Combining Diverse Feature Priors

Saachi Jain

Dimitris Tsipras

Aleksander Mądry

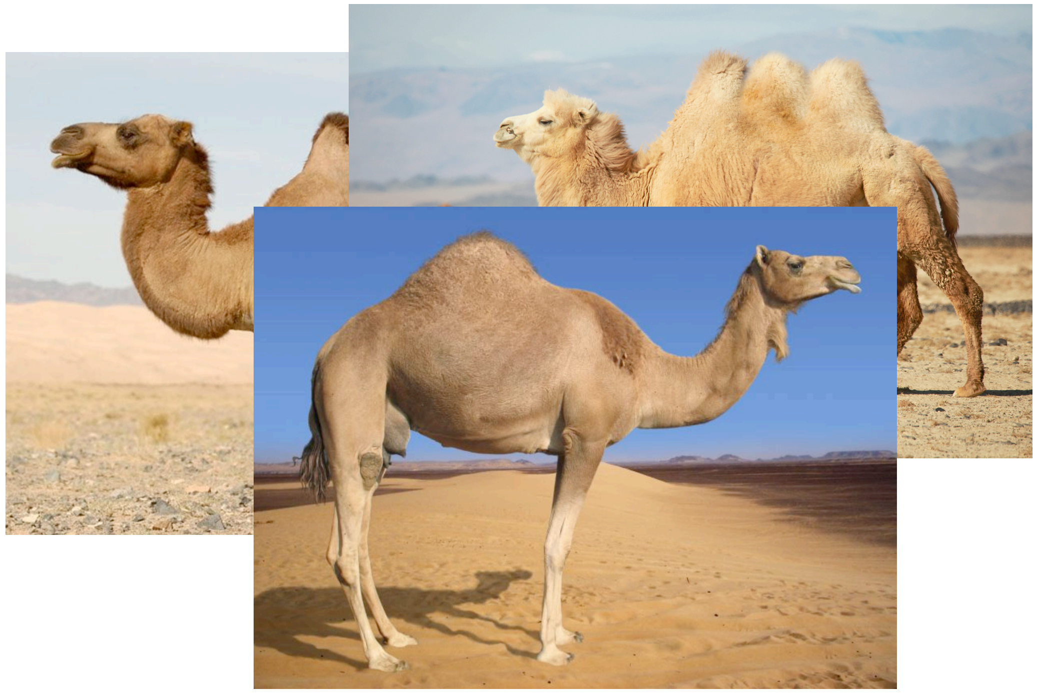
MIT CSAIL

gradient-science.org

Features and generalization

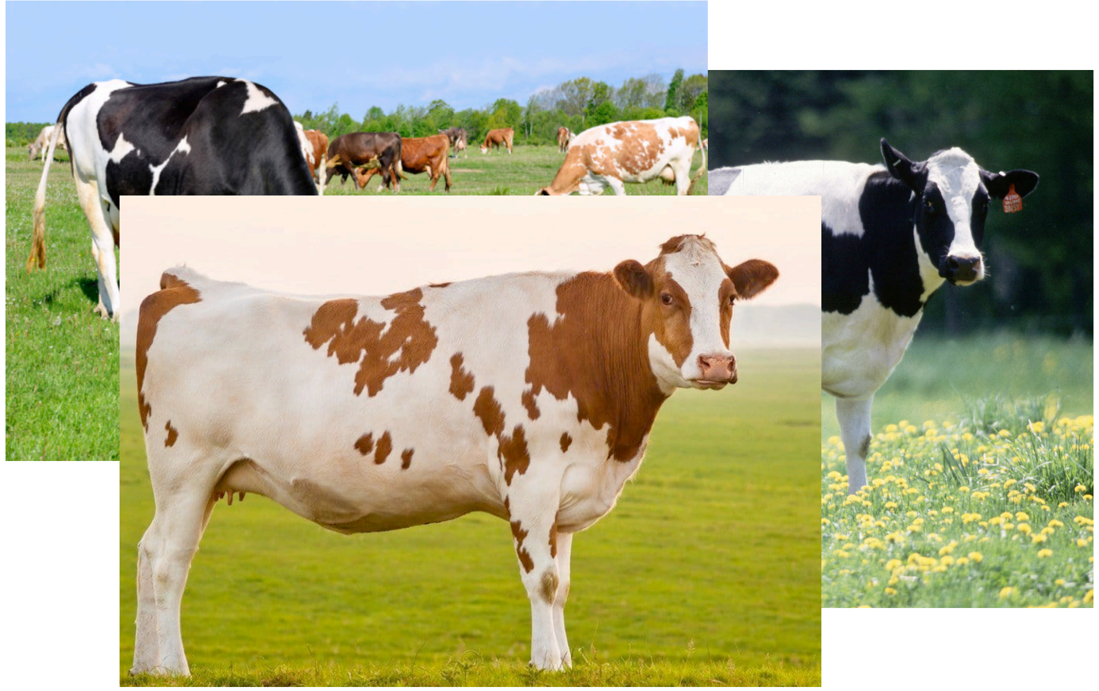
Features and generalization

"Camel"

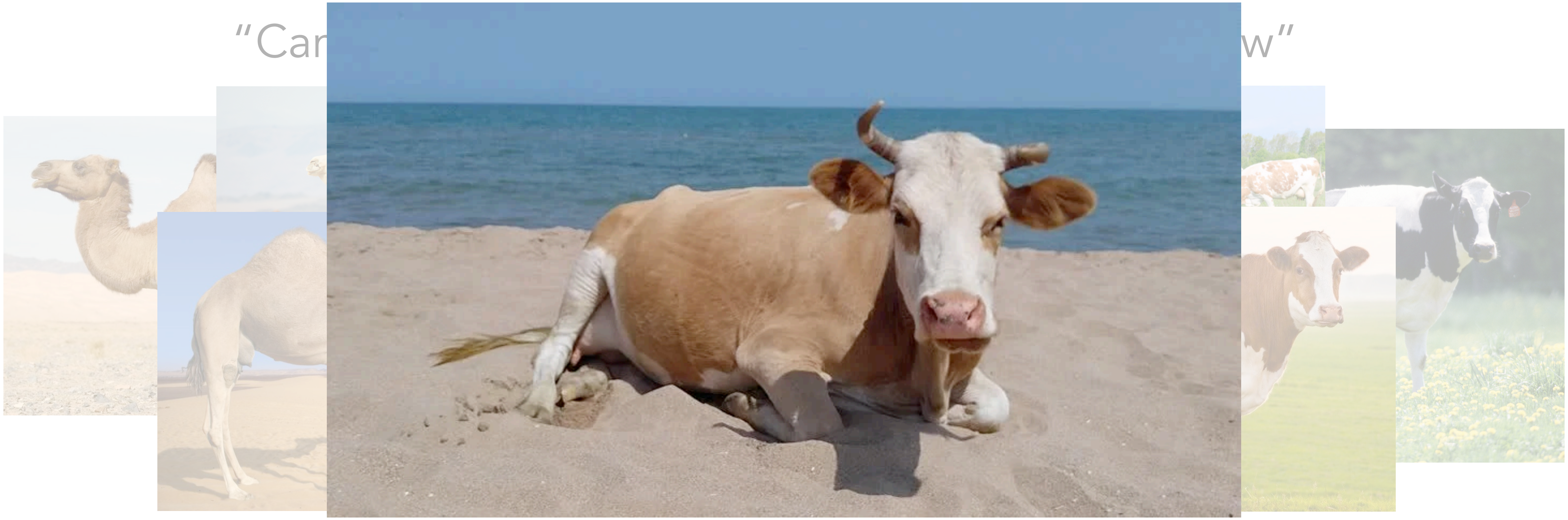


vs.

"Cow"



Features and generalization



Can it recognize a cow on the beach?

Features and generalization

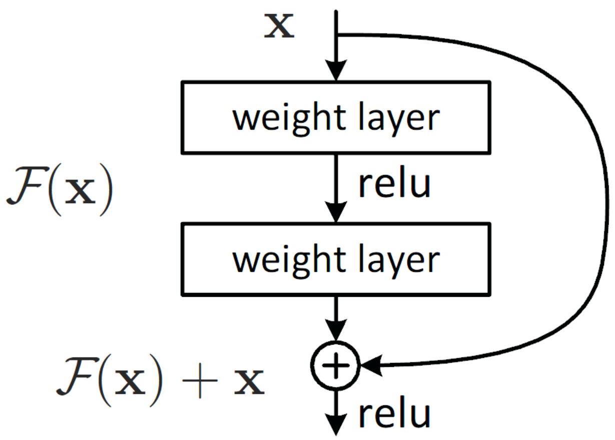
Can it recognize a cow on the beach?

Generalization is driven by **feature priors**

What factors influence learned features?

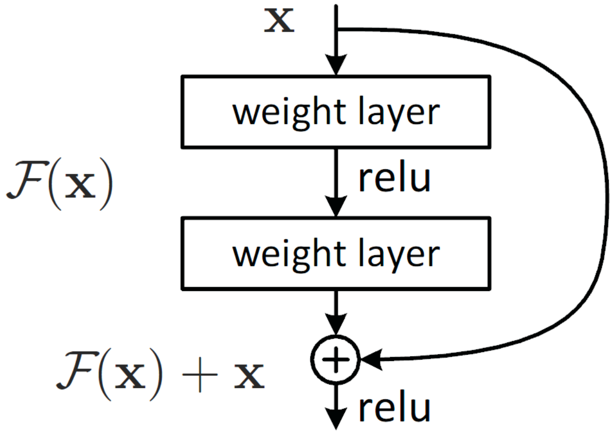
What factors influence learned features?

Architecture

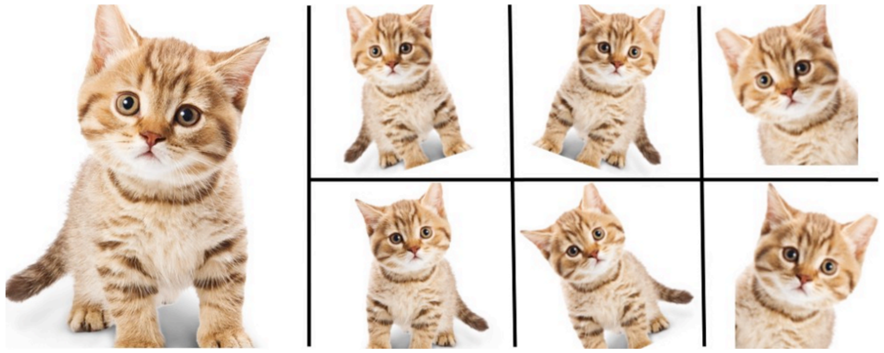

$\mathbf{x}$
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

What factors influence learned features?

Architecture

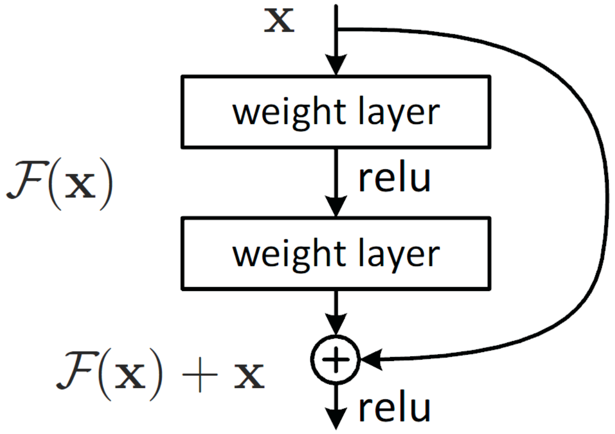

$\mathbf{x}$
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Augmentations

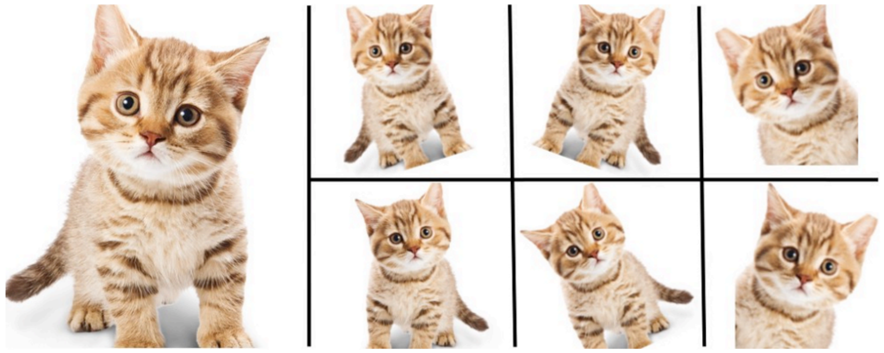


What factors influence learned features?

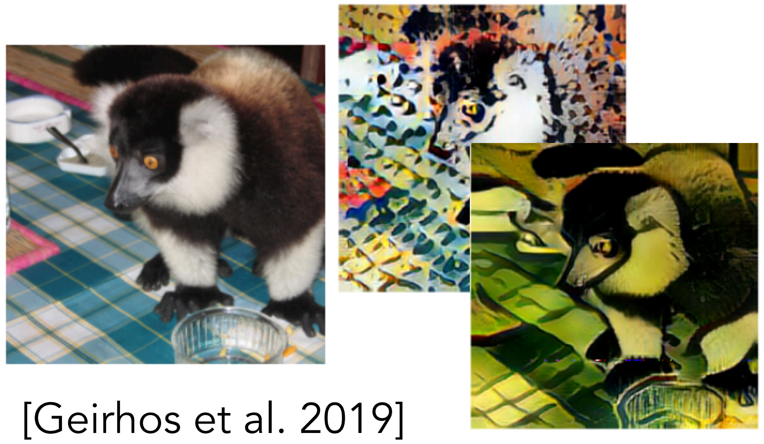
Architecture


$\mathbf{x}$
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Augmentations



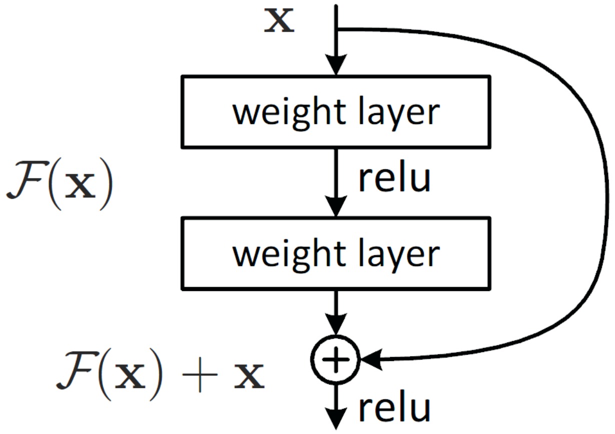
Stylized training



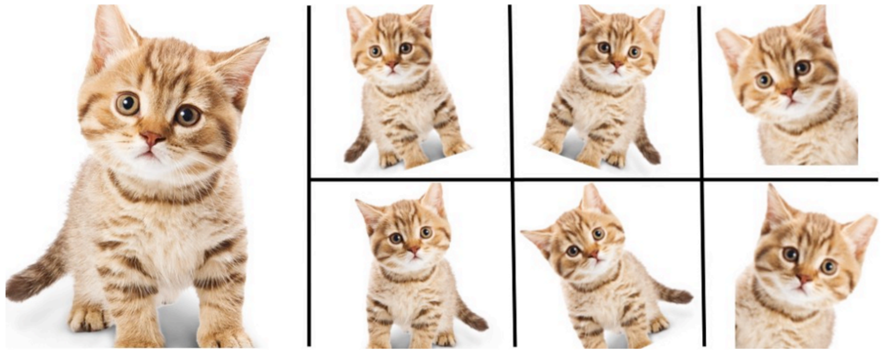
[Geirhos et al. 2019]

What factors influence learned features?

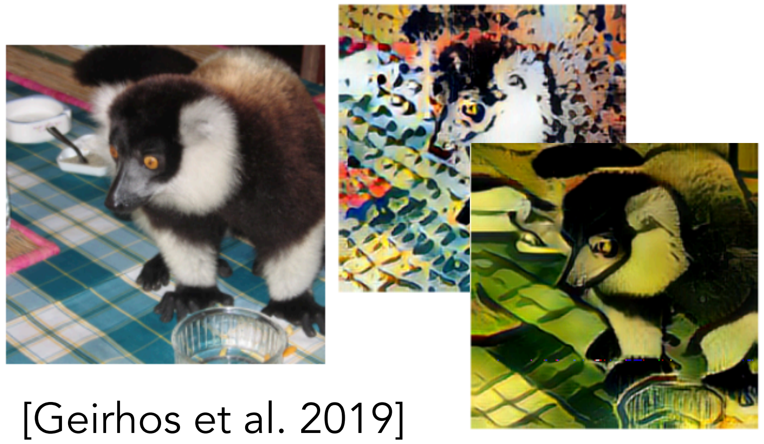
Architecture


$\mathbf{x}$
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Augmentations



Stylized training



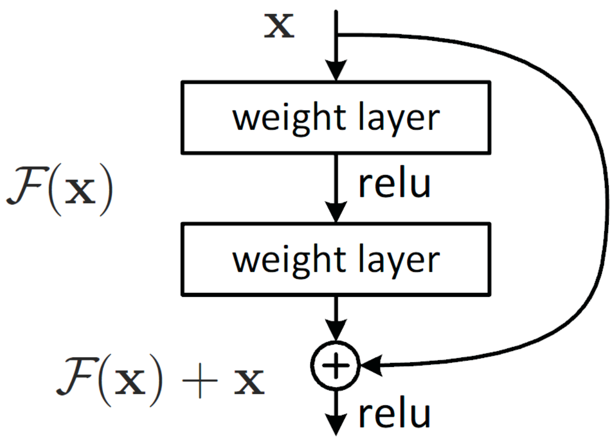
[Geirhos et al. 2019]

Limited receptive field

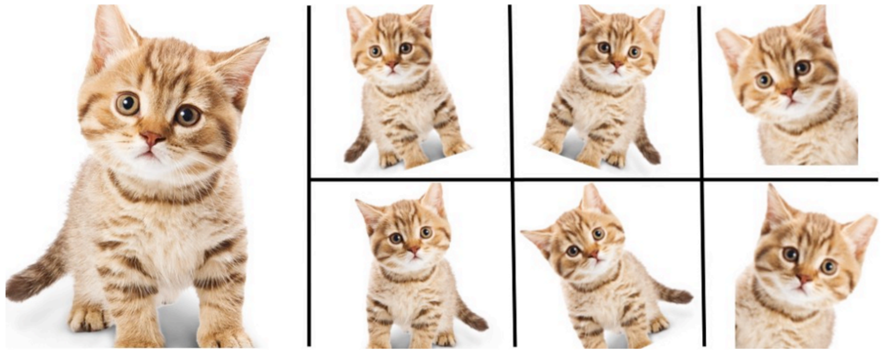
[Brendel Bethge 2019]

What factors influence learned features?

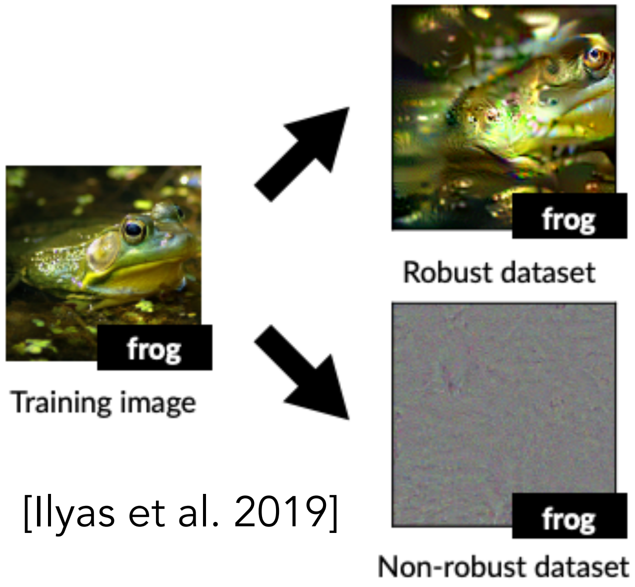
Architecture


$\mathbf{x}$
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Augmentations

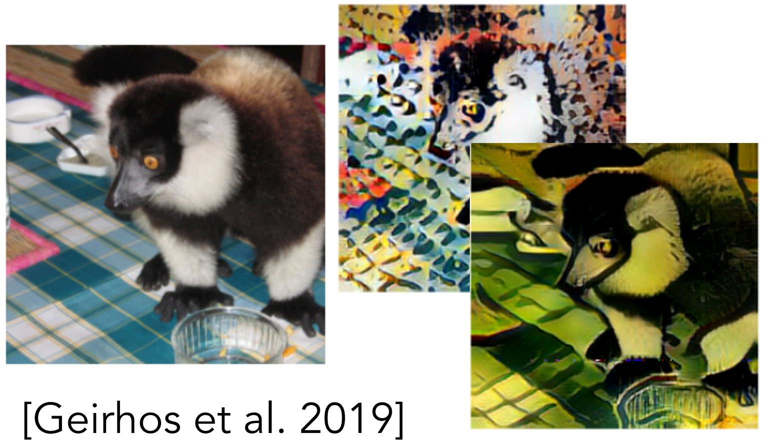


Robust optimization



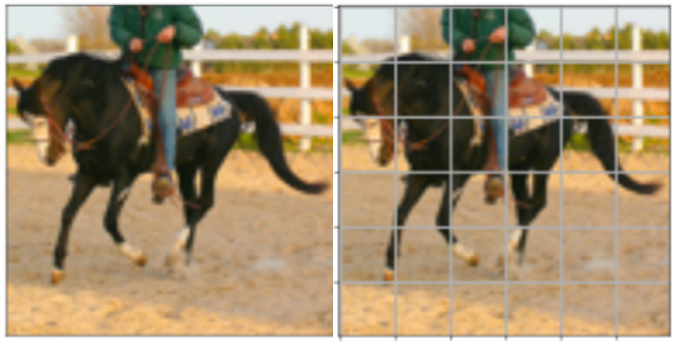
[Ilyas et al. 2019]

Stylized training



[Geirhos et al. 2019]

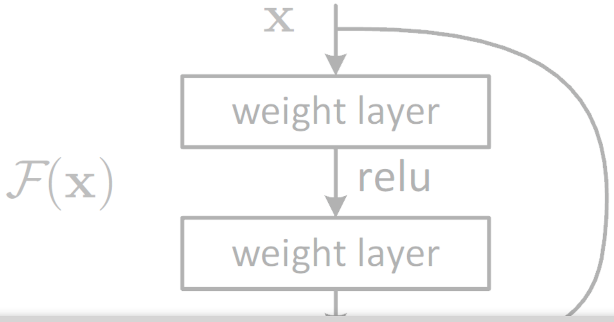
Limited receptive field



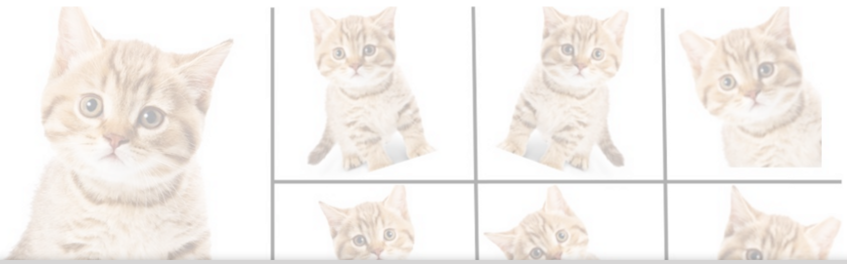
[Brendel Bethge 2019]

What factors influence learned features?

Architecture



Augmentations



How can we harness this emerging space of **feature priors**?

on

Stylized training

[Geirhos et al. 2019]

Limited receptive field

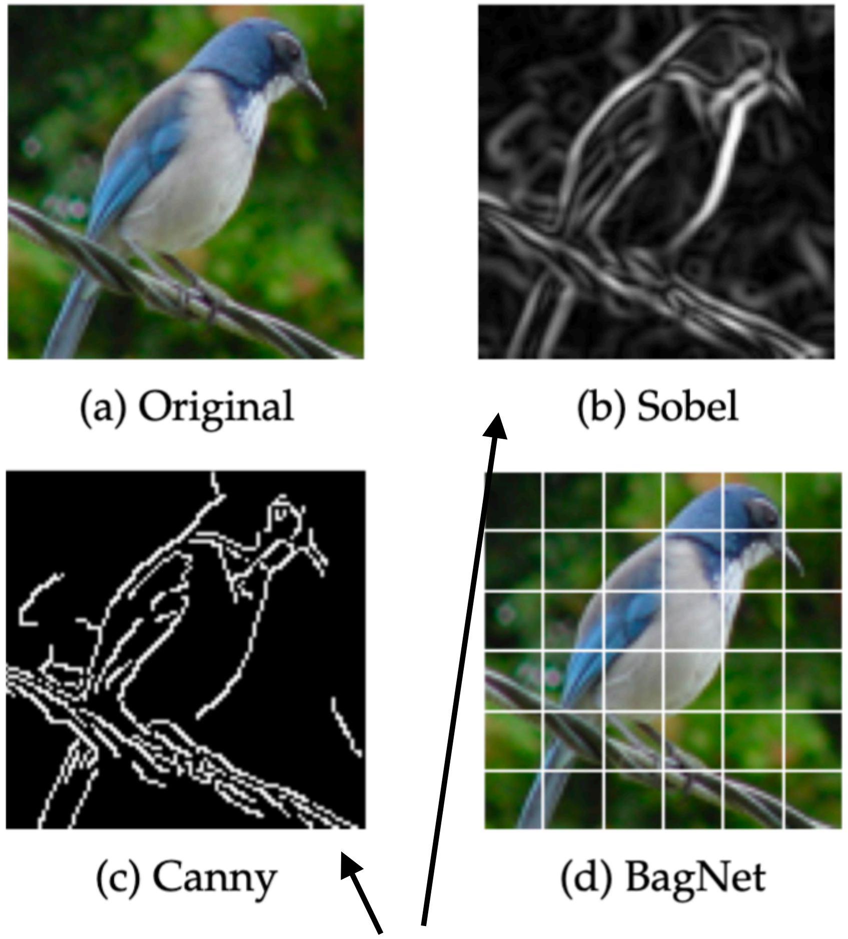
[Brendel Bethge 2019]

[Ilyas et al. 2019]

Feature priors as distinct perspectives on data

Feature priors as distinct perspectives on data

Train models with
diverse feature priors



(a) Original

(b) Sobel

(c) Canny

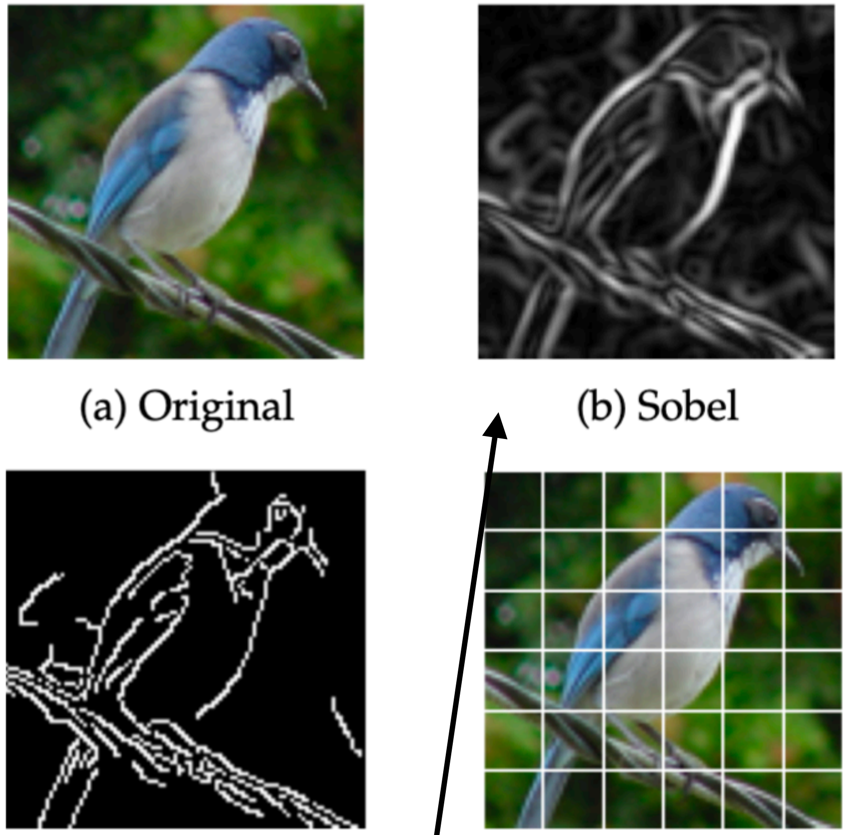
(d) BagNet

shape-biased

texture-biased

Feature priors as distinct perspectives on data

Train models with
diverse feature priors



(a) Original

(b) Sobel

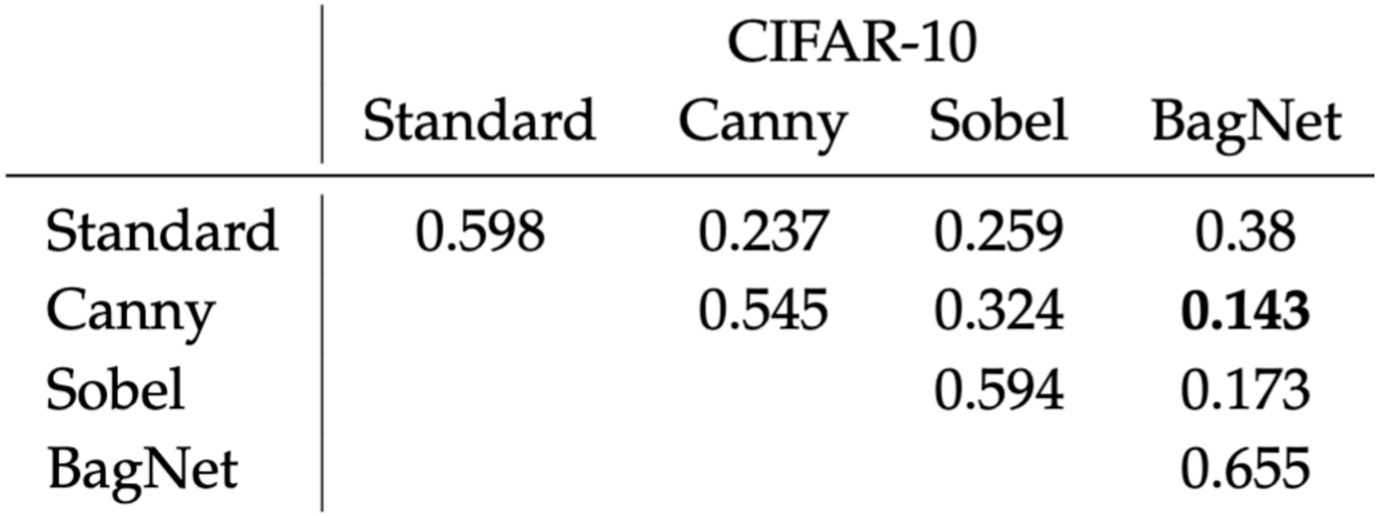
(c) Canny

(d) BagNet

shape-biased

texture-biased

Correlation of correct predictions

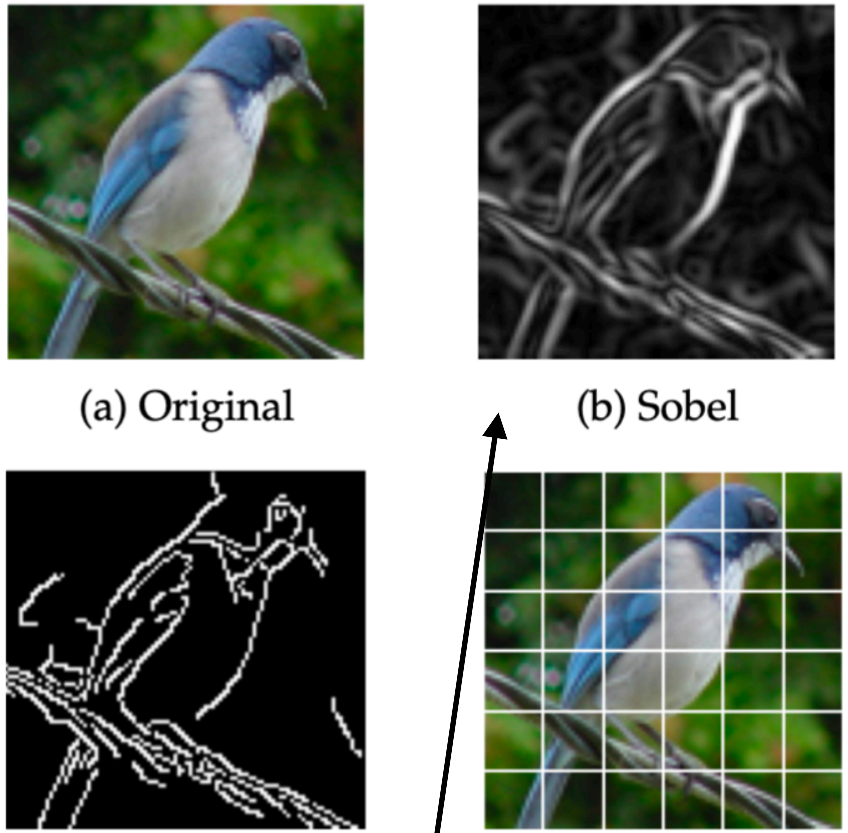


	CIFAR-10			
	Standard	Canny	Sobel	BagNet
Standard	0.598	0.237	0.259	0.38
Canny		0.545	0.324	**0.143**
Sobel			0.594	0.173
BagNet				0.655

Models with different feature priors make **different mistakes**

Feature priors as distinct perspectives on data

Train models with
diverse feature priors



(a) Original

(b) Sobel

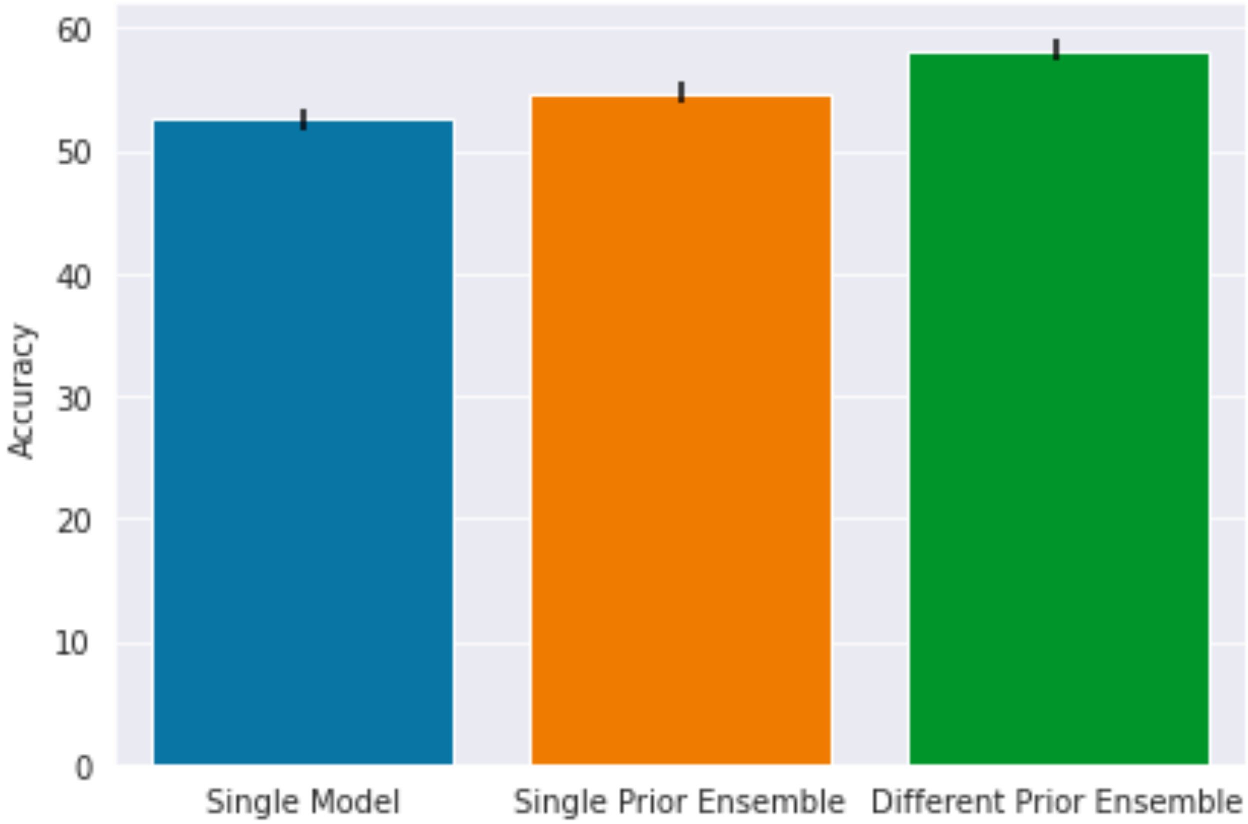
(c) Canny

(d) BagNet

shape-biased

texture-biased

Diverse ensembles perform better



Models with different feature priors make **different mistakes**

Feature priors as distinct perspectives on data

Train models with
diverse feature priors

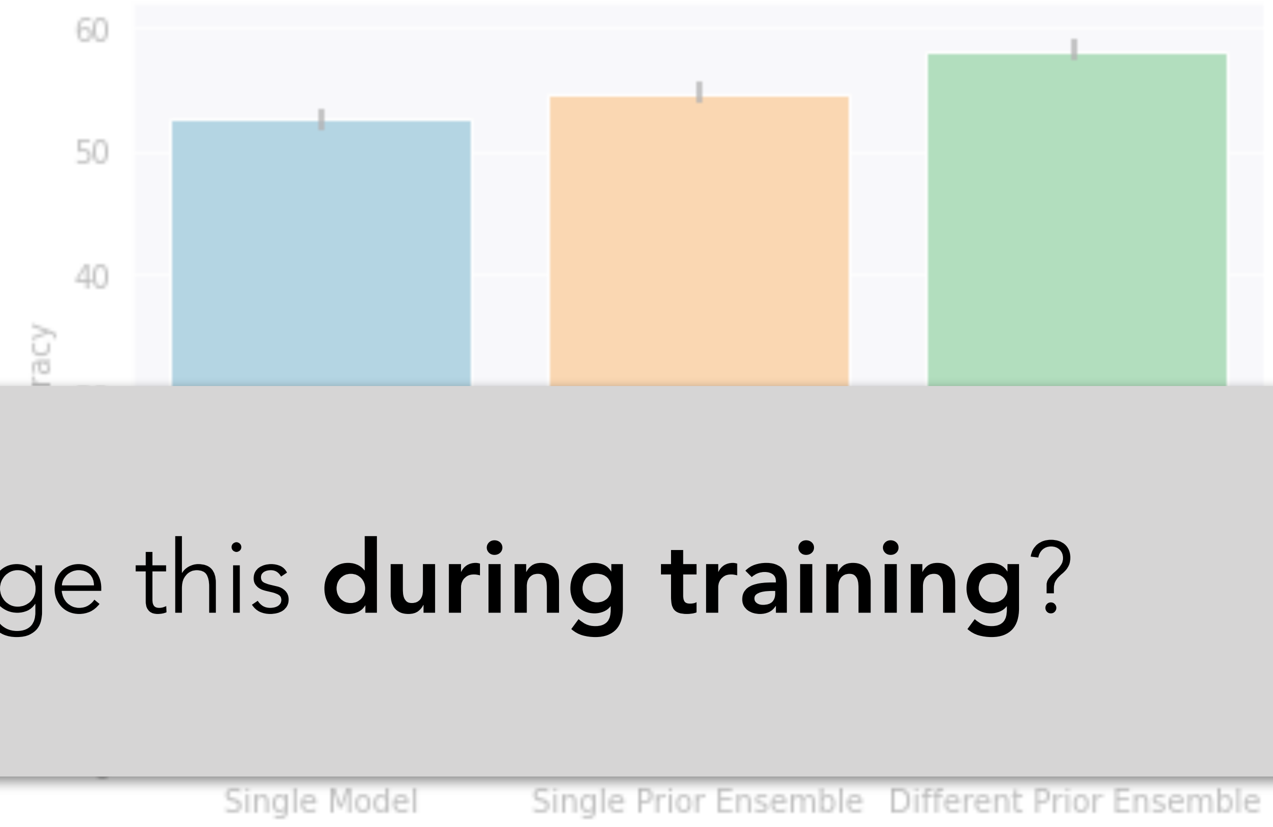
(c) Canny

(d) BagNet

shape-biased

texture-biased

Diverse ensembles perform better



How do we leverage this **during training**?

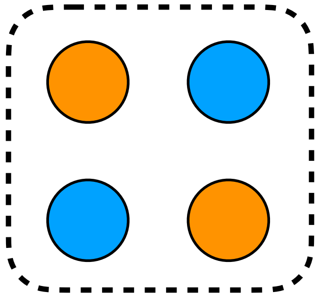
Models with different feature priors make **different mistakes**

Self-training and confirmation bias

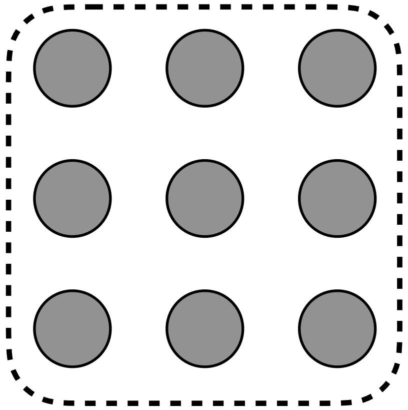
Self-training and confirmation bias

Self-training

Labeled data

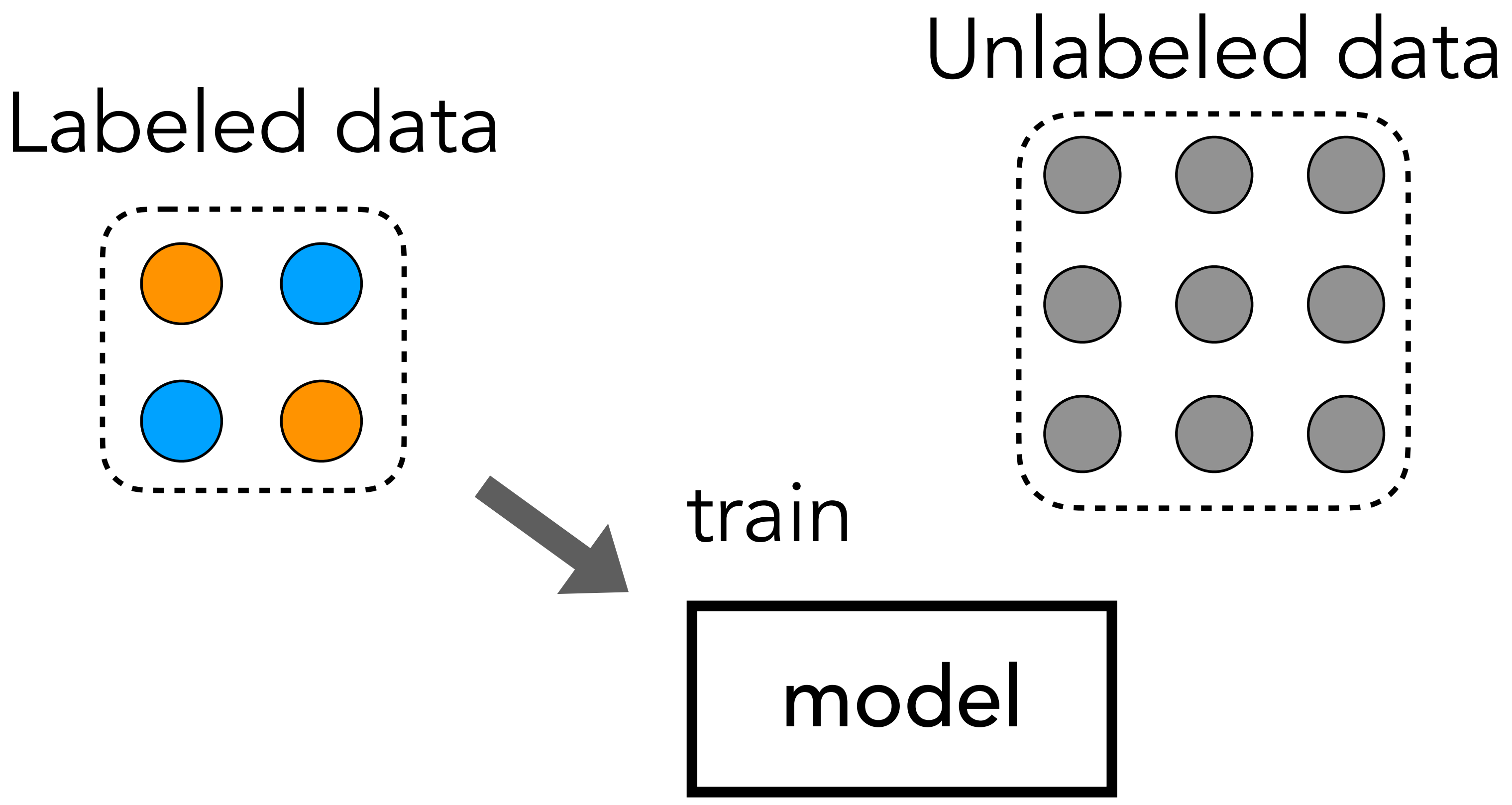


Unlabeled data



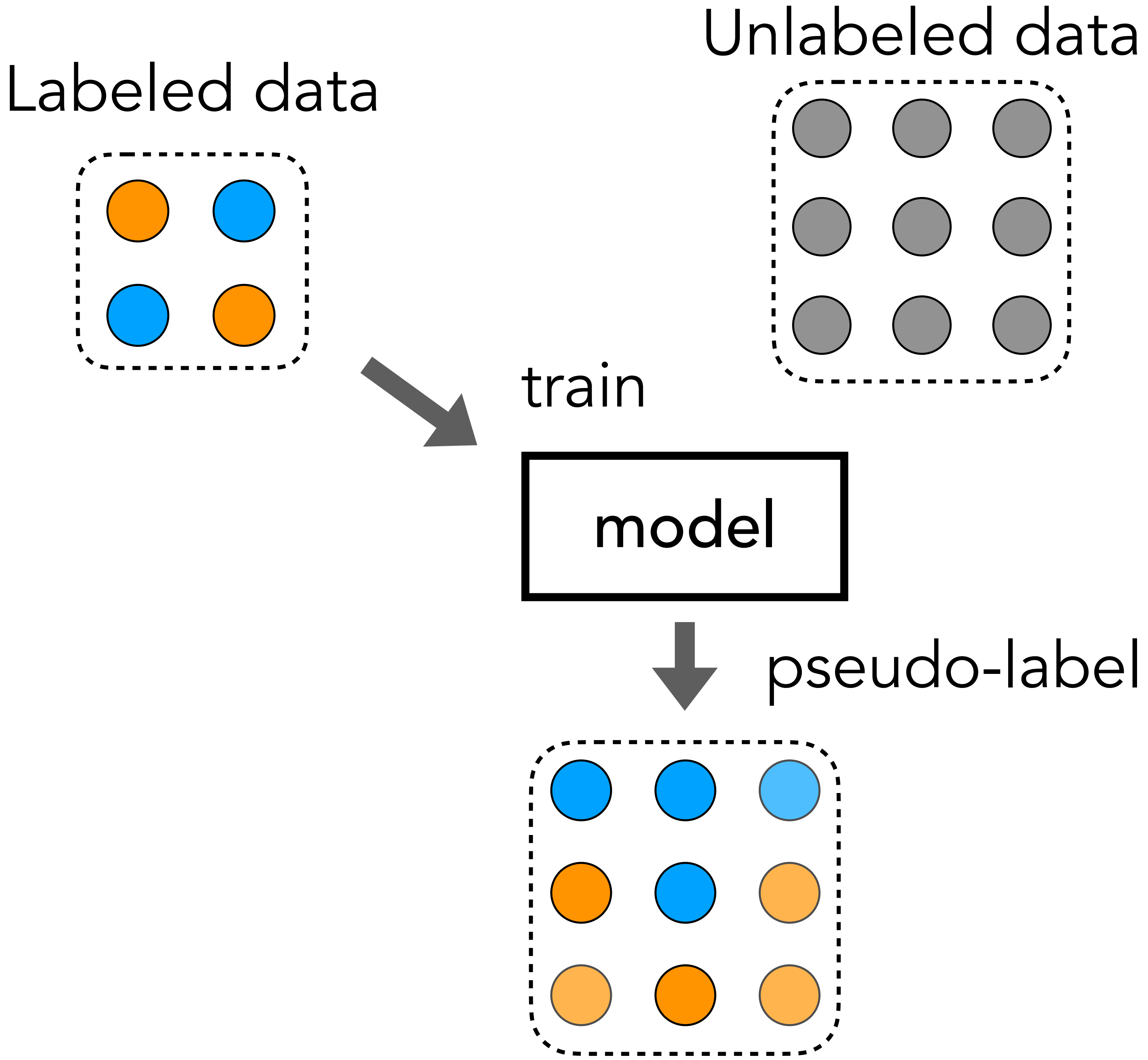
Self-training and confirmation bias

Self-training



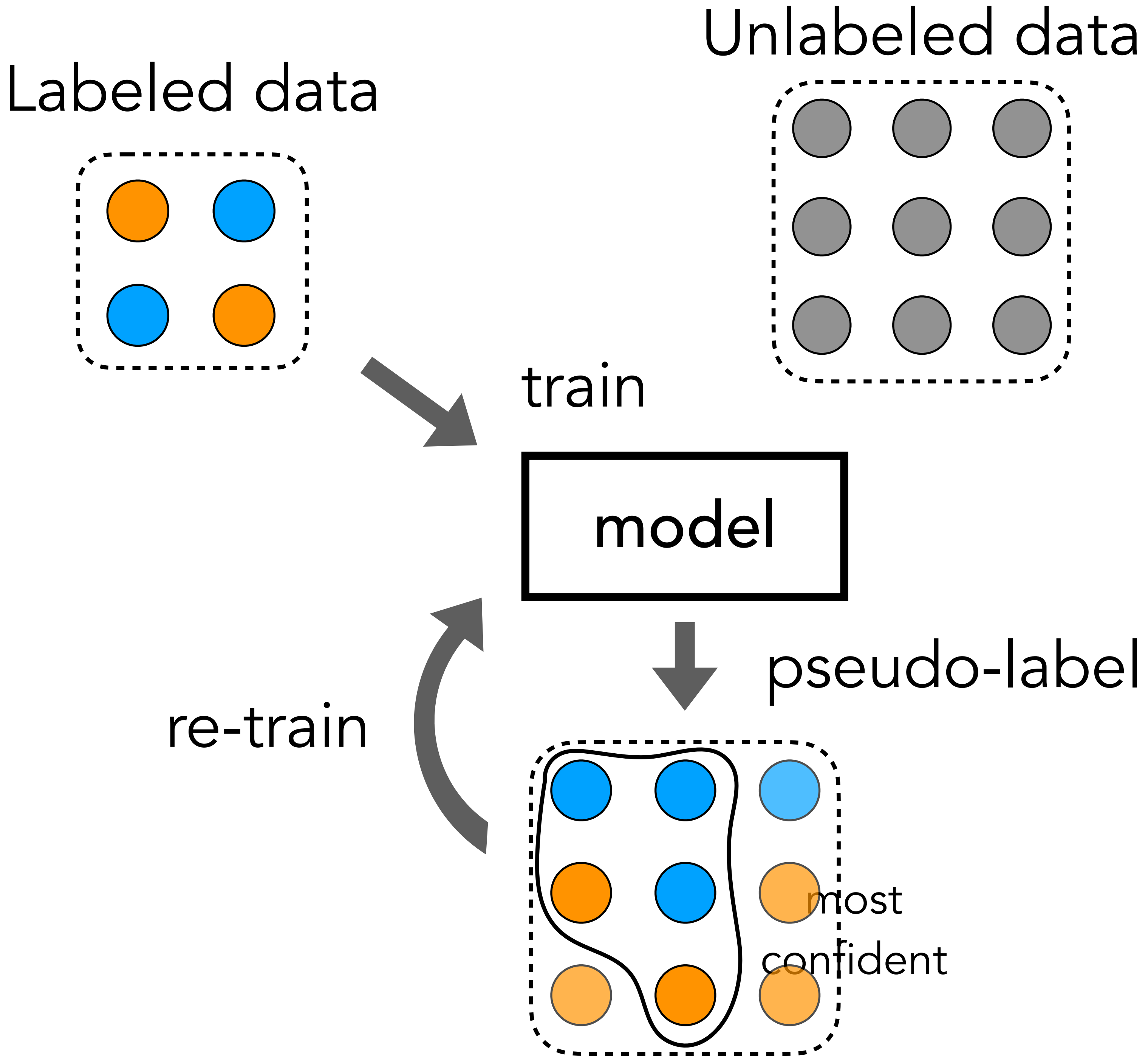
Self-training and confirmation bias

Self-training



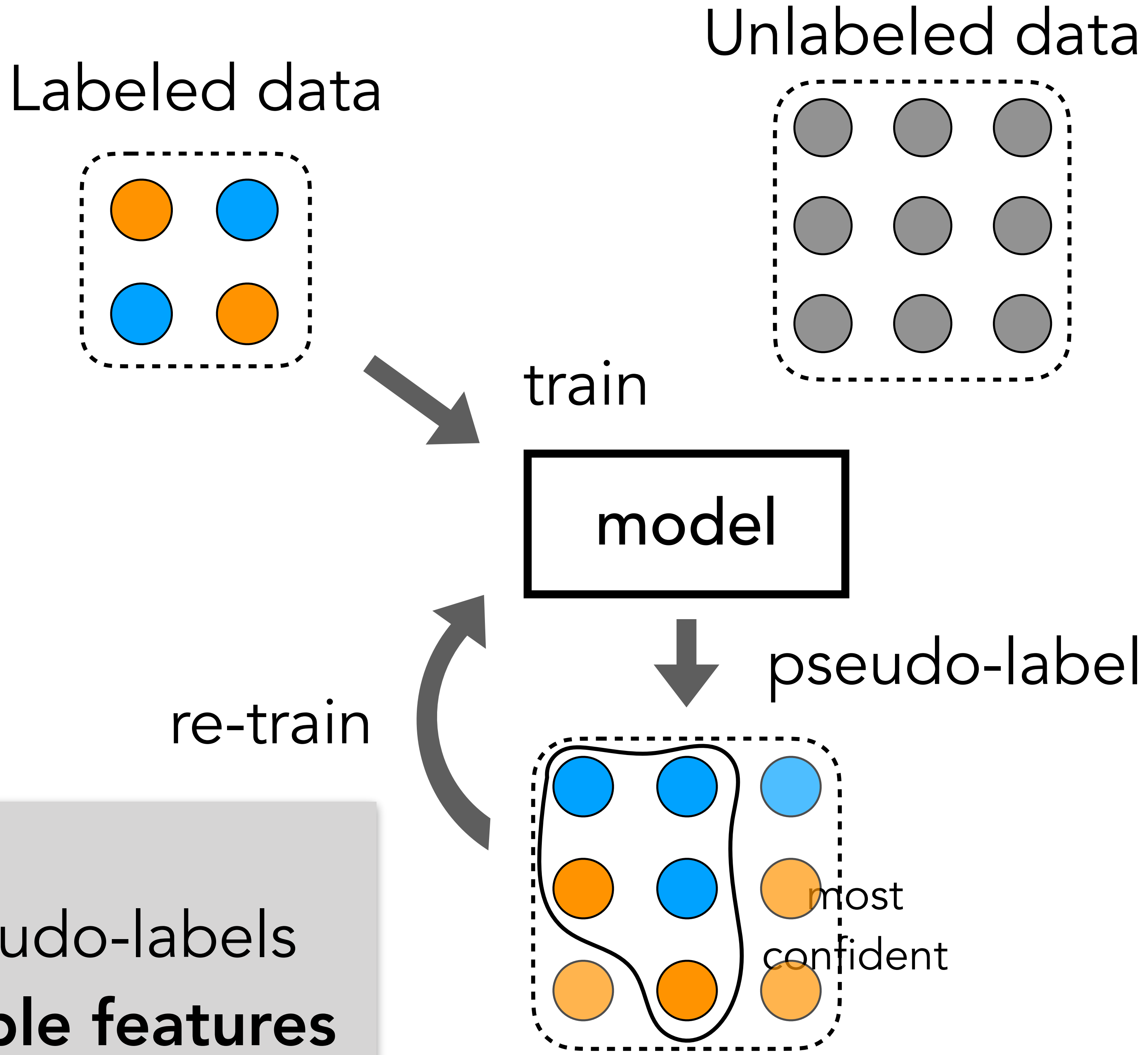
Self-training and confirmation bias

Self-training



Self-training and confirmation bias

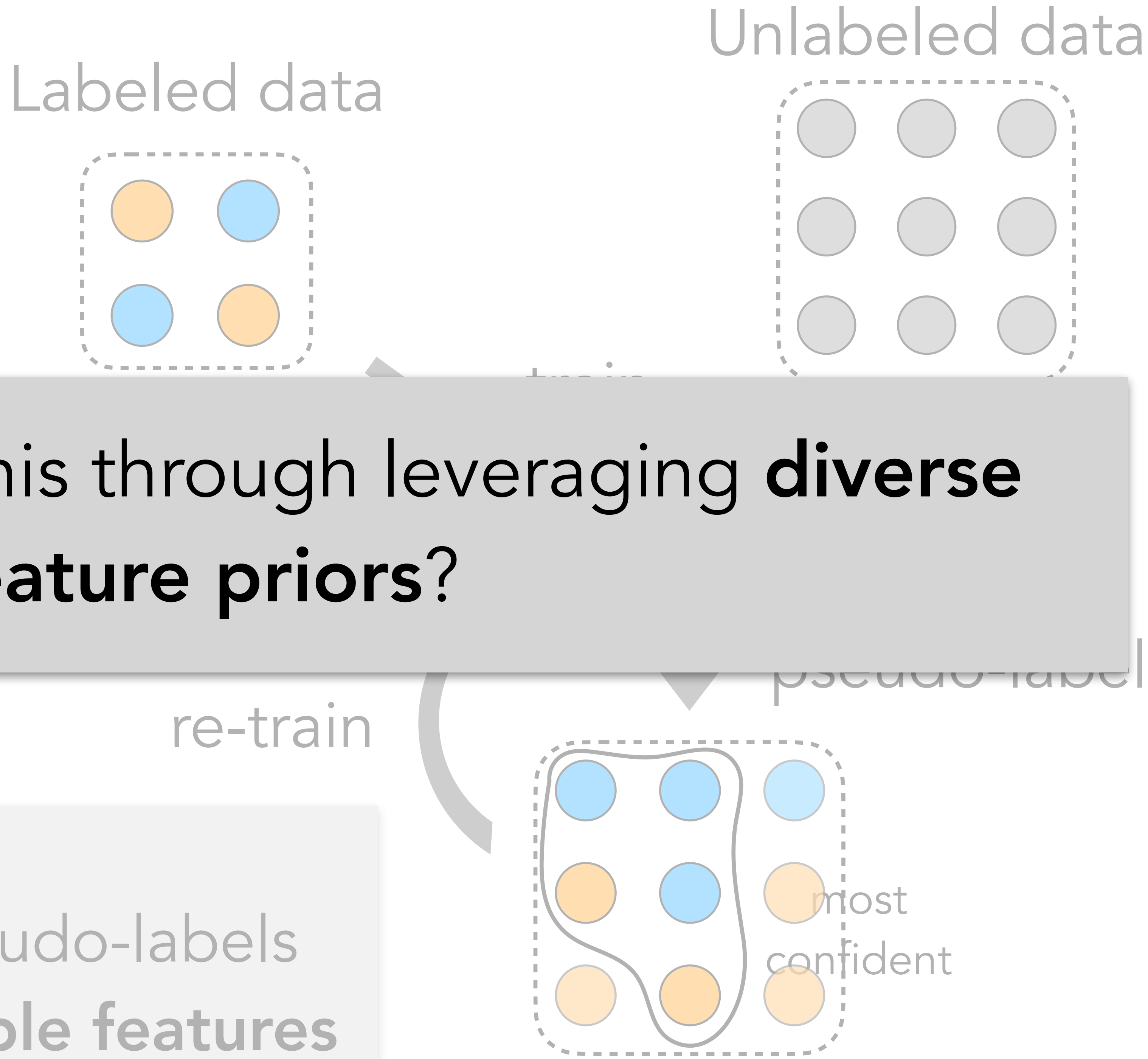
Self-training



Confirmation bias: Pseudo-labels can propagate **undesirable features**

Self-training and confirmation bias

Self-training



Can we mitigate this through leveraging **diverse feature priors**?

Confirmation bias: Pseudo-labels can propagate **undesirable features**

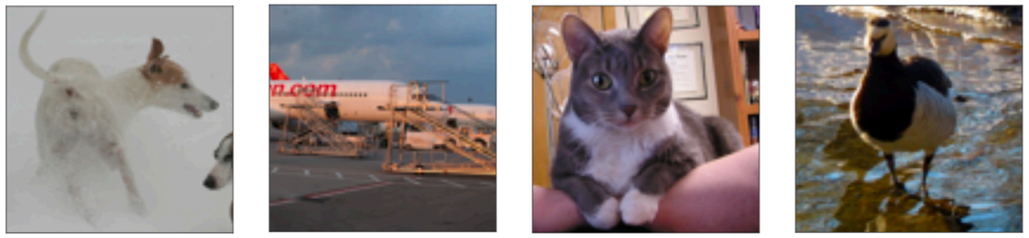
Co-training with diverse feature priors

Key idea: Different feature priors lead to models that learn **different features**

Co-training with diverse feature priors

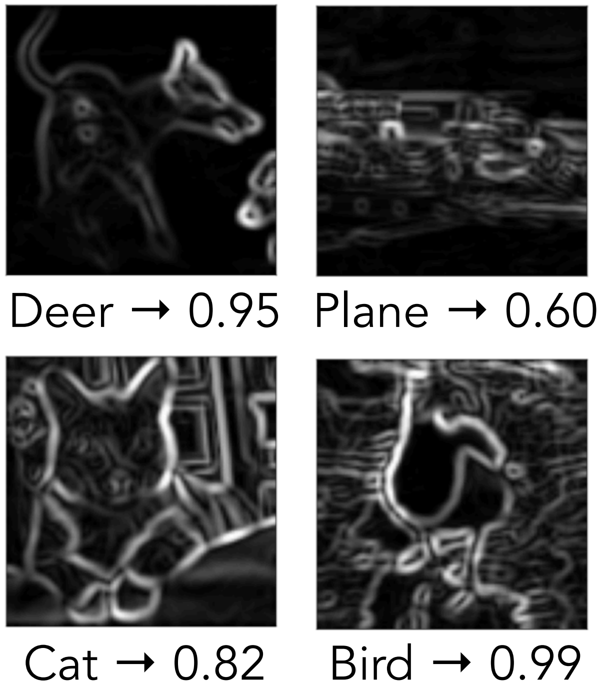
Key idea: Different feature priors lead to models that learn **different features**

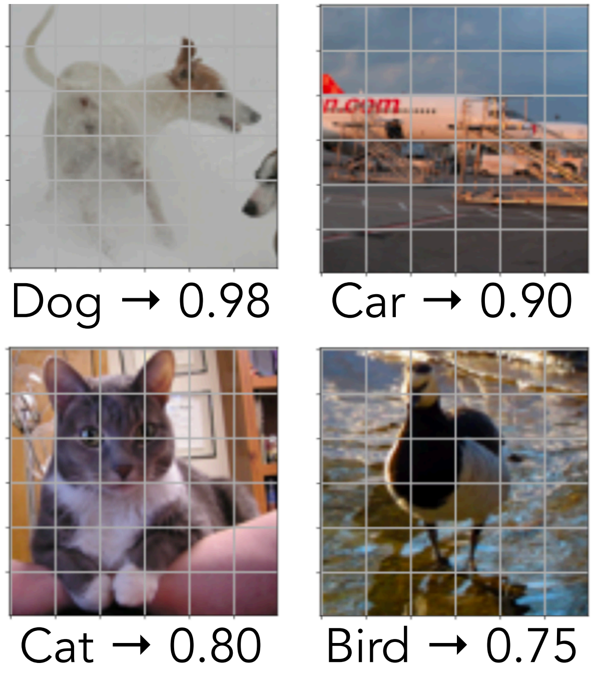
Shape-biased model



Unlabeled data

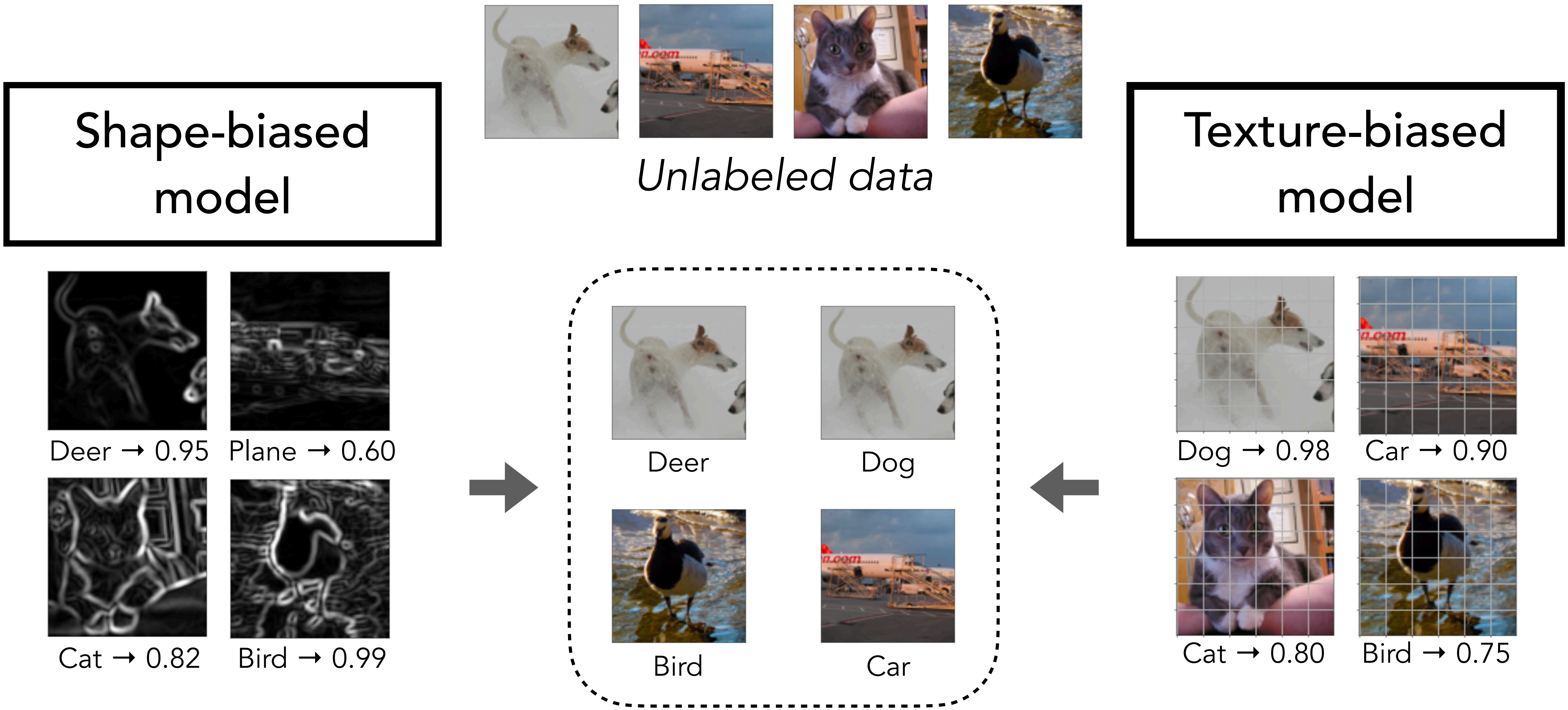
Texture-biased model





Co-training with diverse feature priors

Key idea: Different feature priors lead to models that learn **different features**



Shape-biased model

Unlabeled data

Texture-biased model

Deer → 0.95

Plane → 0.60

Cat → 0.82

Bird → 0.99

Deer

Dog

Bird

Car

Dog → 0.98

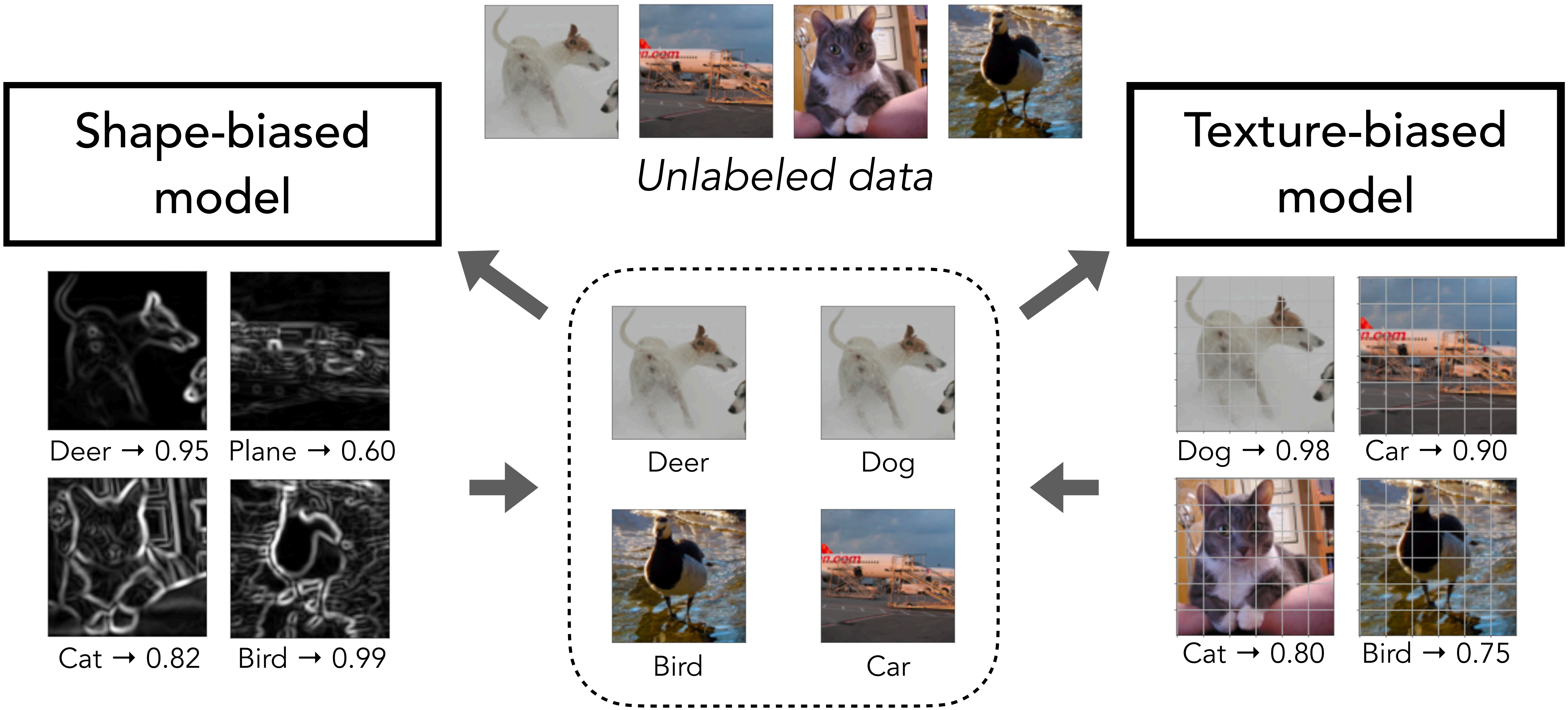
Car → 0.90

Cat → 0.80

Bird → 0.75

Co-training with diverse feature priors

Key idea: Different feature priors lead to models that learn **different features**



Shape-biased model

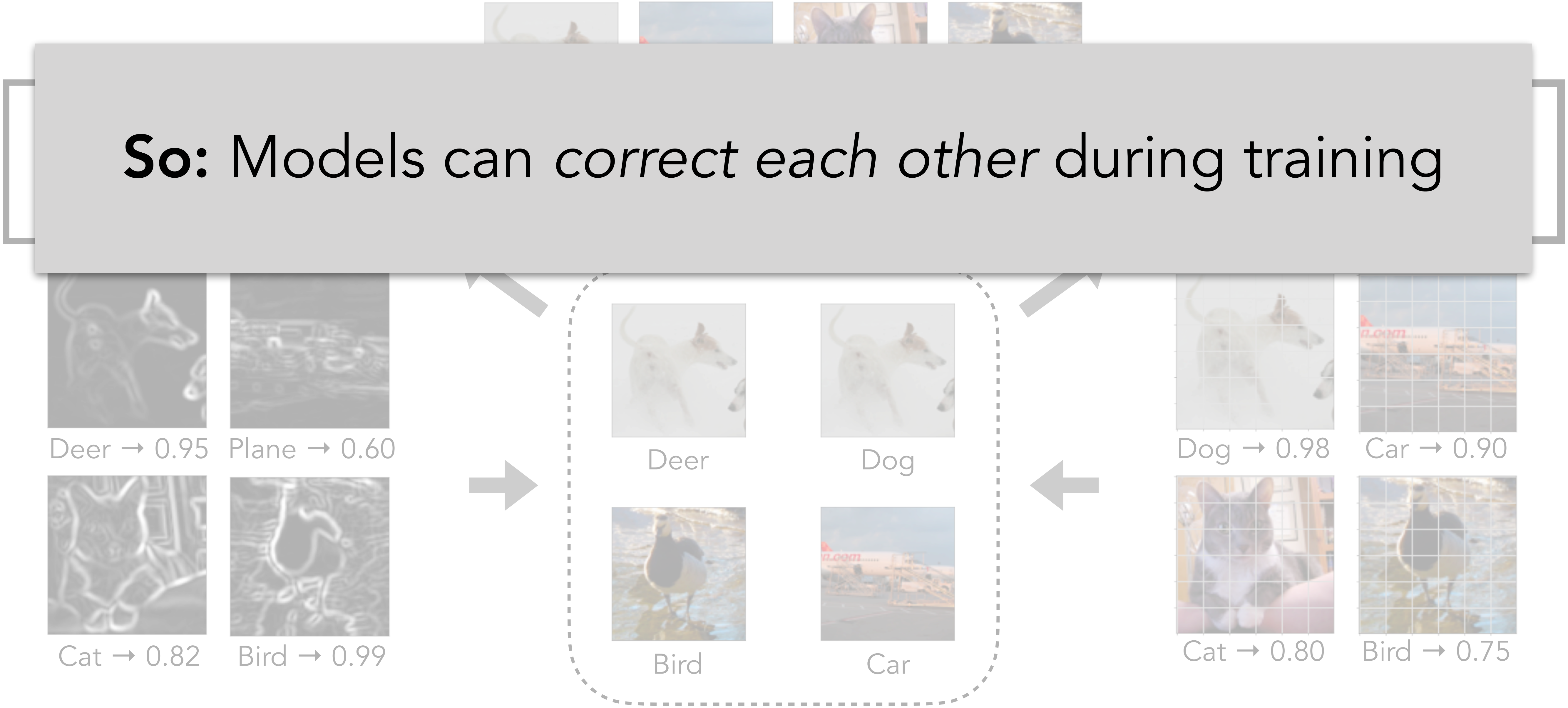
Unlabeled data

Texture-biased model

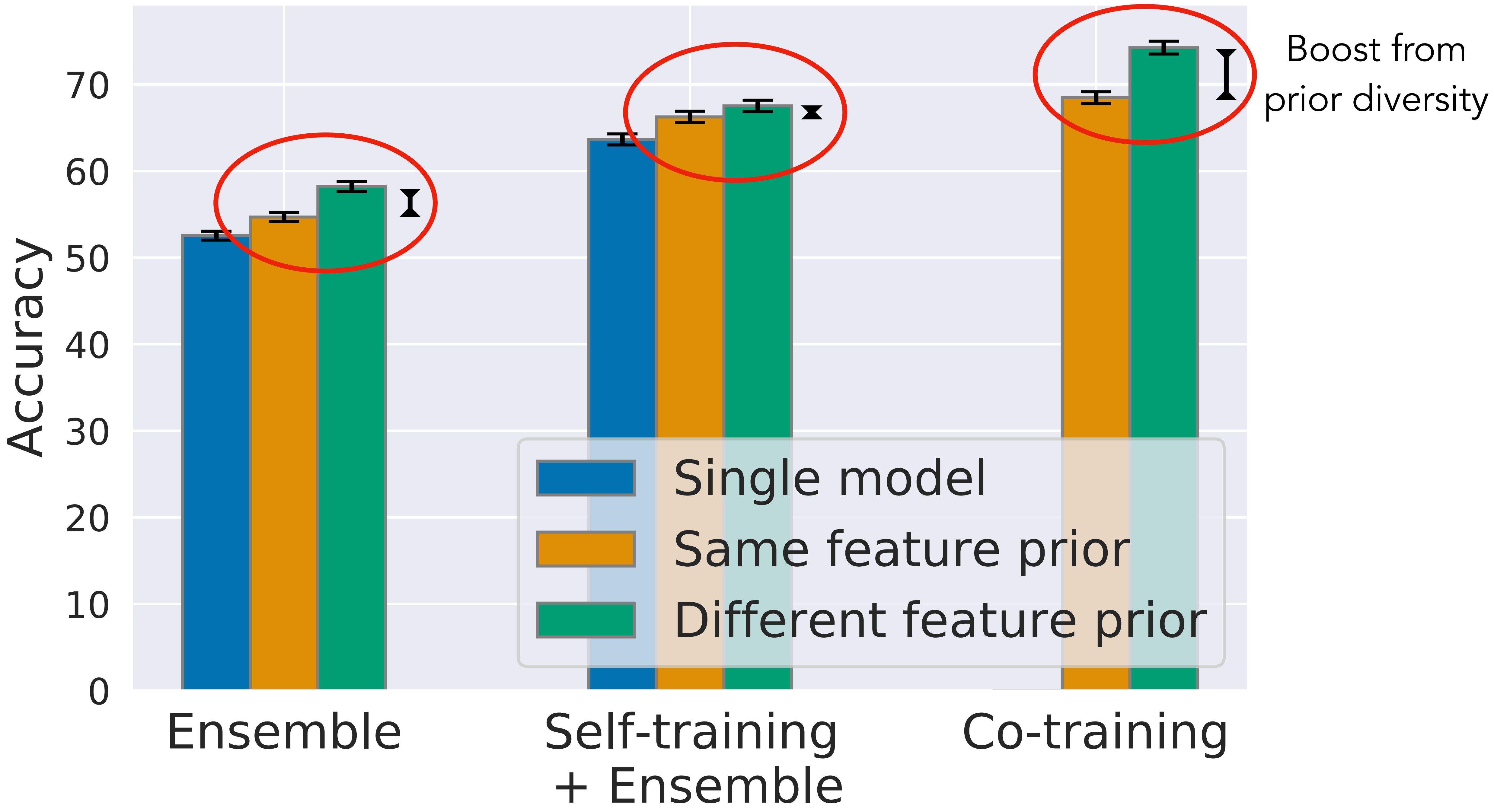
Co-training with diverse feature priors

Key idea: Different feature priors lead to models that learn **different features**

So: Models can *correct each other* during training



Indeed: Co-training with diverse features helps



Boost from prior diversity

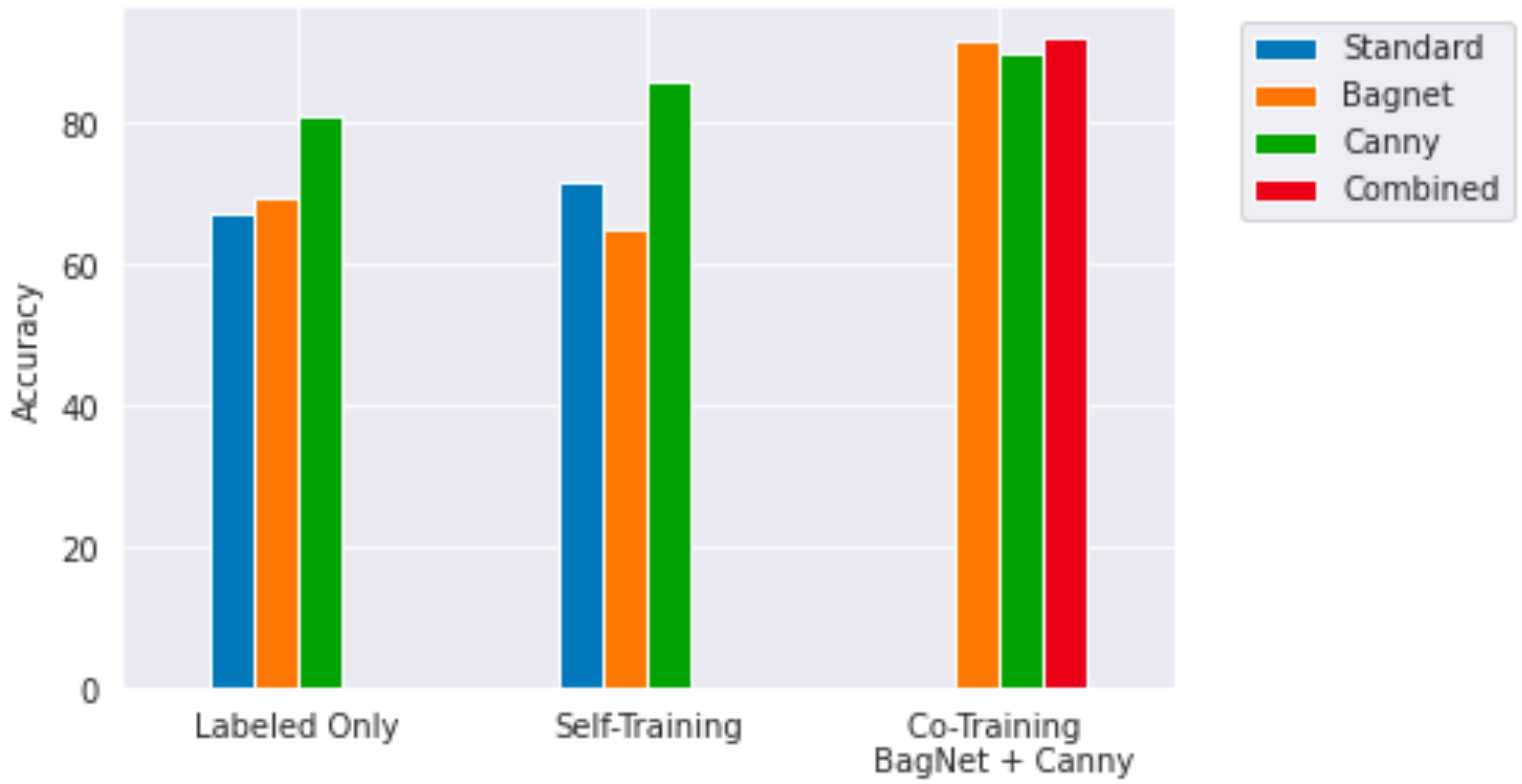
Also: Helps with avoiding learning spurious correlations

Also: Helps with avoiding learning spurious correlations

Task: CelebA gender, but all **women are blonde** during training

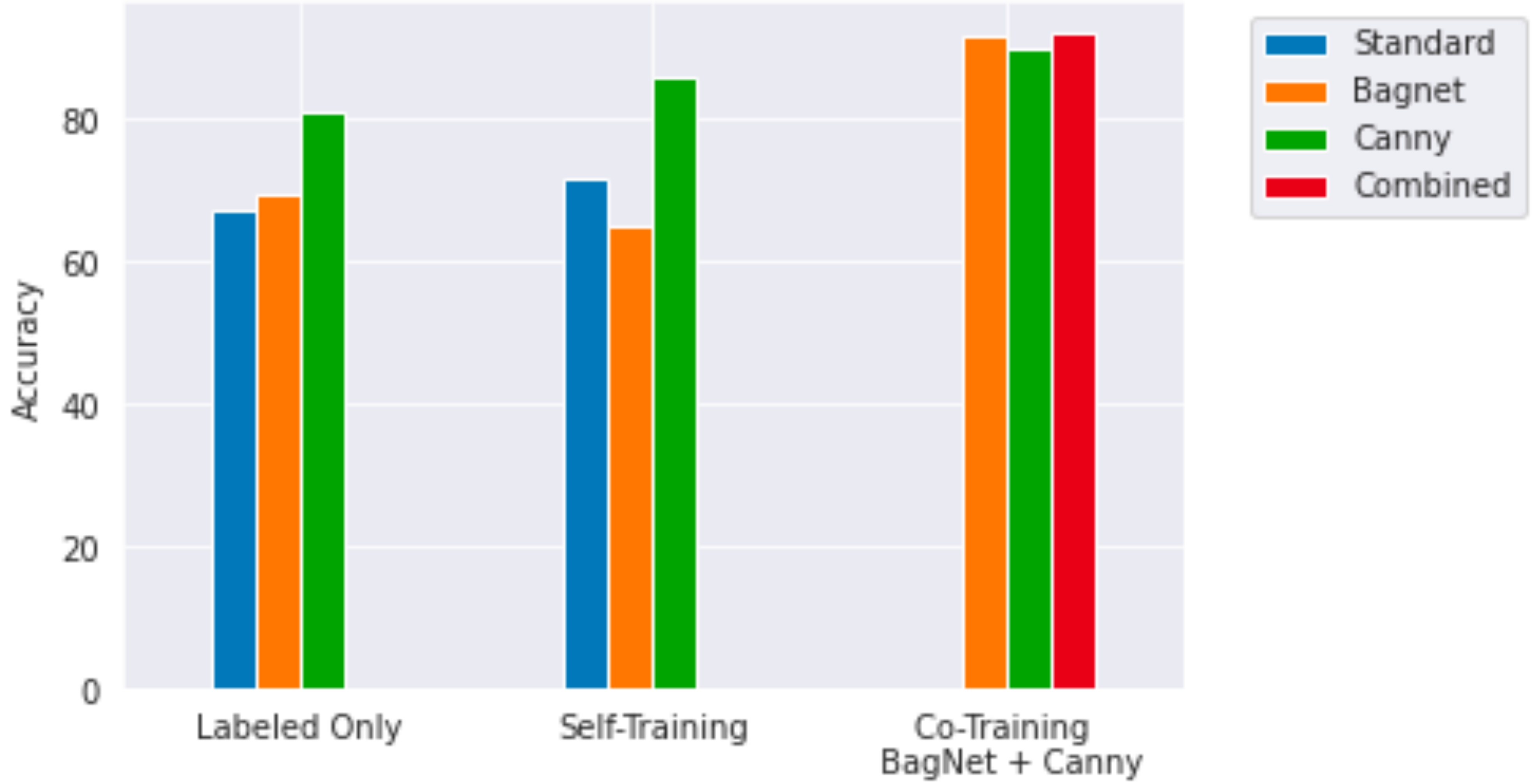
Also: Helps with avoiding learning spurious correlations

Task: CelebA gender, but all **women are blonde** during training



Also: Helps with avoiding learning spurious correlations

Task: CelebA gender, but all **women are blonde** during training



→ Models can steer each other away from **misleading features**

Key takeaway

Key takeaway

Incorporating **diverse feature priors** into training can improve generalization and help avoid spurious correlations

Key takeaway

Incorporating **diverse feature priors** into training can improve generalization and help avoid spurious correlations

Going forward

What other feature priors can we use here?

What are other ways to combine feature priors?