of the most popular datasets for evaluating visual abstract reasoning. However, a few recent studies [Hu *et al.*, 2021a; Benny *et al.*, 2021; Spratley *et al.*, 2020] found that a model trained only on the 8 answer images can achieve good results in RAVEN, which is suspicious and contradicts the spirit of abstract visual reasoning, since a correct answer must be reasoned from context images. These results point out the potential bias in the original RAVEN dataset. For a more fair evaluation, we run our method on two other variants of RAVEN, as depicted below.

I-RAVEN [Hu *et al.*, 2021a] and **RAVEN-FAIR** [Benny *et al.*, 2021] are two recently developed datasets to control for potential bias in the original RAVEN dataset. Both datasets have 7 image configurations – Center, 2x2Grid, 3x3Grid, Left-Right, Up-Down, Out-InCenter, and Out-InGrid. Each of them contains 10,000 problems, for a total of 70,000 RPM questions and 1,112,000 images. In addition, both RAVEN-FAIR and I-RAVEN contain the same sets of relationships, including progression, constant, union, and arithmetic. These relationships are thought to be very challenging for current deep networks. The only difference between the two datasets is the method used to generate negative answers. RAVEN-FAIR starts with the correct answer image in each RPM question and iteratively generates one negative example at a time by randomly changing one visual attribute of the correct answer. In I-RAVEN, a bisection tree is constructed to change one attribute at a time but in two different attribute directions. Moreover, recent studies [Benny *et al.*, 2021; Hu *et al.*, 2021a] have demonstrated that a model merely trained on the eight answer images can only perform slightly

Network	Training	I-RAVEN			RAVEN-FAIR		
		5000 (11.9%)	3000 (7.1%)	1000 (2.4%)	5000 (11.9%)	3000 (7.1%)	1000 (2.4%)
MRNet	Supervised	53.42±0.54	46.93±0.32	34.19±0.30	59.41±0.72	50.72±0.53	37.66±1.97
	MeanTeacher	60.60±3.35	52.10±0.88	32.22±1.66	68.19±2.24	61.32±0.72	**38.53±1.98**
	FixMatch	62.66±2.87	50.15±1.22	33.12±1.38	74.82±1.98	62.37±1.29	38.19±1.57
	RuleMatch	**69.85±1.84**	**57.82±0.91**	**34.18±1.11**	**78.15±1.82**	**69.11±0.99**	38.34±0.97
RelBase	Supervised	70.05±6.45	57.93±0.55	33.32±1.92	77.85±2.95	65.92±2.17	37.67±0.94
	MeanTeacher	80.30±1.52	74.10±2.12	31.59±0.85	87.92±1.86	83.94±2.28	38.34±3.14
	FixMatch	85.38±0.57	73.33±2.64	31.88±0.83	91.15±0.27	83.45±2.82	39.25±1.34
	RuleMatch	**88.14±0.45**	**77.33±1.08**	**34.21±0.98**	**92.11±1.61**	**86.43±1.21**	**39.33±1.27**

Table 1: Results of all compared training algorithms, including purely supervised training (denoted as Supervised), Mean Teacher Model, FixMatch, and the proposed RuleMatch. All results are reported under the early stopping control (20 epochs) and as the average over 3 experimental runs. The best results for each k-labeled setting are highlighted as **Bold.**

better than chance on both datasets, further confirming the validity of the two datasets.

To test whether our RuleMatch can be broadly applied to different kinds of reasoning networks. We choose two competing models – MRNet [Spratley *et al.*, 2020] and RelBase [Benny *et al.*, 2021]. MRNet explicitly infers row-wise and column-wise rules, and also proposes to discover rules in features at different spatial scales. RelBase uses frame-wise convolutions to integrate all context images and each answer image in order to discover their relationships. We generally follow their original implementations except that here 4 GPUs were used for model training here. In addition, MRNet uses a two-stage training pipeline. As our aim is not to demonstrate the state-of-the-art performance of MRNet *per se*, we remove the second stage for simplicity.

Both I-RAVEN and RAVEN-FAIR are divided into training, validation, and test splits. The validation set is used to select the best training checkpoint for evaluation. All models 80×80 images as input. We also flip images horizontally as weak augmentation. Optimization is done by the Adam solver [Kingma and Ba, 2015] with a learning rate of 1e-3 and a batch size of 32. Weight decay is 1e-5 for both datasets. To further reduce the risk of overfitting, we add a dropout layer with a probability of 0.1 on the residual branches. All of our methods are trained for 100 epochs or when the validation accuracy is not improved in the last 20 epochs.

4.2 Main Results

In this section, we compare the proposed RuleMatch with several well-established semi-supervised learning algorithms, including MeanTeacher [Tarvainen and Valpola, 2017] and FixMatch [Sohn *et al.*, 2020]. We also include the purely supervised learning regime for comparison. Experiments are conducted on the two RAVEN datasets mentioned above, with different numbers of N_l labeled examples, *i.e.*, 1000, 3000, and 5000. All unlabeled examples will be included for sample discovery in all semi-supervised learning.

Table 1 shows all the results. We draw three main conclusions. First, the proposed RuleMatch algorithm achieves the best performance on both datasets in most N_l settings. Specifically, both MRNet and RelBase networks trained with our RuleMatch obtain significant improvements over those trained with FixMatch in almost all 5000 and 3000 conditions. Second, all methods do not perform well on very limited training data (see results obtained with 2.4% data). Some methods even obtain diminished performance than the supervised regime. This is mainly because the models have to learn noisy rules and cannot fully utilize unlabeled data with such scarce labeled data, resulting in suboptimal training. Indeed, this phenomenon shows that abstract relationships in RPMs cannot be easily generalized. Third, RelBase and MRNet are two competing models. Rel-Base directly extracts relationships in all eight context images without special designs for row-wise and column-wise rules. In contrast, MRNet deliberately includes row-wise and column-wise relation modules. Our RuleMatch can improve both in most experimental settings, demonstrating good flexibility in dealing with different forms of reasoning networks. All these results show that RuleMatch is an efficient algorithm that can discover useful information by combining both labeled and unlabeled data.

4.3 Ablation Studies

We run several ablation experiments with our RuleMatch on both I-RAVEN and RAVEN-FAIR. The number of labeled data in both datasets is set to 5000. For simplicity, we do not change the augmentation method of labeled data as this setup has been well-demonstrated in FixMatch [Sohn *et al.*, 2020]. The main goal here is to investigate whether our *RuleAug* is useful for using unlabeled data.

To this end, we propose two variants of our RuleMatch. First, we use two weakly-augmented views of an unlabeled question to check the consistency of the prediction. If their prediction is consistent, we then use the proposed *RuleAug* to augment that unlabeled question for network training. The results of this approach on MRNet and RelBase are shown in the 2nd row of Table 2. Compared to our default setting of using all three *RuleAug* (1st row), including two for label consistency checking and one for network training, this variant

Network	Augmentations $\mathcal{X}$	$\mathcal{U}_1$	$\mathcal{U}_2$	$\mathcal{U}_3$	I-RAVEN 5000 labels	RAVEN-FAIR 5000 labels
MRNet	WeakAug	RuleAug	RuleAug	RuleAug	**69.85±1.84**	**78.15±1.82**
	WeakAug	RuleAug	WeakAug	WeakAug	68.01±2.39	75.84±1.74
	WeakAug	WeakAug	RuleAug	RuleAug	68.82±2.39	77.55±1.74
RelBase	WeakAug	RuleAug	RuleAug	RuleAug	**88.14±0.45**	**92.11±1.61**
	WeakAug	RuleAug	WeakAug	WeakAug	87.28±1.88	90.65±1.66
	WeakAug	WeakAug	RuleAug	RuleAug	88.32±1.49	91.89±1.69

Table 2: Results of the data augmentation applied to the unlabeled data. Since FixMatch demonstrates the use of the weakly-augmentation (WeakAug) method on the labeled data, we do not explore different types of augmentation for simplicity. Note that $\mathcal{U}_2$ and $\mathcal{U}_3$ are used to investigate rule consistency, $\mathcal{U}_1$ is used for model training together with pseudo labels. All results are reported under the early stopping control (20 epochs) and as the average of 3 experimental runs. The best results for each dataset are highlighted in **Bold.**

obtains worse performance. Second, we use two views augmented by our *RuleAug* for label consistency checking, and use weakly-augmented view of this data for network training. This variant achieves slightly better performance than the first choice, but still produces weaker performance than our full version. All of these results lead to two conclusions. First, typical image augmentation, such as random flipping, cannot systematically disrupt abstract rules, and may lead to less effective consistent checking. Our *RuleAug* is a more robust augmentation approach for RPM problems. Second, as presented in FixMatch, it is important to align the predicted distribution over two different levels of augmentation, *i.e., weakly- and strong-augmentation*. In this work, we align the two predicted distributions using randomly perturbed rules, which are more suitable for RAVEN problems due to the statistical characteristics of this task. Overall, all these results confirm the positive contribution of our *RuleAug* and the rule-based sample selection method.

5 Conclusions

In this work, we present RuleMatch, a simple but specifically tailored for semi-supervised learning of abstract visual reasoning tasks. The key components of our algorithm are the rule-based augmentation approach and the rule consistency loss function. The goal of our proposed rule-based augmentation is to disrupt abstract rules in a RAVEN question, allowing the network to learn more high-level rule-wise features. Furthermore, our algorithm automatically selects useful unlabeled data through a rule consistency mechanism applied to two views of the augmented unlabeled question. This consistent examination allows our algorithm to discover more reliable unlabeled data, resulting in surprisingly-high accuracy with just $\sim 10\%$ labeled data.

Our work has three limitations. First, our algorithm uses more augmented data, which inevitably increases training cost. Second, our RuleMatch cannot improve model performance when the model is trained on a very limited number of labeled data. This is mainly because extracting relationships across multiple images in RPMs is much more difficult than recognizing objects in a single image. Third, it is debatable whether humans use semi-supervised learning or few-shot learning when taking RAVEN tests. Although humans may not encounter over many unlabeled RPM questions, we argue that humans have already learned a vast number of abstract rules (*e.g.*, progression, constant, *etc*) through other contexts in everyday life. Another factor is that humans have also already possessed a powerful pre-trained visual system that can almost perfectly recognize low-level visual attributes of objects in an RPM question. Here, the image encoder part of these networks still requires extensive training. We highlight the differences between machines and humans in learning RPM problems. But our SSL work here is at least an important step towards improving data-efficient learning of abstract analogical reasoning in machines, and serves as a cornerstone for future work such as few-shot learning.

In summary, we believe that our proposed RuleMatch will help to enable the use of machine learning algorithms in such reasoning tasks, where labels of these tasks are usually expensive or difficult to obtain.

Acknowledgments

This work was partially supported by the NSFC Projects (62206316, 32100901, 32171094), the Guangdong NSF Project (2022A1515011254), the Shanghai NSF Project (21ZR1434700), the Sichuan NSF Project (2022NSFSC0527), the Shanghai Pujiang Program (21PJ1407800), the Research Project of Shanghai Science and Technology Commission (20dz2260300) and the Fundamental Research Funds for the Central Universities.

Contribution Statement

Y.X and L.Y, as co-first authors (*), made equal contributions to this study. Y.X., L.Y., and R-Y. Z. conceived and designed the research. Y.X., L.Y. implemented the models and performed the experiments. R-Y.Z. is the corresponding author ($^\#$). H.Y., Z.Z., D-H, W., X.W., and X.X provided valuable feedback during the research. X.W. and X.X. also provided computing resources for model implementation. Y.X., L.Y., and R-Y. Z. wrote the first draft. All the authors revised the manuscript and designed figures.

References

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

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

[Benny *et al.*, 2021] Yaniv Benny, Niv Pekar, and Lior Wolf. Scale-localized abstract reasoning. In *CVPR*, pages 12557–12565, 2021.

[Berthelot *et al.*, 2019] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, volume 32, pages 5049–5059, 2019.

[Bodla *et al.*, 2018] Navaneeth Bodla, Gang Hua, and Rama Chellappa. Semi-supervised fusedgan for conditional image generation. In *ECCV*, pages 669–683, 2018.

[Conneau *et al.*, 2016] Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*, 2016.

[Donahue *et al.*, 2014] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, pages 647–655. PMLR, 2014.

[Gibson *et al.*, 2013] Bryan R Gibson, Timothy T Rogers, and Xiaojin Zhu. Human semi-supervised learning. *Topics in cognitive science*, 5(1):132–172, 2013.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hu *et al.*, 2021a] Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. Stratified rule-aware network for abstract visual reasoning. In *AAAI*, volume 35, pages 1567–1574, 2021.

[Hu *et al.*, 2021b] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. Simple: similar pseudo label exploitation for semi-supervised classification. In *CVPR*, pages 15099–15108, 2021.

[Jahrens and Martinetz, 2020] Marius Jahrens and Thomas Martinetz. Solving raven's progressive matrices with multi-layer relation networks. In *IJCNN*, pages 1–6. IEEE, 2020.

[Jeong *et al.*, 2019] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *NeurIPS*, volume 32, pages 10759–10768, 2019.

[Kalish *et al.*, 2011] Charles W Kalish, Timothy T Rogers, Jonathan Lang, and Xiaojin Zhu. Can semi-supervised learning explain incorrect beliefs about categories? *Cognition*, 120(1):106–118, 2011.

[Karpathy *et al.*, 2014] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.

[Katsumata *et al.*, 2022] Kai Katsumata, Duc Minh Vo, and Hideki Nakayama. Ossgan: Open-set semi-supervised image generation. In *CVPR*, pages 11185–11193, 2022.

[Kaul *et al.*, 2022] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Label, verify, correct: A simple few shot object detection method. In *CVPR*, pages 14237–14247, 2022.

[Kim *et al.*, 2016] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, volume 25, pages 1097–1105, 2012.

[Laine and Aila, 2016] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[Lovett and Forbus, 2017] Andrew Lovett and Kenneth Forbus. Modeling visual problem solving as analogical reasoning. *Psychological review*, 124(1):60, 2017.

[Lovett *et al.*, 2010] Andrew Lovett, Kenneth Forbus, and Jeffrey Usher. A structure-mapping model of raven's progressive matrices. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32, 2010.

[Misra *et al.*, 2015] Ishan Misra, Abhinav Shrivastava, and Martial Hebert. Watch and learn: Semi-supervised learning for object detectors from video. In *CVPR*, pages 3593–3602, 2015.

[Raven and Court, 1938] John C Raven and JH Court. *Raven's progressive matrices*. Western Psychological Services Los Angeles, CA, 1938.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015.

[Sharif Razavian *et al.*, 2014] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR workshops*, pages 806–813, 2014.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Sohn *et al.*, 2020] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, volume 33, pages 596–608, 2020.

[Spratley *et al.*, 2020] Steven Spratley, Krista Ehinger, and Tim Miller. A closer look at generalisation in raven. In *ECCV*, pages 601–616. Springer, 2020.

[Tang *et al.*, 2021] Peng Tang, Chetan Ramaiah, Yan Wang, Ran Xu, and Caiming Xiong. Proposal learning for semi-supervised object detection. In *WACV*, pages 2291–2301, 2021.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, volume 30, pages 1195–1204, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 6000–6010, 2017.

[Wang *et al.*, 2020] Duo Wang, Mateja Jamnik, and Pietro Lio. Abstract diagrammatic reasoning with multiplex graph networks. *arXiv preprint arXiv:2006.11197*, 2020.

[Welling and Kipf, 2016] Max Welling and Thomas N Kipf. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2016.

[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.*, 2021] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *ICCV*, pages 3060–3069, 2021.

[Yamins *et al.*, 2014] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[Yang *et al.*, 2016] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *ICML*, pages 40–48. PMLR, 2016.

[Zhang *et al.*, 2019a] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *CVPR*, pages 5317–5327, 2019.

[Zhang *et al.*, 2019b] Chi Zhang, Baoxiong Jia, Feng Gao, Yixin Zhu, Hongjing Lu, and Song-Chun Zhu. Learning perceptual inference by contrasting. In *NeurIPS*, pages 1075–1087, 2019.

[Zhang *et al.*, 2021] Chi Zhang, Baoxiong Jia, Song-Chun Zhu, and Yixin Zhu. Abstract spatial-temporal reasoning via probabilistic abduction and execution. In *CVPR*, pages 9736–9746, 2021.

[Zheng *et al.*, 2019] Kecheng Zheng, Zheng-Jun Zha, and Wei Wei. Abstract reasoning with distracting features. In *NeurIPS*, pages 5842–5853, 2019.

[Zhu and Goldberg, 2009] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

[Zhu *et al.*, 2007] Xiaojin Zhu, Timothy Rogers, Ruichen Qian, and Chuck Kalish. Humans perform semi-supervised classification too. In *AAAI*, volume 2007, pages 864–870, 2007.

[Zhu *et al.*, 2010] Xiaojin Zhu, Bryan R Gibson, Kwang-Sung Jun, Timothy T Rogers, Joseph Harrison, and Chuck Kalish. Cognitive models of test-item effects in human category learning. In *ICML*, pages 1247–1254, 2010.

[Zhu, 2005] Xiaojin Zhu. Semi-supervised learning literature survey. *Computer Sciences, University of Wisconsin-Madison*, 10:1–60, 2005.

[Zhuang *et al.*, 2021] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

[Zhuo and Kankanhalli, 2022] Tao Zhuo and Mohan Kankanhalli. Effective abstract reasoning with dual-contrast network. *arXiv preprint arXiv:2205.13720*, 2022.