Robust Visual Understanding

Knowledge-Guided and Multimodal Reasoning

Tejas Gokhale

Assistant Professor

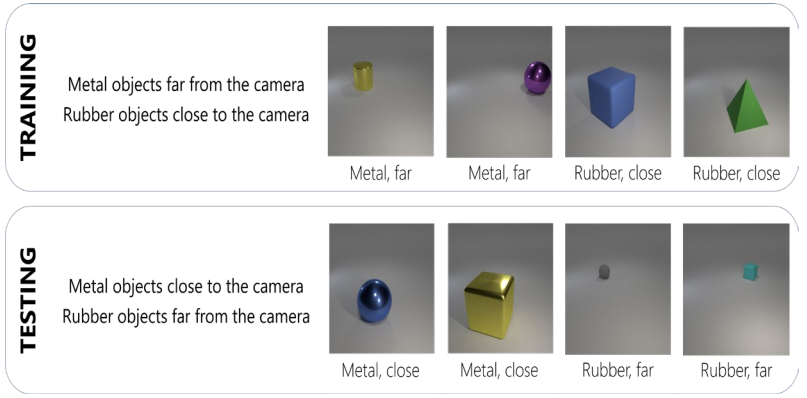
University of Maryland, Baltimore County

My lab's focus: Perception & Reasoning with Robustness

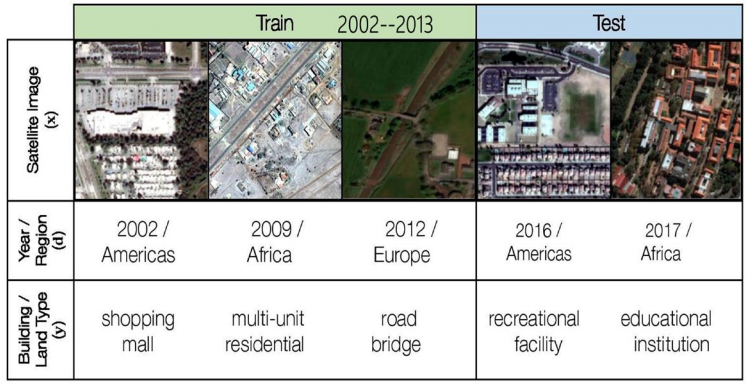
Robust Image Recognition

Dealing with Style Shift

Dealing with Attribute-Level Shift

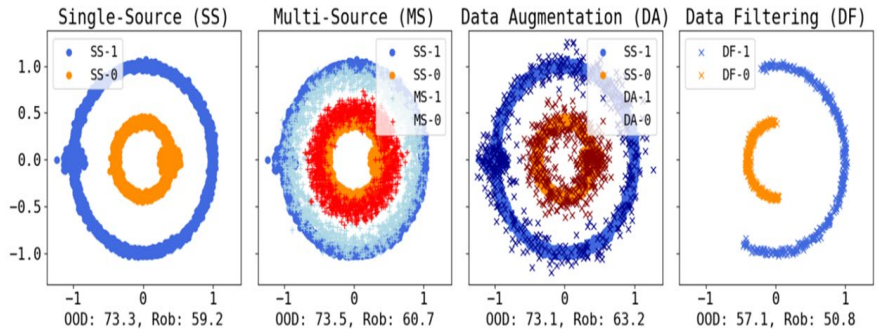


Robust Overhead image recognition

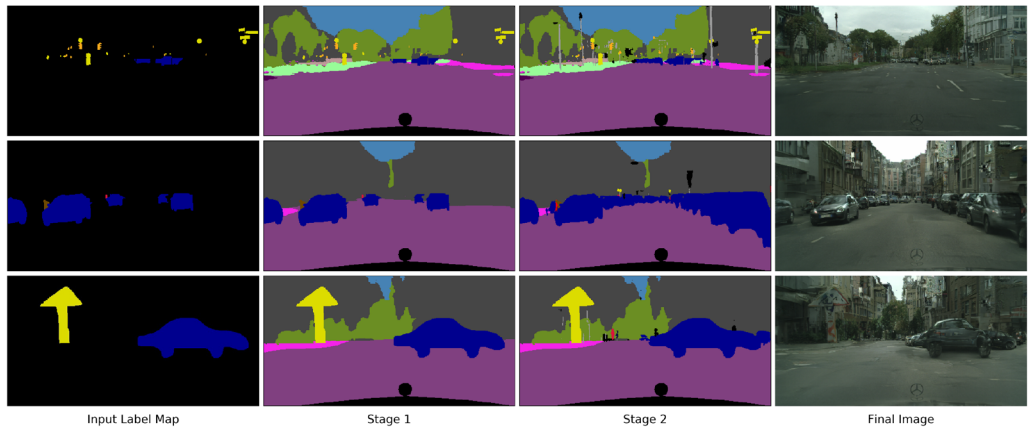


Gokhale AAAI'21;
Gokhale ACL'22;
Gokhale WACV'23;
Cheng ICCV '23;
Wisdom arxiv 2023;
Kulkarni CVPR-W'21

Effects of multiple data sources on OOD and adversarial robustness



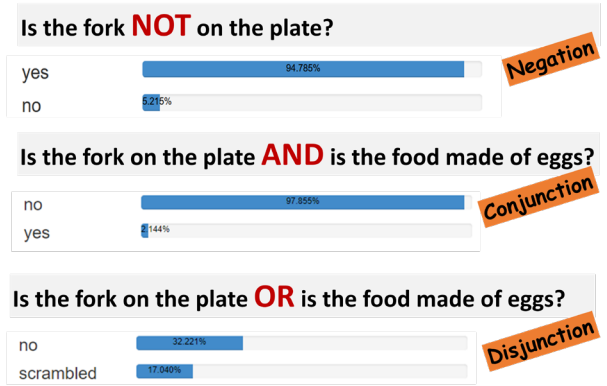
Scene Completion for Missing Sensor/Modality



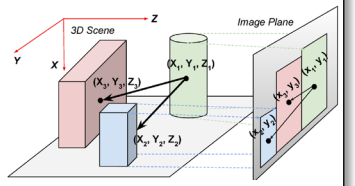
My lab's focus: Perception & Reasoning with Robustness

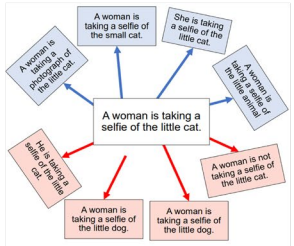
Robust Visual Reasoning (Visual QA, Video Captioning, V&L Inference)

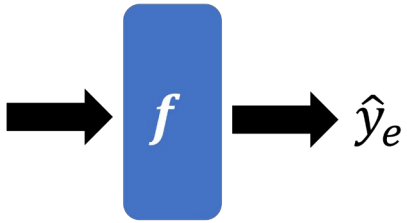
V&L Robustness: Logical, Semantic, Spatial (use additional knowledge sources and sensors)



Question	Answer
Is that a giraffe or an elephant?	Giraffe
Who is feeding the giraffe *behind* the man?	Lady
Is there a fence *near* the animal *behind* the man?	Yes
On which side of the image is the man?	*Right*
Is the giraffe *behind* the man?	Yes

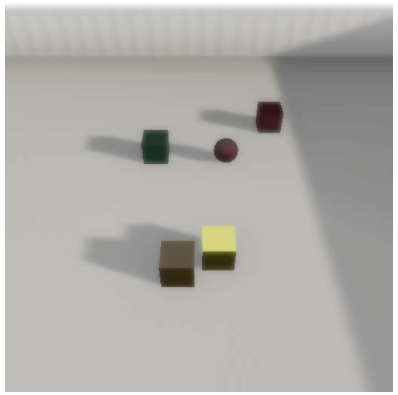






$\hat{y}_e$

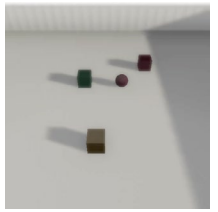
Understanding Agent Actions in Videos with Commonsense, Counterfactual and Physics-Based Reasoning



Counterfactual Question

What will happen if the yellow cube is **removed** ?

(A) Purple Cube will collide with brown cube



Planning Question

How can the collision between yellow and purple cube be stopped?

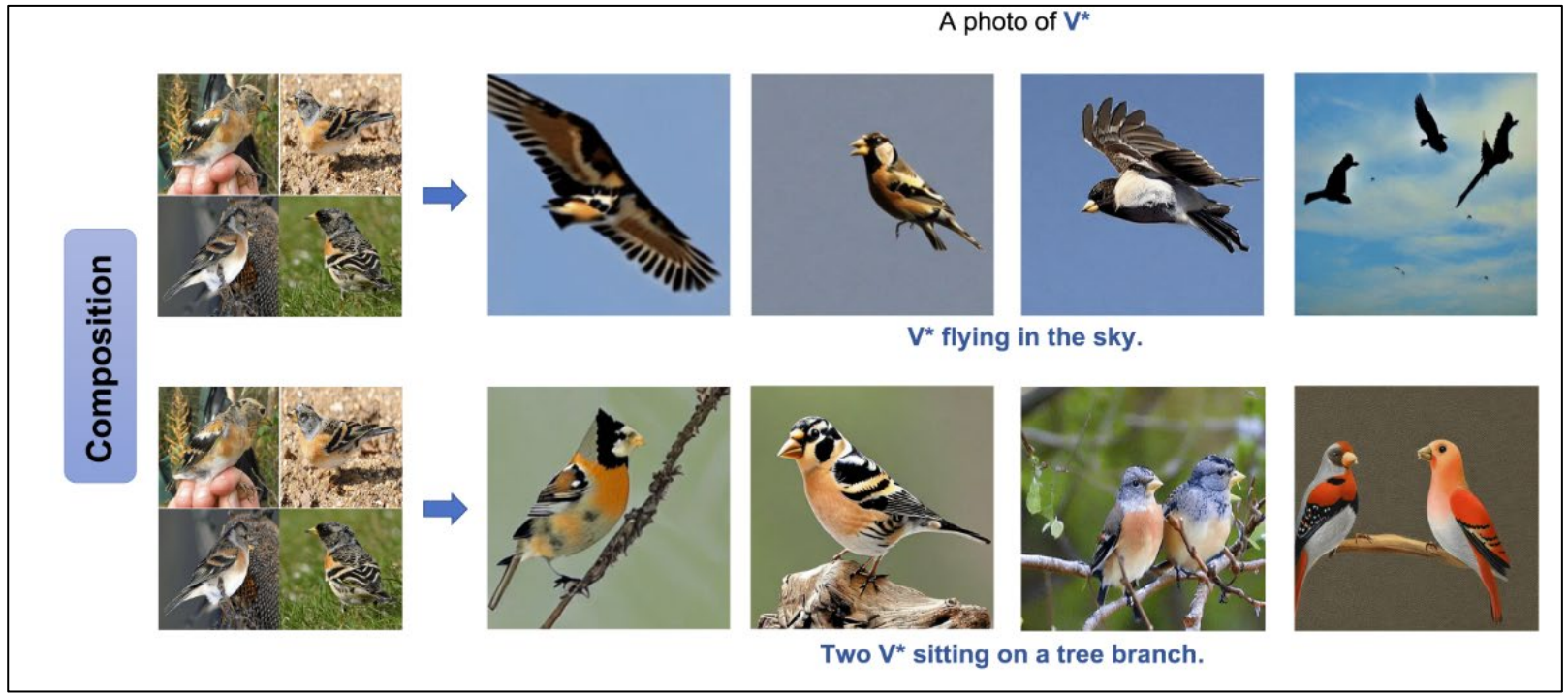
(A) **Add** teal sphere to the right of purple sphere

Conventional Caption	Group of runners get prepared to run a race.
Commonsense-Enriched Caption	In order to win a medal, a group of runners get prepared to run a race. As a result they are congratulated at the finish line. They are athletic.
Commonsense Question Answering	What happens next to the runners? { Are congratulated at the finish line / become tired

Gokhale ECCV '20; Gokhale EMNLP'20; Gokhale ACL'21; Fang EMNLP'20; Banerjee ICCV'21; Patel EMNLP'22

Novel Vision+Language Concept Description

- OOD detection: detect novel (unseen / unknown) objects in videos
- Few-Shot Concept Learning
  - learn that concept
  - assign semantic meaning (in latent space)
  - Reproduce the concept (novel view synthesis)



Patel et al. AAAI 2024

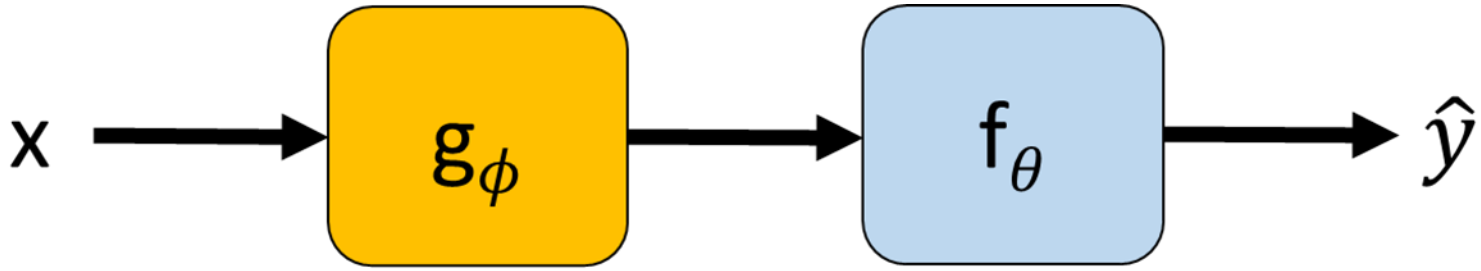
Key Capabilities relevant to Video-LINCS

Visual Data Engineering & Image Augmentation

Model-in-the-loop, dynamic, learnable & evolving

data augmentation
&
novel view-synthesis

Language-based, attribute-based, knowledge-guided, adversarial augmentation



Learnable Image Transformation

$g()$ *With Parameters* ϕ

Visual Recognition Model

$f()$ *With Parameters* θ

$$p = f(\text{  })$$

$$p_g = f(\text{  })$$

$$p_r = f(\text{  })$$

$$p_{mix} = \frac{p_c + p_g + p_r}{3}$$

$$L_{consistency} = D_{KL}(p_{mix}|p) + D_{KL}(p_{mix}|p_g) + D_{KL}(p_{mix}|p_r)$$

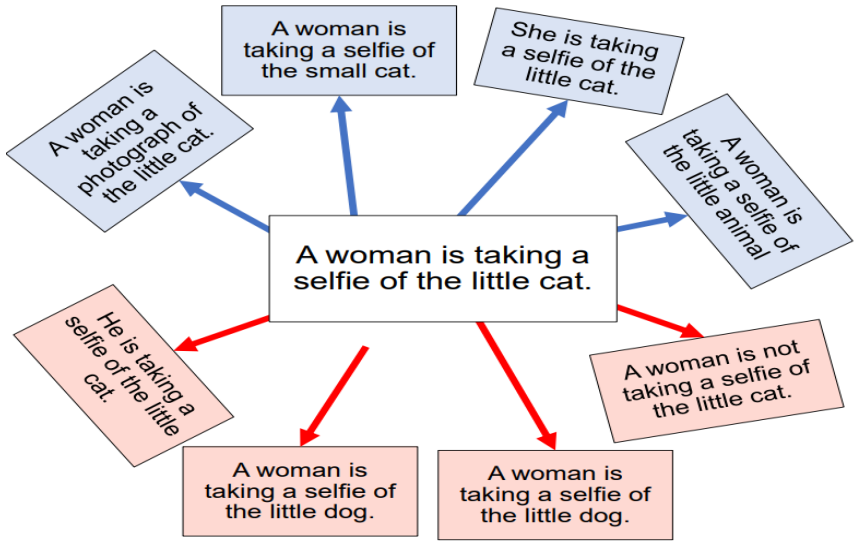
Key Capabilities relevant to Video-LINCS

Natural Language as a Visual "Sensor"

Humans (ordinary/domain-expert) describe visual scenes in natural language
(e.g. English, Hindi, Chinese, Arabic)

Vision-Language Alignment helps for reasoning "beyond pixels"

Commonsense inferences crucial when some sensors malfunction/uncertain/compromised



Conventional Caption	Group of runners get prepared to run a race.
Commonsense-Enriched Caption	In order to win a medal, a group of runners get prepared to run a race. As a result they are congratulated at the finish line. They are athletic.
Commonsense Question Answering	What happens next to the runners? { Are congratulated at the finish line / become tired

Key Capabilities relevant to Video-LINCS

Visual Data Engineering & Image Augmentation

Model-in-the-loop, dynamic, learnable & evolving

data augmentation & novel view-synthesis

Language-based, attribute-based, knowledge-guided, adversarial augmentation

Natural Language as a Visual "Sensor"

Humans (ordinary/domain-expert) describe visual scenes in natural language
(e.g. English, Hindi, Chinese, Arabic)

Vision-Language Alignment helps for reasoning "beyond pixels"

Commonsense inferences crucial when some sensors malfunction/uncertain/compromised

Task Expertise

Image Classification
Object Detection
Vision-Language Alignment
Visual Question Answering
Video Question Answering
Video Captioning
Video-based Reasoning

Robustness Expertise

Domain Adaptation
Few-Shot / Weak Supervision
Adversarial Robustness
OOD Detection