A-I-RAVEN and I-RAVEN-Mesh: Two New Benchmarks for Abstract Visual Reasoning

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

We study generalization and knowledge reuse capabilities of deep neural networks in the domain of abstract visual reasoning (AVR), employing Raven's Progressive Matrices (RPMs), a recognized benchmark task for assessing AVR abilities. Two knowledge transfer scenarios referring to the I-RAVEN dataset are investigated. Firstly, inspired by generalization assessment capabilities of the PGM dataset and popularity of I-RAVEN, we introduce *Attributeless-I-RAVEN* (A-I-RAVEN), a benchmark with 10 generalization regimes that allow to systematically test generalization of abstract rules applied to held-out attributes at various levels of complexity (primary and extended regimes). In contrast to PGM, A-I-RAVEN features compositionality, a variety of figure configurations, and does not require substantial computational resources. Secondly, we construct *I-RAVEN-Mesh*, a dataset that enriches RPMs with a novel component structure comprising line-based patterns, facilitating assessment of progressive knowledge acquisition in transfer learning setting. We evaluate 13 strong models from the AVR literature on the introduced datasets, revealing their specific shortcomings in generalization and knowledge transfer.

1 Introduction

Generalization, the ability of a model to perform well on unseen data, remains a fundamental challenge in deep learning (DL). While DL methods have demonstrated remarkable achievements in various domains, their generalization capabilities are often questioned, particularly in tasks that demand abstract problem-solving and reasoning skills [Chollet, 2019]. One such domain is abstract visual reasoning (AVR) [Mitchell, 2021; Stabinger *et al.*, 2021; van der Maas *et al.*, 2021; Małkiński and Mańdziuk, 2023] that encompasses tasks requiring (human) fluid intelligence – an aspect of human cognition believed to be crucial for reasoning in never-encountered settings [Carpenter *et al.*, 1990]. The most popular AVR tasks are Raven's Progressive Matrices (RPMs) [Raven, 1936; Raven and Court, 1998], which constitute a common problem found in human IQ tests. Typical RPMs comprise two components – the context panels arranged in a 3×3 grid with the bottom-right panel missing and up to 8 answer panels, out of which only one correctly completes the matrix. Solving an RPM instance requires identification of underlying abstract rules applied to certain attributes of the objects composing the instance (see Fig. 1 for an illustrative example).

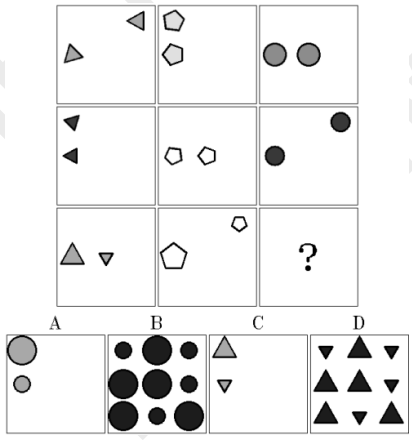


Figure 1: **RPM example.** The correct answer is A.

Design of computational methods capable of tackling RPMs has for decades been an active area of research [Evans, 1964; Foundalis, 2006; Lovett *et al.*, 2007; Kunda *et al.*, 2010]. Consequently, a number of works considered automatic creation of RPM datasets [Matzen *et al.*, 2010; Wang and Su, 2015; Mańdziuk and Żychowski, 2019] and a wide suite of predictive models [Hernández-Orallo *et al.*, 2016; Hernández-Orallo, 2017] were proposed, with DL methods showing the most promising performance [Yang *et al.*, 2022; Małkiński and Mańdziuk, 2025b]. While this rapid progress led to exceeding the human level in particular problem setups [Wu *et al.*, 2020; Mondal *et al.*, 2023], a fundamental challenge of generalization to novel problem settings remains largely unattained.

Initial works designed several RPM datasets [Matzen *et al.*, 2010; Wang and Su, 2015; Hoshen and Werman, 2017], however, measuring generalization was not their focus. While some works explored knowledge transfer between related tasks [Mańdziuk and Żychowski, 2019; Tomaszewska *et al.*,

2022], the complexity of the datasets was limited and consequently they didn't pose a challenge for contemporary DL methods. To measure generalization in modern DL models, the PGM dataset was introduced [Barrett *et al.*, 2018]. PGM defines eight generalization regimes, each specifying the distribution of objects, rules and attributes in train and test splits. For instance, in the `Held-out Triples` split, a given rule–object–attribute triplet (e.g. Progression on Object's Size) was assigned only to one of the two splits. In effect, the models were tested on triplet combinations different from training ones, allowing to assess their generalization capabilities. A subsequent work proposed RAVEN [Zhang *et al.*, 2019a], another RPM dataset with enriched perceptual complexity of matrices instantiated in seven visual configurations (`Center`, `2x2Grid`, `3x3Grid`, `Left-Right`, `Up-Down`, `Out-InCenter`, `Out-InGrid`). Moreover, the benchmark is characterized by a moderate sample size, i.e. 70K instances, compared to 1.42M RPMs per each of the eight regimes in PGM. Due to this size disparity, subsequent research gravitated towards RAVEN and its revised variants (I-RAVEN [Hu *et al.*, 2021] and RAVEN-Fair [Benny *et al.*, 2021]), which didn't require substantial computational resources to train DL models.

Contribution. Drawing inspiration from the broad adoption of RAVEN and the generalization assessment capabilities of PGM, this paper proposes a novel suite of generalization challenges stemming from I-RAVEN [Hu *et al.*, 2021] (a revised variant of RAVEN that removes a bias in RAVEN's answer panels). However, unlike I-RAVEN, the proposed suite of benchmarks allows for a direct assessment of the generalization and knowledge transfer of AVR models. Compared to PGM, our datasets feature compositionality and variety of figure configurations, and their processing doesn't require substantial computational resources. Furthermore, they include structural annotations, which are utilized, for example, in recent neuro-symbolic approaches [Zhang *et al.*, 2021; Zhang *et al.*, 2022].

First, we introduce *Attributeless-I-RAVEN* (A-I-RAVEN), comprising 10 generalization regimes. The 4 primary regimes correspond to specific held-out attributes ({`Position`, `Type`, `Size`, `Color`}), resp. The training matrices in these regimes adhere to the `Constant` rule for the respective attribute, whereas test matrices employ a rule different from `Constant` for this attribute (i.e., `Progression`, `Arithmetic`, or `Distribute Three`). Moreover, we propose 6 extended regimes: 3 of them feature a held-out attribute pair, while another 3 replace the `Constant` rule in the training set with each remaining rule. In effect, each regime comprises different distributions of training and test data.

Next, we propose *I-RAVEN-Mesh*, a variant of I-RAVEN with a new grid-like structure overlaid on the matrices. The dataset enables assessing generalization to incrementally added structures and progressive knowledge acquisition in a transfer learning (TL) setting.

Investigations involving 13 contemporary AVR DL models reveal that the introduced benchmarks present a substantial challenge for the tested methods, raising the need for further advancements in this area.

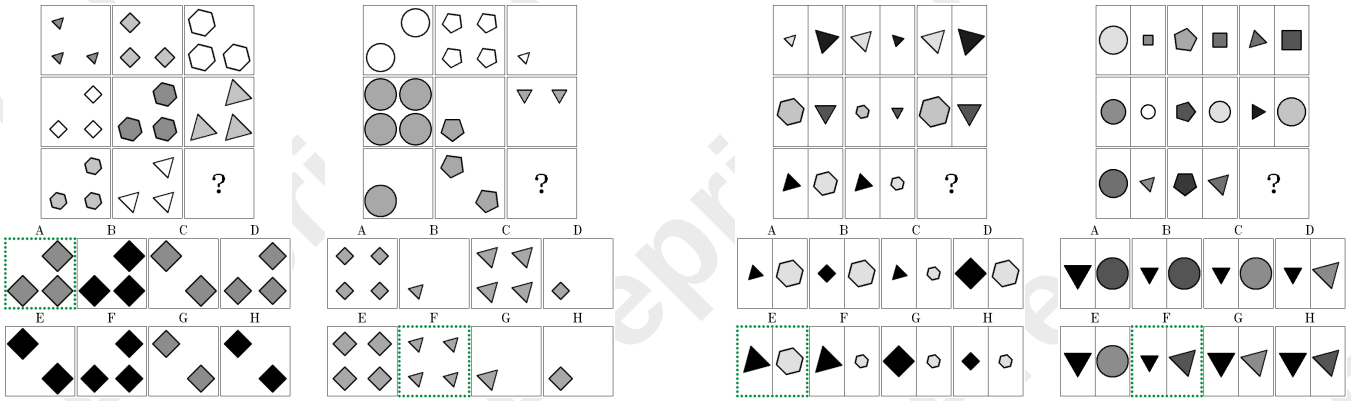
The key contributions of the paper are summarized below.

- We introduce the A-I-RAVEN dataset that enables measuring generalization across 10 regimes.
- We construct *I-RAVEN-Mesh*, an extension of I-RAVEN with a new component structure that facilitates assessment of progressive knowledge acquisition in a TL setting.
- We evaluate the performance of state-of-the-art AVR models on the introduced benchmarks, uncovering their limitations in terms of generalization to novel problem settings.

2 Related Work

Generalization in AVR. In recent years, a variety of AVR problems and corresponding datasets have emerged [Nie *et al.*, 2020; Fleuret *et al.*, 2011; Qi *et al.*, 2021; Shanahan *et al.*, 2020; Hill *et al.*, 2019; Zhang *et al.*, 2020] and several attempts have been made to measure generalization in contemporary AVR models based on the introduced benchmarks. In particular, distinct visual configurations were employed in RAVEN to assess how a model trained on one configuration performs on the remaining ones [Zhang *et al.*, 2019a; Spratley *et al.*, 2020; Zhuo and Kankanhalli, 2021]. Although in such a setting the visual aspects of train/test matrices come from different distributions, the underlying rules and attributes remain the same. In contrast, A-I-RAVEN enables studying the generalization of rules applied to held-out attributes, shifting the focus from perception towards reasoning. Besides RPMs, the limits of generalization have been explored in other AVR tasks as well. Visual Analogy Extrapolation Challenge evaluates model's capacity for extrapolation [Webb *et al.*, 2020]. However, such specialized datasets might favor models that explicitly embed the notion of extrapolation in their design and aim for being invariant only to specific attributes such as object size or location. Differently, our benchmarks allow verifying the model's capacity to learn a given concept from the data and generalize it to novel settings. This perspective links our work to the recent literature on concept learning [Moskvichev *et al.*, 2023]. However, the concept-oriented benchmarks that originate from ARC [Chollet, 2019] remain largely unsolved by DL models and pose a significant challenge even for leading multi-modal large language models [Mitchell *et al.*, 2023]. In contrast, both benchmarks proposed in this work are attainable by DL models, though further advances in generalization abilities of the models are necessary to consider them solved.

Model architectures. Preliminary attempts to solve RPMs with DL models involve WReN [Barrett *et al.*, 2018] that reasons over object relations using Relation Network [Santoro *et al.*, 2017], or SRAN [Hu *et al.*, 2021] that relies on a hierarchical architecture with panel encoders devoted to particular image groups. A common theme enabling generalization in DL models is to explicitly identify RPM objects. To this end, RelBase [Spratley *et al.*, 2020] employs Attend-Infer-Repeat, an unsupervised scene decomposition method, STSN [Mondal *et al.*, 2023] utilizes Slot attention [Locatello *et al.*, 2020]



(a) `A/Position` train (b) `A/Position` test (c) `A/Color` train (d) `A/Color` test

Figure 2: **A-I-RAVEN**. Left: Matrices from the `A/Position` regime belonging to the `2×2 Grid` configuration. In (a), object position is constant across rows, while in (b) object numerosity is governed by `Distribute Three`. Right: Matrices from the `A/Color` regime belonging to the `Left-Right` configuration. In (c), object color is constant across rows in left and right image parts, while in (d) it's governed by `Progression`. Correct answers are marked in a green dotted border. Please refer to [Małkiński and Mańdziuk, 2025a, Appendix A] for examples from other generalization regimes.

to decompose matrix to slots containing particular objects and Temporal Context Normalization (TCN) [Webb *et al.*, 2020] to normalize latent matrix panel representations in a task-specific context, DRNet [Zhao *et al.*, 2024] relies on a dual-stream design, and MRNet [Benny *et al.*, 2021] presents a multi-scale architecture. SCL [Wu *et al.*, 2020] proposes the scattering transformation, CoPINet [Zhang *et al.*, 2019b] and CPCNet [Yang *et al.*, 2023b] rely on contrastive architectures, PredRNet [Yang *et al.*, 2023a] learns to minimize the prediction error, ALANS [Zhang *et al.*, 2021] and PrAE [Zhang *et al.*, 2022] employ neuro-symbolic architectures, and SCAR [Małkiński and Mańdziuk, 2024b] adapts its computation to the structure of the considered matrix. Despite the high variety of AVR models, experiments on the introduced benchmarks reveal their shortcomings in terms of generalization and knowledge transfer.

3 Proposed Datasets

The set of attributes in I-RAVEN is $\mathcal{A} = \{$`Position`, `Number`, `Type`, `Size`, `Color`$\}$ and the set of rules is $\mathcal{R} = \{$`Constant`, `Progression`, `Arithmetic`, `Distribute Three`$\}$. For attribute $a \in \mathcal{A}$ and a dataset split $s \in \mathcal{S}$, where $\mathcal{S} = \{$train, val., test$\}$, we define the set of rules applicable to a in split s by $R(a, s) \subseteq \mathcal{R}$. In I-RAVEN all rule–attribute pairs are valid in all splits:

$$R(a, s) = \mathcal{R}, \quad \forall a \in \mathcal{A} \wedge \forall s \in \mathcal{S} \tag{1}$$

3.1 Attributeless-I-RAVEN (A-I-RAVEN)

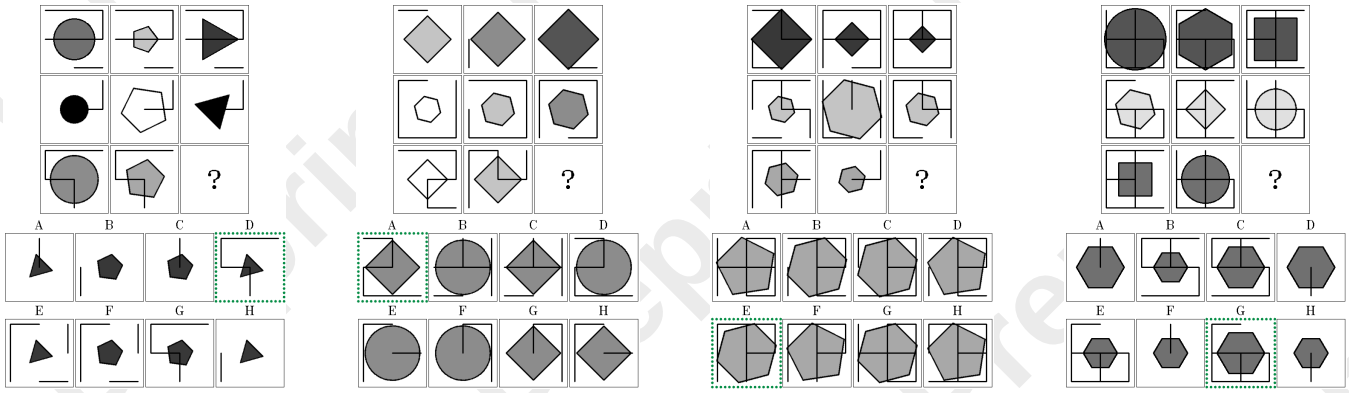
To probe generalization in DL models, we present A-I-RAVEN, a benchmark composed of 10 generalization regimes. Example matrices are illustrated in Fig. 2, with additional samples provided in [Małkiński and Mańdziuk, 2025a, Appendix A]. Each regime defines a set of held-out attributes A^*, each with a corresponding rule $r^*(a), a \in A^*$. In train and validation splits, held-out attribute $a \in A^*$ is governed by $r^*(a)$. In the test split, $a \in A^*$ is governed by a different rule sampled from $\mathcal{R} - \{r^*(a)\}$. In effect, during training, the model doesn't see rule–attribute combinations required to solve test matrices. There are no rule-related constraints on the remaining attributes. In summary, we have:

$$R(a, s) = \begin{cases} \{r^*(a)\} & \text{if } a \in A^* \wedge s \in \{\text{train, val.}\}, \\ \mathcal{R} - \{r^*(a)\} & \text{if } a \in A^* \wedge s = \text{test}, \\ \mathcal{R} & \text{if } a \notin A^*. \end{cases} \tag{2}$$

We define 4 primary regimes with $r^*(a) =$ `Constant` that correspond to individual held-out attributes ($|A^*| = 1$), denoted as `A/<Attribute>` (e.g., `A/Type`). Since `Position` and `Number` attributes are tightly coupled (e.g., it's impossible to increase cardinality of objects while keeping their position constant), we allocate a single generalization regime, `A/Position`, to cover both attributes. In addition, we define 6 extended regimes as supplementary generalization challenges. In the first group a pair of attributes is held-out in the training set, i.e. $|A^*| = 2$. Specifically, we introduce 3 new regimes: `A/ColorSize`, `A/ColorType`, and `A/SizeType`, based on the respective attribute pairs. In the second group, `Constant` rule in $r^*(a)$ is replaced with each of the 3 remaining rules, leading to `A/Color-Progression`, `A/Color-Arithmetic`, and `A/Color-DistributeThree` regimes. While this modification could be applied to all the described regimes, we focus on the `Color` attribute due to its broad range of possible values.

3.2 I-RAVEN-Mesh

The other of the proposed benchmarks is designed to probe progressive knowledge acquisition in a TL setting. I-RAVEN-Mesh extends I-RAVEN by introducing a novel visual component overlaid on top of the existing I-RAVEN components (see Fig. 3). Though the dataset can serve as a learning challenge on its own, the main motivation behind its introduction is to employ models pre-trained on I-



(a) `Constant` (b) `Progression` (c) `Arithmetic` (d) `Distribute Three`

Figure 3: **I-RAVEN-Mesh.** Matrices with the `Position` attribute of the mesh component governed by all applicable rules. For the sake of readability, we present examples belonging to the `Center` configuration. (a) Line position is constant in each row. (b) The line pattern displayed in the first column is rotated by 90 degrees in subsequent columns. (c) The union set operator applied to the first and the second column produces line positions in the third column. (d) Each row contains lines arranged in one out of three available patterns. Correct answers are marked in a green dotted border. Please refer to [Małkiński and Mańdziuk, 2025a, Appendix A] for examples concerning the `Number` attribute.

Attribute	Rule	Description
`Number`	`Constant`	Each image in a given row contains the same number of lines.
	`Progression`	The count of lines in a given row changes by a constant factor (e.g. $2, 4, 6$).
	`Arithmetic`	The number of lines in the third column is determined based on an arithmetic operation applied to the preceding columns (e.g. $3 - 1 = 2$).
	`Distribute Three`	Three line counts are sampled and spread among images in a given row.
`Position`	`Constant`	Each image in a given row contains the same position of lines.
	`Progression`	A panel arrangement is sampled in each row and rotated by 90 degrees in subsequent columns.
	`Arithmetic`	The position of lines in the third column is computed based on a set operation (union or difference) applied to the preceding columns.
	`Distribute Three`	Three line arrangements are sampled and spread among images in a given row.

Table 1: Description of rule–attribute pairs in I-RAVEN-Mesh.

RAVEN and fine-tune them on I-RAVEN-Mesh with a configurable train sample size, facilitating analysis of their TL performance. The mesh grid comprises from 1 to 12 lines placed in predefined locations. The set of available lines covers the inner and outer edges of a 2×2 grid (12 lines in total). The mesh component has two attributes: $\mathcal{A}^{\text{mesh}} = \{\texttt{Number}, \texttt{Position}\}$, which govern the count and location of lines, respectively. To each attribute a rule $r \in \mathcal{R}$ can be applied. Table 1 describes the effect of applying a given rule–attribute pair to the mesh component. To generate the mesh component of an I-RAVEN-Mesh matrix, we sample one of the two attributes $a \in \mathcal{A}^{\text{mesh}}$ and a corresponding rule $r \in \mathcal{R}$ that governs its values. As the attributes often depend on each other (e.g., it's impossible to increase the number of lines while keeping their position constant), we don't constrain the value of the other attribute. The rule–attribute pairs for the base I-RAVEN components are generated in the same way as in the original dataset. To generate answers to the matrix, we follow the impartial algorithm proposed in I-RAVEN [Hu *et al.*, 2021]. In addition, each matrix contains at least one incorrect answer that differs from the correct one only in the mesh component, ensuring that the solver has to identify the correct rule governing the mesh component in order to solve the matrix. To facilitate training with an auxiliary loss, in which the model additionally predicts the representation of rules governing the matrix [Barrett *et al.*, 2018], we extend the base set of rule annotations with ones concerning the Mesh component.

4 Experiments

We assess generalization of state-of-the-art models for solving RPMs on A-I-RAVEN and evaluate progressive knowledge acquisition on I-RAVEN-Mesh.

Experimental setup. In all experiments we use the Adam optimizer [Kingma and Ba, 2014] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and a batch size set to 128. Learning rate is initialized to 0.001 and reduced 10-fold (at most 3 times) if no progress is seen in the validation loss in 5 subsequent epochs, and training stops early in the case of 10 epochs without progress. Unless stated otherwise, each model configuration was trained 3 times with a different seed, and we report mean and standard deviation for these runs. In each experiment, we utilize $42\,000$ training, $14\,000$ valida-

	I-RAVEN†	I-RAVEN (ours)	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	59.0 (±18.5)	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)

Table 2: **Single-task learning.** 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† provides results on I-RAVEN reported by model authors in the corresponding papers, while I-RAVEN (ours) presents results obtained with our experimental setup, which utilizes a typical configuration of an optimizer and learning rate scheduler without model-specific tuning, and doesn't involve data augmentation, see "Experimental setup" in Section 4 for details.

	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)

Table 3: **A-I-RAVEN extended regimes.** P, A, and D3 denote Progression, Arithmetic, and Distribute Three, resp.

tion, and 14 000 test matrices, following the standard data split protocol taken in prior works [Zhang *et al.*, 2019a; Hu *et al.*, 2021]. All models are trained with the auxiliary loss with sparse encoding [Małkiński and Mańdziuk, 2024a] and $\beta = 1$. Experiments were run on a worker with a single NVIDIA DGX A100 GPU.

Models. In addition to the simple CNN-LSTM baseline [Barrett *et al.*, 2018], we assess generalization of SOTA AVR models including WReN [Barrett *et al.*, 2018], CoPINet [Zhang *et al.*, 2019b], RelBase [Spratley *et al.*, 2020], SCL [Wu *et al.*, 2020], MRNet [Benny *et al.*, 2021], ALANS [Zhang *et al.*, 2021], SRAN [Hu *et al.*, 2021], PrAE [Zhang *et al.*, 2022], CPCNet [Yang *et al.*, 2023b], PredRNet [Yang *et al.*, 2023a], STSN [Mondal *et al.*, 2023], and DRNet [Zhao *et al.*, 2024]. For direct comparison, we evaluate all models on I-RAVEN following the above-described experimental setup.

Reproducibility. To guarantee reproducibility of experiments, we use a fixed set of random seeds and turn off hardware and framework features concerning indeterministic computation wherever possible. Together with the code, we provide the full training script that can be used to run all training jobs. The training job is packaged as a Docker image with fixed dependencies to isolate the configuration of the training environment. The released code allows for generation of all datasets from scratch, eliminating the dependency on file-hosting services required to distribute the data. The code for reproducing all experiments is publicly accessible at: https://github.com/mikomel/raven

4.1 Generalization on A-I-RAVEN

Main regimes. In the first set of experiments we evaluate all considered models on 4 primary generalization regimes of A-I-RAVEN. The results are presented in Table 2, along with the reference results on I-RAVEN and I-RAVEN-Mesh. The best outcomes on `A/Color` and `A/Position` are achieved by DRNet, followed by RelBase and SCL that perform comparably. In `A/Size` and `A/Type`, SCL outperforms other models, with DRNet and RelBase taking the second and third place, resp. Interestingly, the top 3 models present a mix of architectures comprising large models, such as DRNet that includes a Vision Transformer backbone (24.7M params), as well as small models, such as SCL and RelBase that include mainly convolutional and feed-forward layers (0.6M

and 1.3M params, resp.). This suggests that various architectural approaches may be taken to achieve reasonable generalization performance in solving RPMs.

Extended regimes. Table 3 shows the aggregated performance of all considered models on 6 extended A-I-RAVEN regimes. Similarly to the main regimes, the best results are achieved by SCL, RelBase, and DRNet. Overall, replacing the `Constant` rule in the training set of the `A/Color` regime with `Progression` yields a dataset of slightly lower complexity, as the best model on `A/Color-Progression` achieved 75.6% accuracy, a 5.6 p.p. increase compared to the best result on `A/Color`. Conversely, using the `Arithmetic` and `Distribute Three` rules increases the difficulty, as measured by the drop of the max accuracy by 3.3 p.p. and 4.3 p.p., resp. Furthermore, using a pair of held-out attributes significantly increases the complexity. For instance, in `A/ColorType`, the most challenging regime, the best result is only 32.0%. We conclude that A-I-RAVEN provides a suite of challenging regimes of variable complexity, in which even the best-performing models are far from solving all test matrices.

Dataset difficulty. Across all A-I-RAVEN regimes, the highest average result was achieved by SCL (56.3%), followed by DRNet (54.8%) and RelBase (54.1%). While SCL achieved 83.4% test accuracy on I-RAVEN, on A-I-RAVEN regimes it scored from 32.0% on `A/ColorType` to 76.7% on `A/Position`. Similar differences can be observed for all remaining models, which shows that generalization regimes of A-I-RAVEN pose a bigger challenge than the base dataset.

Fig. 4 displays the difference in performance of top-3 models on test and validation splits. On I-RAVEN and I-RAVEN-Mesh the difference is negligible, as in these datasets both splits follow the same distribution. However, the difference in attributeless regimes is significant, indicating the need for further research on generalization.

Tables 4 – 15 in [Małkiński and Mańdziuk, 2025a, Appendix C] present the results of all considered models on test and validation splits and the difference between these two splits for particular datasets/regimes. The difference in model performance between test and validation splits in I-RAVEN (Table 4) and I-RAVEN-Mesh (Table 5) is negligible. In A-I-RAVEN regimes, however, the difference is significant, showing limitations of all evaluated models in terms of generalization. Across 4 primary regimes (Tables 6 – 9), the biggest difference concerns the `A/Type` regime, suggesting that generalization of rules applied to novel shape types constitutes a real challenge for the contemporary models. In all 3 extended regimes concerning held-out attribute pairs (`A/ColorSize`, `A/ColorType`, and `A/SizeType`) the performance difference on test and validation splits is bigger than in the primary regimes (see Tables 10 – 12). This drop stems from overall weaker performance on the test split, confirming high difficulty of these regimes. Model performance on the next 3 regimes concerning the `Color` attribute and rules other than `Constant` (`A/Color-Progression`, `A/Color-Arithmetic`, and `A/Color-DistributeThree`) is better, though further progress in generalization is required to fully close the performance gap between test and validation splits (see Tables 13 – 15).

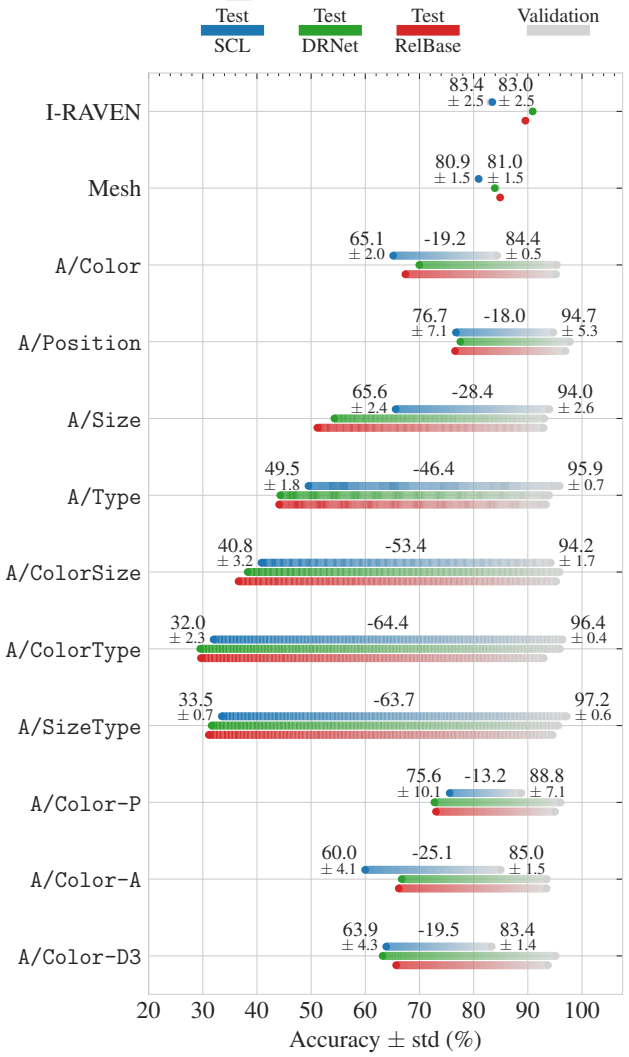


Figure 4: **Dataset difficulty.** Performance of top-3 models on test and validation splits. Numerical values refer to SCL scores.

Per-configuration results. Tables 16 – 27 in [Małkiński and Mańdziuk, 2025a, Appendix C] present the detailed results of all considered models for all matrix configurations. The most challenging configurations in I-RAVEN and I-RAVEN-Mesh are `3x3Grid` and `Out-InGrid`, in which image panels contain more objects than in the remaining configurations. Apparently, such setups require stronger reasoning capabilities to correctly identify the rules applied to multiple objects. Also, the results on the `Left-Right` and `Up-Down` configurations are relatively weaker in most regimes. In these configurations, rules may be applied to both matrix components (left/right and up/down, resp.), increasing the task complexity. This also concerns the `Out-InGrid` configuration in the `A/Size` regime, and the `Out-InCenter` configuration in the `A/SizeType` regime. Results in the `A/Position` regime are close-

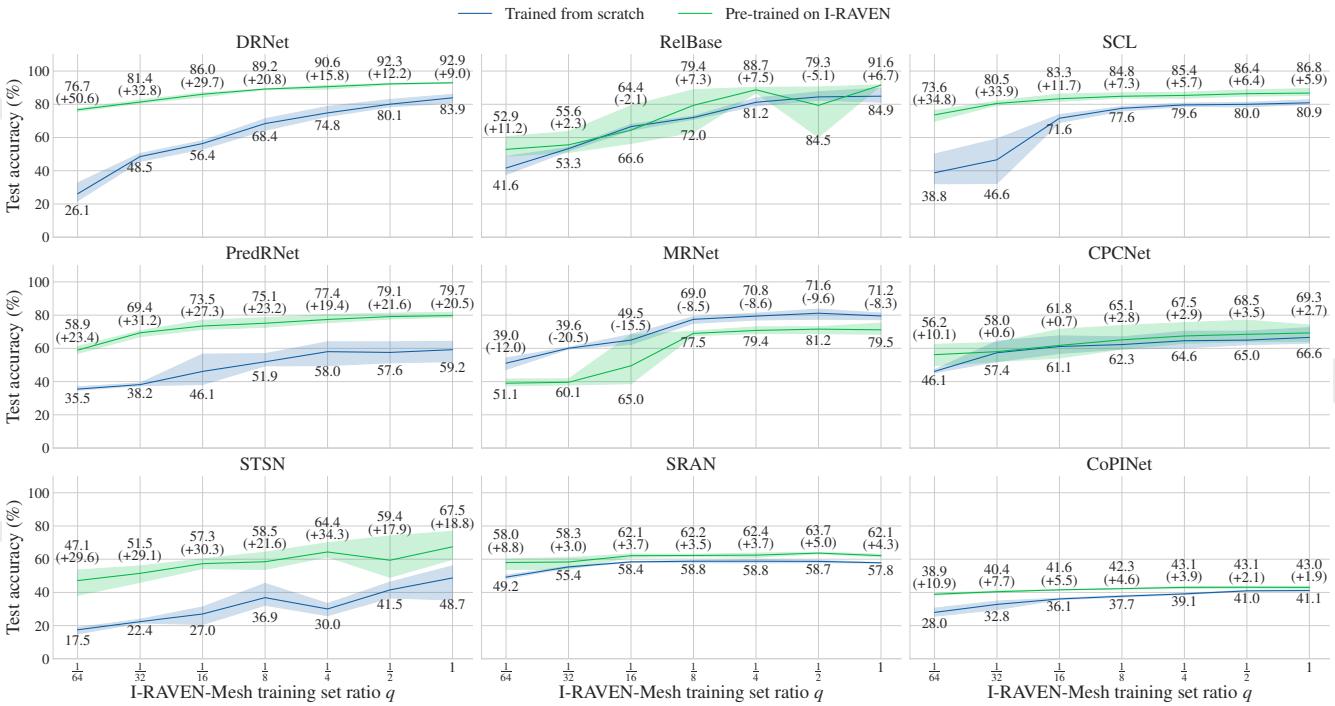


Figure 5: **Transfer learning.** Mean and standard deviation of test accuracy on I-RAVEN-Mesh across three random seeds. Models were trained in two setups: 1) from scratch on I-RAVEN-Mesh with variable sample size; 2) pre-trained on full I-RAVEN and fine-tuned on I-RAVEN-Mesh with variable sample size. Results for setups 1) and 2) are shown below and above the plot lines, resp.

to-perfect in configurations comprising a single object in each component (`Center`, `Left-Right`, `Up-Down`, and `Out-InCenter`) and weaker in the remaining configurations (`2x2Grid`, `3x3Grid` and `Out-InGrid`). This performance drop can be attributed to the fact that `Position` attribute can only be effectively applied to the `2x2Grid`, `3x3Grid` and `Out-InGrid` configurations allowing modification of the object's position. In the remaining configurations its application does not introduce any changes.

4.2 Progressive Knowledge Acquisition on I-RAVEN-Mesh

In the second set of experiments we employ I-RAVEN-Mesh to examine the TL ability of the best performing models. To this end, we consider variants of partial I-RAVEN-Mesh dataset with a fraction $q \in \{\frac{1}{64}, \ldots, 1\}$ of the training set and compare the performance of a model trained from scratch on a partial dataset to that of a model pre-trained on full I-RAVEN and fine-tuned on the respective part of I-RAVEN-Mesh. Fig. 5 shows that for $q = \frac{1}{64}$ pre-training RelBase, MRNet, CPCNet, SRAN, and CoPINet on I-RAVEN leads to gains smaller than 15 p.p., whereas pre-training DRNet, SCL, PredRNet, and STSN improved their accuracy by 50.6, 34.8, 23.4 and 29.6 p.p., resp. In addition, TL clearly improved performance of DRNet, SCL, PredRNet, and STSN in all considered settings. In particular for $q = 1$ by 9.0, 5.9, 20.5, and 18.8 p.p., resp., indicating the models' capacity for knowledge reuse.

5 Conclusion

We investigate generalization capabilities of DL models in the AVR domain. To accelerate research in this area, we propose two RPM benchmarks. A-I-RAVEN introduces 10 generalization regimes of variable complexity that assess model's capability to solve matrices with rules applied to novel attributes at various levels of complexity (primary and extended regimes). Contrary to the existing PGM dataset, A-I-RAVEN features compositionality, offers a variety of figure configurations, and above all does not require substantial computational resources. I-RAVEN-Mesh overlays line-based patterns on top of the RPM, facilitating TL studies. Experiments on 13 strong literature AVR models reveal their limitations in terms of generalization. We believe that the introduced datasets complement existing RPM benchmarks and will foster progress in the AVR area.

Limitations and future work. In this work we study generalization and knowledge transfer in contemporary AVR models employing RPM datasets. While RPMs are by far the most popular AVR tasks, the AVR domain also includes other types of problems not covered in the paper [Małkiński and Mańdziuk, 2023]. The Machine Number Sense dataset presents visual arithmetic problems [Zhang *et al.*, 2020], VAEC defines an extrapolation challenge [Webb *et al.*, 2020], while ARC proposes a set of diverse tasks in a few-shot learning setting [Chollet, 2019]. Similar studies could be performed on problems other than RPMs to test the performance and knowledge transfer abilities of AVR models in other problem settings.

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. This paper builds on the MSc thesis titled "Transfer learning in abstract visual reasoning domain" by Adam Kowalczyk from the Warsaw University of Technology, Warsaw, Poland.

References

[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.

[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.

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

[Evans, 1964] Thomas G Evans. A heuristic program to solve geometric-analogy problems. In *Proc. of the April 21-23, 1964, spring joint computer conference*, pages 327–338, 1964.

[Fleuret *et al.*, 2011] François Fleuret, Ting Li, Charles Dubout, Emma K Wampler, Steven Yantis, and Donald Geman. Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, 108(43):17621–17625, 2011.

[Foundalis, 2006] Harry E Foundalis. *Phaeaco: A cognitive architecture inspired by Bongard's problems.* PhD dissertation, Indiana University, 2006.

[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.

[Hernández-Orallo, 2017] José Hernández-Orallo. *The measure of all minds: evaluating natural and artificial intelligence*. Cambridge University Press, 2017.

[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.

[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.

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

[Kunda *et al.*, 2010] Maithilee Kunda, Keith McGreggor, and Ashok Goel. Taking a look (literally!) at the raven's intelligence test: Two visual solution strategies. In *Proc. of the Annual Meeting of the Cognitive Science Society*, volume 32, 2010.

[Locatello *et al.*, 2020] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *NeurIPS*, 33:11525–11538, 2020.

[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. *arXiv:2406.11061*, 2025.

[Małkiński and Mańdziuk, 2025b] 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 *et al.*, 2023] Melanie Mitchell, Alessandro B Palmarini, and Arseny Moskvichev. Comparing humans, GPT-4, and GPT-4V on abstraction and reasoning tasks. *arXiv:2311.09247*, 2023.

[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.

[Moskvichev *et al.*, 2023] Arsenii Kirillovich Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The conceptARC benchmark: Evaluating understanding and generalization in the ARC domain. *Transactions on Machine Learning Research*, 2023.

[Nie *et al.*, 2020] Weili Nie, Zhiding Yu, Lei Mao, Ankit B Patel, Yuke Zhu, and Anima Anandkumar. Bongard-LOGO: A new benchmark for human-level concept learning and reasoning. *NeurIPS*, 33:16468–16480, 2020.

[Qi *et al.*, 2021] Yonggang Qi, Kai Zhang, Aneeshan Sain, and Yi-Zhe Song. PQA: Perceptual question answering. In *CVPR*, pages 12056–12064, 2021.

[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.

[Shanahan *et al.*, 2020] Murray Shanahan, Kyriacos Nikiforou, Antonia Creswell, Christos Kaplanis, David Barrett, and Marta Garnelo. An explicitly relational neural network architecture. In *ICML*, pages 8593–8603. PMLR, 2020.

[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.*, 2022] Yuan Yang, Deepayan Sanyal, Joel Michelson, James Ainooson, and Maithilee Kunda. A conceptual chronicle of solving raven's progressive matrices computationally. In *Proceedings of the 8th International Workshop on Artificial Intelligence and Cognition*, 2022.

[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.*, 2020] Wenhe Zhang, Chi Zhang, Yixin Zhu, and Song-Chun Zhu. Machine number sense: A dataset of visual arithmetic problems for abstract and relational reasoning. In *AAAI*, volume 34, pages 1332–1340, 2020.

[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.*, 2022] 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.

[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.

[Zhuo and Kankanhalli, 2021] Tao Zhuo and Mohan Kankanhalli. Effective abstract reasoning with dual-contrast network. In *ICLR*, 2021.