SMU SINGAPORE MANAGEMENT UNIVERSITY | School of Computing and Information Systems

ICLR 2022 Workshop on the Elements of Reasoning:
Objects, Structure, and Causality (OSC)

Causality in Computer Vision

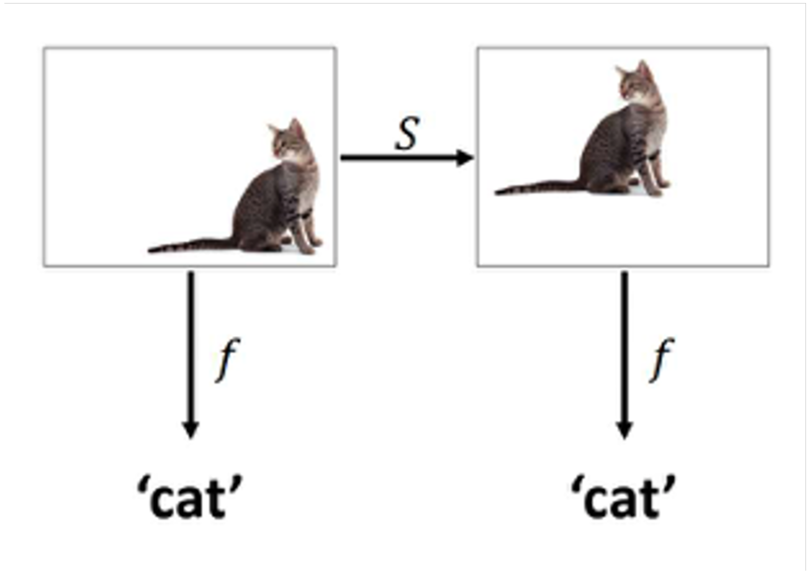
Invariant Learning with Insufficient Data

Qianru SUN

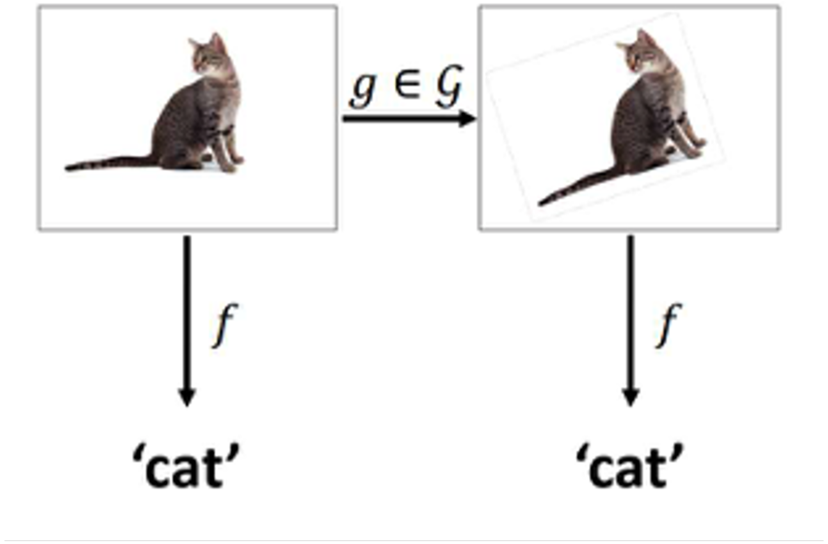
29 Apr 2022

Invariant Learning?

Invariant to Shift

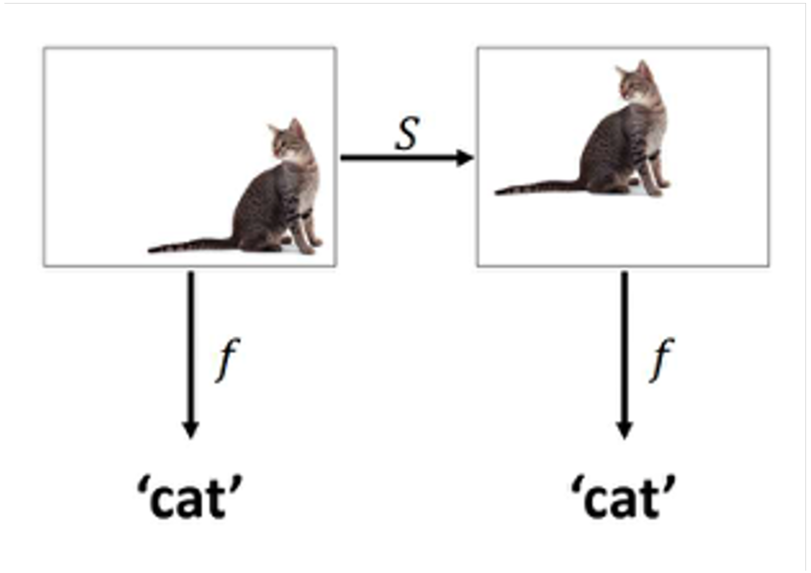


Invariant to Rotation



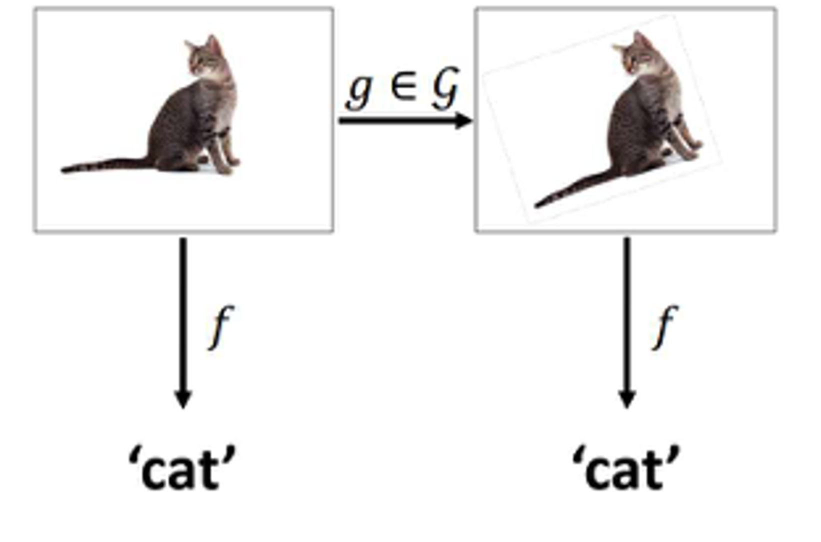
Invariant Learning with Sufficient Data?

Invariant to Shift



Range of Shift covers every position

Invariant to Rotation

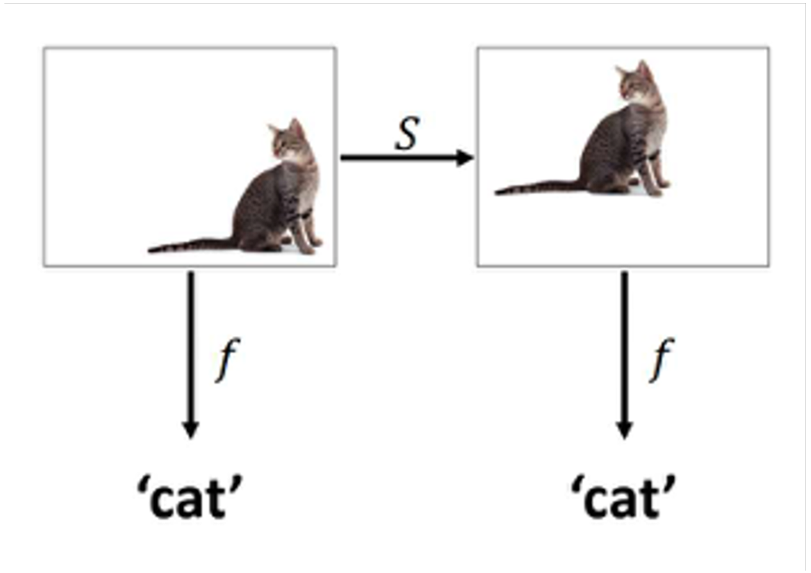


Range of Rotation covers every view angle

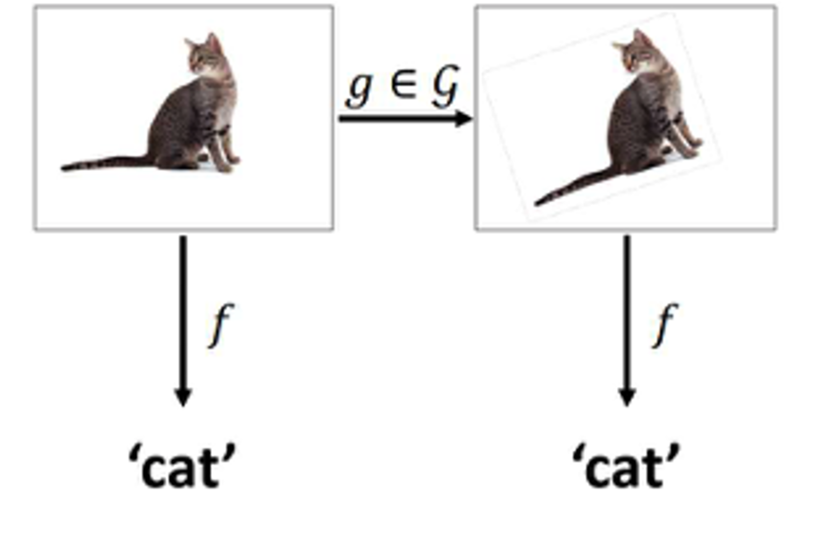
Invariant Learning with Sufficient Data?

Key Property: Disentangling Class, Position, and View Angle

Invariant to Shift



Invariant to Rotation



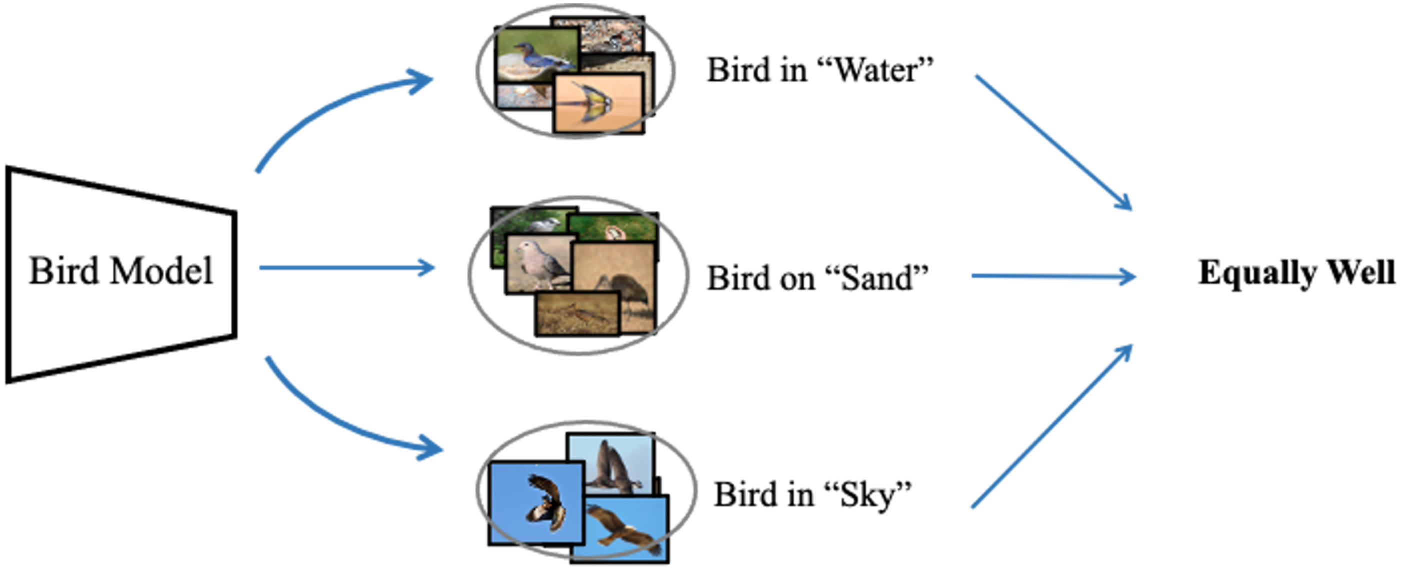
Range of Shift covers every position

Range of Rotation covers every view angle

Invariant Learning with Sufficient Data?

Key Property: Disentangling Class, Position, and View Angle

This is, however, not easy to achieve in real-world data …



not always true

Invariant Learning with Sufficient Data?

Key Property: Disentangling Class, Position, and View Angle

This is, however, not easy to achieve in real images of complex foregrounds and backgrounds …

missing "Sand" data in training, while can be tested

for example

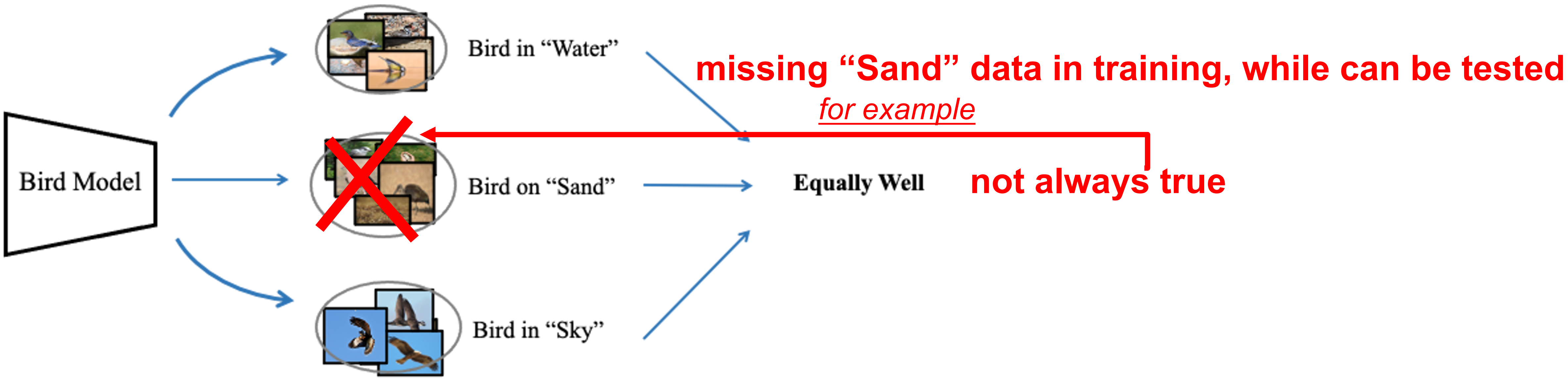
not always true

Training data is often Insufficient.

Invariant Learning with Insufficient Data

Key Property: Disentangling Class, Position, and View Angle

This is, however, not easy to achieve in real images of complex foregrounds and backgrounds …



Training data is often Insufficient.

Invariant Learning with Insufficient Data (OOD Data)

When Model Fails in OOD Generalization

For example, the model's prediction is misled by biased backgrounds:

Training

Camel + Sand

Cow + Grass

Testing

Camel + Grass

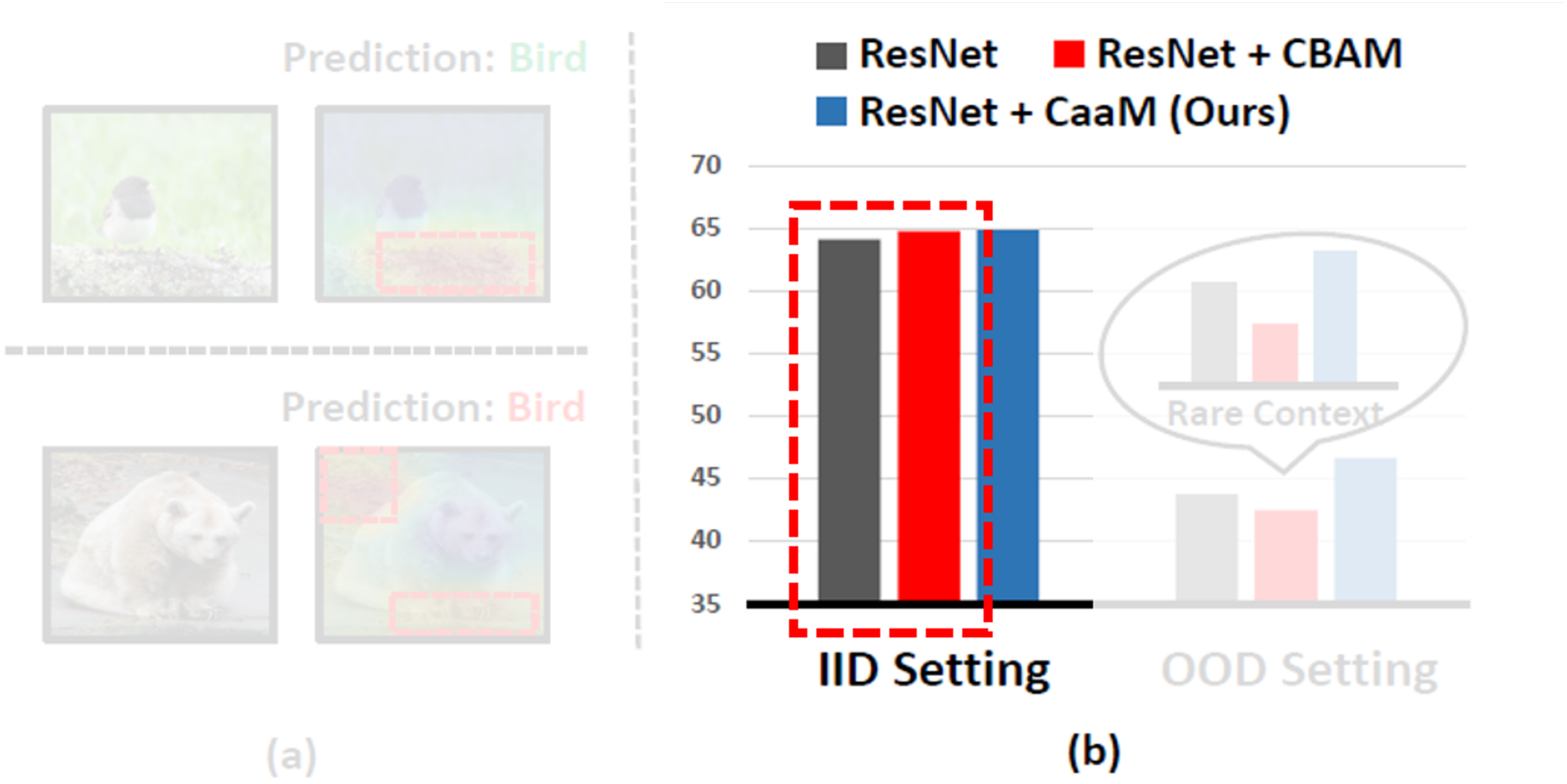
Prediction: Cow

Cow + Sand

Prediction: Camel

When Model Fails in OOD Generalization

Another example is the model attention is biased to backgrounds:

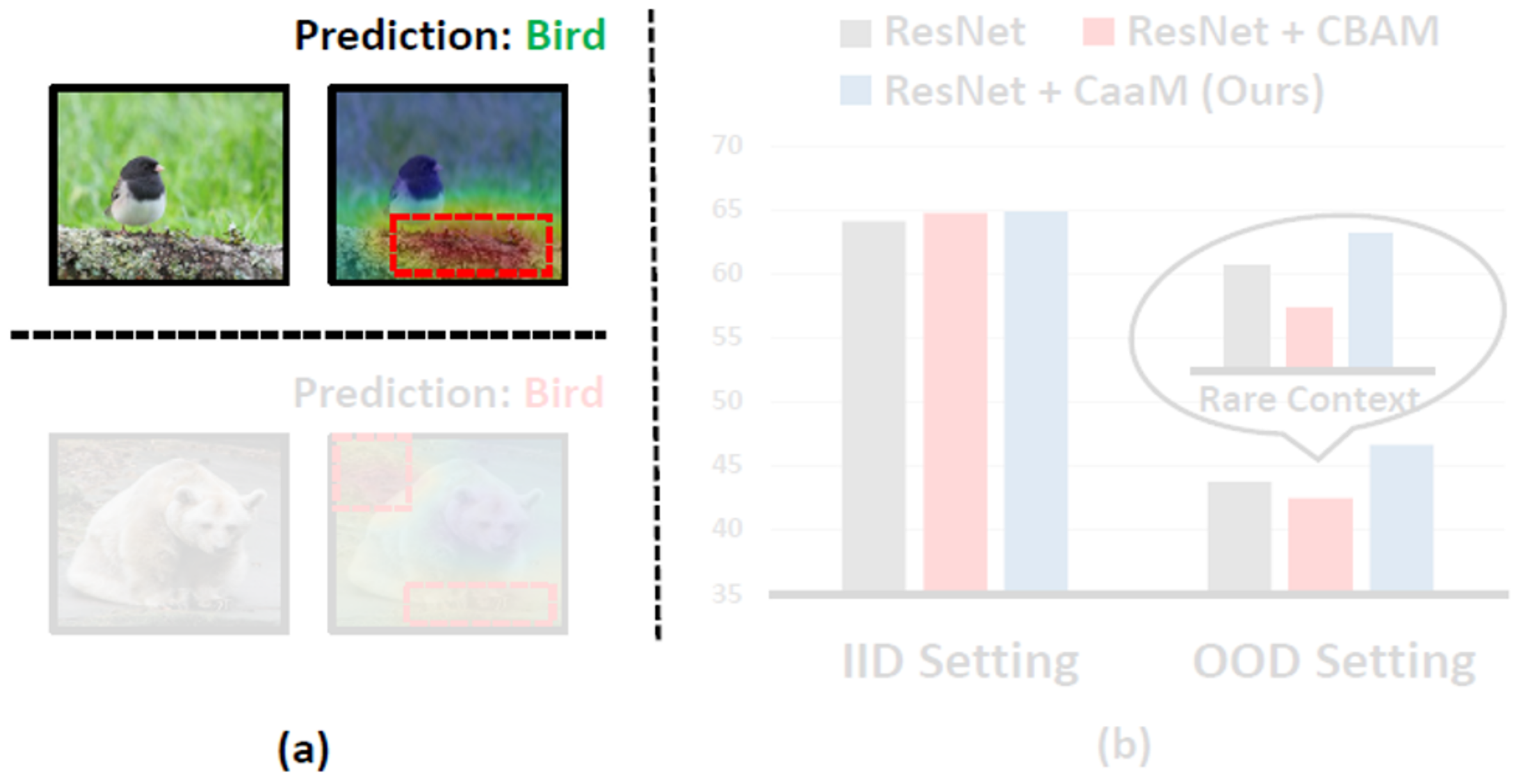


"Attention is all you need"

When Model Fails in OOD Generalization

Another example is the model attention is biased to backgrounds:

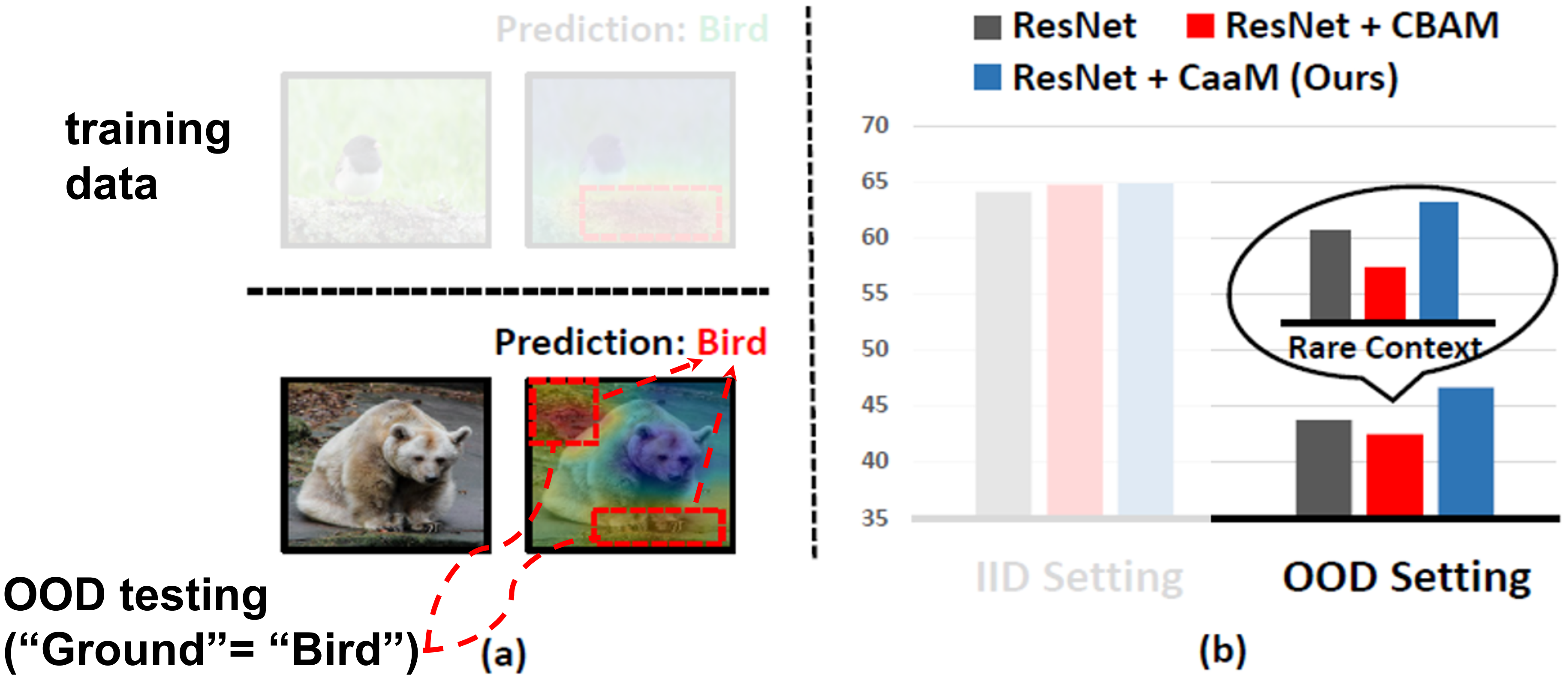
training data



"Attention is all you need"?

When Model Fails in OOD Generalization

Another example is the model attention is biased to backgrounds:



More failures in rare backgrounds

“Attention is all you need”?

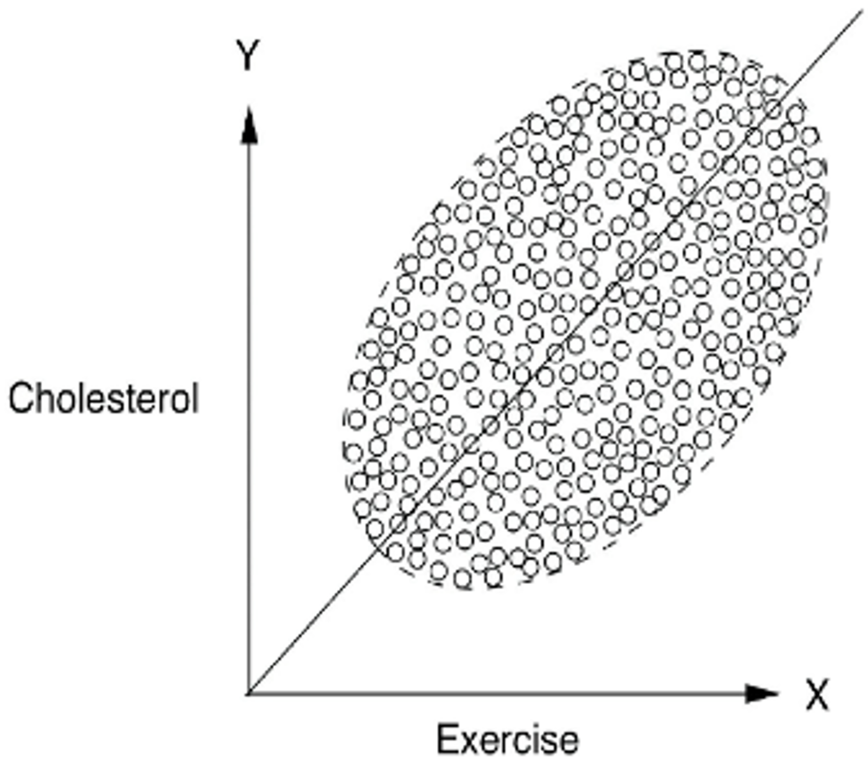
What is the reason?

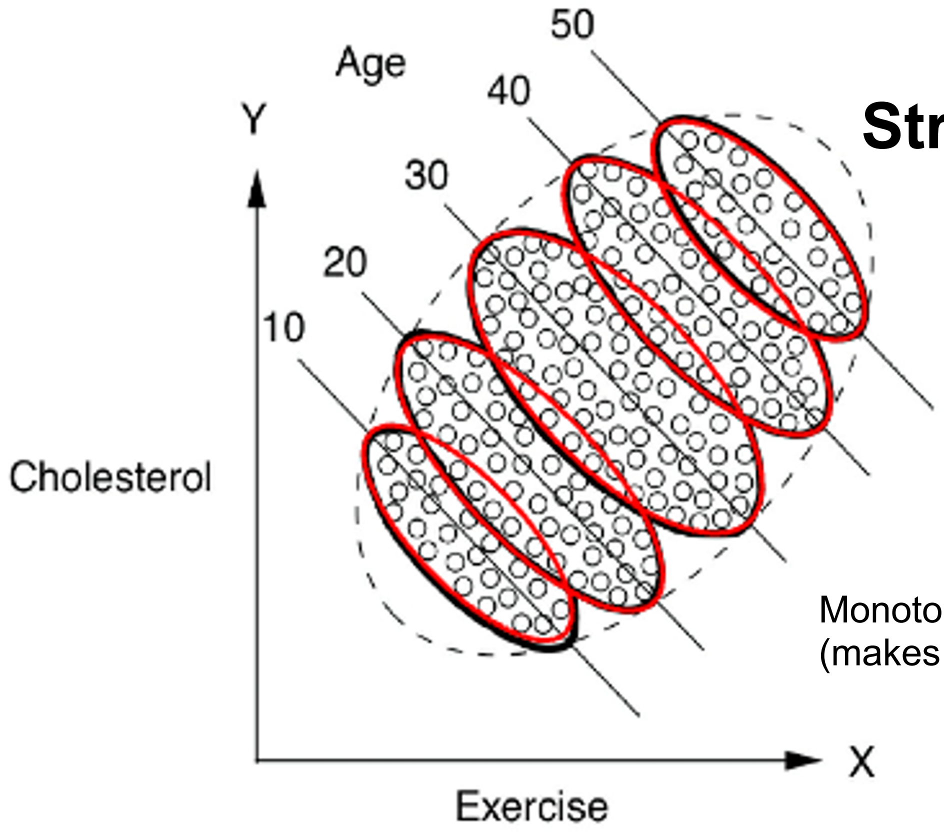
Models learn only correlations $P(Y|X)$

How to solve the issue?

Causal Intervention (from correlation to causality):
$P(Y|X) \rightarrow P(Y|do(X))$

Causal Intervention: P(Y│X)→P(Y|do(X))





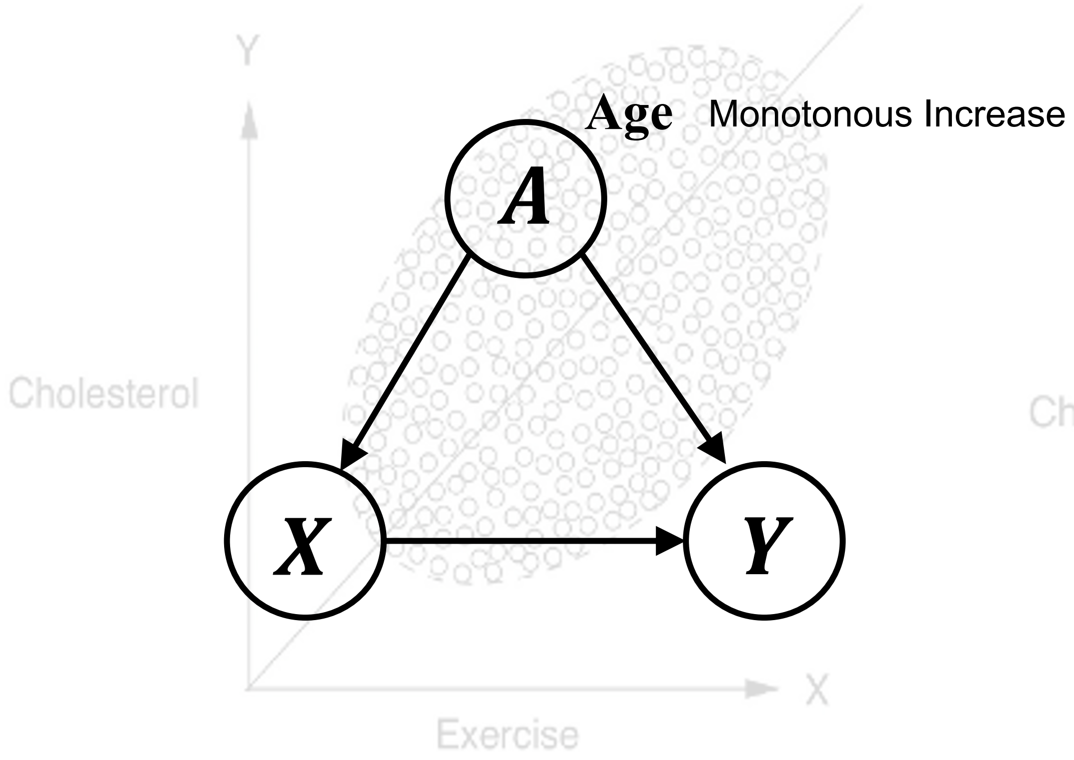
Stratification

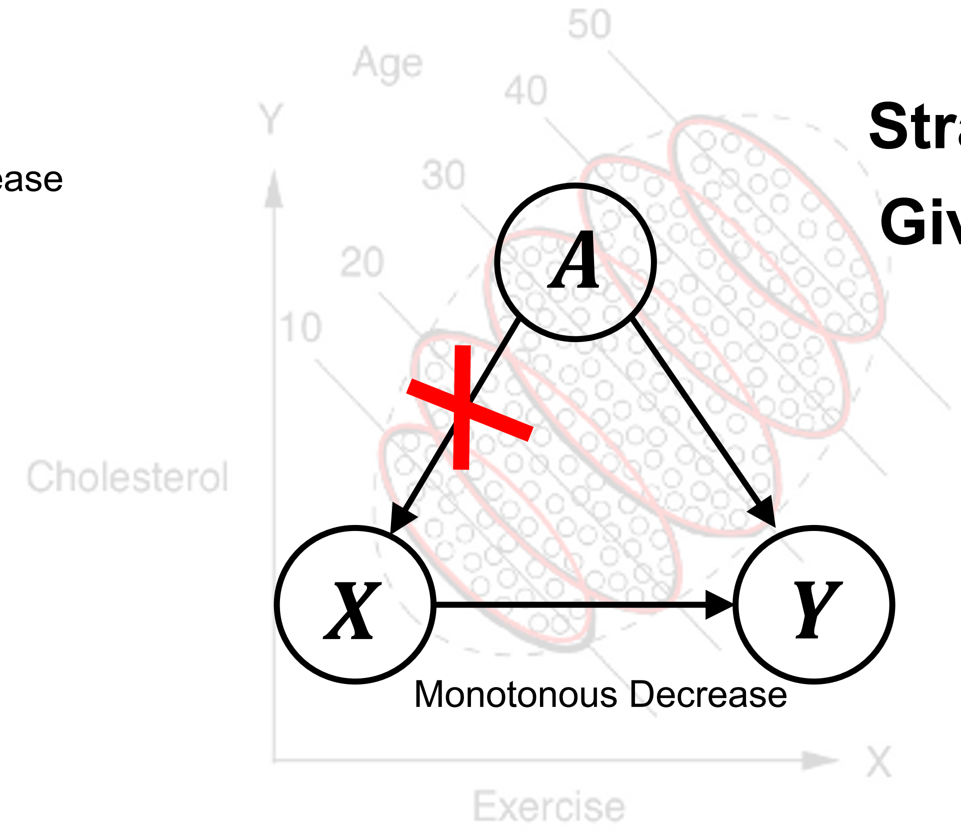
Monotonous Decrease
(makes more sense)

P(Y│X)

P(Y|do(X)) ✔

Causal Intervention: P(Y | X)→P(Y|do(X))





Stratification:

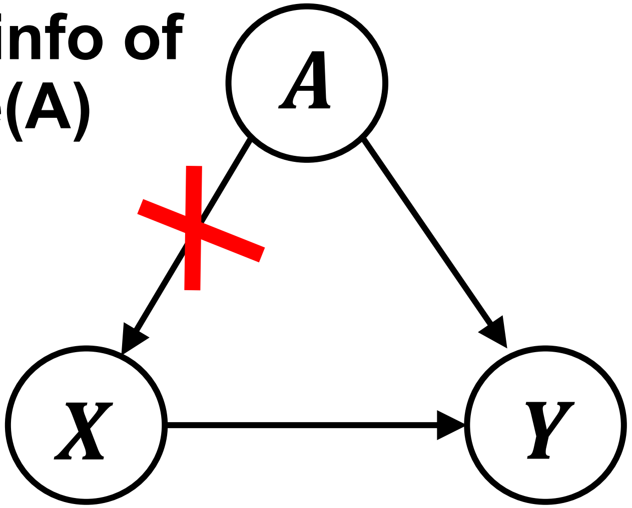
Given info of Age(A)

P(Y | X) **P(Y|do(X))**

Causal Intervention: P(Y│X)→P(Y|do(X))

Stratification:

Given info of Age(A)



$$P(Y|do(x)) = \sum_{\boxed{a}} P(Y|X, A = a)\underline{P(A = a)}$$

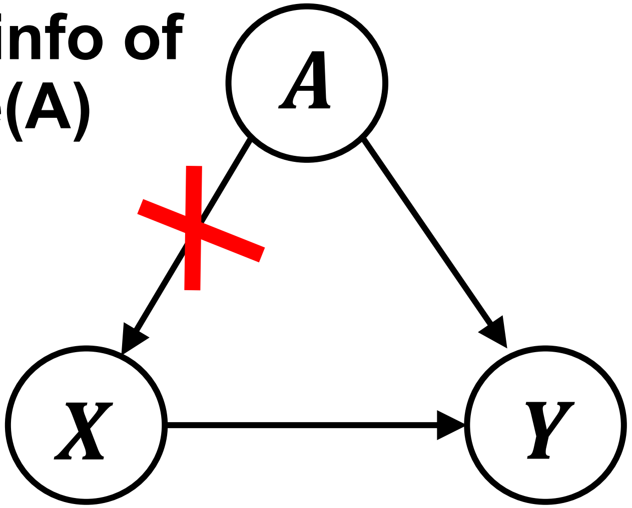
Stratify the Confounder

Weighted by a sample-agnostic prior

Causal Intervention: P(Y | X)→P(Y|do(X))

Stratification:

Given info of Age(A)



How to implement Stratification in Computer Vision?

$$P(Y|do(x)) = \sum_{\boxed{a}} P(Y|X, A = a)\underline{P(A = a)}$$

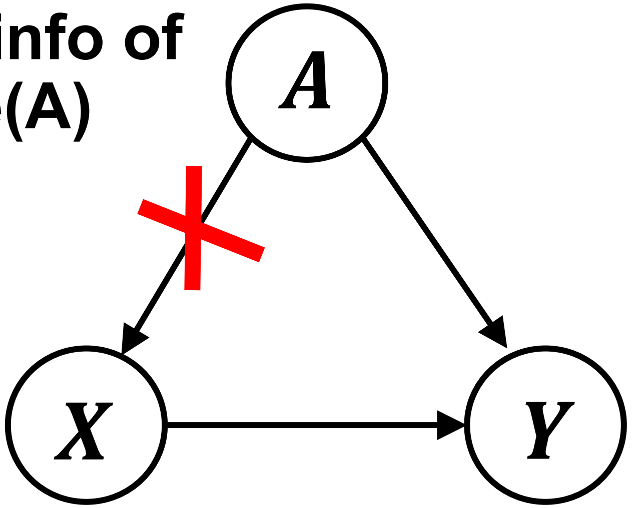
Stratify the Confounder

Weighted by a sample-agnostic prior

Causal Intervention: P(Y|X)→P(Y|do(X))

Stratification:

Given info of Age(A)



How to implement Stratification in Computer Vision?

-- where we do not have the definition and representation of confounders

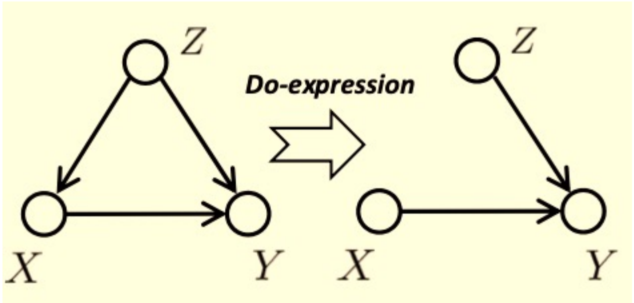
$$P(Y|do(x)) = \sum_{\boxed{a}} P(Y|X, A = a)\underline{P(A = a)}$$

Stratify the Confounder

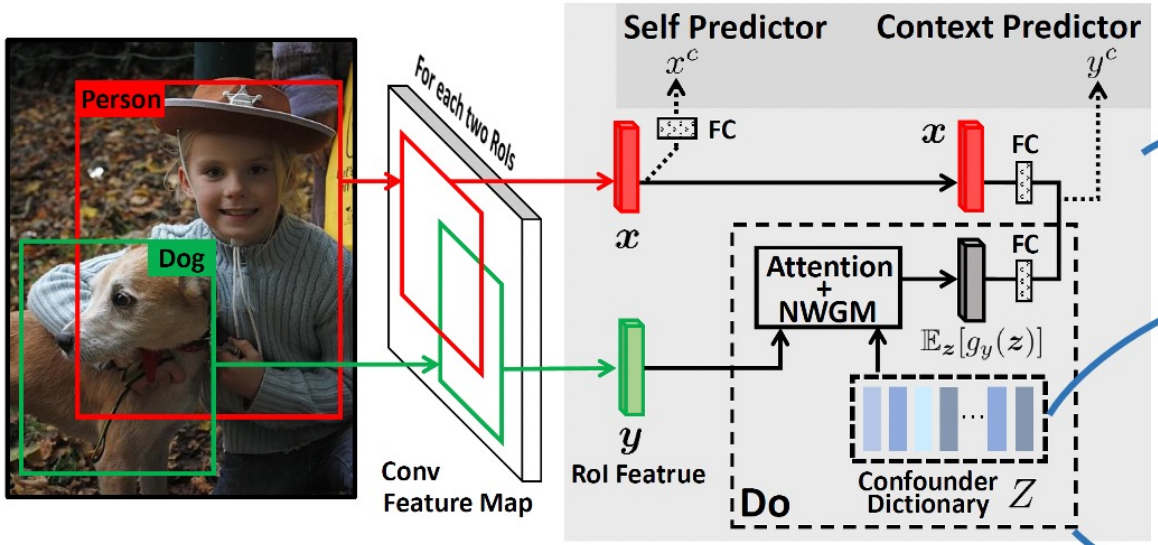
Weighted by a sample-agnostic prior

Causal Intervention: P(Y | X)→P(Y|do(X)) in Computer Vision

Causal Intervention by
debiasing from all Contextual Objects---a Confounder set



X and Y are both image areas, Z is the confounder



Context Label Prediction Proxy Task

Use an object dictionary to denote confounder Z

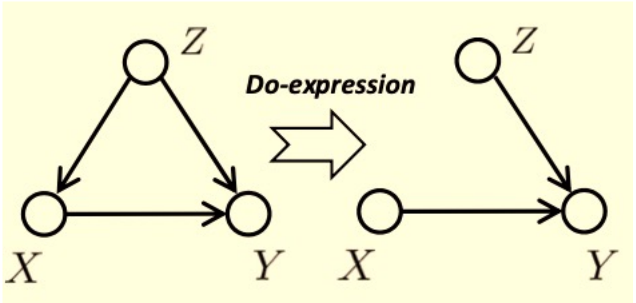
$$P(\text{Dog}|\text{do}(\text{Person})) = \sum_{z \in Z} P(\text{Dog}|\text{Person}, z)P(z)$$

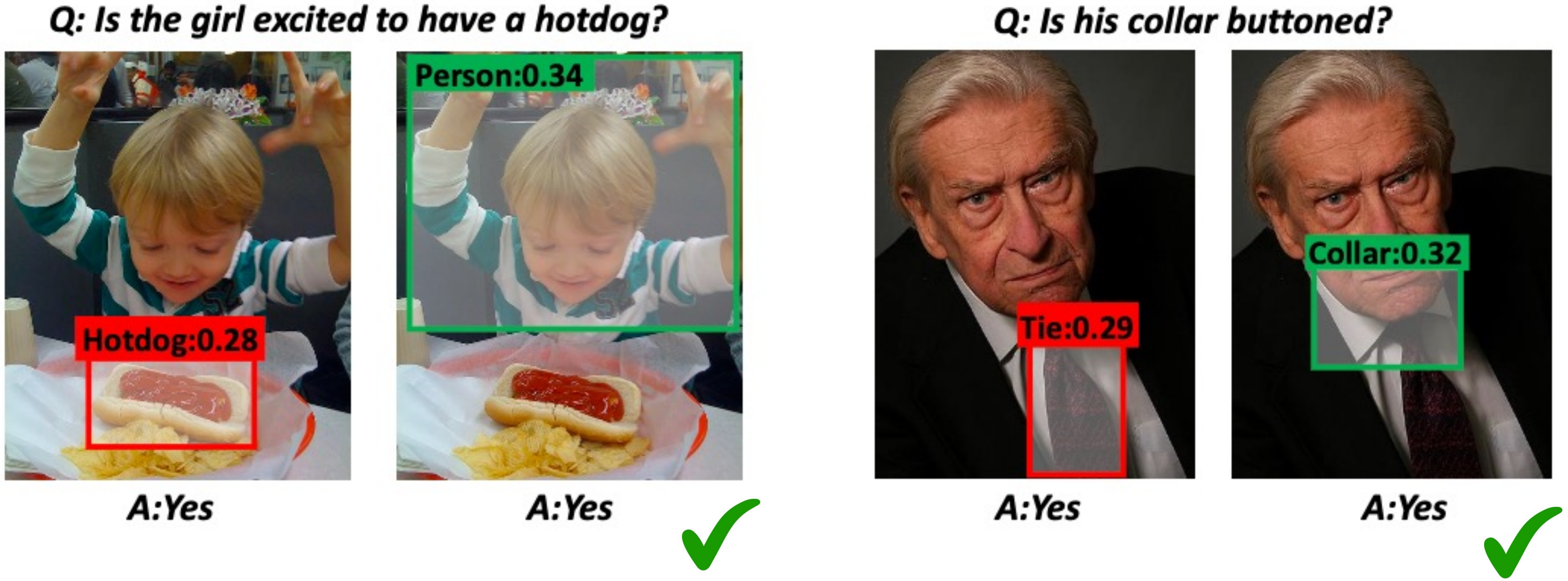
Wang, et al. "Visual commonsense r-cnn." CVPR 2020.

Causal Intervention: P(Y│X)→P(Y|do(X)) in Computer Vision

Causal Intervention by

debiasing from all Contextual Objects---a Confounder set





Wang, et al. "Visual commonsense r-cnn." CVPR 2020.

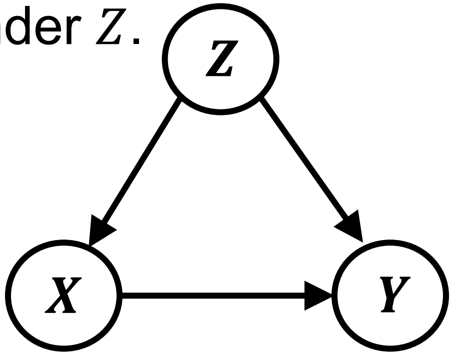
Causal Intervention: $P(Y \mid X) \rightarrow P(Y|do(X))$ in Computer Vision?

Causal Intervention: any issues?

Causal Intervention is designed for inference. When used for training:

- Non-positivity: X may never appear with some Z in training set.
  e.g., how would you reweight the "black swan" if there is even no "black swan" sample?

- The derivation assumes $P(Y|X)$ to be independent causal mechanisms. That means Z should not influence the $P(Y|X)$ in training. The model should perform **equally** well under different confounder Z.



Parascandolo, et al. "Learning independent causal mechanisms." ICML 2018.

Causal Intervention: P(Y│X)→P(Y|do(X)) in Computer Vision?

Causal Intervention: any issues?

Causal Intervention is designed for inference. When used for training:

- Non-positivity: X may never appear with some Z in training set.
  e.g., how would you reweight the “black swan” if there is even no “black swan” sample?

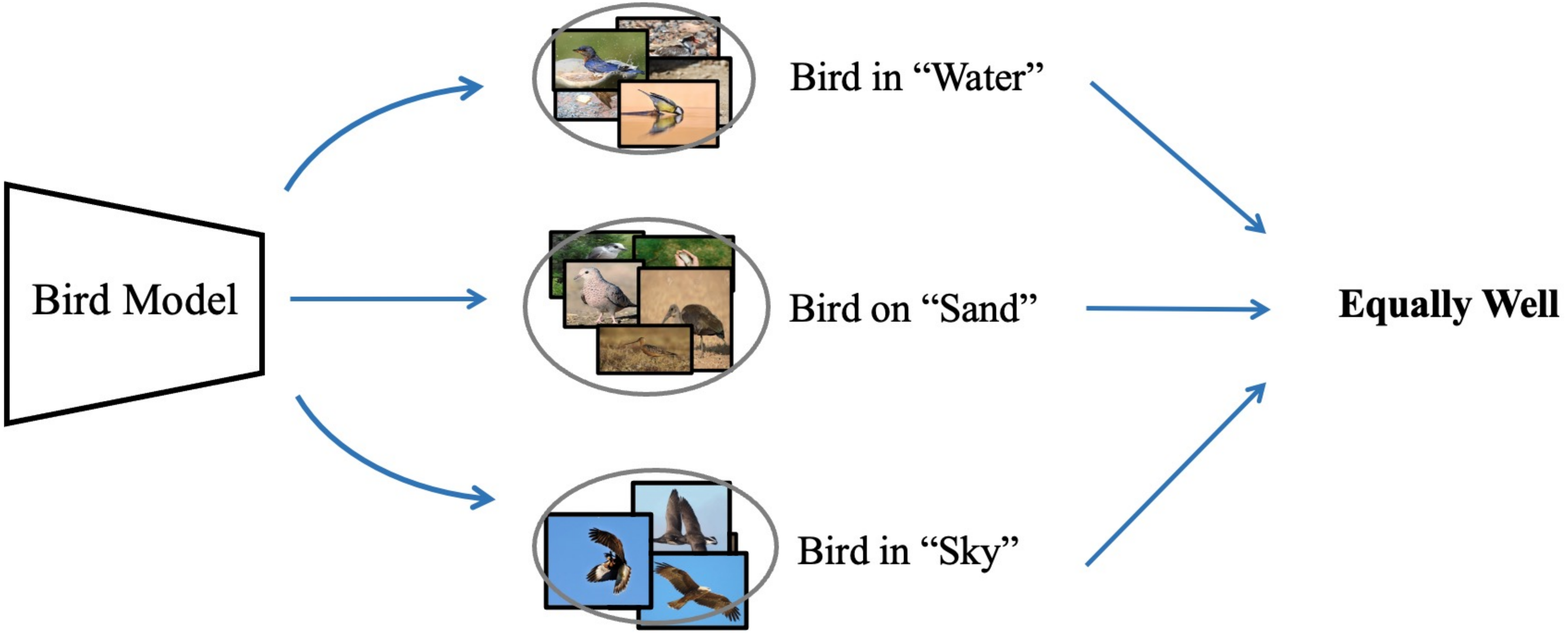
- The derivation assumes $P(Y|X)$ to be independent causal mechanisms. That means Z should not influence the $P(Y|X)$ in training. The model should perform **equally** well under different confounder Z.

Invariant Learning   Insufficient Data

Parascandolo, et al. "Learning independent causal mechanisms." ICML 2018.

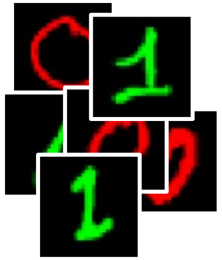
Invariant Learning in Insufficient Data

The model should perform **equally** well under different confounder Z.



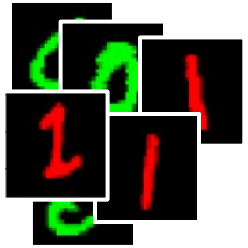
Invariant Learning in Insufficient Data

A toy dataset:



Training Dataset

$$P(Y = 1|\text{Green}) = 0.8$$



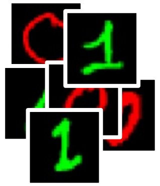
Testing Dataset

$$P(Y = 1|\text{Red}) = 0.8$$

OOD Generalization Problem !

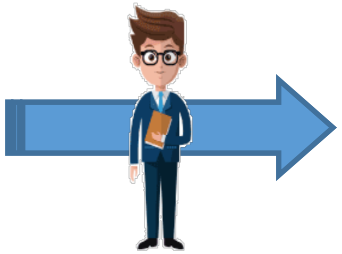
Invariant Learning in Insufficient Data: Invariant Risk Minimization (IRM)

IRM Loss for tackling this problem:

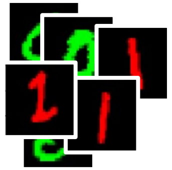


Training Dataset

$P(Y = 1|\text{Green}) = 0.8$



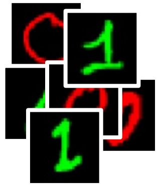
Human Expert



Testing Dataset

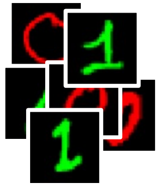
$P(Y = 1|\text{Red}) = 0.8$

Environment e_1



$P(Y = 1|\text{Green}) = 0.7$

Environment e_2



$P(Y = 1|\text{Green}) = 0.9$

Training with IRM Loss

IRM Loss will only find the ***invariant*** factor and thus remove the color factor

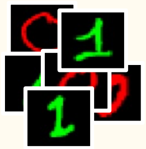
Arjovsky, et al. "Invariant risk minimization." *arXiv preprint arXiv:1907.02893* (2019).

Invariant Learning in Insufficient Data: Invariant Risk Minimization (IRM)

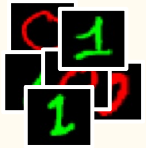
IRM Loss for tackling this problem: by constructing good env

Environment e_1 Environment e_2

Manually-constructed Environment



$P(Y = 1|\text{Green}) = 0.7$

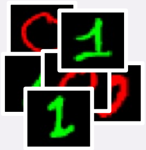


$P(Y = 1|\text{Green}) = 0.9$

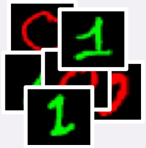
IRM Loss → **Digit Model** ✓

Good: Split the data w.r.t the Bias Factor (Color)

Random Environment



$P(Y = 1|\text{Green}) = 0.8$



$P(Y = 1|\text{Green}) = 0.8$

IRM Loss → **Digit & Color Model** ✗

Arjovsky, et al. "Invariant risk minimization." *arXiv preprint arXiv:1907.02893* (2019).

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

IRM vs. CaaM:

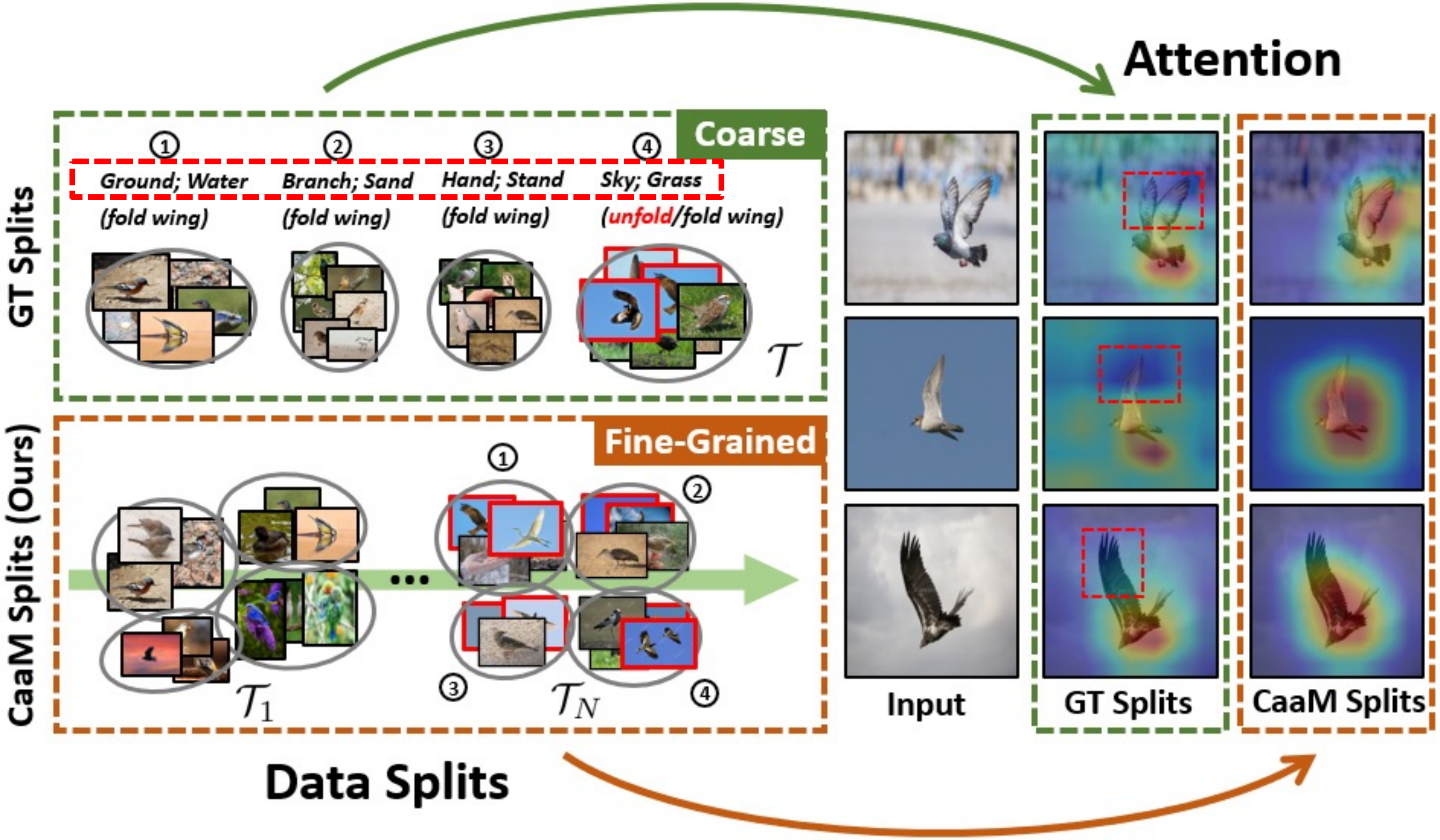
IRM	CaaM
Single	Multiple
Coarse	Fine-Grained
Manually-constructed	Automatically-constructed
Environments	***Environments***

To learn invariances across environments, find a data representation such that the optimal classifier on top of that representation matches for all environments.

Wang, et al. "Causal attention for unbiased visual recognition." *ICCV* (2021).

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

How to understand CaaM:

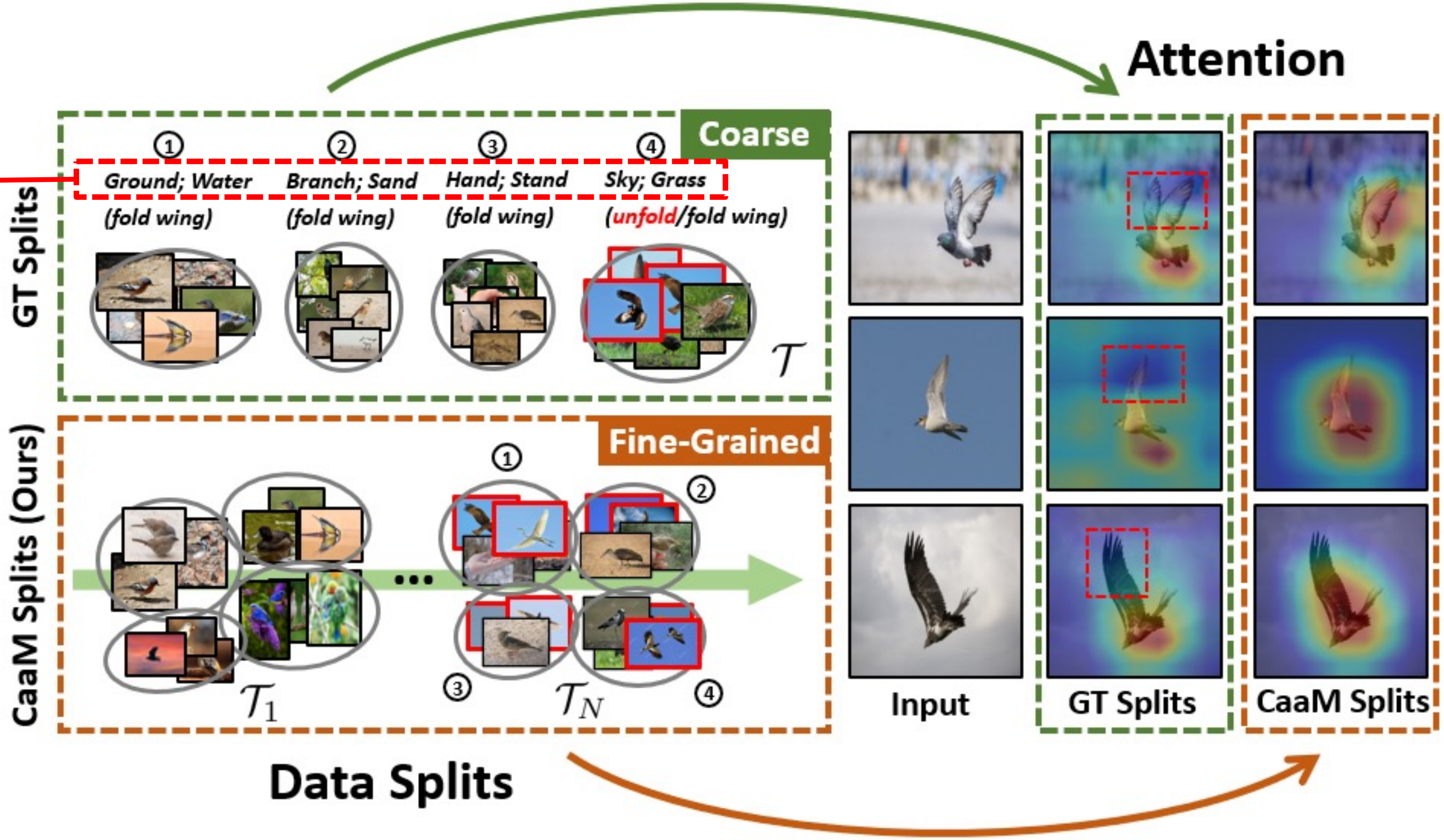


Wang, et al. "Causal attention for unbiased visual recognition." *ICCV* (2021).

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

How to understand CaaM:

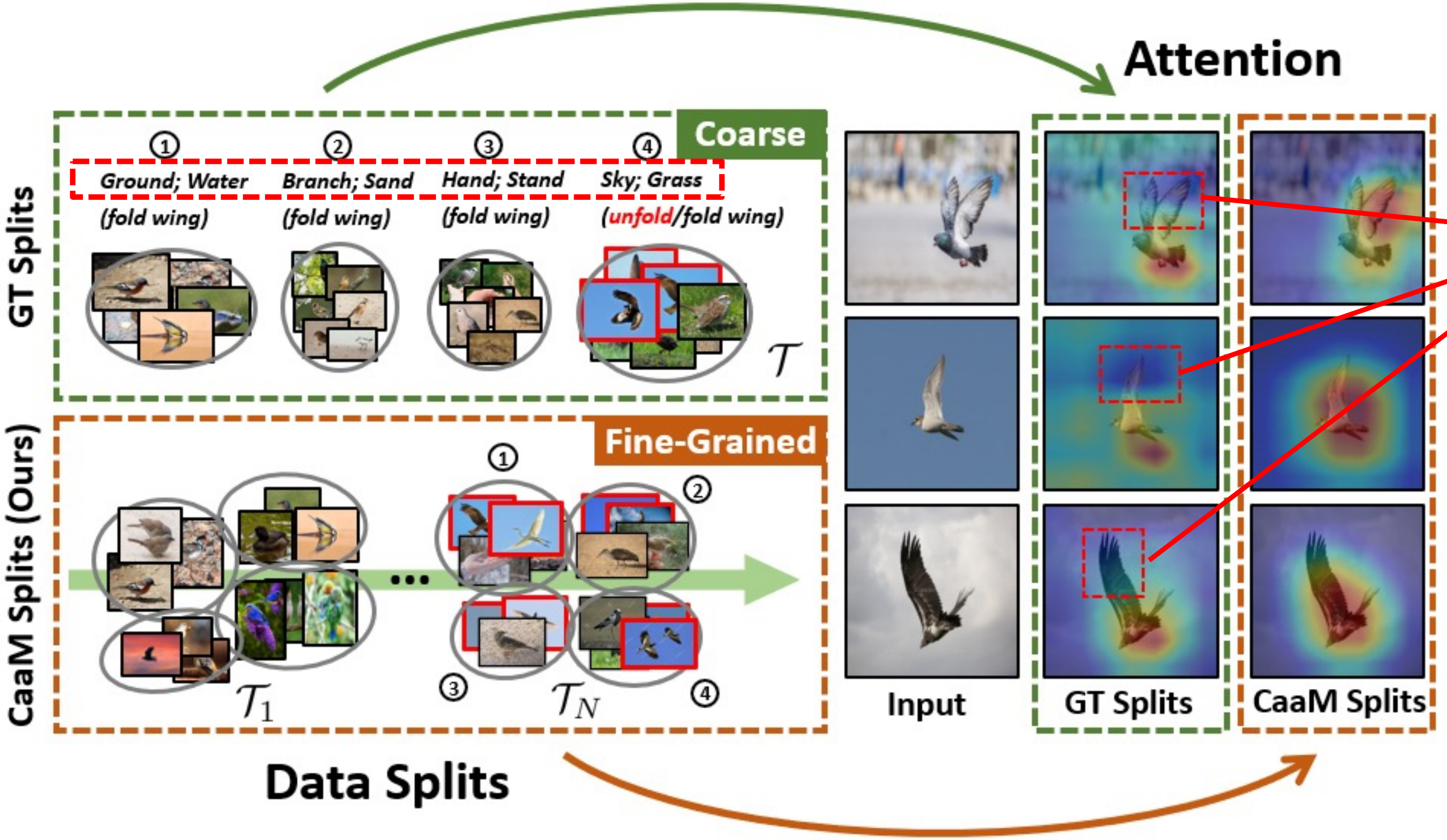
IRM: Group the images according to the ground truth context (single split)



Wang, et al. "Causal attention for unbiased visual recognition." *ICCV* (2021).

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

How to understand CaaM:

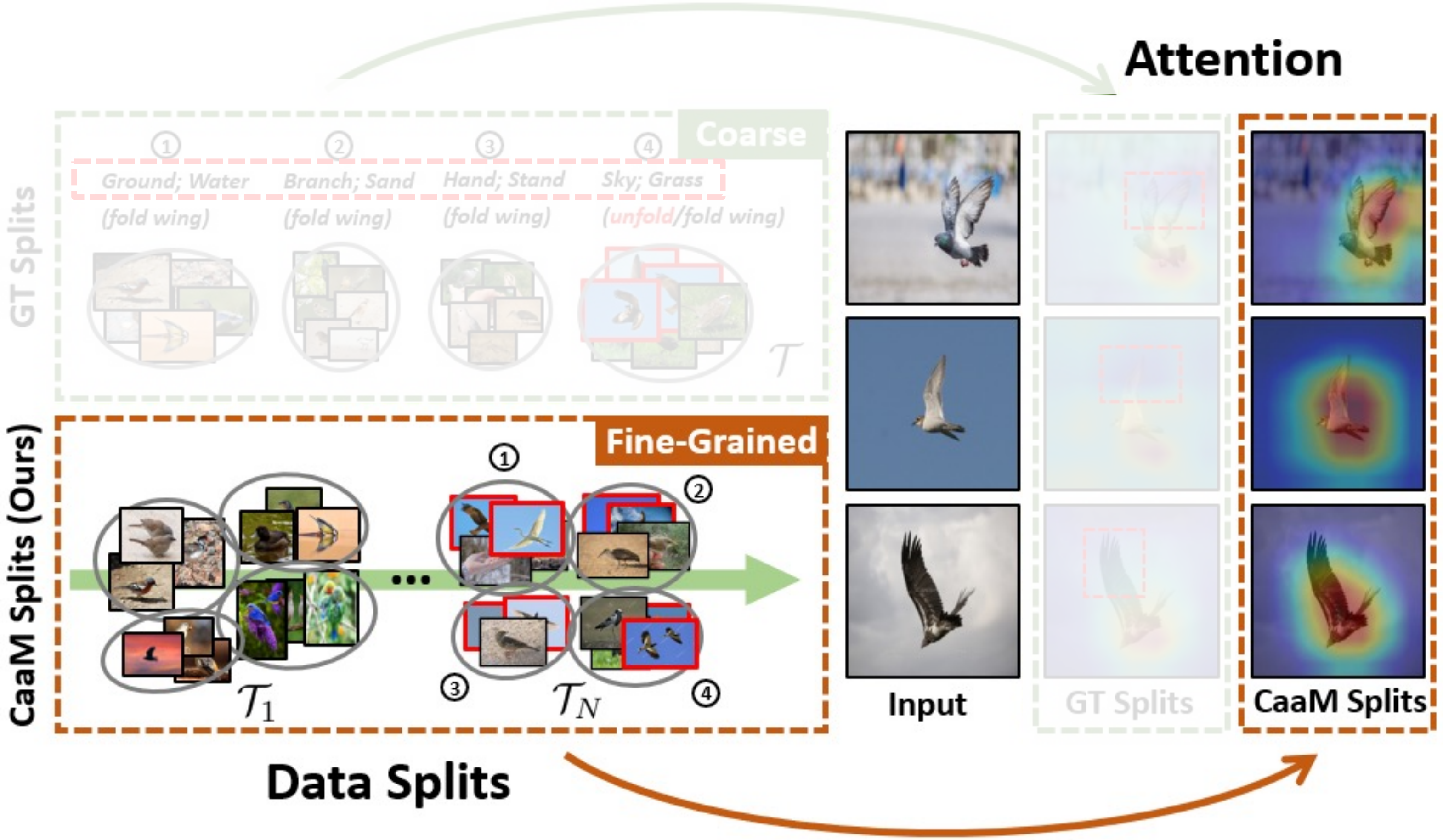


The causal feature of "bird" – "wing" – is also removed

Wang, et al. "Causal attention for unbiased visual recognition." *ICCV* (2021).

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

How to understand CaaM:

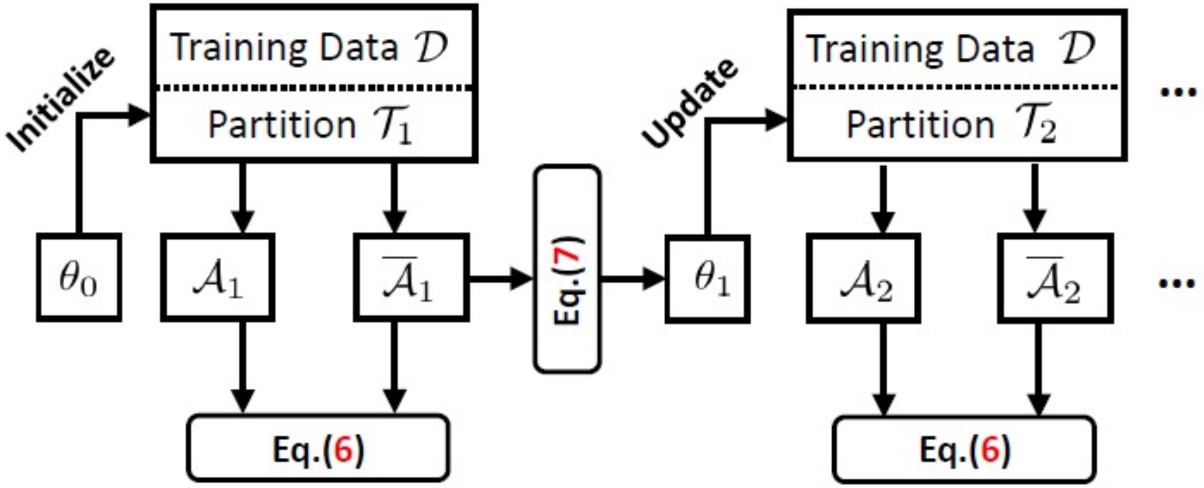


More accurate attention

Wang, et al. "Causal attention for unbiased visual recognition." *ICCV* (2021).

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

How to implement CaaM:



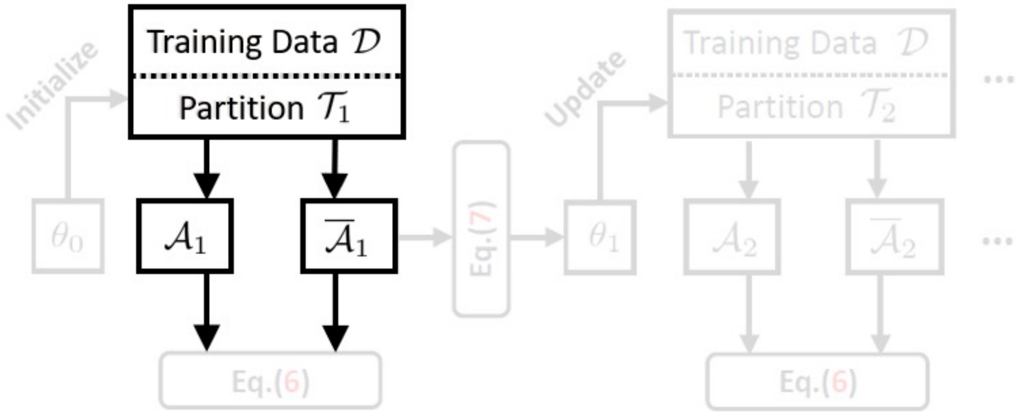
A: Causal Effect

$\bar{A}$: Confounding Effect

Wang, et al. "Causal attention for unbiased visual recognition." *ICCV* (2021).

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

How to implement CaaM:



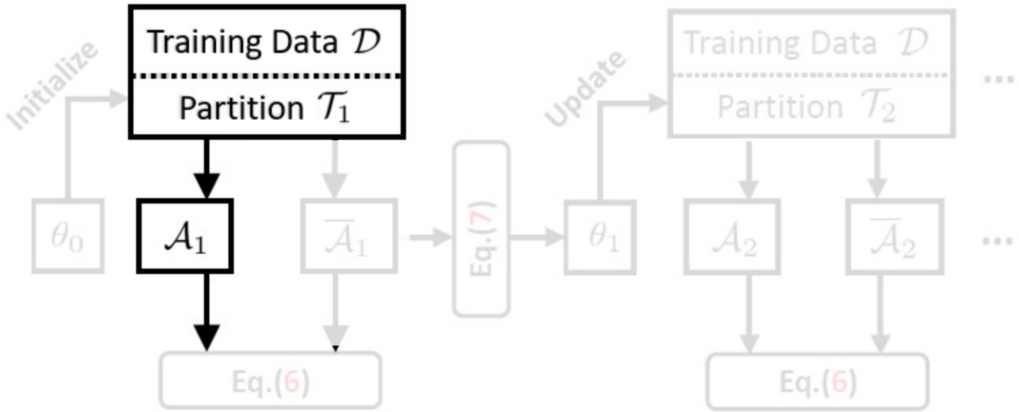
Cross-Entropy Loss (XE): This loss is to ensure that A and $\bar{A}$ combined will capture the total effects

$$\mathrm{XE}(f, \tilde{x}, \mathcal{D}) = \mathbb{E}_{(x,y)\in\mathcal{D}}\, \ell\left(f(\tilde{x}), y\right) \qquad \tilde{x} = \mathcal{A}(x) \circ \overline{\mathcal{A}}(x)$$

Wang, et al. "Causal attention for unbiased visual recognition." *ICCV* (2021).

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

How to implement CaaM:



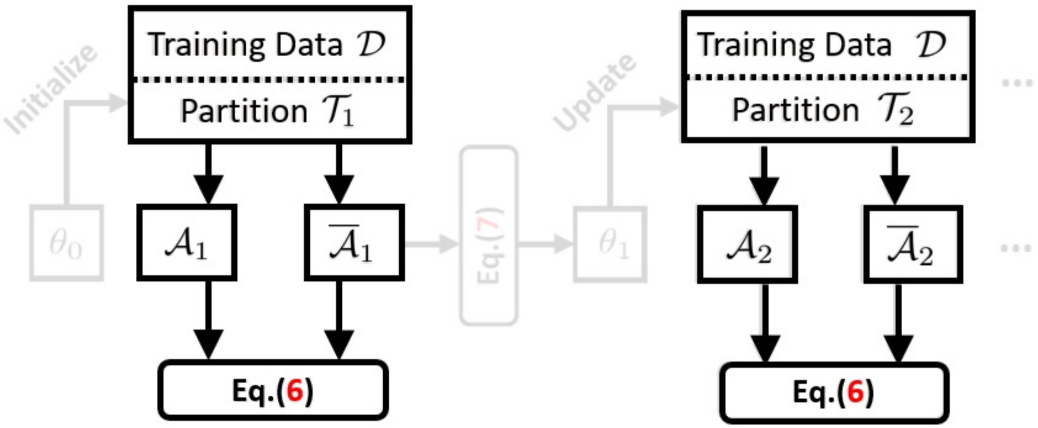
Invariant Loss (IL): This loss is for learning A that is split invariant made by the causal intervention with the data partition T_i.

$$\mathrm{IL}(g, \mathcal{A}(x), \mathcal{T}_i) = \sum_{t \in \mathcal{T}_i} \mathrm{XE}(g, \mathcal{A}(x), t) + \underbrace{\lambda \| \nabla_{\mathbf{w}=\mathbf{1.0}} \mathrm{XE}(\mathbf{w}, \mathcal{A}(x), t) \|_2^2,}_{\textit{Follow the original IRM}}$$

Wang, et al. "Causal attention for unbiased visual recognition." *ICCV* (2021).

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

How to implement CaaM:



Mini-Game: Optimize the Model

$$\min_{\mathcal{A},\overline{\mathcal{A}},f,g,h} \underline{\mathrm{XE}(f,\widetilde{x},\mathcal{D}) + \mathrm{IL}(g,\mathcal{A}(x),\mathcal{T}_i)} + \underline{\mathrm{XE}(h,\overline{\mathcal{A}}(x),\mathcal{D})},$$

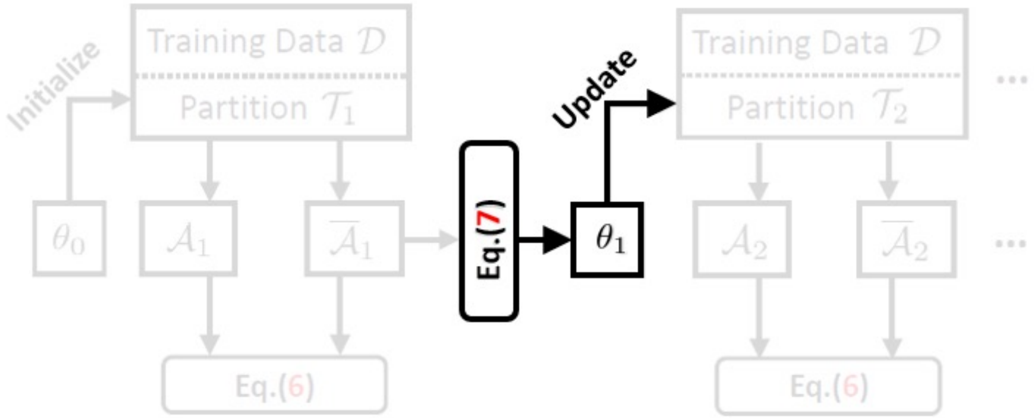
Learn the causal feature $A(x)$.

Learn the bias classifier h by fitting XE loss on $\bar{A}(x)$.

Wang, et al. "Causal attention for unbiased visual recognition." *ICCV* (2021).

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

How to implement CaaM:



$$\max_{\theta} \ \mathrm{IL}(h, \overline{\mathcal{A}}(x), \mathcal{T}_i(\theta))$$

A **good** partition update should capture the bias factor that is currently **NOT** split invariant.

Maxi-Game: Optimize a New Partition

Wang, et al. "Causal attention for unbiased visual recognition." *ICCV* (2021).

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

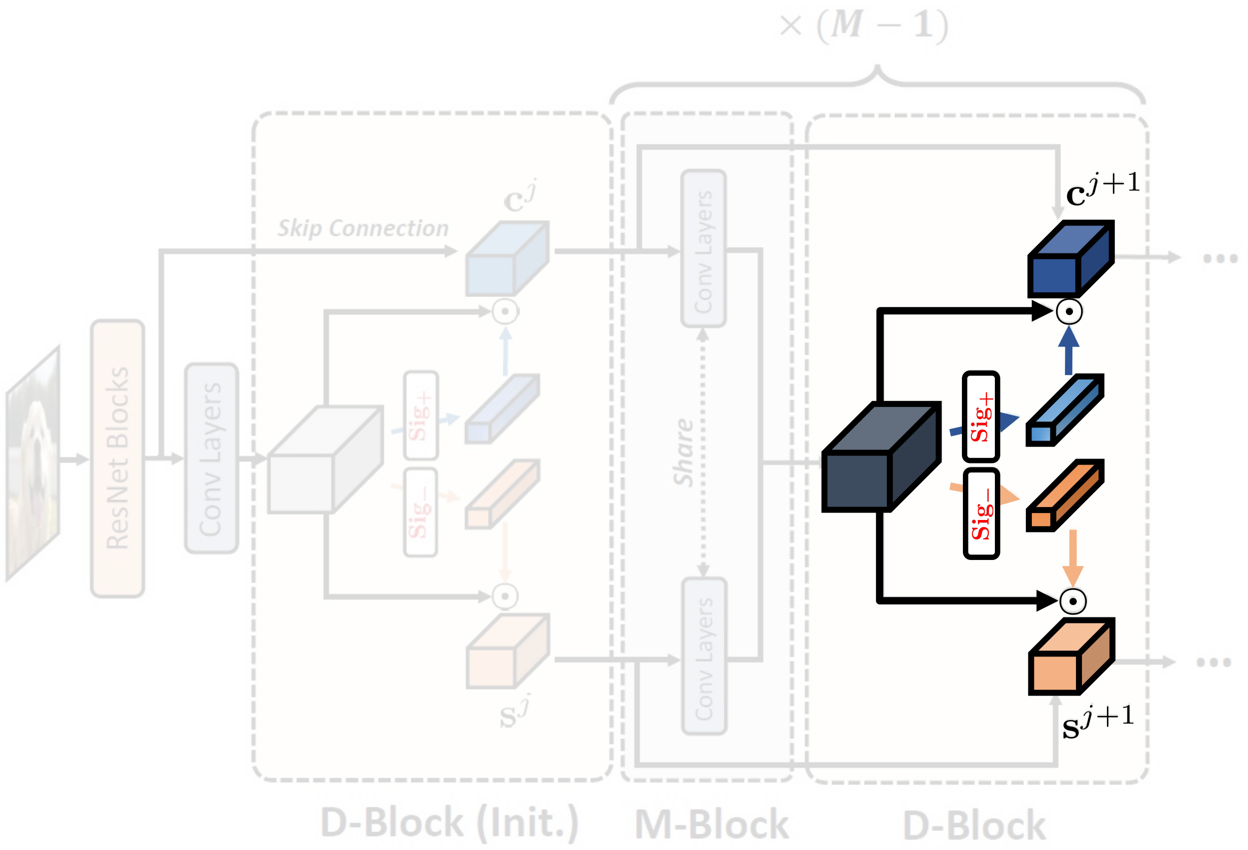
How to implement CaaM:

CaaM Attention Calculus for CNN

$$
\text{CaaM}: \begin{cases} \mathbf{z} = \text{CBAM}(\mathbf{x}), \\ \mathbf{c} = \boxed{\text{Sigmoid}(\mathbf{z})} \odot \mathbf{x}, \\ \mathbf{s} = \boxed{\text{Sigmoid}(-\mathbf{z})} \odot \mathbf{x} \end{cases}
$$

$$Sigmoid(z) = 1 - Sigmoid(-z)$$

Complementary Attention

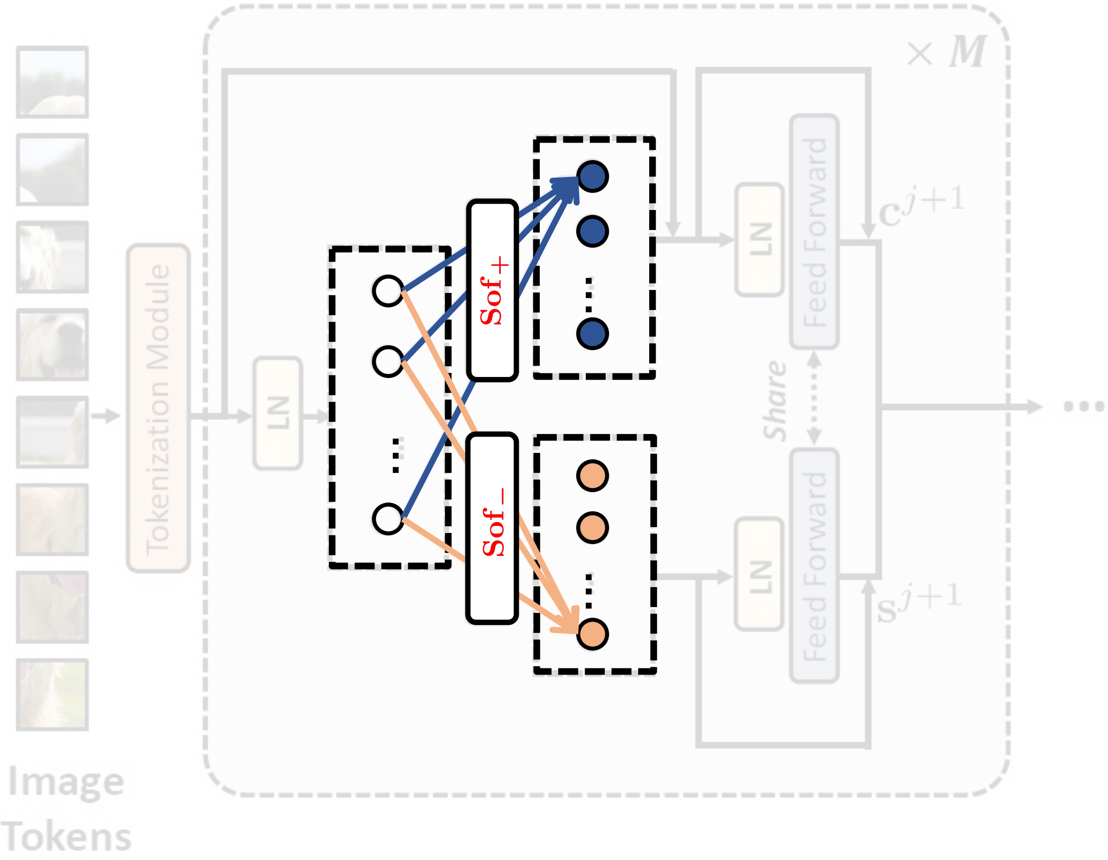


Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

How to implement CaaM:

CaaM Attention Calculus for ViT

$$\mathbf{CaaM}: \begin{cases} \mathbf{q}, \mathbf{k}, \mathbf{v} = \mathbf{W}_q\mathbf{x}, \mathbf{W}_k\mathbf{x}, \mathbf{W}_v\mathbf{x}, \\ \mathbf{c} = \mathrm{Softmax}(\frac{\mathbf{q}\mathbf{k}^{\mathrm{T}}}{\sqrt{d_K}})\mathbf{v}, \\ \mathbf{s} = \mathrm{Softmax}(-\frac{\mathbf{q}\mathbf{k}^{\mathrm{T}}}{\sqrt{d_K}})\mathbf{v} \end{cases}$$



Wang, et al. "Causal attention for unbiased visual recognition." *ICCV* (2021).

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

Evaluate CaaM:

	Model	CNN-Based					ViT-Based				
		NICO		ImageNet-9 [31]		ImageNet-A [23]	NICO		ImageNet-9 [31]		ImageNet-A [23]
		Val	Test	Biased	Unbiased [5]	Test	Val	Test	Biased	Unbiased [5]	Test
Conv.	ResNet18 [20]	43.77	42.61	95.00	94.40	33.67	–	–	–	–	–
Conv.	ResNet18+CBAM [57]	42.15	42.46	94.81	94.09	34.31	–	–	–	–	–
Conv.	T2T-ViT7 [63]	–	–	–	–	–	36.23	35.62	88.76	88.35	31.28
Conv.	RUBi [8]	43.86	44.37	94.81	94.27	34.13	35.27	34.15	87.95	87.48	29.90
Conv.	ReBias [5]	44.92	45.23	95.20	94.89	34.26	35.28	35.74	88.99	88.32	29.33
Conv.	Cutout [15]	43.69	43.77	95.24	94.81	34.68	35.31	33.69	87.52	86.47	27.97
Conv.	Mixup [67]	44.85	41.46	95.43	94.79	37.71	37.85	34.31	89.72	88.66	30.73
w/ H.A. T	IRM [4]	40.62	41.46	94.13	94.41	33.52	36.46	34.38	89.43	88.87	30.17
w/ H.A. T	REx [34]	41.00	41.15	94.15	94.28	33.18	36.23	33.46	88.52	87.26	29.18
w/ H.A. T	Unshuffle [49]	43.15	43.00	94.71	94.33	34.41	37.38	36.00	87.38	86.86	28.61
w/ H.A. T	**CaaM (Ours)**	**45.46**	**45.77**	**95.52**	**94.96**	35.60	**38.08**	**37.54**	**90.05**	**89.35**	**32.01**
w/o H.A. T	IRM [4]	40.54	41.23	94.09	94.32	33.39	33.76	33.77	89.62	88.98	29.25
w/o H.A. T	REx [34]	40.85	41.52	93.26	93.79	32.84	35.62	34.00	88.68	87.01	28.72
w/o H.A. T	Unshuffle [49]	41.69	41.61	94.81	94.30	34.04	33.62	32.92	88.38	87.39	28.52
w/o H.A. T	**CaaM (Ours)**	**46.38**	**46.62**	**96.19**	**95.83**	**38.55**	**38.00**	**37.61**	**90.33**	**90.01**	**32.38**

Wang, et al. "Causal attention for unbiased visual recognition." *ICCV* (2021).

School of Computing and Information Systems

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

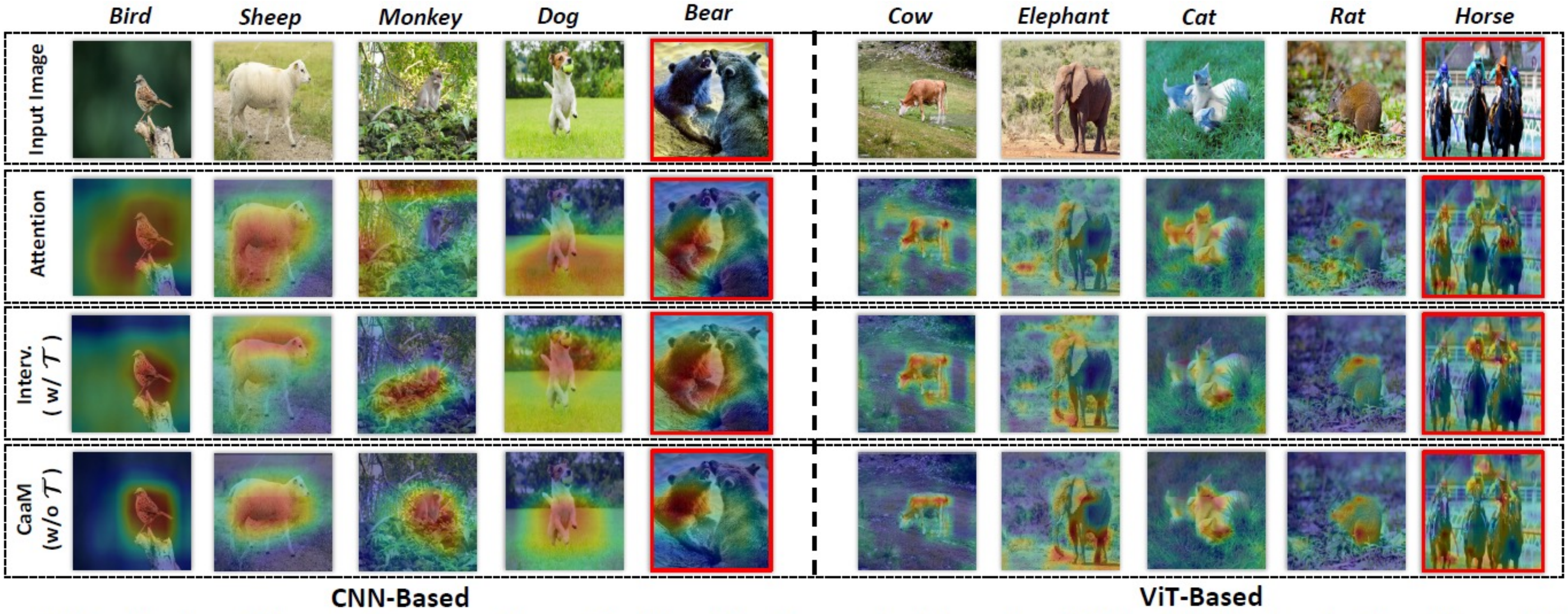
Evaluate CaaM:

	Model	CNN-Based NICO Val	CNN-Based NICO Test	CNN-Based ImageNet-9 [31] Biased	CNN-Based ImageNet-9 [31] Unbiased [5]	CNN-Based ImageNet-A [23] Test	ViT-Based NICO Val	ViT-Based NICO Test	ViT-Based ImageNet-9 [31] Biased	ViT-Based ImageNet-9 [31] Unbiased [5]	ViT-Based ImageNet-A [23] Test
Conv.	ResNet18 [20]	43.77	42.61	95.00	94.40	33.67	–	–	–	–	–
Conv.	ResNet18+CBAM [57]	42.15	42.46	94.81	94.09	34.31	–	–	–	–	–
Conv.	T2T-ViT7 [63]	–	–	–	–	–	36.23	35.62	88.76	88.35	31.28
Conv.	RUBi [8]	43.86	44.37	94.81	94.27	34.13	35.27	34.15	87.95	87.48	29.90
Conv.	ReBias [5]	44.92	45.23	95.20	94.89	34.26	35.28	35.74	88.99	88.32	29.33
Conv.	Cutout [15]	43.69	43.77	95.24	94.81	34.68	35.31	33.69	87.52	86.47	27.97
Conv.	Mixup [67]	44.85	41.46	95.43	94.79	37.71	37.85	34.31	89.72	88.66	30.73
w/ H.A. $\mathcal{T}$	IRM [4]	40.62	41.46	94.13	94.41	33.52	36.46	34.38	89.43	88.87	30.17
w/ H.A. $\mathcal{T}$	REx [34]	41.00	41.15	94.15	94.28	33.18	36.23	33.46	88.52	87.26	29.18
w/ H.A. $\mathcal{T}$	Unshuffle [49]	43.15	43.00	94.71	94.33	34.41	37.38	36.00	87.38	86.86	28.61
w/ H.A. $\mathcal{T}$	**CaaM (Ours)**	**45.46**	**45.77**	**95.52**	**94.96**	35.60	**38.08**	**37.54**	**90.05**	**89.35**	**32.01**
w/o H.A. $\mathcal{T}$	IRM [4]	40.54	41.23	94.09	94.32	33.39	33.76	33.77	89.62	88.98	29.25
w/o H.A. $\mathcal{T}$	REx [34]	40.85	41.52	93.26	93.79	32.84	35.62	34.00	88.68	87.01	28.72
w/o H.A. $\mathcal{T}$	Unshuffle [49]	41.69	41.61	94.81	94.30	34.04	33.62	32.92	88.38	87.39	28.52
w/o H.A. $\mathcal{T}$	**CaaM (Ours)**	**46.38**	**46.62**	**96.19**	**95.83**	**38.55**	**38.00**	**37.61**	**90.33**	**90.01**	**32.38**

Wang, et al. "Causal attention for unbiased visual recognition." *ICCV* (2021).

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

Evaluate CaaM:



Wang, et al. "Causal attention for unbiased visual recognition." *ICCV* (2021).

Invariant Learning in Insufficient Data: Causal Attention Module (CaaM)

Improved CaaM (our ongoing work):

CaaM

Generate environments for all the classes

The environments may still be **inaccurate**

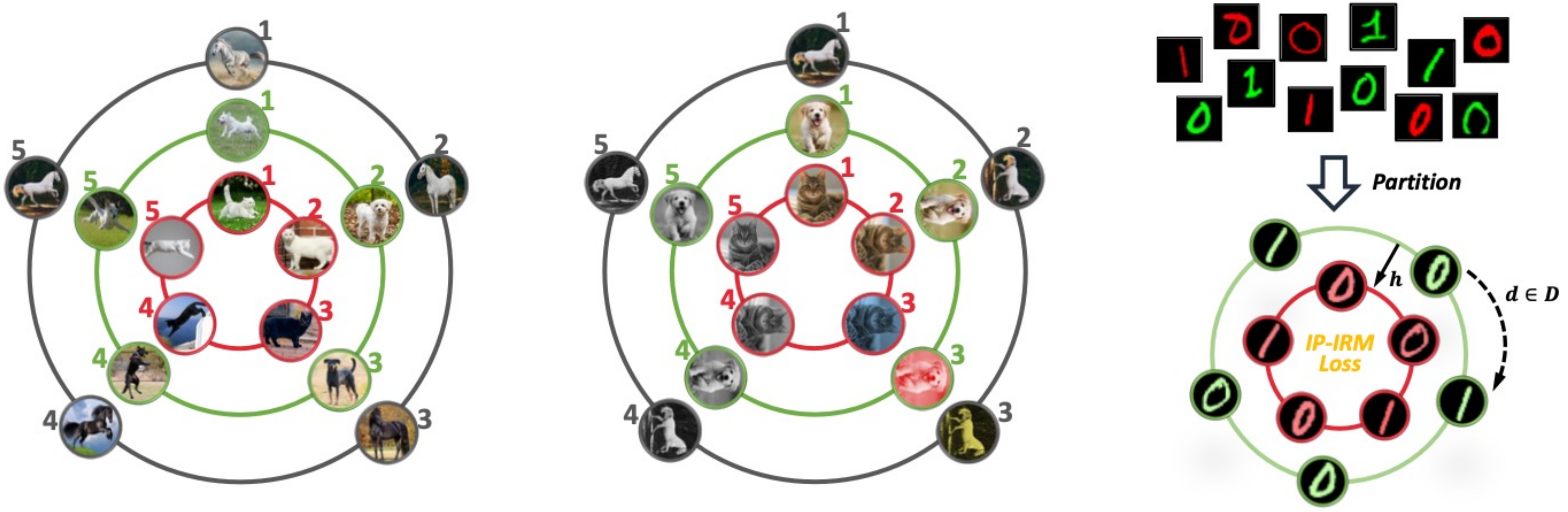
Improved CaaM

Class-wise IRM

Each environment corresponds to one class

Invariant Learning in Insufficient Data: Representation Learning

Representation Learning (our published work):



(a) Supervised Learning (b) Self-Supervised Learning (c) Our IP-IRM

Wang, et al. "Self-Supervised Learning Disentangled Group Representation as Feature." *NeurIPS* (2021) Spotlight.

School of Computing and Information Systems

Invariant Learning in Insufficient Data

An Interesting Challenge: NICOCHALLENGE (also in ECCV'22 workshop)

https://nicochallenge.com/

What is NICO CHALLENGE?

The goal of NICO Challenge is to facilitate the OOD (Out-of-Distribution) generalization in visual recognition through promoting the research on the intrinsic learning mechanisms with native invariance and generalization ability. The training data is a mixture of several observed contexts while the test data is composed of unseen contexts. Participants are tasked with developing reliable algorithms across different contexts (domains) to improve the generalization ability of models.