Toward Trustworthy AI-diagnosis of Alzheimer's Disease from MRI

Haohan Wang

(in transition to)

School of Information Sciences

University of Illinois Urbana-Champaign

Outline

- The challenges of AI diagnosis in robustness / confounding factors
  - Central hypothesis of robust machine learning
- Principled understanding trustworthy machine learning
  - Data augmentation and alignment regularization
- (early-stage) Alzheimer's disease diagnosis

Outline

- The challenges of AI diagnosis in robustness / confounding factors
  - Central hypothesis of robust machine learning

- Principled understanding trustworthy machine learning
  - Data augmentation and alignment regularization

- (early-stage) Alzheimer's disease diagnosis

Background

Alzheimer’s Disease and AI Diagnosis

Alzheimer's Disease

- a neurodegenerative disorder
- destroys memory and other mental functions

- The burdens are significant

- 6 million Americans
- By 2050, 13 million

- 1 in 3 seniors
- more than breast cancer and prostate cancer combined.

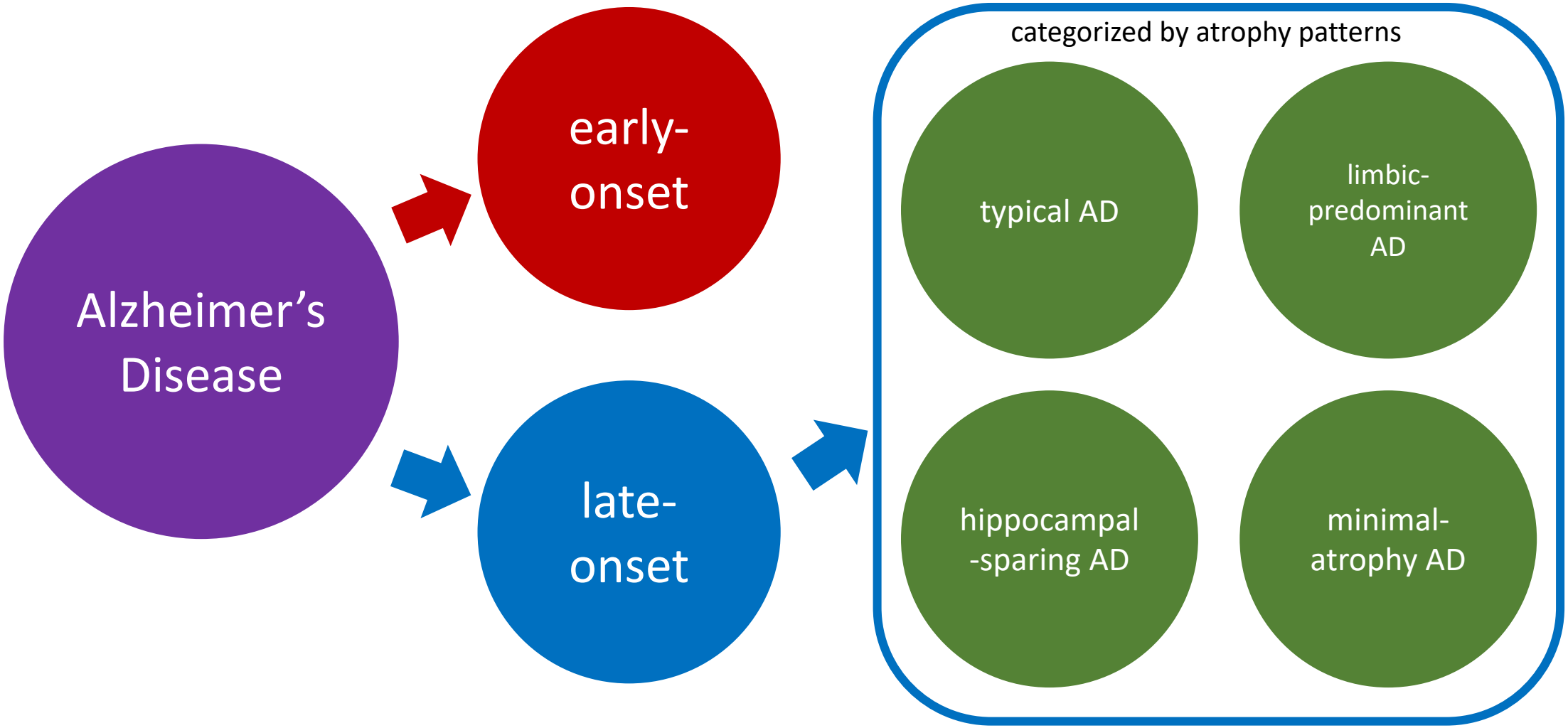
- In 2021 $355 billion
- By 2050 $1.1 trillion

- 47 % of minorities

Source: Costs of Alzheimer's to Medicare and Medicaid, https://alzimpact.org/

Source: Facts and Figures of Alzheimer, https://www.alz.org/alzheimers-dementia/

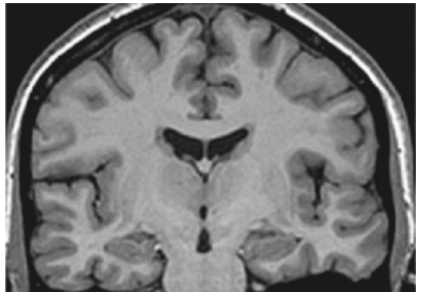
Alzheimer's Disease Is Heterogeneous



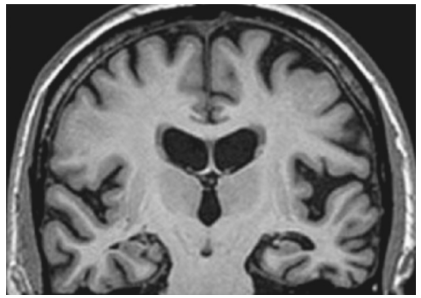
Ferreira *et al*. 2020. Neurology

Diagnosis of Alzheimer's Disease

- brain imaging is one of the tests
  - MRI: 3D image helps reveal loss of brain mass (atrophy pattern)



healthy brain



diseased brain

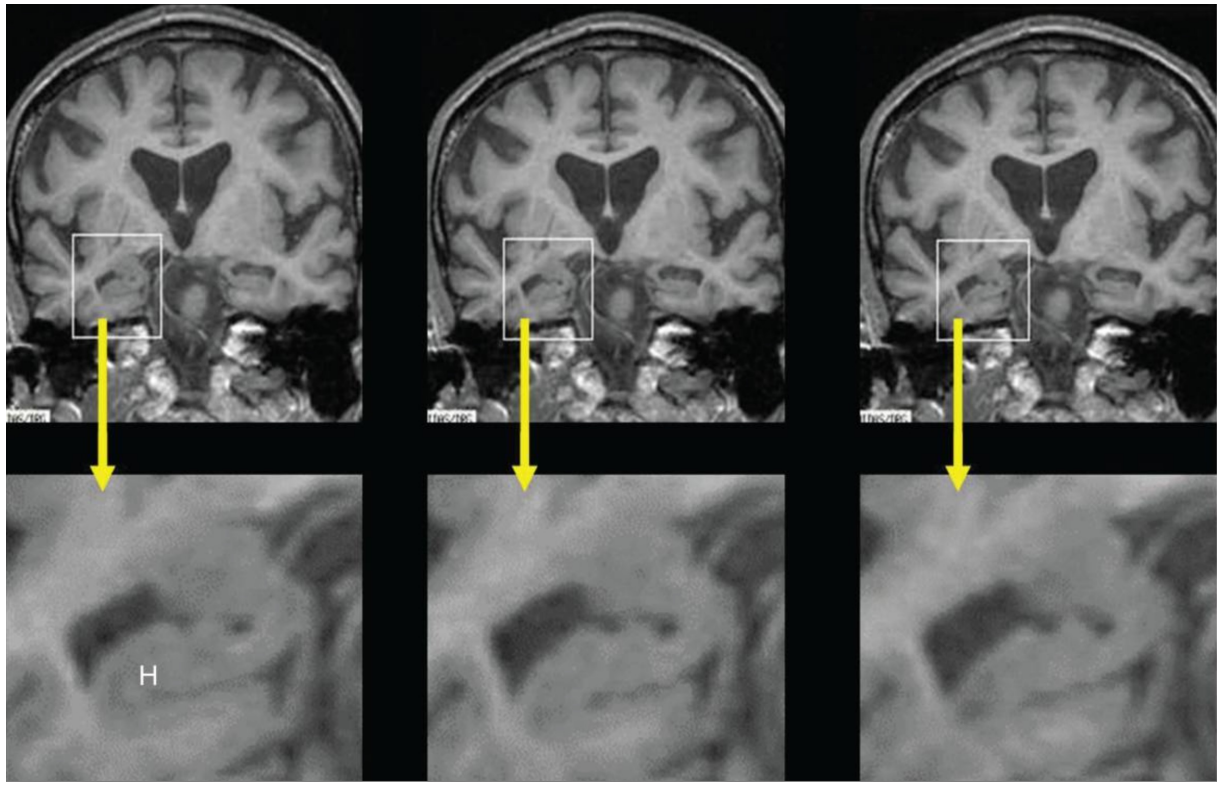
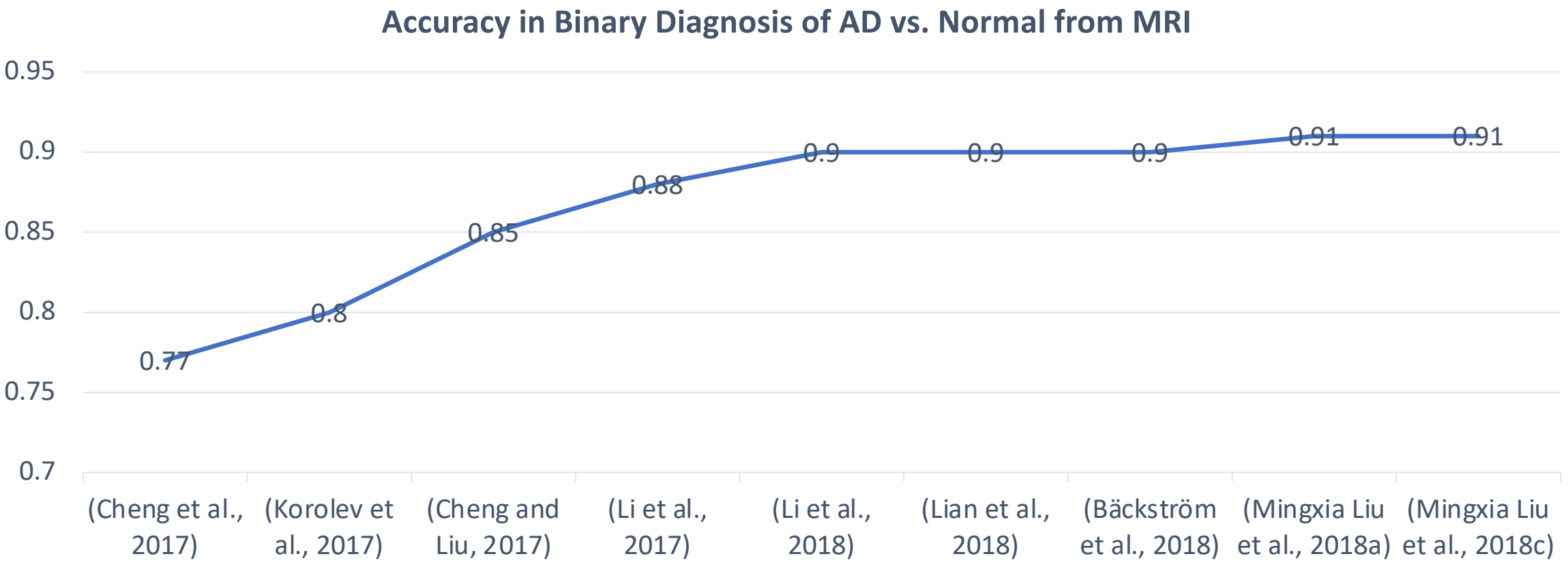


Figure: Johnson, Keith A., et al. *Cold Spring Harbor perspectives in medicine* 2012

Deep Learning Methods over Medical Imaging

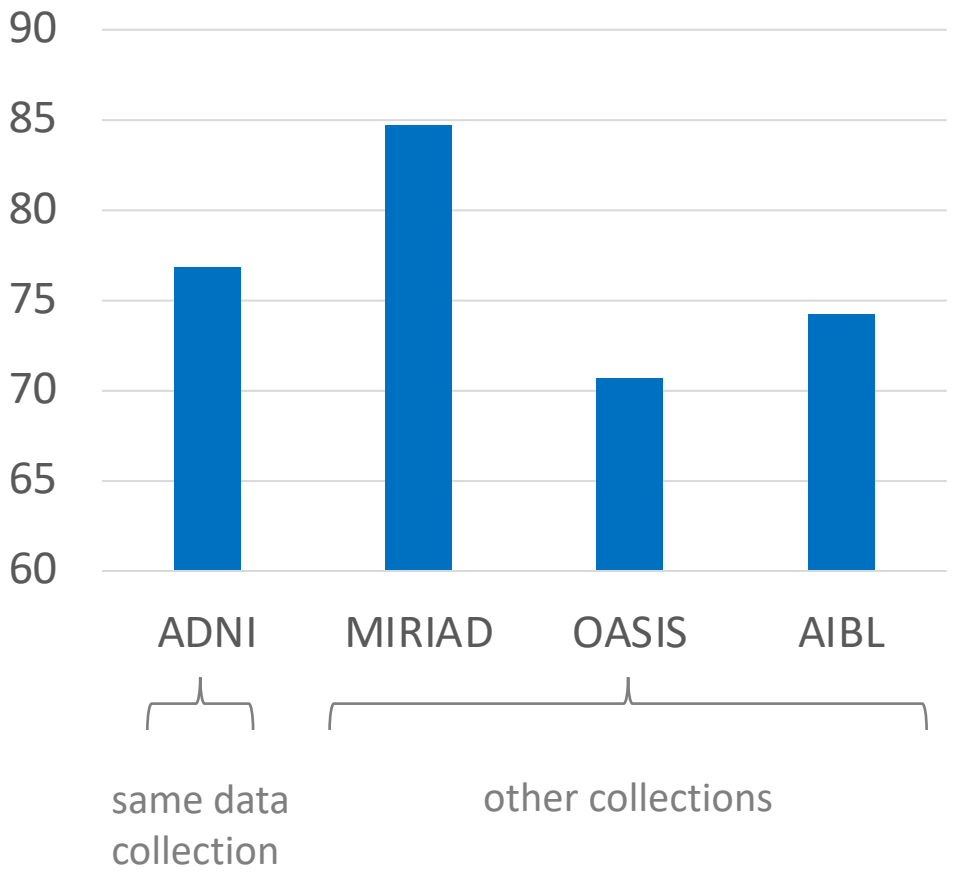
- increasing performances in deep learning diagnosis



Wen et al 2020, Convolutional Neural Networks for Classification of Alzheimer's Disease: Overview and Reproducible Evaluation

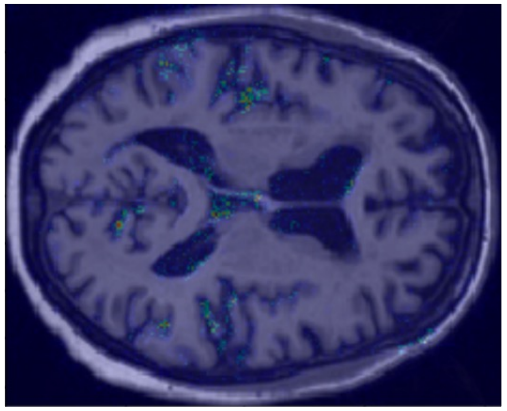
Diagnosing Alzheimer from MRI

- Best model from (Wen et al 2020)
  - 3D convolutional neural network

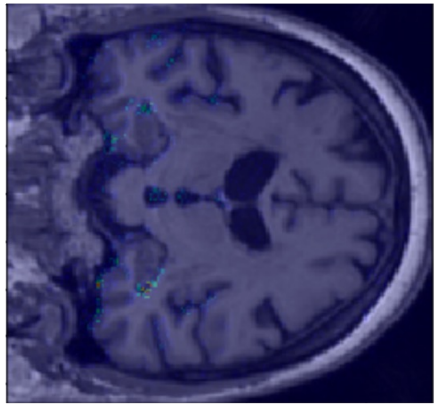


Visualizations of the Model's Decisions

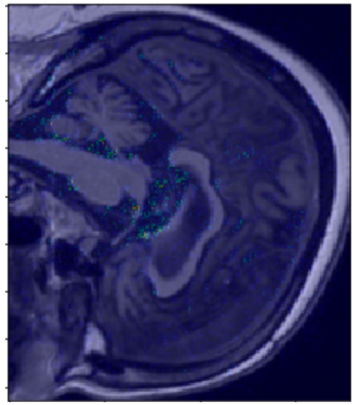
- saliency map to understand the model's diagnosis strategy
  - The model does not seem to focus on the atrophic areas of the brain



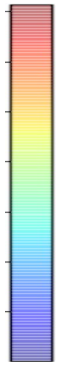
axial view



coronal view



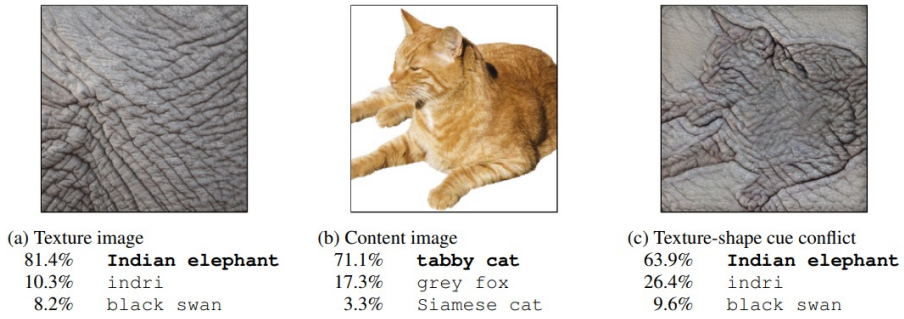
sagittal view



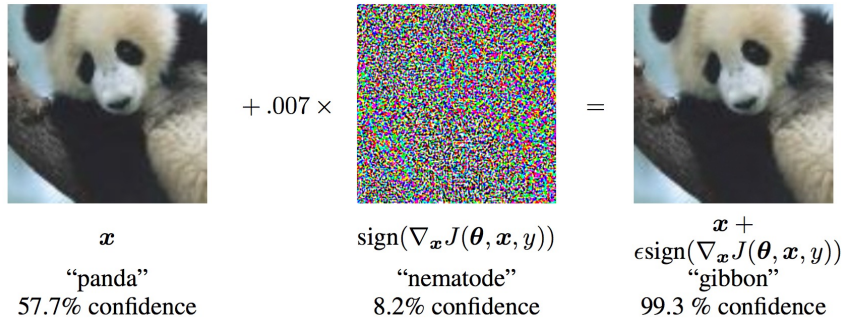
heatmap of the gradient

Challenges of image classification

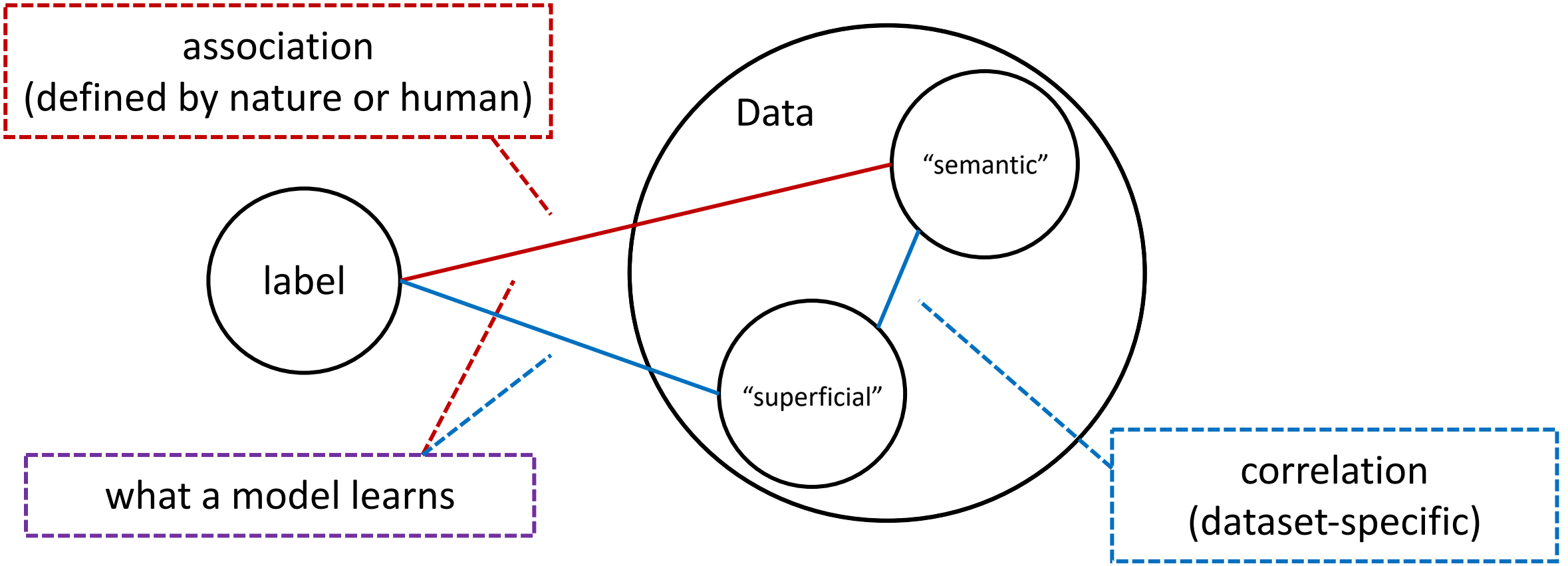
- CNN's tendency in superficial statistics
  - ImageNet-trained CNNs are biased towards texture
    - (Geirhos *et al* 2019)



  - Adversarial examples
    - (Szegedy *et al* 2013)



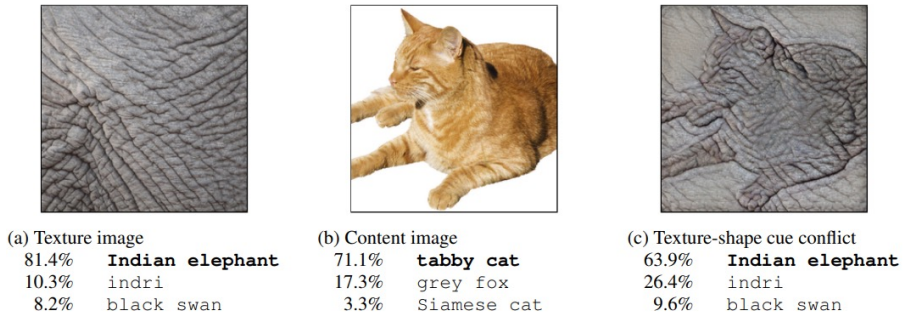
One Hypothesis for the Non-robust Behaviors

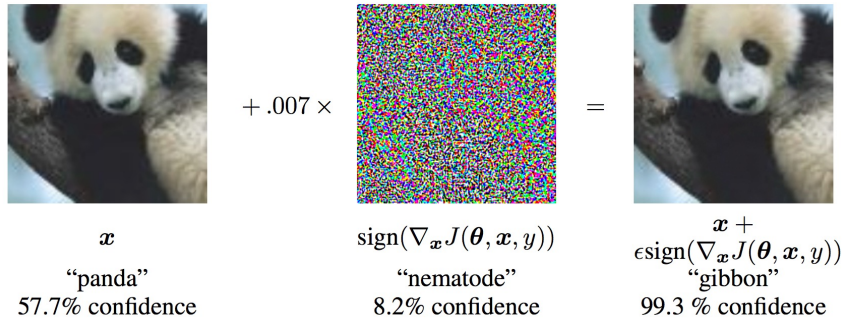


Wang, *et al*., High-frequency component helps explain the generalization of convolutional neural networks. CVPR 2020 Oral

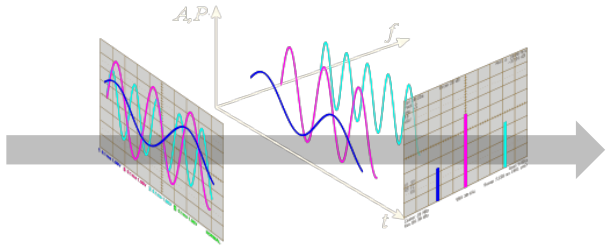
Challenges of image classification

- CNN's tendency in superficial statistics
  - ImageNet-trained CNNs are biased towards texture
    - (Geirhos *et al* 2019)
  - The Origins of texture bias are in the data
    - (Hermann *et al* 2020)
  - Adversarial examples
    - (Szegedy *et al* 2013)
  - Adversarial examples are features
    - (Ilyas *et al* 2020)

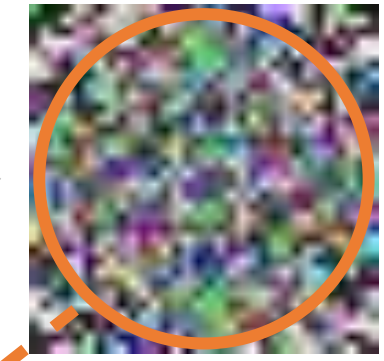




High/Low-frequency Reconstructed Images



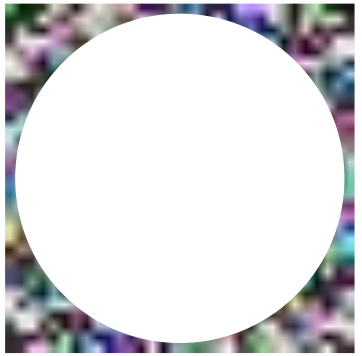
Fourier transform



a predetermined radius

low-frequency

reconstructed



high-frequency

reconstructed

Misalignment between Human and Model

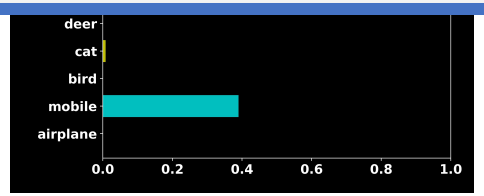
original image

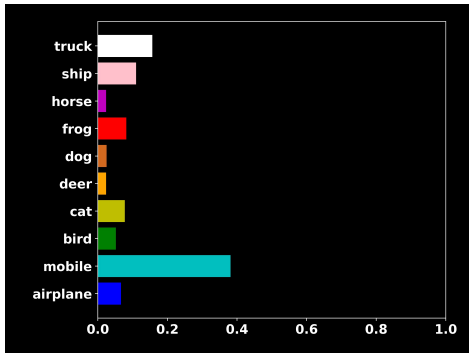
low-frequency reconstructed image

high-frequency reconstructed image

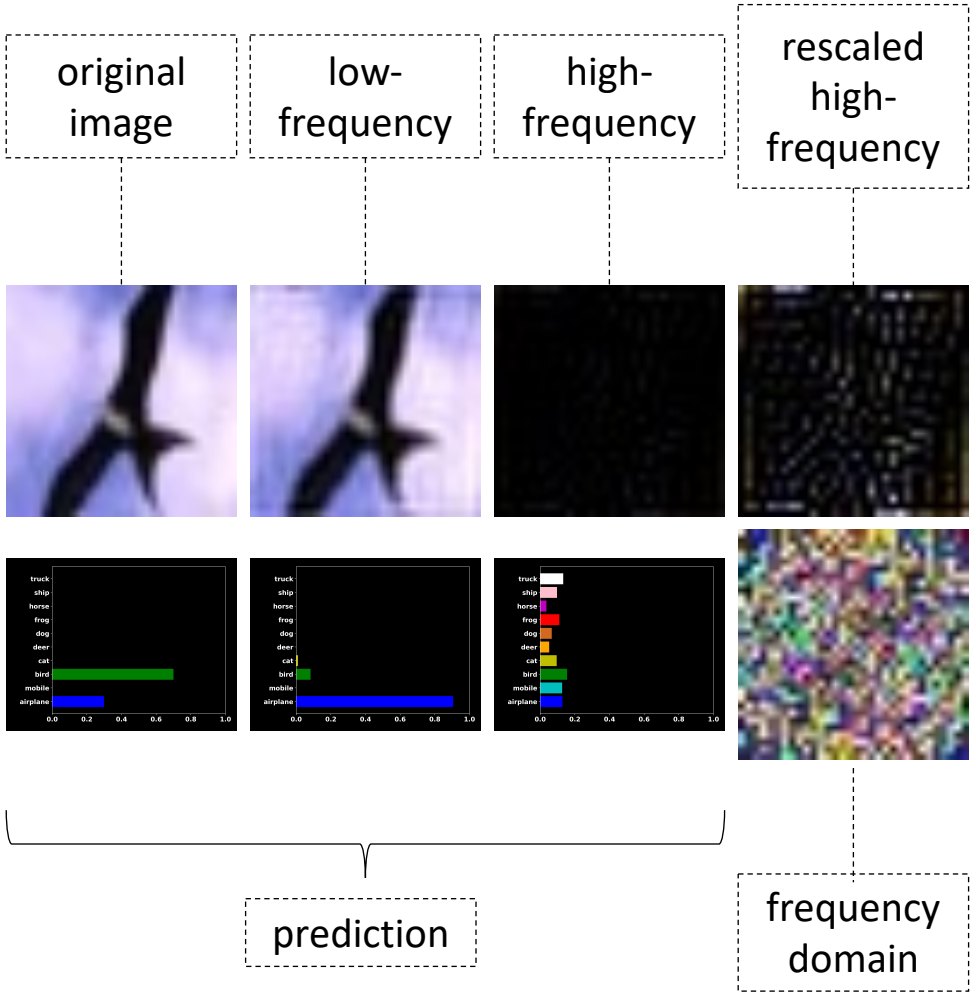
pixel values rescaled to be 0-255

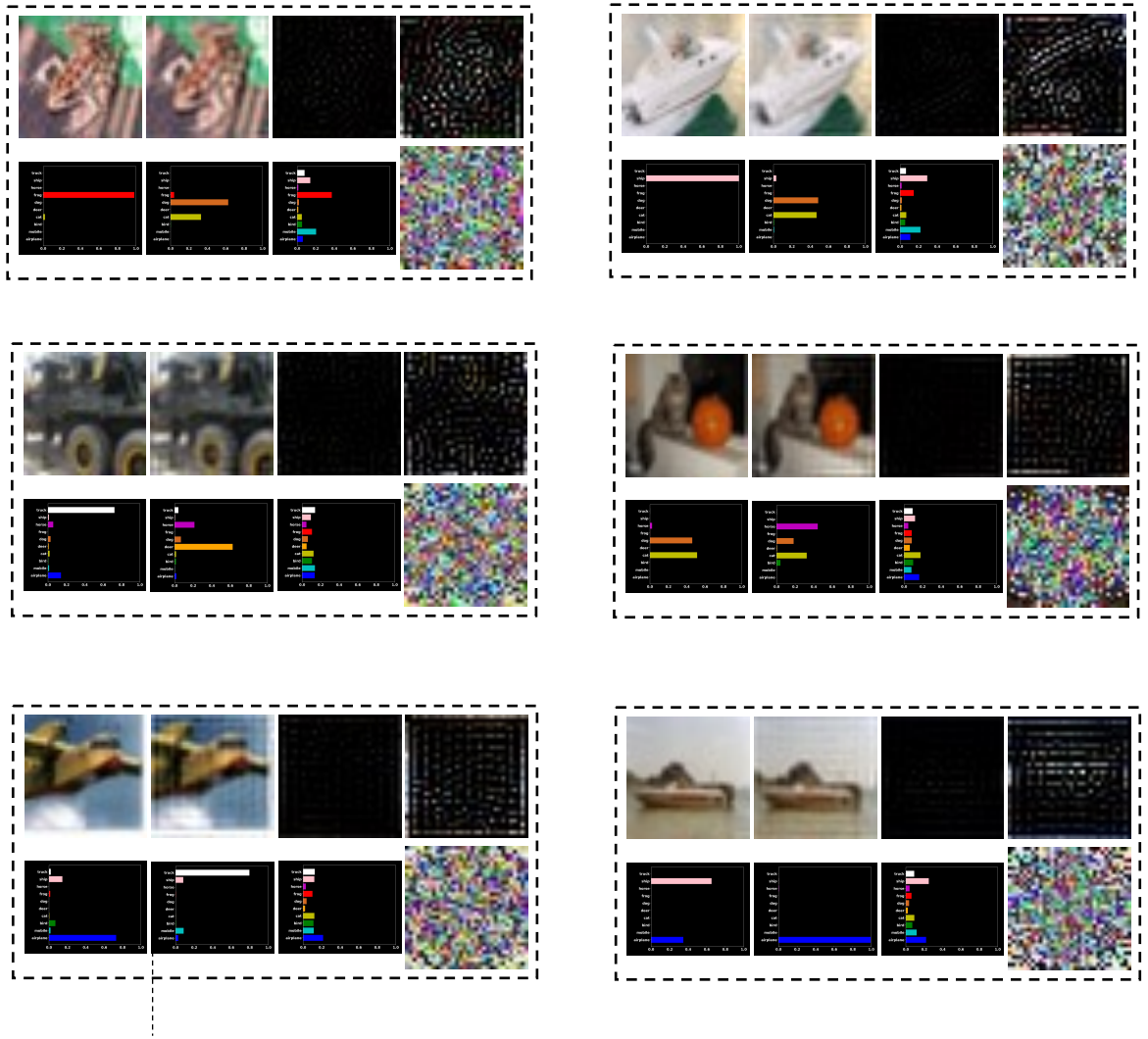
model trained on original images





Additional Examples





Wang, *et al*., High-frequency component helps explain the generalization of convolutional neural networks. CVPR 2020 Oral

Challenges in Medical Imaging

- Models tend to learn different superficial features

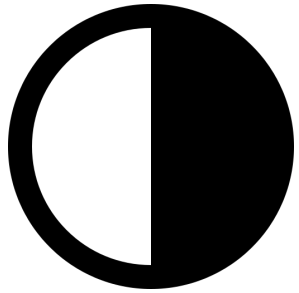
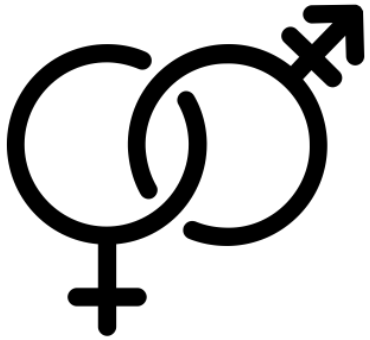
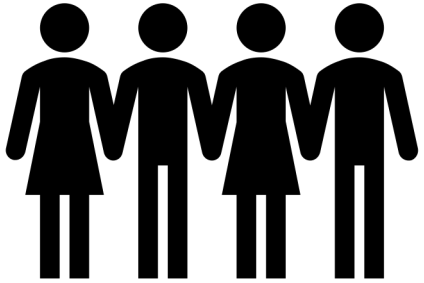


image contrast



gender



individual traits

Wang, *et al*., 2020. PSB (oral)

Outline

- The challenges of AI diagnosis in robustness / confounding factors
  - Central hypothesis of robust machine learning

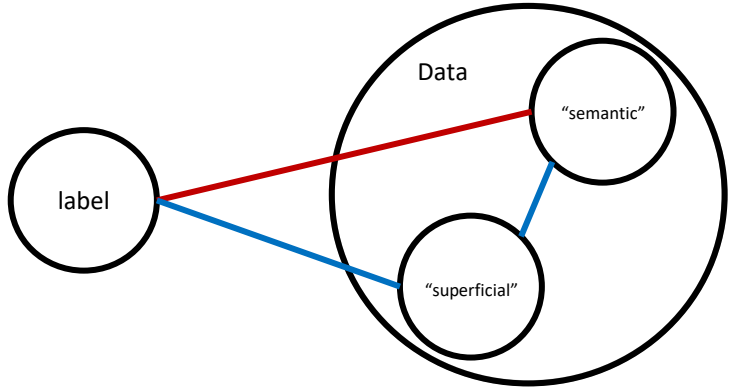
- Principled understanding trustworthy machine learning
  - Data augmentation and alignment regularization

- (early-stage) Alzheimer's disease diagnosis

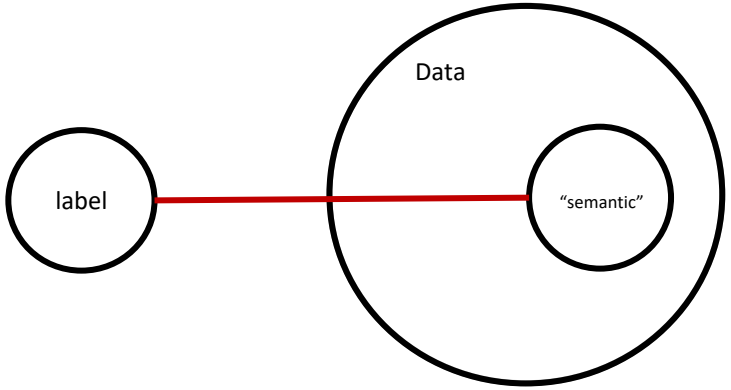
Principled Understanding

Machine Learning across Datasets

training domain
(when the model is trained in lab)



testing domain
(when the model is used)

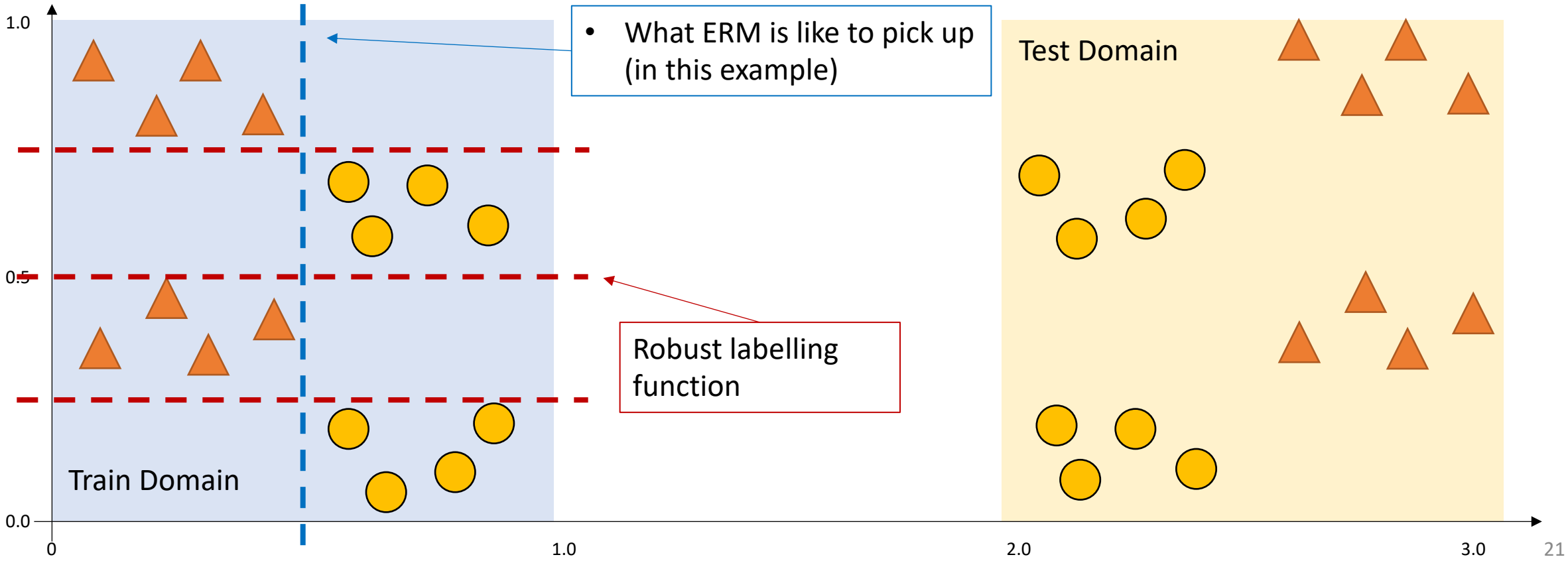


Test domain is similar to, but different from train domain

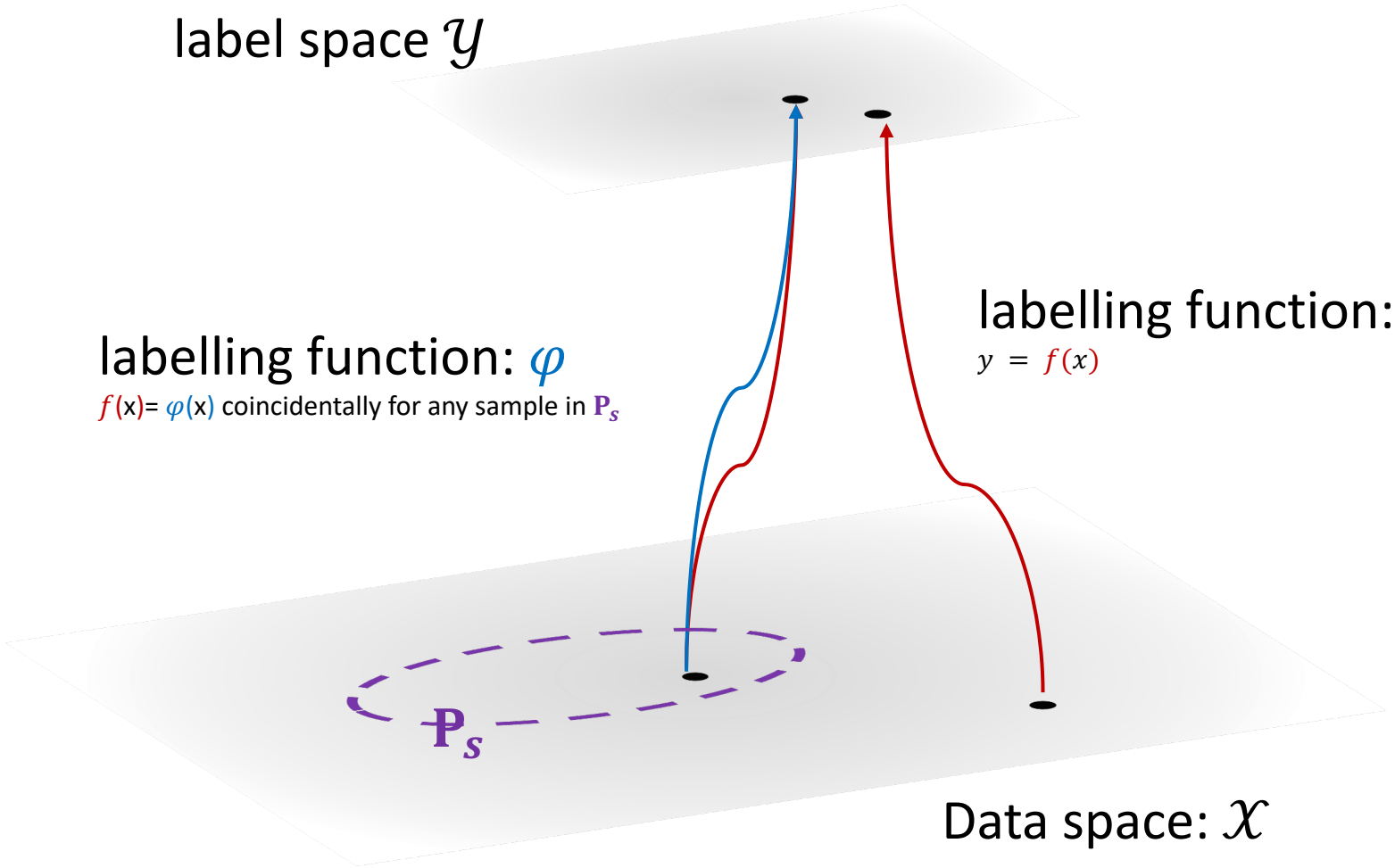
domain adaptation (Ben-David 2007), domain generalization (Muandet et al., 2013)

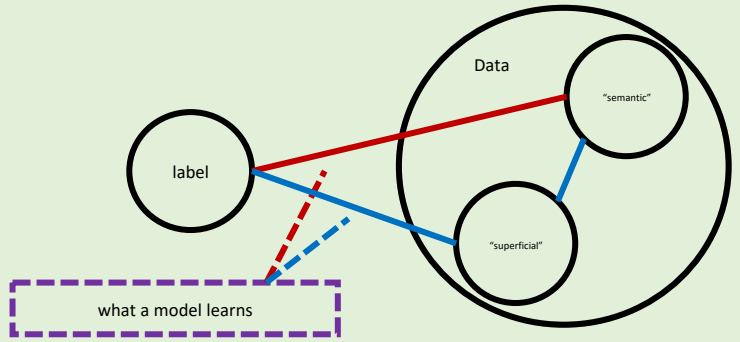
Similar but Different: intuitively

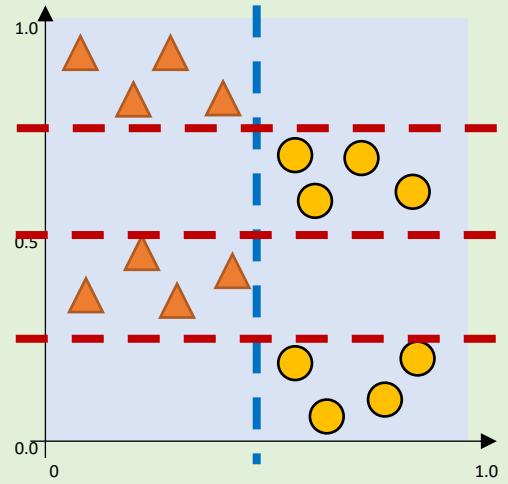
- Similar: there is a shared labelling function
- Different: the training domain has an additional labelling function



Similar but Different: formally







Generalization Error Bound in I.I.D

$$\varepsilon_{\mathbf{P}_t}(\theta) \leq \hat{\varepsilon}_{\mathbf{P}_s}(\theta) + \phi(\Theta, n, \delta)$$

Generalization Error Bound with Superficial Features

$$\varepsilon_{\mathbf{P}_t}(\theta) \leq \hat{\varepsilon}_{\mathbf{P}_s}(\theta) + \phi(\Theta, n, \delta) + c(\theta)$$

$$c(\theta) = \frac{1}{n} \sum_{(x,y)\in(X,Y)_{\mathbf{P}_S}} \mathbb{I}[\theta(x) = y]\, r\big(\theta, \mathcal{S}(\varphi, x)\big)$$

$c(\theta)$: the accuracy gain because θ learns φ

usually guided by additional domain knowledge

a robust model

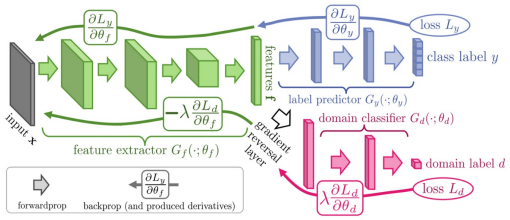
- small $\hat{\varepsilon}_{\mathbf{P}_s}(\theta)$
- small $c(\theta)$
  - $\mathcal{S}(\varphi, x)$ (i.e., the features the superficial function uses) is given
  - φ is given

The Principled Understanding Connects to Other Methods

$\mathcal{S}(\varphi, x)$ { $\mathbb{I}[\theta(x) = y] \leq 1$ → upper bound of $c(\theta)$ } ⇨ minimizing $\hat{\varepsilon}_{\mathbf{P}_S}(\theta)$ and bound ⇔ worst-case data augmentation $\min_{\theta} \sum_{(x,y)\in(X,Y)_{\mathbf{P}_S}} \max_{z_{\mathcal{S}(\varphi,x)}\in\Omega(x)_{\mathcal{S}(\varphi,x)}} l(\theta(z), y)$

$$\varepsilon_{\mathbf{P}_t}(\theta) \leq \hat{\varepsilon}_{\mathbf{P}_s}(\theta) + c(\theta) + \phi(\Theta, n, \delta),$$

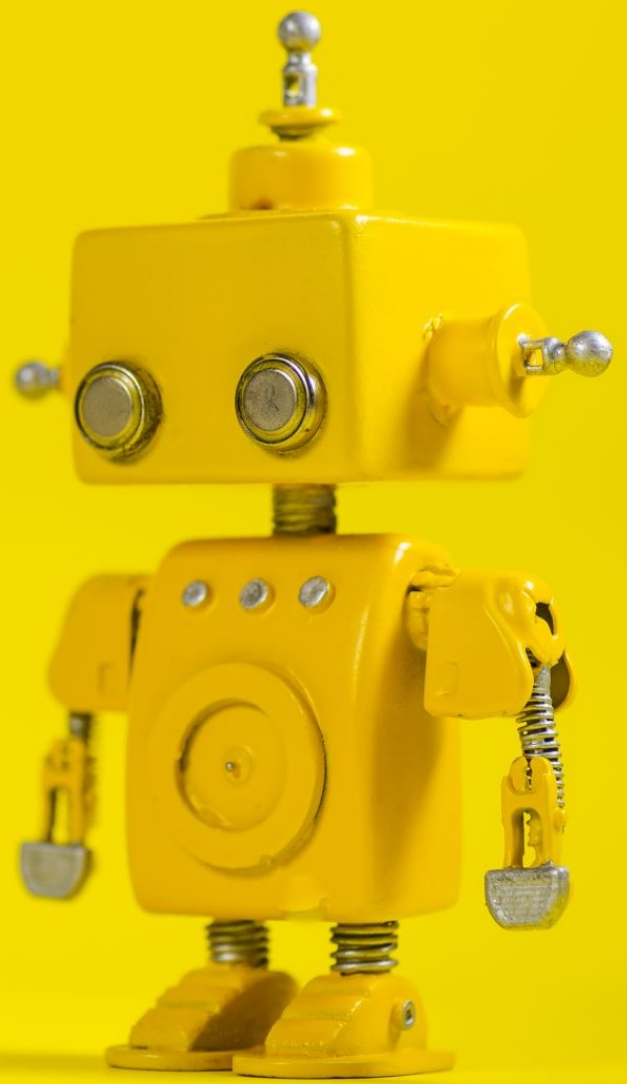
(Wang, et al. 2019, ICLR)
(Wang, et al. 2019, NeurIPS)
(He, Zhang, Wang 2019, EMNLP, WS)



(Wang, et al. 2020, ECCV)
(Wang, et al. 2022, CVPR)

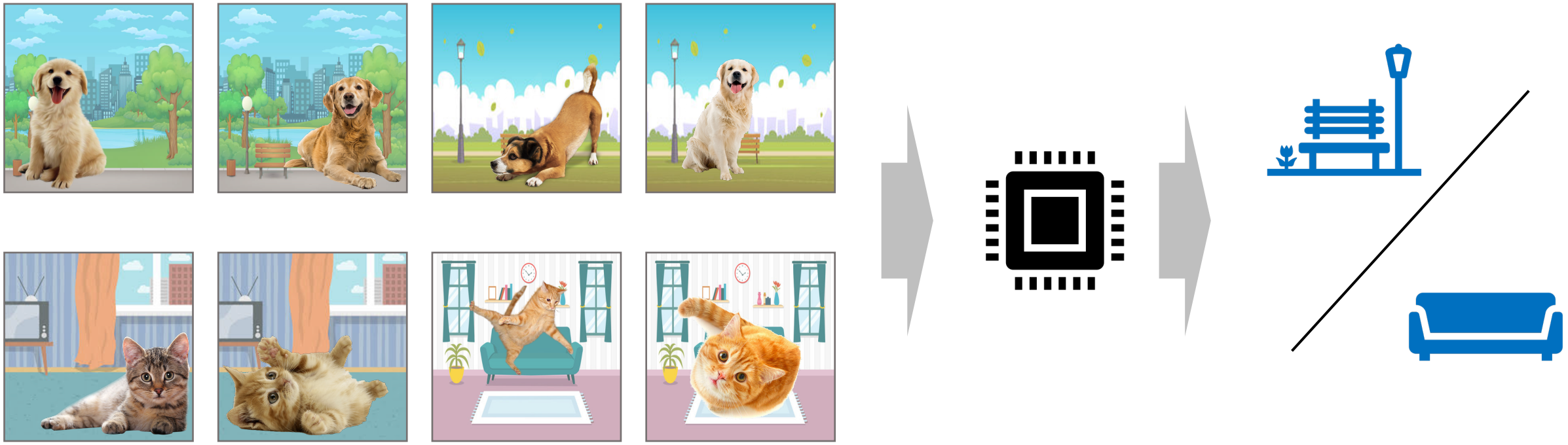
Method Development

consistency regularization
w. data augmentation



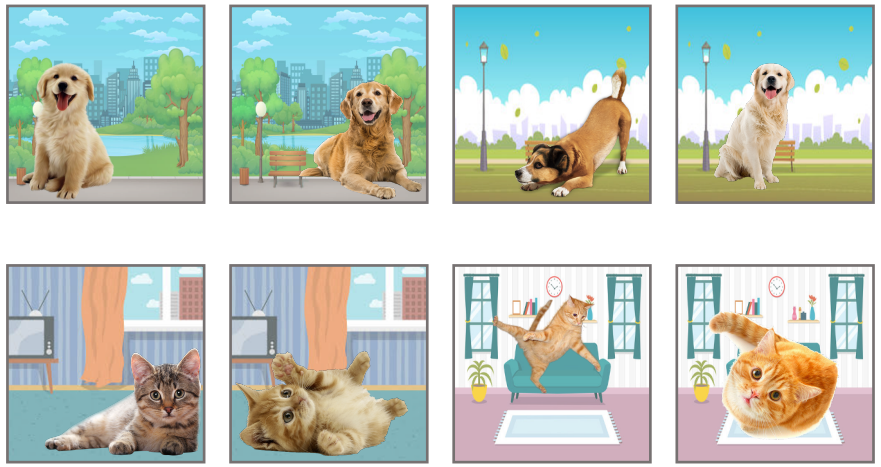
The Outdoor Dog vs. Indoor Cat Classification

- The model may learn the superficial features (the background).

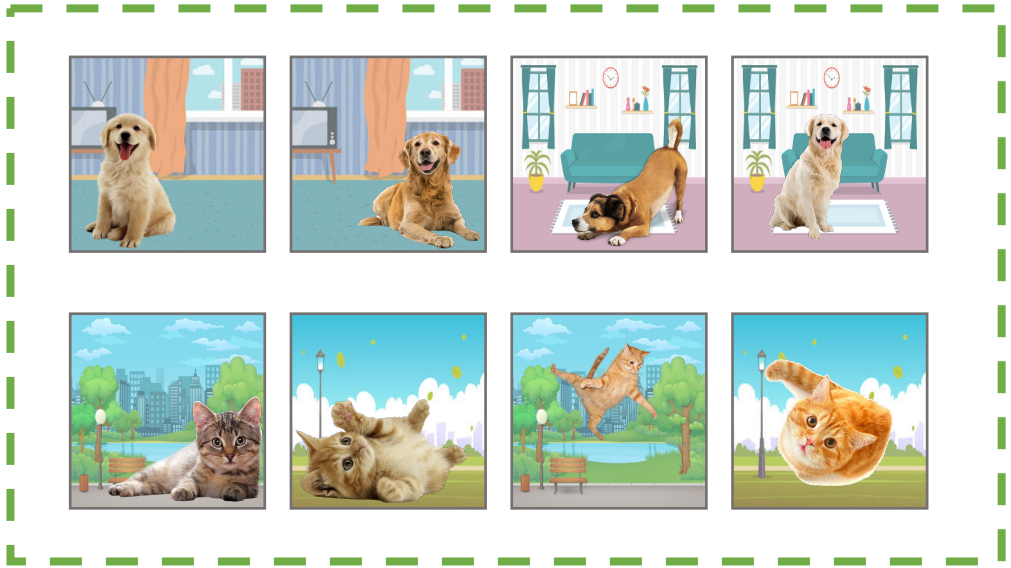


Data Augmentation

- Augmented data are usually generated before the training process,
  - with a predefined set of augmentations over the superficial features.



original data

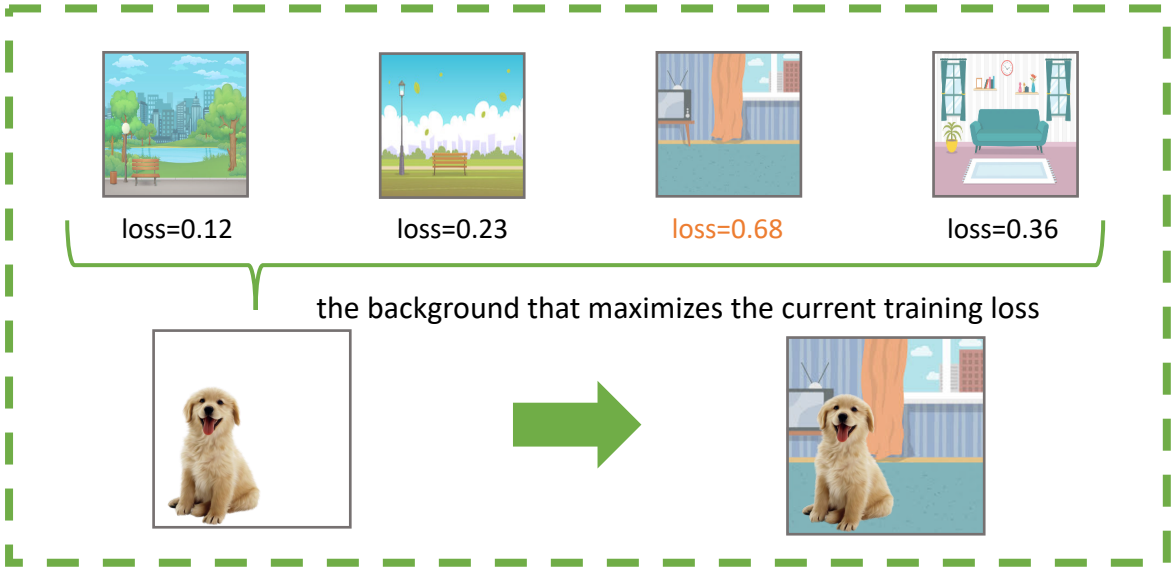


augmented data

Worst-case Data Augmentation

- Augmented data are generated during training to maximize the training loss,
  - with a predefined set of augmentations over the superficial features.
  - also known as adversarial training

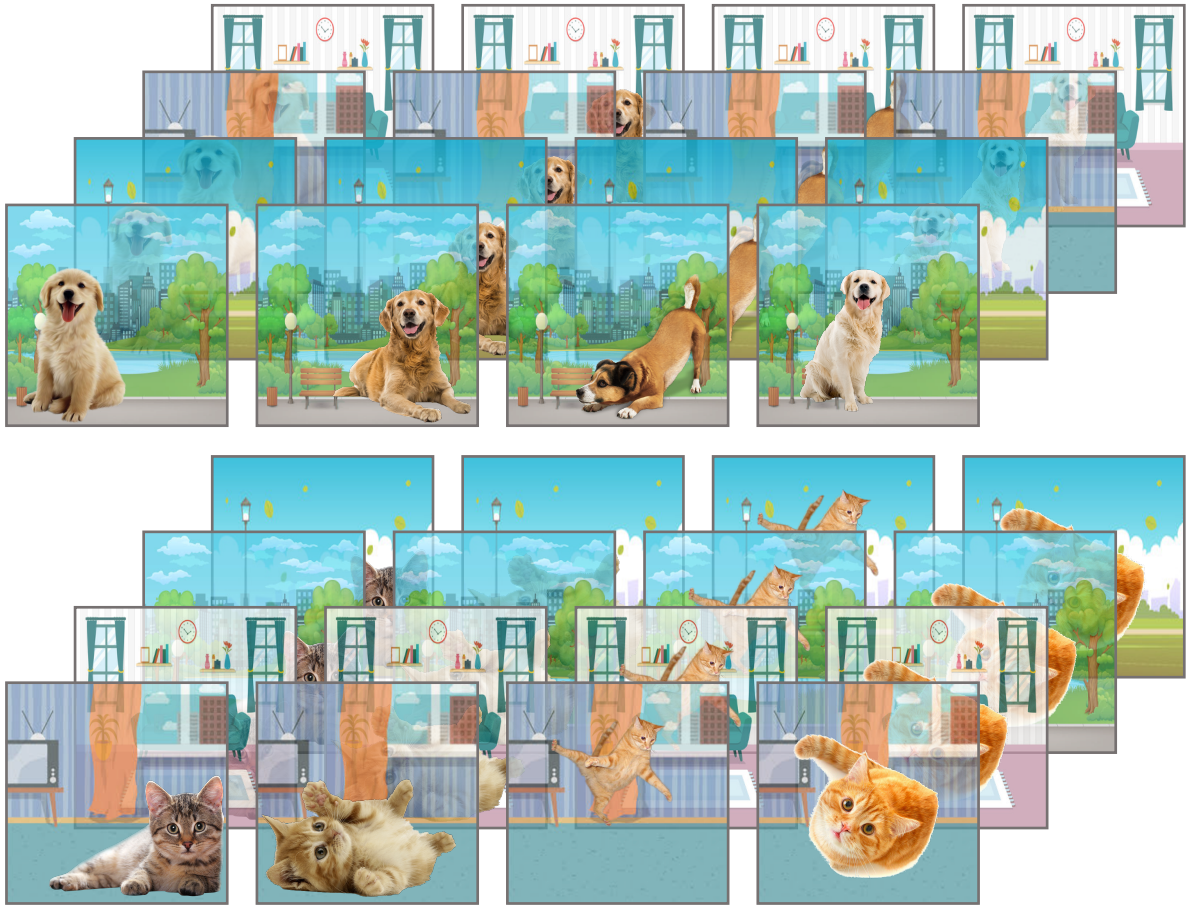
original data

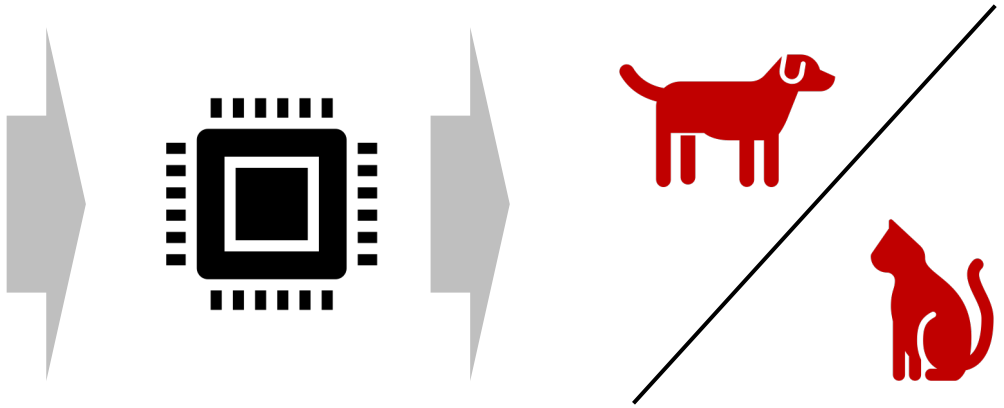


augmented data

Training with Augmented Data

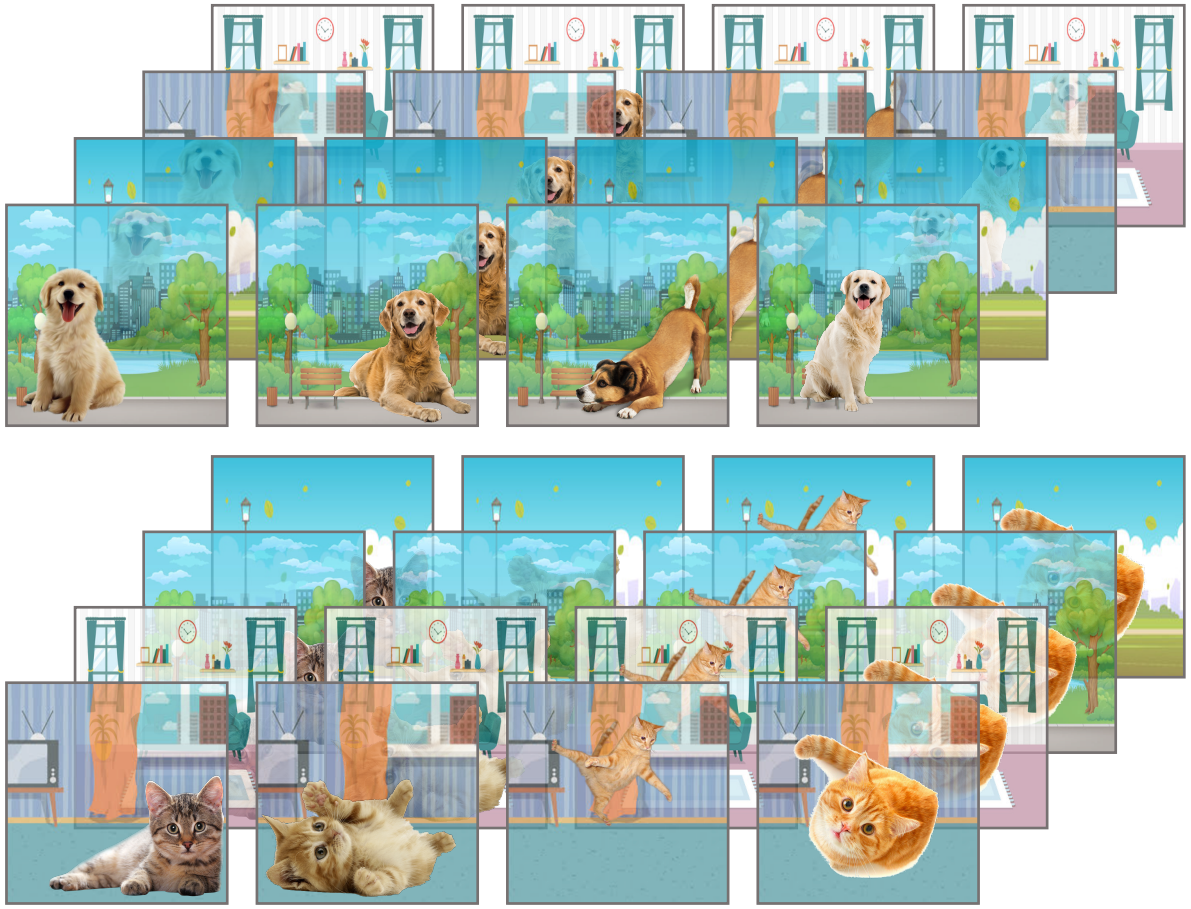
- It might push the model to discard superficial features and learn semantic features.

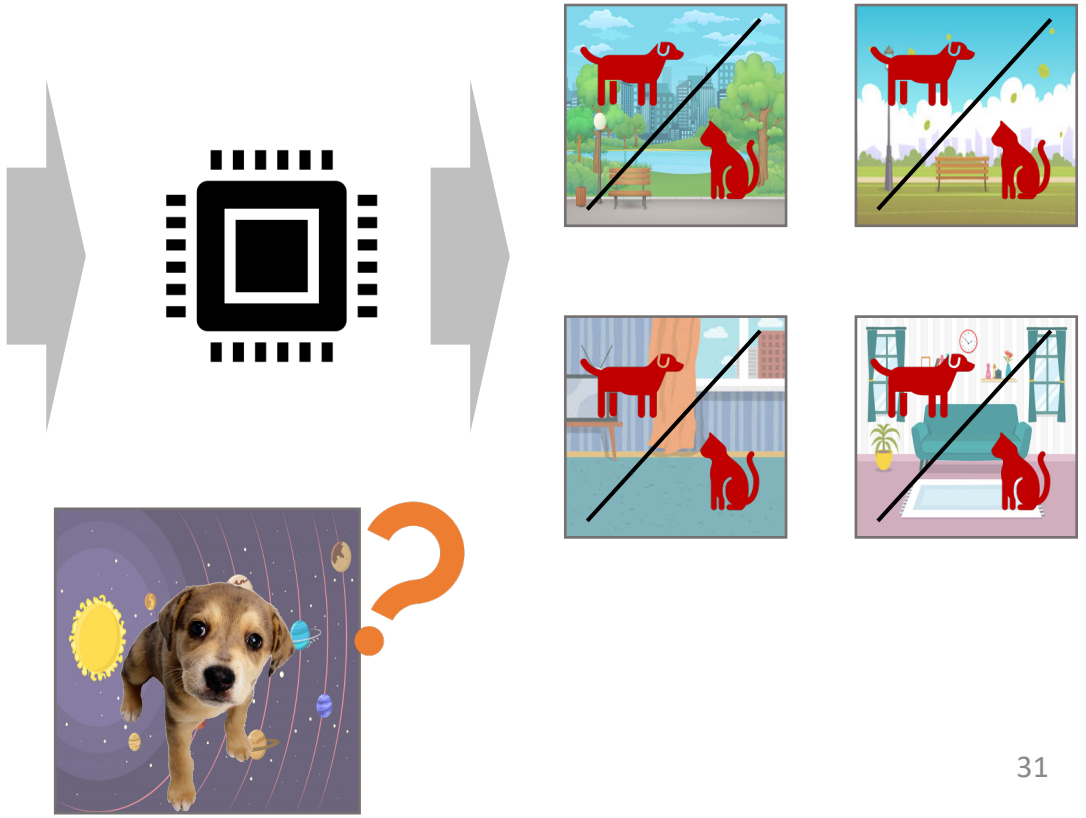




Training with Augmented Data: A Potential Limitation

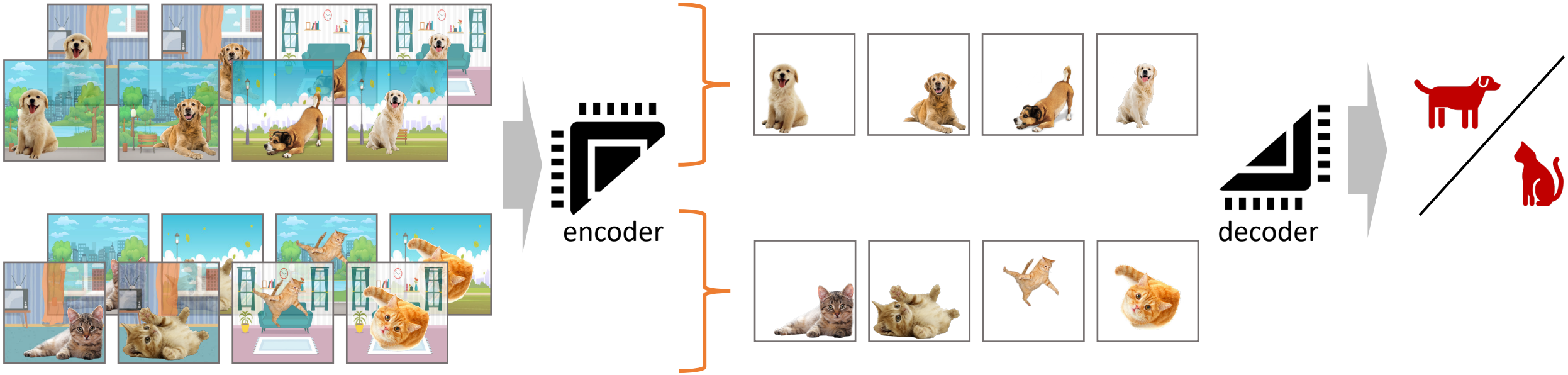
- It might instead push the model to learn semantic features conditioning on the superficial features.





A Classifier Might Not Be Necessary

- Alignment regularization pushes the model to learn the same representations from an image and its augmented counterpart.



Alignment Regularization

- $\ell 2$ distance and cosine similarities
  - internal representations
  - speech recognition
  - (Liang et al., 2018)
- Squared $\ell 2$ distance
  - logits
  - adversarial robust vision models
  - (Kannan et al., 2018)
- KL divergence
  - softmax outputs
  - adversarial robust vision models
  - (Zhang et al., 2019a)
- Jensen–Shannon divergence
  - embeddings
  - texture invariant image classification
  - (Hendrycks et al., 2020)
- and many others…

If there is a general method that can work well across applications, and enjoys some theoretical support?

Our Solution:
Squared l2 Norm as Alignment Regularization

- We recommend using squared l2 norm as regularization

Empirically

We conduct a set of experiments and find out squared l2 norm is the best choice

Theoretically

We complement our empirical study with a formal proof

- We show a bounded worst-case error (robustness)
- We connect the regularization to the measure of invariance

Comparison to the Top-performing Methods: Experiment Setup

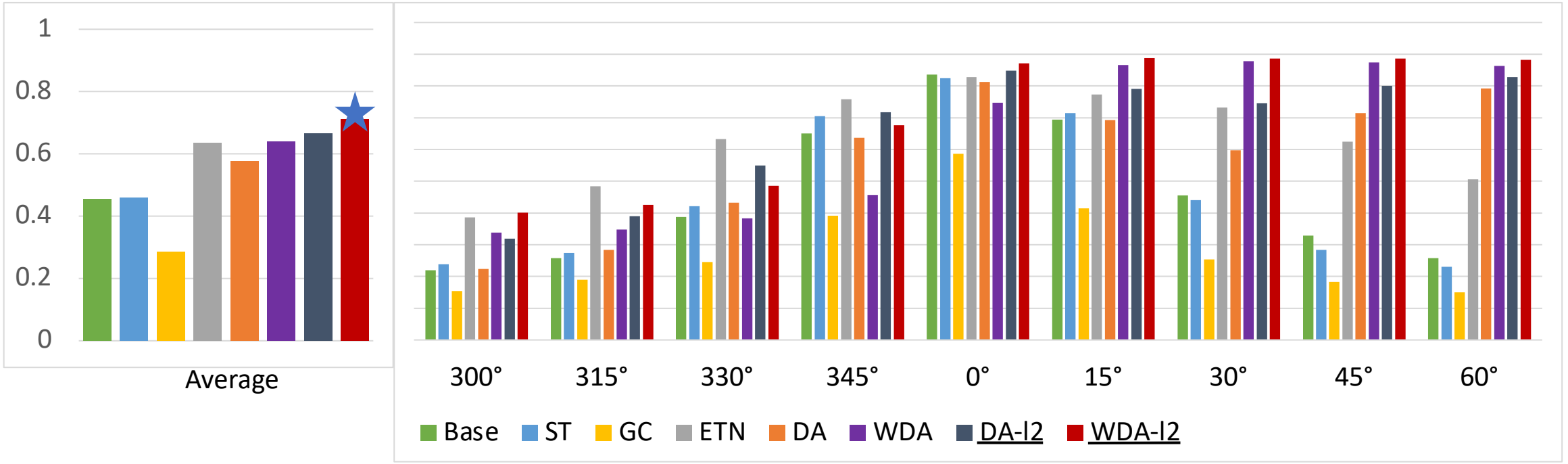
- We experiment in three different tasks
  - rotation-invariant classification
  - texture-perturbed image classification
  - cross-domain image classification

- For each of these three tasks,
  - we compare to the top-performing methods specially designed for the task.

- Our method will use the same generic approach:
  - augmentation functions are the same as in the synthetic experiment.

Comparison to the Top-performing Methods: Results

- Rotation-invariant Classification



The simple method we identified can compete with top-performing methods specially designed for each task.

AlignReg Package

AlignReg

Alignment Regularization with Data Augmentation for Robust and Invariant Machine Learning

- The method can be used through a single line of code
  - In both PyTorch and TensorFlow
- Installation
  - Pip install alignreg
  - Or from https://github.com/jyanln/AlignReg
- Usage

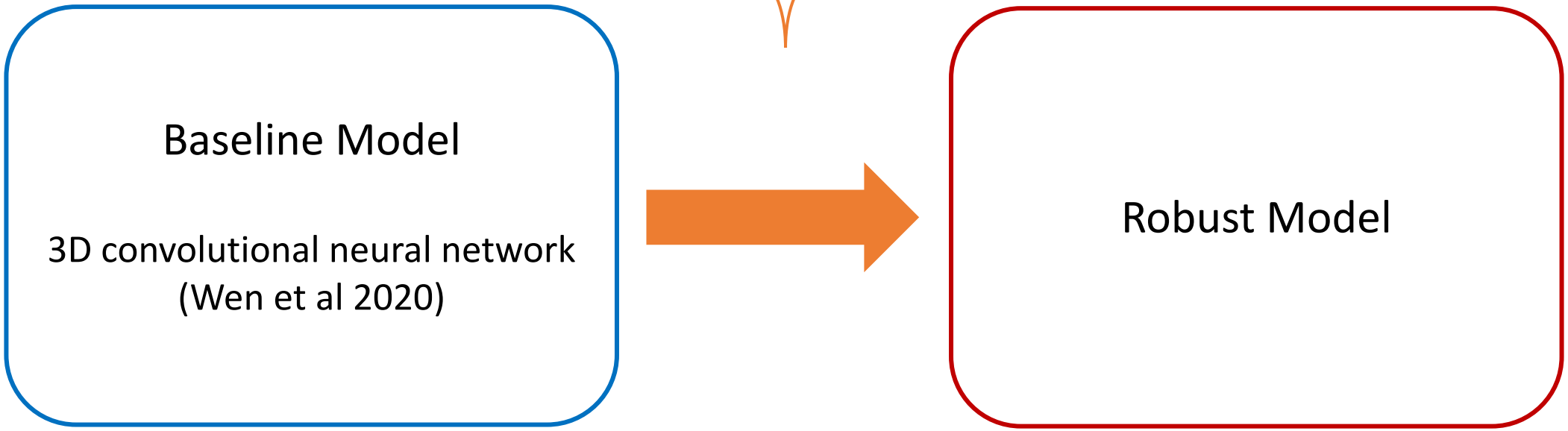
```
# Set regularization factor
l2_lambda = 0.01

# Set augmentation functions
# Here we use the default list from the augment module
augmentations = tf_default_augmentations

# Pass variables into training function
acc, loss_hist = tf_train(ds_train,
                          ds_test,
                          model,
                          optimizer,
                          epochs,
                          loss,
                          l2_lambda,
                          augmentations=tf_default_augmentations,
                          lazy_augmentation=True)
```

Outline

- The challenges of AI diagnosis in robustness / confounding factors
  - Central hypothesis of robust machine learning
- Principled understanding trustworthy machine learning
  - Data augmentation and alignment regularization
- (early-stage) Alzheimer's disease diagnosis

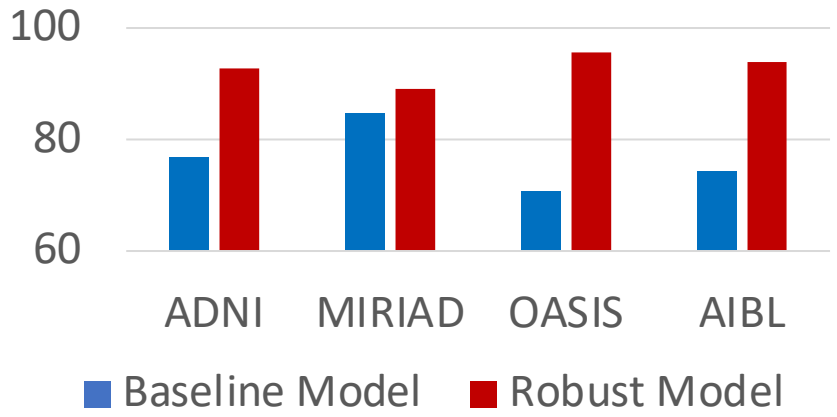
Back to Alzheimer's Diagnosis

$$\varepsilon_{\mathbf{P}_t}(\theta) \leq \hat{\varepsilon}_{\mathbf{P}_s}(\theta) + c(\theta) + \phi(\Theta, n, \delta),$$

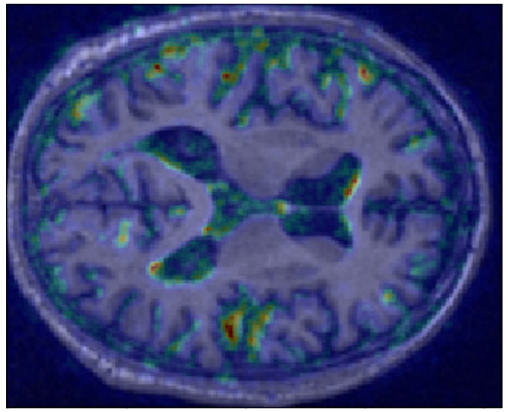


Results

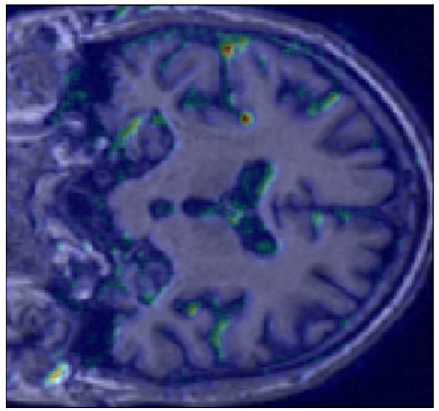
- The robust method can achieve much more stable performances across the datasets



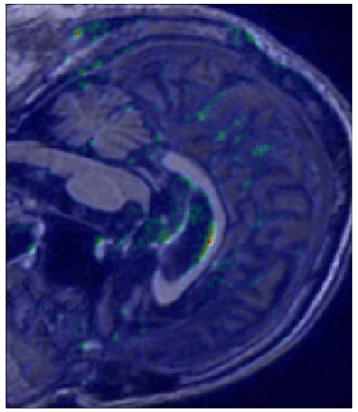
- The model focuses on the atrophic areas of the brain



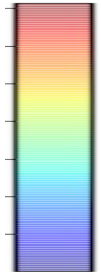
axial view



coronal view



sagittal view



heatmap of the gradient

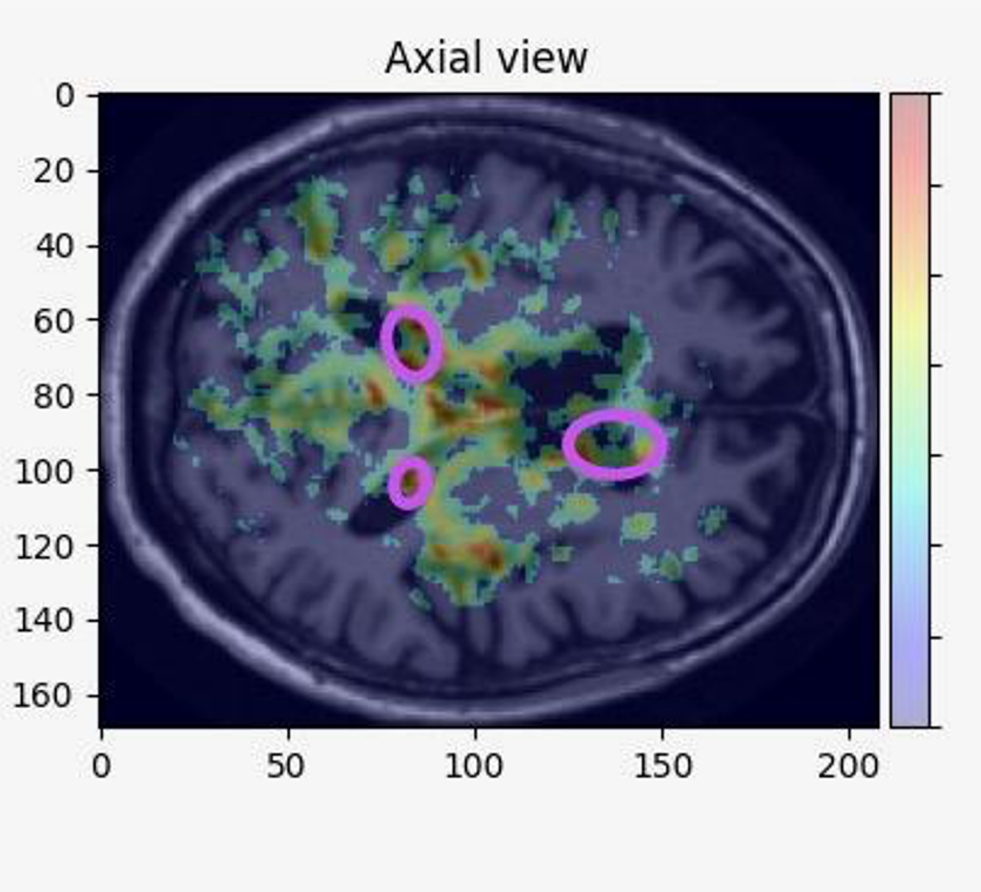
Early-Diagnosis of Alzheimer's Disease through MRI (on-going)

- Predict whether the subject will progress into AD or recover back to normal in the future when the subject is at MCI phase
  - Setup:
    - Subjects must be diagnosed as MCI at first session, and classify the following two cases
      - Subjects are never diagnosed as AD in follow-ups
      - Subjects are diagnosed as AD in one of the follow-ups
    - Prediction over a couple of months to a couple of years
  - Only around 200 samples
- Model
  - Same model, fine-tuned from the previous study
  - Made sure the previous model does not capture subject information

Performances

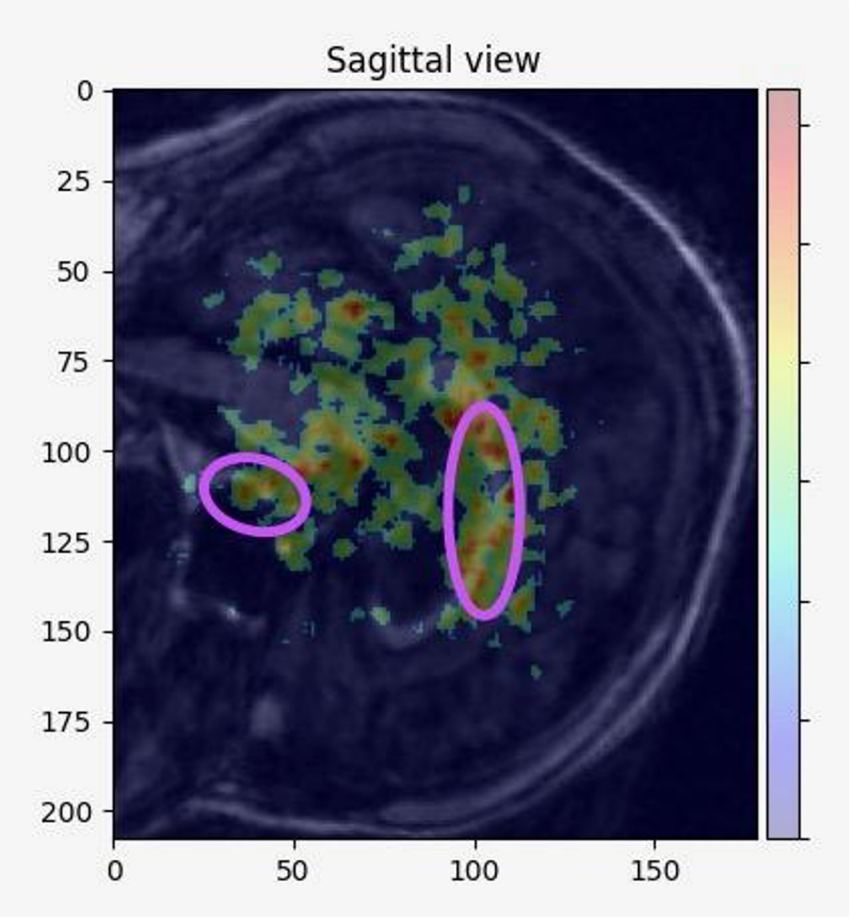
- We obtained 0.95 test accuracies
- With interpretable focuses on certain brain areas
  - Which might help us understand the pathology of AD for early-diagnosis

Toward an Understanding of the Pathology



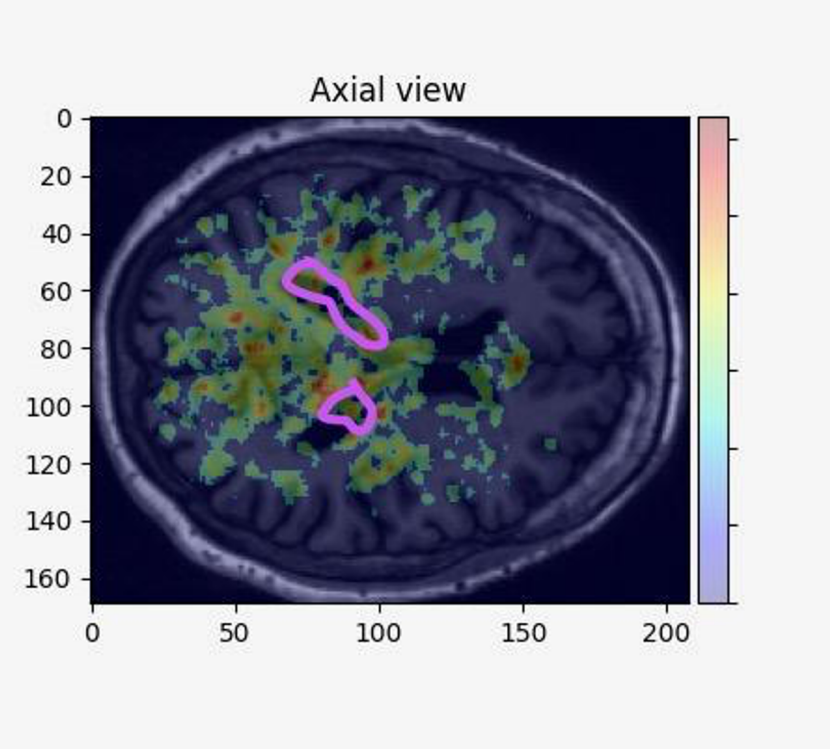
AD: Trigonium Collaterale and Frontal Horn Correlation .44 correlation

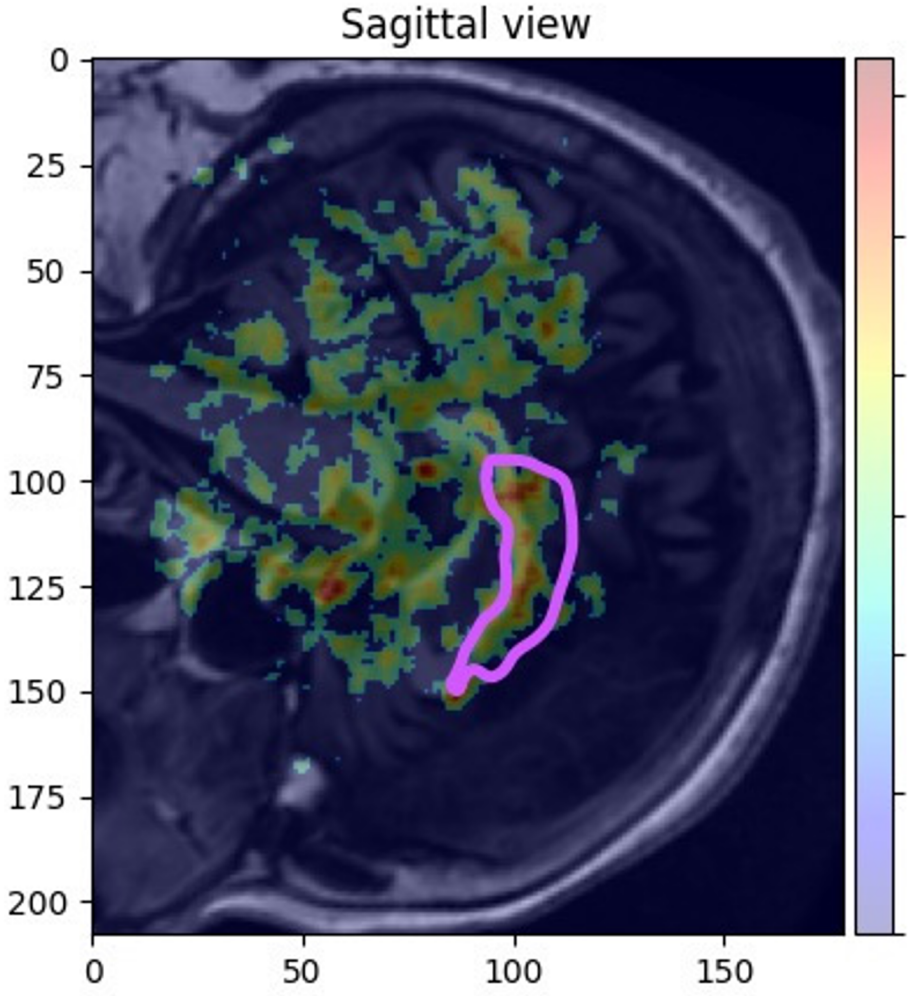
Toward an Understanding of the Pathology



AD: Sella/Hypothalamus and Lateral Ventricle
.3 Correlation

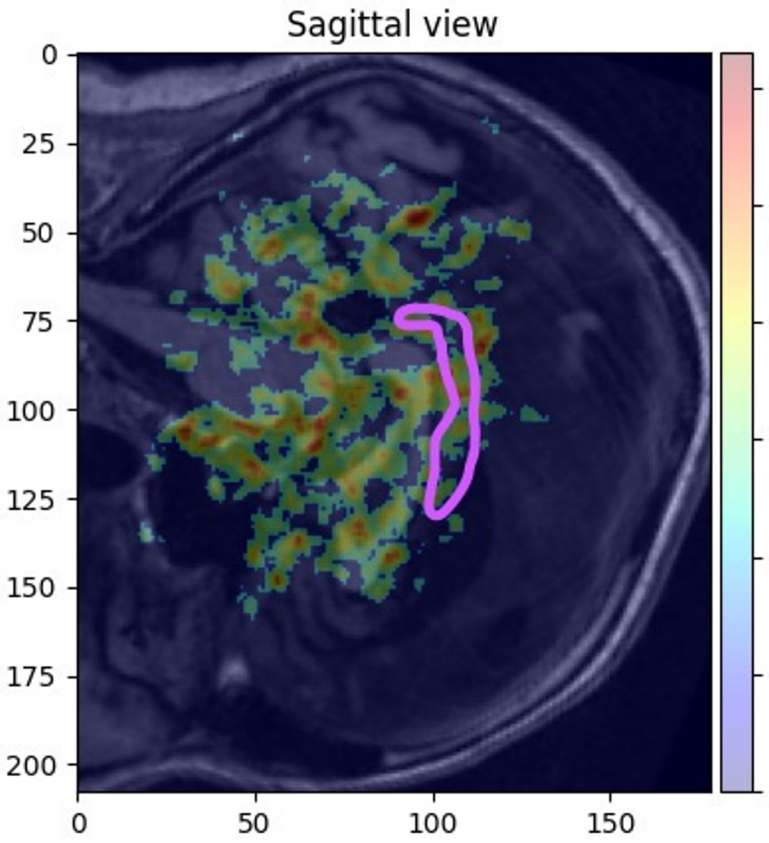
Toward an Understanding of the Pathology

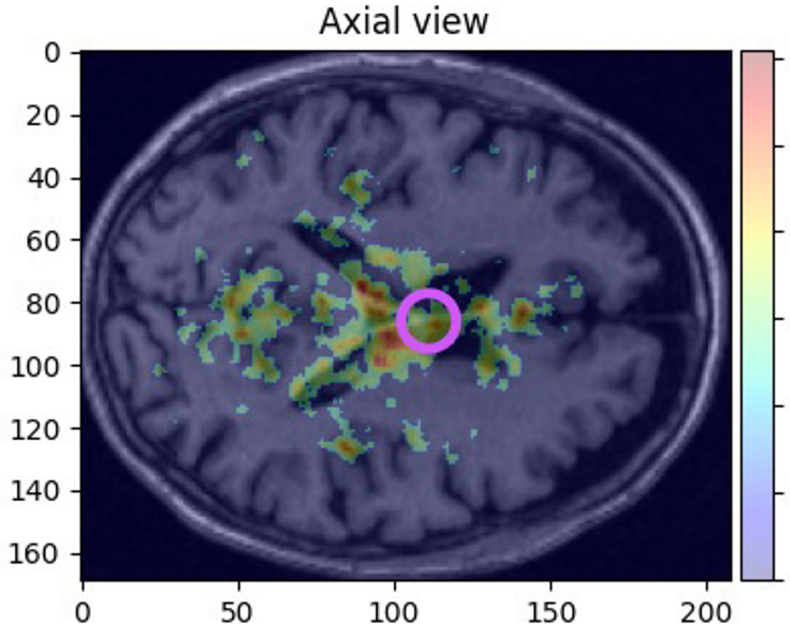




MCI: Lateral ventricle (sagittal view) and Trigonium Collaterale (axial view) . 33 correlation

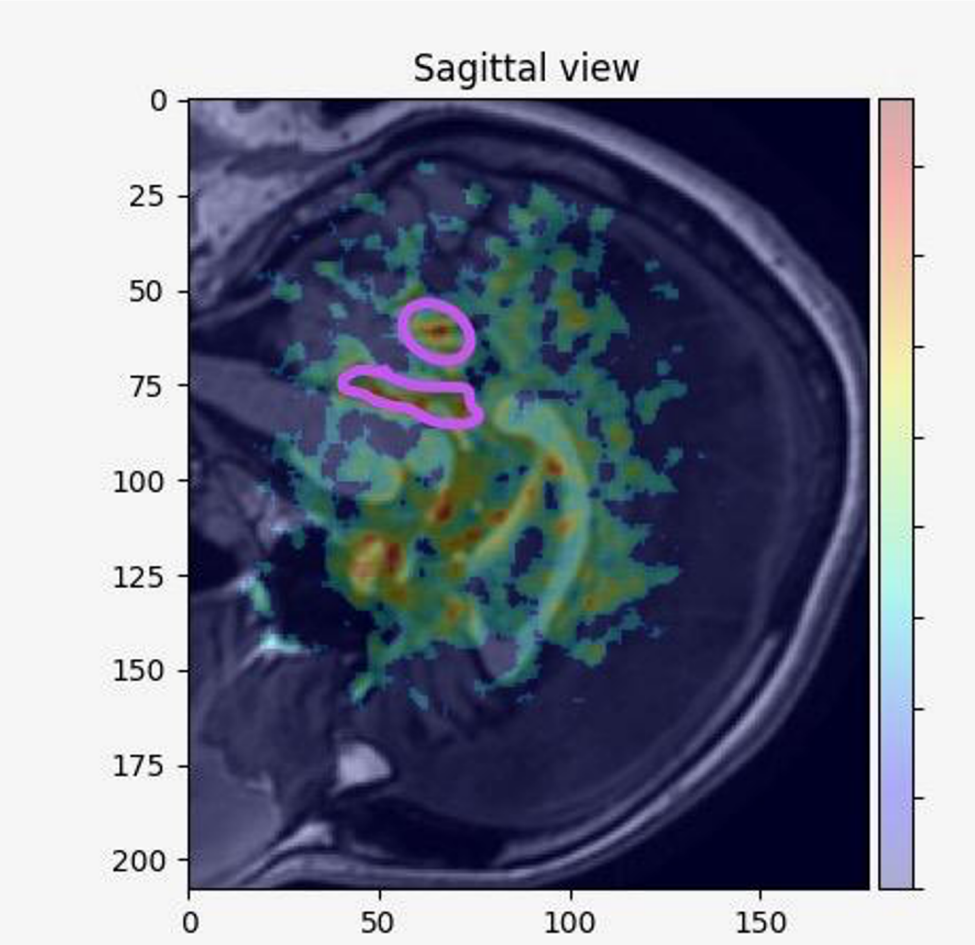
Toward an Understanding of the Pathology





MCI: Corpus Collosum (sagittal view) and Septum Pellucidum area (axial view) .38 correlation

Toward an Understanding of the Pathology



Cerebellum and 4th ventricle
area .32 and .36 correlations

The understanding of function relevance in the cerebellum is still in the early stage (Jacobs et al. 2018)

Thank you

- Contributions:
  - The challenges of AI diagnosis in robustness / confounding factors
    - Central hypothesis of robust machine learning
  - Principled understanding trustworthy machine learning
    - Data augmentation and alignment regularization
  - (early-stage) Alzheimer's disease diagnosis
- Acknowledgement

  - Daniel Huang (Language Technology Institute, Carnegie Mellon University)
  - Stephen Tsou (Biomedical Engineering, Carnegie Mellon University)
  - Thomas Pearce (Department of Pathology, University of Pittsburgh Medical Center)
  - Oscar L. Lopez (Alzheimer's Disease Research Center, University of Pittsburgh Medical Center)
  - Wei Wu (Computational Biology Department, Carnegie Mellon University)
  - Eric P. Xing (Machine Learning Department, Carnegie Mellon University)