CS 1678/2078: Intro to Deep Learning

Bias, domain shifts, attacks

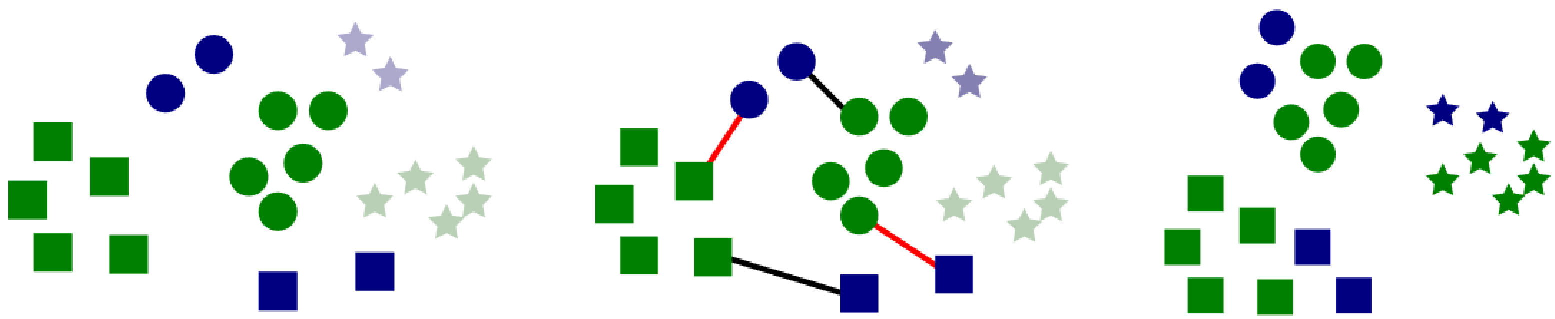
Prof. Adriana Kovashka
University of Pittsburgh
April 15, 2024

Plan for this lecture

- **Domain shifts due to visual style/appearance**
- **Domain shifts due to geography**
- Models inheriting social biases
- Deep fakes and adversarial perturbations

Domain Shifts

Colors = domains, shapes = classes

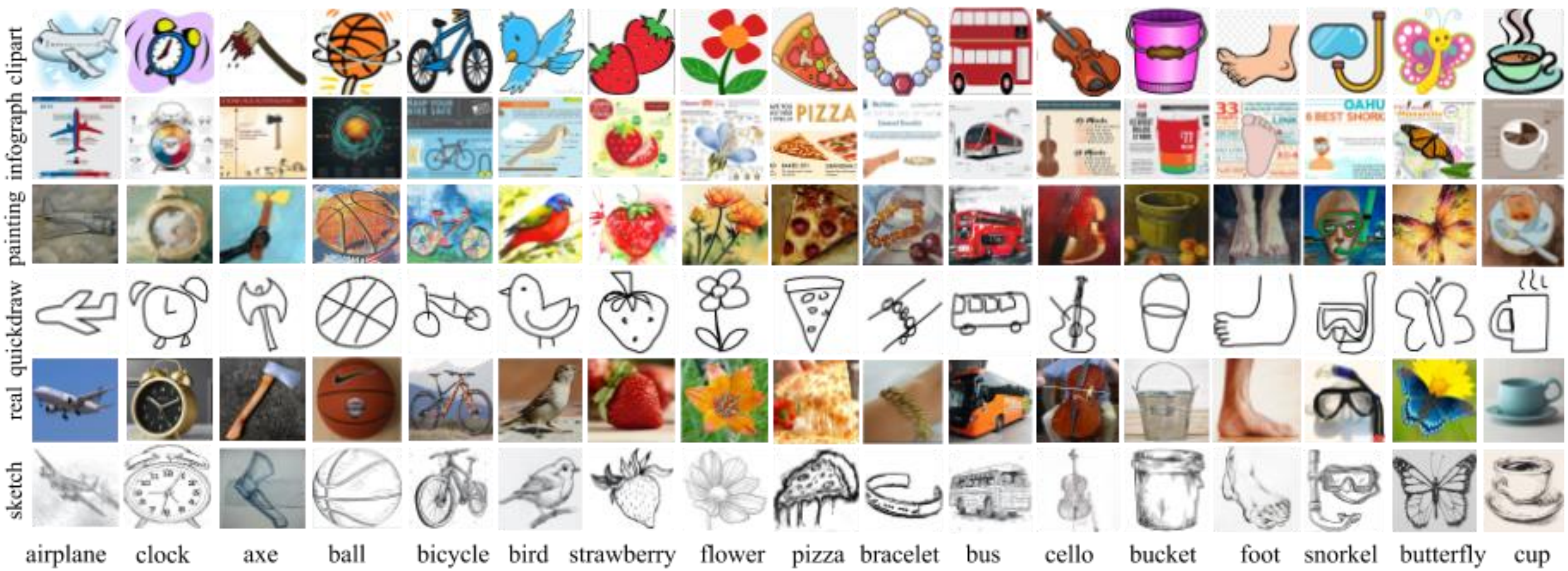


(a) Domain shift problem (b) Pairwise constraints (c) Invariant space

Fig. 2. The key idea of our approach to domain adaptation is to learn a transformation that compensates for the domain-induced changes. By leveraging (dis)similarity constraints (b) we aim to reunite samples from two different domains (blue and green) in a common invariant space (c) in order to learn and classify new samples more effectively across domains. The transformation can also be applied to new categories (lightly-shaded stars). This figure is best viewed in color.

Saenko et al., *Adapting Visual Category Models to New Domains*, ECCV 2010

DomainNet Dataset



Peng et al. *Moment matching for multi-source domain adaptation.* ICCV 2019.

Coping with Domain Shifts

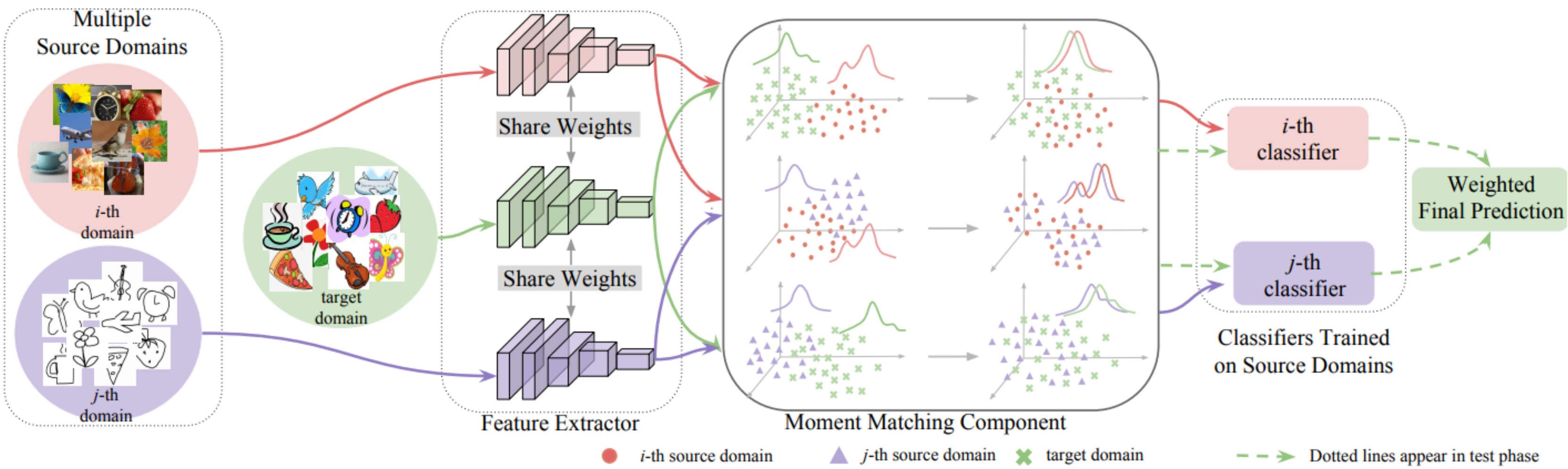


Figure 2. The framework of **Moment Matching for Multi-source Domain Adaptation** (M^3SDA). Our model consists of three components: i) feature extractor, ii) moment matching component, and iii) classifiers. Our model takes multi-source annotated training data as input and transfers the learned knowledge to classify the unlabeled target samples. Without loss of generality, we show the *i*-th domain and *j*-th domain as an example. The feature extractor maps the source domains into a common feature space. The moment matching component attempts to match the *i*-th and *j*-th domains with the target domain, as well as matching the *i*-th domain with the *j*-th domain. The final predictions of target samples are based on the weighted outputs of the *i*-th and *j*-th classifiers. (Best viewed in color!)

Peng et al. *Moment matching for multi-source domain adaptation.* ICCV 2019.

Domain Adversarial Networks

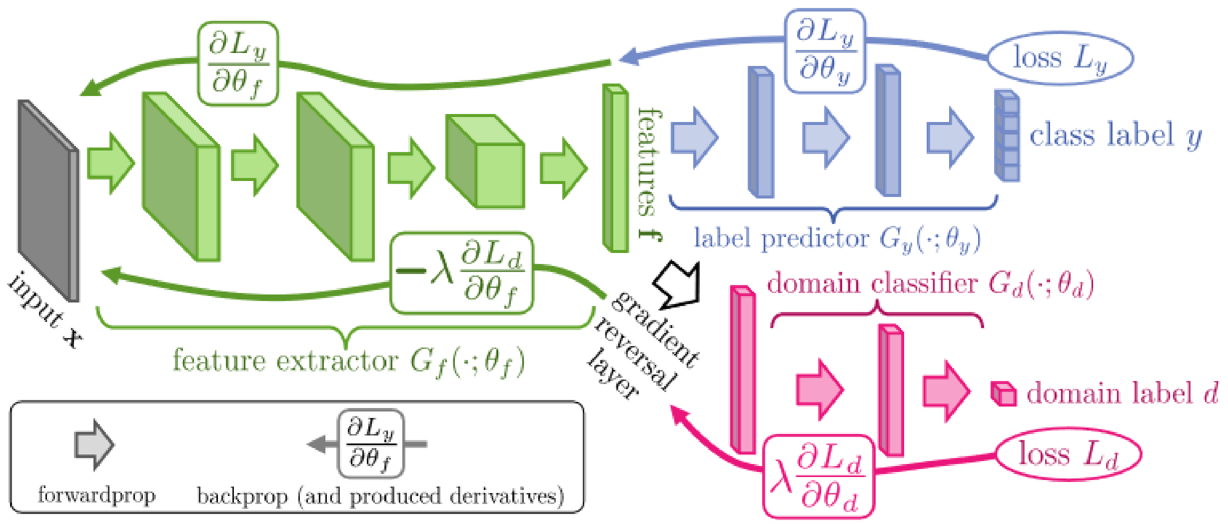
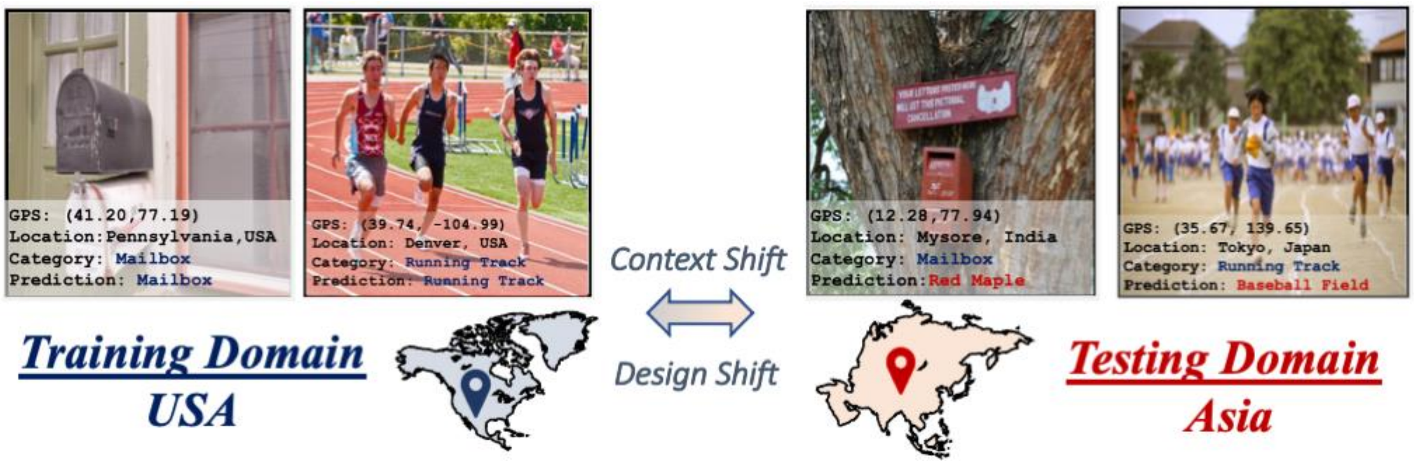


Figure 1: The **proposed architecture** includes a deep *feature extractor* (green) and a deep *label predictor* (blue), which together form a standard feed-forward architecture. Unsupervised domain adaptation is achieved by adding a *domain classifier* (red) connected to the feature extractor via a *gradient reversal layer* that multiplies the gradient by a certain negative constant during the backpropagation-based training. Otherwise, the training proceeds standardly and minimizes the label prediction loss (for source examples) and the domain classification loss (for all samples). Gradient reversal ensures that the feature distributions over the two domains are made similar (as indistinguishable as possible for the domain classifier), thus resulting in the domain-invariant features.

Ganin and Lempitsky. *Unsupervised domain adaptation by backpropagation.* ICML 2015.

Geographic Domain Shifts



"While modern computer vision models yield human-level accuracies when trained and tested on the images from the same geographical domain, the accuracy drops significantly when presented with images from different geographies. Here, images belonging to mailbox and running track are misclassified due to design and context shifts between the domains induced by disparate geographies."

Kalluri et al. *GeoNet: Benchmarking Unsupervised Adaptation across Geographies*. CVPR 2023.

Geographic Domain Shifts

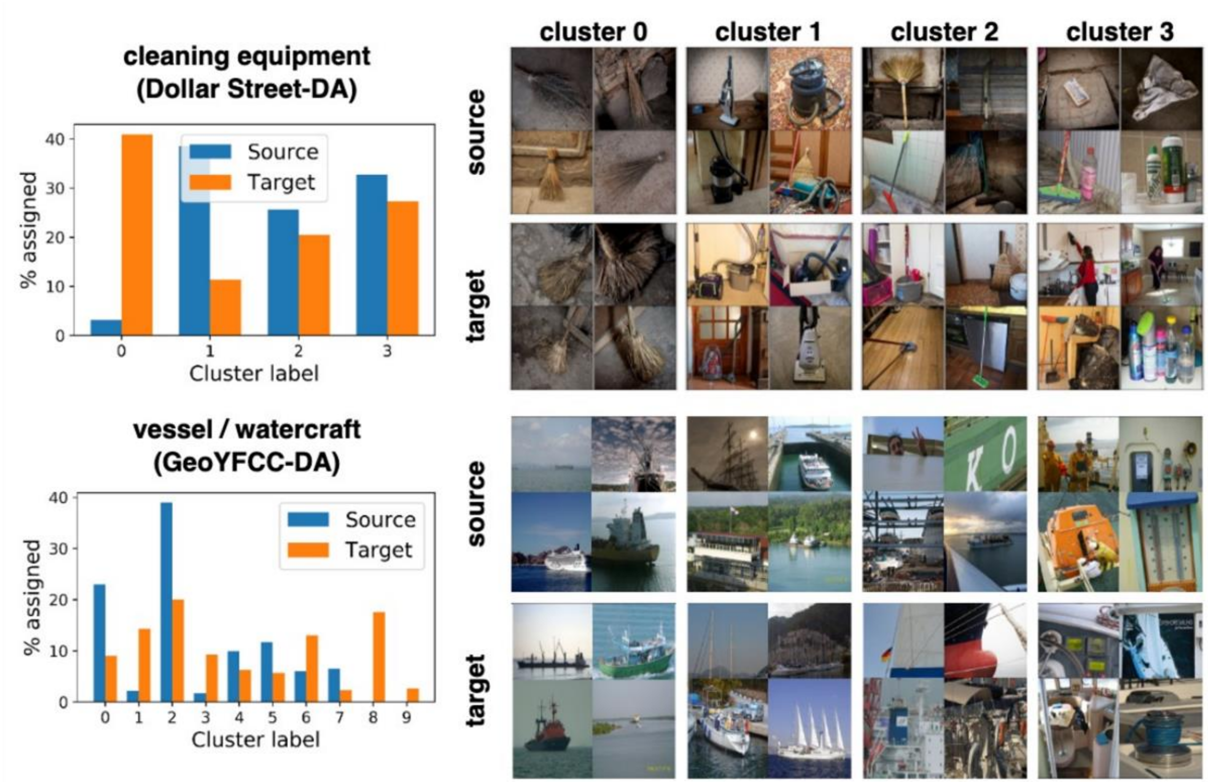
Context Shift

Task-irrelevant information (e.g., background or surroundings)

Prabhu et al. *Can domain adaptation make object recognition work for everyone?* CVPRW 2022.

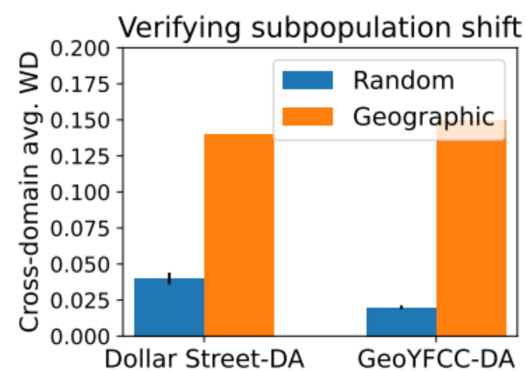
Slide: Marcelo d'Almeida

Geographic Domain Shifts



Subpopulation Shift

Change within category
(e.g., "cleaning equipment"
can be brooms, mops,
vacuum cleaners, etc.)



Prabhu et al. *Can domain adaptation make object recognition work for everyone?* CVPRW 2022.

Slide: Marcelo d'Almeida

Geographic Domain Shifts

Dollar Street-DA

60 countries
(66 in total)

58 categories
(128 in total)

Prabhu et al. *Can domain adaptation make object recognition work for everyone?* CVPRW 2022.

Slide: Marcelo d'Almeida

Geographic Domain Shifts

GeoYFCC-DA

62 countries 68 categories

Prabhu et al. *Can domain adaptation make object recognition work for everyone?* CVPRW 2022.

Slide: Marcelo d'Almeida

Geographic Domain Shifts

Results

Significant performance drops from geographical shifts

Limited improvements from existing DA methods

Method	Dollar Street-DA	GeoYFCC-DA
source	54.66±0.62	42.88
target oracle*	67.73± 0.30	56.78
MMD [10]	55.77±0.75	43.53
DANN [4]	54.80±0.38	42.64
SENTRY [5]	55.73±0.34	42.58
SST	58.71±0.53	45.22

*denotes that the target oracle was trained on target data non-overlapping with the test set (80%) whereas DA methods were adapted without labels on the entire target dataset.

Prabhu et al. *Can domain adaptation make object recognition work for everyone?* CVPRW 2022.

Slide: Marcelo d'Almeida

Incorporating Geo-Diverse Knowledge into Prompting for Increased Geographical Robustness in Object Recognition

Kyle Buettner[1], Sina Malakouti[2], Xiang Lorraine Li[1,2], Adriana Kovashka[1,2]
[1]Intelligent Systems Program, [2]Department of Computer Science, University of Pittsburgh, PA, USA
{buettnerk, sem238}@pitt.edu, {xianglli, kovashka}@cs.pitt.edu

Abstract

Existing object recognition models have been shown to lack robustness in diverse geographical scenarios due to significant domain shifts in design and context. Class representations need to be adapted to more accurately reflect an object concept under these shifts. In the absence of training data from target geographies, we hypothesize that geography-specific descriptive knowledge of object categories can be leveraged to enhance robustness. For this purpose, we explore the feasibility of probing a large-language model for geography-specific object knowledge, and we investigate integrating knowledge in zero-shot and learnable soft prompting with the CLIP vision-language model. In particular, we propose a geography knowledge regularization method to ensure that soft prompts trained on a source set of geographies generalize to an unseen target set of ge-

Figure 1. **Descriptive knowledge can bridge concept shifts across geographies.** Observe the wide range of object designs and contexts in the DollarStreet [11] category *tools* around the world. Our work's premise is that textual representations for classes in vision-language models can be tailored to better suit diverse object representations across geographies. Map made with [16].

Buettner et al., CVPR 2024

Geographic Domain Shifts

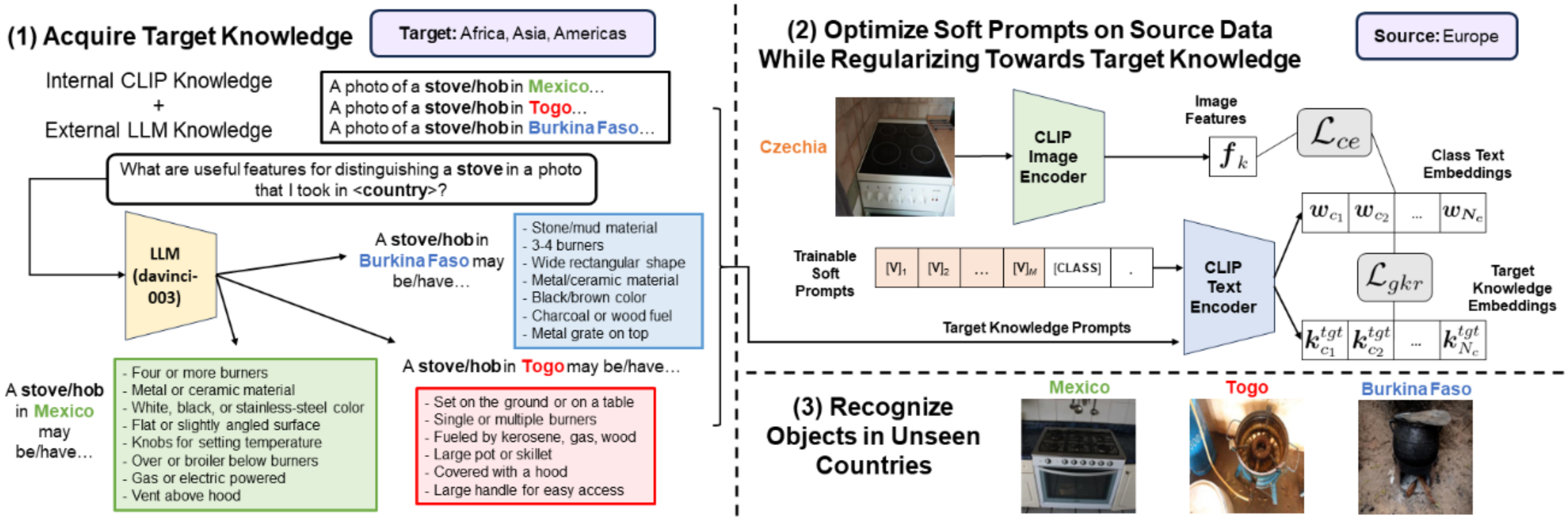


Figure 2. **Geography knowledge regularization.** To ensure robustness in soft prompt learning, we (1) incorporate complementary knowledge internal to CLIP and externally obtained from an LLM. (2) This descriptive knowledge regularizes class representations when training on a specific source geography (*e.g.* Europe), thus (3) increasing robustness when generalizing to unseen geographies (*e.g.* Togo).

Buettner et al., CVPR 2024

Geographic Domain Shifts

Encoder	Prompting Method	*Top-1 Accuracy*										*Top-3 Accuracy*									
		Europe		**Africa**		**Asia**		**Americas**		**Total**		**Europe**		**Africa**		**Asia**		**Americas**		**Total**	
		Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ
ViT-B/32	Zero-Shot CLIP [36]	59.1	-	43.7	-	50.8	-	**55.3**	-	51.7	-	81.1	-	64.8	-	72.3	-	**77.4**	-	73.7	-
	GeneralLLM [30]	57.3	-1.8	44.3	+0.6	50.9	+0.1	54.6	-0.7	51.4	-0.3	78.8	-2.3	64.5	-0.3	72.1	-0.2	75.7	-1.7	73.0	-0.7
	CountryInPrompt	57.5	-1.6	45.2	+1.5	51.9	+1.1	55.0	-0.3	52.1	+0.4	80.2	-0.9	65.5	+0.7	73.3	+1.0	76.9	-0.5	73.9	+0.2
	CountryLLM	59.4	+0.3	45.2	+1.5	52.1	+1.3	**55.3**	0.0	52.6	+0.9	80.9	-0.2	66.4	+1.6	**73.6**	+1.3	**77.4**	0.0	74.6	+0.9
	CountryInPrompt+LLM	**60.8**	+1.7	**45.3**	+1.6	**52.2**	+1.4	55.0	-0.3	**52.8**	+1.1	**81.5**	+0.4	**67.4**	+2.6	**73.6**	+1.3	76.7	-0.7	**74.7**	+1.0
ViT-B/16	Zero-Shot CLIP [36]	64.3	-	46.9	-	53.9	-	**60.1**	-	55.5	-	84.3	-	69.3	-	75.9	-	81.1	-	77.2	-
	GeneralLLM [30]	64.2	-0.1	48.8	+1.9	**56.0**	+2.1	58.5	-1.6	56.8	+1.3	83.9	-0.4	71.1	+1.8	76.3	+0.4	80.4	-0.7	77.9	+0.7
	CountryInPrompt	63.9	-0.4	49.6	+2.7	55.7	+1.8	59.3	-0.8	56.6	+1.1	84.0	-0.3	71.3	+2.0	76.5	+0.6	80.0	-1.1	77.7	+0.5
	CountryLLM	65.2	+0.9	49.6	+2.7	55.6	+1.7	59.7	-0.4	57.0	+1.5	84.3	0.0	71.8	+2.5	**77.5**	+1.6	**81.5**	+0.4	**78.8**	+1.6
	CountryInPrompt+LLM	**65.5**	+1.2	**50.8**	+3.9	**56.0**	+2.1	59.7	-0.4	**57.4**	+1.9	**85.5**	+1.2	**72.5**	+3.2	77.0	+1.1	80.9	-0.2	78.7	+1.5
RN50	Zero-Shot CLIP [36]	53.0	-	38.0	-	44.4	-	49.8	-	45.7	-	76.5	-	60.2	-	66.4	-	**72.7**	-	68.1	-
	GeneralLLM [30]	55.5	+2.5	40.9	+2.9	46.9	+2.5	50.3	+0.5	47.9	+2.2	76.0	-0.5	61.2	+1.0	67.7	+1.3	71.1	-1.6	68.6	+0.5
	CountryInPrompt	54.5	+1.5	**43.4**	+5.4	47.0	+2.6	50.8	+1.0	48.4	+2.7	76.0	-0.5	**64.0**	+3.8	68.7	+2.3	**72.7**	0.0	**70.0**	+1.9
	CountryLLM	56.2	+3.2	41.1	+3.1	47.3	+2.9	50.4	+0.6	48.3	+2.6	**77.2**	+0.7	62.5	+2.3	**68.8**	+2.4	72.4	-0.3	**70.0**	+1.9
	CountryInPrompt+LLM	**56.4**	+3.4	43.0	+5.0	**48.0**	+3.6	**50.9**	+1.1	**49.1**	+3.4	76.7	+0.2	63.1	+2.9	68.3	+1.9	71.1	-1.6	69.4	+1.3

Table 1. **Zero-shot CLIP with descriptive knowledge prompts, top-1/3 balanced accuracy (Acc) on DollarStreet.** Our strategies to capture CLIP's internal country knowledge (CountryInPrompt), external LLM country knowledge (CountryLLM), and their combination (CountryInPrompt+LLM), improve the zero-shot CLIP baseline (prompt "a photo of a"), especially on Africa (exemplified in light blue) and Asia; gains in green, drops in red. Our strategies also outperform the GeneralLLM [30] baseline.

Buettner et al., CVPR 2024

Plan for this lecture

- Domain shifts due to visual style/appearance
- Domain shifts due to geography
- **Models inheriting social biases**
- **Deep fakes and adversarial perturbations**

Bias in Language

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Figure 1: The most extreme occupations as projected on to the *she*−*he* gender direction on g2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded.

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Figure 2: **Analogy examples**. Examples of automatically generated analogies for the pair *she-he* using the procedure described in text. For example, the first analogy is interpreted as *she*:*sewing* :: *he*:*carpentry* in the original w2vNEWS embedding. Each automatically generated analogy is evaluated by 10 crowd-workers are to whether or not it reflects gender stereotype. Top: illustrative gender stereotypic analogies automatically generated from w2vNEWS, as rated by at least 5 of the 10 crowd-workers. Bottom: illustrative generated gender-appropriate analogies.

Bolukbasi et al., *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, NIPS 2016

Bias in Language

he (158)

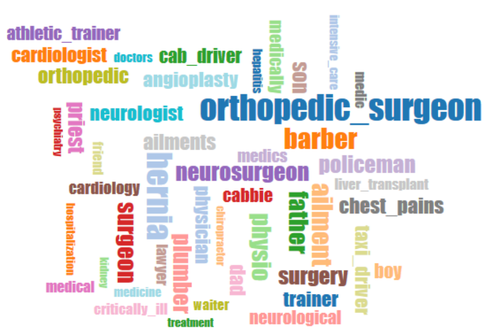
she (42)

Adjectives

Or type your own words...

doctor

he (47)



she (153)

http://wordbias.umiacs.umd.edu/

Bias in Vision

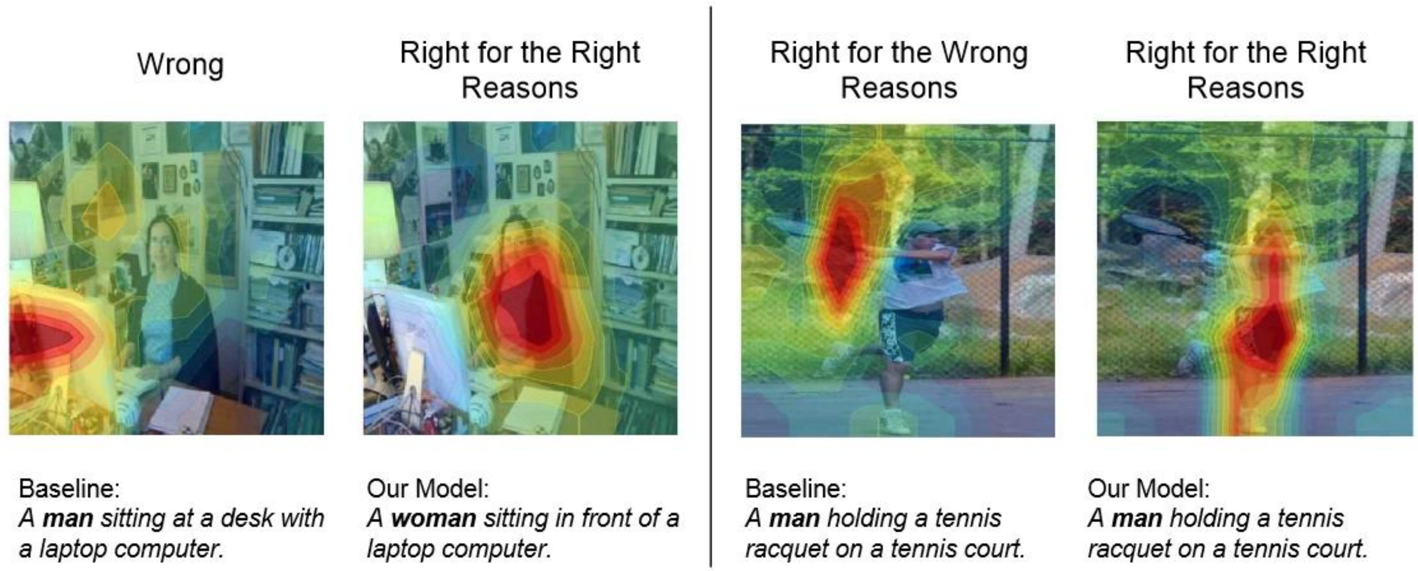


Fig. 1: Examples where our proposed model (Equalizer) corrects bias in image captions. The overlaid heatmap indicates which image regions are most important for predicting the gender word. On the left, the baseline predicts gender incorrectly, presumably because it looks at the laptop (not the person). On the right, the baseline predicts the gender correctly but it does not look at the person when predicting gender and is thus not acceptable. In contrast, our model predicts the correct gender word and correctly considers the person when predicting gender.

Burns et al., *Women also Snowboard: Overcoming Bias in Captioning Models*, ECCV 2018

Human Reporting Bias

The **frequency** with which **people write** about actions, outcomes, or properties is **not a reflection of real-world frequencies** or the degree to which a property is characteristic of a class of individuals

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
- Bunches of bananas with stickers on them on shelves in a store

...We don’t tend to say
Yellow Bananas

What do you see?

Green Bananas

Unripe Bananas

What do you see?

Ripe Bananas

Bananas with spots

Bananas good for banana bread

What do you see?

Yellow Bananas?

Yellow is prototypical for bananas

Prototype Theory

One purpose of categorization is to **reduce the infinite differences** among stimuli **to** behaviourally and **cognitively usable proportions**

There may be some central, prototypical notions of items that arise from stored typical properties for an object category (Rosch, 1975)

May also store exemplars (Wu & Barsalou, 2009)

Fruit

Bananas
“Basic Level”

Unripe Bananas,
Cavendish Bananas

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

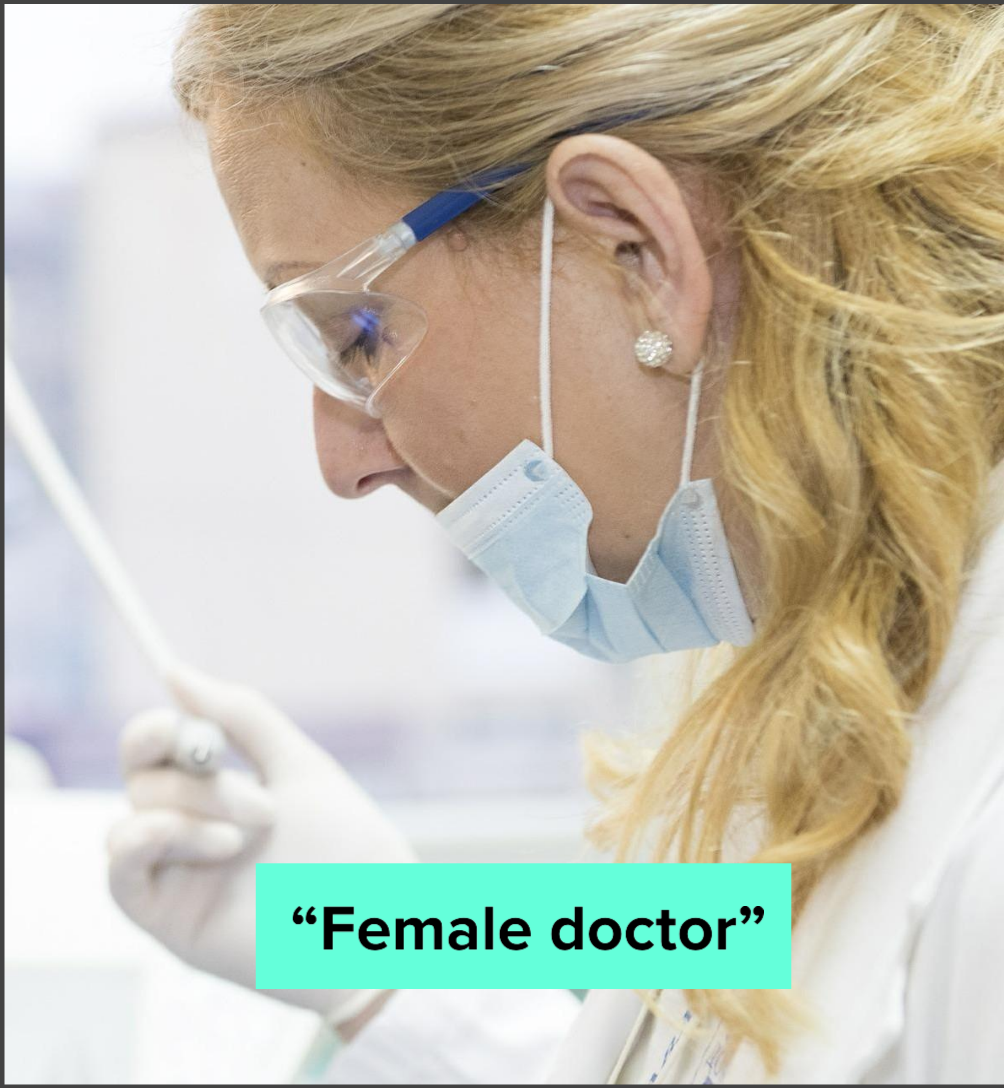
The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

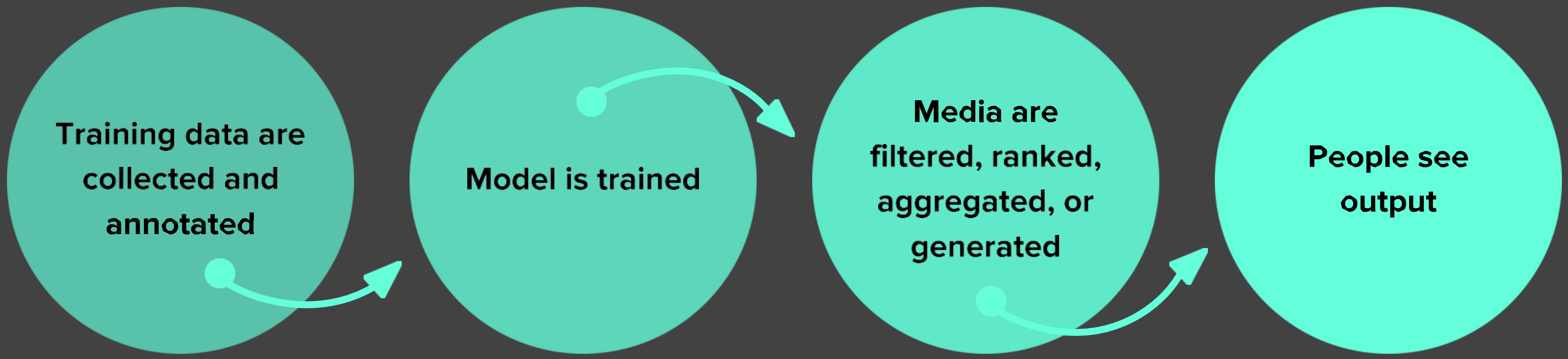
The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?




“Doctor”


“Female doctor”



Margaret Mitchell

Biases in Data

Selection Bias: Selection does not reflect a random sample

CREDIT

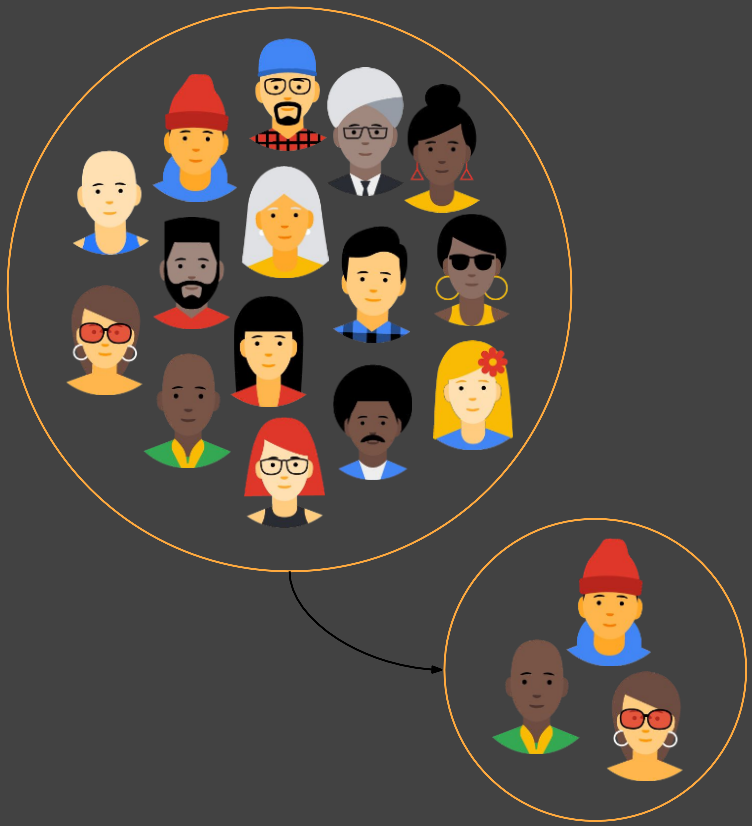
[© 2013–2016 Michael Yoshitaka Erlewine and Hadas Kotek]

Biases in Data

Out-group homogeneity bias: Tendency to see outgroup members as more alike than ingroup members

Biases in Data →
Biased Data Representation

It's possible that you have an appropriate amount of data for every group you can think of but that some groups are represented less positively than others.



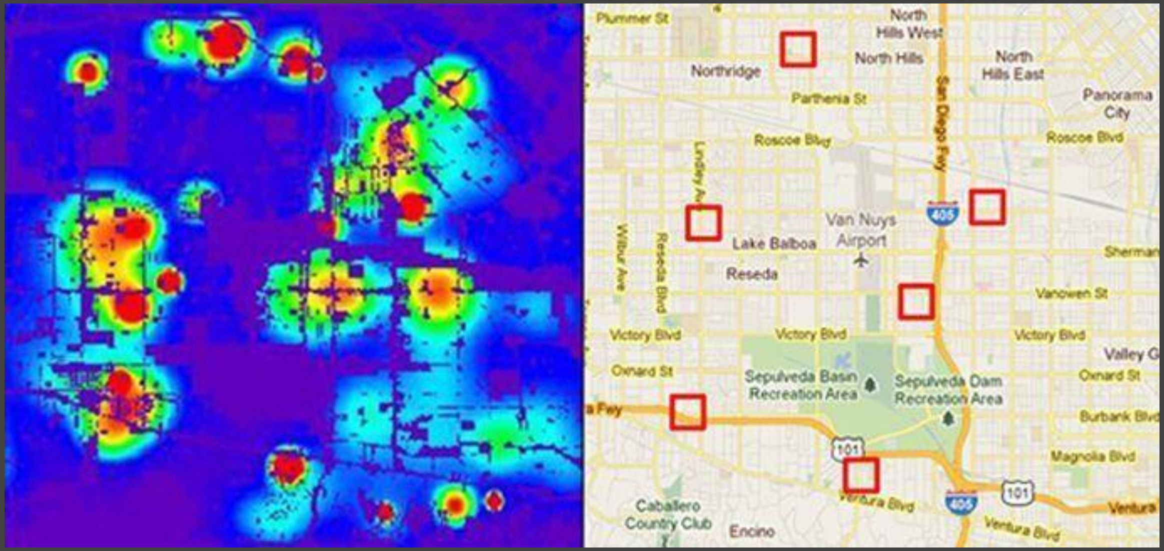
Biases in Data → Biased Labels

Annotations in your dataset will reflect the worldviews of your annotators.

https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

Predicting Policing

- Algorithms identify potential crime hot-spots
- Based on where crime is previously reported, not where it is known to have occurred
- Predicts future events from past



CREDIT

[Smithsonian. *Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?* 2018]

Predicting Sentencing

- Prater (who is white) rated **low risk** after shoplifting, despite two armed robberies; one attempted armed robbery.
- Borden (who is black) rated **high risk** after she and a friend took (but returned before police arrived) a bike and scooter sitting outside.
- Two years later, Borden has not been charged with any new crimes. Prater serving 8-year prison term for grand theft.

CREDIT

ProPublica. Northpointe: Risk in Criminal Sentencing. 2016.

Predicting Criminality

Israeli startup, Faception

> *"Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for profiling people and **revealing their personality based only on their facial image**."*

Offering specialized engines for recognizing "High IQ", "White-Collar Offender", "Pedophile", and "Terrorist" from a face image.

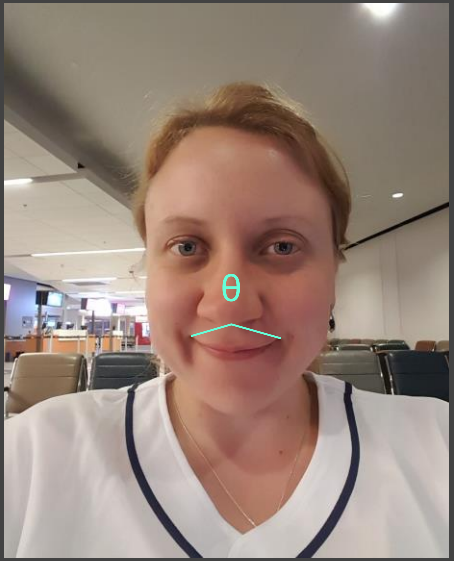
Main clients are in homeland security and public safety.

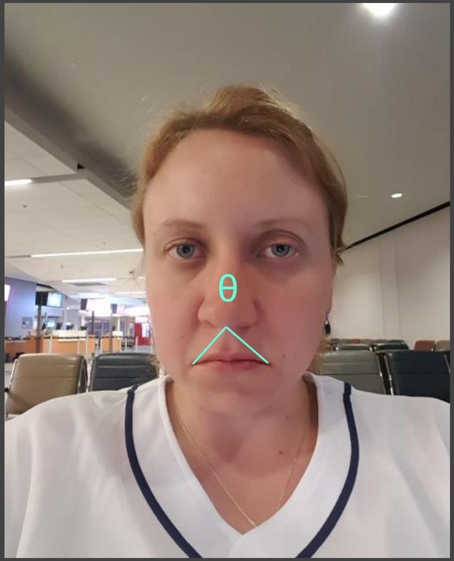
Predicting Criminality

"Automated Inference on Criminality using Face Images" Wu and Zhang, 2016. arXiv

1,856 closely cropped images of faces; Includes "wanted suspect" ID pictures from specific regions.

"[…] angle θ from nose tip to two mouth corners is on average 19.6% smaller for criminals than for non-criminals …"





See our longer piece on Medium, "Physiognomy's New Clothes"

“Deepfakes”

https://www.technologyreview.com/s/611726/the-defense-department-has-produced-the-first-tools-for-catching-deepfakes/
https://www.niemanlab.org/2018/11/how-the-wall-street-journal-is-preparing-its-journalists-to-detect-deepfakes/

Expected Threats

Targeted Personal Attacks

Peele 2017

AI Multimedia Algorithms

⇩

Highly realistic video

Generated Events at Scale

AI Multimedia Algorithms

⇩

1000s +

On a rainy spring day, a vast, violent group gathered in front of the US Capitol to protest recent cuts in Social Security.

Text

Video & Audio

Image

Believable fake events

Ransomfake concept: Identity Attacks as a service (IAaaS)

Bricman 2019

AI Multimedia Algorithms

⇩

Forged Evidence

⇩

Identity Attacks

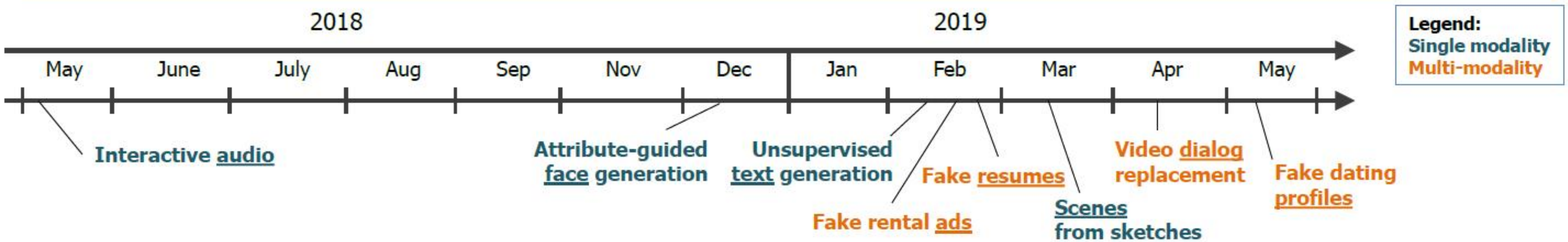
Examples of possible fakes:

- Substance abuse
- Foreign contacts
- Compromising events
- Social media postings
- Financial inconsistencies
- Forging identity

Undermines key individuals and organizations

Matt Turek

Incredible Pace of Synthetic Media Generation



ENTIRE GUEST SUITE

Luxury Condo 3 Bed + 3 Bath
Port Melbourne

○ 8 guests ○ 3 bedrooms ○ 4beds ○ 2 baths

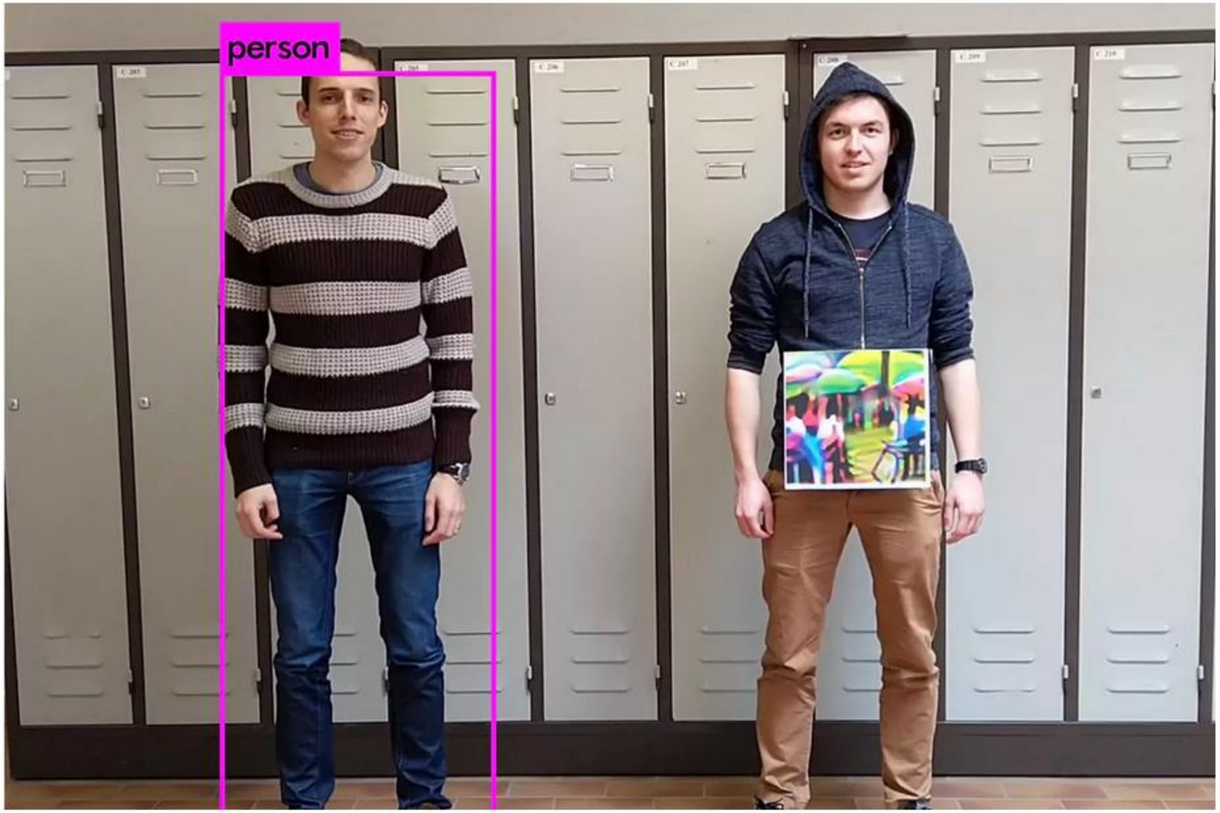
Bathroom (with seating for 2 more people), basin and eclectic French garden and kitchen. 24/7 carpeted charc. Laundrymemberly : More balcony – Garden – Metro, Liverpool Street (15 min walk) Walking distance to Wyckofferdon

Matt Turek

Adversarial Attacks

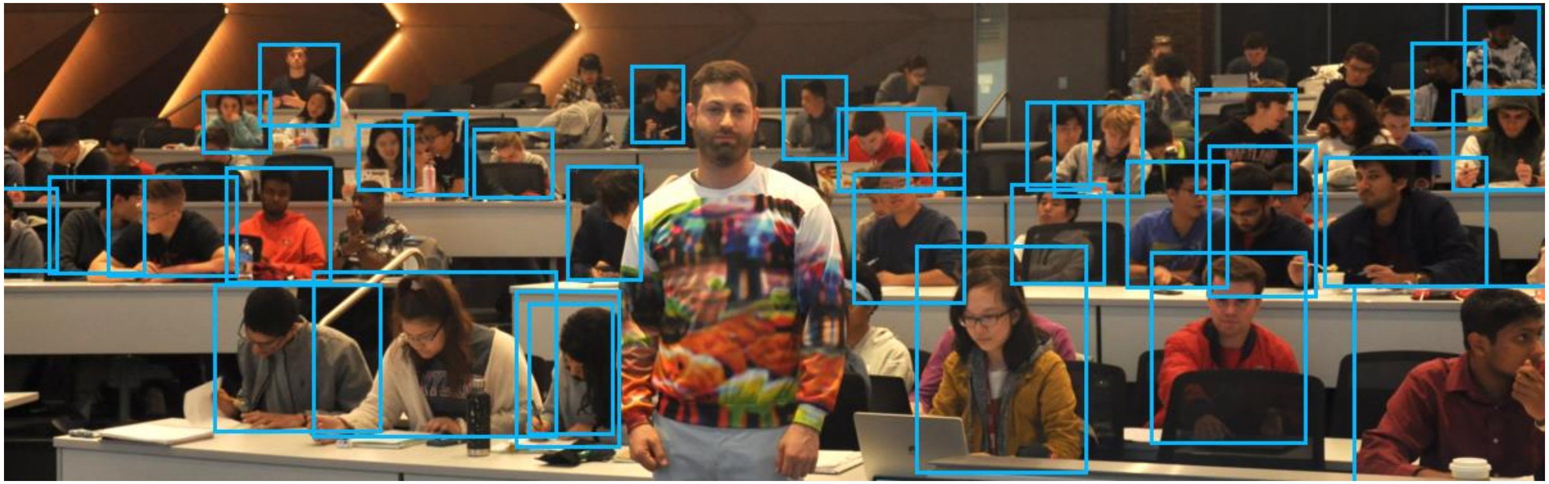
https://bair.berkeley.edu/blog/2017/12/30/yolo-attack/

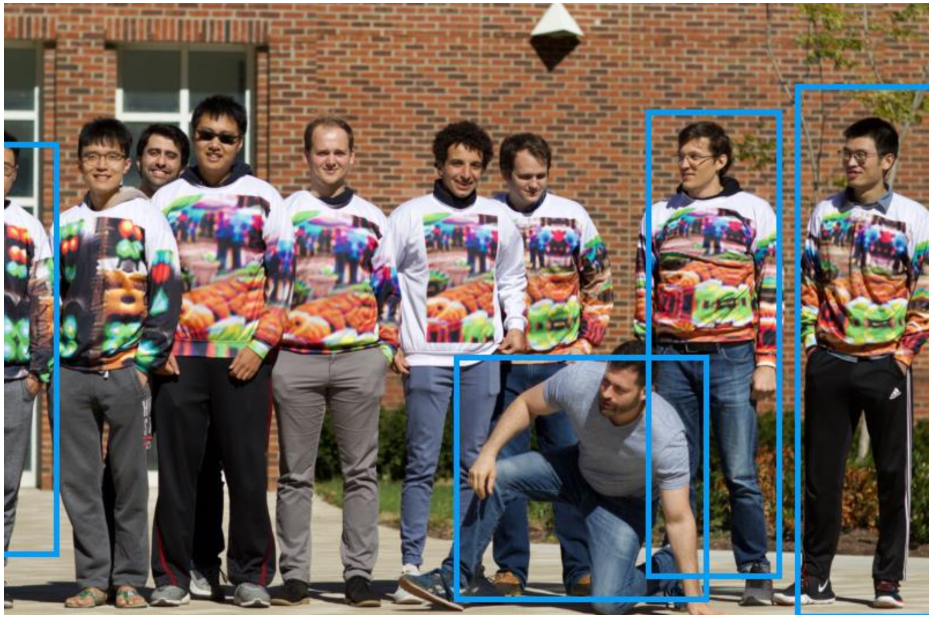
Adversarial Attacks



https://www.theverge.com/2019/4/23/18512472/fool-ai-surveillance-adversarial-example-yolov2-person-detection

Adversarial Attacks





Tom Goldstein https://www.cs.umd.edu/~tomg/projects/invisible/

Adversarial Attacks

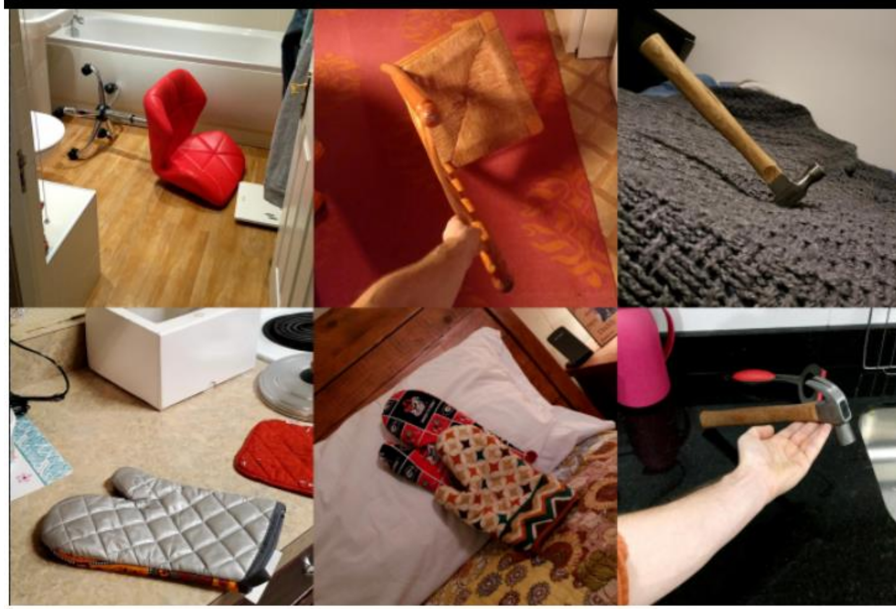
Massachusetts Institute of Technology

NEWS | VIDEO | SOCIAL | FOLLOW MIT

MIT News

ON CAMPUS AND AROUND THE WORLD

Browse or Search



ObjectNet, a dataset of photos created by MIT and IBM researchers, shows objects from odd angles, in multiple orientations, and against varied backgrounds to better represent the complexity of 3D objects. The researchers hope the dataset will lead to new computer vision techniques that perform better in real life.

Photo collage courtesy of the researchers.

This object-recognition dataset stumped the world's best computer vision models

Objects are posed in varied positions and shot at odd angles to spur new AI techniques.

http://news.mit.edu/2019/object-recognition-dataset-stumped-worlds-best-computer-vision-models-1210