July 24, 2025

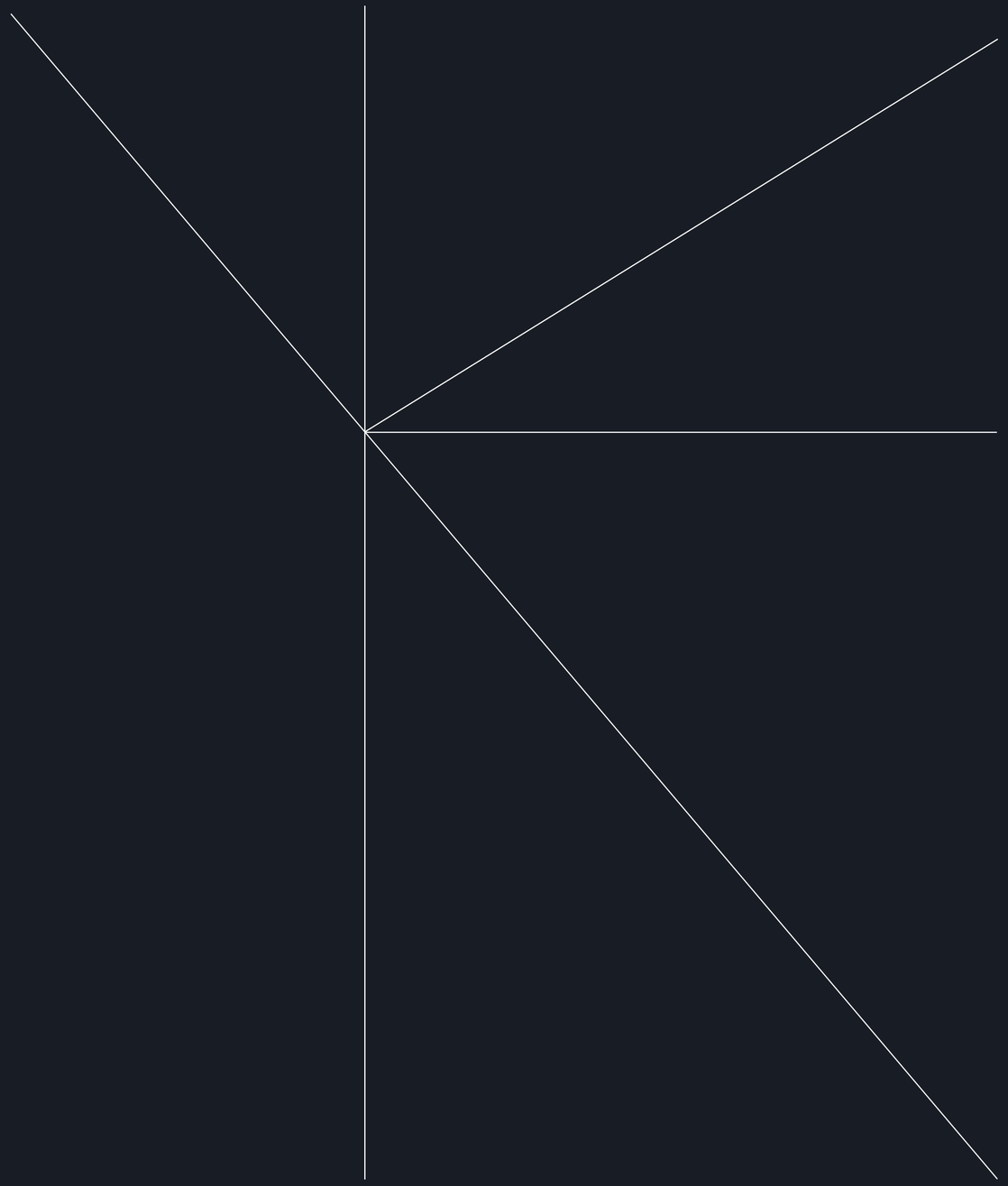
Re-identification risk of medical imaging-based deep learning models

Sadie Lee

THE UNIVERSITY
OF BRITISH COLUMBIA

Contents

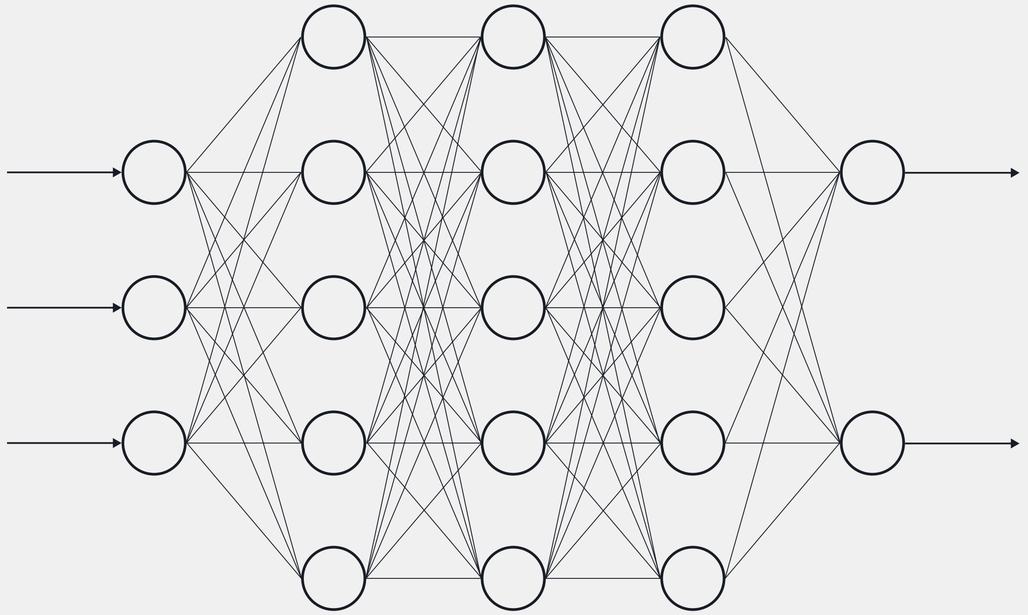
1. Introduction



Motivation

SCENARIO

With access to model parameters alone for a neural network trained on medical images, e.g. an exported checkpoint, is there a risk that patients can be re-identified?



Research questions

RESEARCH QUESTION 1

What patient re-identification risks are present in training AI models on radiology image data?

RESEARCH QUESTION 2

What is the magnitude of these risks?

RESEARCH QUESTION 3

What mitigations can be taken to reduce these risks?

Re-identification

The extent to which an image or its features can be traced back to a real patient, following de-identification [1-3].

Assumptions

01

We know the imaging modality and the anatomical region of the target model's training data.

02

We only have access to the target model's parameters through a frozen `state_dict` checkpoint.

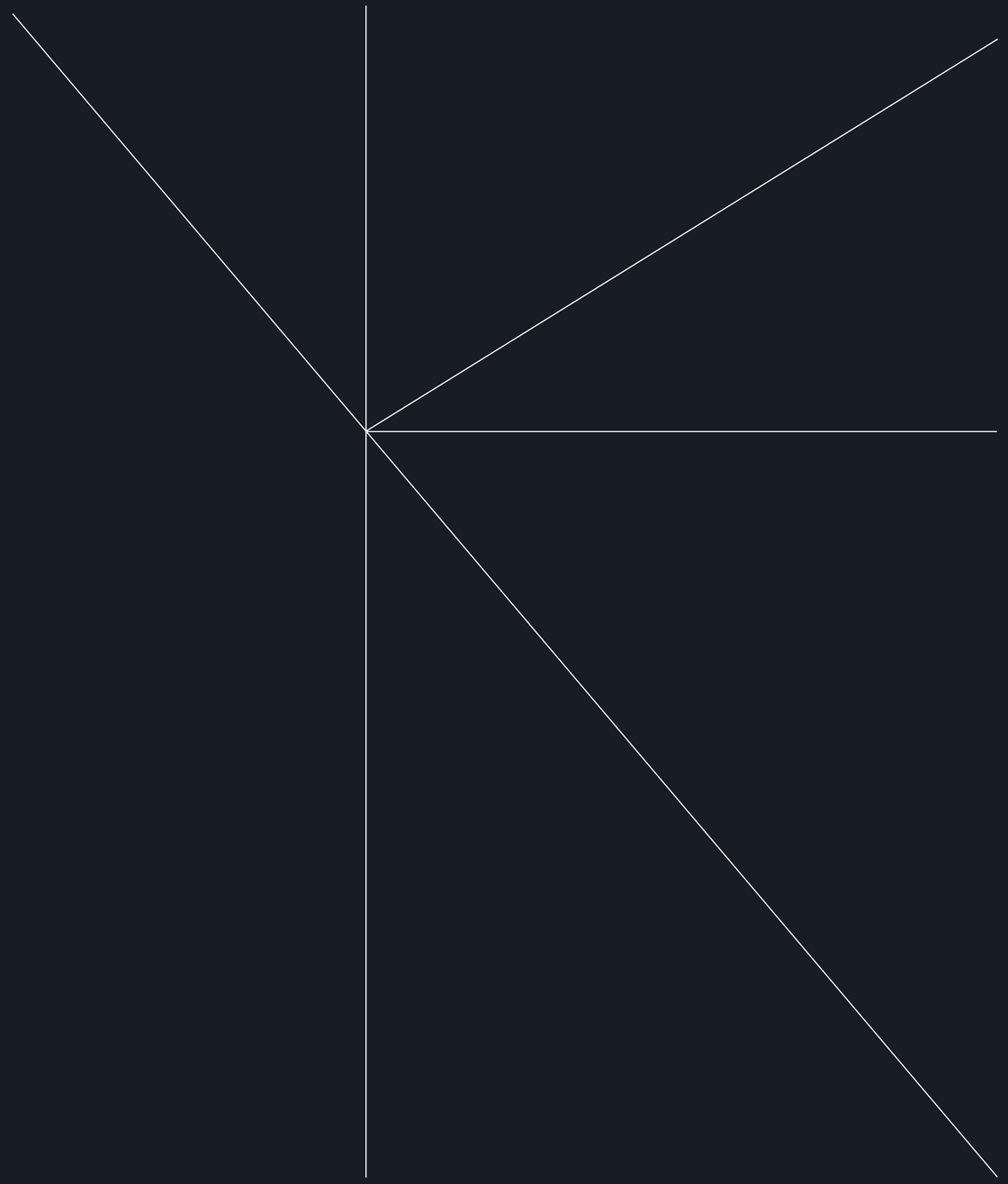
03

We can infer the target model's architecture from the checkpoint by inspecting its layers.

04

We do not have any of the target model's training images in practice.

2. Theory and related work



Memory and memorization

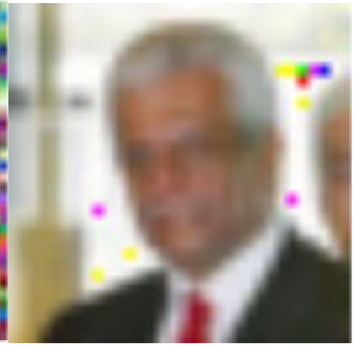
Image models	Memorizing specific features from the training data → similarity between original and reconstruction [4].
Memorization & privacy	Co-occurence with inadvertent privacy leakage and training data reconstruction [5]. Overfitting is one marker of memorization [6].
Architectural differences	ViTs have been shown to memorize more and be more vulnerable to reconstruction and privacy leakage [5].

Image reconstruction (inversion) attacks

Gradient-based inversion [7]

Initialization

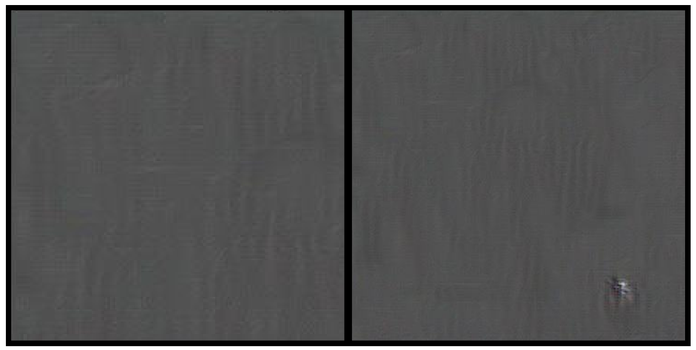
Optimized



Pixel-based inversion [8]

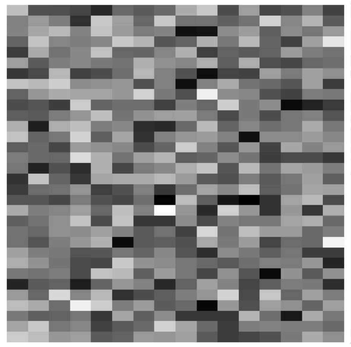
Initialization

Optimized



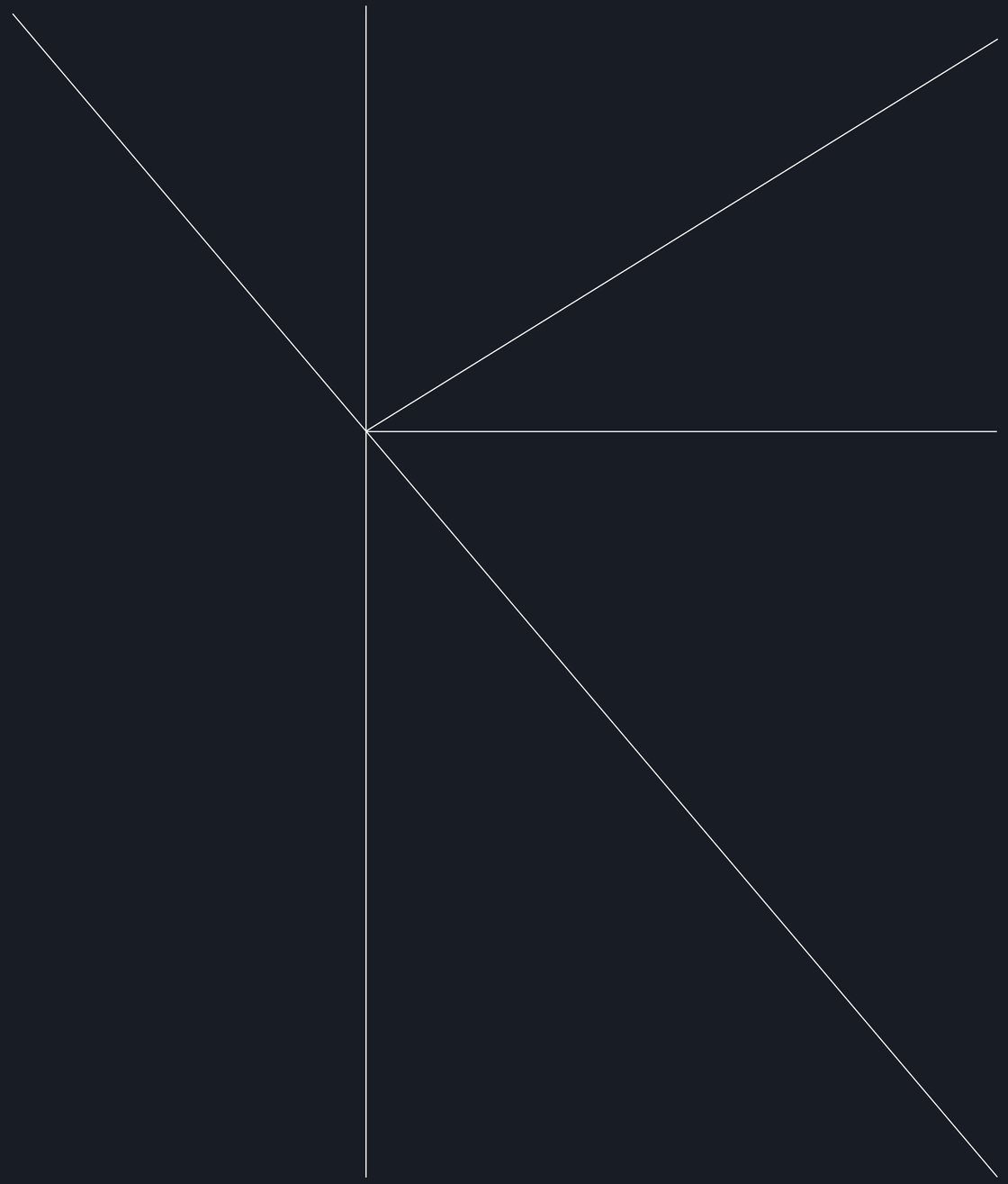
Latent-based inversion

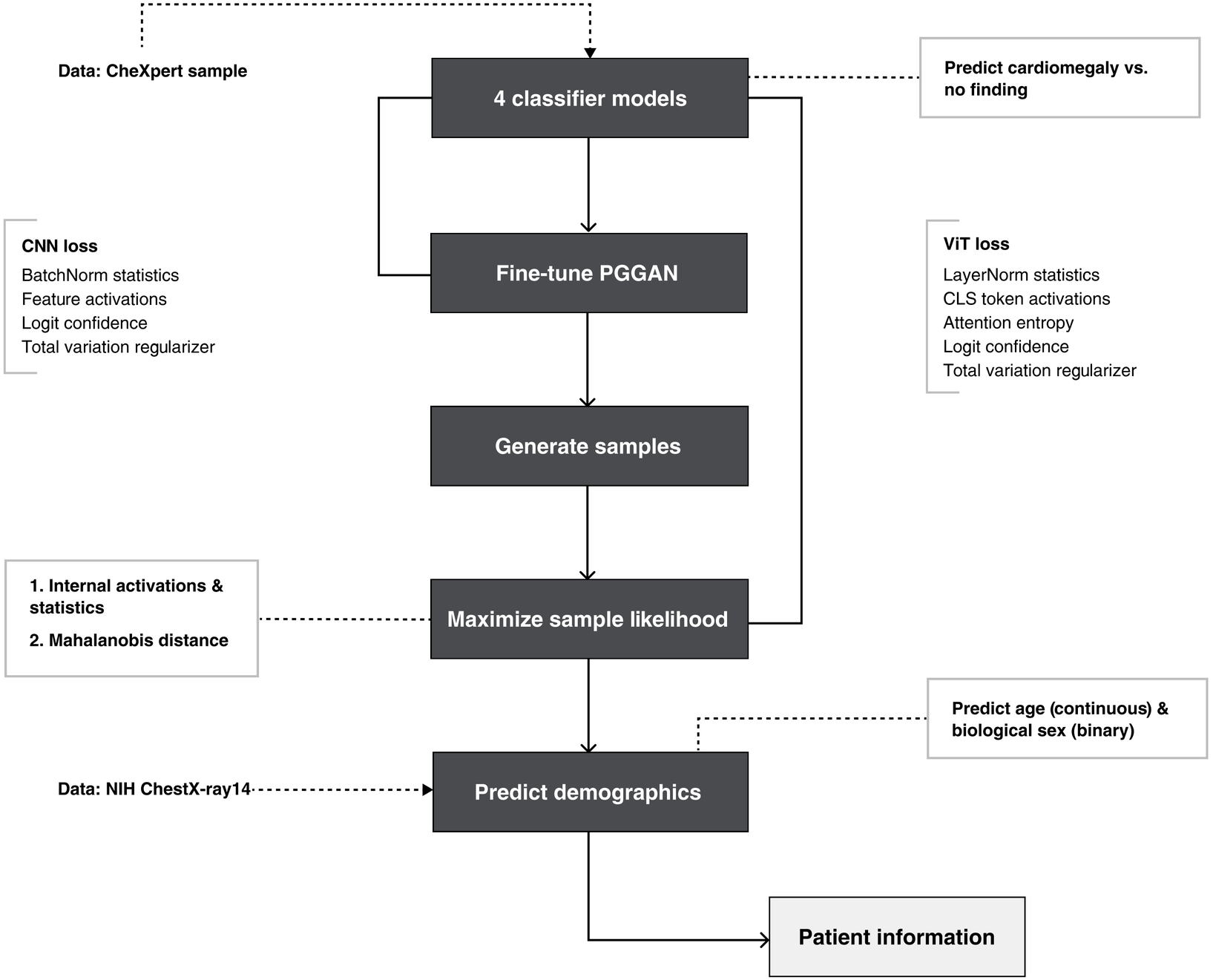
Initialization



Optimized

3. Methods overview




Data: CheXpert sample
4 classifier models
Predict cardiomegaly vs. no finding
CNN loss
BatchNorm statistics
Feature activations
Logit confidence
Total variation regularizer
Fine-tune PGGAN
ViT loss
LayerNorm statistics
CLS token activations
Attention entropy
Logit confidence
Total variation regularizer
Generate samples
1. Internal activations & statistics
2. Mahalanobis distance
Maximize sample likelihood
Predict age (continuous) & biological sex (binary)
Data: NIH ChestX-ray14
Predict demographics
Patient information

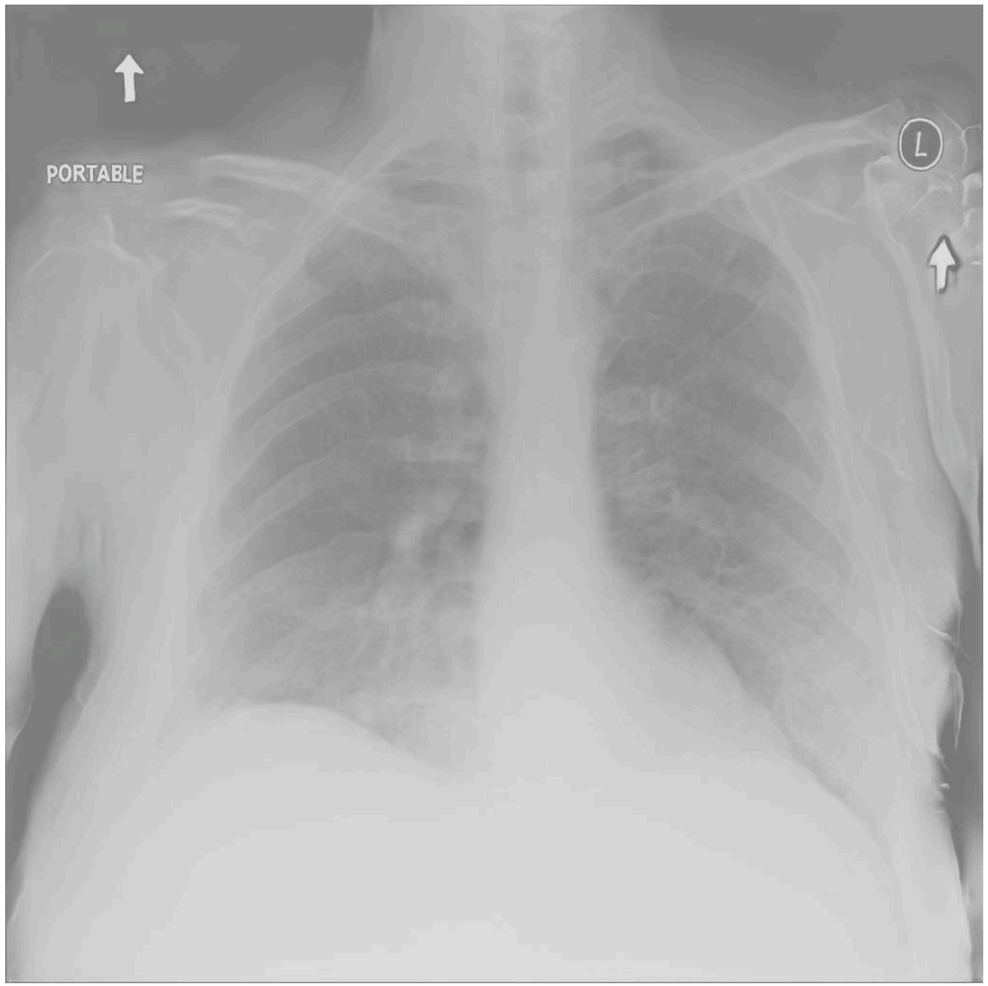
4. Two-stage reconstruction

Stage 1: Approximate target data manifold

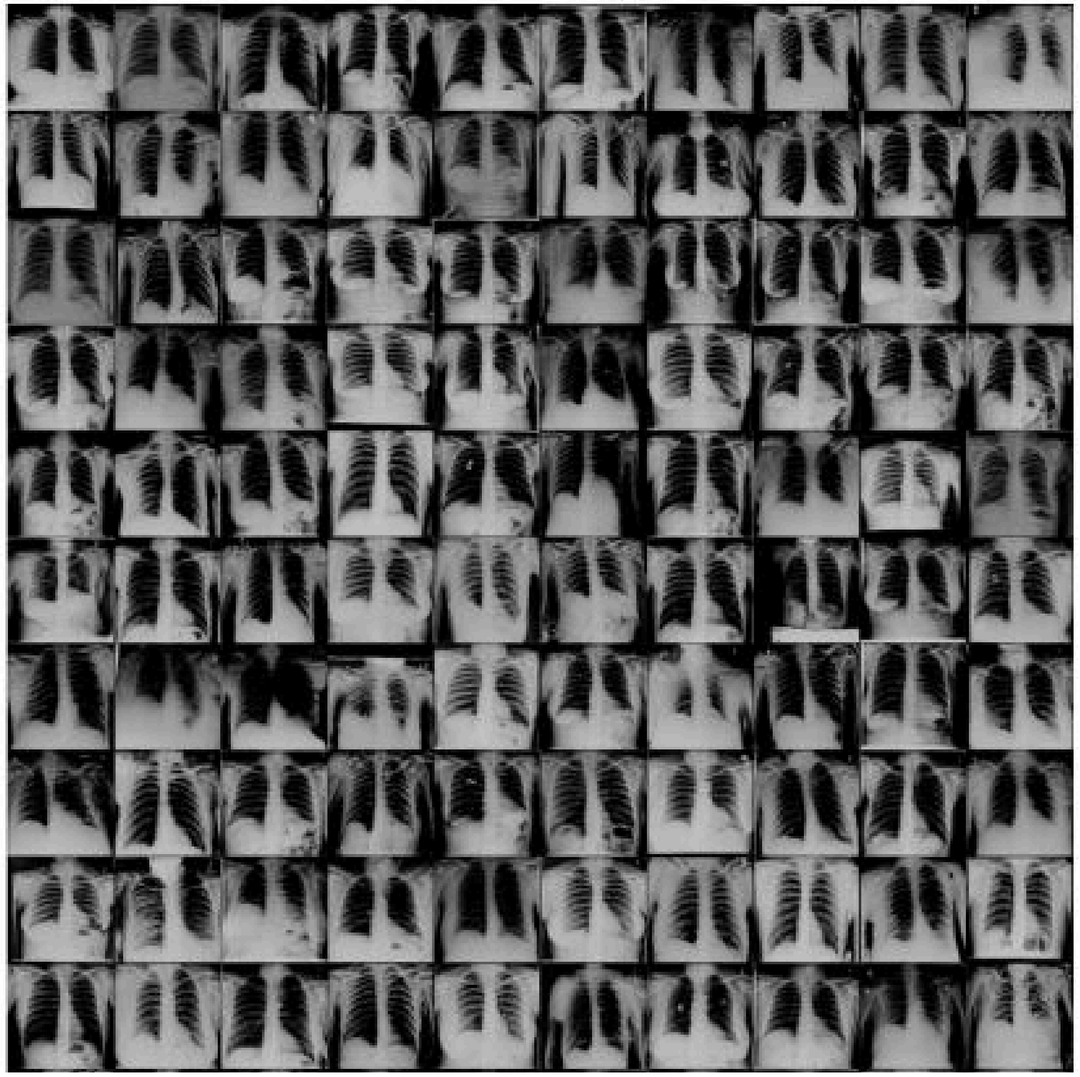
Assume a target model's parameters, e.g. BatchNorm statistics, are **compressed representations** of the training data [8].

Fine-tune a pre-trained PGGAN's generator to approximate the training data manifold by matching the target model's parameters.

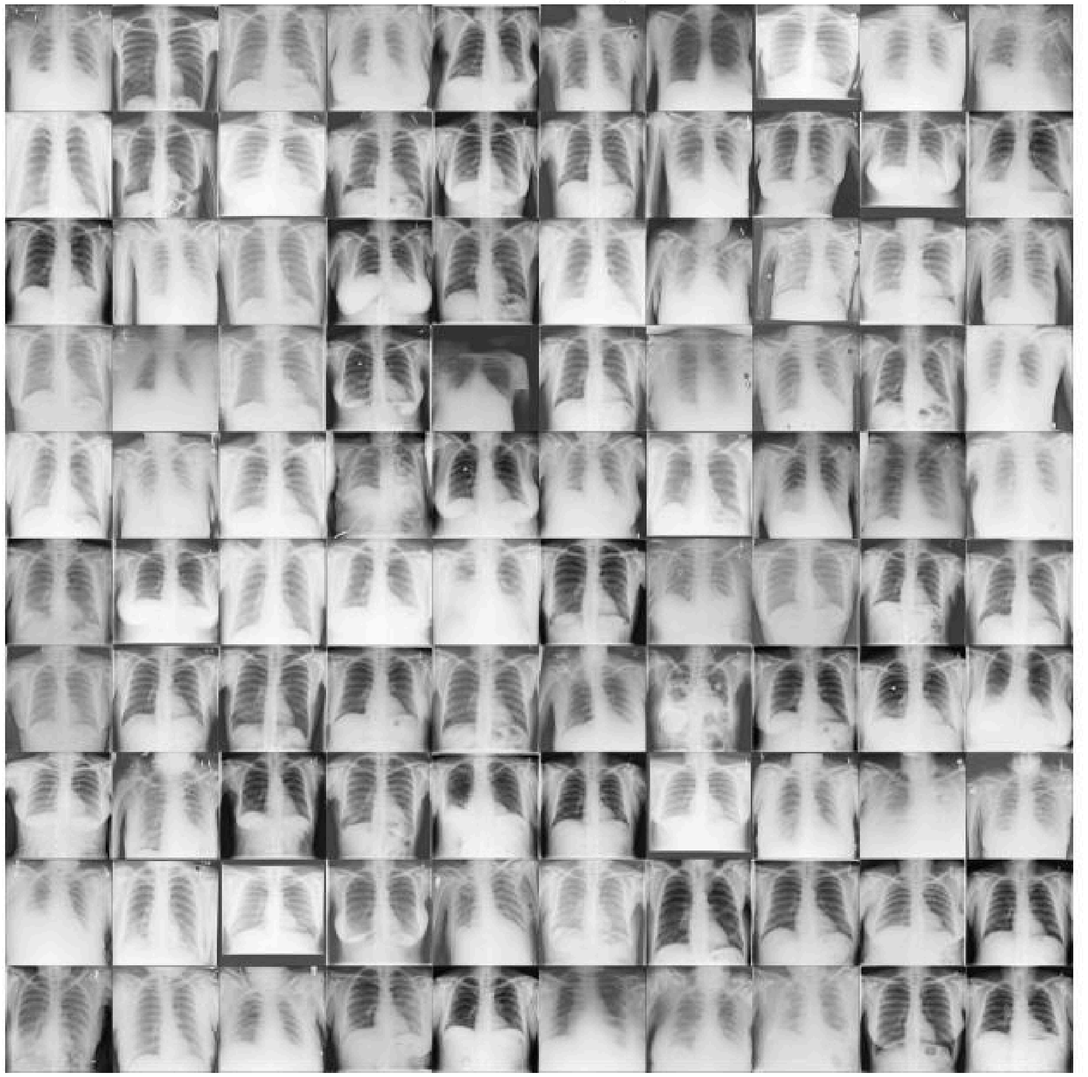
Instead of optimizing on the images directly, we optimize the generator itself.



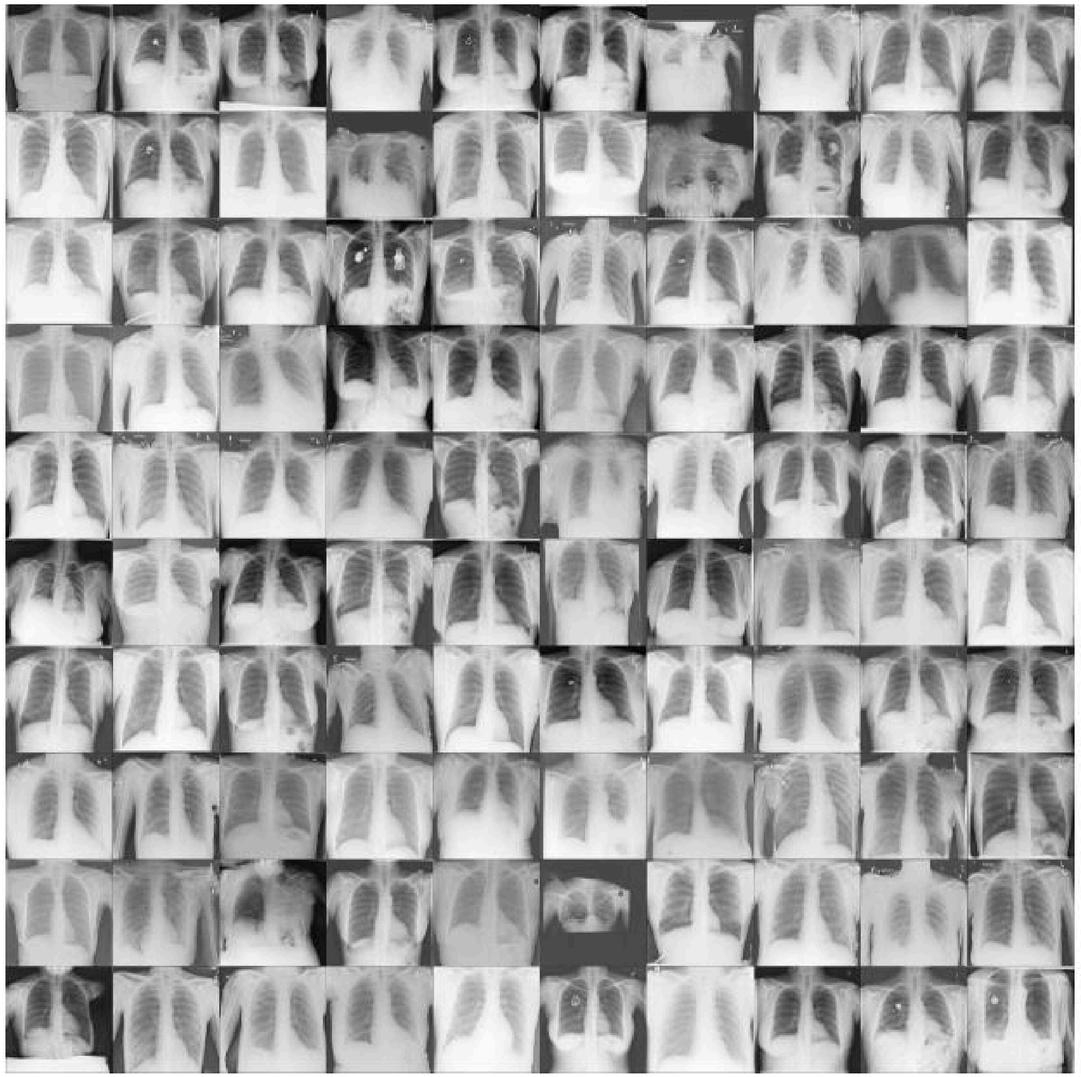
CNN Generated Samples With Classes



Overfit CNN Generated Samples

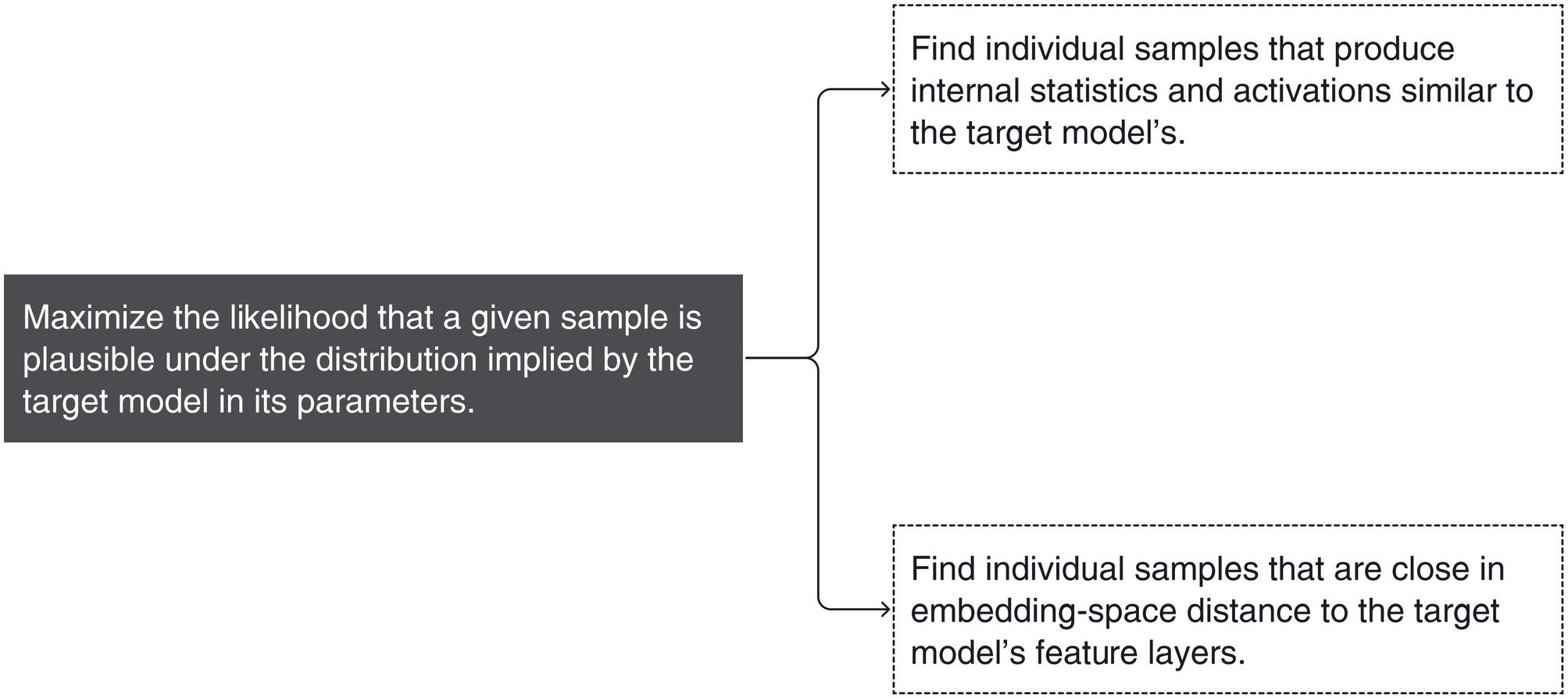


ViT Generated Samples

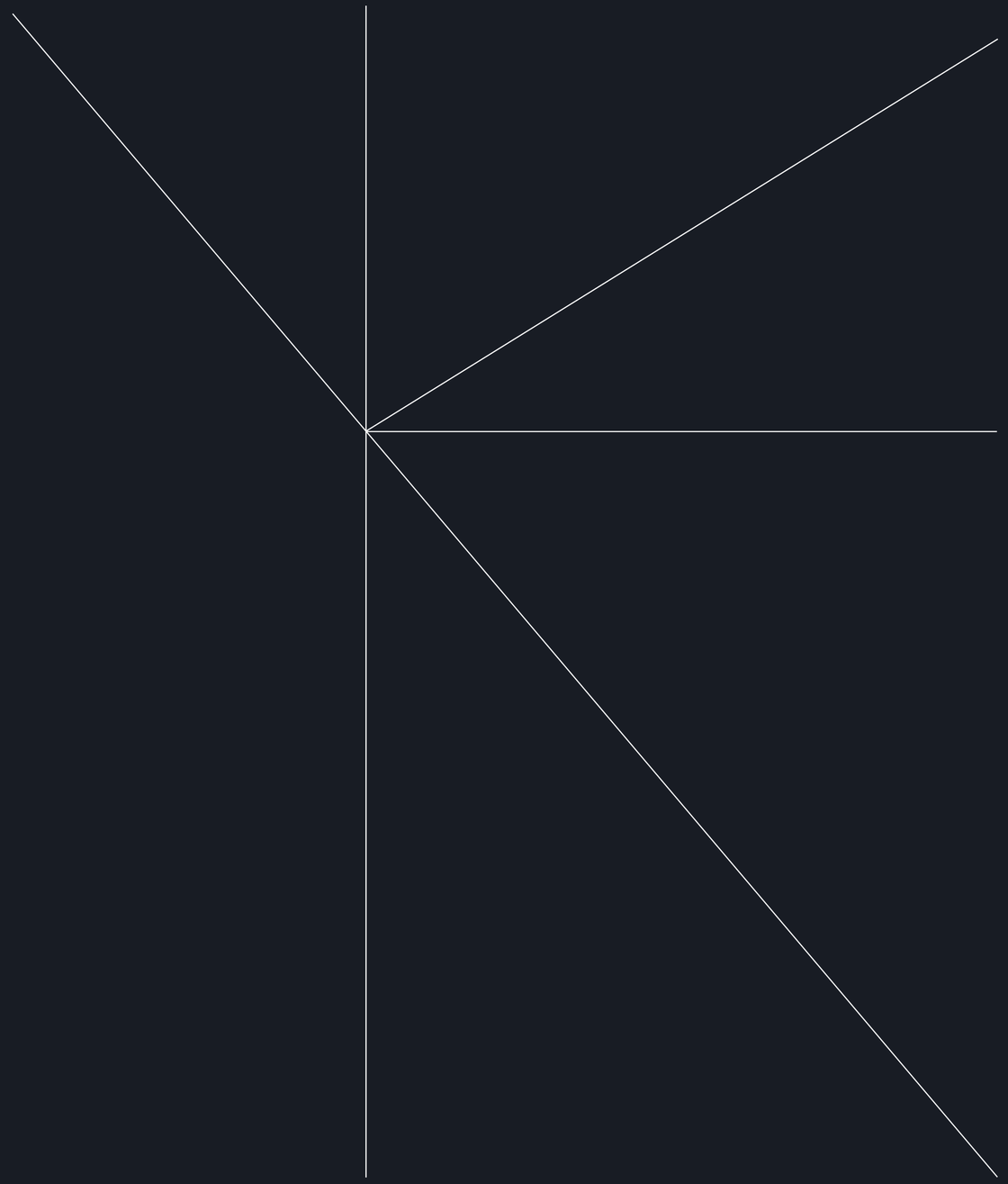


Overfit ViT Generated Samples

Stage 2: Maximize sample likelihood

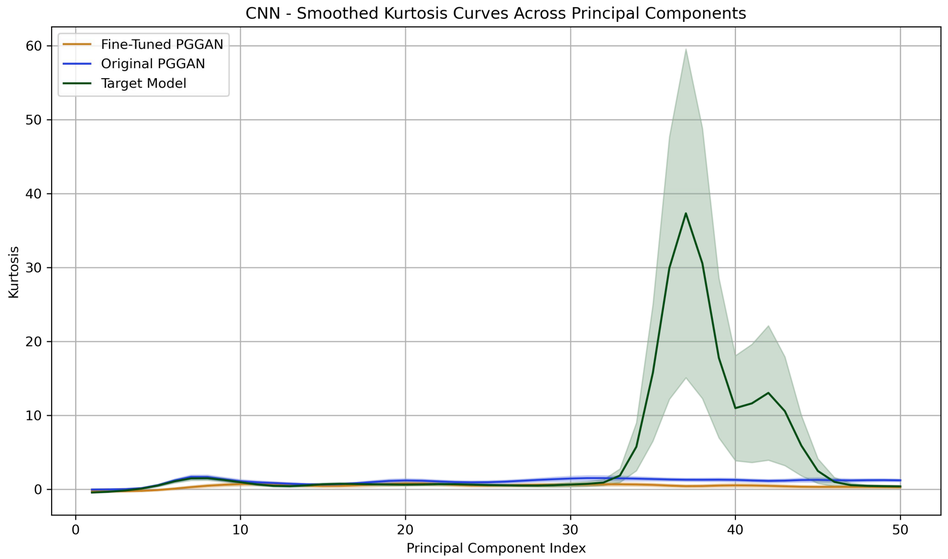


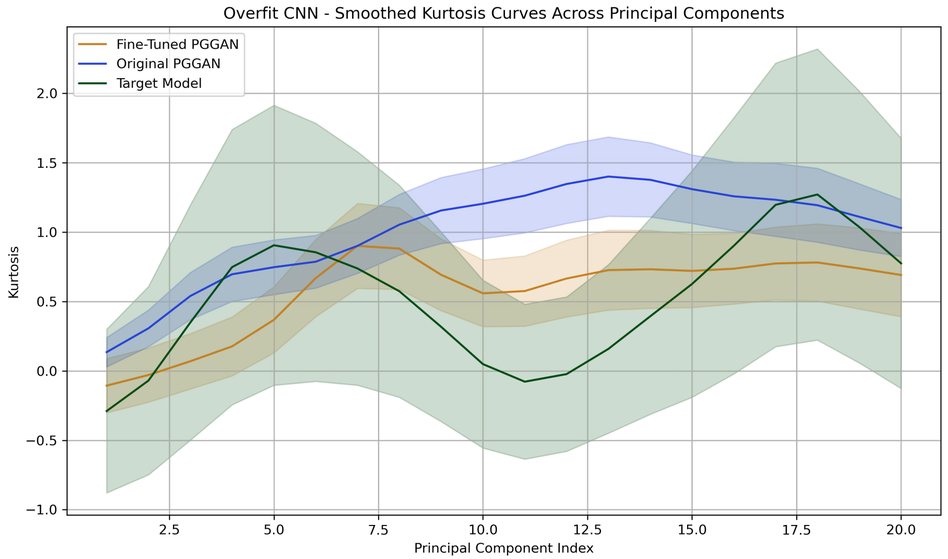
5. Metrics



Distribution shifts - CNN models

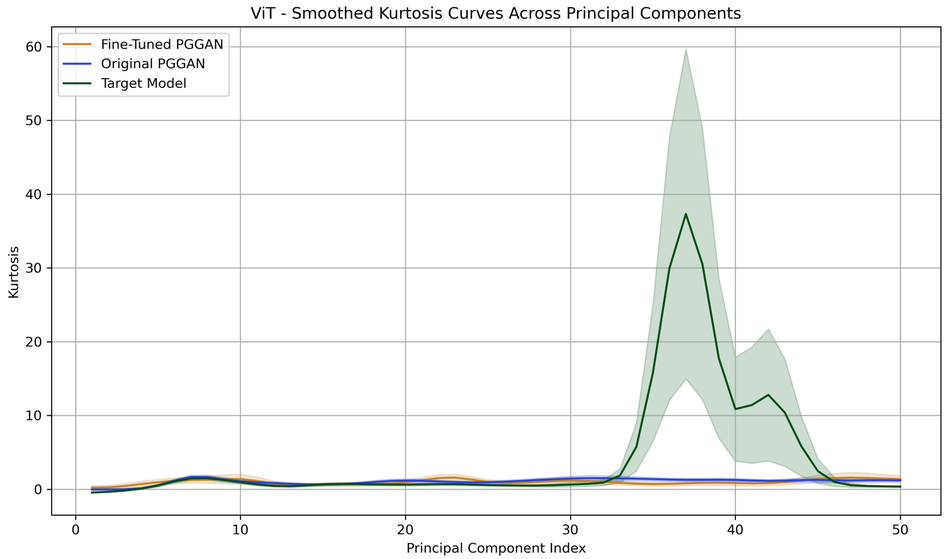
	CNN		Overfit CNN		ViT		Overfit ViT	
	Original → Fine-tuned	Target CNN	Original → Fine-tuned	Target Overfit CNN	Original → Fine-tuned	Target ViT	Original → Fine-tuned	Target Overfit ViT
Pixel entropy	7.402 → 6.486	7.963	7.402 → 6.176	7.965	7.402 → 6.176	7.963	7.402 → 6.125	7.962
PCA kurtosis	1.065 → 0.436	4.306	1.065 → 0.566	0.532	1.065 → 1.001	4.299	1.065 → 0.58	0.396
FID score	159.9088	178.9562	95.5836	198.1082	74.7530	104.9406	90.0923	171.7221

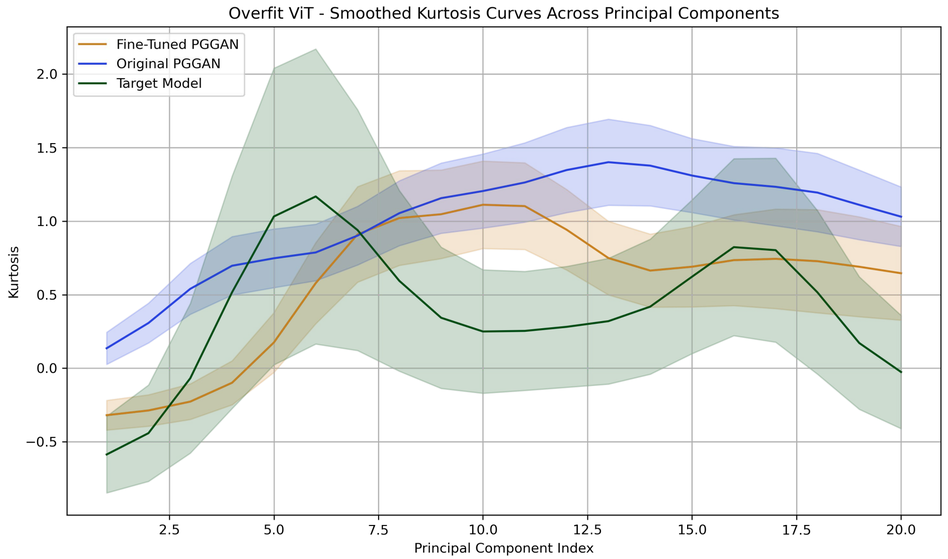




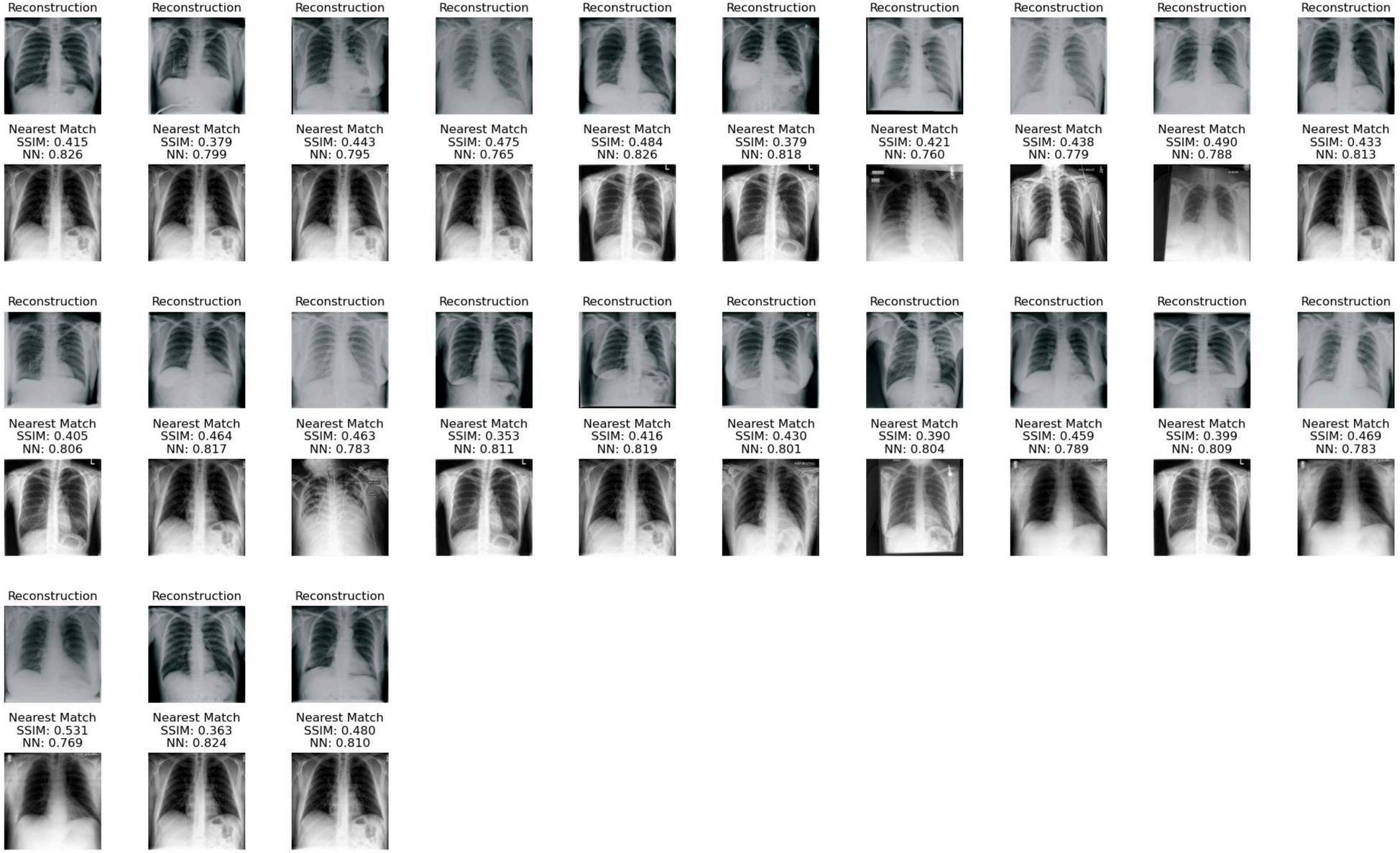
Distribution shifts - ViT models

	CNN		Overfit CNN		ViT		Overfit ViT	
	Original → Fine-tuned	Target CNN	Original → Fine-tuned	Target Overfit CNN	Original → Fine-tuned	Target ViT	Original → Fine-tuned	Target Overfit ViT
Pixel entropy	7.402 → 6.486	7.963	7.402 → 6.176	7.965	7.402 → 6.176	7.963	7.402 → 6.125	7.962
PCA kurtosis	1.065 → 0.436	4.306	1.065 → 0.566	0.532	1.065 → 1.001	4.299	1.065 → 0.58	0.396
FID score	159.9088	178.9562	95.5836	198.1082	74.7530	104.9406	90.0923	171.7221

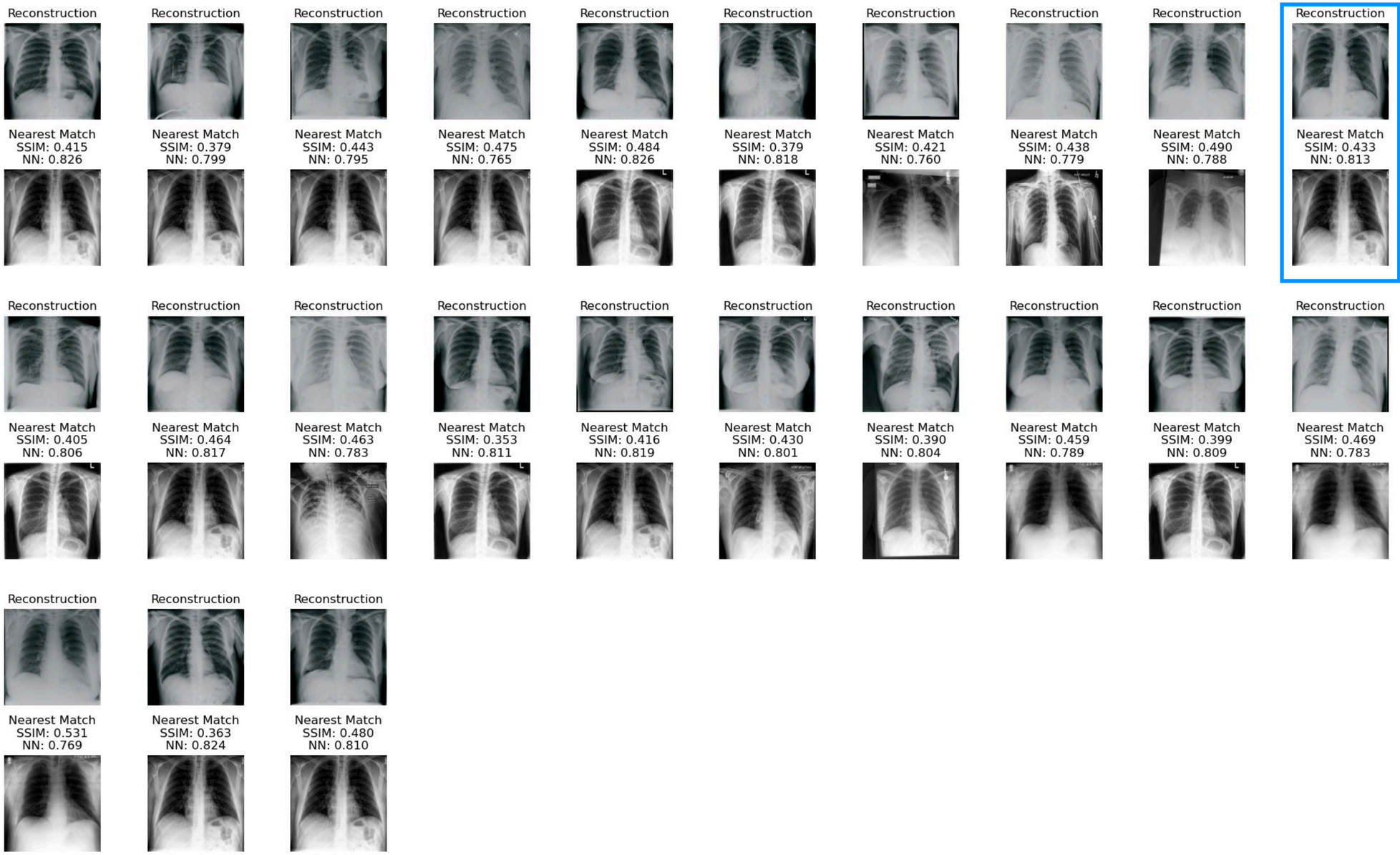


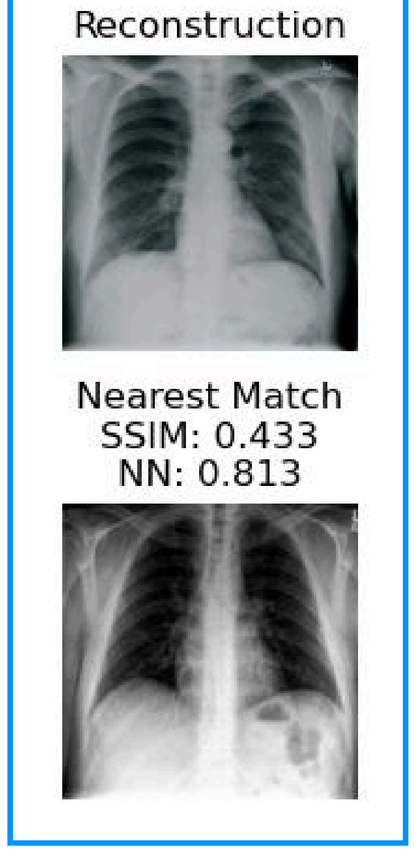


Overfit CNN Likely Samples (n=23)



Overfit CNN Likely Samples (n=23)



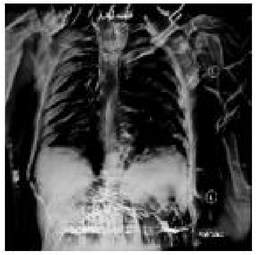


CNN Likely Samples (n=2)

Reconstruction

Nearest Match
SSIM: 0.125
NN: 0.964

Reconstruction



Nearest Match
SSIM: 0.093
NN: 0.914

ViT Likely Samples (n=1)

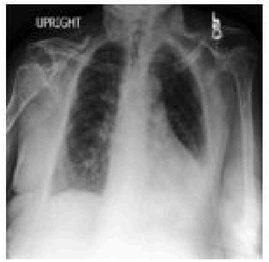
Reconstruction

Nearest Match
SSIM: 0.501
NN: 0.978

Overfit ViT Likely Samples (n=2)

Reconstruction

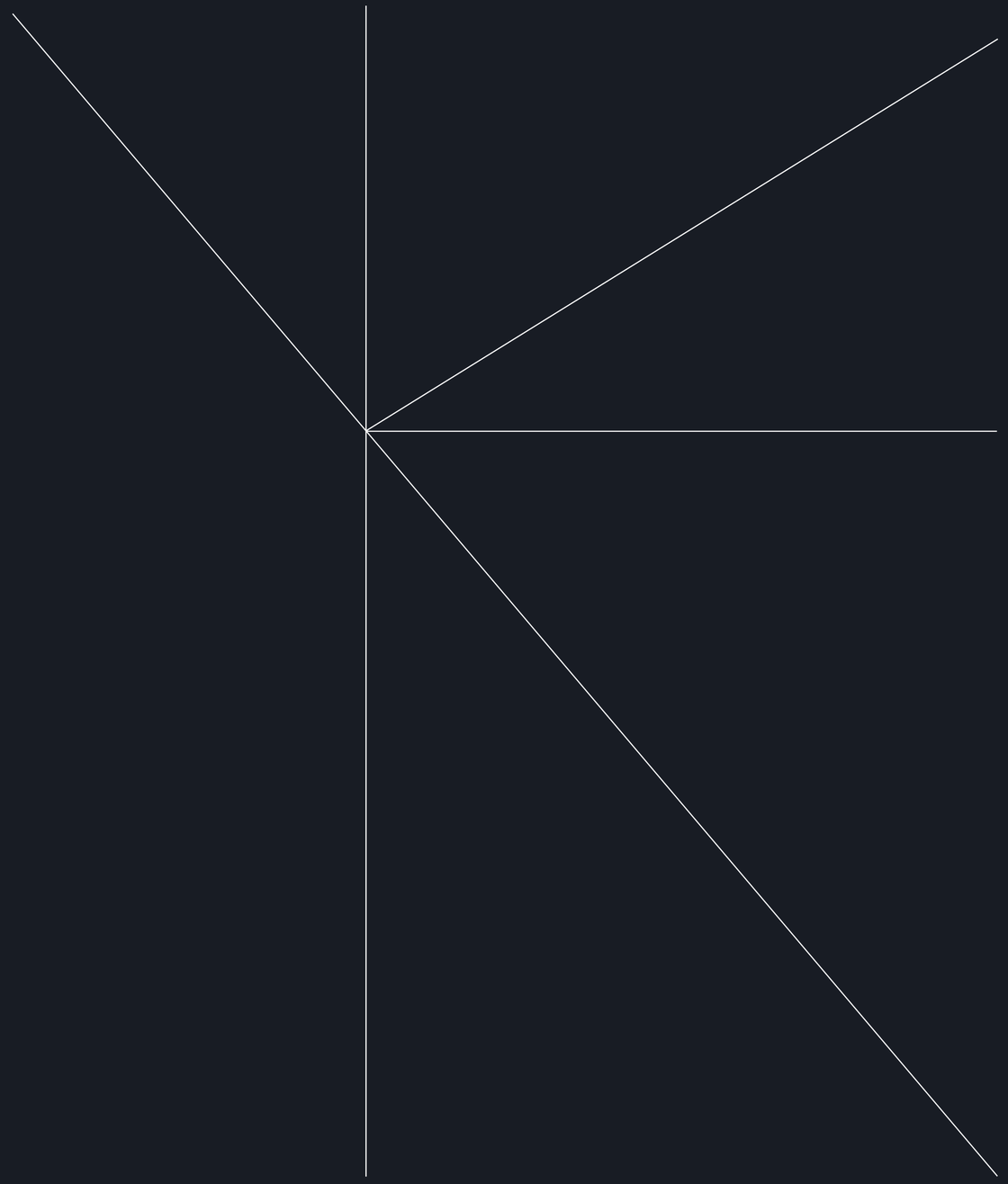
Nearest Match
SSIM: 0.532
NN: 0.860



Reconstruction

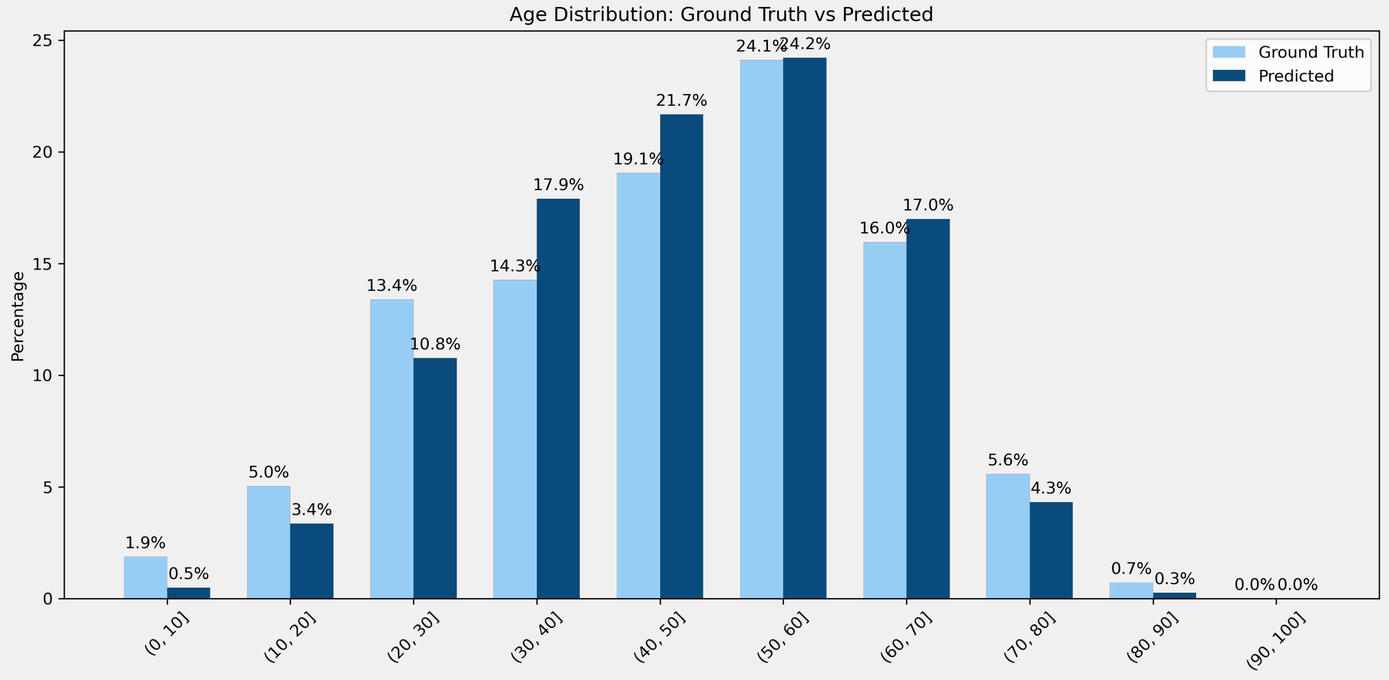
Nearest Match
SSIM: 0.473
NN: 0.903

6. Demographic prediction



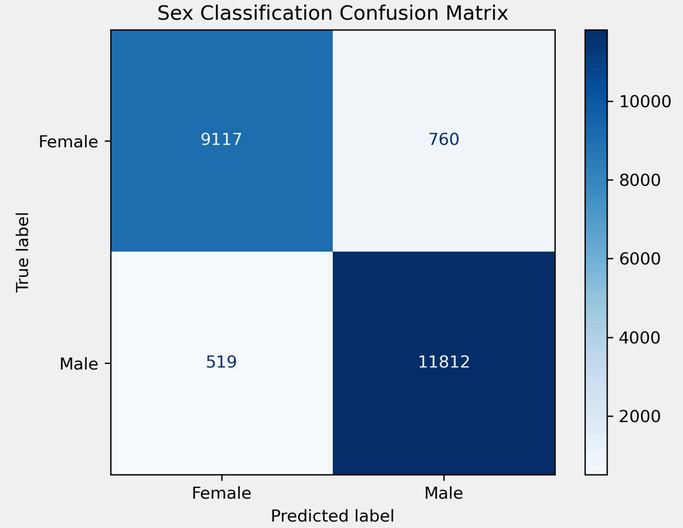
Model

Regression task



Best val MAE: 5.12 Best val RMSE: 6.63 Best val R^2: 0.845

Classification task



Best val AUC: 0.991

Best val accuracy: 0.950

Demographic predictions

Overfit CNN

	PredAge_Recon	PredAge_NN	Age_GT	AbsoluteError_Recon	AbsoluteError_NN	PredSex_Recon	PredSex_NN	Sex_GT	Correct_Recon	Correct_NN
0	51.083	36.347	38.0	13.083	1.653	Male	Male	Male	True	True
1	40.074	36.347	38.0	2.074	1.653	Male	Male	Male	True	True
2	57.408	36.347	38.0	19.408	1.653	Male	Male	Male	True	True
3	51.036	36.347	38.0	13.036	1.653	Male	Male	Male	True	True
4	66.833	38.156	36.0	30.833	2.156	Female	Female	Female	True	True
5	56.344	38.156	36.0	20.344	2.156	Male	Female	Female	False	True
6	42.078	54.446	60.0	17.922	5.554	Male	Male	Male	True	True
7	51.957	48.092	49.0	2.957	0.908	Male	Female	Female	False	True
8	61.132	43.36	31.0	30.132	12.36	Male	Female	Female	False	True
9	61.883	36.347	38.0	23.882	1.653	Male	Male	Male	True	True
10	56.758	38.156	36.0	20.758	2.156	Male	Female	Female	False	True
11	56.860	36.347	38.0	18.860	1.653	Male	Male	Male	True	True
12	47.131	37.415	48.0	0.869	10.585	Female	Male	Male	False	True
13	61.876	38.156	36.0	25.876	2.156	Female	Female	Female	True	True
14	54.889	36.348	38.0	16.889	1.653	Female	Male	Male	False	True
15	49.812	44.66	32.0	17.712	12.66	Female	Male	Male	False	True
16	49.670	22.176	28.0	21.670	5.824	Male	Female	Female	False	True
17	63.451	55.218	44.0	19.451	11.218	Male	Male	Male	True	True
18	38.040	38.156	36.0	2.040	2.156	Female	Female	Female	True	True
19	42.856	55.218	44.0	1.144	11.218	Male	Male	Male	True	True
20	65.286	55.218	44.0	21.286	11.218	Male	Male	Male	True	True
21	45.753	36.347	38.0	7.753	1.653	Male	Male	Male	True	True
22	67.026	36.347	38.0	29.026	1.653	Male	Male	Male	True	True

Demographic predictions

CNN

	PredAge_Recon	PredAge_NN	Age_GT	AbsoluteError_Recon	AbsoluteError_NN	PredSex_Recon	PredSex_NN	Sex_GT	Correct_Recon	Correct_NN
0	53.775	27.856	21.0	32.775	6.856	Female	Male	Male	False	True
1	56.651	64.735	39.0	17.651	25.735	Female	Male	Male	False	True

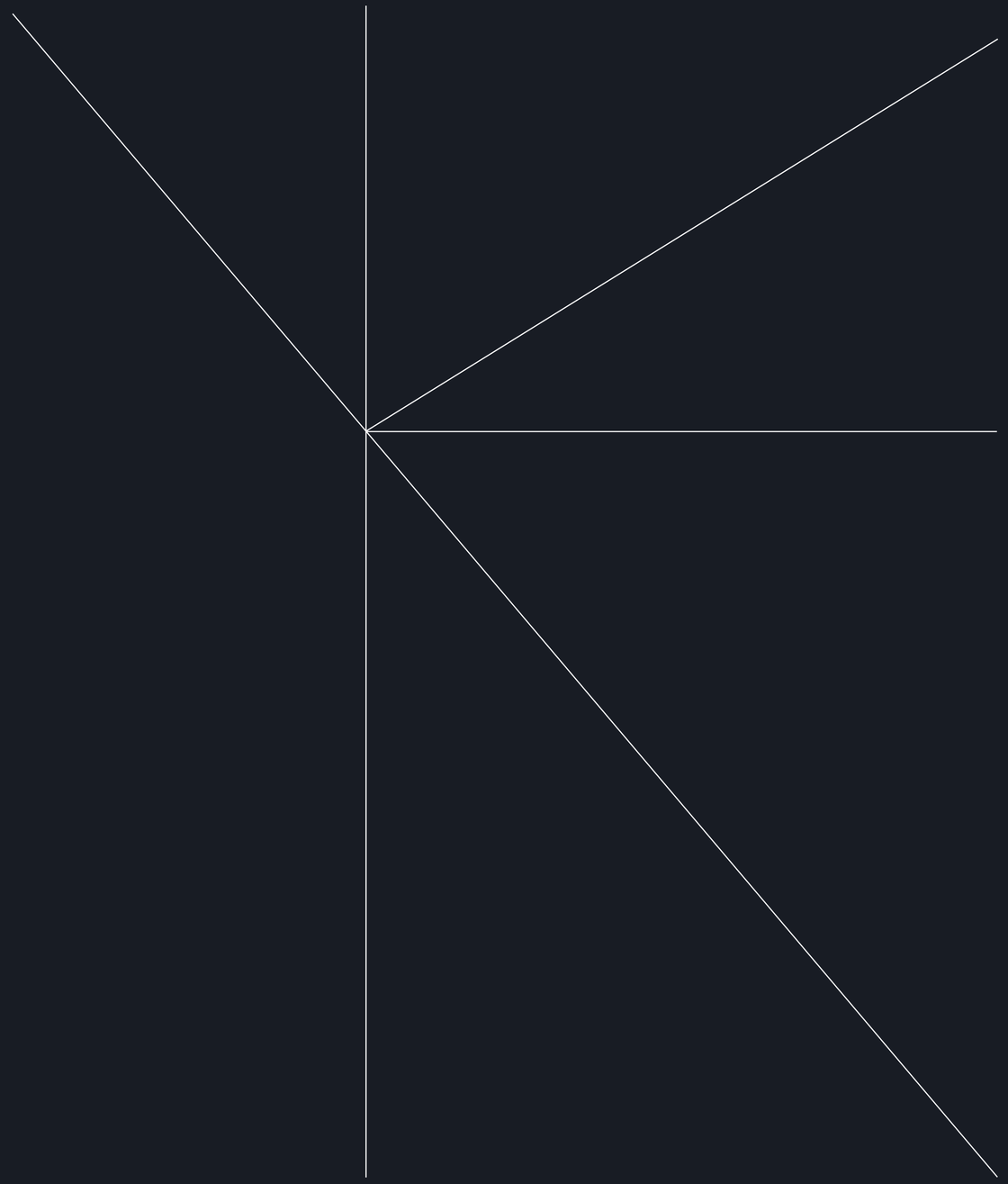
ViT

	PredAge_Recon	PredAge_NN	Age_GT	AbsoluteError_Recon	AbsoluteError_NN	PredSex_Recon	PredSex_NN	Sex_GT	Correct_Recon	Correct_NN
0	45.282	73.385	69.0	23.718	4.385	Male	Male	Male	True	True

Overfit ViT

	PredAge_Recon	PredAge_NN	Age_GT	AbsoluteError_Recon	AbsoluteError_NN	PredSex_Recon	PredSex_NN	Sex_GT	Correct_Recon	Correct_NN
0	43.817	30.173	33.0	10.817	2.827	Female	Female	Female	True	True
1	57.571	60.202	54.0	3.571	6.202	Male	Male	Male	True	True

7. Discussion



Architectural and training differences

CNN & Overfit CNN

Farther distribution distance from PGGAN.

More overlap between similarly activated images and images close in distance to the target model.

Leaves the original manifold more aggressively but sacrifices realism.

ViT & Overfit ViT

Closer distribution distance to both PGGAN and target model.

Does not leave the original manifold as much but still produces semantically plausible images.

Less amount of likely samples but more accurate demographic predictions.

Privacy implications

With a model's parameters from a frozen checkpoint, the ViT may still pose a greater risk to patient re-identification.

Yet, the CNN leads to reconstructions that are more closely aligned with the target distribution.

This may be, in part, due to how memory is encoded in each architecture's parameters.

Considerations & limitations

01

Only uses 2D images

02

Focus is on chest x-rays, and does not account for other modalities or regions

03

Classification models only, no segmentation, regression, etc.

04

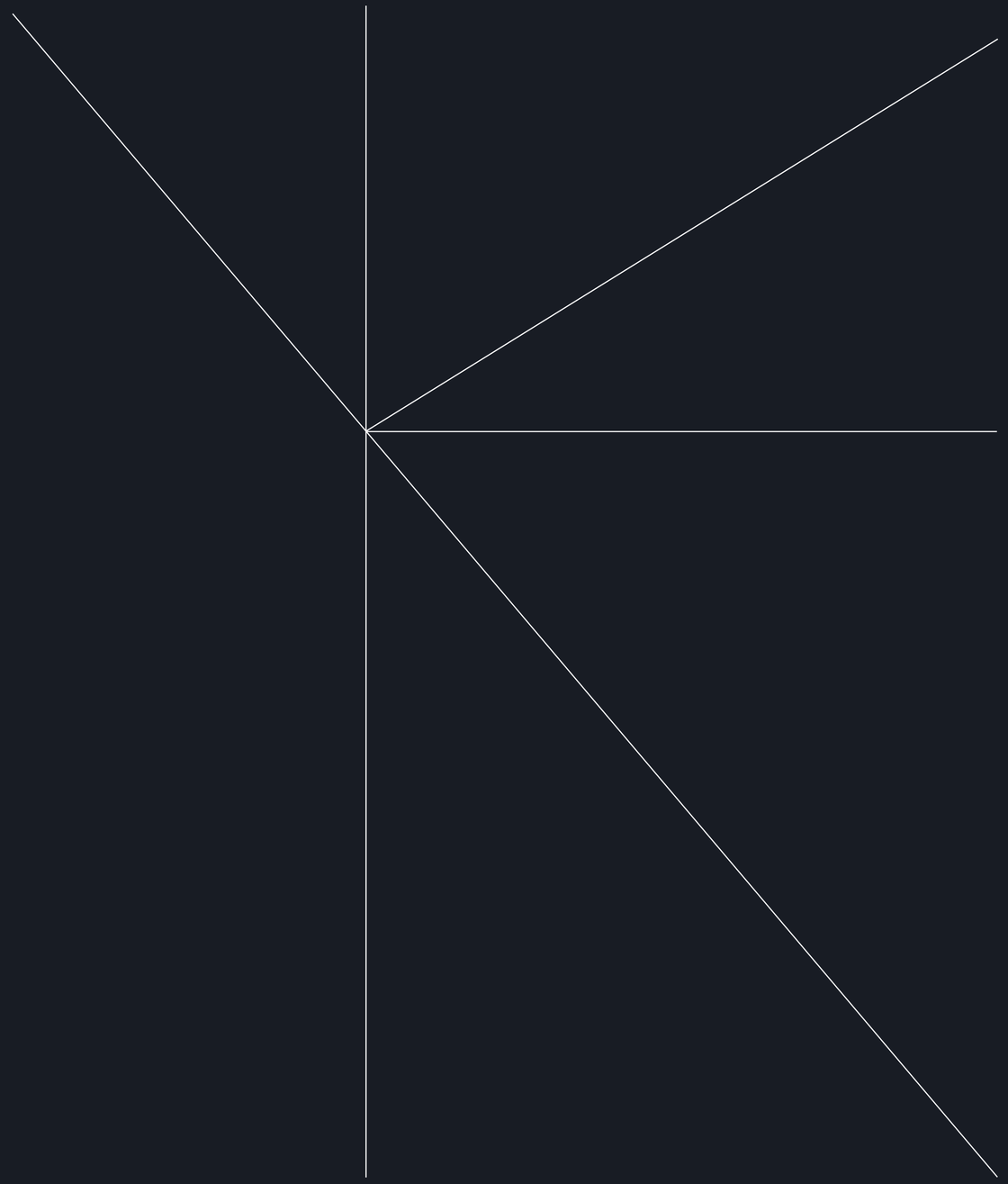
Study does not attempt to link reconstructions and demographics to real patient identities

Conclusion

Images and their predicted demographics have the potential to re-identify a patient, depending on the circumstances.

However, ensuring patient privacy involves a tradeoff with enabling innovation and technical advancement.

8. References



1. Clunie DA, Flanders A, Taylor A, et al. Report of the Medical Image De-Identification (MIDI) Task Group - Best Practices and Recommendations. ArXiv. Published online April 1, 2023:arXiv:2303.10473v2.
2. Xia W, Liu Y, Wan Z, et al. Enabling realistic health data re-identification risk assessment through adversarial modeling. J Am Med Inform Assoc. 2021;28(4):744-752. doi:10.1093/jamia/ocaa327
3. Fernandez V, Sanchez P, Pinaya WHL, Jacenków G, Tsaftaris SA, Cardoso MJ. Privacy Distillation: Reducing Re-identification Risk of Diffusion Models. In: Mukhopadhyay A, Oksuz I, Engelhardt S, Zhu D, Yuan Y, eds. Deep Generative Models. Springer Nature Switzerland; 2024:3-13. doi:10.1007/978-3-031-53767-7_1
4. Carlini N, Hayes J, Nasr M, et al. Extracting Training Data from Diffusion Models. In: ; 2023:5253-5270. Accessed March 31, 2025. https://www.usenix.org/conference/usenixsecurity23/presentation/carlini
5. Zhang G, Liu B, Tian H, Zhu T, Ding M, Zhou W. How Does a Deep Learning Model Architecture Impact Its Privacy? A Comprehensive Study of Privacy Attacks on {CNNs} and Transformers. In: ; 2024:6795-6812. Accessed March 23, 2025. https://www.usenix.org/conference/usenixsecurity24/presentation/zhang-guangsheng
6. Carlini N, Tramèr F, Wallace E, et al. Extracting Training Data from Large Language Models. In: ; 2021:2633-2650. Accessed March 23, 2025. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting
7. Zhu L, Liu Z, Han S. Deep Leakage from Gradients. In: Advances in Neural Information Processing Systems. Vol 32. Curran Associates, Inc.; 2019. Accessed March 20, 2025. https://proceedings.neurips.cc/paper/2019/hash/60a6c4002cc7b29142def8871531281a-Abstract.html
8. Yin H, Molchanov P, Li Z, et al. Dreaming to Distill: Data-free Knowledge Transfer via DeepInversion. Published online June 16, 2020. doi:10.48550/arXiv.1912.08795

Acknowledgements