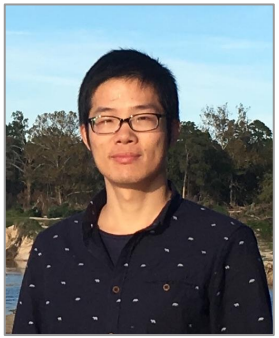
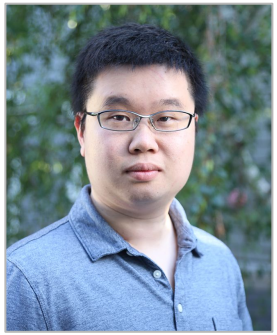
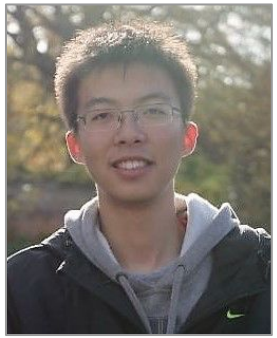
BONGARD-LOGO: A NEW BENCHMARK FOR HUMAN-LEVEL CONCEPT LEARNING AND REASONING



Weili Nie

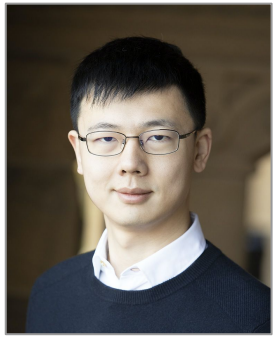


Zhiding Yu



Lei Mao

Ankit Patel



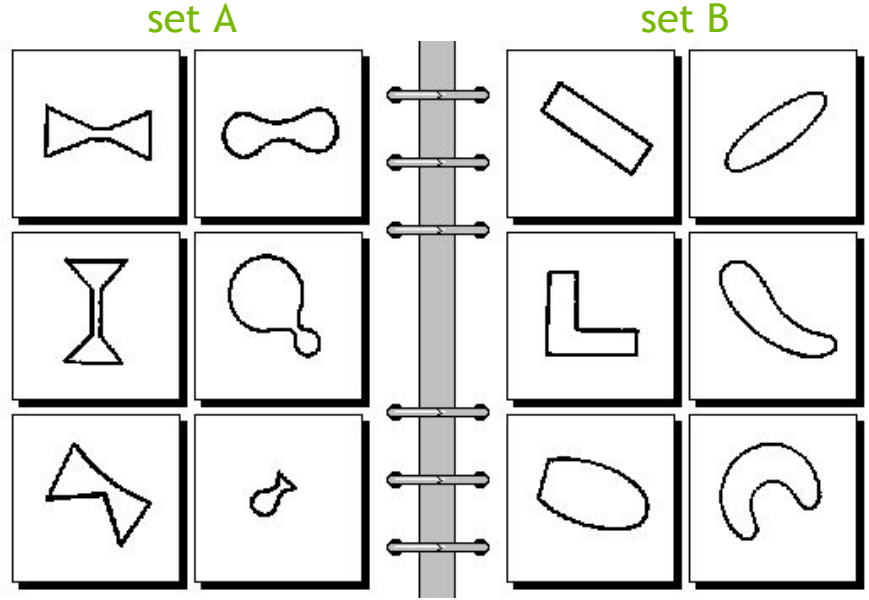
Yuke Zhu

Anima Anandkumar

BACKGROUND: BONGARD PROBLEMS

One Hundred puzzles originally invented by M. M. Bongard in 1967

- Bongard aimed to demonstrate the key properties of *human* visual cognition capabilities.
- Given a set A of six images (positive examples) and another set B of six images (negative examples),
- the *objective* is to discover the concept that the images in set A obey and images in set B violate.



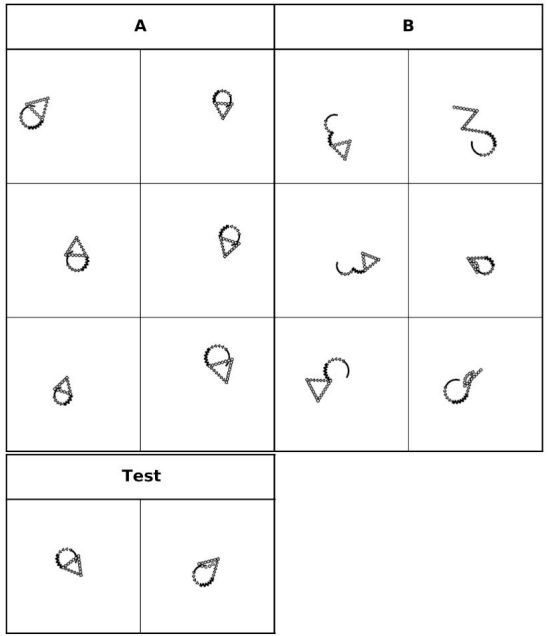
Problem #13 (A neck)

AN OVERVIEW OF BONGARD-LOGO

A benchmark inspired by original BPs for human-level visual concept learning and reasoning

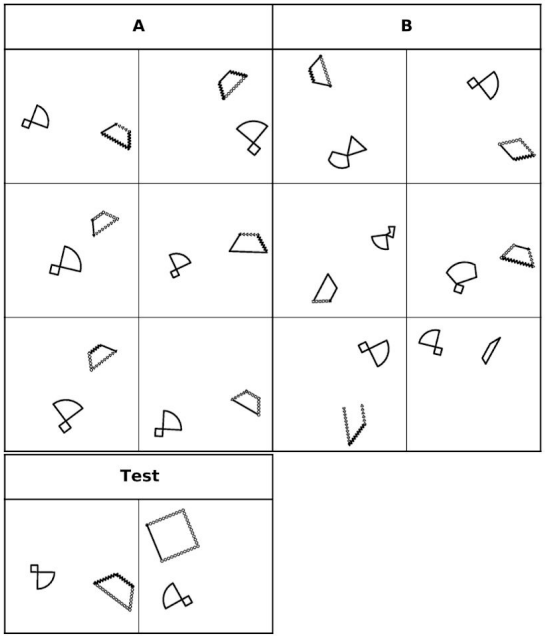
- It transforms concept learning into a *few-shot binary classification* problem
- It consists of *12,000* problem instances
  - The large scale makes it digestible by advanced machine learning methods in modern AI
- The problems in Bongard-LOGO belong to *three types* based on the concept categories:
  - *3,600 Free-form shape problems*
  - *4,000 Basic shape problems*
  - *4,400 Abstract shape problems*

THREE TYPES OF BONGARD-LOGO PROBLEMS



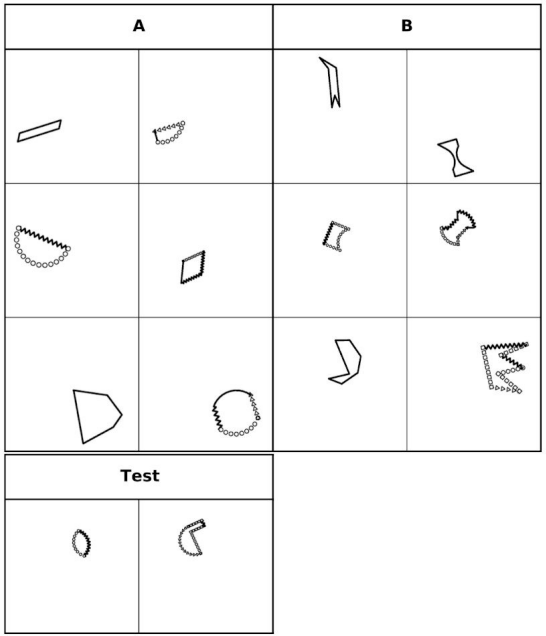
(a) free-from shape problem

(Concept: “ice cream cone”-like shape)



(b) basic shape problem

(Concept: A combination of “fan”-like shape and “trapezoid”)



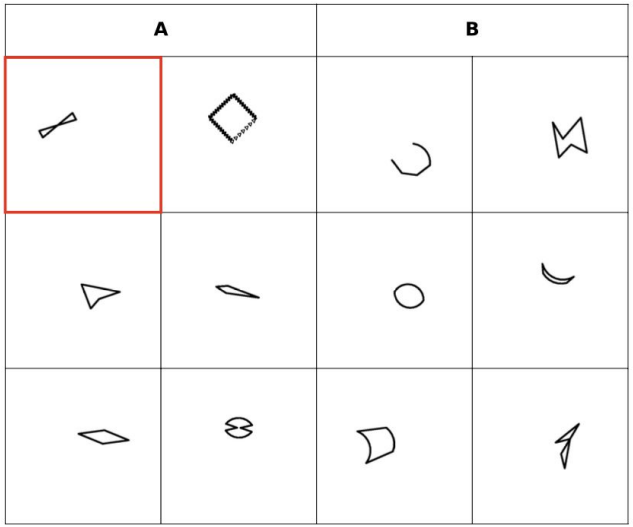
(c) abstract shape problem

(Concept: “convex”)

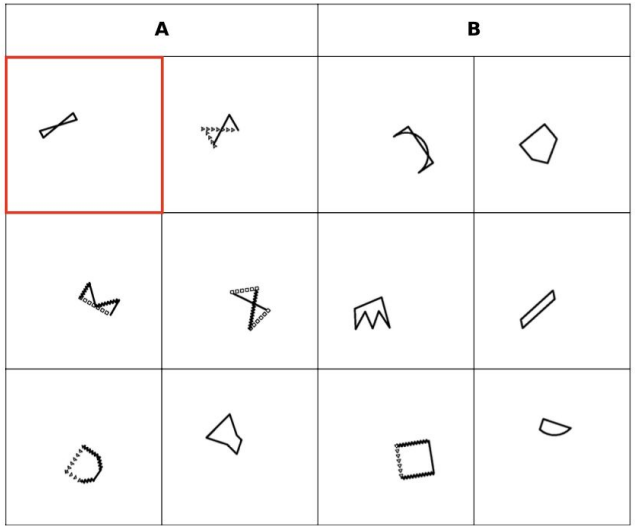
KEY PROPERTIES OF BONGARD-LOGO

It captures three core properties of human cognition exhibited in original BPs

- **Context-dependent perception**
  - The same shape pattern has fundamentally opposite interpretations depending on the context



(a) `have_four_straight_lines`

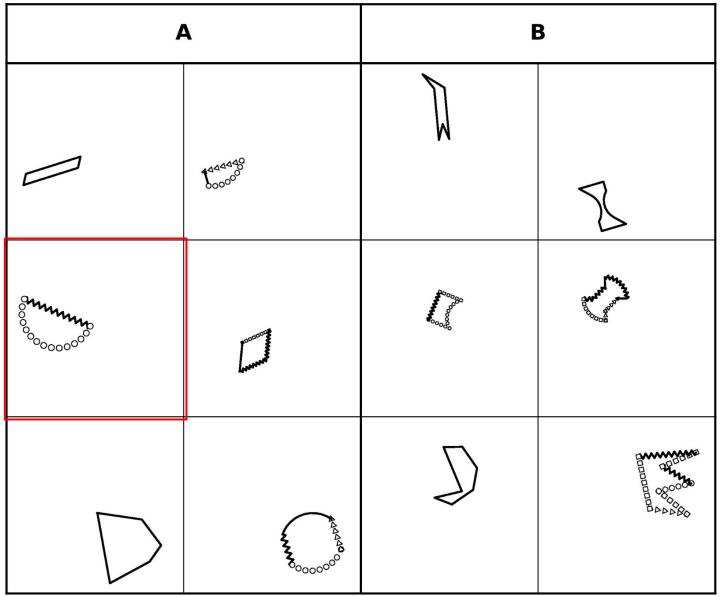


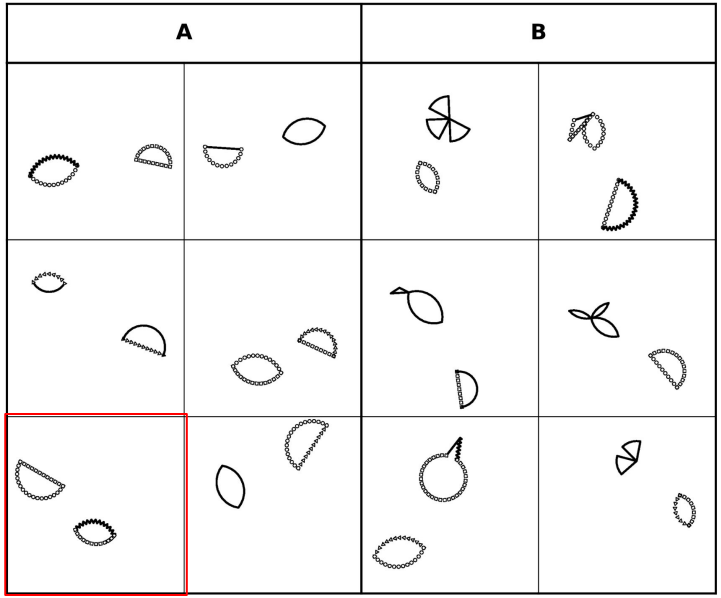
(b) `have_six_straight_lines`

KEY PROPERTIES OF BONGARD-LOGO

It captures three core properties of human cognition exhibited in original BPs

- **Analogy-making perception**
  - Some meaningful structures (i.e., zigzags or a set of circles) can be projected onto another meaningful ones (i.e., straight lines or arcs) for underlying concepts

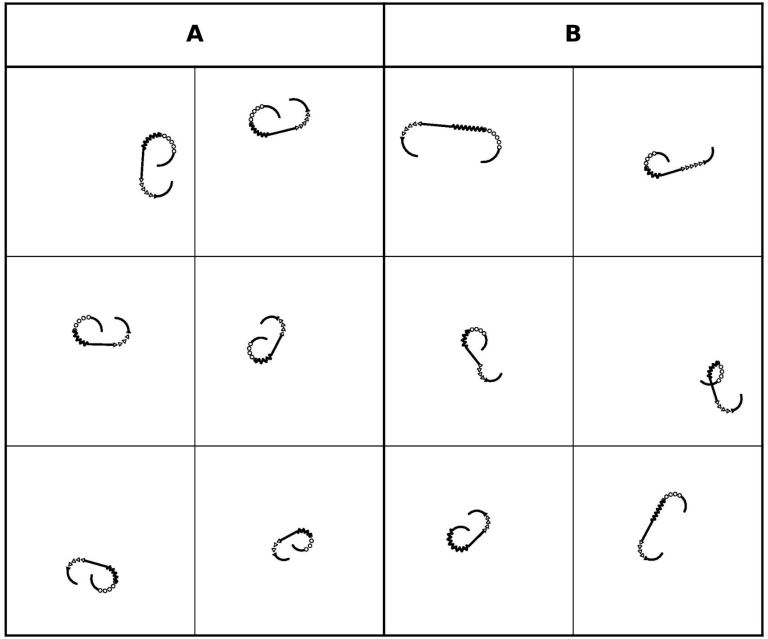


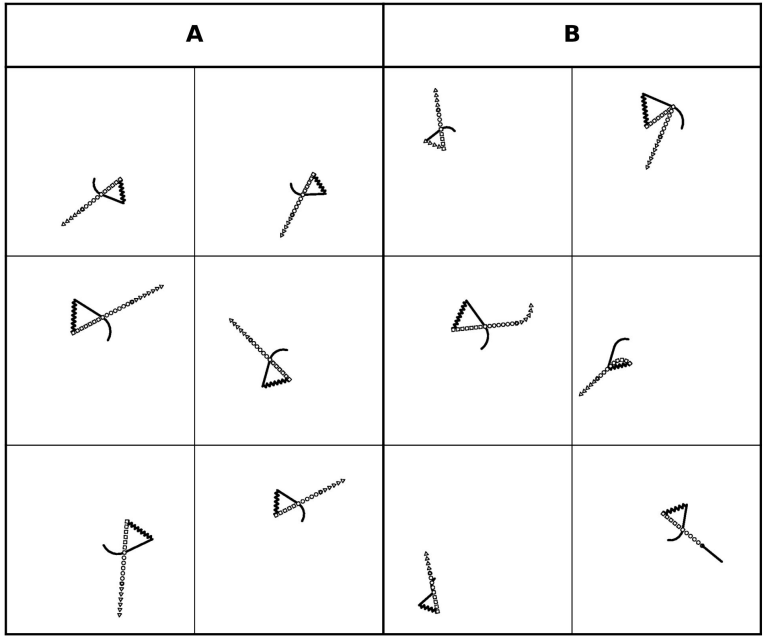


KEY PROPERTIES OF BONGARD-LOGO

It captures three core properties of human cognition exhibited in original BPs

- **Perception with a few examples but infinite vocabulary**
  - There is *no finite set of categories* to name and describe the geometrical arrangements



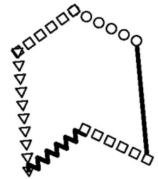


PROBLEM GENERATION

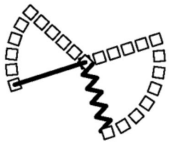
Automatically generating problems with action-oriented language

- We use *LOGO* language for procedural generation:
  - The *procedural commands* for drawing each shape form its ground-truth *action program*
  - Each action program is a list of actions and each action is depicted by a function:

    [Action name] (*[moving type], [moving length]* , *[moving angle]*)
- Two benefits:
  - Easily generate arbitrary shapes and precisely control the shape variation in a human-interpretable way
  - Provide a useful supervision in guiding symbolic reasoning in the action space



line(normal, 1.000, 0.500),
line(circle, 0.583, 0.664),
line(square, 0.583, 0.672),
line(triangle, 1.000, 0.664),
line(zigzag, 0.583, 0.836),
line(square, 0.583, 0.328)



line(zigzag, 0.500, 0.500),
arc(square, 0.625, 0.750),
line(square, 0.500, 0.750),
line(square, 0.500, 0.333),
arc(square, 0.583, 0.750),
line(normal, 0.500, 0.750)

Action Programs

BENCHMARKING ON BONGARD-LOGO

Comparing SOTA few-shot learning methods with human performance

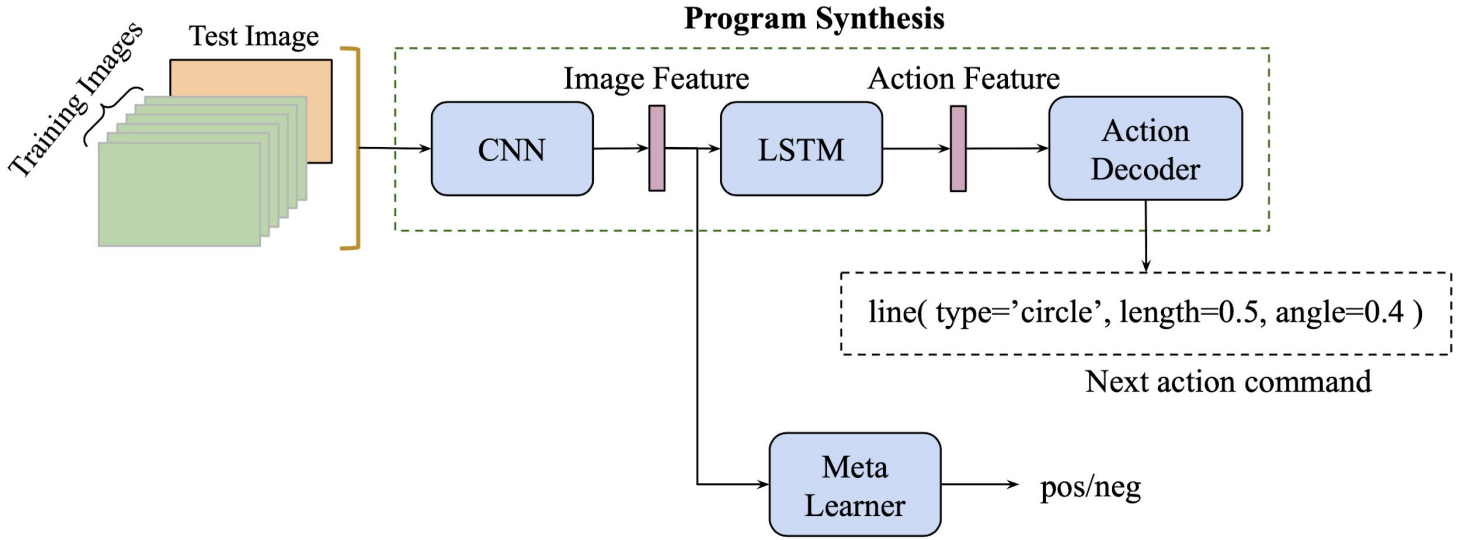
Methods	Train Acc	Test Acc (FF)	Test Acc (BA)	Test Acc (CM)	Test Acc (NV)
SNAIL [19]	59.2 ± 1.0	56.3 ± 3.5	60.2 ± 3.6	60.1 ± 3.1	61.3 ± 0.8
ProtoNet [20]	73.3 ± 0.2	64.6 ± 0.9	72.4 ± 0.8	62.4 ± 1.3	$\mathbf{65.4 \pm 1.2}$
MetaOptNet [21]	75.9 ± 0.4	60.3 ± 0.6	71.7 ± 2.5	61.7 ± 1.1	63.3 ± 1.9
ANIL [22]	69.7 ± 0.9	56.6 ± 1.0	59.0 ± 2.0	59.6 ± 1.3	61.0 ± 1.5
Meta-Baseline-SC [23]	75.4 ± 1.0	$\mathbf{66.3 \pm 0.6}$	$\mathbf{73.3 \pm 1.3}$	63.5 ± 0.3	63.9 ± 0.8
Meta-Baseline-MoCo [23]	$\mathbf{81.2 \pm 0.1}$	65.9 ± 1.4	72.2 ± 0.8	$\mathbf{63.9 \pm 0.8}$	64.7 ± 0.3
WReN-Bongard [17]	78.7 ± 0.7	50.1 ± 0.1	50.9 ± 0.5	53.8 ± 1.0	54.3 ± 0.6
CNN-Baseline	61.4 ± 0.8	51.9 ± 0.5	56.6 ± 2.9	53.6 ± 2.0	57.6 ± 0.7
Human (Expert)	-	92.1 ± 7.0	99.3 ± 1.9	90.7 ± 6.1	
Human (Amateur)	-	88.0 ± 7.6	90.0 ± 11.7	71.0 ± 9.6	

Test accuracy (%) on free-form shape test set (**FF**), basic shape test set (**BA**), combinatorial abstract shape test set (**CM**), and novel abstract shape test set (**NV**). Human (Expert) refers to human subjects who carefully follow our instructions while Human (Amateur) do not. The chance performance is 50%.

There is a significant gap between model and human performance

INCORPORATING SYMBOLIC INFORMATION

Meta-baseline based on program synthesis (Meta-Baseline-PS)



Stage I: Train the program synthesis module to predict action programs

Stage II: Use the pre-trained image feature to fine-tune the meta-learner

INCORPORATING SYMBOLIC INFORMATION

Meta-baseline based on program synthesis (Meta-Baseline-PS)

Methods	Train Acc	Test Acc (FF)	Test Acc (BA)	Test Acc (CM)	Test Acc (NV)
SNAIL [19]	59.2 ± 1.0	56.3 ± 3.5	60.2 ± 3.6	60.1 ± 3.1	61.3 ± 0.8
ProtoNet [20]	73.3 ± 0.2	64.6 ± 0.9	72.4 ± 0.8	62.4 ± 1.3	$\mathbf{65.4 \pm 1.2}$
MetaOptNet [21]	75.9 ± 0.4	60.3 ± 0.6	71.7 ± 2.5	61.7 ± 1.1	63.3 ± 1.9
ANIL [22]	69.7 ± 0.9	56.6 ± 1.0	59.0 ± 2.0	59.6 ± 1.3	61.0 ± 1.5
Meta-Baseline-SC [23]	75.4 ± 1.0	$\mathbf{66.3 \pm 0.6}$	$\mathbf{73.3 \pm 1.3}$	63.5 ± 0.3	63.9 ± 0.8
Meta-Baseline-MoCo [23]	$\mathbf{81.2 \pm 0.1}$	65.9 ± 1.4	72.2 ± 0.8	$\mathbf{63.9 \pm 0.8}$	64.7 ± 0.3
WReN-Bongard [17]	78.7 ± 0.7	50.1 ± 0.1	50.9 ± 0.5	53.8 ± 1.0	54.3 ± 0.6
CNN-Baseline	61.4 ± 0.8	51.9 ± 0.5	56.6 ± 2.9	53.6 ± 2.0	57.6 ± 0.7
Meta-Baseline-PS	$\mathbf{85.2 \pm 1.0}$	$\mathbf{68.2 \pm 0.3}$	$\mathbf{75.7 \pm 1.5}$	$\mathbf{67.4 \pm 0.3}$	$\mathbf{71.5 \pm 0.5}$
Human (Expert)	-	92.1 ± 7.0	99.3 ± 1.9	90.7 ± 6.1	
Human (Amateur)	-	88.0 ± 7.6	90.0 ± 11.7	71.0 ± 9.6	

Test accuracy (%) on free-form shape test set (**FF**), basic shape test set (**BA**), combinatorial abstract shape test set (**CM**), and novel abstract shape test set (**NV**). Human (Expert) refers to human subjects who carefully follow our instructions while Human (Amateur) do not. The chance performance is 50%.

Meta-Baseline-PS clearly outperforms previous SOTA methods

SUMMARY

A new benchmark for human-level visual concept learning and reasoning

- Bongard-LOGO scales up one Hundred original Bongard problems to a large dataset
- Bongard-LOGO demands a new form of human-like perception that is context-dependent, analogical, and of infinite vocabulary
- We developed a program-guided shape generation technique to produce Bongard-LOGO shapes in action-oriented LOGO language
- Large performance gap between human and machine in Bongard-LOGO reveals a failure of today's pattern recognition systems in capturing the core properties of human cognitive learning and reasoning.
- We showed that incorporating symbolic information into neural networks improves the overall performance, suggesting the advantages of neuro-symbolic methods on Bongard-LOGO