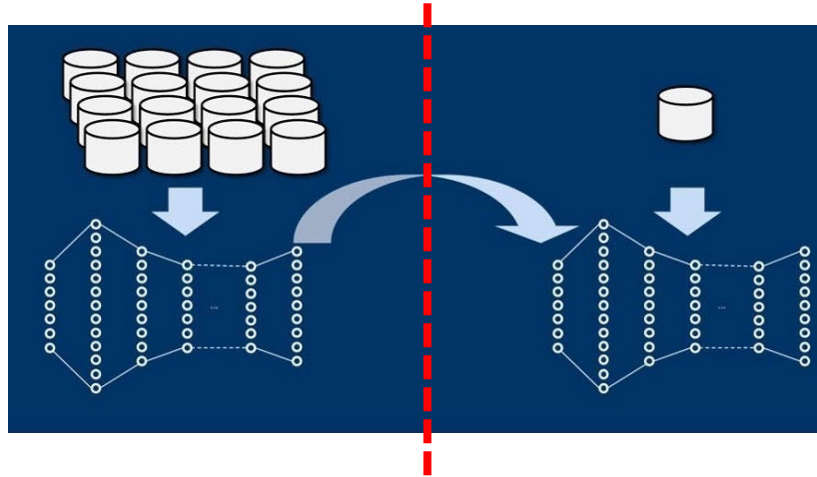# Adversarial-Robust Transfer Learning for Medical Imaging via Domain Assimilation

Xiaohui Chen, Tie Luo*

Department of Computer Science
Missouri University of Science and Technology, USA

The Ubiquitous Transfer Learning

- Transfer learning has been extremely successful and almost ubiquitous in DL
  - ❑ It takes a base model <u>pretrained</u> on large datasets and then <u>fine-tunes</u> it on a small dataset pertaining to a specific (downstream) task

- Medical AI (esp. medical imaging) is <u>no exception</u>
  - ❑ Due to the lack of reliably annotated public large medical datasets
    - ✓ *"So yeah, it makes sense to use TL."*
  - ❑ Plus, "Everybody is using it!" and "It works well!"   Hence, …

Xiaohui Chen and Tie Luo, "*Adversarial-Robust Transfer Learning for Medical Imaging via Domain Assimilation.*"

The Fall of Panacea

(In the following, we focus on CV, but the same principle should apply similarly to other domains such as NLP and audio.)

A crucial factor is overlooked

- *"Not all images are created equal"*

- Pretrained models are all trained on natural images (e.g. ImageNet), but medical images have some <u>distinct</u> properties
  - ❑ which lead to higher vulnerability of medical AI models to adversaries

A little background: Adversarial Attacks in medical domain

In this paper, we focus on Adversarial Examples, which is a popular type of attacks and can be formulated as

$$\underset{\|\delta\|_p \le \epsilon}{\operatorname{argmax}} \, l(\theta, x + \delta, y)$$

Original Input + carefully crafted noise

Model

Wrong prediction

where an adversary aims to find a *small perturbation δ* (constrained by ϵ-ball around original input x) to maximize the loss, such that the model would makea wrong prediction for the new input $x'=x+\delta$

$\text{Loss}(x + \delta, y; \theta)$

$\delta = 0$

Δ

3
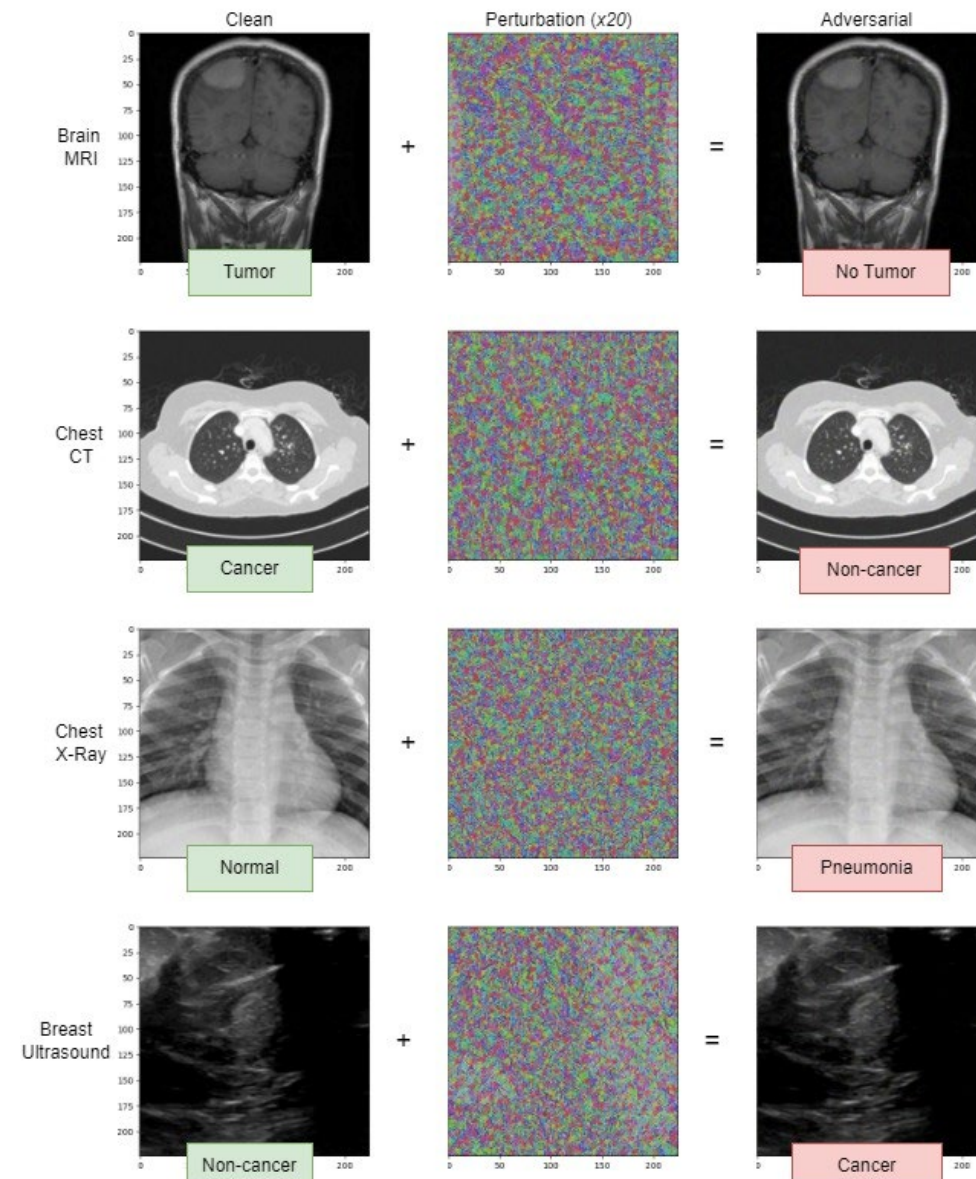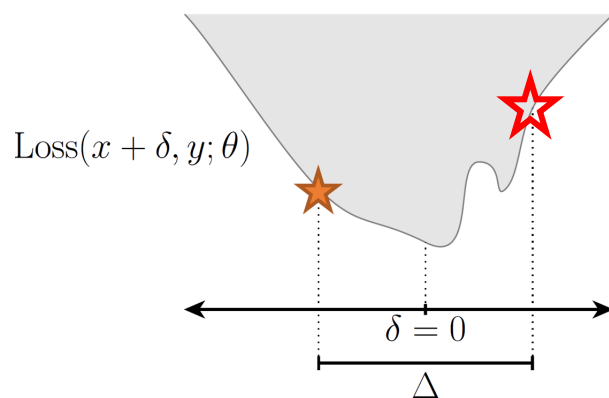
Why are medical AI models more vulnerable?

- **Large attention region:**
  - ❑ Unlike natural images, medical images typically have monotonic biological texture, which tends to mislead DNNs to pay attention to areas irrelevant to diagnosis
  - ❑ In these attention regions, it is easier for small perturbations to cause significant changes in output
- **Sharp loss landscape:**
 - ❑ Medical images have rather simple representations (only features related to lesion) as compared to natural images
  - ❑ DNNs are typically overparameterized (e.g., ResNet50); using them to learn simple patterns from a large attention region tend to create sharp loss landscape
  - ❑ On a sharp loss landscape, a small perturbation can cause drastic changes
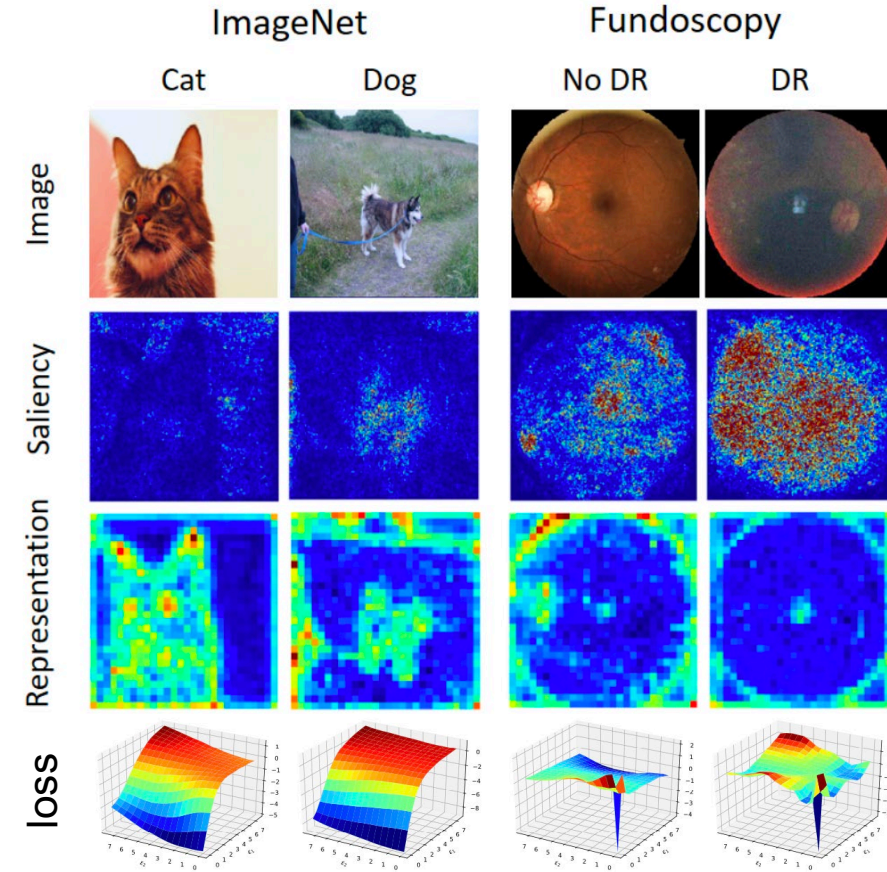
Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F.: Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition* (2021).

Bridging The Gap

- "Domain Discrepancy": substantial gap between natural & medical images
- We propose a Domain Assimilation strategy to bridge this gap:-
  - ❏ **Colorize and adapt texture** of medical images to resemble nature images
  - ❏ Retain essential texture (over-adaptation can lead to misdiagnoses!)
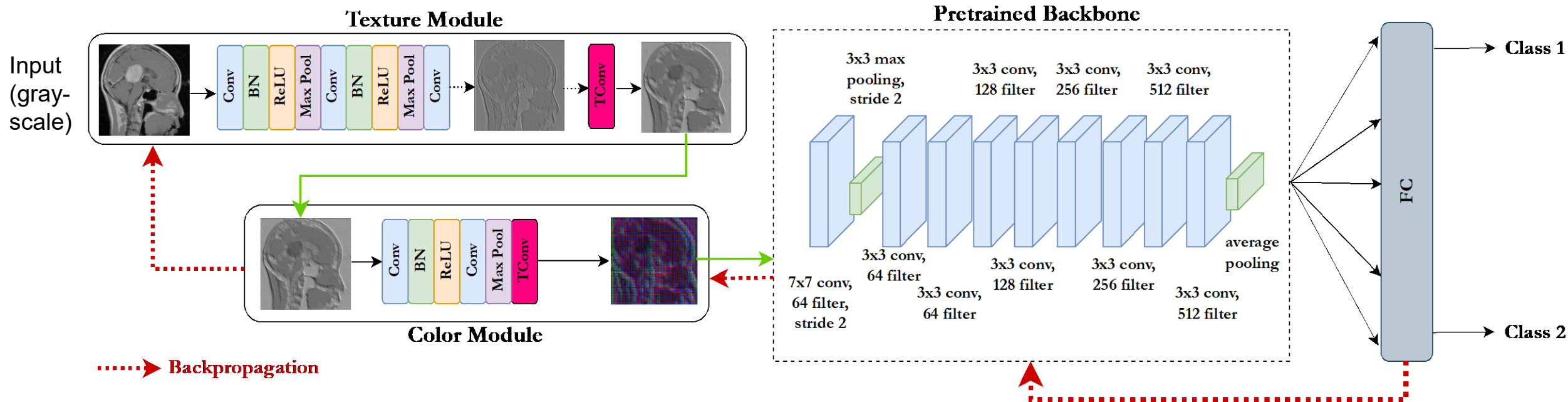
- **Texture and color adaptation:**

$$< \theta_T, \theta_C, \theta_B, \theta_F > = \operatorname*{argmin}_{\theta_T, \theta_C, \theta_B, \theta_F} L\Big(F\Big(B\Big(C(T(X, \theta_T), \theta_C), \theta_B\Big), \theta_F\Big), y\Big)$$

  - ❑ T: texture module, C: color module, B: pretrained backbone, F: final classifier.

- **Retain essential texture:** restrict distortion using GLCM loss

  - ❑ Gray Level Co-occurrence Matrix (GLCM) is a texture descriptor quantifying texture features and describing local spatial relationships at intensity levels.

$$< \theta_T, \theta_C, \theta_B, \theta_F > = \operatorname*{argmin}_{\theta_T, \theta_C, \theta_B, \theta_F} \Big(\alpha \times \text{CrossEntropyLoss} \Big(F(B(C(T(X, \theta_T), \theta_C), \theta_B), \theta_F), y\Big) +$$

$$(1 - \alpha) \times \text{GLCMLoss}\Big(C(X, \theta_C), X\Big) \Big) \qquad (18)$$

Measure distortion using second-order texture features:

$$\text{GLCMLoss} = \max_{1 \leqslant i \leqslant m} \sum_{j=1}^{n} \Big| SOT\Big(grayscale\big(C(X) \big) \Big)_{i,j} - SOT(X)_{i,j} \Big|$$

SOT: second-order texture feature matrix

GLCM

- Given an image I, let i and j represent the grayscale values, (x, y) the spatial locations of pixels, and $(\nabla x, \nabla y)$ the offset determined by a predefined distance and orientation (angle)

$$P_{\nabla x, \nabla y}(i,j) = \sum_{x=1}^{w} \sum_{y=1}^{h} \begin{cases} 1, & if\ I(x,y) = i\ and\ I(x + \nabla x, y + \nabla y) = j \\ 0, & otherwise \end{cases}$$

Think of GLCM as a frequency matrix

- Each combination of distance and orientation will generate a GLCM (8 in our case)

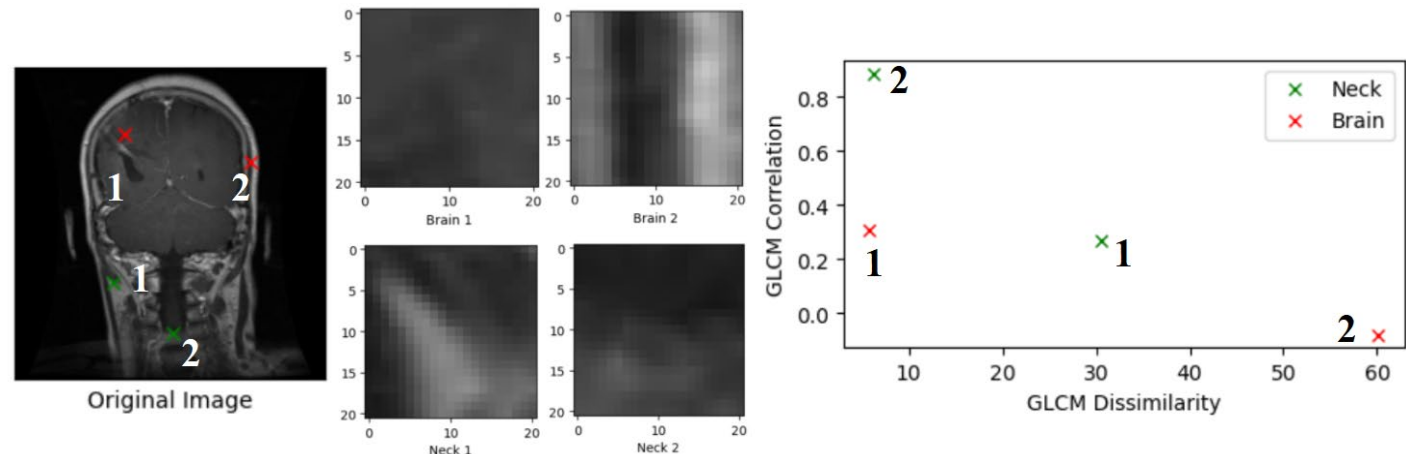- For each GLCM, we further extract second-order texture features for each small area (20x20):

  ❑ $Angular\ Second\ Moment (ASM) = \sum_{i,j} P(i,j)^2$

  ❑ $Contrast = \sum_{i,j} |i - j|^2 P(i,j)$

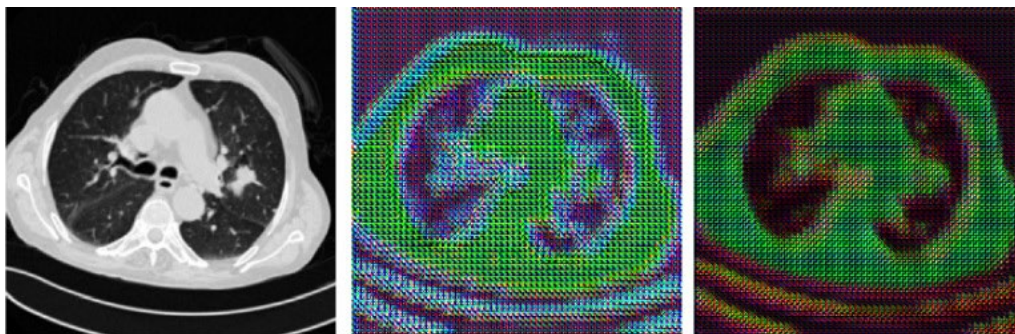 ❑ $Homogeneity = \sum_{i,j} \frac{P(i,j)}{1 + |i-j|^2}$

  ❑ $Correlation = \sum_{i,j} \frac{(i - \mu_i)(j - \mu_j) P(i,j)}{\sqrt{\sigma_i^2 \sigma_j^2}}$

  ❑ $Disimilarity = \sum_{i,j} P(i,j)|i - j|$
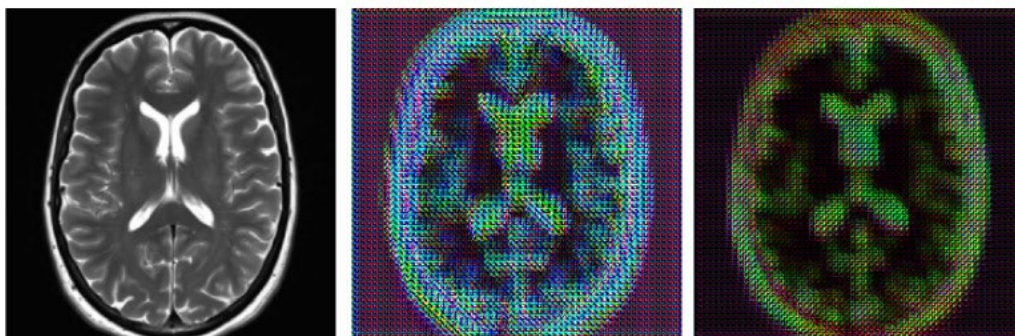
Xiaohui Chen and Tie Luo, "*Adversarial-Robust Transfer Learning for Medical Imaging via Domain Assimilation.*"

Intermediate Result:
Post Texture and Color adaptation

Chest CT

Brain MRI

Original   w/o GLCM loss   w/ GLCM loss

Experiments

- Datasets:

	Classes	Class Size
Brain MRI	no-tumor, tumor(glioma/meningioma/pituitary)	1595, 1595
Chest Xray	normal, pneumonia	1583, 1583
Chest CT	no-cancer(normal/benign), cancer	536, 536
Breast UltraSnd	no-cancer(normal/benign), cancer	210, 210

- Models and training:

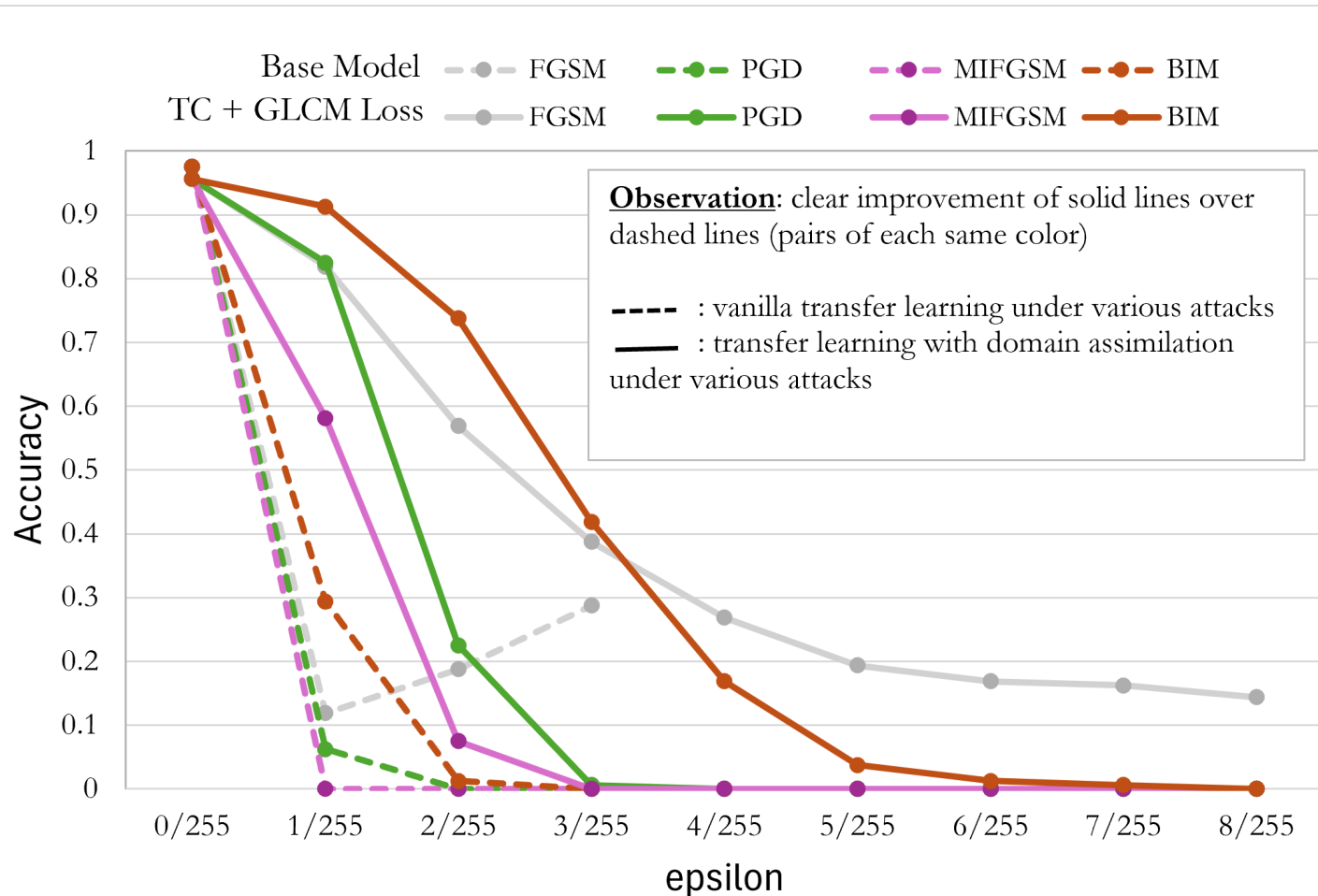| Pretrained Model Selection		Parameters
ResNet18, ResNet50, DenseNet121	Epochs	300
	Learning Rate	1e-4
	Batch Size	32
	EarlyStopping	30
	Input Size	(224,224,1)

- GLCM parameters: Distance = 3, Orientation = 0/45/90/135/180/225/270/315 (degrees)
- **Attacks: FGSM, BIM, MIFGSM, PGD**
  - ❑ Perturbation size $\epsilon \in \{\frac{1}{255}, \frac{2}{255}, \frac{3}{255}, \frac{4}{255}, \frac{5}{255}, \frac{6}{255}, \frac{7}{255}, \frac{8}{255}\}$

Xiaohui Chen and Tie Luo, "*Adversarial-Robust Transfer Learning for Medical Imaging via Domain Assimilation.*"

Results

- Model accuracy vs. attack strength (Chest CT)

**Clear winning margin of solid over dashed lines** (same-color pairs)

Dashed lines: vanilla transfer learning under various attacks

Solid lines: transfer learning with domain assimilation under various attacks

Vanilla FGSM (dashed gray line) is an <u>outlier</u>, because it moves gradient only <u>one step</u>, so increasing epsilon does not help but may result in it moving a <u>big step</u> toward a <u>worse</u> place (while all the other attacks do <u>iterative</u> gradient ascent). So for FGSM, one should focus on <u>small espilon values</u>.

Xiaohui Chen and Tie Luo, "*Adversarial-Robust Transfer Learning for Medical Imaging via Domain Assimilation.*"

Take-Home Message

- **Transfer Learning should be used with caution** on "Domain Discrepancy"
  - ❑ More vulnerable to adversarial attacks
- Domain Assimilation as a solution:
  - ❑ **Colorization and texture adaptation** (lightweight)
  - ❑ **Retain essential texture** (GLCM to avoid over-adaptation)
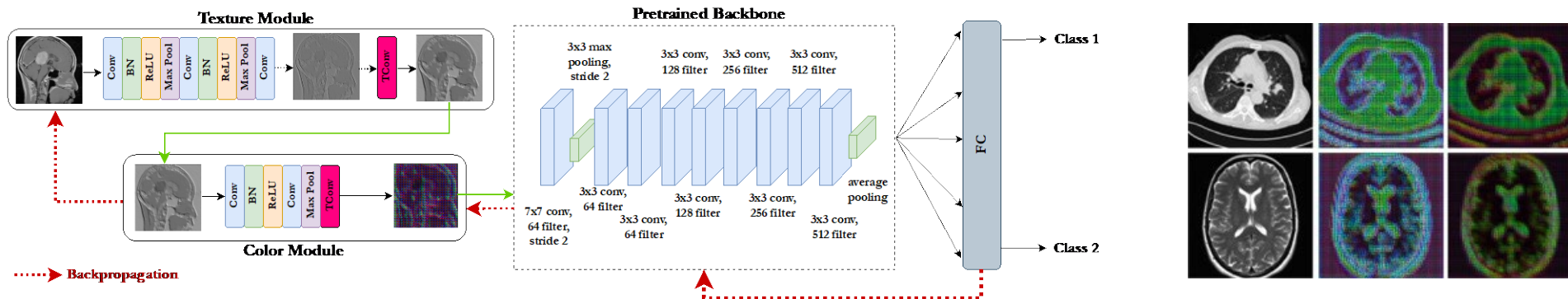
- **Future work:** more in-depth texture analysis and enhancement

Thank you!