Advancing Generalization Across a Variety of Abstract Visual Reasoning Tasks

Mikołaj Małkiński[1] and **Jacek Mańdziuk**[1,2]
[1]Warsaw University of Technology, Warsaw, Poland
[2]AGH University of Krakow, Krakow, Poland
mikolaj.malkinski.dokt@pw.edu.pl, jacek.mandziuk@pw.edu.pl

Abstract

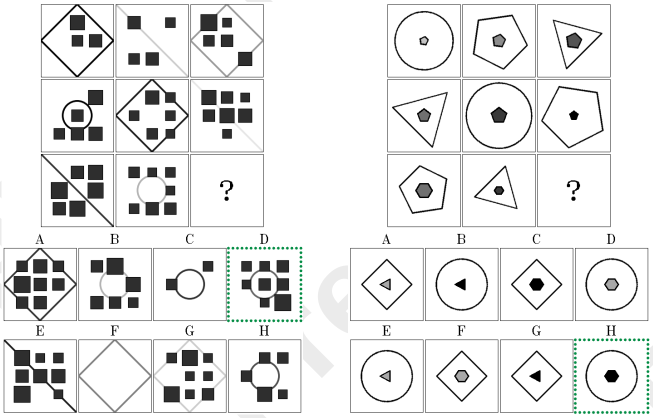
The abstract visual reasoning (AVR) domain presents a diverse suite of analogy-based tasks devoted to studying model generalization. Recent years have brought dynamic progress in the field, particularly in i.i.d. scenarios, in which models are trained and evaluated on the same data distributions. Nevertheless, o.o.d. setups that assess model generalization to new test distributions remain challenging even for the most recent models. To advance generalization in AVR tasks, we present the *Pathways of Normalized Group Convolution* model (PoNG), a novel neural architecture that features group convolution, normalization, and a parallel design. We consider a wide set of AVR benchmarks, including Raven's Progressive Matrices and visual analogy problems with both synthetic and real-world images. The experiments demonstrate strong generalization capabilities of the proposed model, which in several settings outperforms the existing literature methods.

1 Introduction

The abstract visual reasoning (AVR) domain encompasses visual tasks requiring reasoning about abstract patterns expressed through image-based analogies. A classical AVR task, Raven's Progressive Matrices (RPMs) [Raven, 1936; Raven and Court, 1998], illustrated in Fig. 1, consists of a 3×3 grid of panels with the bottom-right panel missing. Panels in the first two rows are designed according to some number of abstract rules that govern objects and attributes in the images. The task is to complete the grid by selecting the correct answer from the eight provided choices. Another AVR task, visual analogies, shown in Fig. 2, involves two rows of images. The top row presents an abstract relation that must be instantiated in the bottom row by selecting one of the four answer panels that correctly completes the analogy.

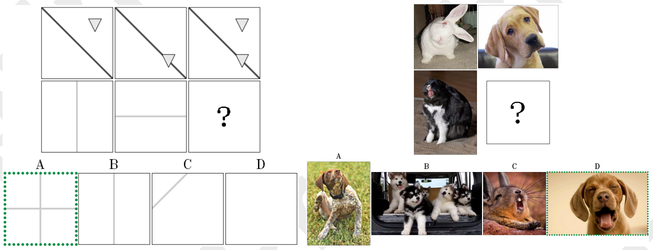
Solving AVR tasks involves detecting rule patterns across images, abstracting them into crisp concepts, and applying these concepts to novel scenarios. For example, matrices in the visual analogy problems (VAP) dataset [Hill *et al.*, 2019] present different domains in matrix rows (e.g., `shape type` (top) and `line type` (bottom) in Fig. 2a), emphasizing the importance of forming domain-independent concept representations. Such analogy-making abilities are closely tied to fluid intelligence [Snow *et al.*, 1984; Carpenter *et al.*, 1990; Lake *et al.*, 2017], a cornerstone of human cognition. Replicating these capabilities in learning systems has been a long-standing goal of research in the field [Gentner, 1980; Hofstadter, 1995; French, 2002; Lovett *et al.*, 2007; Gentner and Forbus, 2011].

A key aspect of AVR tasks, central to our work, are their systematic problem generation methods. Underneath each AVR task design lies a precise definition of its abstract structure, which defines the rule patterns expressed in the matrices.



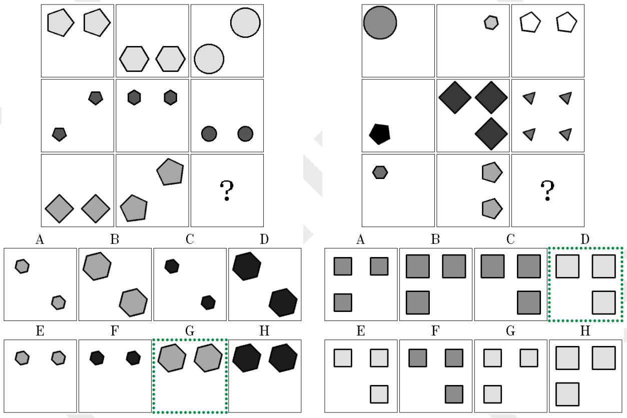
(a) PGM [Barrett *et al.*, 2018] (b) I-RAVEN [Hu *et al.*, 2021]

Figure 1: Raven's Progressive Matrices (RPMs).



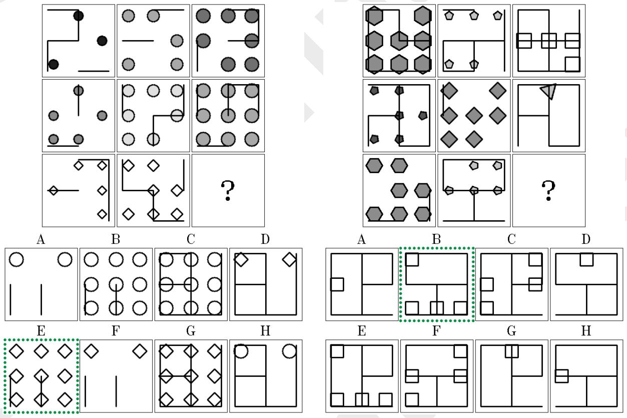
(a) VAP [Hill *et al.*, 2019] (b) VASR [Bitton *et al.*, 2023]

Figure 2: Visual analogies.



(a) A/ColorSize train (b) A/ColorSize test

Figure 3: A-I-RAVEN [Małkiński and Mańdziuk, 2025a].



(a) Number DistributeThree (b) Position Arithmetic

Figure 4: I-RAVEN-Mesh [Małkiński and Mańdziuk, 2025a].

For instance, each PGM matrix [Barrett *et al.*, 2018] has a corresponding abstract structure $\mathcal{S} = \{(r, o, a) \mid r \in \mathcal{R}, o \in \mathcal{O}, a \in \mathcal{A}\}$, where $\mathcal{R} = \{$progression, XOR, OR, AND, consistent union$\}$, $\mathcal{O} = \{$shape, line$\}$, and $\mathcal{A} = \{$size, type, color, position, number$\}$ are the sets of rules, objects, and attributes, resp. This formal specification facilitates defining dataset splits with varying feature distributions, enabling the evaluation of generalization by training models on matrices with specific abstract structures and testing them on matrices with different structures.

Motivation. Several AVR studies have addressed the i.i.d. problem formulation, where models are trained and tested on matrices sampled from a shared feature distribution. Continuous improvements have produced methods that surpass human performance on tasks like RPMs, given sufficient amount of training data [Hernández-Orallo *et al.*, 2016; Małkiński and Mańdziuk, 2025c]. Other research lines have demonstrated the benefits of knowledge transfer [Mańdziuk and Żychowski, 2019; Tomaszewska *et al.*, 2022] and multi-task learning [Małkiński and Mańdziuk, 2024b]. Despite these achievements, o.o.d. problem formulations, where models are evaluated on matrices sampled from a different feature distribution than the one used for training, remain a major challenge even for state-of-the-art (SOTA) deep learning (DL) models. Moreover, existing approaches in the AVR domain primarily target synthetic tasks with simple 2D geometric shapes, without considering their applicability to problems with real-world data. In this work, we strive to develop a model architecture that not only performs well in i.i.d. tasks but also excels in o.o.d. settings. Additionally, we consider both synthetic and real-world setups to broaden the applicability of the proposed approach.

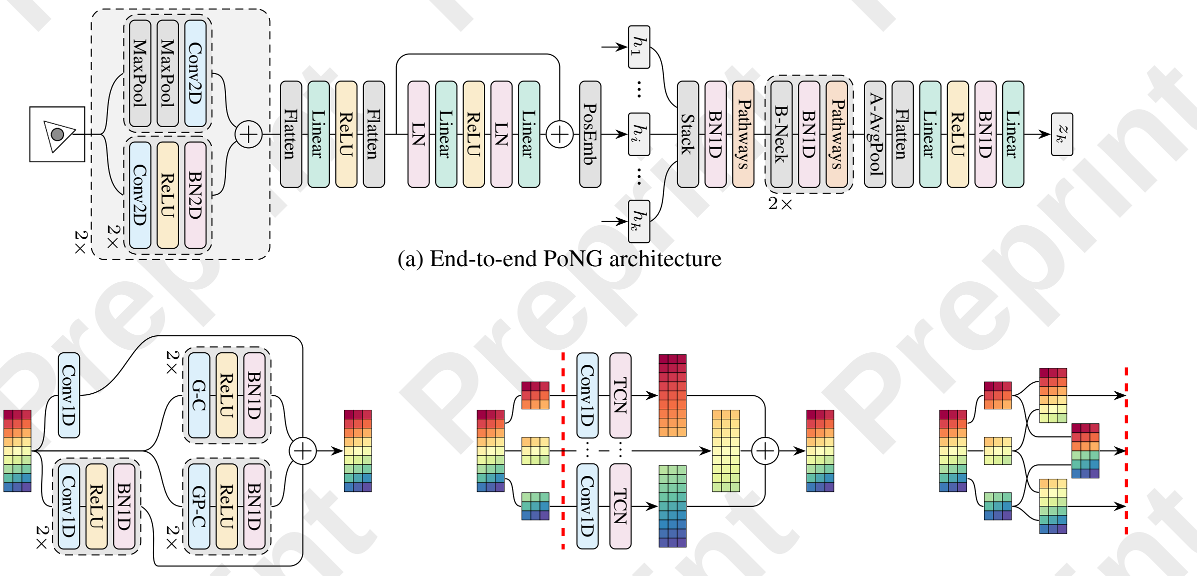
Contribution. To tackle these open challenges, we introduce the following contributions:

- We propose *Pathways of Normalized Group Convolution* (PoNG), a new neural model for AVR tasks that integrates group convolution, normalization, and a parallel design.
- We perform a comprehensive evaluation of PoNG against a wide range of SOTA models. The experiments show PoNG's versatility in both i.i.d. and o.o.d. problem setups, spanning RPMs and VAPs in synthetic and real-world scenarios.
- We conduct an ablation study to analyze the contributions of PoNG's modules, providing deeper insights into its design.

2 Related Work

AVR tasks. The AVR domain comprises a wide range of challenges [Mitchell, 2021; van der Maas *et al.*, 2021; Stabinger *et al.*, 2021; Małkiński and Mańdziuk, 2023]. Most relevant to our work are tasks involving RPMs and visual analogies. After the introduction of early RPM datasets [Matzen *et al.*, 2010; Wang and Su, 2015; Hoshen and Werman, 2017], two large-scale benchmarks were developed and broadly adopted in the DL literature. PGM [Barrett *et al.*, 2018] (Fig. 1a) introduced 8 regimes to measure generalization of DL models. RAVEN [Zhang *et al.*, 2019a] presented matrices with hierarchical structures across 7 figure configurations. Subsequent works further expanded the RAVEN dataset line: I-RAVEN [Hu *et al.*, 2021] (Fig. 1b) mitigated a bias in RAVEN's answer generation method, A-I-RAVEN [Małkiński and Mańdziuk, 2025a] (Fig. 3) defined 10 generalization regimes of varying complexity, and I-RAVEN-Mesh [Małkiński and Mańdziuk, 2025a] (Fig. 4) overlayed line-based patterns on top of the matrices. A parallel research stream introduced the VAP dataset [Hill *et al.*, 2019] (Fig. 2a), which similarly to PGM enables measuring generalization on matrices with a different structure. VASR [Bitton *et al.*, 2023] (Fig. 2b) introduced analogies with real-world images requiring understanding of rich visual scenes. A detailed description of the datasets used in this work is provided in Section 4.1.

AVR solvers. Early attempts to solve AVR tasks with DL models were prompted by the development of Relation Network (RN) [Santoro *et al.*, 2017]. WReN [Barrett *et al.*, 2018] applies RN to panel embeddings, CoPINet [Zhang *et al.*, 2019b] integrates RN with contrastive mechanisms, and



(a) End-to-end PoNG architecture

(b) Pathways block

(c) Group convolution (P3)

(d) Group-pair convolution (P4)

Figure 5: **PoNG.** (a) The panel encoder embeds each input image x_i independently, producing h_i. Context panel embeddings $\{h_i\}_{i=1}^{8}$ together with the embedding of k'th answer h_k are stacked and processed with the reasoner, leading to z_k. (b) The pathways block, a key component of PoNG, comprises four parallel pathways P1 – P4. (c) P3 and (d) P4 employ novel normalized group convolution operators. PosEmb denotes position embedding, G-C the group convolution module used in P3, and GP-C the group-pair convolution module used in P4. The red dashed line marks the point after which G-C and GP-C perform analogous computation.

MRNet [Benny *et al.*, 2021] embeds RN into a multi-scale architecture. Differently, SRAN [Hu *et al.*, 2021] processes distinct panel groups with dedicated ResNet encoders, SCL [Wu *et al.*, 2020] splits embeddings into groups processed by a shared neural layer, RelBase [Spratley *et al.*, 2020] primarily relies on convolutional layers, ARII [Zhang *et al.*, 2022b] learns robust rule representations via internal inferences, CPCNet [Yang *et al.*, 2023b] utilizes a self-contrasting learning process to align perceptual and conceptual input representations, PredRNet [Yang *et al.*, 2023a] mimics the prediction and matching process, and DRNet [Zhao *et al.*, 2024] merges panel representations of two independent encoders. Other works develop neuro-symbolic approaches [Zhang *et al.*, 2021; Zhang *et al.*, 2022a] or perform explicit object recognition prior to reasoning [Mondal *et al.*, 2023; Mondal *et al.*, 2024]. Neural Structure Mapping [Shekhar and Taylor, 2022] decouples perception from reasoning to solve visual analogies. Bitton *et al.* [2023] formulate several zero-shot and supervised methods to solve real-world analogies using a frozen pre-trained Vision Transformer [Dosovitskiy *et al.*, 2021] as the panel encoder. Although diverse approaches have been tried to tackle AVR benchmarks, contemporary models continue to exhibit limitations in generalization. In this context, we propose PoNG, a new versatile AVR model that performs well across diverse tasks.

3 Pathways of Normalized Group Convolution (PoNG) Model

We introduce PoNG (Fig. 5), a novel model that outcompetes baselines across a number of problem settings. The model follows a typical two-stage design. Firstly, it generates an embedding of each image panel. Then, it aggregates representations of matrix panels to predict the index of the correct answer. The details are described in [Małkiński and Mańdziuk, 2025b, Appendix A].

Let (X, y, r) denote an AVR matrix, where $X = \{x_i\}_{i=1}^{n}$ is the set of n image panels comprising n_c context panels $\{x_i\}_{i=1}^{n_c}$ and n_a answer panels $\{x_i\}_{i=n_c+1}^{n}$, $x_i \in [0,1]^{h \times w}, i = 1, \ldots, n$ is a grayscale image of height h and width w, $y \in \{0,1\}^{n_a}$ is the one-hot encoded index of the correct answer, $r \in \{0,1\}^{d_r}$ is the multi-hot encoded representation of matrix rules of dimensionality d_r using sparse encoding [Małkiński and Mańdziuk, 2024a]. For RPMs $n_c = n_a = 8$, for VAP $n_c = 5, n_a = 4$, and for VASR $n_c = 3, n_a = 4$. In each experiment $h = w = 80$, while d_r is determined by the number of different abstract structures in the corresponding dataset ($d_r = 40$ for I-RAVEN and A-I-RAVEN, $d_r = 48$ for I-RAVEN-Mesh, $d_r = 50$ for PGM, and $d_r = 28$ for VAP).

Panel encoder. The first component of the model has the form $\mathcal{E} : x \rightarrow h$, where $h \in \mathbb{R}^{d_h}$ is the input panel embedding of dimensionality d_h. Following RelBase [Spratley *et al.*, 2020], the module comprises 2 blocks of the same architecture. Each block includes 2 parallel pathways that build high-level and low-level features, resp. The first one contains 2 convolutional blocks, each with 2D convolution, ReLU, and Batch Normalization (BN) [Ioffe and Szegedy, 2015]. The second one contains 2D max pooling followed by 2D convolution. The sum of both pathway results forms the block output. Differently from RelBase, we flatten the height and width dimensions of the resultant embedding, pass it through a linear layer with ReLU, flatten the channel and spatial dimensions, and pass the tensor through a feed-forward residual block with Layer Normalization (LN) [Ba *et al.*, 2016]. Finally, we concatenate the tensor with a position embedding (a learned 25-dimensional vector for each panel in the context grid), leading to h.

Reasoner. The second component of the model has the form $\mathcal{R} : \{h_i\}_{i=1}^{8} \cup h_k \rightarrow z_k$, where h_k is the panel embedding of k'th answer. For each answer panel, the reasoner produces embedding z_k that describes how well the considered answer fits into the matrix context. Panel embeddings $\{h_i\}_{i=1}^{8} \cup h_k$ are stacked and processed by a sequence of 3 reasoning blocks interleaved with 2 bottleneck layers for dimensionality reduction. Each reasoning block comprises BN and 4 parallel pathways, outputs of which are added together to form the output of the block. Next, the latent representation is passed through adaptive average pooling, flattened, processed with a linear layer with ReLU, passed through BN and projected with a linear layer to $z_k \in \mathbb{R}^{128}$.

Pathways. The key aspect of the reasoner module are its pathways. Each takes an input tensor of shape (B, C, D), where B is the batch size, C is the number of channels, and D is the feature dimension. In the first reasoning block $D = d_h$ and C corresponds to the number of panel embeddings in the considered group ($C = 9$ for RPMs, $C = 6$ for VAP, and $C = 4$ for VASR). Pathways are described as follows: P1 – a pointwise 1D convolution layer that mixes panel features at each spatial location; P2 – a sequence of 2 blocks, each comprising 1D convolution, ReLU, and BN, that builds higher level features spanning neighbouring spatial locations; P3 – analogous to P2, but 1D convolution is replaced with a group 1D convolution that splits the tensor into several groups along the channel dimension, applies a 1D convolution with shared weights to each group, and adds together the representations of each group; P4 – analogous to P3, but groups are arranged into pairs concatenated along the channel dimension and processed with a 1D convolution with shared weights. In contrast to [Krizhevsky *et al.*, 2012], the proposed group convolution layers in both P3 and P4 apply TCN [Webb *et al.*, 2020] to the outputs in each group. In the first layer P3 and P4 split the input tensor into 3 groups for RPMs and visual analogies, and into 2 groups for VASR, which allows for producing embeddings of each matrix row and each pair of rows, resp. Though we apply the pathways block in the AVR context, we envisage it as a generic module, also applicable to other settings involving a set of vector representations of shape (B, C, D).

Answer prediction. Representations of the context matrix filled-in with the respective answer, $\{z_k\}_{k=1}^{n_a}$, are processed with three prediction heads. The target head $\mathcal{P}^y : z_k \rightarrow \widehat{y_k}$ employs two linear layers interleaved with ReLU to produce score $\widehat{y_k} \in \mathbb{R}$ describing how well the answer k aligns with the matrix context. The aggregate rule head $\mathcal{P}_1^r : \{z_k\}_{k=1}^{n_a} \rightarrow \widehat{r_1}$ computes the sum of inputs and processes it with two linear layers interleaved with ReLU, producing a latent prediction of matrix rules $\widehat{r_1} \in \mathbb{R}^{d_r}$. We also introduce a novel target-conditioned rule head $\mathcal{P}_2^r : \{z_k\}_{k=1}^{n_a} \rightarrow \widehat{r_2}$, which processes its input through a linear layer and computes a weighted sum of the resultant embeddings with weights given by the predicted probability distribution over the set of possible answers $\sigma(\{\widehat{y_k}\}_{k=1}^{n_a})$, where σ denotes softmax. The model is trained with a joint loss function $\mathcal{L} = \text{CE}(\sigma(\{\widehat{y_k}\}_{k=1}^{n_a}), y) + \beta\text{BCE}(\zeta(\widehat{r_1}, r)) + \gamma\text{BCE}(\zeta(\widehat{r_2}, r))$, where ζ denotes sigmoid, CE cross-entropy, BCE binary cross-entropy, $\beta = 25$ and $\gamma = 5$ are balancing coefficients.

4 Experiments

We employ a set of diverse AVR tasks to evaluate PoNG's generalization capabilities. Section 4.1 introduces the selected datasets, Section 4.2 details the experimental setup, and Section 4.3 presents the results.

4.1 AVR Datasets

AVR models are typically evaluated on RPM benchmarks, a problem set well-established in the literature. We utilize four RPM datasets: PGM [Barrett *et al.*, 2018], I-RAVEN [Hu *et al.*, 2021], I-RAVEN-Mesh and A-I-RAVEN [Małkiński and Mańdziuk, 2025a]. We extend the evaluation of PoNG beyond RPMs, to two benchmarks comprising visual analogies with both synthetic [Hill *et al.*, 2019] and real-world [Bitton *et al.*, 2023] images.

PGM. The PGM dataset was the first large-scale RPM benchmark designed to evaluate the AVR capabilities of deep learning models. In PGM each matrix is defined by an abstract structure encompassing its rules, objects, and attributes. To assess generalization, the dataset is divided into 8 generalization regimes. In the Neutral regime, the train, validation, and test splits share the same feature distribution, constituting an i.i.d. learning challenge. In the remaining regimes, the train and validation splits share a common distribution, while the test split relies on a different distribution, enabling the evaluation of generalization to unseen feature combinations. Each regime contains 1.42M RPMs, where 1.2M, 20K, and 200K belong to the train, validation, and test splits, resp.

I-RAVEN. The RAVEN dataset [Zhang *et al.*, 2019a] was constructed to expand the range of visual configurations in RPMs. It incorporates 7 configurations that define object locations within the matrices. For instance, in the `Left-Right` configuration, each panel is divided into left and right parts that can be governed by distinct rules. A subsequent study identified a bias in the RAVEN's answer generation method, enabling models to learn shortcut solutions [Hu *et al.*, 2021]. To alleviate this issue, the I-RAVEN dataset was proposed, which employs an impartial answer generation method. We utilize I-RAVEN in the experiments to avoid learning shortcut solutions. The benchmark consists of 10K matrices per configuration, totaling 70K matrices, split into the train, validation, and test with a 60/20/20 ratio.

I-RAVEN-Mesh. The I-RAVEN-Mesh dataset builds upon I-RAVEN by rendering a grid of 1 to 12 lines on the underlying matrices. The grid is defined by two attributes, the number of lines and their position. While the dataset was originally introduced to assess knowledge acquisition in transfer learning settings, we use it for standard supervised learning. In this setup, the model is trained directly on the dataset, analogously to I-RAVEN, to expand the scope of the considered i.i.d. tasks.

	i.i.d. tasks: I-RAVEN and I-RAVEN-Mesh			o.o.d. tasks: A-I-RAVEN			
	I-RAVEN†	I-RAVEN	I-RAVEN-Mesh	`A/Color`	`A/Position`	`A/Size`	`A/Type`
ALANS	–	27.0 (±8.4)	15.9 (±2.6)	15.2 (±1.4)	16.0 (±1.0)	23.3 (±6.5)	19.0 (±3.4)
CPCNet	**98.5**	70.4 (±6.4)	66.6 (±5.1)	51.2 (±3.8)	68.3 (±4.0)	43.5 (±3.5)	38.6 (±4.3)
CNN-LSTM	18.9	27.5 (±1.5)	28.9 (±0.4)	17.0 (±3.1)	24.0 (±2.9)	13.6 (±1.4)	14.5 (±0.8)
CoPINet	46.1	43.2 (±0.1)	41.1 (±0.3)	32.5 (±0.2)	41.3 (±1.6)	21.8 (±0.2)	19.8 (±0.9)
DRNet	97.6	90.9 (±1.1)	83.9 (±2.7)	70.0 (±1.6)	77.5 (±0.9)	54.3 (±3.0)	44.3 (±0.8)
MRNet	83.5	86.7 (±2.3)	79.5 (±2.0)	33.6 (±8.2)	62.6 (±2.6)	20.6 (±5.0)	19.4 (±0.3)
PrAE	77.0	19.5 (±0.4)	33.2 (±0.4)	47.9 (±0.9)	68.2 (±3.3)	41.3 (±1.8)	37.0 (±1.7)
PredRNet	96.5	88.8 (±1.8)	59.2 (±6.4)	59.4 (±1.0)	73.7 (±0.7)	47.5 (±1.3)	40.2 (±1.3)
RelBase	91.1	89.6 (±0.6)	84.9 (±4.4)	67.4 (±2.7)	76.6 (±0.3)	51.1 (±2.4)	44.1 (±1.0)
SCL	95.0	83.4 (±2.5)	80.9 (±1.5)	65.1 (±2.0)	76.7 (±7.1)	65.6 (±2.4)	49.5 (±1.8)
SRAN	60.8	58.2 (±1.6)	57.8 (±0.2)	38.3 (±1.0)	56.9 (±0.7)	34.4 (±3.0)	30.7 (±2.2)
STSN	95.7	51.0 (±24.8)	48.7 (±11.5)	39.3 (±6.9)	36.1 (±19.9)	38.4 (±16.6)	39.1 (±5.0)
WReN	23.8	18.4 (±0.0)	25.7 (±0.2)	16.9 (±0.5)	17.3 (±0.4)	12.4 (±0.5)	15.1 (±0.7)
PoNG (ours)	95.9	**95.9** (±0.7)	**89.3** (±2.4)	**80.3** (±4.3)	**79.3** (±0.7)	**73.5** (±3.1)	**59.4** (±6.9)

Table 1: **RAVEN-related datasets.** Mean and standard deviation of test accuracy for three random seeds. Best dataset results are marked in bold and the second best are underlined. I-RAVEN† denotes results on I-RAVEN reported by model authors in the corresponding papers.

	`A/ColorSize`	`A/ColorType`	`A/SizeType`	`A/Color-P`	`A/Color-A`	`A/Color-D3`
ALANS	15.1 (±3.3)	17.7 (±3.2)	15.7 (±3.2)	24.8 (±18.8)	18.3 (±6.6)	22.4 (±7.7)
CPCNet	33.0 (±5.3)	25.0 (±0.9)	24.1 (±1.2)	50.5 (±0.6)	45.9 (±2.7)	37.8 (±0.9)
CNN-LSTM	13.4 (±0.9)	14.7 (±1.7)	13.0 (±0.1)	17.2 (±1.5)	17.1 (±3.7)	20.6 (±6.7)
CoPINet	18.3 (±0.3)	17.2 (±0.1)	19.7 (±0.7)	35.8 (±0.6)	35.2 (±0.5)	26.9 (±0.5)
DRNet	38.3 (±0.5)	29.5 (±0.5)	31.6 (±1.2)	72.8 (±1.3)	66.7 (±1.2)	63.2 (±0.3)
MRNet	18.7 (±1.1)	20.0 (±2.6)	28.2 (±0.9)	34.4 (±3.4)	35.7 (±5.9)	18.6 (±0.1)
PrAE	30.0 (±1.1)	26.7 (±0.7)	25.6 (±0.8)	62.3 (±0.9)	43.0 (±26.5)	55.1 (±0.8)
PredRNet	31.0 (±1.6)	28.0 (±0.7)	27.9 (±0.5)	62.3 (±2.2)	56.9 (±1.4)	48.5 (±0.9)
RelBase	36.6 (±0.8)	29.7 (±0.6)	31.1 (±1.0)	73.0 (±1.8)	66.2 (±1.0)	65.7 (±4.6)
SCL	40.8 (±3.2)	32.0 (±2.3)	**33.5** (±0.7)	75.6 (±10.1)	60.0 (±4.1)	63.9 (±4.3)
SRAN	22.7 (±1.1)	20.9 (±0.9)	23.3 (±0.3)	42.1 (±2.3)	39.9 (±2.7)	34.6 (±3.6)
STSN	27.3 (±4.6)	21.9 (±4.6)	12.3 (±0.1)	39.9 (±14.7)	25.7 (±10.6)	20.7 (±7.7)
WReN	13.5 (±0.1)	13.8 (±0.7)	14.1 (±0.2)	18.0 (±0.4)	17.1 (±0.2)	17.7 (±0.6)
PoNG (ours)	**44.7** (±2.1)	**34.3** (±0.8)	32.1 (±2.1)	**81.4** (±3.1)	**70.0** (±4.1)	**81.3** (±1.6)

Table 2: **A-I-RAVEN extended regimes.** P, A, and D3 denote Progression, Arithmetic, and Distribute Three, resp.

A-I-RAVEN. The A-I-RAVEN dataset was introduced to combine the generalization assessment capabilities of PGM with the broad adoption of RAVEN-like benchmarks. Drawing from PGM, A-I-RAVEN defines 10 generalization regimes. In each regime, a subset of attributes follows specific rules in the train and validation splits, while being governed by different rules in the test split. For example, in the `A/ColorSize` regime, the Color and Size attributes adhere to the Constant rule in the train and validation splits and are governed by a rule other than Constant in the test split. This approach enables the evaluation of models on RPMs with novel rule–attribute combinations that were not seen during training. Each regime contains 70K matrices, analogously to I-RAVEN.

VAP. The VAP benchmark was introduced to assess the analogy-making capabilities of learning systems. Each VAP matrix consists of a 2×3 grid of panels. The task is to identify a concept in the source domain (top row) and instantiate it in the target domain (bottom row) by selecting the correct answer panel to complete the matrix. The dataset defines five generalization regimes: Novel Domain Transfer, Novel Target Domain: Colour of Shapes, Novel Target Domain: Type of Lines, Novel Attribute Values: Interpolation, and Novel Attribute Values: Extrapolation, which test the model's generalization to novel domains or attribute values. Each regime contains 710K matrices, with 600K, 10K, and 100K devoted to the train, validation, and test splits, resp. In all experiments we use the learning analogies by contrasting (LABC) dataset variant, which constructs the answer set using semantically plausible images that consistently complete the target domain with some relation.

VASR. The VASR dataset features visual analogies involving real-world images, requiring the learner to understand complex real-world scenes before solving the analogy problem. Each matrix consists of a 2×2 panel grid, with the bottom-right image missing. The task is to complete the matrix by selecting the correct image from the 4 provided choices. VASR follows the classical analogy problem formulation, which aims to complete the following relation: A is to B, as C is to D. We use Silver data for training, which includes

Model	Neutral	Interpolation	HO-AP	HO-TP	HO-Triples	HO-LT	HO-SC	Extrapolation	Average
SCL	87.1	56.0	79.6	76.6	23.0	14.1	12.6	19.8	46.1
MRNet	93.4	68.1	38.4	55.3	25.9	**30.1**	**16.9**	19.2	43.4
ARII	88.0	57.8	50.0	64.1	32.1	16.0	12.7	29.0	43.7
PredRNet	97.4	70.5	63.4	67.8	23.4	27.3	13.1	19.7	47.8
DRNet	**99.1**	83.8	**93.7**	78.1	**48.8**	27.9	13.1	22.2	**58.3**
Slot-Abstractor	91.5	**91.6**	63.3	78.3	20.4	16.7	14.3	**39.3**	51.9
PoNG (ours)	98.1	75.2	92.1	**97.7**	46.1	16.9	12.6	19.9	57.3

Table 3: **PGM**. Test accuracy of PoNG in all regimes of the PGM dataset. The Held-out Attribute Pairs regime is denoted as HO-AP, Held-out Triple Pairs as HO-TP, Held-out Triples as HO-Triples, Held-out Attribute line-type as HO-LT, and Held-out Attribute shape-colour as HO-SC. For reference, we provide results of SCL [Wu *et al.*, 2020; Małkiński and Mańdziuk, 2024a], MRNet [Benny *et al.*, 2021], ARII [Zhang *et al.*, 2022b], PredRNet [Yang *et al.*, 2023a], DRNet [Zhao *et al.*, 2024], and Slot-Abstractor [Mondal *et al.*, 2024].

	ND Transfer	NTD LineType	NTD ShapeColor	NAV Interpolation	NAV Extrapolation	Average
LBC	0.87 ± 0.005	0.76 ± 0.020	0.78 ± 0.004	0.93 ± 0.004	0.62 ± 0.020	0.79
NSM	0.88	0.79	0.78	0.93	**0.74**	0.82
PredRNet	0.96 ± 0.003	$\mathbf{0.82} \pm 0.010$	0.80 ± 0.010	0.97 ± 0.002	0.72 ± 0.060	**0.85**
PoNG (ours)	$\mathbf{0.98} \pm 0.001$	0.78 ± 0.006	$\mathbf{0.81} \pm 0.006$	$\mathbf{0.98} \pm 0.000$	0.68 ± 0.007	0.84

Table 4: **Visual Analogy Problems [Hill *et al.*, 2019].** Results of LBC, NSM, and PredRNet come from [Yang *et al.*, 2023a, Table 2d]. For PoNG, we present mean and std of test accuracy for three random seeds. ND denotes Novel Domain, NTD — Novel Target Domain, NAV — Novel Attribute Values.

150K, 2.25K, and 2.55K matrices in the train, validation, and test splits, resp. Experiments are conducted on both dataset variants, featuring random and difficult distractors, resp.

4.2 Experimental Setting

We assess PoNG's generalization by comparing its performance to SOTA models on the respective datasets. PoNG is trained using a standard training strategy involving the Adam optimizer [Kingma and Ba, 2014] with default hyperparameters ($\lambda = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$). Learning rate λ is reduced by a factor of 10 after 5 epochs without improvement in validation loss. Early stopping is applied after 10 epochs without validation loss reduction. We use batch size $B = 128$ for experiments on RAVEN-like datasets and $B = 256$ in the remaining cases to reduce training time on large datasets. All experiments are performed on a single GPU (NVIDIA DGX A100).

To ensure reproducibility, we use a set of fixed random seeds, provide a list of commands for running training jobs, and explicitly list static dependencies in configuration files. The code for reproducing all experiments is publicly accessible at: https://github.com/mikomel/raven

4.3 Results

Results on I-RAVEN and I-RAVEN-Mesh. We begin with evaluating PoNG in the i.i.d. setting on the I-RAVEN and I-RAVEN-Mesh datasets, comparing it to 13 SOTA baselines. As presented in Table 1, on I-RAVEN, using our experimental setup, PoNG achieves a test accuracy of 95.9%, outperforming all other models. When compared to results obtained with model-specific experimental setups (I-RAVEN†), PoNG is placed just behind CPCNet, DRNet, and PredRNet, which achieve slightly better scores. PoNG also secures the 1st place on I-RAVEN-Mesh, demonstrating high capacity to handle matrices with rules that span a large number of objects. Unlike many baseline models that rely on deeper architectures such as DRNet, SRAN or STSN, PoNG presents competitive performance despite its parameter-efficient design. These results demonstrate PoNG's strong ability to solve i.i.d. RPM tasks.

Results on A-I-RAVEN. To assess generalization, we evaluate PoNG on the 4 primary regimes of the A-I-RAVEN dataset, where the training and test distributions differ significantly. As shown in Table 1, PoNG outperforms all baselines across all settings, achieving test accuracies ranging from 59.4% on `A/Type` to 80.3% on `A/Color`, surpassing the best reference models by 10.3 and 9.9 p.p., resp. Additionally, Table 2 shows PoNG's performance across 6 extended regimes, which cover more challenging generalization tasks. Similarly, PoNG achieves superior performance in all but one regimes. Notably, PoNG outperforms the 2nd best model in the `A/Color-D3` regime by 15.6 p.p. Overall, the results on A-I-RAVEN highlight PoNG's ability to perform well across a wide range of generalization tasks with varying levels of complexity. However, certain regimes such as the 3 extended regimes with held-out attribute pairs (`A/ColorSize`, `A/ColorType`, `A/SizeType`) continue to pose a significant challenge for all models (including PoNG), raising the need for further advances in generalization.

Results on PGM. Table 3 presents PoNG's results across PGM regimes. The model achieves strong results in several settings, particularly excelling in the Held-out Triple Pairs regime, where it surpasses the best reference model by 19.4 p.p. On average, PoNG scored 57.3% accuracy securing the 2nd place, just behind DRNet with 58.3%. These results confirm PoNG's ability to perform well on RPM-based general-

Distractors	Zero-Shot ViT	Zero-Shot Swin	Supervised Concat	PoNG (best-of-3)	PoNG (mean ± std)
Random	86.0	86.0	84.1	**92.0**	91.8 ± 0.3
Difficult	50.3	52.9	54.9	**70.5**	69.5 ± 1.1

Table 5: **Visual Analogies of Situation Recognition (VASR)** [Bitton *et al.*, 2023]. Results of selected baselines come from [Bitton *et al.*, 2023, Table 3]. For PoNG, we present mean with std and best-of-3 test accuracy for three random seeds.

	I-RAVEN	I-RAVEN-Mesh	`A/Color`	`A/Position`	`A/Size`	`A/Type`
w/o P1 and P2	92.8 (− 3.1)	74.4 (−14.9)	73.3 (− 7.0)	76.4 (− 2.9)	58.4 (−15.2)	49.5 (− 9.8)
w/o P3 and P4	95.6 (− 0.3)	88.0 (− 1.3)	78.9 (− 1.4)	78.6 (− 0.7)	73.9 (+ 0.4)	53.9 (− 5.5)
w/o TCN	96.0 (+ 0.1)	90.8 (+ 1.4)	75.4 (− 4.9)	80.3 (+ 1.0)	66.6 (− 6.9)	57.5 (− 1.9)
$\beta = 0$	94.2 (− 1.7)	91.4 (+ 2.1)	79.0 (− 1.3)	77.5 (− 1.8)	70.3 (− 3.2)	53.3 (− 6.1)
$\gamma = 0$	95.7 (− 0.1)	88.8 (− 0.5)	74.2 (− 6.1)	79.6 (+ 0.3)	73.0 (− 0.5)	56.9 (− 2.5)
$\beta = 0 \wedge \gamma = 0$	79.7 (−16.2)	32.7 (−56.7)	72.1 (− 8.2)	75.1 (− 4.2)	64.9 (− 8.6)	49.0 (−10.3)
union	81.4 (−14.5)	32.5 (−56.8)	76.2 (− 4.1)	74.1 (− 5.2)	66.9 (− 6.6)	46.0 (−13.4)

Table 6: **PoNG ablations**. Test accuracy averaged across 3 random seeds and a difference to the default model setup (cf. Table 1). Union denotes application of all ablations except for the first one.

ization challenges extending beyond the RAVEN dataset line.

Synthetic visual analogies. Table 4 presents PoNG's results across 5 regimes from the VAP benchmark. PoNG achieves SOTA results in 3 out of 5 settings when compared to PredRNet, the currently leading VAP model. The Novel Attribute Values: Extrapolation regime poses the greatest challenge among VAP regimes, aligning with findings from PGM, where Extrapolation is also one of the most demanding regimes. Overall, PoNG and PredRNet perform competitively, with PredRNet achieving a better average score by 1 p.p. PoNG's strong results on VAP highlight its versatility in generalization tasks that extend beyond RPMs.

Real-World visual analogies. To evaluate PoNG on the VASR dataset we followed the approach proposed by the VASR authors and employed the Vision Transformer (ViT) [Dosovitskiy *et al.*, 2021] as a perception backbone that produces image embeddings. Specifically, we used the same model variant as [Bitton *et al.*, 2023], which is ViT-L/32 pre-trained on ImageNet-21k at resolution 224x224 and fine-tuned on ImageNet-1k at resolution 384x384. We replaced the panel encoder of PoNG with this frozen pre-trained backbone and trained the rest of the model from scratch. The results are presented in Table 5. The three reference methods perform comparably to each other, with Supervised Concat being slightly inferior to Zero-Shot methods on the random distractor split and slightly superior on the difficult split. However, in both dataset variants PoNG significantly outcompetes the strongest reference result with 92.0% vs. 86.0% and 70.5% vs. 54.9%, resp. This suggests that the proposed reasoner block is much more effective in reasoning over pre-trained embeddings than baseline methods. The results support the claim that PoNG is a versatile model with strong analogical reasoning capabilities, applicable to both synthetic and real-world domains.

Ablation study. We performed an ablation study on the RAVEN dataset line to evaluate the contributions of different PoNG components. Table 6 summarizes the results. The removal of P1 and P2 (cf. Fig. 5) leads to performance drop, in particular on I-RAVEN-Mesh (-14.9 p.p.) and `A/Size` (-15.2 p.p.). Similarly, removing P3 and P4 reduces model performance, especially on `A/Type` (-5.5 p.p.). Disabling TCN leads to generally worse results, primarily on `A/Color` (-4.9 p.p.) and `A/Size` (-6.9 p.p.). As shown in [Małkiński and Mańdziuk, 2025b, Appendix B], PoNG w/o TCN may fail to generalize rules to held-out attributes. Training without $\mathcal{P}_1^r$ ($\beta = 0$) or $\mathcal{P}_2^r$ ($\gamma = 0$) typically reduces model performance, but training with one of these rule-based prediction heads compensates to some degree the lack of the other. However, the removal of both ($\gamma = 0 \wedge \beta = 0$) deteriorates results across all datasets, signifying high relevance of the auxiliary training signal in PoNG's training. Overall, the ablation study demonstrates that all employed design choices contribute to the model performance.

5 Conclusion

Generalization to novel problem types is an active and open area of DL research. In this work, we introduced PoNG, a novel AVR model that leverages group convolution, parallel design, weight sharing, and normalization. To evaluate its effectiveness and versatility, we conducted experiments on four RPM benchmarks and two visual analogy datasets comprising both synthetic and real-world images. PoNG demonstrates strong performance across all considered problems, often surpassing the state-of-the-art reference methods.

Future directions. We believe that the proposed pathways block, a key component of PoNG, is a generic module also applicable to other tasks that require reasoning over a set of objects (vector embeddings). Nevertheless, the presented experimental evaluation of PoNG is focused on variable RPM benchmarks, including I-RAVEN, I-RAVEN-Mesh, A-I-RAVEN, and PGM, and two visual analogy datasets, i.e. VAP and VASR. Assessing the model's performance on problems outside the AVR domain constitutes an interesting continuation of this work.

Acknowledgments

This research was carried out with the support of the Laboratory of Bioinformatics and Computational Genomics and the High Performance Computing Center of the Faculty of Mathematics and Information Science Warsaw University of Technology. Mikołaj Małkiński was funded by the Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) programme.

References

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv:1607.06450*, 2016.

[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.

[Bitton *et al.*, 2023] Yonatan Bitton, Ron Yosef, Eliyahu Strugo, Dafna Shahaf, Roy Schwartz, and Gabriel Stanovsky. VASR: Visual analogies of situation recognition. In *AAAI*, volume 37, pages 241–249, 2023.

[Carpenter *et al.*, 1990] Patricia A Carpenter, Marcel A Just, and Peter Shell. What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychological review*, 97(3):404, 1990.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[French, 2002] Robert M French. The computational modeling of analogy-making. *Trends in cognitive Sciences*, 6(5):200–205, 2002.

[Gentner and Forbus, 2011] Dedre Gentner and Kenneth D Forbus. Computational models of analogy. *Wiley interdisciplinary reviews: cognitive science*, 2(3):266–276, 2011.

[Gentner, 1980] Dedre Gentner. *The structure of analogical models in science*. Bolt Beranek and Newman Cambridge, 1980.

[Hernández-Orallo *et al.*, 2016] José Hernández-Orallo, Fernando Martínez-Plumed, Ute Schmid, Michael Siebers, and David L Dowe. Computer models solving intelligence test problems: Progress and implications. *Artificial Intelligence*, 230:74–107, 2016.

[Hill *et al.*, 2019] Felix Hill, Adam Santoro, David Barrett, Ari Morcos, and Timothy Lillicrap. Learning to make analogies by contrasting abstract relational structure. In *ICLR*, 2019.

[Hofstadter, 1995] Douglas R Hofstadter. *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought*. Basic books, 1995.

[Hoshen and Werman, 2017] Dokhyam Hoshen and Michael Werman. IQ of neural networks. *arXiv:1710.01692*, 2017.

[Hu *et al.*, 2021] 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.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456. PMLR, 2015.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2014.

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

[Lake *et al.*, 2017] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[Lovett *et al.*, 2007] Andrew Lovett, Kenneth Forbus, and Jeffrey Usher. Analogy with qualitative spatial representations can simulate solving raven's progressive matrices. In *Proc. of the Annual Meeting of the Cognitive Science Society*, volume 29, 2007.

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

[Małkiński and Mańdziuk, 2024a] Mikołaj Małkiński and Jacek Mańdziuk. Multi-label contrastive learning for abstract visual reasoning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):1941–1953, 2024.

[Małkiński and Mańdziuk, 2024b] Mikołaj Małkiński and Jacek Mańdziuk. One self-configurable model to solve many abstract visual reasoning problems. In *AAAI*, volume 38, pages 14297–14305, 2024.

[Małkiński and Mańdziuk, 2025a] Mikołaj Małkiński and Jacek Mańdziuk. A-I-RAVEN and I-RAVEN-Mesh: Two new benchmarks for abstract visual reasoning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, 2025. (Accepted).

[Małkiński and Mańdziuk, 2025b] Mikołaj Małkiński and Jacek Mańdziuk. Advancing generalization across a variety of abstract visual reasoning tasks. *arXiv:2505.13391*, 2025.

[Małkiński and Mańdziuk, 2025c] Mikołaj Małkiński and Jacek Mańdziuk. Deep learning methods for abstract visual reasoning: A survey on raven's progressive matrices. *ACM Computing Surveys*, 57(7):1–36, 2025.

[Mańdziuk and Żychowski, 2019] Jacek Mańdziuk and Adam Żychowski. DeepIQ: A human-inspired AI system for solving IQ test problems. In *2019 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2019.

[Matzen *et al.*, 2010] Laura E Matzen, Zachary O Benz, Kevin R Dixon, Jamie Posey, James K Kroger, and Ann E Speed. Recreating raven's: Software for systematically generating large numbers of raven-like matrix problems with normed properties. *Behavior research methods*, 42(2):525–541, 2010.

[Mitchell, 2021] Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101, 2021.

[Mondal *et al.*, 2023] Shanka Subhra Mondal, Taylor Whittington Webb, and Jonathan Cohen. Learning to reason over visual objects. In *ICLR*, 2023.

[Mondal *et al.*, 2024] Shanka Subhra Mondal, Jonathan D. Cohen, and Taylor Whittington Webb. Slot abstractors: Toward scalable abstract visual reasoning. In *ICML*, volume 235, pages 36088–36105. PMLR, 2024.

[Raven and Court, 1998] John C Raven and John Hugh Court. *Raven's progressive matrices and vocabulary scales*. Oxford pyschologists Press Oxford, England, 1998.

[Raven, 1936] James C Raven. Mental tests used in genetic studies: The performance of related individuals on tests mainly educative and mainly reproductive. *Master's thesis, University of London*, 1936.

[Santoro *et al.*, 2017] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *NeurIPS*, 30:4967–4976, 2017.

[Shekhar and Taylor, 2022] Shashank Shekhar and Graham W. Taylor. Neural structure mapping for learning abstract visual analogies, 2022.

[Snow *et al.*, 1984] Richard E Snow, Patrick C Kyllonen, and Brachia Marshalek. The topography of ability and learning correlations. *Advances in the psychology of human intelligence*, 2(S 47):103, 1984.

[Spratley *et al.*, 2020] Steven Spratley, Krista Ehinger, and Tim Miller. A closer look at generalisation in RAVEN. In *European Conference on Computer Vision*, pages 601–616. Springer, 2020.

[Stabinger *et al.*, 2021] Sebastian Stabinger, David Peer, Justus Piater, and Antonio Rodríguez-Sánchez. Evaluating the progress of deep learning for visual relational concepts. *Journal of Vision*, 21(11):8–8, 2021.

[Tomaszewska *et al.*, 2022] Paulina Tomaszewska, Adam Żychowski, and Jacek Mańdziuk. Duel-based deep learning system for solving IQ tests. In *International Conference on Artificial Intelligence and Statistics*, pages 10483–10492. PMLR, 2022.

[van der Maas *et al.*, 2021] Han LJ van der Maas, Lukas Snoek, and Claire E Stevenson. How much intelligence is there in artificial intelligence? a 2020 update. *Intelligence*, 87:101548, 2021.

[Wang and Su, 2015] Ke Wang and Zhendong Su. Automatic generation of raven's progressive matrices. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[Webb *et al.*, 2020] Taylor Webb, Zachary Dulberg, Steven Frankland, Alexander Petrov, Randall O'Reilly, and Jonathan Cohen. Learning representations that support extrapolation. In *ICML*, pages 10136–10146. PMLR, 2020.

[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:2007.04212*, 2020.

[Yang *et al.*, 2023a] Lingxiao Yang, Hongzhi You, Zonglei Zhen, Dahui Wang, Xiaohong Wan, Xiaohua Xie, and Ru-Yuan Zhang. Neural prediction errors enable analogical visual reasoning in human standard intelligence tests. In *ICML*, volume 202, pages 39572–39583. PMLR, 2023.

[Yang *et al.*, 2023b] Yuan Yang, Deepayan Sanyal, James Ainooson, Joel Michelson, Effat Farhana, and Maithilee Kunda. A cognitively-inspired neural architecture for visual abstract reasoning using contrastive perceptual and conceptual processing. *arXiv:2309.10532*, 2023.

[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. *NeurIPS*, 32:1075–1087, 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.

[Zhang *et al.*, 2022a] Chi Zhang, Sirui Xie, Baoxiong Jia, Ying Nian Wu, Song-Chun Zhu, and Yixin Zhu. Learning algebraic representation for systematic generalization in abstract reasoning. In *European Conference on Computer Vision*, pages 692–709. Springer, 2022.

[Zhang *et al.*, 2022b] Wenbo Zhang, Site Mo, Xianggen Liu, Sen Song, et al. Learning robust rule representations for abstract reasoning via internal inferences. *NeurIPS*, 35:33550–33562, 2022.

[Zhao *et al.*, 2024] Kai Zhao, Chang Xu, and Bailu Si. Learning visual abstract reasoning through dual-stream networks. In *AAAI*, volume 38, pages 16979–16988, 2024.