Counterfactual Analysis：
A Great Tool for unbiased learning

Guangyi Chen

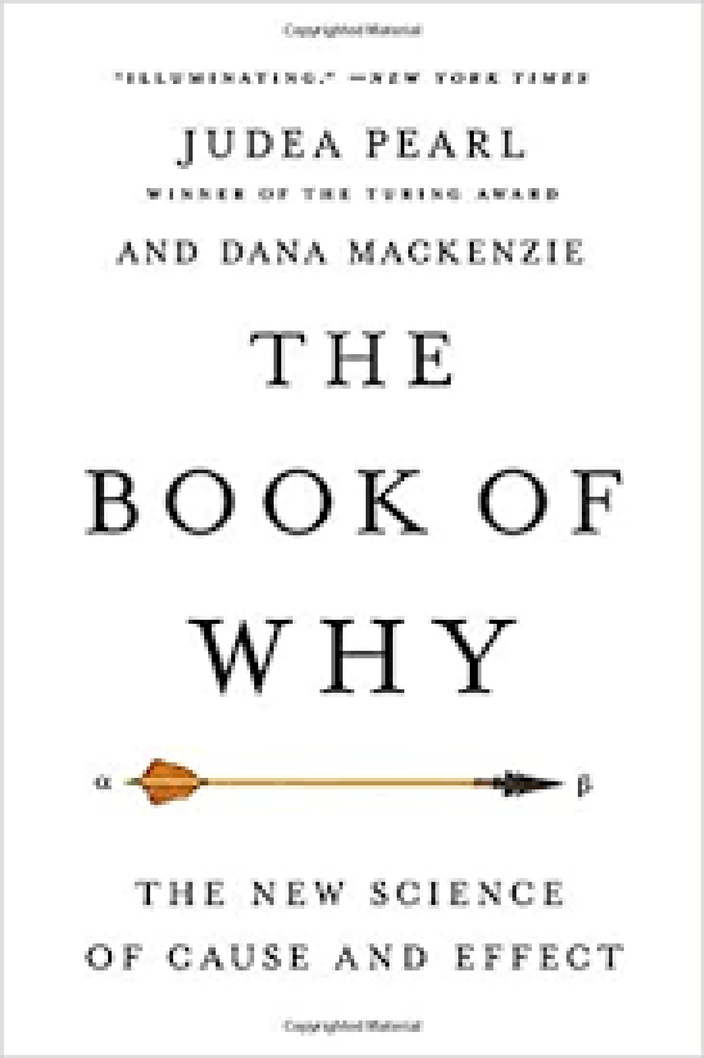
MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

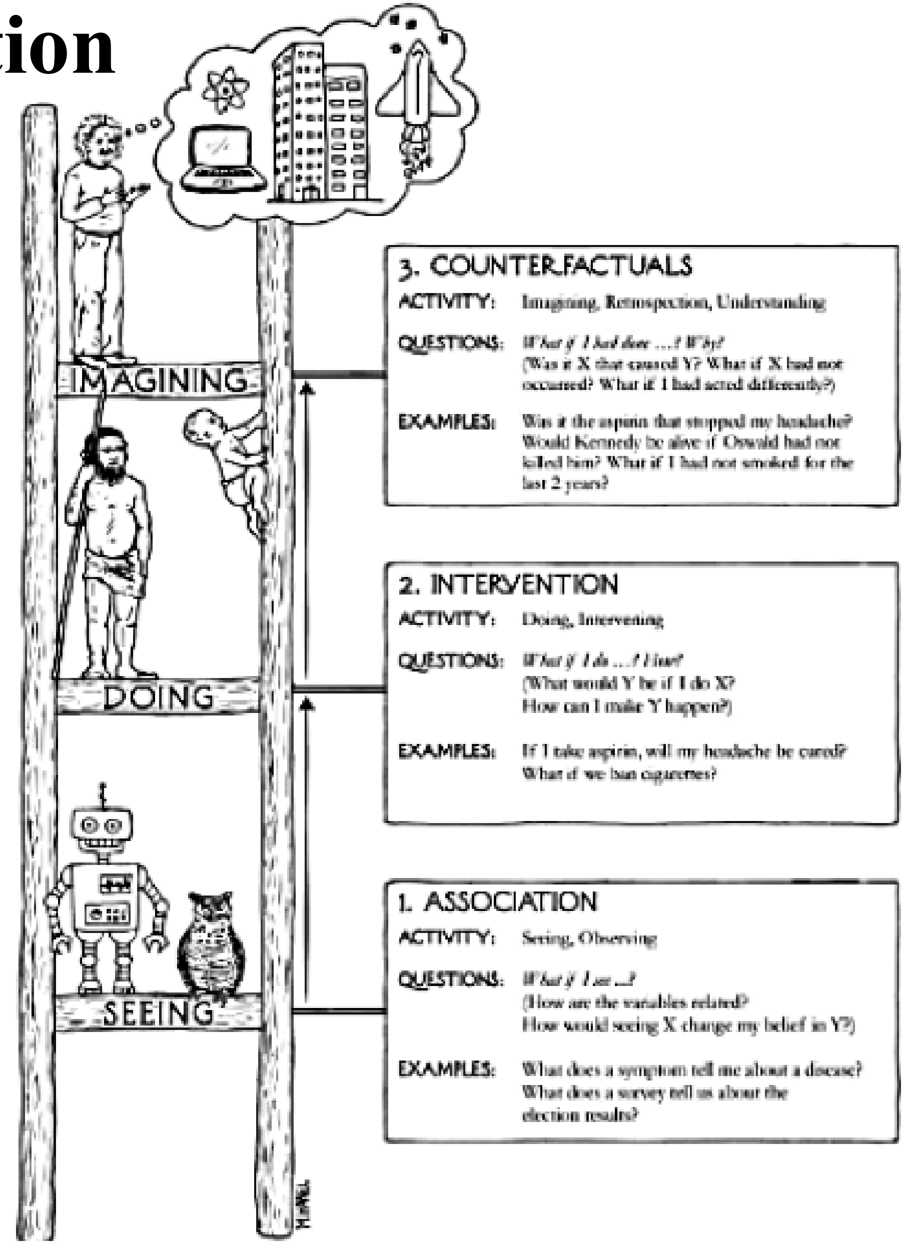
Carnegie
Mellon
University

Contents

Contents

The Ladder of Causation





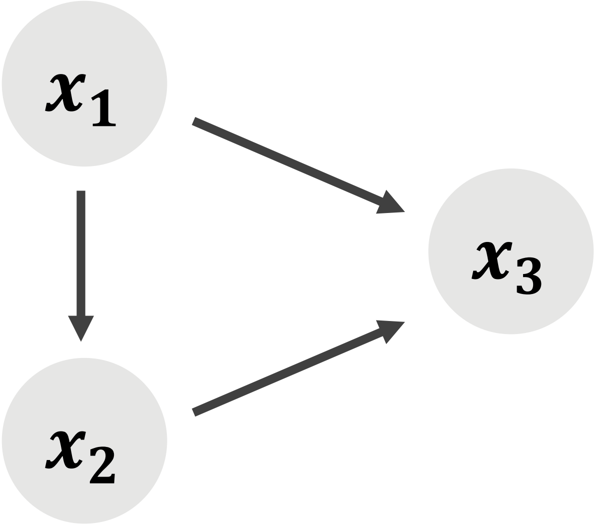
Counterfactuals: imagining
What if I had not done …

Intervention： do actions
What if I do …

Association: likelihood
What if I see …

Pearl J, Mackenzie D. The book of why: the new science of cause and effect, 2018.

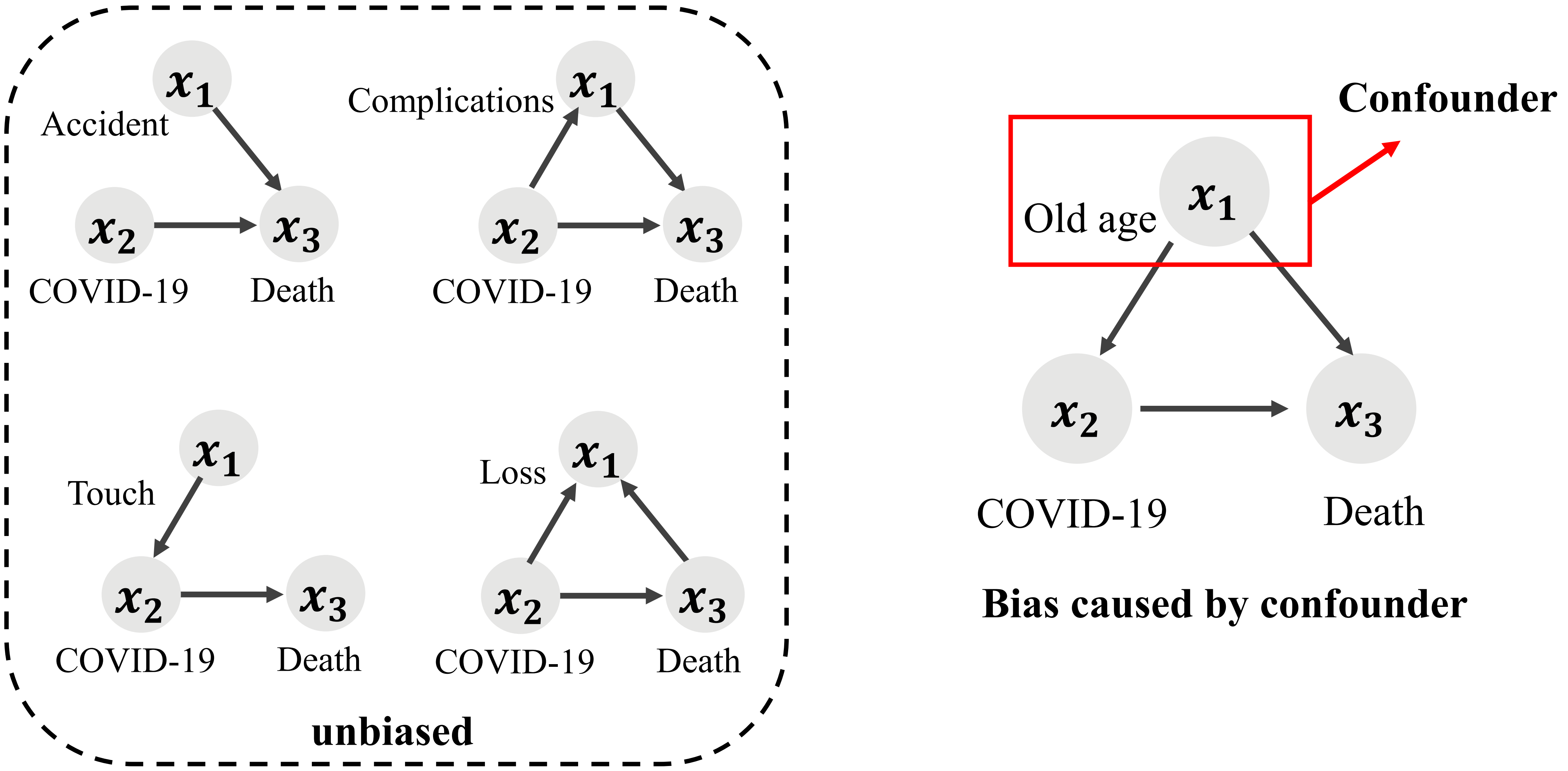
Causal Graph



- Causal Graph is a directed acyclic graph
- Node: Variables x_1, x_2, x_3
- Edge: The link from causal to effect $x_1 \to x_3$
- Path: From one variable to another $x_1 \to x_2 \to x_3$

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

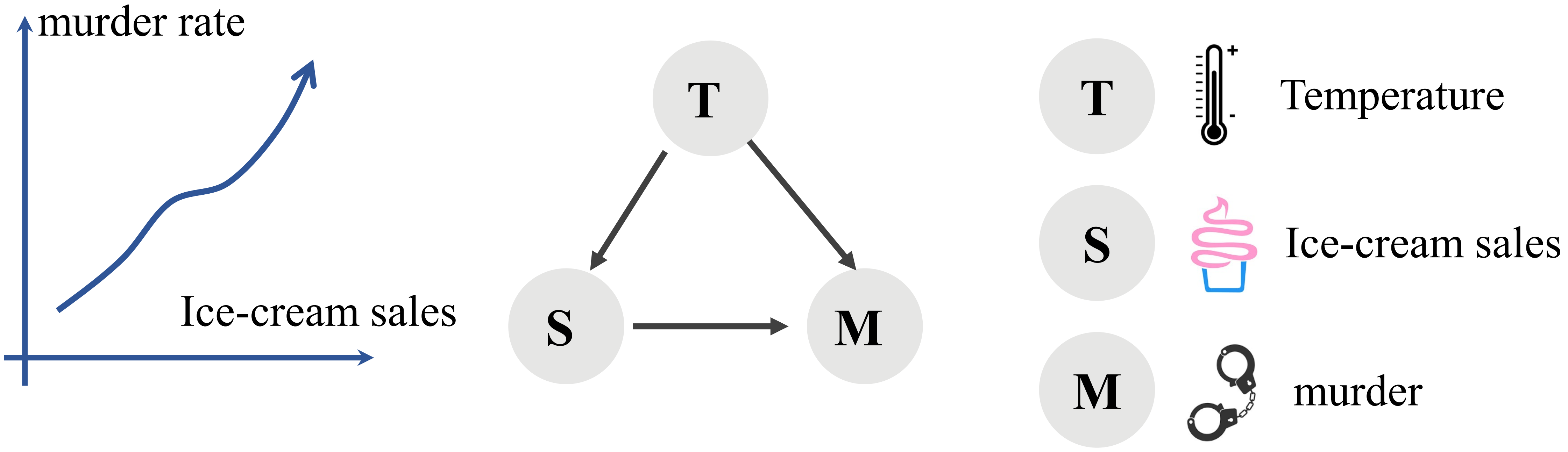
Confounder



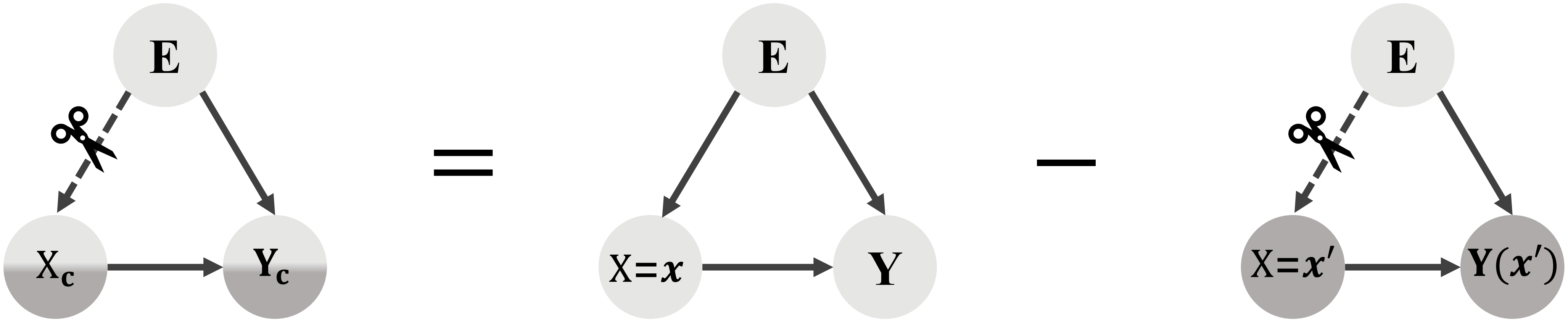
Accident
x_1
x_2
x_3
COVID-19
Death
Complications
x_1
x_2
x_3
COVID-19
Death
Touch
x_1
x_2
x_3
COVID-19
Death
Loss
x_1
x_2
x_3
COVID-19
Death
unbiased
Confounder
Old age
x_1
x_2
x_3
COVID-19
Death
Bias caused by confounder

A Typical Example of the Confounder

The consumption of ice cream and the number of murders in New York are positively correlated. It is biased because of the confounder of confounder temperature.



Counterfactual Analysis



Causal prediction | Factual prediction | Counterfactual prediction

$$\hat{Y}_{causal} = \hat{Y}_x - \hat{Y}_{X_i=x'}$$

$$\hat{Y}_x = \mathcal{F}_\theta(X_i = x)$$

$$\hat{Y}_{X_i=x'} = \mathcal{F}_\theta(do(X_i = x'))$$

Effect of Treatment on the Treated $ETT = E[Y_x - Y_{x'} | X = x]$

- ➢ Reducing the bias of environment
- ➢ Enhancing the causation between outputs and main clues

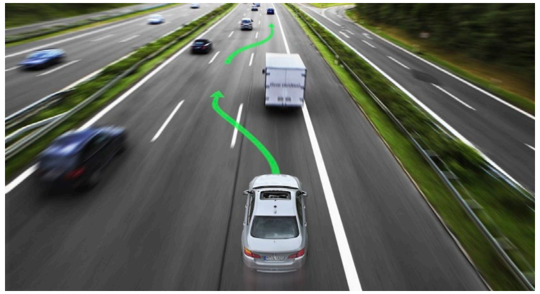
Contents

The Goal of Trajectory Prediction

Inputs: observed previous positions

Outputs: one/multiple reasonable future predictions

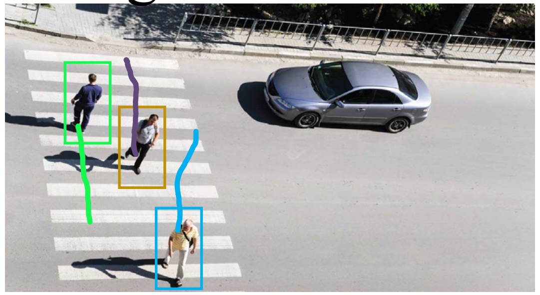
Autonomous vehicles



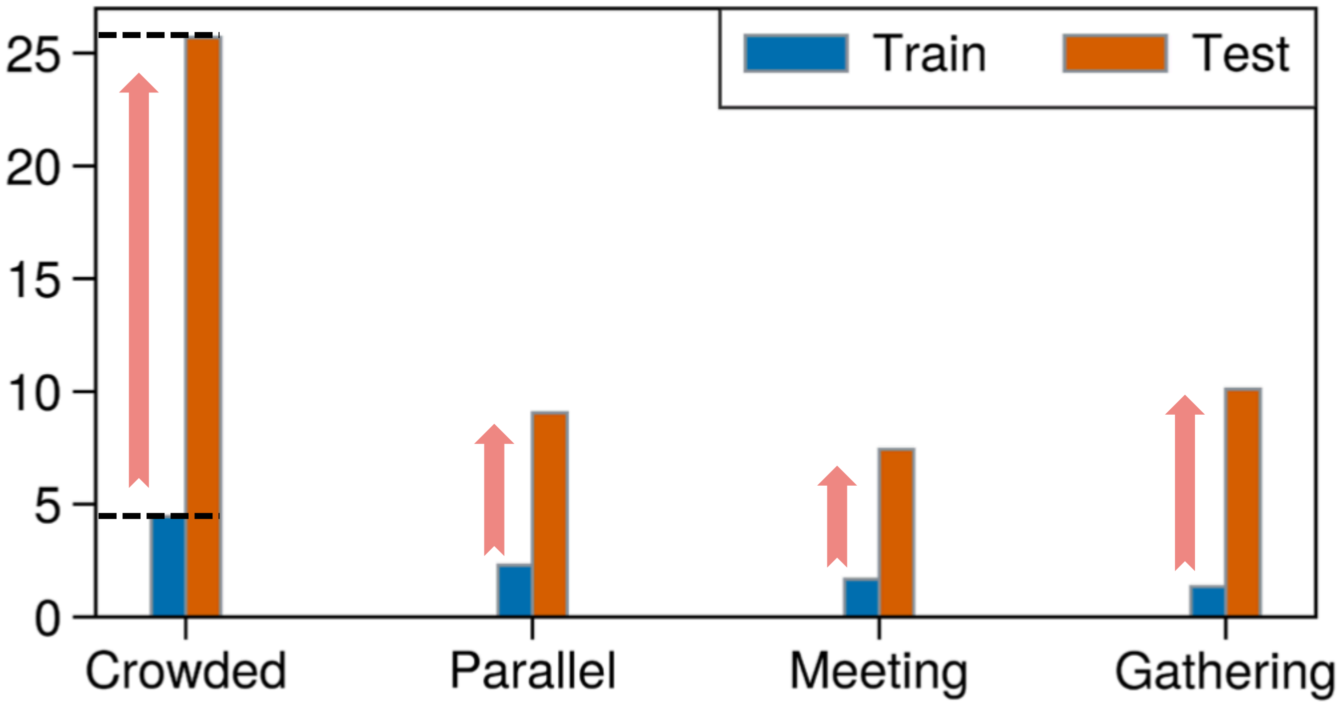
Social robotics

Delivery bots

Intelligent surveillance

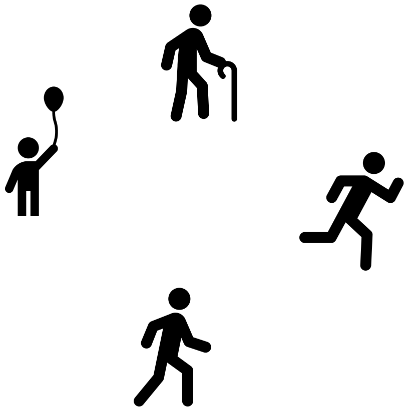


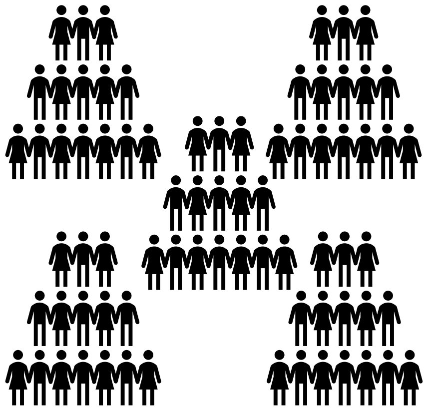
Environment Bias



Statistical bias between training and testing environments.

E.g., number of neighbors



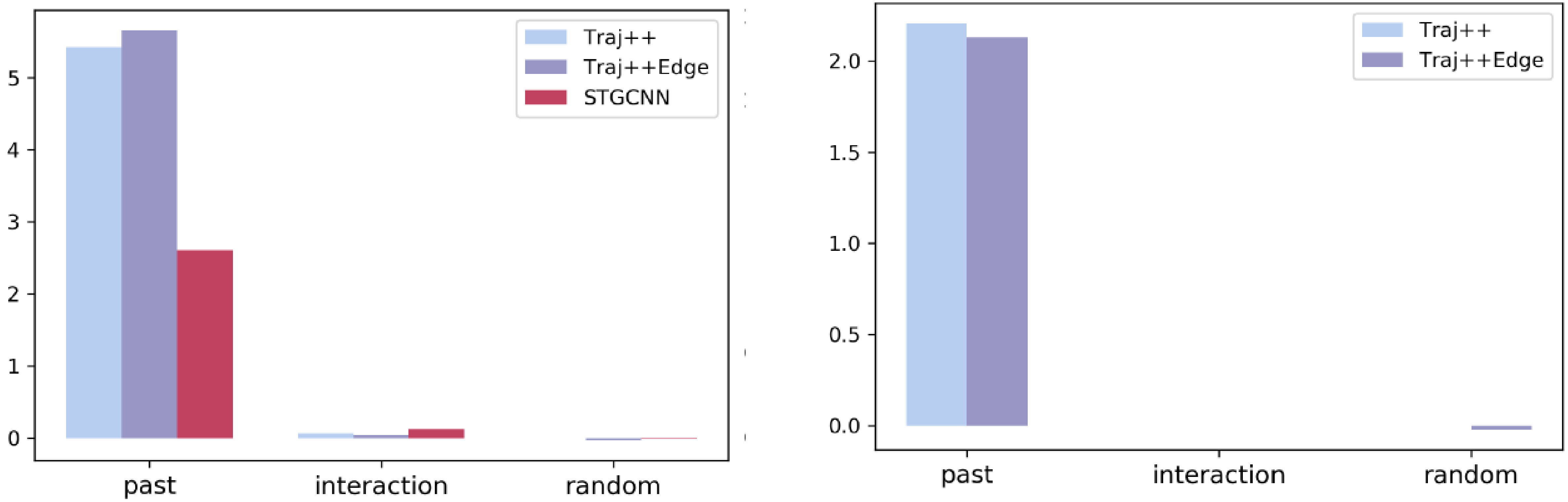


Environment Bias

Visualization of the obvious environment difference.

Available area

Other Evidences of Interaction Bias.

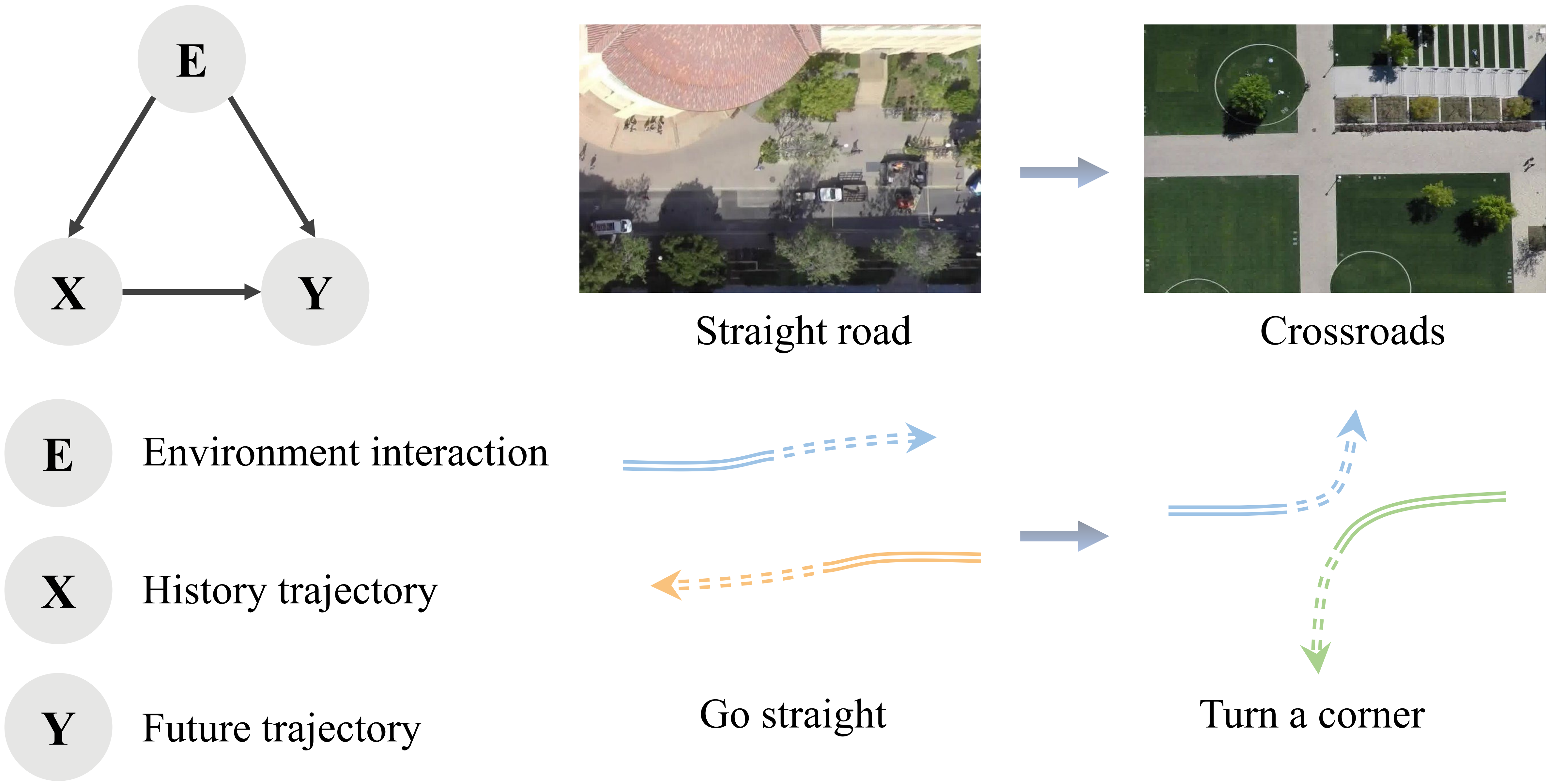


Shapley values of Trajectron++ variants and STGCNN on ETH/UCY and SDD datasets

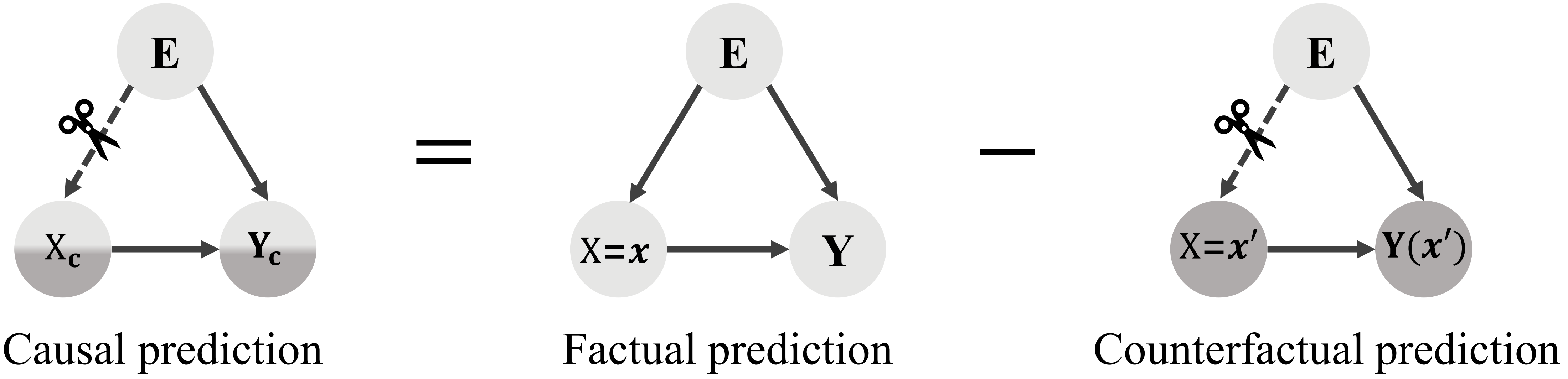
Shapley values of past trajectory is far higher than the others

Makansi, Osama, et al. "You Mostly Walk Alone: Analyzing Feature Attribution in Trajectory Prediction." arXiv, 2021.

Causal Graph



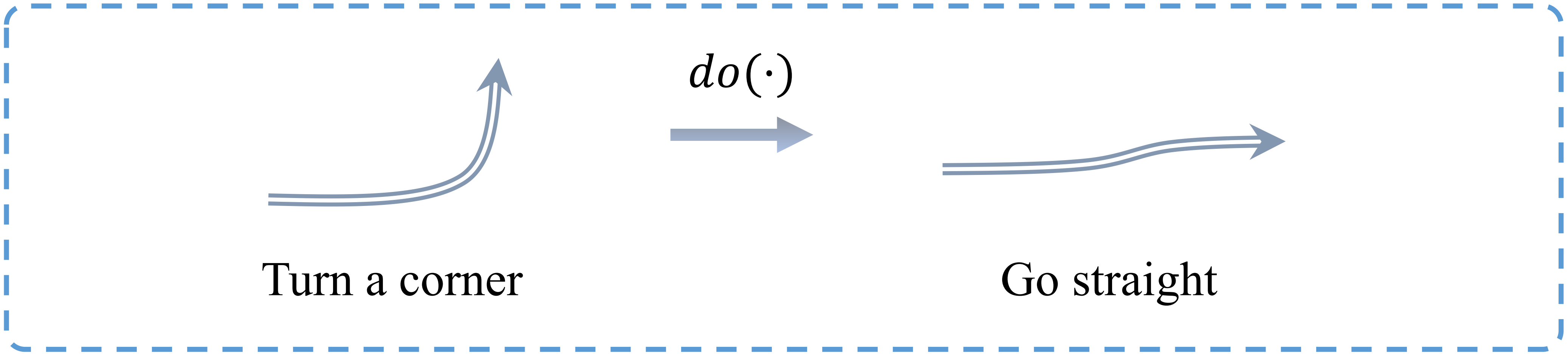
Key idea



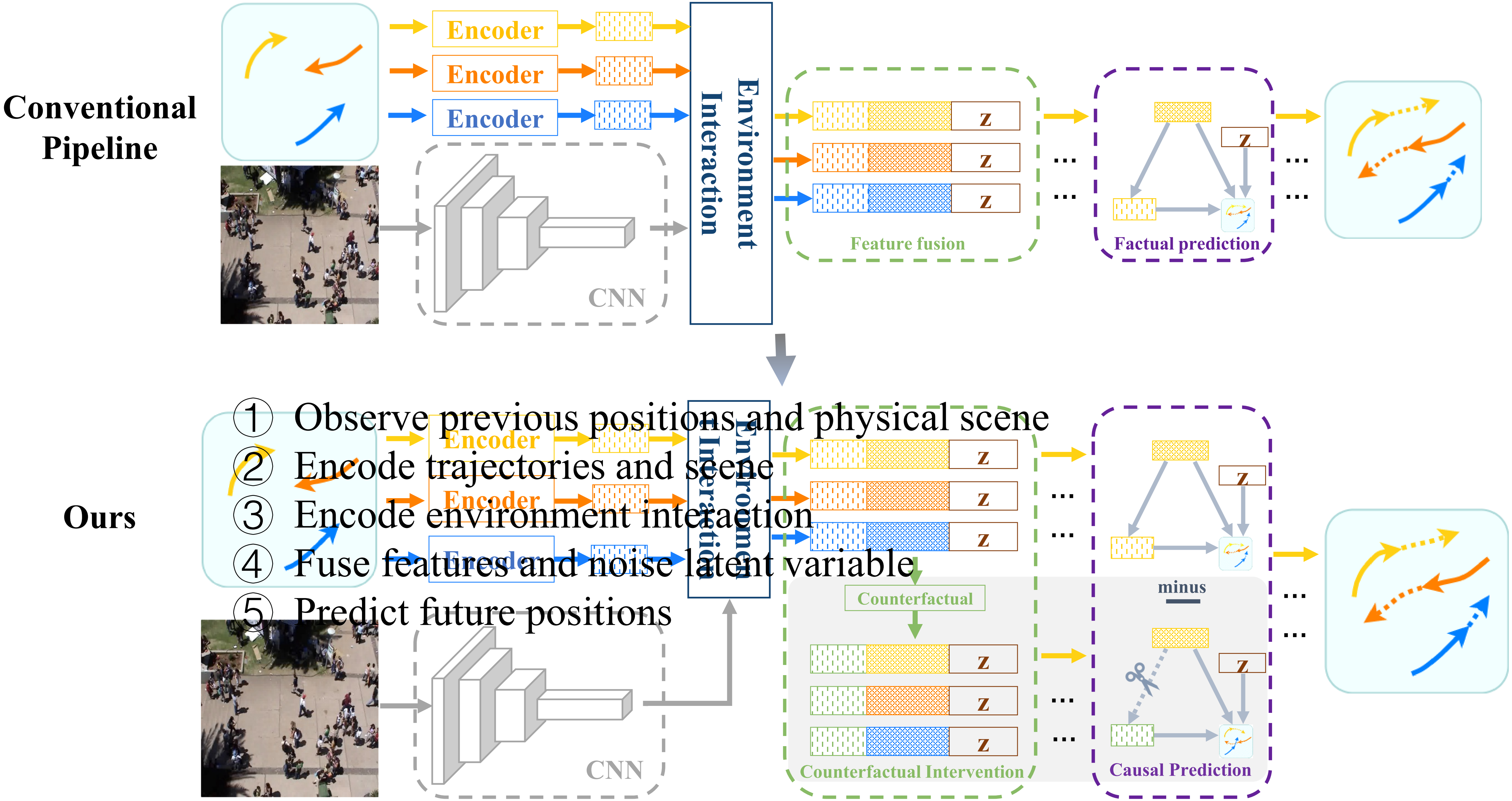
$$\hat{Y}_{causal} = \hat{Y}_x - \hat{Y}_{X_i=x'}$$

$$\hat{Y}_x = \mathcal{F}_\theta(X_i = x)$$

$$\hat{Y}_{X_i=x'} = \mathcal{F}_\theta(do(X_i = x'))$$

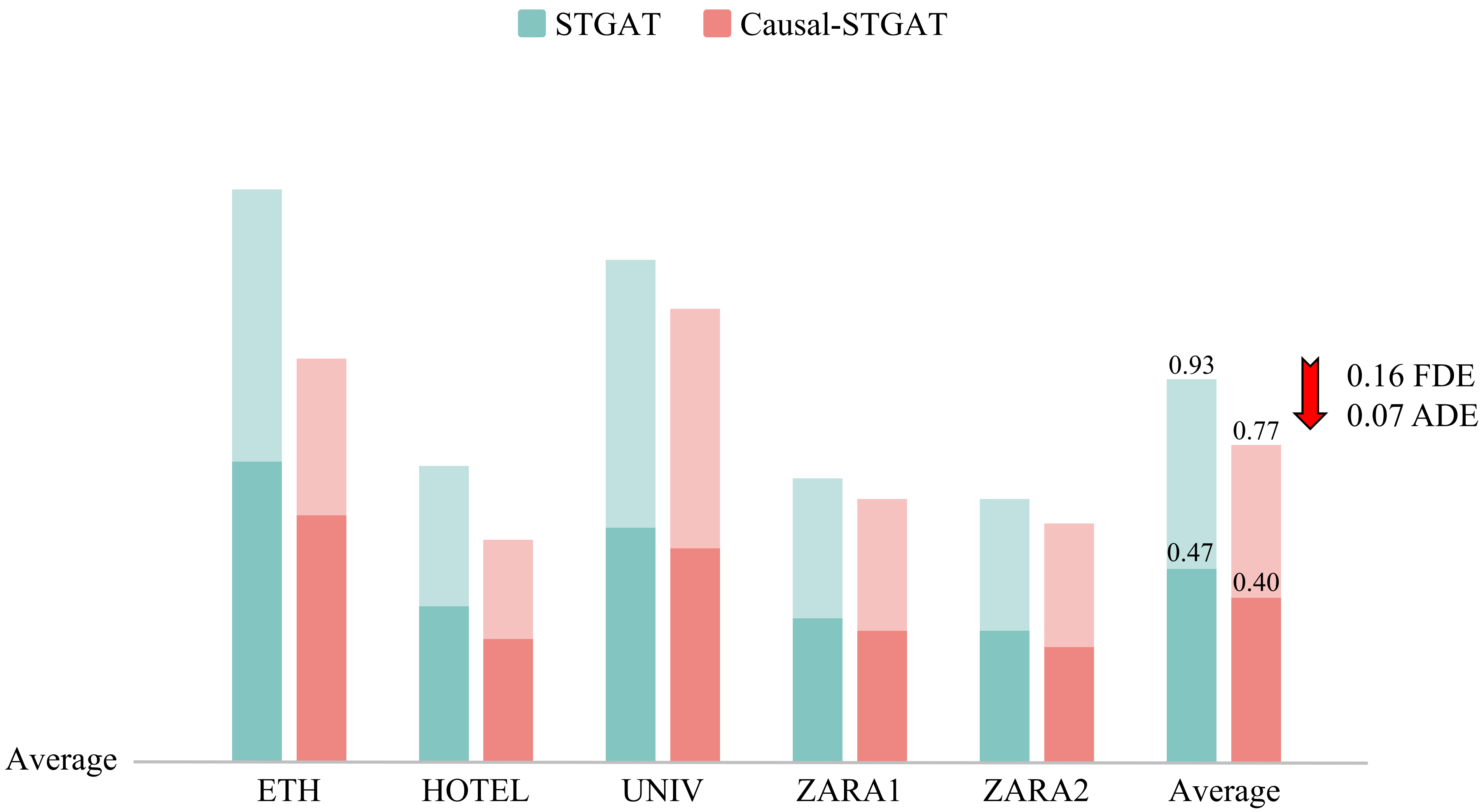


Pipeline



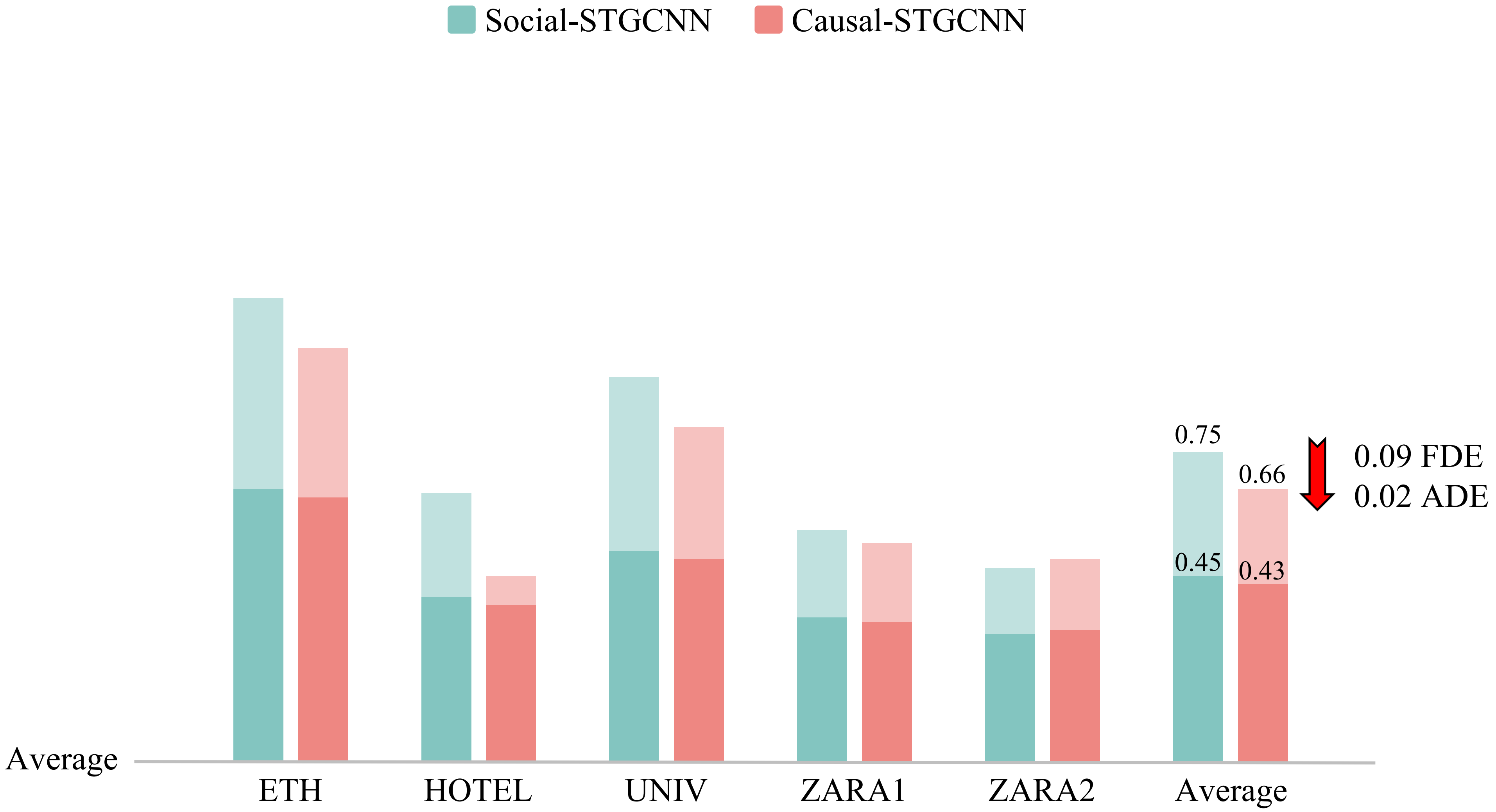
① Observe previous positions and physical scene
② Encode trajectories and scene
③ Encode environment interaction
④ Fuse features and noise latent variable
⑤ Predict future positions

Quantitative Evaluation on STGAT



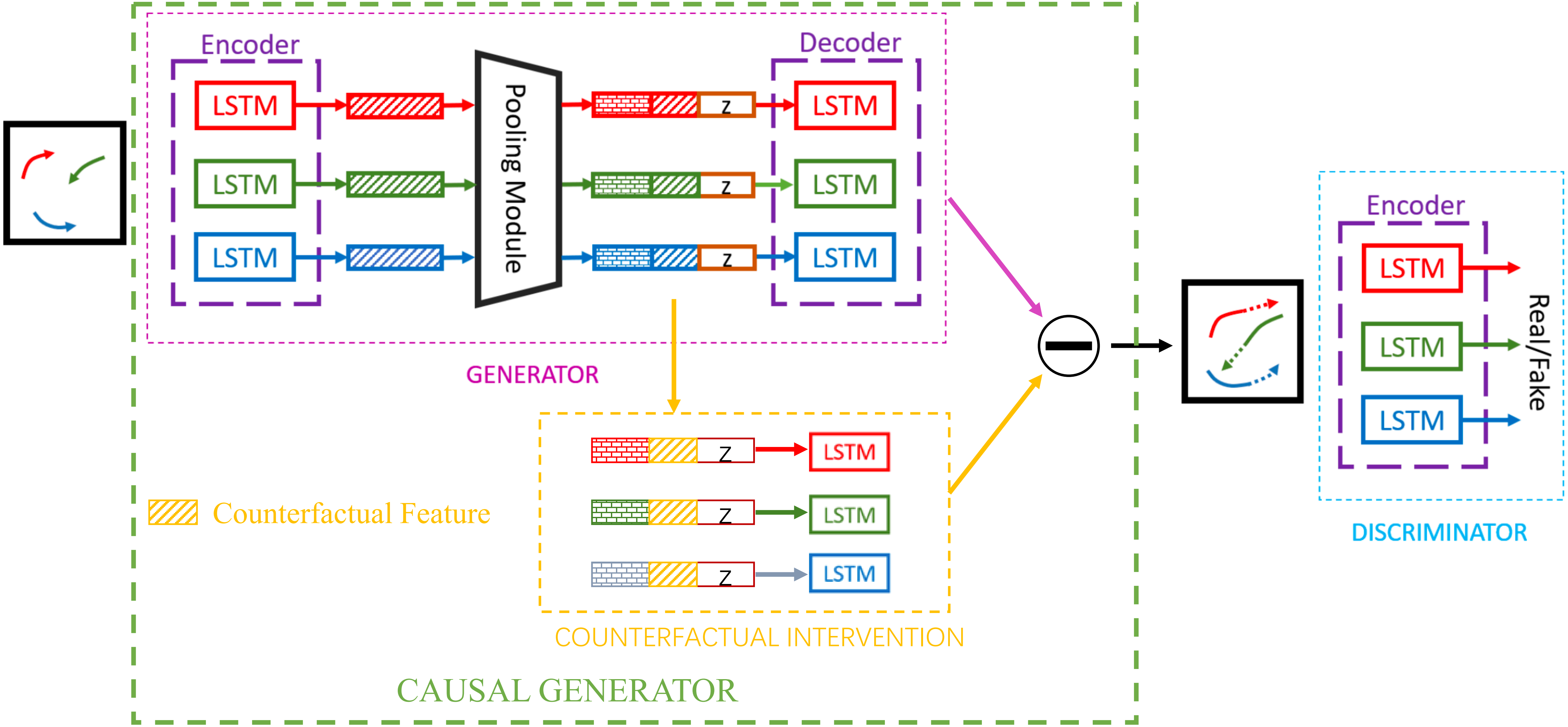
Our causal-based methods improve performance on both ADE (below) & FDE (stacked on top).

Quantitative Evaluation on Social STGCNN

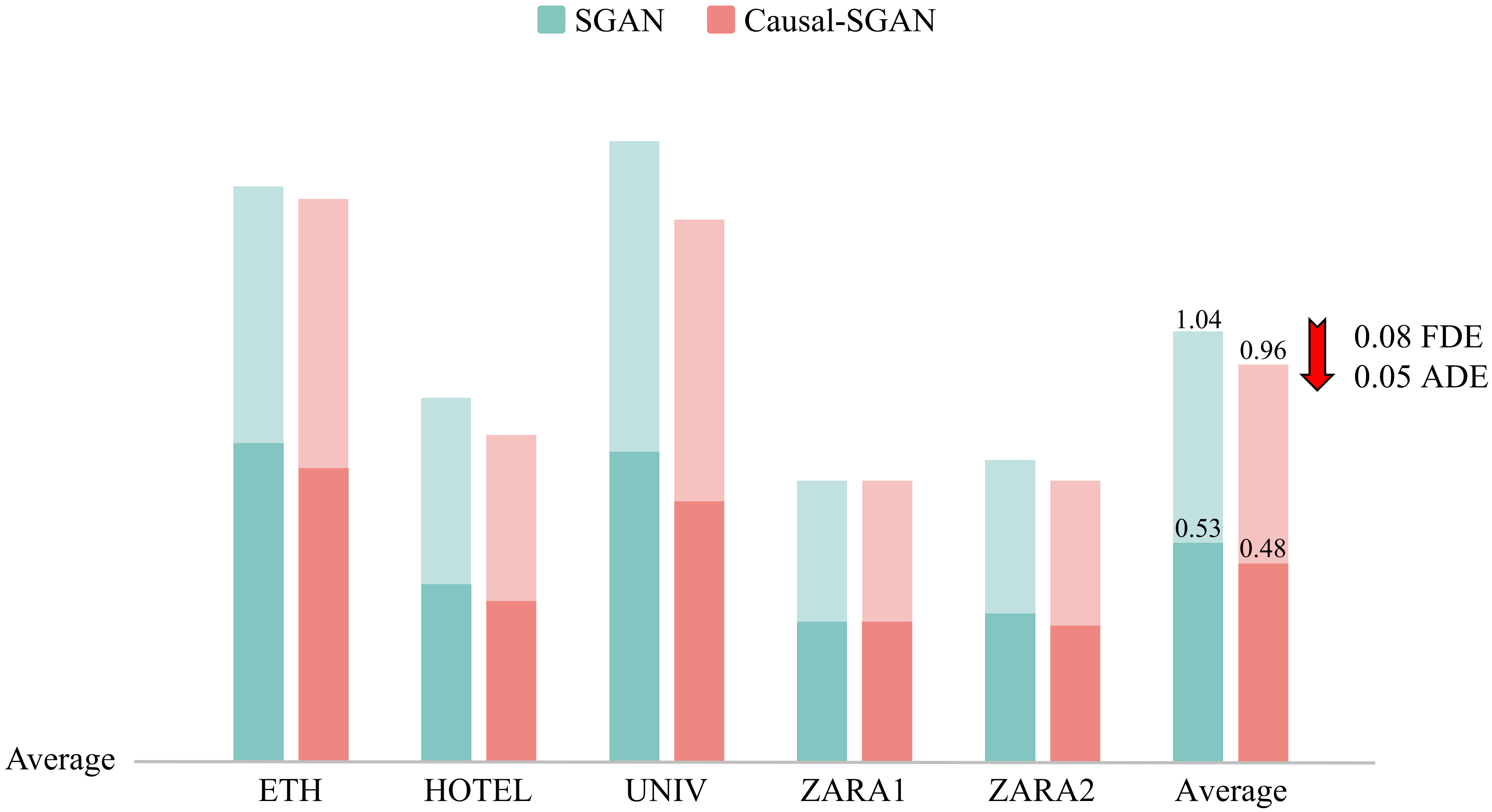


Our causal-based methods improve performance on both ADE (below) & FDE (stacked on top).

Counterfactuals on Social GAN

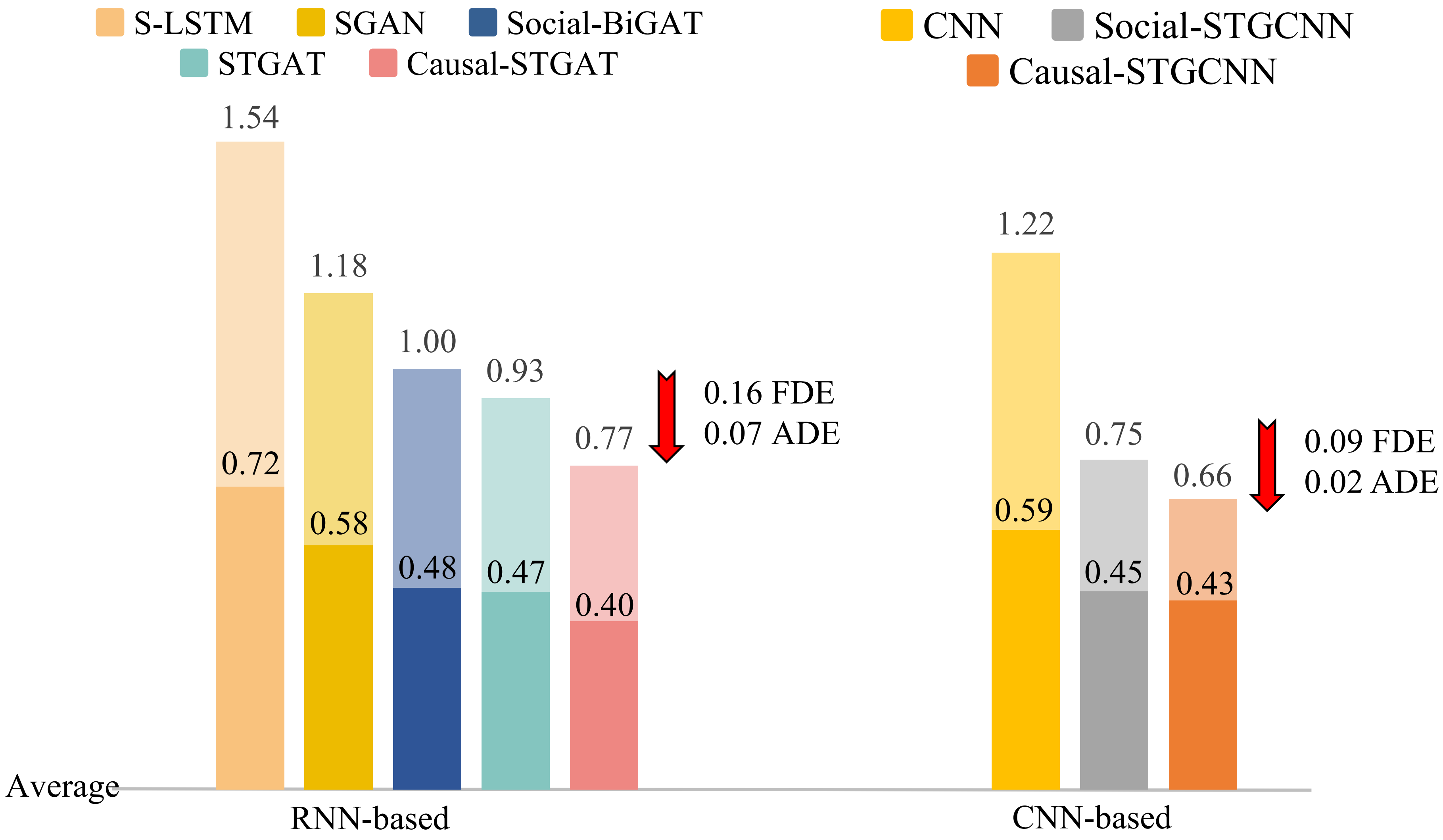


Quantitative Evaluation on Social GAN



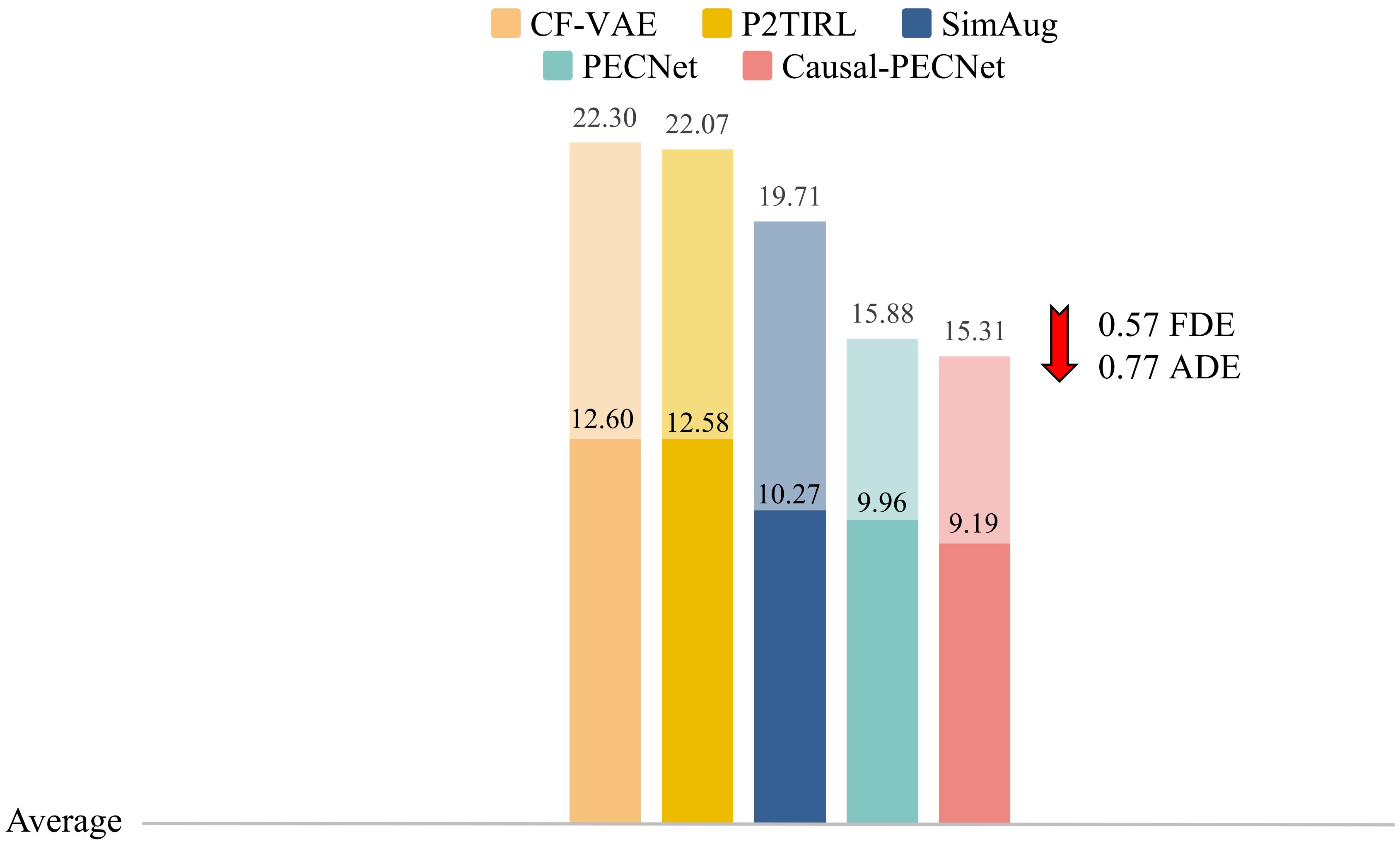
Our causal-based methods improve performance on both ADE (below) & FDE (stacked on top).

Quantitative Evaluation on ETH/UCY



Our causal-based methods improve performance on both ADE (below) & FDE (stacked on top).

Quantitative Evaluation on SDD



Our causal-based methods improve performance on both ADE (below) & FDE (stacked on top).

Counterfactual Implementation

Method	ADE	FDE
STGAT(Baseline)	0.47	0.93
Causal-STGAT-Zero	0.40	0.77
Causal-STGAT-Mean	0.44	0.84
Causal-STGAT-Random	0.42	0.80
Causal-STGAT-Generate	0.40	0.76

Zero: a zero vector
Mean: the mean of all feature vectors
Random: a random vector sampled from a uniform distribution with [-0.1, 0.1]
Generate: a vector learned by a generative model (GAN or VAE)

Only interaction is biased?

Method	ADE	FDE
STGAT	0.47	0.93
Causal-STGAT-X	0.40	0.77
Causal-STGAT-S	0.46	0.86

X: history trajectory

S: social interaction

Interactions are biased

Past trajectories are biased too

Inference Speed and Model Size

Method	Social-STGCNN		Causal-STGCNN	STGAT		Causal-STGAT
Parameters Count	7.6k	≈	7.6k	56k	≈	56k
Inference Speed	0.0116		0.0124	0.3343		0.3418

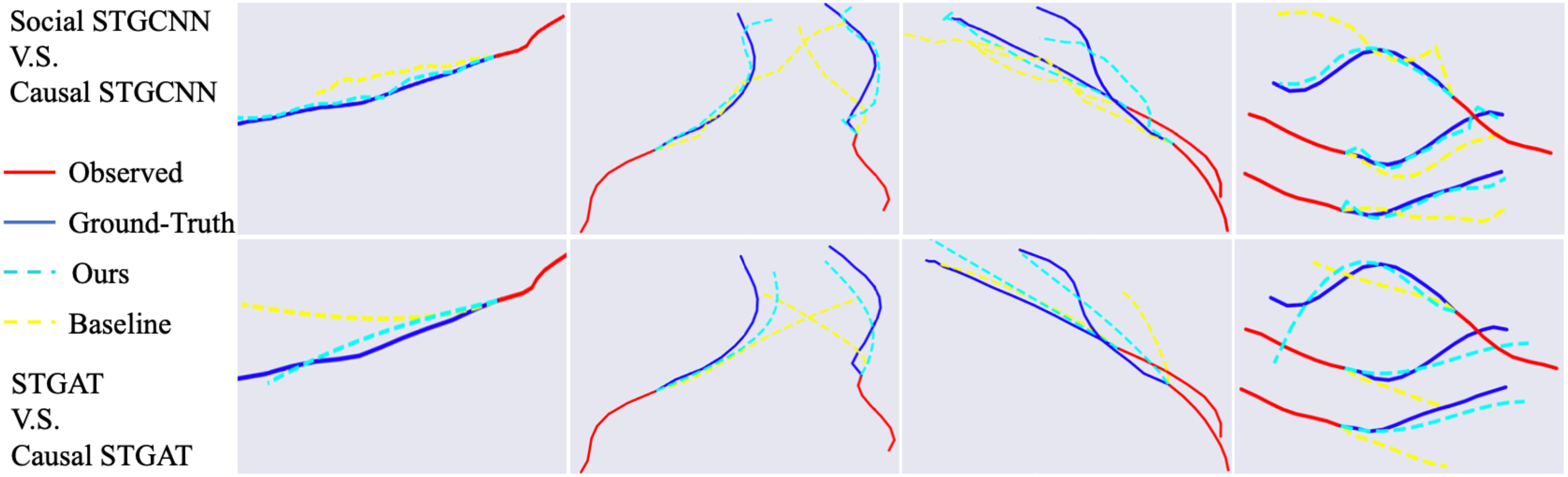
Our counterfactual analysis method does not need any extra parameters

Δ≈ 7%

Δ≈ 2%

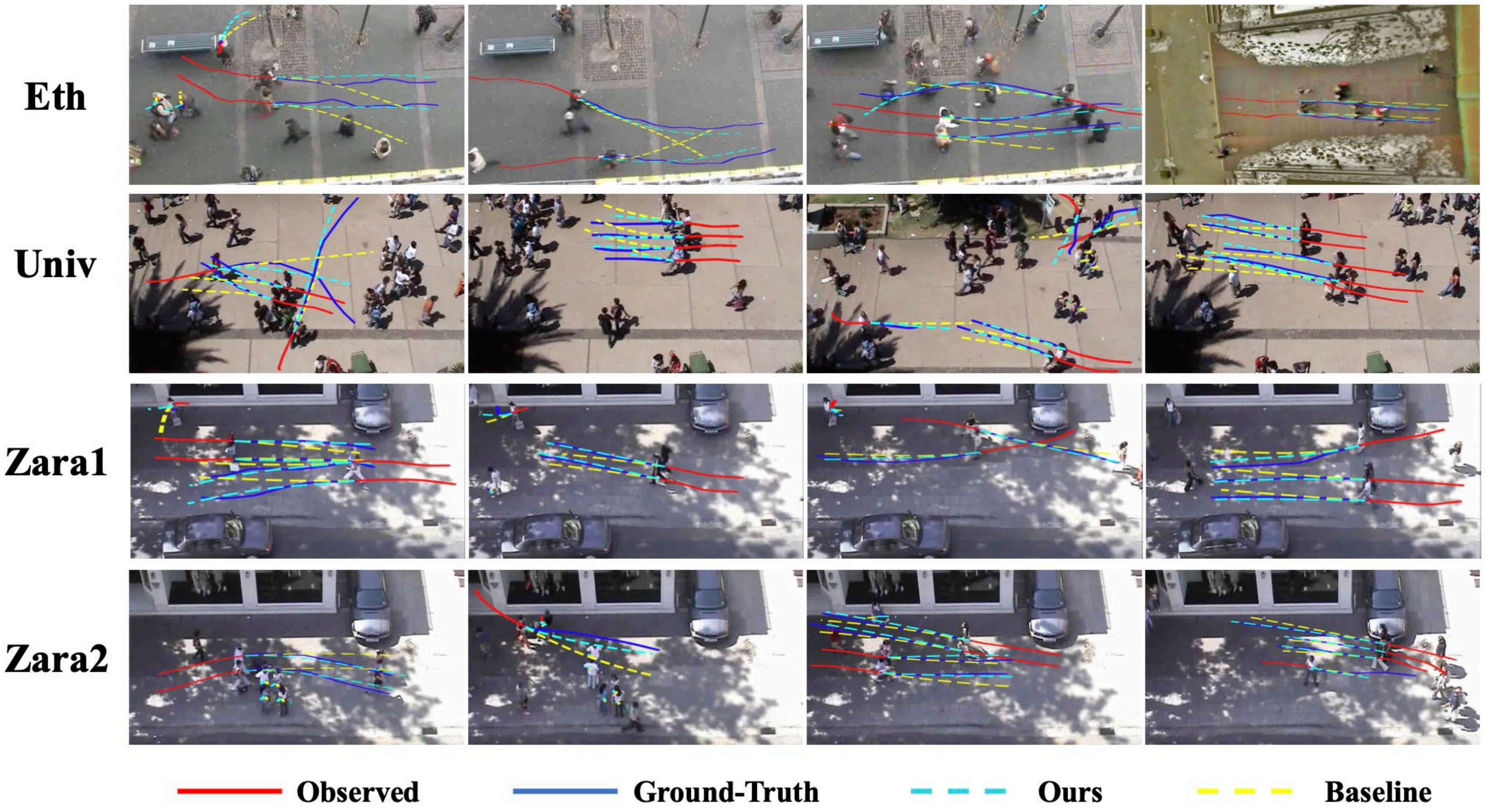
The extra speed cost of our counterfactual analysis method is not heavy

Qualitative Evaluation



(a) Walking alone (b) Gathering (c) Meeting from the same direction (d) Meeting from different directions

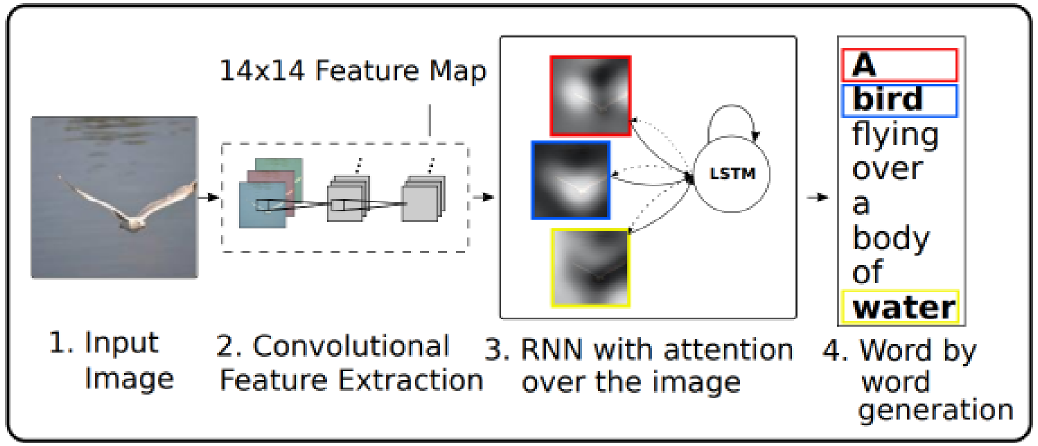
Qualitative Evaluation



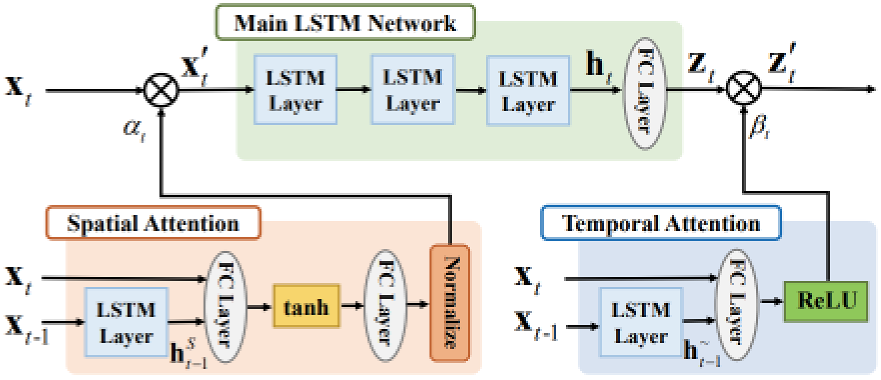
Contents

- Introduction of Counterfactual Analysis
- Approach 1: Human Trajectory Prediction via Counterfactual Analysis
- Approach 2: Counterfactual Attention Learning
- Approach 3: Benchmarking Fairness of Image Recognition Models
- Future Work

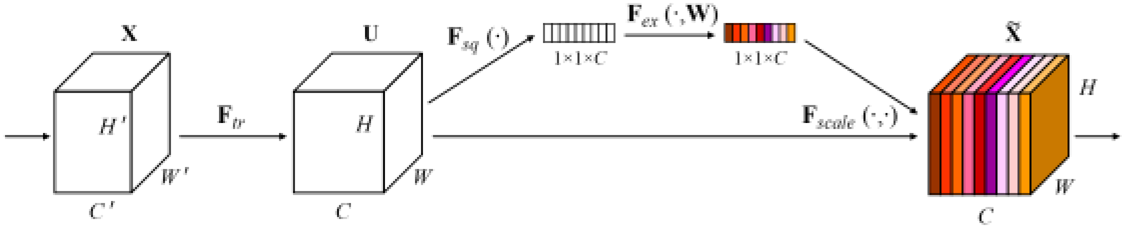
Attention Learning



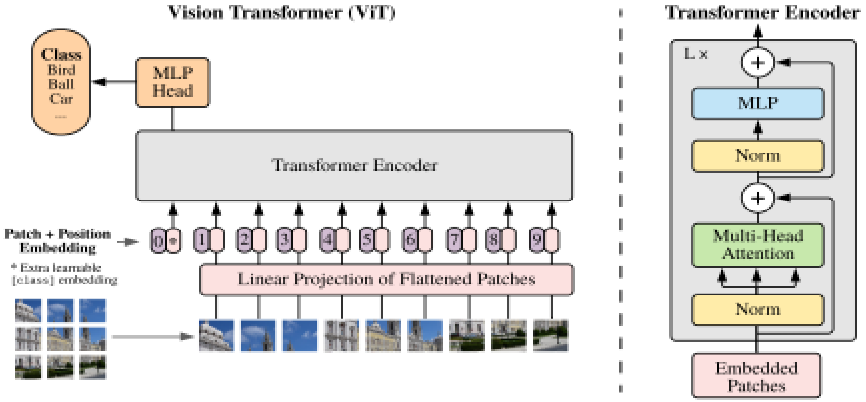
Show, attend and tell, 2015



Spatio-temporal attention, 2017



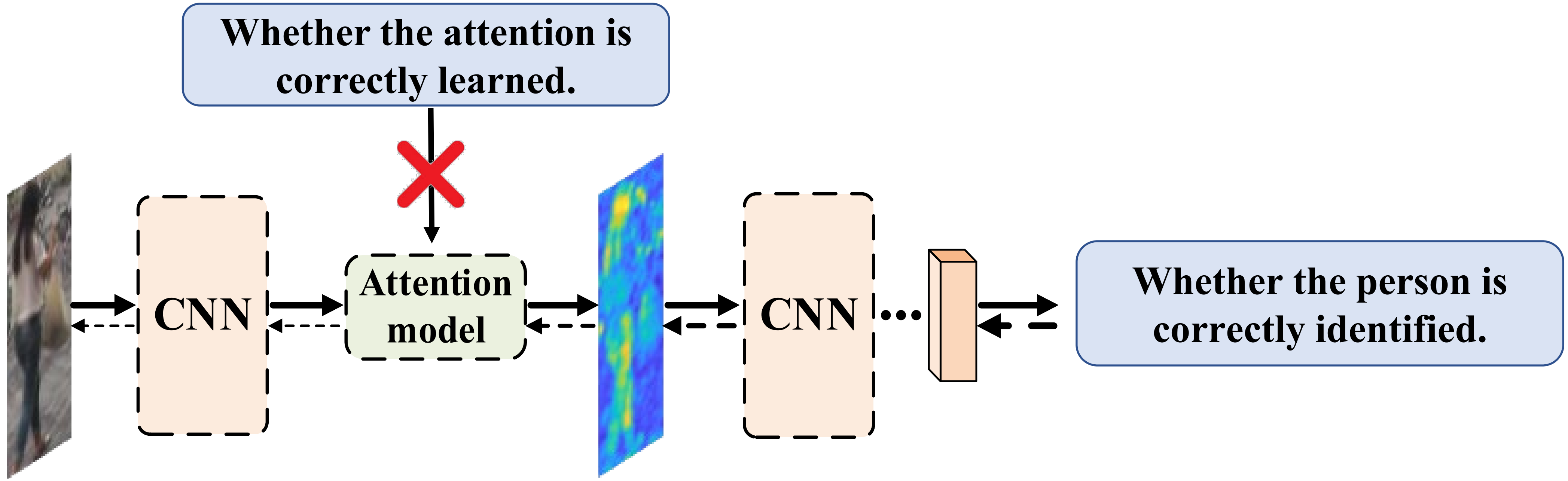
SENet, 2017

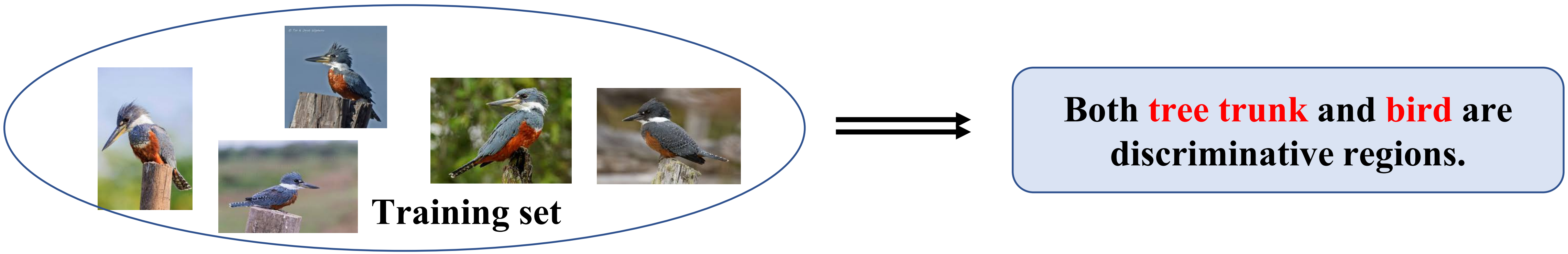


Vision Transformer, 2020

Motivation

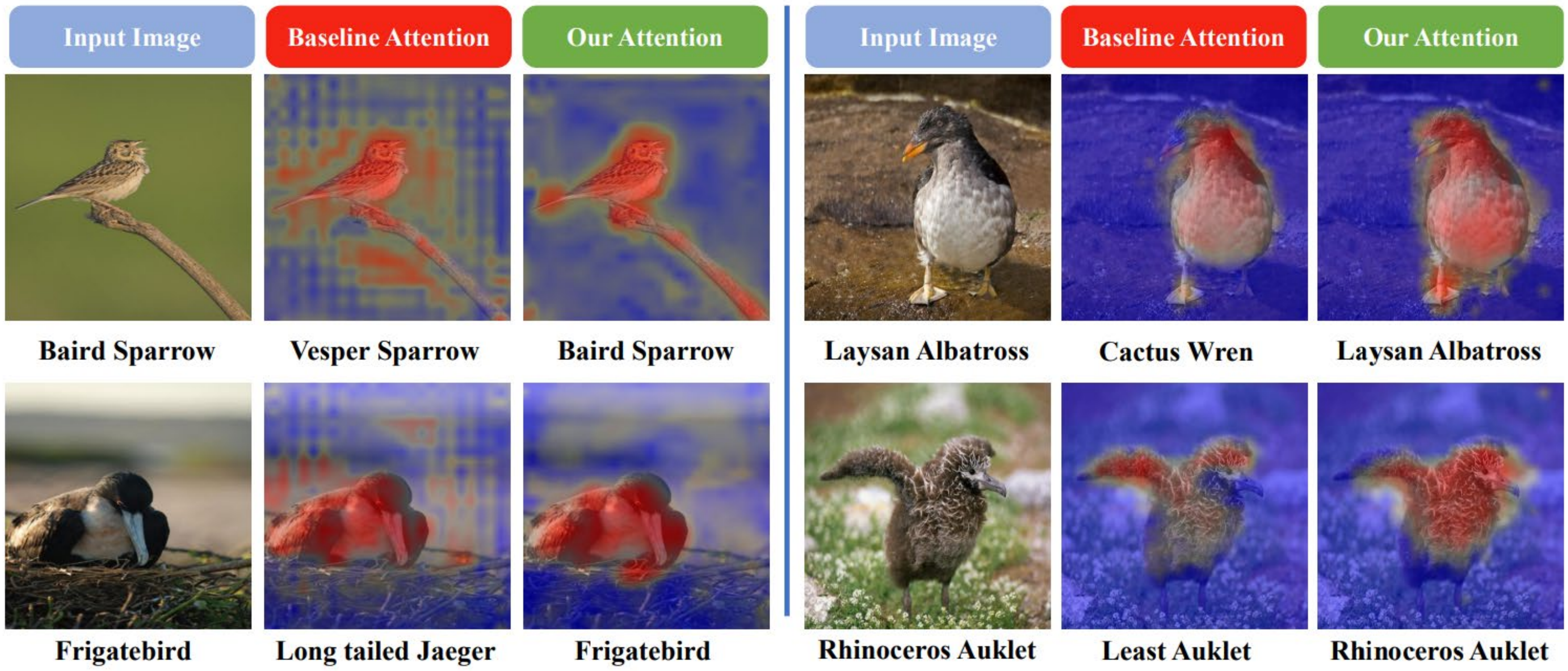
➢ Attention is always learned in a weakly-supervised manner, which ignores the causality between the prediction and attention.





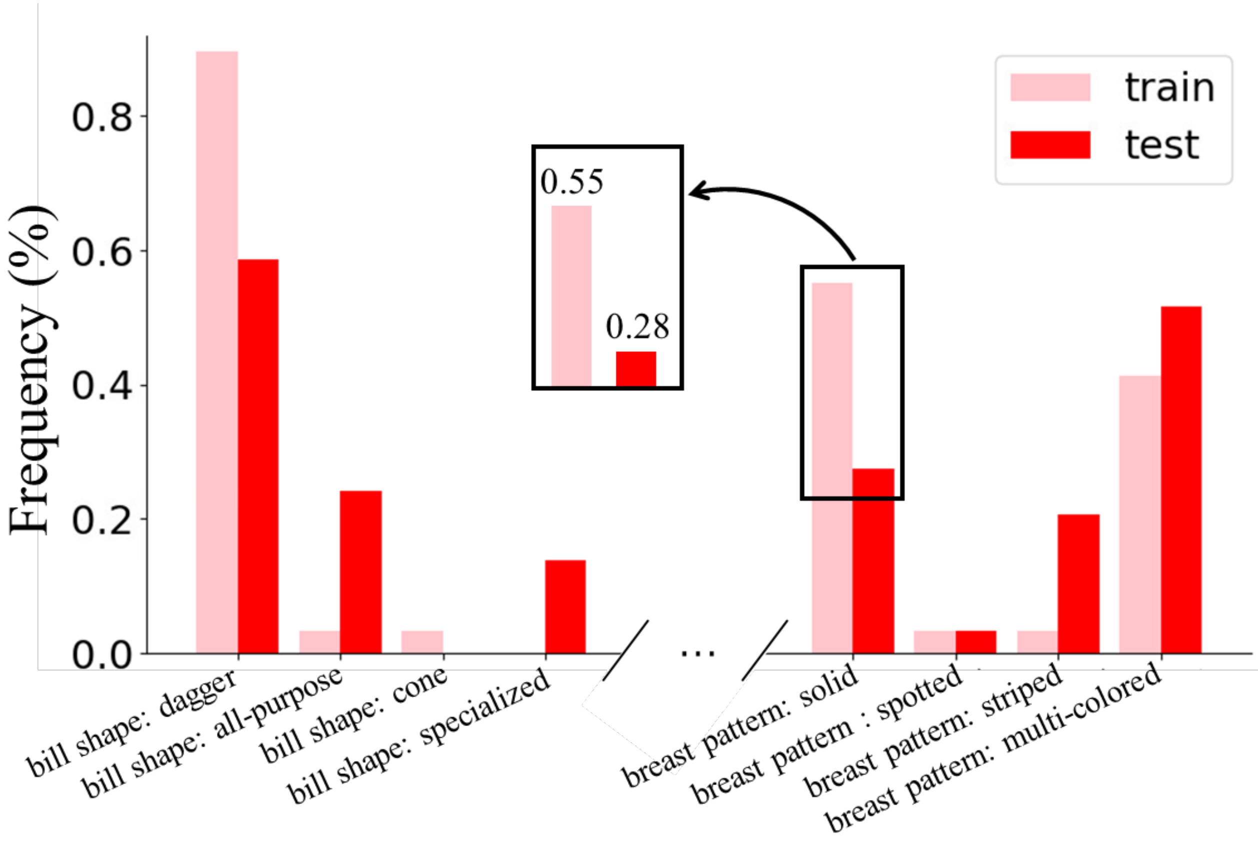
Motivation

➢ Misleading and scattered attentions can still be observed from a well-trained attention model

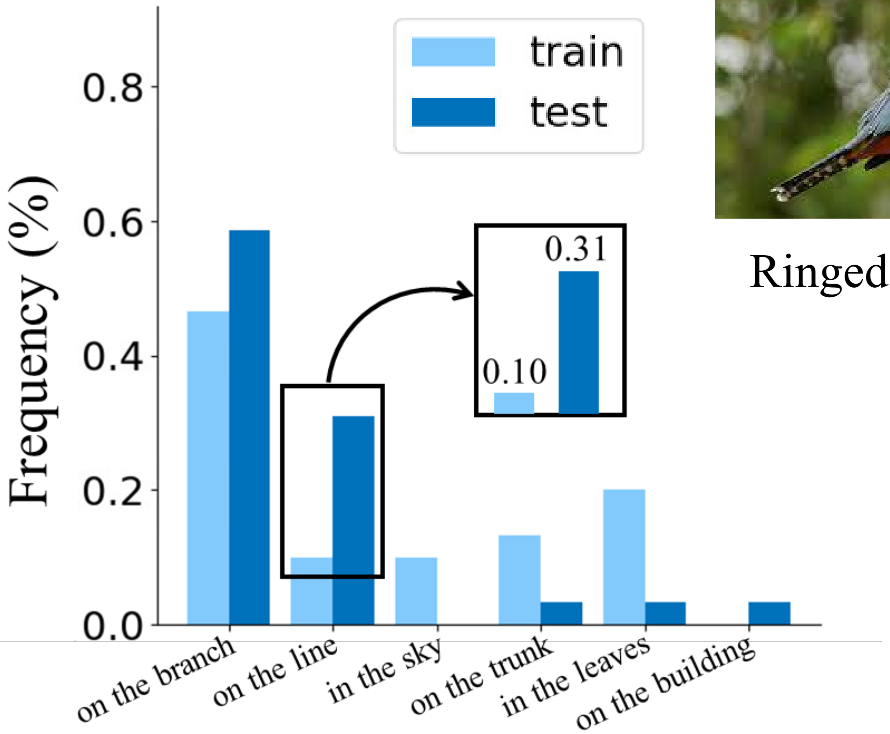


Biases in Fine-Grained Visual Recognition

- In the task of fine-grained categorization, both intrinsic attributes and external environments show the dataset bias in the statistics.



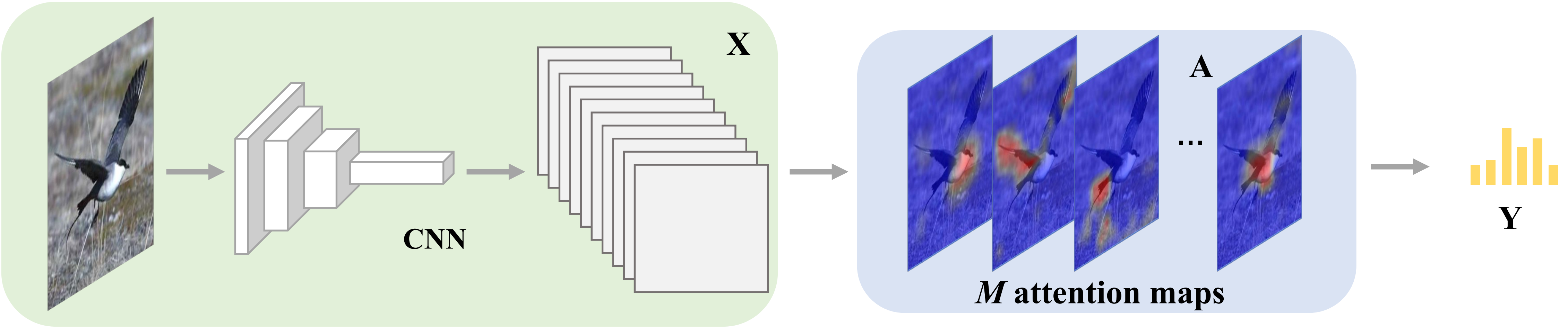
(a) The attributes

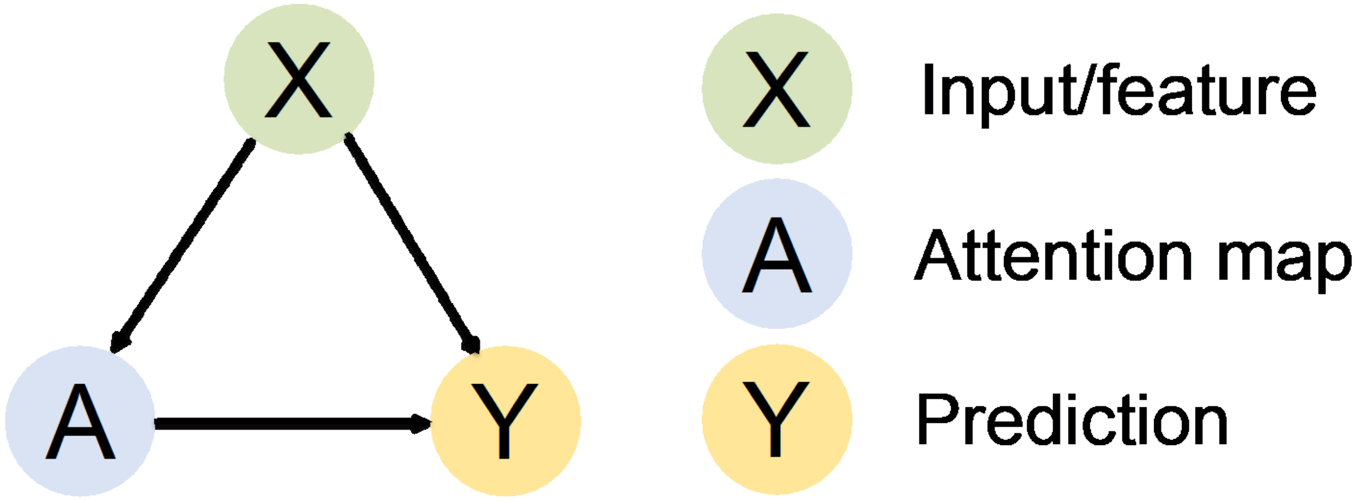


(b) The environments

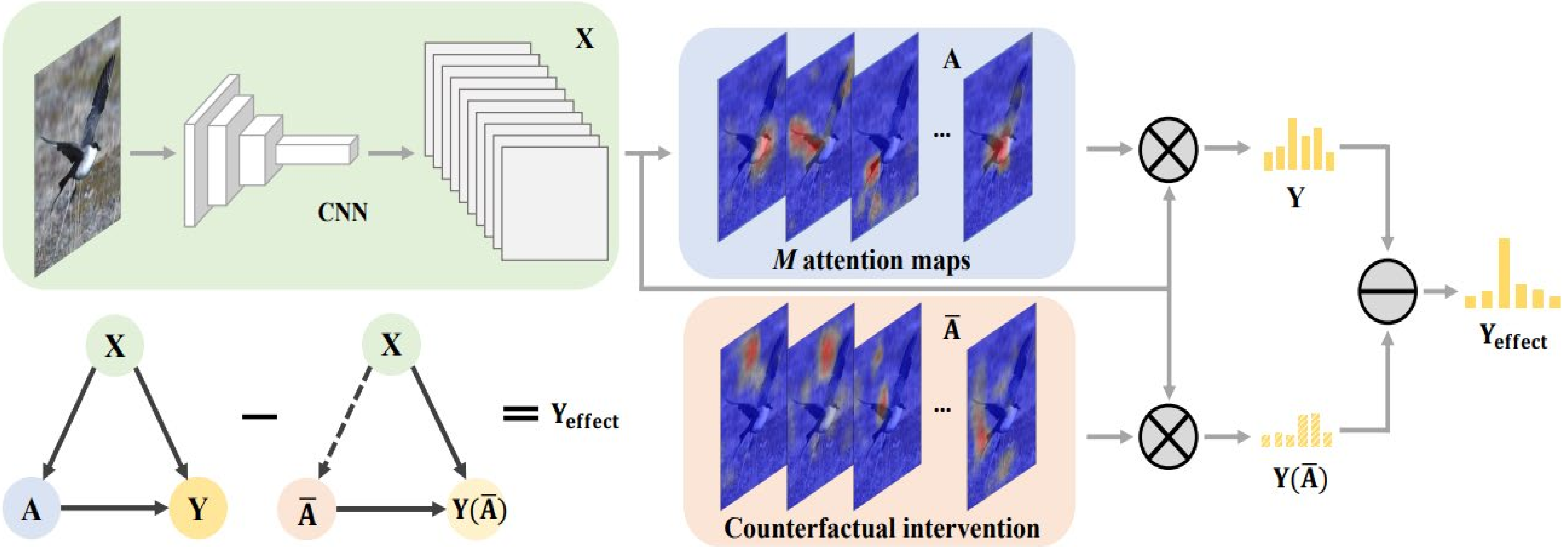
Ringed Kingfisher

Attention Models for Fine-Grained Recognition





Counterfactual Attention Learning



$$Y_{\text{effect}} = \mathbb{E}_{\bar{\mathbf{A}} \sim \gamma}\left[Y(A=\mathbf{A}, X=\mathbf{X}) - Y(do(A=\bar{\mathbf{A}}), X=\mathbf{X})\right]$$

Results on Fine-grained Image Categorization

Method	CUB	Cars	Aircraft
RA-CNN [12]	85.3	92.5	-
MA-CNN [65]	86.5	92.8	89.9
MAMC [50]	86.5	93.0	-
NTS-Net [63]	87.5	93.9	91.4
WS-DAN [19]	89.4	94.5	93.0
DCL [9]	87.8	94.5	93.0
Stacked LSTM [13]	90.4	-	-
API-Net [70]	90.0	95.3	93.9
Baseline	89.3	94.0	93.6
Baseline + CAL	**90.6**	**95.5**	**94.2**

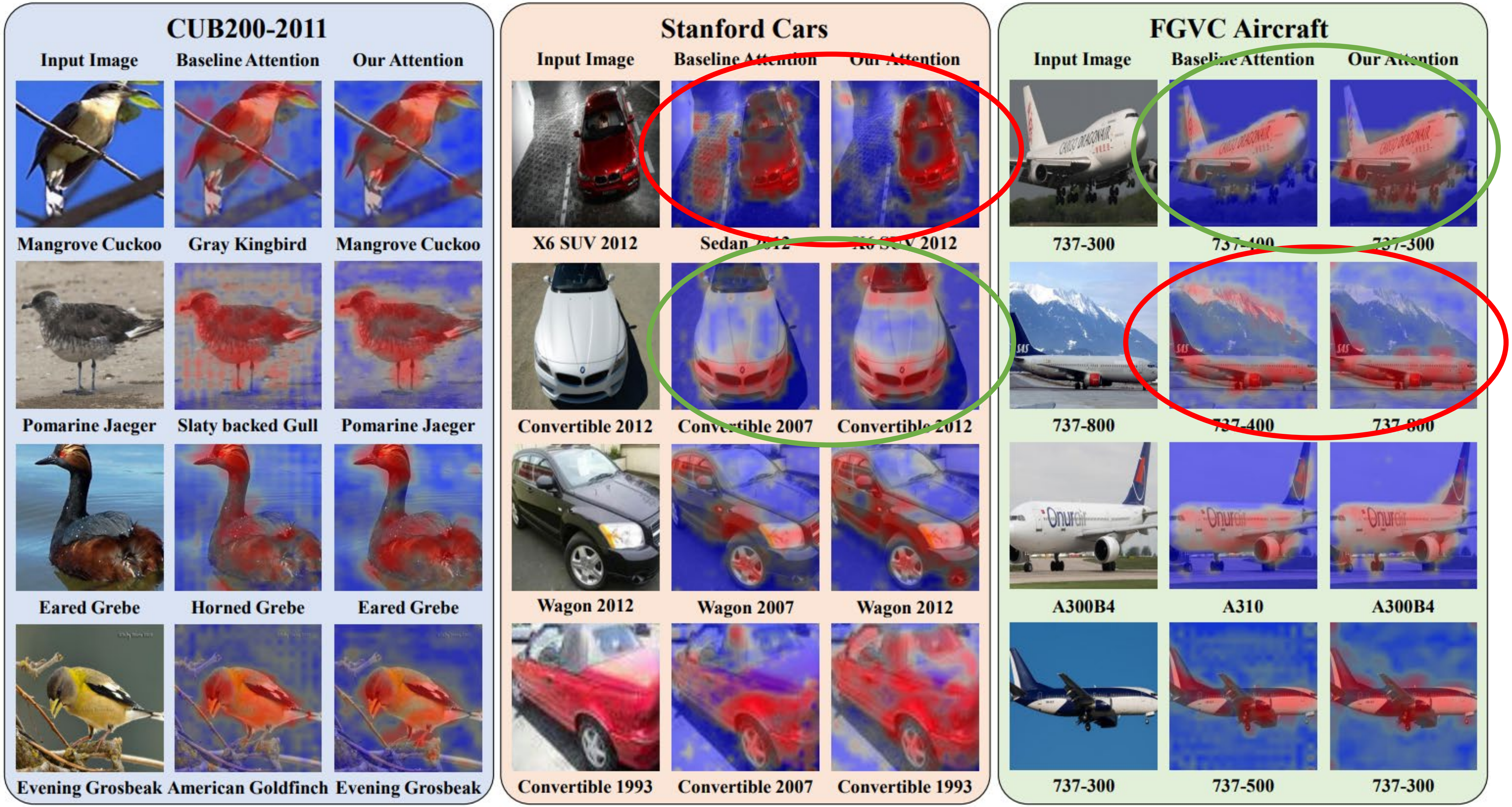
Results on Person Re-identification

Method	Market1501			DukeMTMC-ReID			MSMT17		
	R1	R5	mAP	R1	R5	mAP	R1	R5	mAP
HA-CNN [27]	91.2	-	75.7	80.5	-	63.8	-	-	-
Part-aligned [49]	91.7	96.9	79.6	84.4	92.2	69.3	-	-	-
Mancs [56]	93.1	-	82.3	84.9	-	71.8	-	-	-
PCB+RPP [52]	93.8	97.5	81.6	83.3	-	69.2	68.2	-	40.4
IANet [17]	94.4	-	83.1	87.1	-	73.4	75.5	85.5	46.8
JDGL [67]	94.8	-	86.0	86.6	-	74.8	77.2	-	52.3
SCAL [5]	95.8	**98.7**	89.3	88.9	95.2	79.1	-	-	-
MHN [4]	95.1	98.1	85.0	89.1	94.6	77.2	-	-	-
SFT [37]	93.4	-	82.7	86.9	-	73.2	73.6	-	47.6
OSNet [68]	94.8	-	84.9	88.6	-	73.5	78.7	-	52.9
BAT-Net [11]	95.1	98.2	87.4	87.7	94.7	77.3	79.5	89.1	56.8
Auto-ReID [43]	94.5	-	85.1	-	-	-	78.2	88.2	52.5
MGN+circleloss [51]	**96.1**	-	87.4	-	-	-	76.9	-	52.1
Baseline	94.0	97.7	85.9	85.7	93.6	74.0	75.3	86.4	50.5
Baseline + CAL	94.5	97.9	87.0	87.2	94.1	76.4	79.5	89.0	56.2
Baseline†	94.9	98.3	89.0	88.7	94.7	78.2	81.4	90.3	59.3
Baseline† + CAL	95.5	98.5	**89.5**	**90.0**	**96.1**	**80.5**	**84.2**	**92.0**	**64.0**

Results on Vehicle Re-identification

Method	Veri-776			VehicleID								
	Test 11587			Test 800			Test 1600			Test 2400		
	R1	R5	mAP	R1	R5	mAP	R1	R5	mAP	R1	R5	mAP
GSTE [3] -	-	59.4	87.1	-	-	82.1	-	-	79.8	-	-	
AAMI [69]	85.9	91.8	61.3	63.1	83.3	-	52.9	75.1	-	47.3	70.3	-
FDA-NeT [34]	84.3	92.4	55.5	-	-	-	59.8	77.1	65.3	55.5	74.7	61.8
VAML* [10]	89.8	96.0	66.3	88.1	97.3	-	83.2	95.1	-	80.4	93.0	-
AAVER [21]	88.7	94.1	58.5	72.5	93.2	-	66.9	89.4	-	60.2	84.9	-
EALN [35]	84.4	94.1	57.4	75.1	88.1	77.5	71.8	83.9	74.2	69.3	81.4	71.0
DFLNet [2]	93.2	97.6	73.3	78.8	**95.1**	82.8	-	-	-	69.8	**90.6**	75.4
ResNet50	94.5	97.2	72.0	76.7	93.5	84.1	74.9	89.5	81.4	71.0	84.9	78.0
ResNet50 + CAL	**95.4**	**97.9**	**74.3**	**82.5**	94.7	**87.8**	**78.2**	**91.0**	**83.8**	**75.1**	88.5	**80.9**

Qualitative Results



Contents

- Introduction of Counterfactual Analysis
- Approach 1: Human Trajectory Prediction via Counterfactual Analysis
- Approach 2: Counterfactual Attention Learning
- Approach 3: Benchmarking Fairness of Image Recognition Models
- Future Work

Fairness crisis

➢ Most of current models are unfairly biased against certain subpopulations

Joy Buolamwini, Algorithmic Justice League

Joy Buolamwini, Algorithmic Justice League

Joy Buolamwini. Artificial Intelligence Has a Problem With Gender and RacialBias, 2019.

Some challenging examples

FairNet

Table 1: Comparisons of the proposed *FairNet* against existing image classification benchmarks.

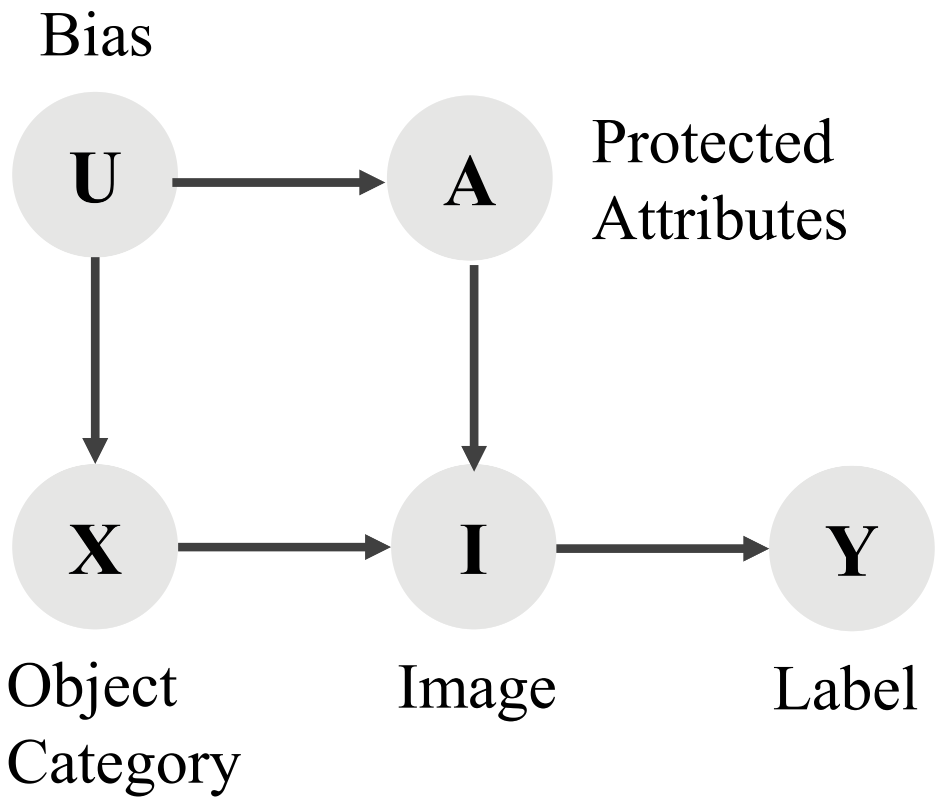
Dataset	Size	#Category	Source	Bias	Metric	Attribute	Year
ImageNet val [26]	50,000	1,000	Internet	-	accuracy	✗	2012
ImageNet v2 [24]	10,000	1,000	Internet	natural distribution shift	accuracy	✗	2019
ObjectNet [1]	50,000	313	self-collected	backgrounds, rotations, viewpoints	accuracy	3	2019
ImageNet-C [12]	-	-	-	image corruptions	relative mCE	✗	2019
ImageNet-A [13]	7,500	200	Internet	-	accuracy	✗	2019
ImageNet-R [11]	30,000	200	Internet	textures, styles	accuracy	✗	2020
FairNet (Ours)	**50,000**	**1,000**	**Internet**	**comprehensive**	**fairness**	**26**	

- More comprehensive attributes (external and internal)
- New evaluation metric to measure fairness

FairNet


moving
night
painting
human
snow
indoor
water
dead
rain
hen
wallaby
cut
human
many
plate
painting
rotten
night
bitten
occlusion
lemon
banana

Structural Causal Model



$X \rightarrow I \rightarrow Y$: models predict the label of the object with the observed image.

$X \rightarrow I \leftarrow$A: image is determined the object and its protected attributes

$A \leftarrow U \rightarrow X$: dataset bias causes the spurious correlation between attributes and objects.

Fairness metric

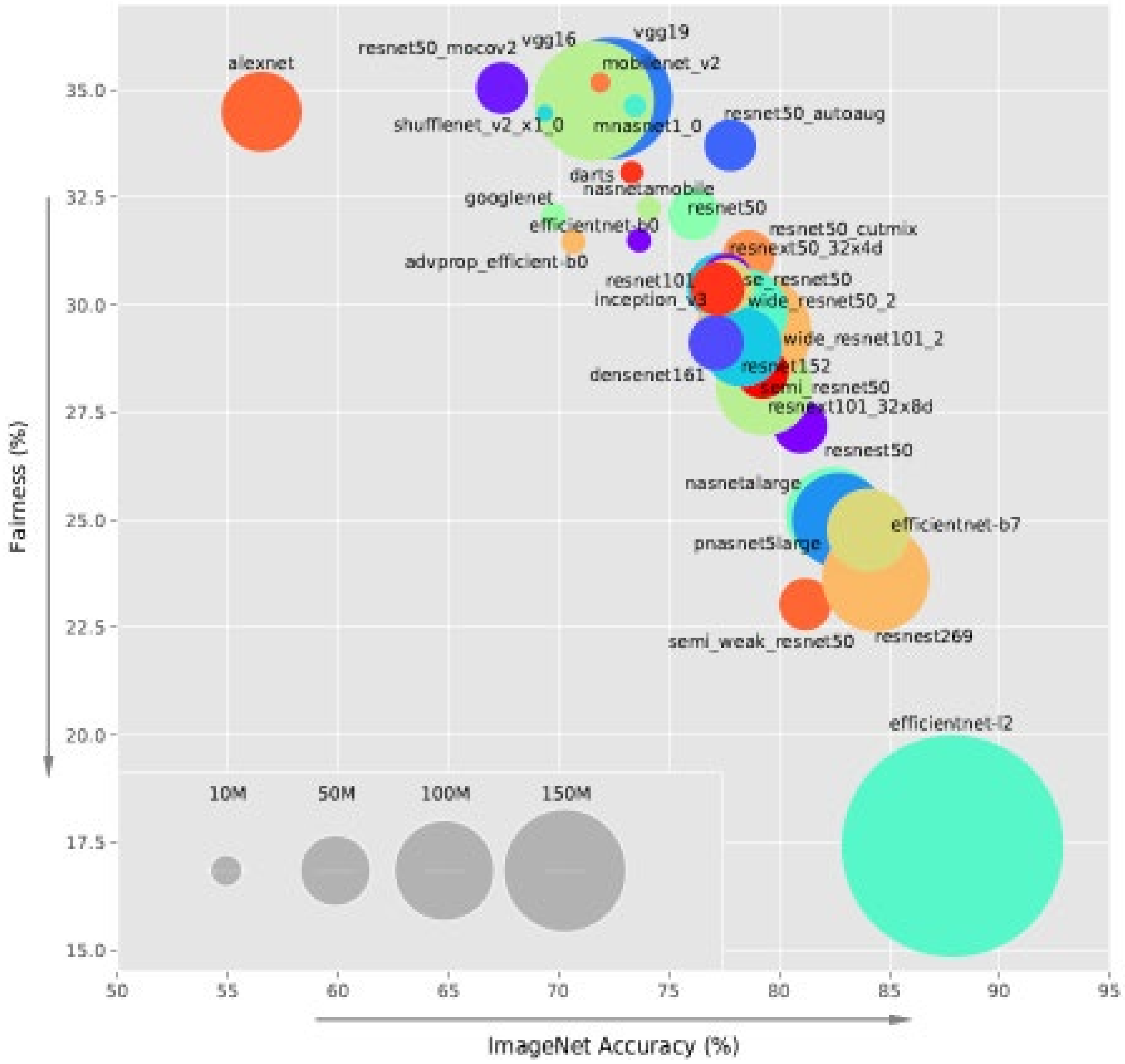
$$P(Y|I(do(A=a),do(X=x))) \\ = P(Y|I(do(A=\bar{a}),do(X=x)))$$

$$F(x,A) = \mathrm{Acc}_{\mathrm{I}}(x) - \mathrm{Acc}_{\mathrm{F}}(x, A=\bar{a})$$

$$F_{\mathrm{avg}} = \mathbb{E}_x \frac{1}{|\mathcal{A}_x|} \sum_{A \in \mathcal{A}_x} F(x,A),$$

$$F_{\max} = \mathbb{E}_x \max_{A \in \mathcal{A}_x} F(x,A)$$

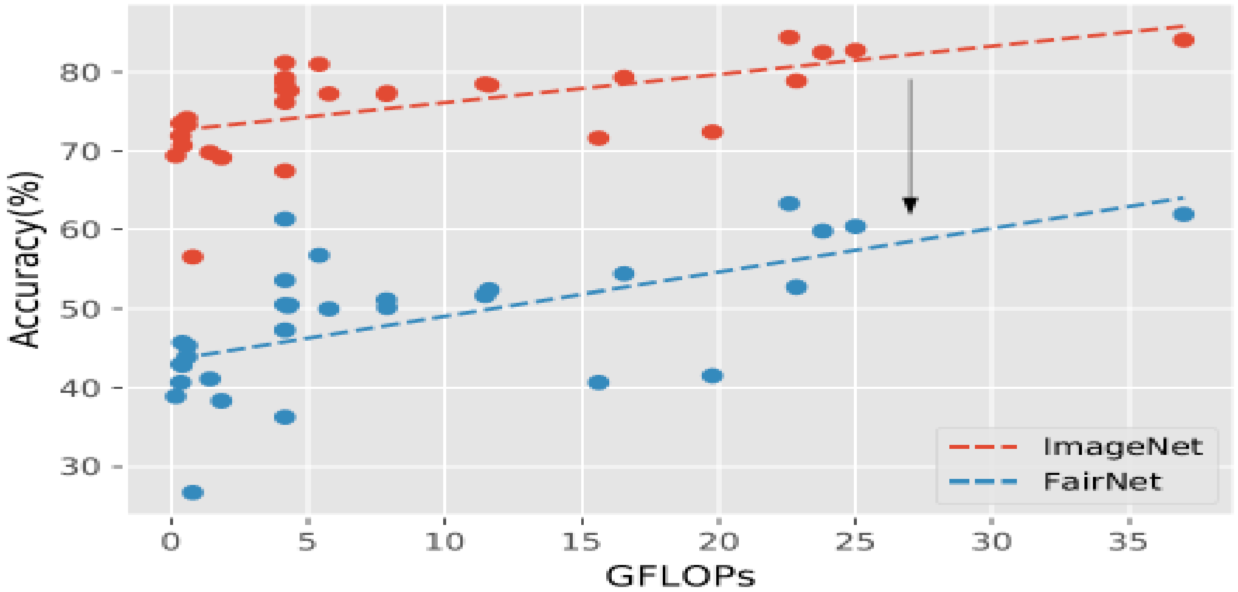
Benchmark on FairNet

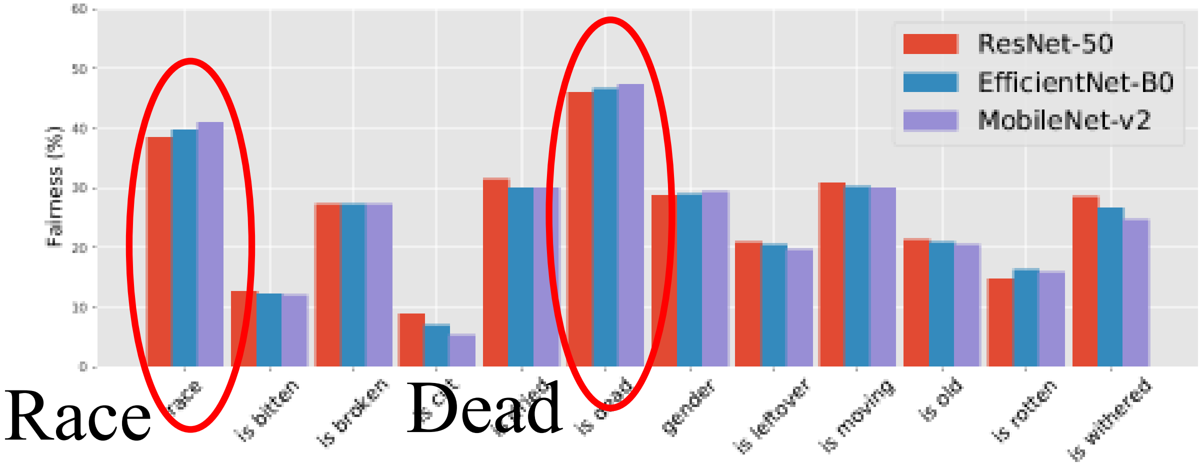


Benchmark on FairNet

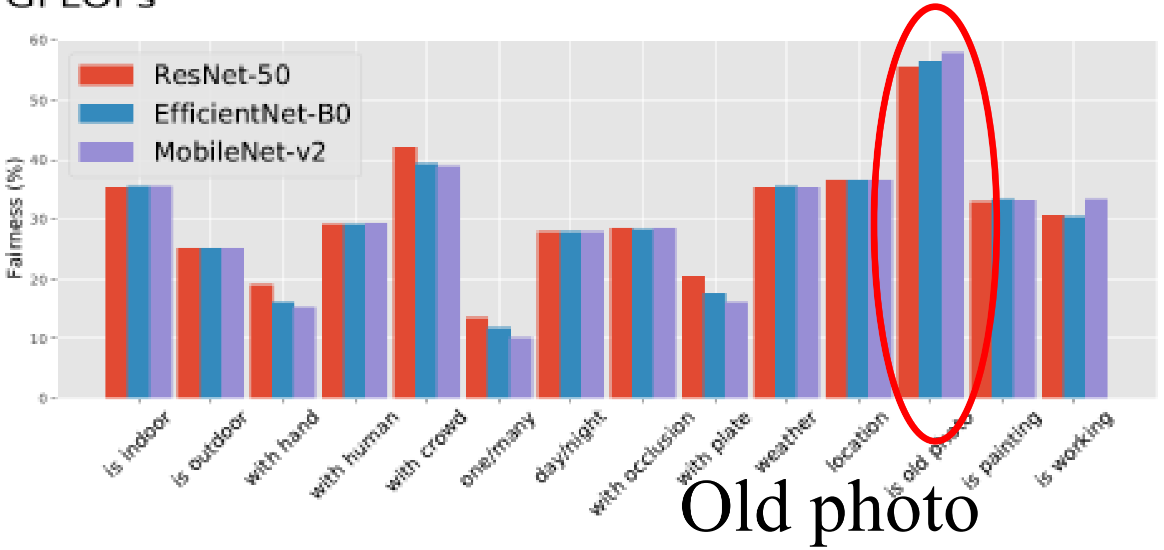
$Rank_I$	Model	Acc_I ↑	Acc_F ↑	Gap ↓	$Rank_F$	Δ Rank	F_{avg} ↓	F_{max} ↓
1	EfficientNet-L2 [32]	87.9	72.5	15.4	1	0	17.4	47.0
4	PNASNet-5-Large [18]	82.7	60.4	22.3	5	-1	25.0	55.7
5	NASNet-A-Large [46]	82.5	59.8	22.7	6	-1	25.1	55.5
6	ResNet-50 (IG-1B-Targeted) [40]	81.2	61.3	19.9	4	+2	23.0	53.1
11	ResNet-50 + CutMix [41]	78.6	50.5	28.1	14	-3	31.1	60.2
13	ResNet-152 [9]	78.3	52.4	26.0	11	+2	29.0	58.7
14	ResNet-50 + AutoAug [5]	77.8	47.3	30.5	19	-5	33.7	61.7
15	SE-ResNet-50 [14]	77.6	50.3	27.3	16	-1	30.4	59.5
16	ResNeXt-50 (32×4d) [39]	77.6	50.5	27.1	15	+1	30.6	59.0
19	DenseNet-161 [15]	77.1	51.1	26.0	13	+6	29.1	57.7
20	ResNet-50 [9]	76.1	47.3	28.8	19	+1	32.1	60.1
21	NASNet-A-Mobile [46]	74.1	45.3	28.8	22	-1	32.2	59.1
22	EfficientNet-b0 [32]	73.6	45.7	27.9	21	+1	31.5	58.2
23	MNASNet 1.0 [31]	73.5	43.0	30.5	24	-1	34.6	60.3
24	DARTS [19]	73.3	43.9	29.4	23	+1	33.1	60.2
25	VGG-19 [27]	72.4	41.5	30.9	26	-1	34.8	60.7
26	MobileNet-v2 [26]	71.9	40.7	31.2	28	-2	35.2	59.9
27	VGG-16 [27]	71.6	40.6	30.9	29	-2	34.7	59.9
28	EfficientNet-B0 + AdvProp [38]	70.7	42.8	27.8	25	+3	31.5	57.4
29	GoogleNet [30]	69.8	41.1	28.7	27	+2	32.1	56.9

Fairness Analysis





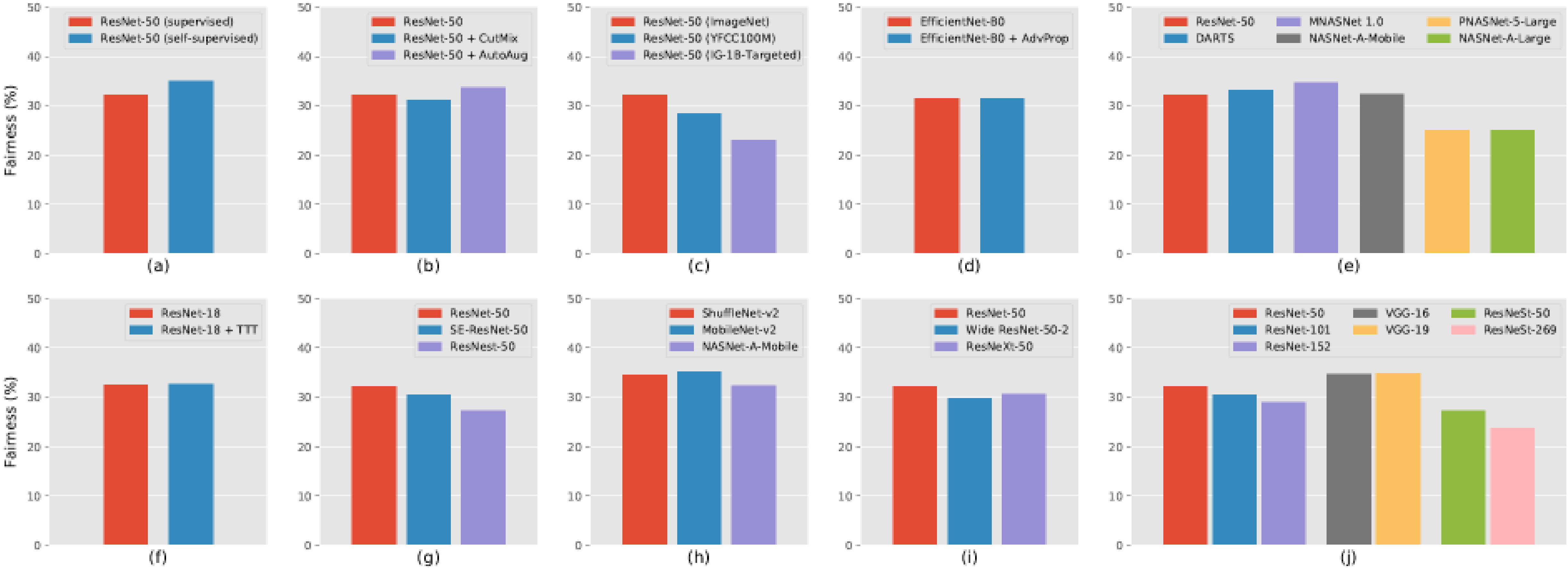
(a) Fairness of the internal attributes



(b) Fairness of the external attributes

What Can Help to Improve Fairness

- Self supervised ✗
- Data argumentation ✗
- Adversarial training ✗
- Larger training datasets ✓
- Neural architecture search ✗
- Test-time training ✗
- Self-attention ✓
- Big model ✓



Contents

Future Work

- Rely on high-level semantic representation
  - Extend the causal learning into representation learning
  - Learn latent causal representation
- Rely on the strong human prior
  - Adaptively discover the causal relations
  - Jointly learn to discover, represent, and analyze
- Without the commonsense knowledge
  - Use the commonsense knowledge to build counterfactuals

Thanks for your listening

https://chengy12.github.io/