Learning Visual Abstract Reasoning through Dual-Stream Networks

Kai Zhao[1], **Chang Xu**[1], **Bailu Si**[1, 2*]

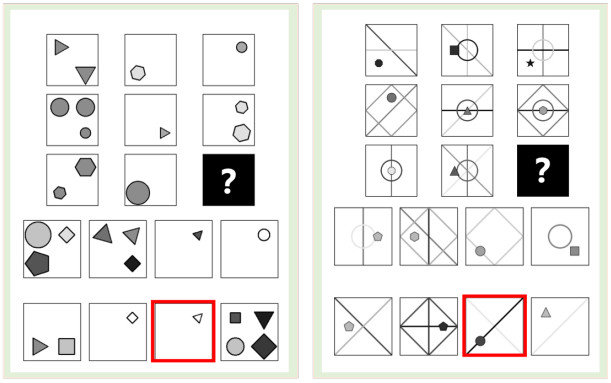
[1]School of Systems Science, Beijing Normal University
[2]Chinese Institute for Brain Research, Beijing
zhaokai_id@foxmail.com, changxu@mail.bnu.edu.cn, bailusi@bnu.edu.cn

Abstract

Visual abstract reasoning tasks present challenges for deep neural networks, exposing limitations in their capabilities. In this work, we present a neural network model that addresses the challenges posed by Raven's Progressive Matrices (RPM). Inspired by the two-stream hypothesis of visual processing, we introduce the Dual-stream Reasoning Network (DRNet), which utilizes two parallel branches to capture image features. On top of the two streams, a reasoning module first learns to merge the high-level features of the same image. Then, it employs a rule extractor to handle combinations involving the eight context images and each candidate image, extracting discrete abstract rules and utilizing an multilayer perceptron (MLP) to make predictions. Empirical results demonstrate that the proposed DRNet achieves state-of-the-art average performance across multiple RPM benchmarks. Furthermore, DRNet demonstrates robust generalization capabilities, even extending to various out-of-distribution scenarios. The dual streams within DRNet serve distinct functions by addressing local or spatial information. They are then integrated into the reasoning module, leveraging abstract rules to facilitate the execution of visual reasoning tasks. These findings indicate that the dual-stream architecture could play a crucial role in visual abstract reasoning.

Introduction

One goal of artificial intelligence (AI) is to equip machines with universal reasoning capabilities. Presently, deep learning has emerged as the dominant paradigm in AI, enabling the modeling of data to execute intricate tasks such as image classification (He et al. 2016; Dosovitskiy et al. 2021), object recognition (Girshick et al. 2015; Ronneberger, Fischer, and Brox 2015), and natural language processing (Vaswani et al. 2017). In the field of cognitive science, analogical reasoning has consistently been regarded as the foundation of general intelligence that sets humans apart from animals and is considered the essence of cognition. It is often shaped by the interplay between higher cognitive abilities and the quality of incoming representations (Norman, 1975). However, current deep learning systems still struggle to excel in tasks that demand analogical and relational reasoning.



(a) One example from RAVEN (b) One example from PGM

Figure 1: Examples from RAVEN and Procedurally Generated Matrices (PGM) are shown in (a) and (b) respectively. Both types of problems involve presenting participants with eight context images. They are required to select the correct answer (highlighted in red) from the candidate set of eight images to fill in the blank (denoted by **?**), in order to satisfy specific rules in the row or column direction of the 3×3 matrix.

As a significant assessment tool in the realm of analogy reasoning, the RAVEN test boasts an 80-year history (Court 1982; Prabhakaran et al. 1997; Perfetti et al. 2009). Researchers from various disciplines, including psychology, cognitive science, and artificial intelligence, have extensively explored this area. In recent years, preceding works (Zhang et al. 2019a; Hu et al. 2021; Benny, Pekar, and Wolf 2021) have generated program-controlled RPM datasets, as depicted in Figure 1. These efforts have greatly facilitated deep learning research in this domain and have assessed the analogy reasoning capability of deep learning systems (Małkiński and Mańdziuk 2022a). Serving as a widely accepted benchmark for intelligence evaluation, RPM problem requires participants to identify one or more rules within a 3×3 matrix, and then make a correct choice. This process of abstract reasoning mirrors the attributes of advanced human intelligence (Snow and Lohman 1984; Snow et al. 1984; Jaeggi et al. 2008).

Inspired by the two-stream hypothesis (Goodale and Mil-

*Corresponding author.

ner 1992; Grezes and Decety 2002; Maguire, Burgess, and O'Keefe 1999) in neuroscience, we propose a Dual-stream Reasoning Network (DRNet) to address the RPM problems. DRNet simulates object recognition through the ventral stream and spatial attention through the dorsal stream in the context of dual-stream vision. It extracts high-level visual features from these two streams. After performing fusion on the extracted visual features, DRNet feeds these features into the rule extractor to infer relationships between images, resulting in abstract rule representations. DRNet utilizes these rule representations to predict the correct answers. Codes are available at https://github.com/VecchioID/DRNet.

We conduct comprehensive empirical studies on several RPM benchmarks. To summarize, our contributions include:

- Unlike previous single-stream frameworks, DRNet combines the advantages of local and spatial representations, allowing it to exhibit distinct interpretations of input images. This collective enhancement improves the model's reasoning performance.
- DRNet achieves remarkable generalization performance and outperforms other models on multiple datasets, showcasing the effectiveness of this framework for non-verbal visual abstract reasoning problems.
- Visualization results of the rule representations indicate that the learned representations can be clustered based on rule categories, thereby facilitating visual abstract tasks.

Related Work

Raven's Progressive Matrices

Most previous work on RPMs has focused on single-stream network frameworks, such as ConvNets with inductive bias and ViTs focused on self-attention.

Early attempts at deep learning for RPM used a convolutional neural network (CNN) deep learning model by Hoshen and Werman (Hoshen and Werman 2017). Modern architectures such as WReN (Santoro et al. 2018), CoPINet (Zhang et al. 2019b), Rel-Base (Spratley, Ehinger, and Miller 2020), SRAN (Hu et al. 2021), MRNet (Benny, Pekar, and Wolf 2021), PredRnet (Yang et al. 2023), etc. use variants of convolutional neural networks (LeCun et al. 1998; He et al. 2016) for feature extraction. This suggests that the inductive bias can potentially generalize well in different configurations (Santoro et al. 2017; Jahrens and Martinetz 2020; Zhuo and Kankanhalli 2020; Zhang et al. 2022b; Mondal, Webb, and Cohen 2023; Małkiński and Mańdziuk 2022b). Among all previous studies, SCL (Wu et al. 2020) uses the compositional representation of object attributes and their relations for reasoning. Some symbolically inspired models to incorporate logical rule or object vectors into the unidirectional flow framework, e.g. PrAE (Zhang et al. 2021), ALANS learner (Zhang et al. 2022a) and NVSA (Hersche et al. 2023).

Another line of work focuses on attention mechanisms (Hahne et al. 2019; Rahaman et al. 2021; Mondal, Webb, and Cohen 2023; Sahu, Basioti, and Pavlovic 2022; Ma et al. 2022). A recent study shows that visual transformers retain more spatial information than CNNs (Raghu et al. 2021). Previous studies such as dynamic inference with neural interpreters (Rahaman et al. 2021) and STSN (Mondal, Webb, and Cohen 2023) explored the modular network architecture and image representations for abstract visual reasoning.

There are also some non-single stream studies, such as graph neural networks and reinforcement learning. MXGNet (Wang, Jamnik, and Liò 2020) proposes a multilayer graph neural network for multi-panel diagrammatic reasoning tasks, while LEN (Zheng, Zha, and Wei 2019) demonstrates a reinforcement learning teacher model to guide the training process.

Two Stream Networks

Inspired by the two stream hypothesis, two stream networks are widely explored in the video action recognition field (Simonyan and Zisserman 2014; Carreira and Zisserman 2017; Feichtenhofer, Pinz, and Zisserman 2016; Zolfaghari et al. 2017). I3D (Carreira and Zisserman 2017) builds a two stream 3D-CNN architecture and takes RGB video and optical flow as inputs. SlowFast (Feichtenhofer et al. 2019) is a model that encodes videos with different frame rates. DS-Net (Mao et al. 2021) uses a dual-stream network to explore the representation capacity of local and global pattern features for image classification. Chen et al. employ a dual visual encoder containing two separate streams to model both the raw videos and the key-point sequences for sign language understanding (Chen et al. 2022). We introduce dual-stream networks to abstract visual reasoning tasks. Although our work uses dual-stream networks like previous studies, we have a different implementation. First, dual-stream network was implemented as a dual encoder module to extract high level image features across different images. Second, we introduce a reasoning module in DRNet, which allows it to extract rules from different RPM problems.

How to model the interactions between different streams is non-trivial. I3D (Carreira and Zisserman 2017) uses a late fusion strategy by simply averaging the predictions of two streams. Another way is to fuse the intermediate features of each stream in the early stage by lateral connections (Feichtenhofer et al. 2019), concatenation (Zhou et al. 2021), or addition (Cui, Liu, and Zhang 2019). In this work, our approach directly models local and spatial features via a dual encoder, and then a learnable module is used to fuse the intermediate features of each stream.

Methods

The structure of DRNet is shown in Figure 2. It consists of two components: (1) a dual encoder module to transform each image into two high-level features, and (2) a reasoning module to score each context candidate group's features. The highest score is selected as the final predicted answer.

Dual Encoder Module

This module consists of two parallel streams, where a CNN is used to recognize objects to acquire local features, while ViT is used to play a role in attending to the spatial location of objects.

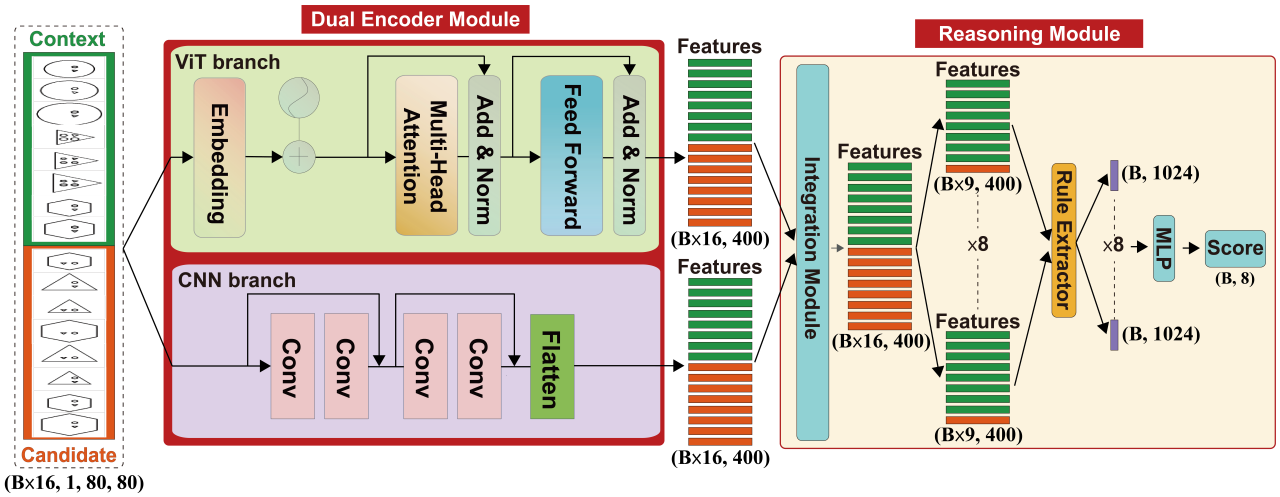


Figure 2: An overview of our DRNet. DRNet consists of a dual encoder module and a reasoning module, where (B×16, 1, 80, 80) represents (batchsize×16, channels, image size, image size). The dual encoder module is used to extract input image features in parallel, after which the features are fed into the reasoning module. The reasoning module first learns to merge the high-level embeddings of the same image. Then, it employs a rule extractor to handle combinations involving the eight context images and each candidate image, extracting abstract rules and utilizing an MLP to make predictions.

CNN branch. Our CNN stream has two ResBlocks, each containing a residual branch and a shortcut connection. Each residual branch has two convolutional layers with kernel sizes of 7. Each convolutional layer down-samples the input feature with a stride of 2, which expands the receptive fields of the neurons and allows for the extraction of higher-level information. The shortcut connection applies the $MaxPool2d$ operation twice to match the output size of the residual branch with a stride of 2. In total, our first ResBlock can be formulated as:

$$x^l = ReLU(BN(Conv_{7\times7}(x^{l-1}))), l \in 1, 2 \quad (1)$$

$$x^l = x^l + Maxpool2d(x^0), l = 2 \quad (2)$$

where x^0 represent input features, l represents the layer index of convolutional layers. x^{cnn_out} can be obtained by treating x^2 in the same way as above through the second ResBlock. We set the filters as [64, 64, 64, 16] from the first to the last convolutional layer.

ViT branch. ViT branch processes each image parallelly. Our ViT has the same network framework as in (Vaswani et al. 2017), except that we employ 1D learnable positional encodings to add them to patch embeddings for retaining positional information. Our ViT has 8 attentional heads with depth of 12. We first split each image into 16 patches, with each patch size of 20×20. One convolutional layer with a kernel of 20 and a stride 20, is applied to transform into a patch embedding with size of 400. After transformer encoder, we finally obtained an averaged feature vector. The ViT branch can be formulated as:

$$x^{vit\_out} = ViT(x^0) \quad (3)$$

where x^{vit_out} represents the output features of ViT branch.

To clarify the process, we describe the data flow illustrated in Figure 2. For each RPM problem, we have 8 context images I_i^c, where $i \in \{1, 2, ..., 8\}$, and 8 candidate images I_i^a, where $i \in \{1, 2, ..., 8\}$, which are combined to create an input denoted as $I = [I_1^c, I_2^c, ..., I_8^c, I_1^a, I_2^a, ..., I_8^a]$, with $I \in \mathcal{R}^{(16\times1\times80\times80)}$. This input I is then simultaneously fed into both the CNN branch and the ViT branch with a batch approach. The result is image features: $x^{cnn\_out} \in \mathcal{R}^{(B\times16,1,20,20)}$ from the CNN branch and $x^{vit\_out} \in \mathcal{R}^{(B\times16,400)}$ from the ViT branch. To ensure compatibility, we reshape x^{cnn_out} to $\mathcal{R}^{(B\times16,400)}$, aligning its shape with that of x^{vit_out}. Here, B represents the batch size. Finally, the outputs x^{vit_out} and x^{cnn_out} are passed to the reasoning module.

Reasoning Module

This reasoning module consists of an integration module and a rule extractor to fuse high-level features and extract abstract rule representations of RPM problems.

Integration Module. Just as the two streams ultimately project to the hippocampus (Huang et al. 2021), DRNet designs an integration module to model the interactions between different streams. To promote order-invariance between the two vectors, a permutation-invariant operator is recommended. One can use the sum operator (**SUM**). This approach has been employed by (Niebur and Koch 1995; Benny, Pekar, and Wolf 2021).

$$\mathbf{SUM}(\cdot) := x^{cnn\_out} + x^{vit\_out} \quad (4)$$

To reduce the variance of **SUM**, a mean (**MEA**) operator is defined as follows:

$$\mathbf{MEA}(\cdot) := (x^{cnn\_out} + x^{vit\_out})/2 \quad (5)$$

As a variation of the above operators, we propose an adaptive attention operator **AUT** to automatically combine two streams,

$$\mathbf{AUT}(\cdot) := w_1 x^{cnn\_out} + w_2 x^{vit\_out} \quad (6)$$

Where w_1 and w_2 are learnable tensors. We provide three methods for determining changes in these two parameters: $\mathbf{AUT-L1}$ normalization, $\mathbf{AUT-L2}$ normalization, and the unrestricted way ($\mathbf{AUT}$).

In the last approach, we concatenate these two streams and implement the learnable attention operator **LIN** using a linear layer:

$$\mathbf{LIN}(\cdot) := concat(x^{cnn\_out}, x^{vit\_out})A^T + b \quad (7)$$

We compare different operators in Figure 3 and adopt the **LIN** operator in DRNet.

After feature fusion, we split the fused features x into two groups: e_i and c_i, where $i \in 1, 2, ..., 8$. We then concatenate each c_i with the 8 context features to form $r_i = [e_1, e_2, ..., e_8, c_i]$, with $r_i \in \mathcal{R}^{(B,i,9,400)}$. Next, we pass r_i into the rule extractor to infer the relationships between the nine feature vectors, as depicted in the reasoning module of Figure 2.

Rule Extractor. The rule extractor consists of two ResBlocks. Each residual branch has two 1D convolutional layers with a kernel size of 7. Each convolutional layer learns to expand the receptive fields of the neurons to extract higher-level relations with a stride of 1. The shortcut connection applies a 1D convolutional layer to the two ResBlocks with a kernel size and stride of 1. In total, our rule extractor can be formulated as:

$$r_i^l = ReLU(BN(Conv_7(r_i^{l-1}))), l \in 1, 2 \quad (8)$$

where l represents the layer index of convolutional layers. We set the filters to $[64, 128]$ for the first and second convolutional layers. After the skip connection, we apply $MaxPool1d(r_i^1)$ to reshape $r_i \in \mathcal{R}^{(B,i,128,400)}$ into $r_i \in \mathcal{R}^{(B,i,128,100)}$. Then, we send r_i to the second ResBlocks as follows:

$$r_i^l = ReLU(BN(Conv_7(r_i^{l-1}))), l \in 3, 4 \quad (9)$$

We set the filters to $[128, 64]$ for the third and fourth layers. After the skip connection, we apply $AdaptiveAvgPool1d(r_i^1)$ to reshape $r_i \in \mathcal{R}^{(B,i,64,100)}$ into $r_i \in \mathcal{R}^{(B,i,64,16)}$. Finally, we flatten $r_i \in \mathcal{R}^{(B,i,64,16)}$ into $r_i \in \mathcal{R}^{(B,i,1024)}$ to obtain 8 embeddings. The embeddings corresponding to the correct labels are both abstract representations of the rules.

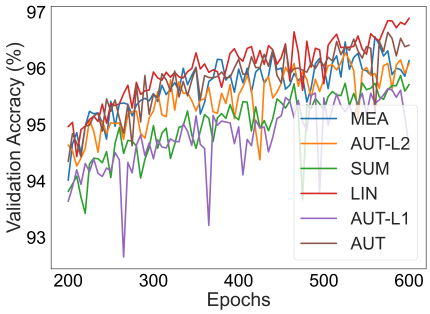


Figure 3: The performance analysis of various operators in DRNet on the RAVEN dataset reveals that the **LIN** operator outperforms its counterparts.

Classifier. Lastly, we use an $\mathbf{MLP}$ consisting of three linear layers to score these features, and the highest score determines the best answer:

$$Answer = \underset{i \in \{1,...,8\}}{\arg\max}[\mathbf{MLP}(r_i)] \quad (10)$$

Between every two linear layers, we have added an ELU function and a $BatchNorm1d$ layer, with a dropout probability of 0.5. For each linear layer, the output dimensions are 512, 256, and 1 respectively.

Experiments

Datasets

The **PGM** dataset (Santoro et al. 2018) comprises 1.2 million training samples, 20 thousand validation samples, and 200 thousand test samples. Each panel within the PGM dataset varies in terms of types, sizes, colors, and shapes. There are 1-4 rules per row or column for each matrix panel. The PGM dataset includes 8 regimes, with 7 of them involving interpolation, extrapolation, held-out attribute pairs (HO AP), held-out pairs of triples (HO TP), held-out triples (HO Triples), held-out line-type (HO LT), and held-out shape-color (HO SC) scenarios. These regimes systematically assess out-of-distribution (OOD) generalization using various approaches.

RAVEN-style Datasets. The RAVEN dataset (Zhang et al. 2019a), along with its variants I-RAVEN (Hu et al. 2021) and RAVEN-FAIR (Benny, Pekar, and Wolf 2021), are compact datasets, each comprising 7 configurations, with each configuration containing 10,000 samples. The distribution ratio across the training, test, and validation sets is 6:2:2. Every problem within these datasets features 4-8 rules in each row/column. Each problem is presented with 8 context panels arranged in an incomplete 3 x 3 matrix, alongside 8 candidate answer panels. We trained DRNet jointly on all configurations in each RAVEN-style dataset.

Implementation Details

All datasets include training, validation, and test sets. We utilize a standard batch size of 256 and evaluate the reported accuracy on the test set using the best validation accuracy checkpoint. The same set of hyperparameters is applied across all benchmarks, employing Adam (Da 2014) optimizer with a learning rate of 3e-4, β values of (0.9, 0.999), and a weight decay of 1e-6. No additional supervision signals, such as metadata, are utilized during training. Additionally, for RAVEN-style datasets, we present the median outcome from 5 distinct runs. Given the computational demands of training on large-scale PGM datasets, we provide a single result, aligning with the approach of prior works (Zhang et al. 2019b; Benny, Pekar, and Wolf 2021; Yang et al. 2023). In the experimental results presented below, when the validation loss no longer decreases within 20 epochs, we perform early stopping.

Method	WReN	CoPINet	MRNet	SCL	MLRN	Rel-Base	ARII	STSN	PredRNet	NVSA	DRNet
PGM-N	62.6	56.4	94.5	88.9	98.0	85.5	88.0	98.2	97.4	-	**99.06**
RAVEN	16.8	91.4	96.6	91.6	12.3†	91.7	-	89.7†	95.8	87.7	**96.89**
I-RAVE	23.8	46.1	83.5†	95.0	12.3†	91.1†	91.1	95.7†	96.5	88.1	**97.62**
RAVE-F	30.3	50.6	88.4	90.1†	29.5†	93.5†	-	95.4	97.1	-	**97.58**
Average	33.4	61.1	90.8	91.4	38.0	90.5	-	94.8	96.7	-	**97.79**

Table 1: Recognition accuracy (%) on PGM Neutral (PGM-N), RAVEN, I-RAVEN (I-RAVE), and RAVEN-FAIR (RAVE-F). For all RAVEN-style datasets, accuracy is obtained by averaging across all seven configurations. † indicates that the results were not reported in the original paper; we obtained these results from Table 1 of (Yang et al. 2023). The best results for each dataset are highlighted in bold font.

Main Results

We conducted experiments on PGM, RAVEN, I-RAVEN and RAVEN-FAIR, all of which have predefined training, validation and test data splits. During training, we used the training set for model training and the test set for evaluation, while the validation set was used to select the optimal checkpoint for evaluation. We used vertical/horizontal flip data augmentation with a probability of 0.3 for RPM training samples.

State-of-the-art Comparisons. We compare DRNet with several previous models, including WReN (Santoro et al. 2018), CoPINet (Zhang et al. 2019b), MRNet (Benny, Pekar, and Wolf 2021), SCL (Wu et al. 2020), MLRN (Jahrens and Martinetz 2020), Rel-Base (Spratley, Ehinger, and Miller 2020), ARII (Zhang et al. 2022b), STSN (Mondal, Webb, and Cohen 2023) and PredRNet (Yang et al. 2023). We have also compared our method with end-to-end symbolic methods such as NVSA (Hersche et al. 2023). Experiments were conducted on PGM *Neutral* and three RAVEN-style datasets.

Table 1 shows the main results on four datasets. First, our DRNet achieves the best average performance on the four datasets compared to single-stream models, such as STSN and PredRNet. STSN introduces slot attention to extract image-wise features and then proposes a transformer-based module to explore relationships between contexts and choices for reasoning. PredRNet introduces prediction error into ConvBlocks to improve reasoning performance. PredRNet provides the best average performance (96.7%) among all compared methods. While DRNet outperforms PredRNet with an average performance of 97.78%. In addition, DRNet actually achieves better performance on the RAVEN (1.09%), I-RAVEN (+1.12%), and RAVEN-FAIR (+0.48%) datasets than PredRNet, respectively. Some recently proposed methods, such as STSN, MLRN, and ARII, only show good results on one or two datasets. For example, CoPINet and MLRN only perform well on RAVEN (91.4%) and PGM-N (98.0%) respectively. In contrast, DRNet shows superior results on all 4 datasets, which suggests the generalization performance on different datasets. Second, compared to certain competitive models like MRet, our approach extracts rules only from combinations of 8 image sets, without the need for column rule learning. If we remove the ViT branch in DRNet, our model is similar to Rel-Base. If we remove the CNN branch, our model completely degenerates into an attention-based model. However, neither branch performed well enough (see ablation experiment), indicating that the two-stream design of our model was critical and helped our model achieve an average result of 97.78%.

Out-of-Distribution Generalization in PGM. In addition to the Neutral dataset of PGM, the remaining seven datasets were employed to evaluate the model's capacity to handle out-of-distribution scenarios. We assessed DRNet's performance across all sub-datasets of PGM while maintaining consistent model settings. The outcomes of these evaluations are meticulously documented in Table 2. A careful examination of the data presented in Table 1 reveals that our proposed model exhibited only a marginal average enhancement of 1.09% for in-distribution datasets. However, when confronting challenges posed in out-of-distribution scenarios, our model showcased a notable enhancement, reaching up to 11.23%. This highlights the robust learning prowess inherent in the dual-stream architecture, enabling it to effectively handle OOD challenges.

Ablation Experiments

We conducted ablation experiments on both the I-RAVEN dataset, representing in-distribution, and the PGM HO AP dataset, representing OOD. Since the three RAVEN-style datasets have similar distributions, we only tested on one.

Different hyper-parameters. We evaluated the impact of different depths of ViTs and various sizes of convolutional kernels on model performance. We selected ViT depths of 4, 8, 12 (DRNet), and 16. Regarding convolution, we investigated common kernel sizes: 3, 5, and 7 (DRNet). The results are presented in Table 3. DRNet demonstrates a gradual improvement in performance with increasing ViT depth. However, when the depth is 16, the performance of DRNet on HO AP decreases from 93.7% to 89.7%. Additionally, DRNet's performance gradually improves with larger convolutional kernel sizes on both datasets. These results indicate that the current hyperparameters for DRNet are optimal.

Single stream *v.s.* Dual stream. Compared with single-stream models like MRNet and PredRNet, DRNet performs very well, especially in the OOD scenario. To help understand our proposed model, DRNet, the first thing we aimed to clarify is the role of each branch. The results are shown in Table 4.

Each stream achieves a recognition accuracy of over 80% on the in-distribution I-RAVEN dataset. Data augmentation (DA) helps the model achieve higher performance on

Method	Neut	Intr	Extr	HO AP	HO TP	HO Tri	HO LT	HO SC	Average
WReN	62.6	62.4	17.2	27.2	41.9	19.0	14.4	12.5	32.4
ARII	88.0	72.0	**29.0**	50.0	64.1	32.1	16.0	12.7	45.49
MXGNet	66.7	65.4	18.9	33.6	43.3	19.9	16.7	16.6	35.14
MRNet	93.4	68.1	19.2	38.4	55.3	25.9	**30.1**	**16.9**	43.41
PredRNet	97.4	70.5	19.7	63.4	67.8	23.4	27.3	13.1	47.1
DRNet	**99.06**	**83.78**	22.22	**93.74**	**78.11**	**48.77**	27.92	13.09	**58.33**

Table 2: Recognition accuracy (%) across all regimes of PGM. PGM comprises 1 Neutral and 7 OOD sub-datasets. Neut: Neutral, Inter: Interpolation; Extr: Extrapolation; HP AP: Held-Out Attribute Pairs; HO TP: Held-Out Triple Pairs; HO Tri: Held-Out Triples; HO LT: Held-Out Line Type; HO SC: Held-Out Shape Color. The best results are highlighted in bold font.

	ViT Depth				Conv kernel		
Dataset	4	8	**12**	16	3	5	**7**
I-RAVE	96.4	96.0	**97.6**	96.2	90.9	95.8	**97.6**
HO AP	93.4	93.4	**93.7**	89.7	84.6	91.5	**93.7**

Table 3: Recognition accuracy (%) of different hyper-parameters on I-RAVE and PGM HO AP. The results, in bold, show that the hyper-parameters used in DRNet are optimal. I-RAVE represents I-RAVEN; HO AP corresponds to PGM Held-Out Attribute Pairs.

ViT	CNN	DA	I-RAVE	HO AP
√	×	×	83.62	67.53
√	×	√	87.52	72.26
×	√	×	90.83	58.73
×	√	√	95.50	62.87
√	√	×	91.68	90.46
√	√	√	**97.62**	**93.74**

Table 4: Recognition accuracy (%) of different branches across various datasets. CNN refers to the CNN branch mentioned earlier; DA stands for Data Augmentation, and we used vertical/horizontal flip data augmentation with a probability of 0.3 for RPM training samples.

such datasets; a single CNN stream with data augmentation achieves a recognition accuracy of 95.50%, surpassing many previous studies. The contribution of the added ViT stream to the performance improvement in DRNet appears to be low. Hence, we also conducted an ablation experiment on the OOD PGM HO AP dataset. As shown in Table 4, the recognition accuracy for ViT+DA is 72.26%, while for CNN+DA, it is only 62.87%. The performance of the single stream is far inferior to the dual-stream architecture of DRNet, indicating that the dual-stream network enhances the model's performance on OOD datasets.

Dual CNN Stream *v.s.* Dual ViT Stream. Subsequently, we embarked on the replacement of the network's branches to ascertain whether any dual-stream model comprising distinct network components possesses adept relational reasoning capabilities. Given that the tensor shapes of input and output for each stream in DRNet remain consistent, our focus is centered on assessing whether the parameter count of the new networks was comparable to that of DRNet. For simplicity of description, we define the CNN branch network in DRNet as CNN-base and the ViT branch as ViT-base.

Method	I-RAVE	HO AP
DCNet(24M)	**98.24**	73.78
DVNet-s(26M)	87.52	92.60
DVNet-h(47M)	97.32	93.24
DRNet-P(3.4M)	96.06	91.23
DRNet(24.6M)	97.62	**93.74**

Table 5: Recognition accuracy (%) of different dual-stream networks on I-RAVEN and PGM HO AP datasets, where DCNet refers to the Dual-CNN network, DVNet refers to the Dual-ViT network, and DRNet-P refers to the patch-based DRNet.

First, we replace the ViT-base in DRNet with ResNet-32 to form a dual-CNN network (DCNet-24M), where M represents learnable parameters, in millions. For the ResBlocks, we set filters to $[64 \times 3, 128 \times 4, 256 \times 6, 512 \times 3]$, where $[3, 4, 6, 3]$ represents the repeat times of ResBlocks, and all convolutional layers in the ResBlocks have a kernel size of 7. The first convolutional layer in ResNet-32 has a kernel size of 3, and the filter is set to 64.

Second, we replace the CNN-base in DRNet with a shallow ViT-small of depth 1 to create a dual-ViT network (DVNet-s-26M). The network architecture of ViT-small is the same as ViT-base.

Third, we use two ViT-b models to construct a larger dual-ViT network (DVNet-h-47M). Fourth, DVNet-h increases the number of parameters drastically. Given that we are applying attention to the sum of ViT-base and CNN-base, we instead utilized a patch-based CNN from (Brutzkus et al. 2022) , replacing ViT-base to create a patch-based DRNet (DRNet-P-3.4M). We tested the four aforementioned network configurations on the I-RAVEN and PGM HO AP datasets, and the results are shown in Table 5.

From Table 5, it can be observed that DCNet performs excellently on the I-RAVEN dataset (98.24%), surpassing DRNet (97.62%). However, DCNet's performance drastically declines on the OOD PGM HO AP dataset, achieving only 73.78%. Nevertheless, DCNet's results still outperform those of previous state-of-the-art models, such as PredRNet. In contrast, DVNet-s performs well in both the I-RAVEN and OOD paradigms, indicating that spatial attentional representations at different scales are crucial for the

abstract reasoning process. As the number of learnable parameters in DVNet-h increases, its performance aligns with that of DRNet, but at a higher computational cost. This suggests that DRNet strikes an effective balance between computational complexity and performance. Compared to other baseline models, DRNet has a larger number of parameters. Therefore, we introduced a small-parameter version called DRNet-P. As shown in Table 5, DRNet-P exhibits a significant reduction in parameters by 86.2% (24.6M→3.4M) compared to DRNet. Despite this reduction, its performance only slightly decreased by 1.56% on the I-RAVE dataset and by 2.51% on the PGM HO AP dataset. This indicates the potential of dual-stream networks in abstract visual reasoning.

Rule Representations

Although DRNet achieves a high recognition accuracy, we still lack an understanding of its reasoning process. All rule extraction takes place within the Rule Extractor. Therefore, we conducted a t-SNE (van der Maaten and Hinton 2008) analysis of the embeddings corresponding to the correct answers of the rule extractor. Due to the variety of rules covered by RAVEN problems, it is challenging to visualize vectors with multiple rule labels. Therefore, we selected the PGM Neutral dataset for rule visualization due to its smaller number of rules. Based on the actual rule types of each problem (AND, OR, XOR, Consistent union, Progression), we visualized 200k test samples, and the results are shown in Figure 4. The rule extractor module can identify and form abstract rule representations for downstream classification task.

Furthermore, we utilized radial basis function kernel principal component analysis to reduce the 200k rule representations from 1024 to 768 dimensions, preserving spatial information and reducing redundancy. Subsequently, we computed pairwise cosine similarities for the 5 rule categories, resulting in 10 sets of scores. The mean values of these scores range from -0.01 to 0.007, indicating near orthogonal representations and this may be a reason for the superior performance of DRNet.

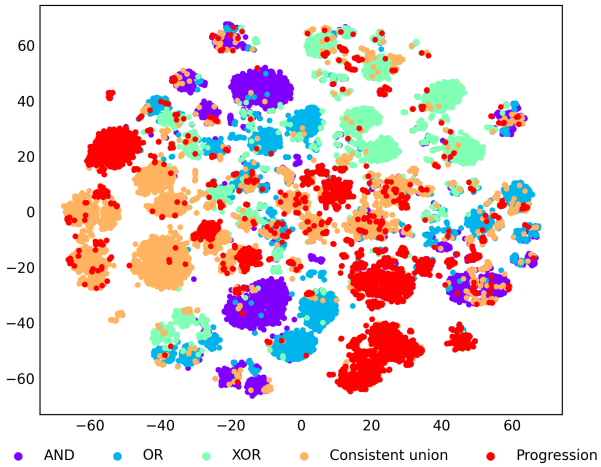


Figure 4: T-SNE visualization of abstract rules in PGM Neutral. The clustered embedding of similar abstract rules indicates that the Reasoning Module in DRNet is adept at discovering rules.

Visualization of Two Streams

Cadieu and Olshausen shows that learning both ventral and dorsal-like representations in a single ANN with two pathways is possible if one forces the two pathways to process separately the phase and amplitude of a complex decomposition of the stimuli (Cadieu and Olshausen 2012).

To begin to understand how the two streams in DRNet process images, we analyze their internal representations. Self-attention allows ViT to integrate information across the entire image even in the lowest layers. We investigate to what degree the network makes use of this capability. For the analysis of ViT, we adopt the approach used in (Vaswani et al. 2017). We find that ViT attends to image regions that are relevant for spatial information, as shown in Figure 5 (a), working like the *where* pathway.

For the visualization of CNN convolutional layers, we obtained the results through a single forward pass, as shown in Figure 5 (b). We found that during the learning process, CNN gradually acquires high-level image representations by combining local features, working like the *what* pathway.

We also investigated the similarity between the emerging representations. We computed the cosine similarity for dual encoder representations on PGM-N and I-RAVE test sets (Figure 6). It can be seen that in both datasets, the learned dual encoder representations exhibit small cosine similarities.

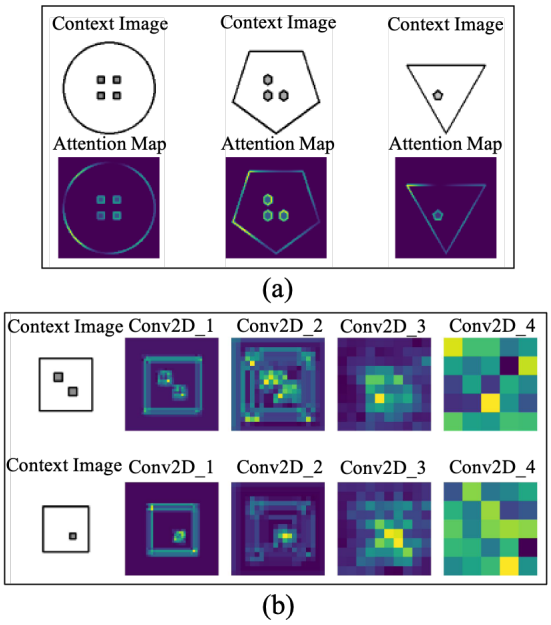


Figure 5: Illustrative examples from ViT and CNN streams. (a) Demonstrative instances of attention mapping from output tokens to the input space. (b) Visualization of the convolutional layer obtained through the forward process of DRNet.

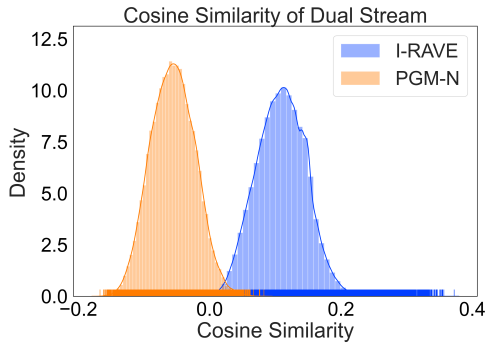


Figure 6: Cosine similarity for dual encoder representations. We calculated the cosine similarity between dual encoder representations for each test sample and displayed the statistical data in the form of a histogram and a rug plot.

Discussion

DRNet shows superior performance on multiple datasets, highlighting the potential of this dual-stream architecture. In our work, we use two backbone networks to mimic the two streams of visual processing in mammalian brains, and the high-level features obtained by these two streams form clearer discrete abstract rules through a rule extractor. With its remarkable generalization performance and accuracy across multiple benchmarks, DRNet provides an effective and powerful baseline in visual abstract reasoning.

To our knowledge, there is no existing research on dual-stream architectures within these benchmarks. To encourage further work, the technical shortcomings of the model are highlighted and explained in detail. The first question is whether a larger convolution kernel affects the model's ability to extract features, and the second is that the visual transformer module we used selected a large patch size to match the size of the CNN branch, which may have affected the visual transformer's ability to generalize global and local information. Currently, our model has poor recognition results in RAVEN-style 3×3 grid configuration. This may be due to the size of the convolution kernel and the patch size of ViT, which needs further investigation in the future.

The two-stream hypothesis involves many brain regions, and rigorous modeling is very challenging. At the framework level, we can extend our model to include more regions, in particular the hippocampal formation, which has long been considered as the basis for memory formation, flexible decision-making and reasoning (Behrens et al. 2018; Whittington et al. 2020, 2022). In (Bakhtiari et al. 2021), the functional specialization of the visual cortex emerges from training parallel pathways with self-supervised predictive learning; the potential to incorporate such functionality into DRNet for similar tasks is still under exploration.

Multimodal perception is necessary to achieve general artificial intelligence. Models combined with language models can exhibit zero-shot or few-shot learning capabilities in such non-verbal reasoning tasks (Huang et al. 2023), and our framework can flexibly integrate multimodal information in the future.

Conclusion

We applied DRNet to multiple datasets and observed that it achieves a high level of recognition accuracy and demonstrates an ability to generalize. Our experiments revealed that the dual-stream framework does not lead to the superposition of effects from the respective branches. In ablation experiments, we discovered that having more learnable parameters in a dual-stream network does not necessarily result in improved performance, and the rule extractor designed in DRNet can learn discrete abstract rule representations. We have illustrated the effectiveness of this work and its potential for abstract visual reasoning.

Acknowledgments

This work is partly supported by National Key R&D Program of China (2019YFA0709503). We thank Xiaohong Wan, Dahui Wang, Hongzhi You, Ruyuan Zhang and Zonglei Zhen for fruitful discussions.

References

Bakhtiari, S.; Mineault, P.; Lillicrap, T.; Pack, C.; and Richards, B. 2021. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, 34: 25164–25178.

Behrens, T. E.; Muller, T. H.; Whittington, J. C.; Mark, S.; Baram, A. B.; Stachenfeld, K. L.; and Kurth-Nelson, Z. 2018. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, 100(2): 490–509.

Benny, Y.; Pekar, N.; and Wolf, L. 2021. Scale-localized abstract reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12557–12565.

Brutzkus, A.; Globerson, A.; Malach, E.; Netser, A. R.; and Shalev-Schwartz, S. 2022. Efficient Learning of CNNs using Patch Based Features. In *International Conference on Machine Learning*, 2336–2356. PMLR.

Cadieu, C. F.; and Olshausen, B. A. 2012. Learning intermediate-level representations of form and motion from natural movies. *Neural computation*, 24(4): 827–866.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chen, Y.; Zuo, R.; Wei, F.; Wu, Y.; Liu, S.; and Mak, B. 2022. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, 35: 17043–17056.

Court, J. 1982. Manual for Raven's Progressive Matrices and Vocabulary Scales.

Cui, R.; Liu, H.; and Zhang, C. 2019. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 21(7): 1880–1891.

Da, K. 2014. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Feichtenhofer, C.; Pinz, A.; and Zisserman, A. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1933–1941.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2015. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1): 142–158.

Goodale, M. A.; and Milner, A. D. 1992. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1): 20–25.

Grezes, J.; and Decety, J. 2002. Does visual perception of object afford action? Evidence from a neuroimaging study. *Neuropsychologia*, 40(2): 212–222.

Hahne, L.; Lüddecke, T.; Wörgötter, F.; and Kappel, D. 2019. Attention on abstract visual reasoning. *arXiv preprint arXiv:1911.05990*.

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

Hersche, M.; Zeqiri, M.; Benini, L.; Sebastian, A.; and Rahimi, A. 2023. A neuro-vector-symbolic architecture for solving Raven's progressive matrices. *Nature Machine Intelligence*, 1–13.

Hoshen, D.; and Werman, M. 2017. Iq of neural networks. *arXiv preprint arXiv:1710.01692*.

Hu, S.; Ma, Y.; Liu, X.; Wei, Y.; and Bai, S. 2021. Stratified Rule-Aware Network for Abstract Visual Reasoning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 1567–1574. AAAI Press.

Huang, C.-C.; Rolls, E. T.; Hsu, C.-C. H.; Feng, J.; and Lin, C.-P. 2021. Extensive cortical connectivity of the human hippocampal memory system: beyond the "what" and "where" dual stream model. *Cerebral Cortex*, 31(10): 4652–4669.

Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O. K.; Patra, B.; Liu, Q.; Aggarwal, K.; Chi, Z.; Bjorck, J.; Chaudhary, V.; Som, S.; Song, X.; and Wei, F. 2023. Language Is Not All You Need: Aligning Perception with Language Models. *CoRR*, abs/2302.14045.

Jaeggi, S. M.; Buschkuehl, M.; Jonides, J.; and Perrig, W. J. 2008. Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, 105(19): 6829–6833.

Jahrens, M.; and Martinetz, T. 2020. Solving raven's progressive matrices with multi-layer relation networks. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–6. IEEE.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Ma, X.; Nie, W.; Yu, Z.; Jiang, H.; Xiao, C.; Zhu, Y.; Zhu, S.; and Anandkumar, A. 2022. RelViT: Concept-guided Vision Transformer for Visual Relational Reasoning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Maguire, E. A.; Burgess, N.; and O'Keefe, J. 1999. Human spatial navigation: cognitive maps, sexual dimorphism, and neural substrates. *Current opinion in neurobiology*, 9(2): 171–177.

Małkiński, M.; and Mańdziuk, J. 2022a. Deep Learning Methods for Abstract Visual Reasoning: A Survey on Raven's Progressive Matrices. *arXiv preprint arXiv:2201.12382*.

Małkiński, M.; and Mańdziuk, J. 2022b. Multi-label contrastive learning for abstract visual reasoning. *IEEE Transactions on Neural Networks and Learning Systems*.

Mao, M.; Zhang, R.; Zheng, H.; Ma, T.; Peng, Y.; Ding, E.; Zhang, B.; Han, S.; et al. 2021. Dual-stream network for visual recognition. *Advances in Neural Information Processing Systems*, 34: 25346–25358.

Mondal, S. S.; Webb, T. W.; and Cohen, J. 2023. Learning to reason over visual objects. In *International Conference on Learning Representations*.

Niebur, E.; and Koch, C. 1995. Control of selective visual attention: Modeling the" where" pathway. *Advances in neural information processing systems*, 8.

Perfetti, B.; Saggino, A.; Ferretti, A.; Caulo, M.; Romani, G. L.; and Onofrj, M. 2009. Differential patterns of cortical activation as a function of fluid reasoning complexity. *Human brain mapping*, 30(2): 497–510.

Prabhakaran, V.; Smith, J. A.; Desmond, J. E.; Glover, G. H.; and Gabrieli, J. D. 1997. Neural substrates of fluid reasoning: an fMRI study of neocortical activation during performance of the Raven's Progressive Matrices Test. *Cognitive psychology*, 33(1): 43–63.

Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; and Dosovitskiy, A. 2021. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34: 12116–12128.

Rahaman, N.; Gondal, M. W.; Joshi, S.; Gehler, P.; Bengio, Y.; Locatello, F.; and Schölkopf, B. 2021. Dynamic inference with neural interpreters. *Advances in Neural Information Processing Systems*, 34: 10985–10998.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Sahu, P.; Basioti, K.; and Pavlovic, V. 2022. SAViR-T: Spatially Attentive Visual Reasoning with Transformers. *CoRR*, abs/2206.09265.

Santoro, A.; Hill, F.; Barrett, D.; Morcos, A.; and Lillicrap, T. 2018. Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning*, 4477–4486.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30.

Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27.

Snow, R. E.; Kyllonen, P. C.; Marshalek, B.; et al. 1984. The topography of ability and learning correlations. *Advances in the psychology of human intelligence*, 2(S 47): 103.

Snow, R. E.; and Lohman, D. F. 1984. Toward a theory of cognitive aptitude for learning from instruction. *Journal of educational psychology*, 76(3): 347.

Spratley, S.; Ehinger, K.; and Miller, T. 2020. A closer look at generalisation in raven. In *European Conference on Computer Vision*, 601–616. Springer.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, D.; Jamnik, M.; and Liò, P. 2020. Abstract Diagrammatic Reasoning with Multiplex Graph Networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Whittington, J. C.; McCaffary, D.; Bakermans, J. J.; and Behrens, T. E. 2022. How to build a cognitive map. *Nature Neuroscience*, 25(10): 1257–1272.

Whittington, J. C.; Muller, T. H.; Mark, S.; Chen, G.; Barry, C.; Burgess, N.; and Behrens, T. E. 2020. The Tolman-Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5): 1249–1263.

Wu, Y.; Dong, H.; Grosse, R.; and Ba, J. 2020. The scattering compositional learner: Discovering objects, attributes, relationships in analogical reasoning. *arXiv preprint arXiv:2007.04212*.

Yang, L.; You, H.; Zhen, Z.; Wang, D.; Wan, X.; Xie, X.; and Zhang, R.-Y. 2023. Neural prediction errors enable analogical visual reasoning in human standard intelligence tests. In *International Conference on Machine Learning*, 39572–39583. PMLR.

Zhang, C.; Gao, F.; Jia, B.; Zhu, Y.; and Zhu, S.-C. 2019a. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5317–5327.

Zhang, C.; Jia, B.; Gao, F.; Zhu, Y.; Lu, H.; and Zhu, S.-C. 2019b. Learning perceptual inference by contrasting. *Advances in Neural Information Processing Systems*, 32.

Zhang, C.; Jia, B.; Zhu, S.-C.; and Zhu, Y. 2021. Abstract spatial-temporal reasoning via probabilistic abduction and execution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9736–9746.

Zhang, C.; Xie, S.; Jia, B.; Wu, Y. N.; Zhu, S.-C.; and Zhu, Y. 2022a. Learning algebraic representation for systematic generalization in abstract reasoning. In *European Conference on Computer Vision*, 692–709. Springer.

Zhang, W.; Tang, L.; Mo, S.; Liu, X.; and Song, S. 2022b. Learning Robust Rule Representations for Abstract Reasoning via Internal Inferences. In *Advances in Neural Information Processing Systems*.

Zheng, K.; Zha, Z.-J.; and Wei, W. 2019. Abstract reasoning with distracting features. *Advances in Neural Information Processing Systems*, 32.

Zhou, H.; Zhou, W.; Zhou, Y.; and Li, H. 2021. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*, 24: 768–779.

Zhuo, T.; and Kankanhalli, M. 2020. Effective abstract reasoning with dual-contrast network. In *International Conference on Learning Representations*.

Zolfaghari, M.; Oliveira, G. L.; Sedaghat, N.; and Brox, T. 2017. Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2904–2913.