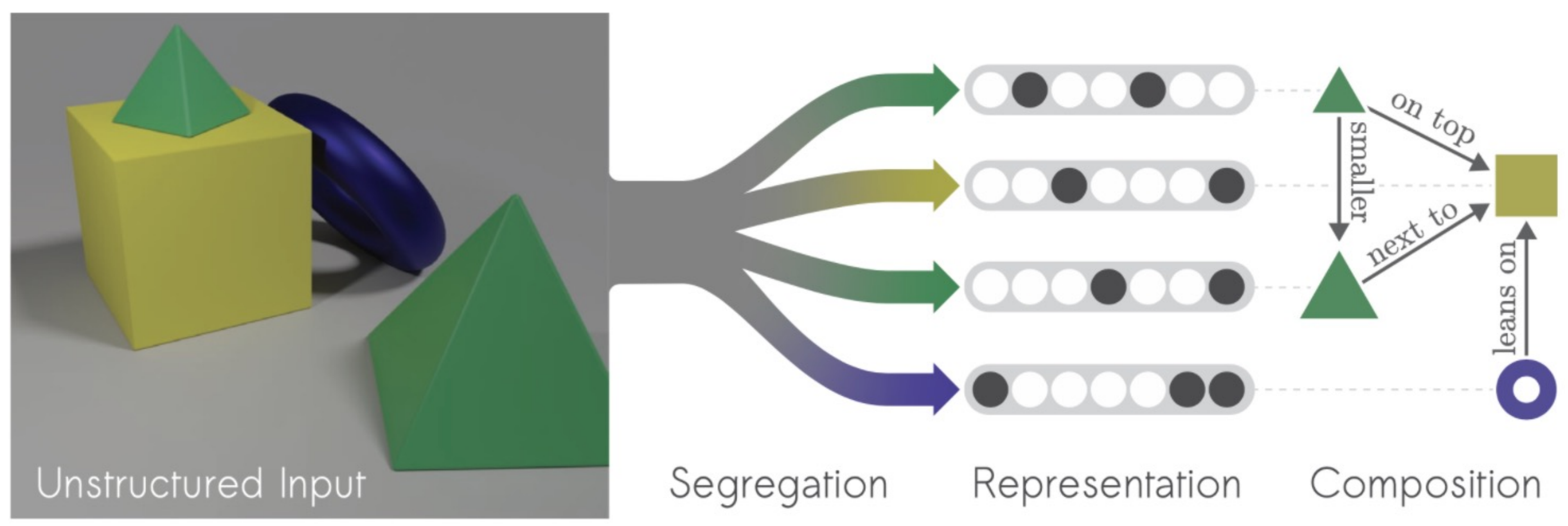
# RelViT: Concept-guided Vision Transformer for Visual Relational Reasoning

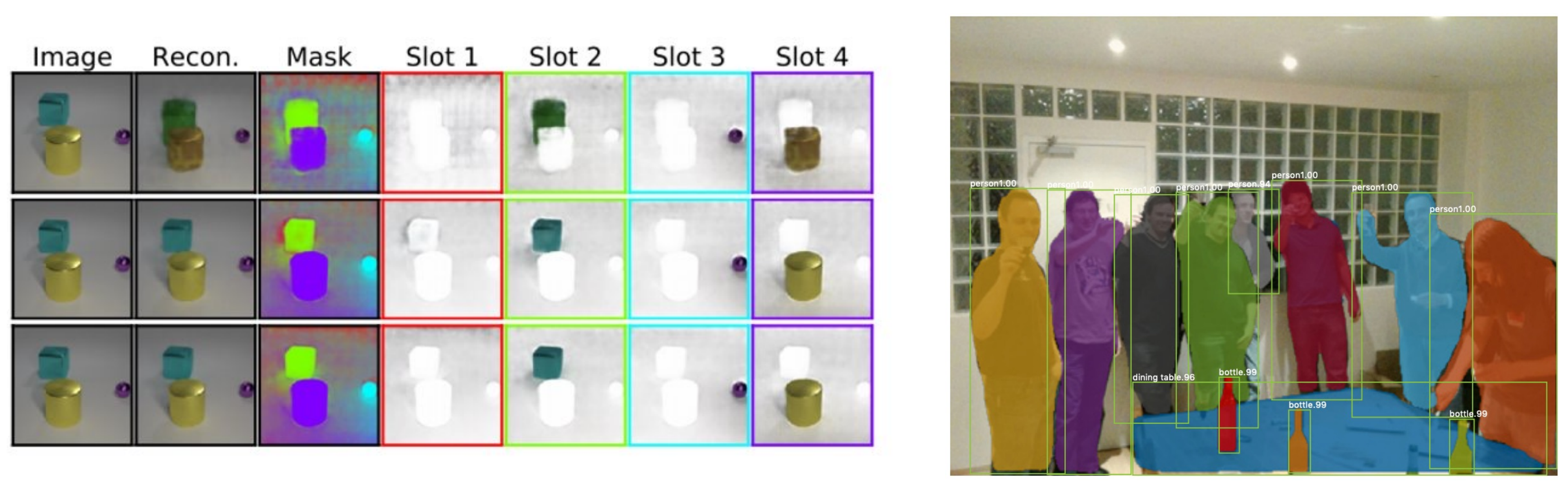
Xiaojian Ma, Weili Nie, Zhiding Yu, Huaizu Jiang, Chaowei Xiao, Yuke Zhu, Song-Chun Zhu, Anima, Anandkumar

## What makes visual reasoning so challenging?

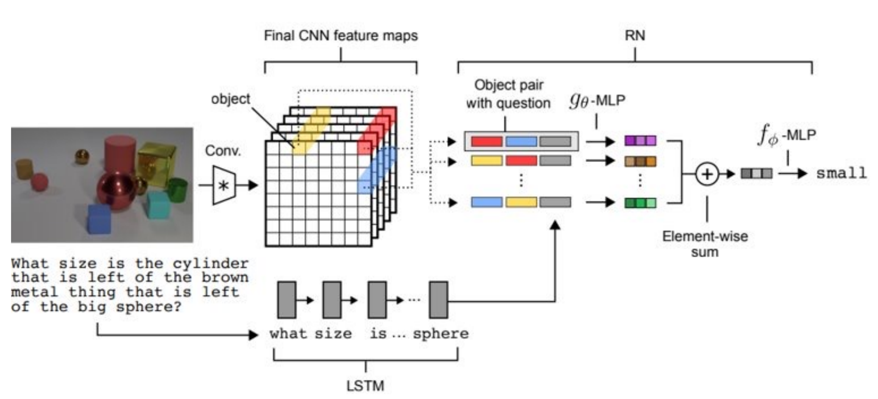
From representations to reasoning in human and AI [1]



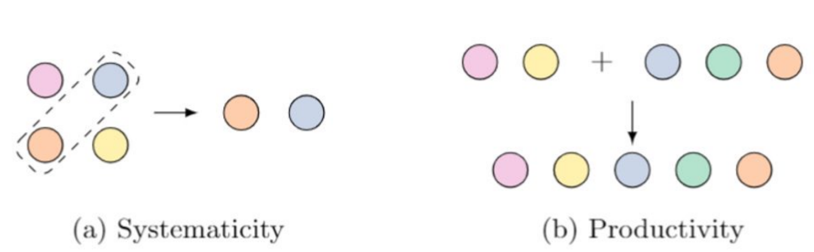

Ingredients for human-level visual reasoning [2,3,4,5]




Object-centric representations [2, 3]

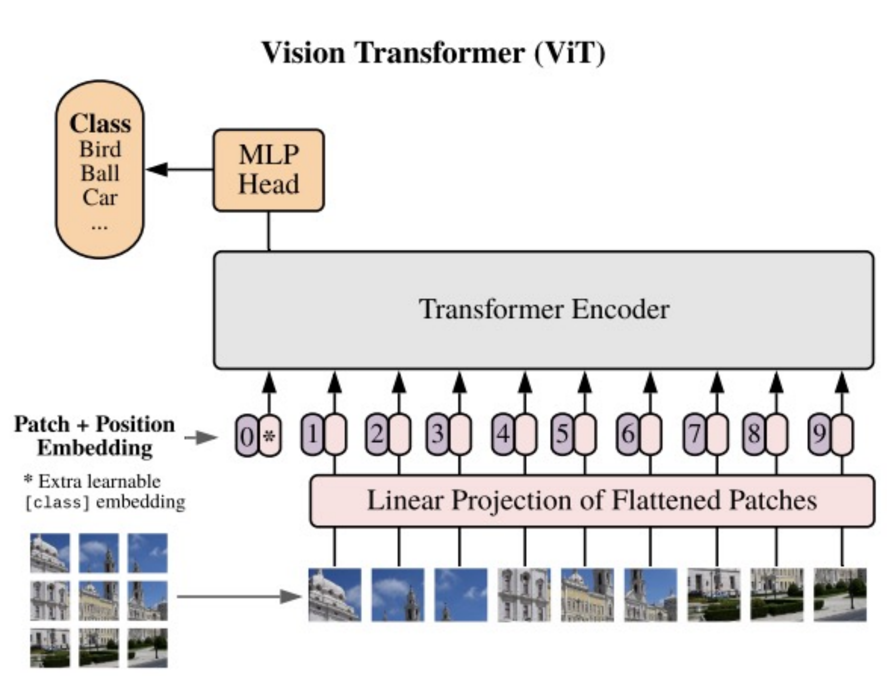


Relational inductive biases [4]




Systematic generalization [5]

## ViTs (partially) offer these ingredients [6]

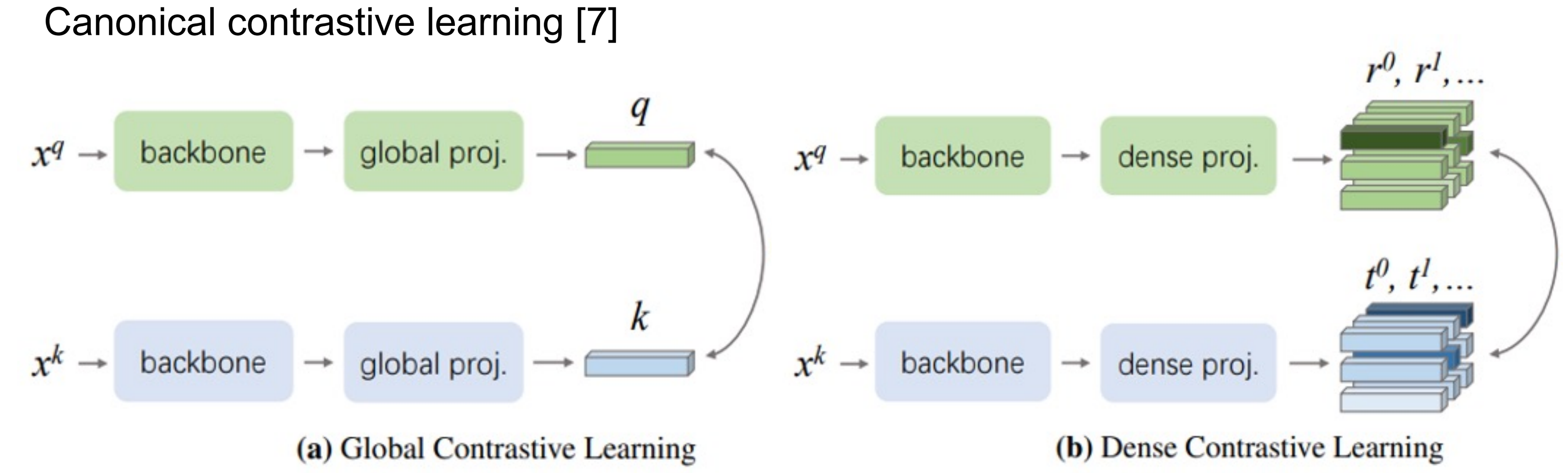



- **Image as patches**: image patches can be viewed as object candidates.
- **Self-attention**: Multi-head self attention (MHSA) in ViT effectively captures the pair-wise relations among input entities.

**Q1:** To which extent ViT helps with visual reasoning?
-- see experiments
**Q2:** Can we make it better?
-- contrastive learning seems helpful, let's give it a try in the regular learning pipeline.

## From contrastive learning to concept-guided contrastive learning

Canonical contrastive learning [7]




- The global CL can help with **relational meaning** and **reasoning**.
- The local CL can help with **object-centric representation** (via unsupervised correspondence learning).
- However, simply contrasting **two views** of the **same** picture could be inefficient, especially when we do know the semantic label of them.

Concept-guided contrastive learning

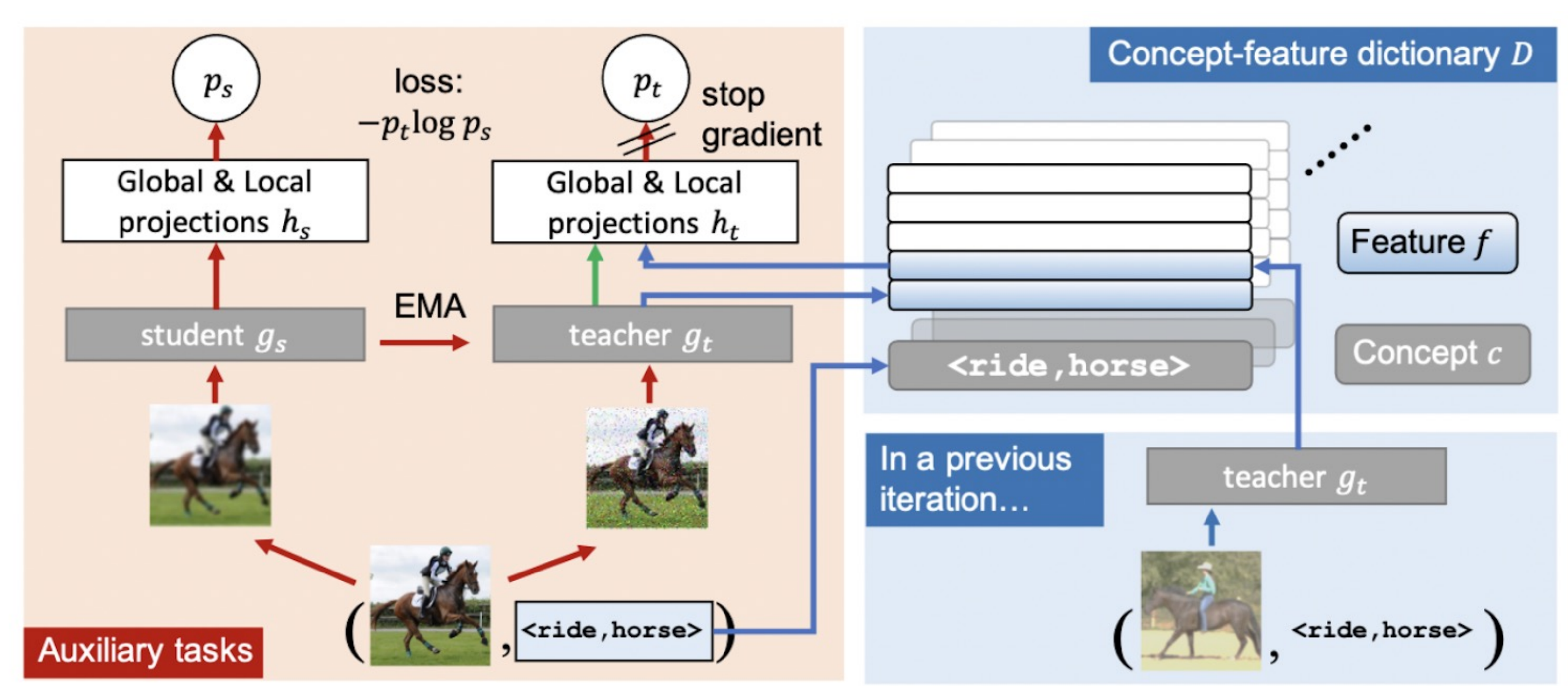



Figure 1: An overview of our method. Red+Green: the learning pipeline of DINO (Caron et al., 2021) and EsViT (Li et al., 2021); Red+Blue: our pipeline.

- We now contrast two (augmented) images with the **same semantics** instead.
- Each image is assumed to be paired with a **concept code** (can be parsed from the data, ex. questions in VQA)
- **Concept-feature dictionary** is introduced for retrieving images with the same concept on-the-fly.
- No significant overhead, **easy** to work with many training pipelines.

## Experiments

HICO

| Method | Ext. superv. | Backbone | Orig. | Systematic-easy Full cls. | Systematic-easy Unseen cls. | Systematic-hard Full cls. | Systematic-hard Unseen cls. |
|---|---|---|---|---|---|---|---|
| Mallya & Lazebnik (2016)* | | ResNet-101 | 33.8 | - | - | - | - |
| Girdhar & Ramanan (2017)* | bbox | ResNet-101 | 34.6 | - | - | - | - |
| Fang et al. (2018)* | pose | ResNet-101 | 39.9 | - | - | - | - |
| Hou et al. (2020)† | | ResNet-101 | 28.57 | 26.65 | 11.94 | 21.76 | 10.58 |
| ViT-only | | PVTv2-b2 | 35.48 | 31.06 | 11.14 | 19.03 | 18.85 |
| EsViT (2021) | | PVTv2-b2 | 38.23 | 35.15 | 11.53 | 22.55 | 21.84 |
| RelViT (Ours) | | PVTv2-b2 | 39.4 | 36.99 | 12.26 | 22.75 | 22.66 |
| RelViT + EsViT (Ours) | | PVTv2-b2 | **40.12** | **37.21** | **12.51** | **23.06** | **22.89** |

GQA

| Method | Bbox feat.* | Backbone | Orig. | Sys. |
|---|---|---|---|---|
| BottomUp (2018) | ✓ | ResNet-101 | 53.21 | - |
| MAC (2018b) | ✓ | ResNet-101 | 54.06 | - |
| MCAN-Small (2019) | ✓ | ResNet-101 | 58.35 | 36.21 |
| MCAN-Small (2019) | | ResNet-101 | 51.1 | 30.12 |
| ViT-only | | PVTv2-b2 | 56.62 | 31.39 |
| EsViT (2021) | | PVTv2-b2 | 56.95 | 31.76 |
| RelViT (Ours) | | PVTv2-b2 | 57.87 | 35.48 |


| GQA overall accuracy | MCAN-Small (w/ bbox) | RelViT (PVTv2-b2) | RelViT (PVTv2-b3) | RelViT (Swin-base) |
|---|---|---|---|---|
| original | 58.35 | 57.87 | 61.41 | **65.54** |
| systematic | 36.21 | 35.48 | 36.25 | **37.51** |

## Takeaway messages

- **Three ingredients for human-level visual reasoning**: object-centric representations, relational inductive bias and systematic generalization.
- **Vision transformer for human-level visual reasoning**: it help eliminate the need for object detection and complex reasoning modules.
- **Concept-guided contrastive learning** can further boost ViT's potentials on solving systematic generalization.

## References


[1] "On the Binding Problem in Artificial Neural Networks" In: arXiv
[2] "Object-Centric Learning with Slot Attention" In: NeurIPS
[3] "Mask R-CNN" In: ICCV
[4] "A simple neural network module for relational reasoning" In: NeurIPS
[5] "Compositionality decomposed: how do neural networks generalise?" In: JAIR
[6] "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" In: ICLR
[7] "Dense Contrastive Learning for Self-Supervised Visual Pre-Training" In: CVPR


Paper

Code