2023

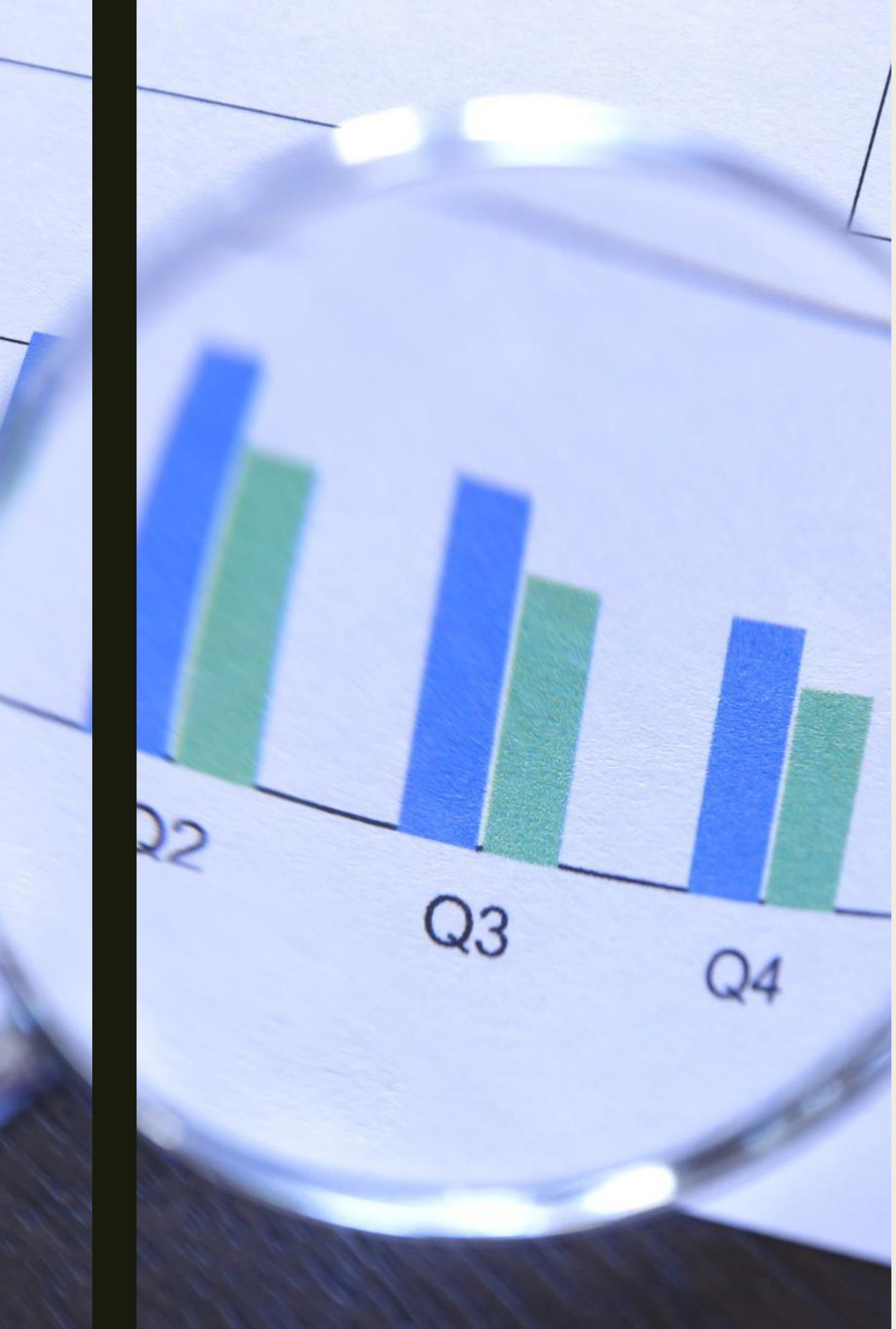
ROBUSTNESS OF VISION TRANSFORMER

Muhammad Huzaifa, Raza Imam, Mohammed El-Amine Azz

ML703

MBZUAI

INTRODUCTION

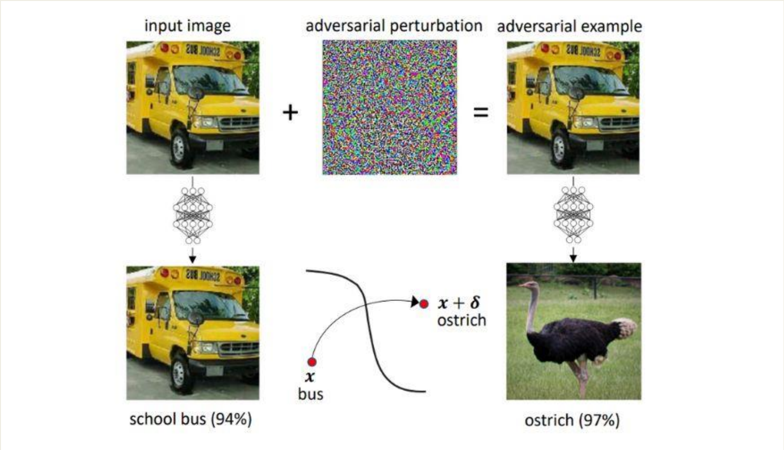


Introduction

- Rise of ViTs (Vision transformers) in computer vision.
- The enhanced performance of ViTs over other methods.
- The problem of adversarial attacks on models.
- Proposed SEViTs (Self-Ensembling ViTs).

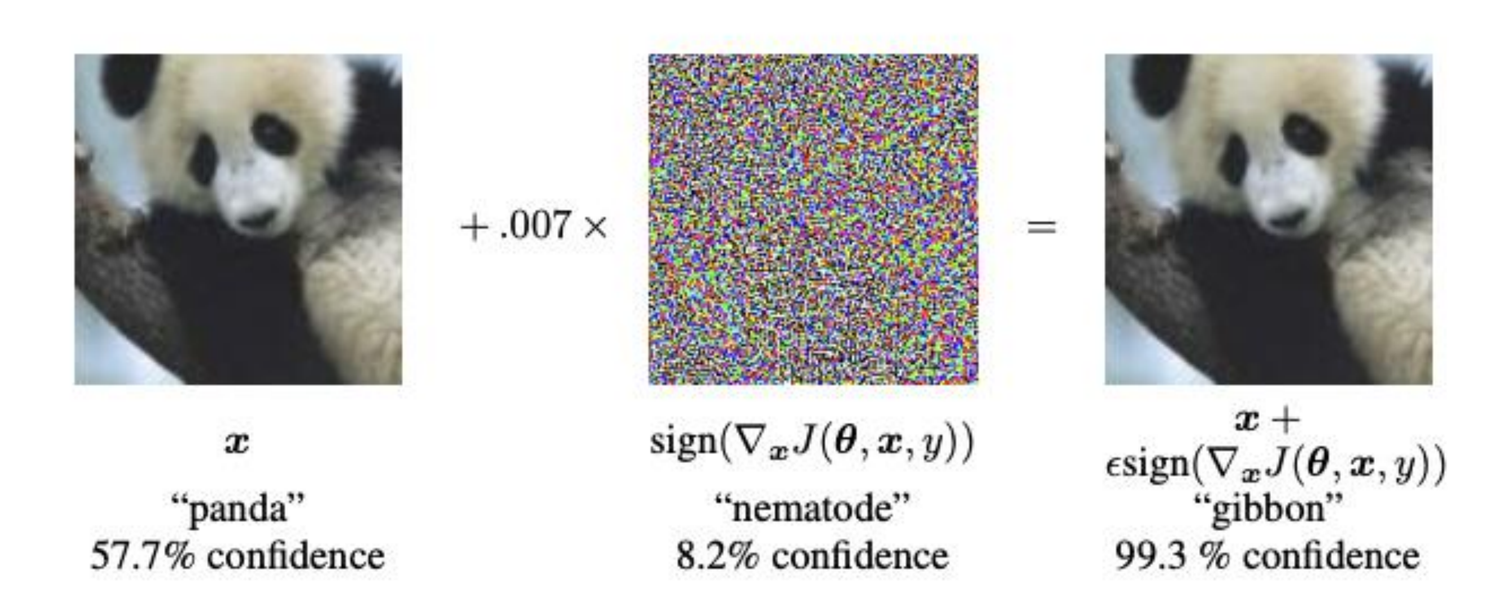
Motivation

- CNNs vulnerability to adversarial attacks.
- Limited literature on the robustness of SEViTs.
- These issues are important in many fields like medical data, and insurance fraud detection.
- Aim:
  - Explore the robustness of ViTs and SEViTs.
  - Explore defensive measures and their effectiveness against adversarial attacks.



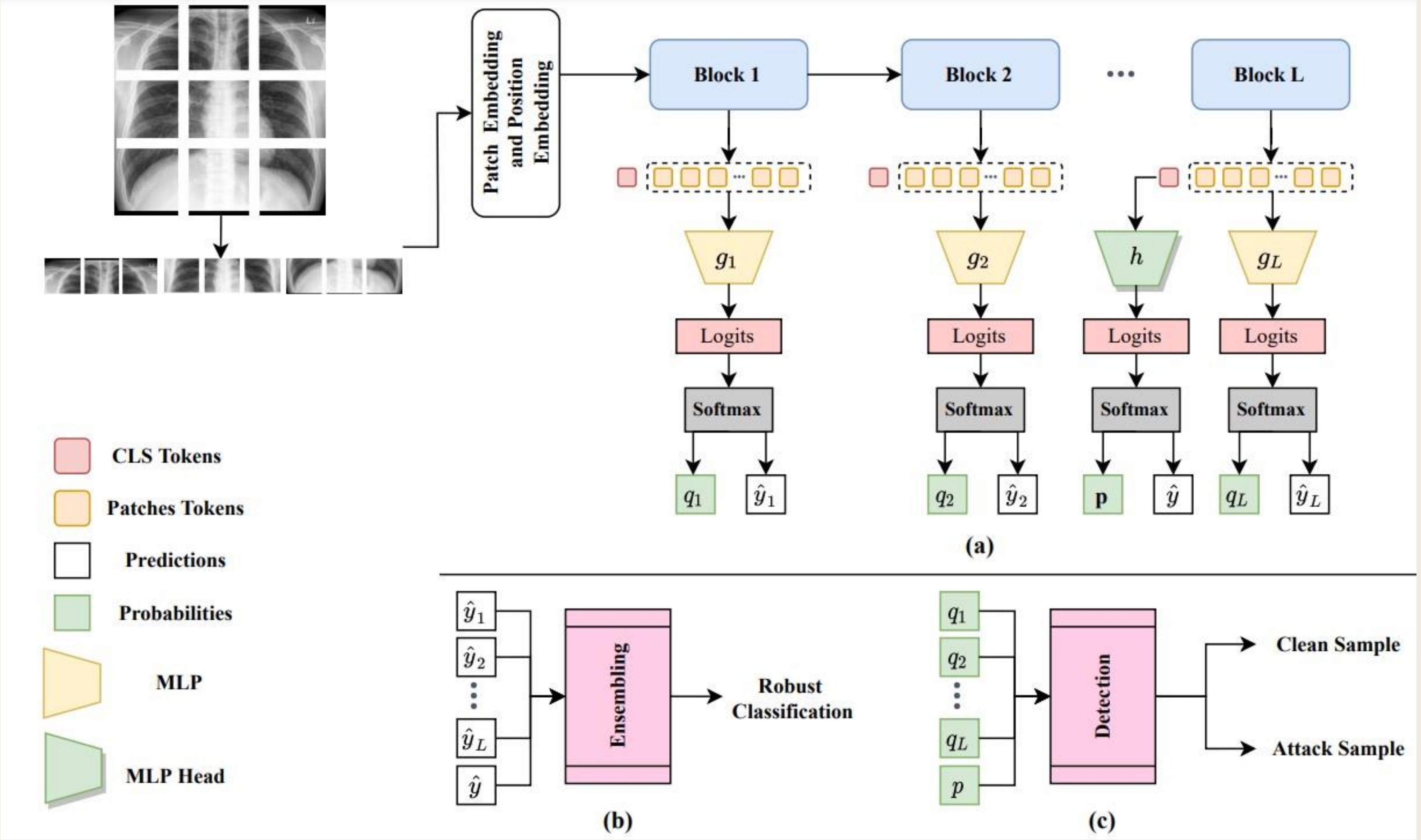
Adversarial Attacks

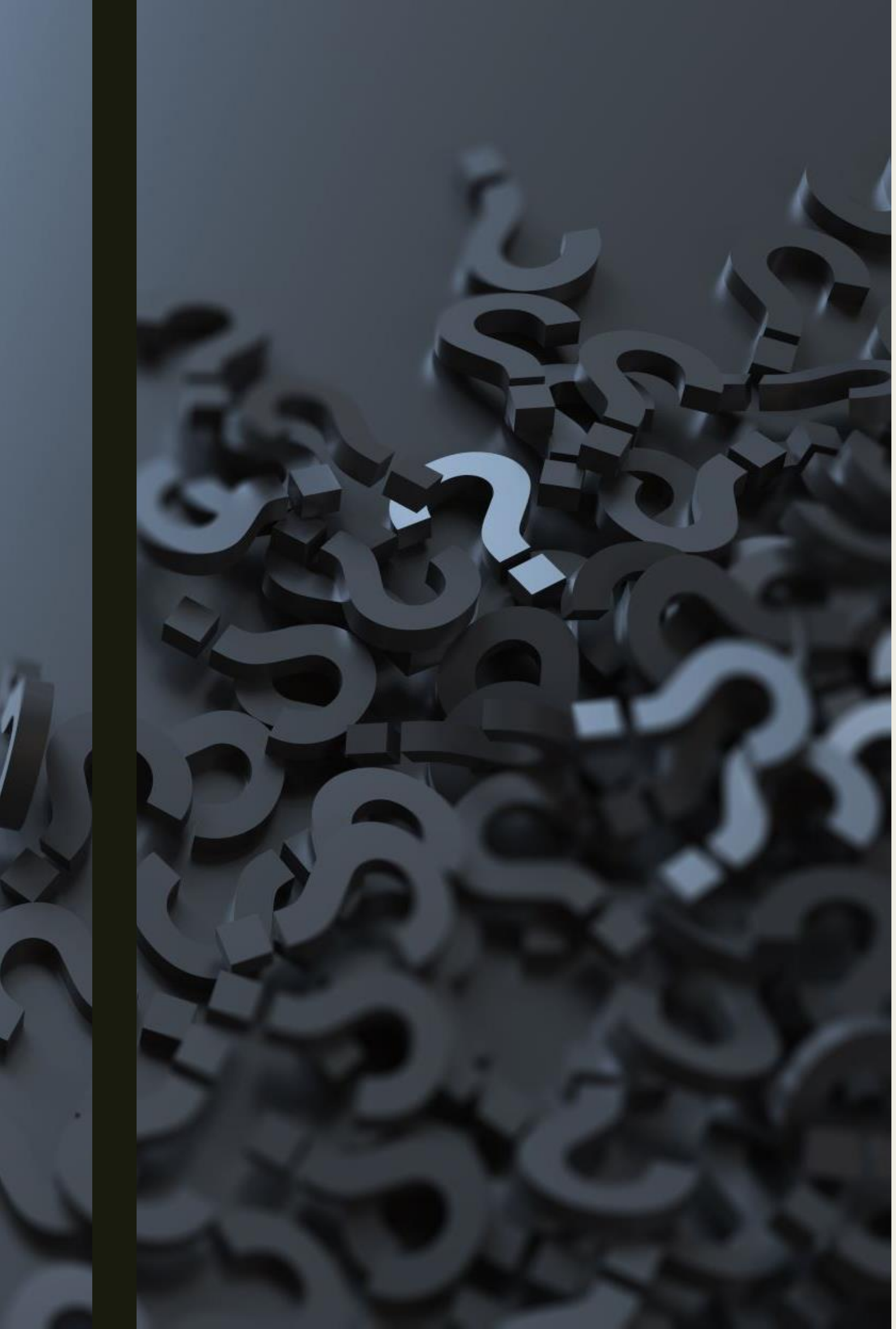
- **FGSM:** $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{r}, \boldsymbol{r} = \epsilon\, sign\, \nabla \boldsymbol{x}\, \mathcal{L}(\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta}), y_s)$
- **BIM:** $\boldsymbol{x}'_0 = \boldsymbol{x}, \boldsymbol{x}_{k+1}' = Clip_{\boldsymbol{x},\epsilon}\{\boldsymbol{x}_k' + \alpha\, sign\, \nabla_{\boldsymbol{x}}\, \mathcal{L}(\boldsymbol{f}(\boldsymbol{x}_k'; \boldsymbol{\theta}), y_s)\}$
- **PGD:** Same as BIM with random initialization ($\boldsymbol{x}'_0 = \boldsymbol{x} + \boldsymbol{r}$, $\boldsymbol{r}$ is random s.t $|\boldsymbol{r}|_\infty \leq \epsilon$)



SEVIT

- Self-Ensembling ViTs:
- Multiple blocks of ViTs where one side output goes to the next as an input.
- Classifies by majority vote of each block's output.
- We can limit the number of blocks.





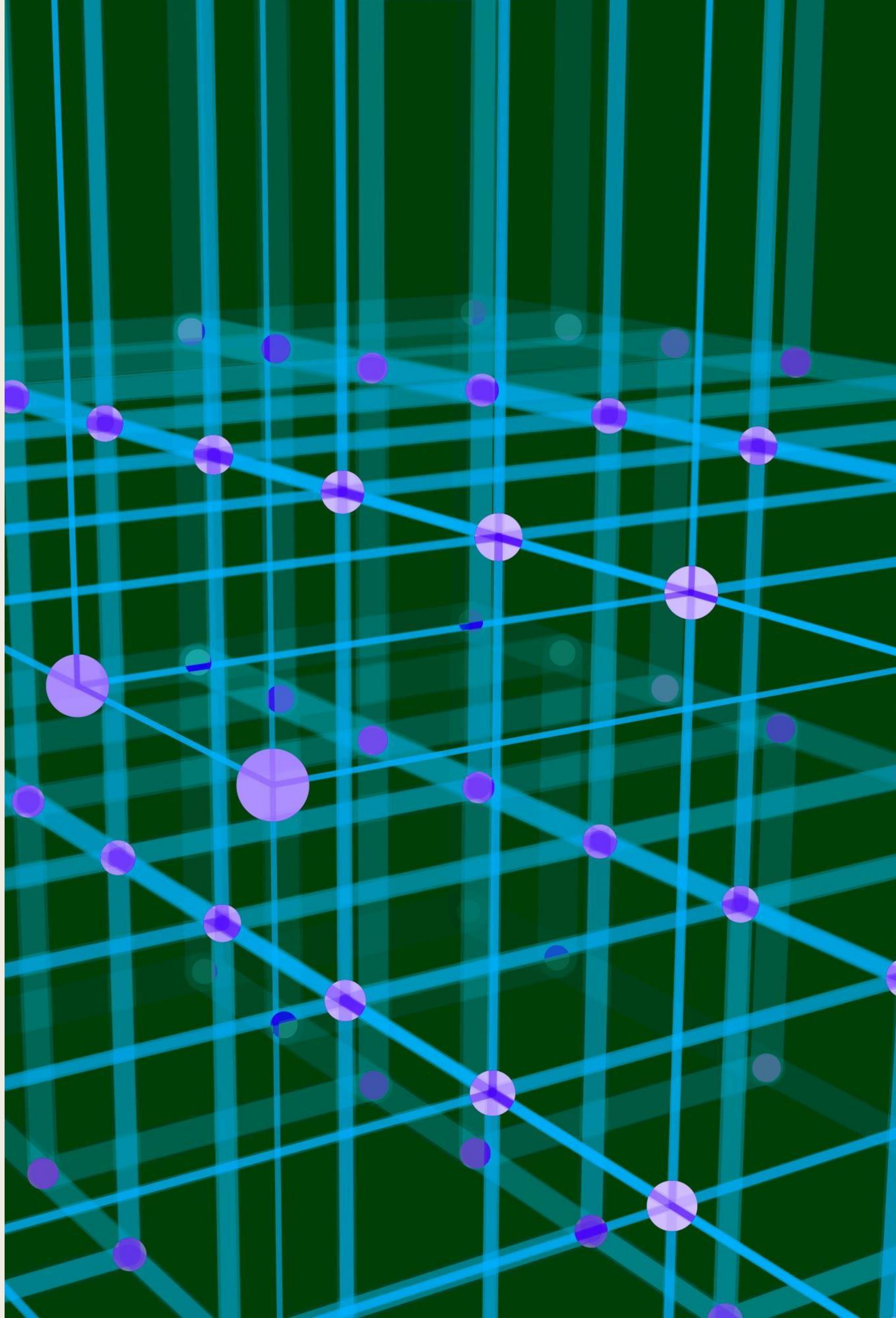
Problem Statement

- We will explore the robustness of ViTs and its variant SEViT, along with defensive measures against adversarial.

EDA

- Dataset: TB dataset (Image/class dataset).
- Classes: 2 (Normal, Tuberculosis).
- Observations: ~6500 (5000 for training and 700 for validation and testing)
- Missing Data: no missing data
- Original Image sizes: 512 x 512 (Images of 3 channels)
- Preprocessing: Resizing Images to 224 x 224

METHOD



Diffusion Model Boom!

- **Diffusion model is SOTA on image generation**
  - Beat BigGAN and StyleGAN on high-resolution images

Figure 1: Selected samples from our best ImageNet 512×512 model (FID 3.85)

Model	FID	sFID	Prec	Rec
LSUN Bedrooms 256×256				
DCTransformer† [42]	6.40	6.66	0.44	**0.56**
DDPM [25]	4.89	9.07	0.60	0.45
IDDPM [43]	4.24	8.21	0.62	0.46
StyleGAN [27]	2.35	6.62	0.59	0.48
ADM (dropout)	**1.90**	**5.59**	**0.66**	0.51
ImageNet 512×512				
BigGAN-deep [5]	8.43	8.13	**0.88**	0.29
ADM	23.24	10.19	0.73	**0.60**
ADM-G (25 steps)	8.41	9.67	0.83	0.47
ADM-G	**7.72**	**6.57**	0.87	0.42

Methodology

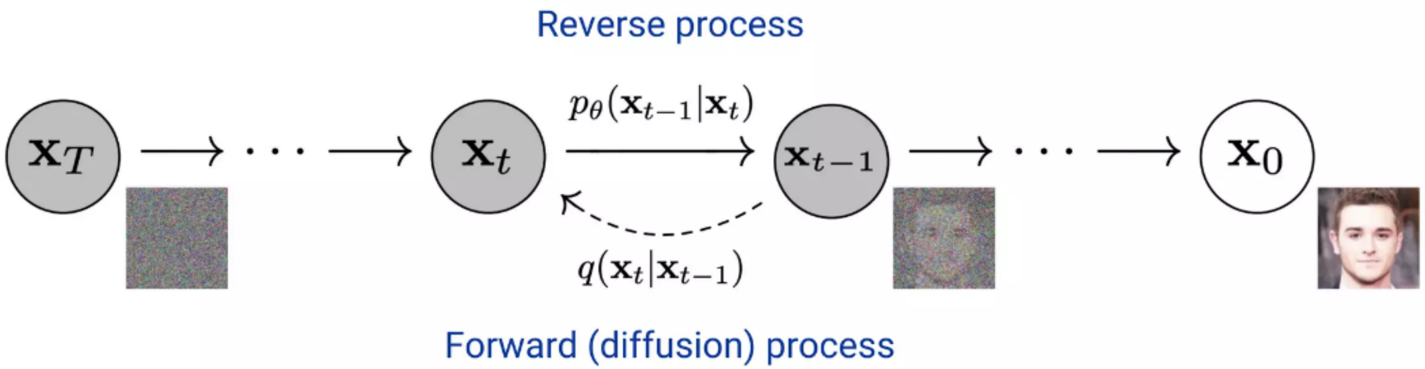
Diffusion Probabilistic Model

- **Diffusion model aims to learn the reverse of noise generation procedure**
  - **Forward step:** (Iteratively) Add noise to the original sample

    → The sample x_0 converges to the complete noise x_T (e.g., $\sim \mathcal{N}(0, I)$)

  - **Reverse step:** Recover the original sample from the noise

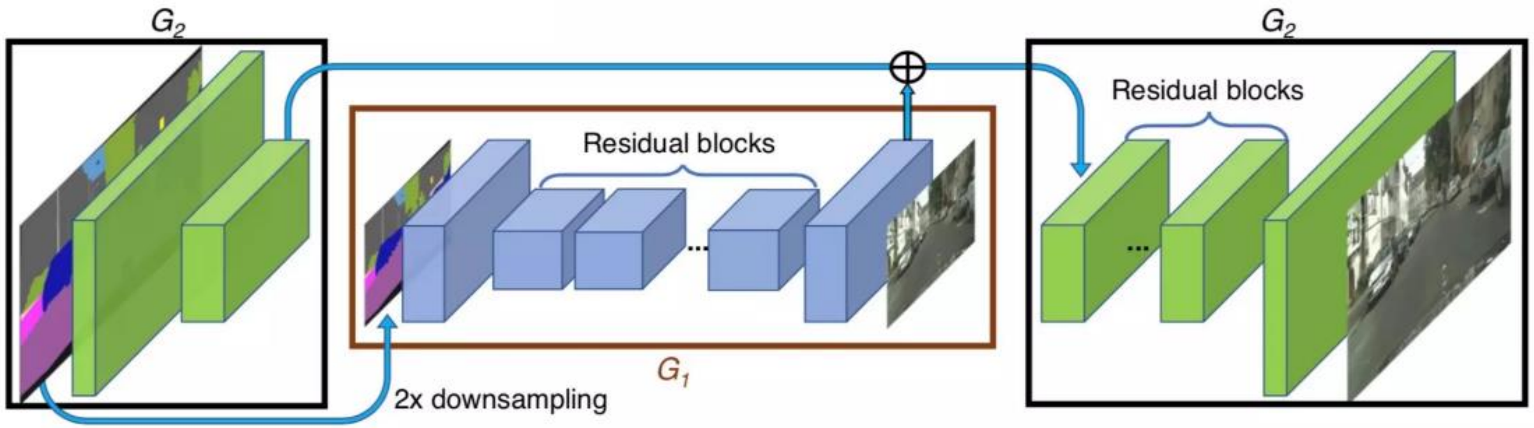
    → Note that it is the "generation" procedure



Methodology

Diffusion Probabilistic Model

- **Diffusion model aims to learn the reverse of noise generation procedure**
  - **Network:** Use the image-to-image translation (e.g., U-Net) architectures
    - Recall that input is $\mathbf{x}_t$ and output is $\mathbf{x}_{t-1}$, both are images
    - It is expensive since both input and output are high-dimensional
    - Note that the denoiser $\mu_\theta(\mathbf{x}_\mathrm{t}, \mathrm{t})$ shares weights, but conditioned by step t



Literature Overview

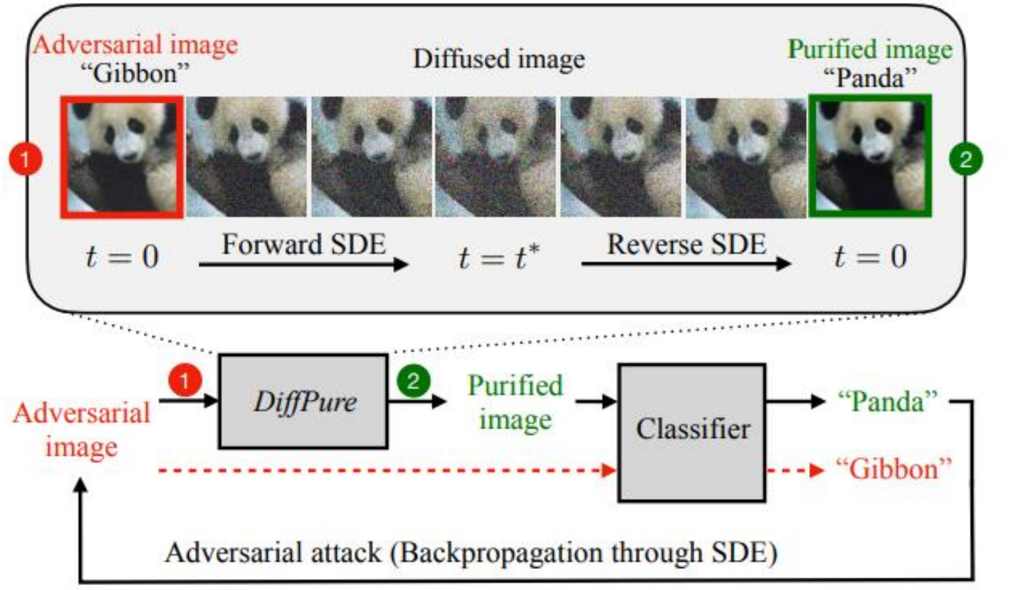


Figure 1. An illustration of *DiffPure*. Given a pre-trained diffusion model, we add noise to adversarial images following the forward diffusion process with a small diffusion timestep t^* to get diffused images, from which we recover clean images through the reverse denoising process before classification. Adaptive attacks backpropagate through the SDE to get full gradients of our defense system.

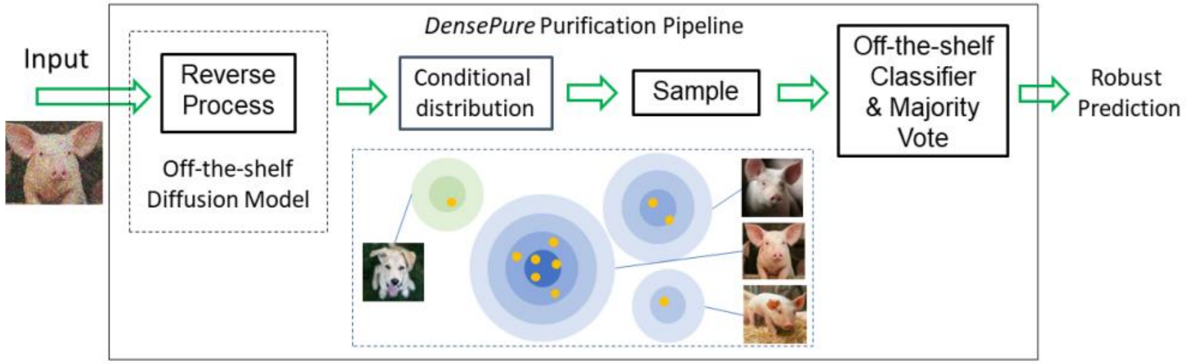
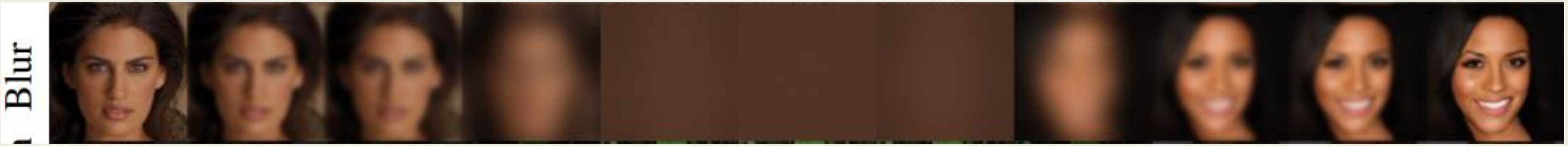


Figure 1: Pipeline of DensePure.

Methodology - Deblurring



$$D(x_0, 0) = x_0.$$

$$R(x_t, t) \approx x_0.$$

$$\min_{\theta} \mathbb{E}_{x \sim \mathcal{X}} \| R_\theta(D(x, t), t) - x \|,$$

Algorithm 1 Naive Sampling

Input: A degraded sample x_t
for $s = t, t-1, \ldots, 1$ **do**
  $\hat{x}_0 \leftarrow R(x_s, s)$
  $x_{s-1} = D(\hat{x}_0, s-1)$
end for
Return: x_0

Methodology - Deblurring

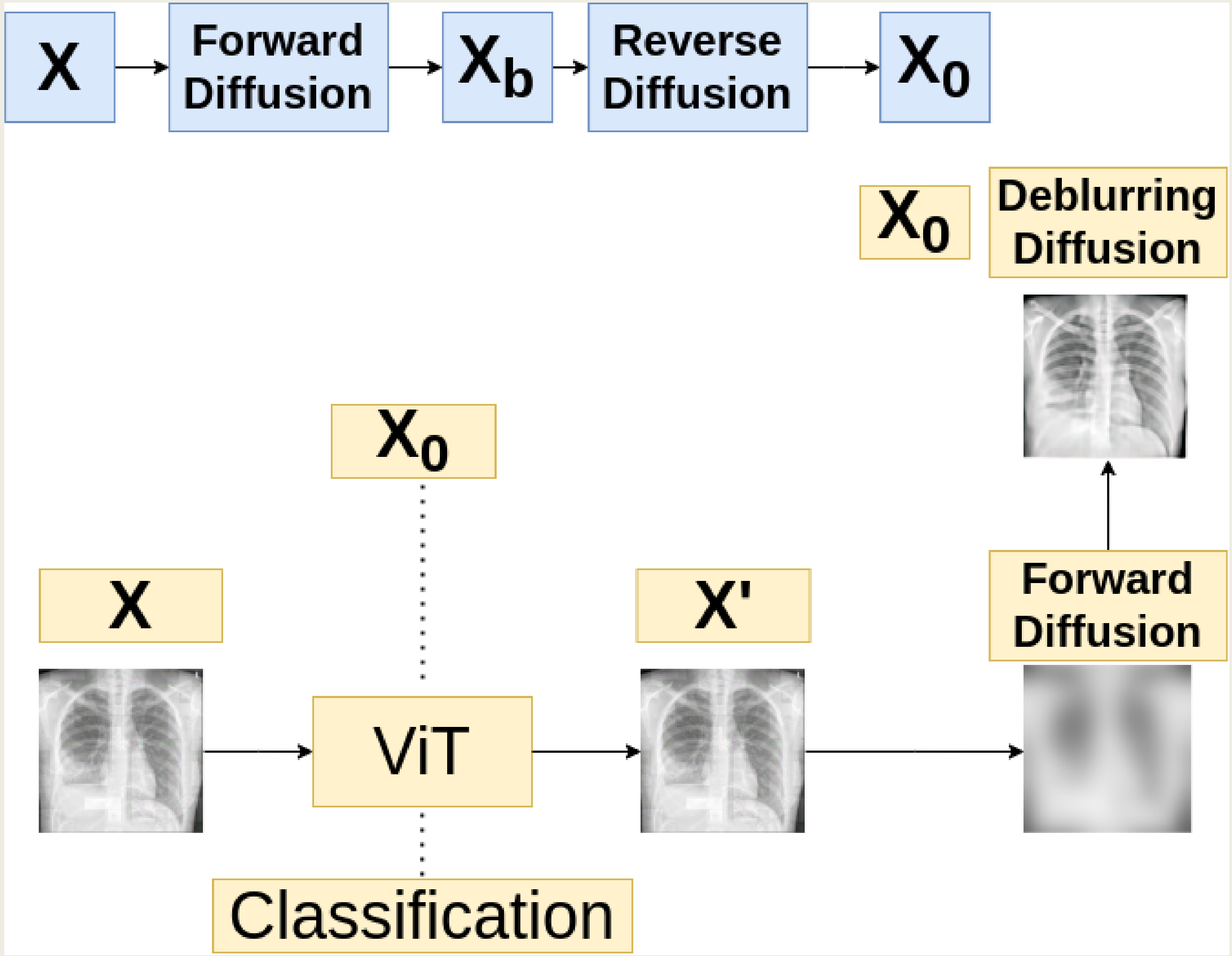
$$x_t = G_t * x_{t-1} = G_t * \ldots * G_1 * x_0 = \bar{G}_t * x_0 = D(x_0, t),$$

Algorithm 2 Improved Sampling for Cold Diffusion

Input: A degraded sample x_t
for $s = t, t-1, \ldots, 1$ **do**
  $\hat{x}_0 \leftarrow R(x_s, s)$
  $x_{s-1} = x_s - D(\hat{x}_0, s) + D(\hat{x}_0, s-1)$
end for

- Deblurring can be thought of as adding frequencies to the image

Our Method – Defensive Diffusion



Methodology Cont.

Natural Images

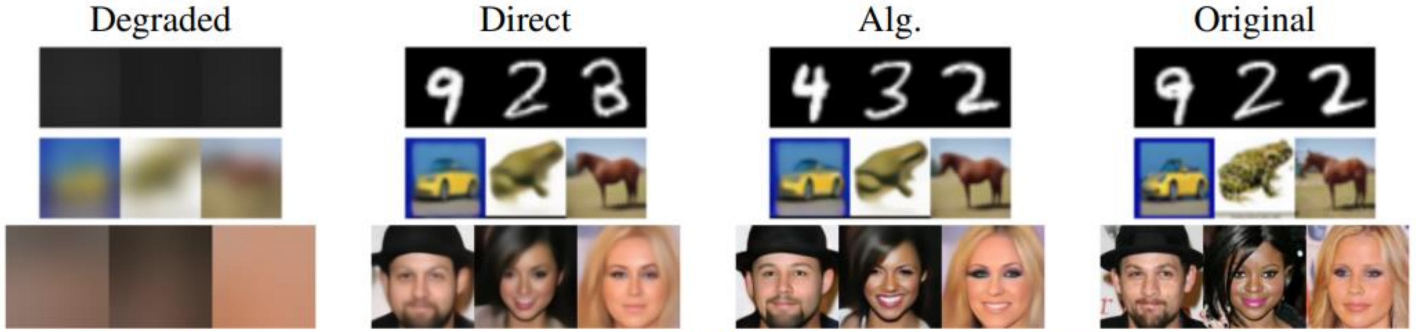
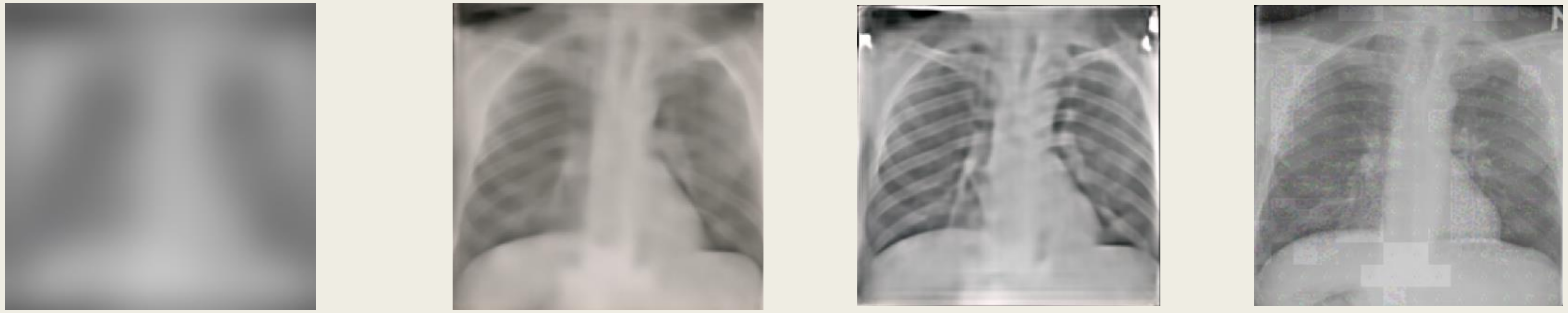


Figure 3: Deblurring models trained on the MNIST, CIFAR-10, and CelebA datasets. **Left to right:** degraded inputs $D(x_0, T)$, direct reconstruction $R(D(x_0, T))$, sampled reconstruction with Algorithm 2, and original image.

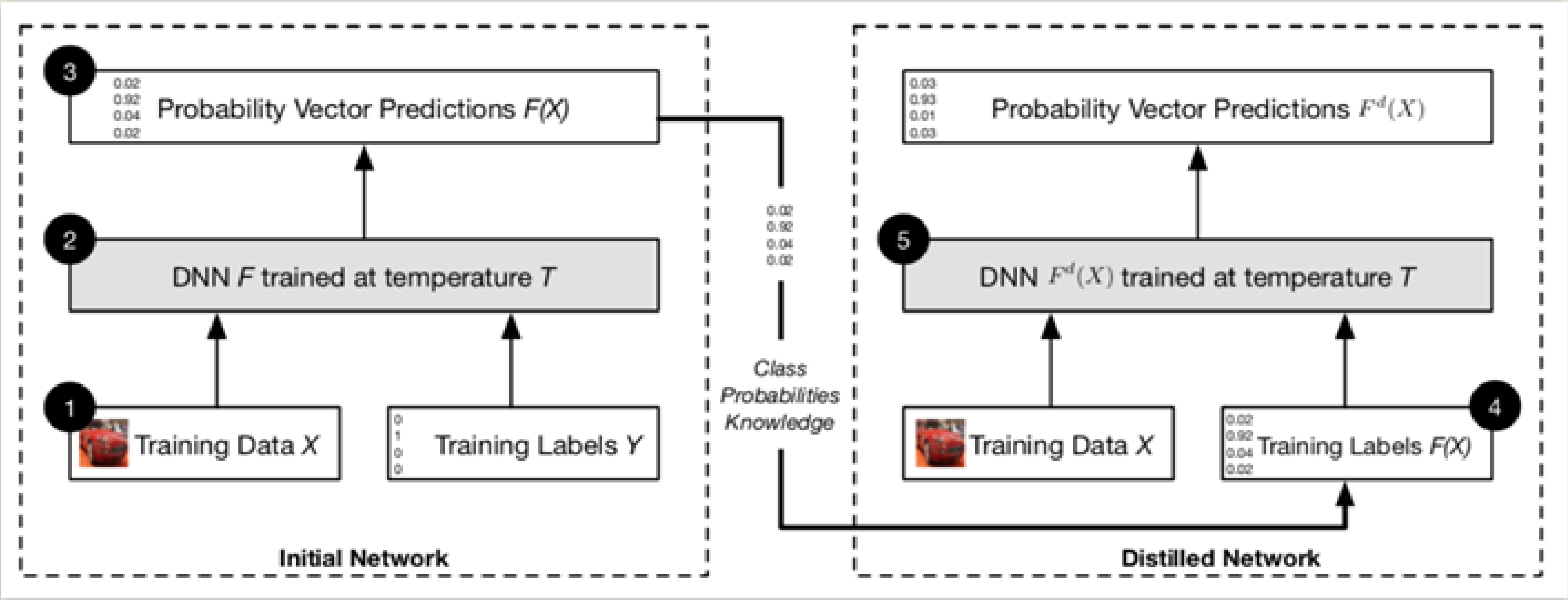
Medical Images



Diffusion Model – Hyperparams

- Time steps = 20
- Train steps = 700,000
- Blur Std = 7.0
- Loss Type = L1
- Batch size = 1
- Unet blocks – 4 up and 4 down (Consisting of ConvNext, Residual and Attention Block)

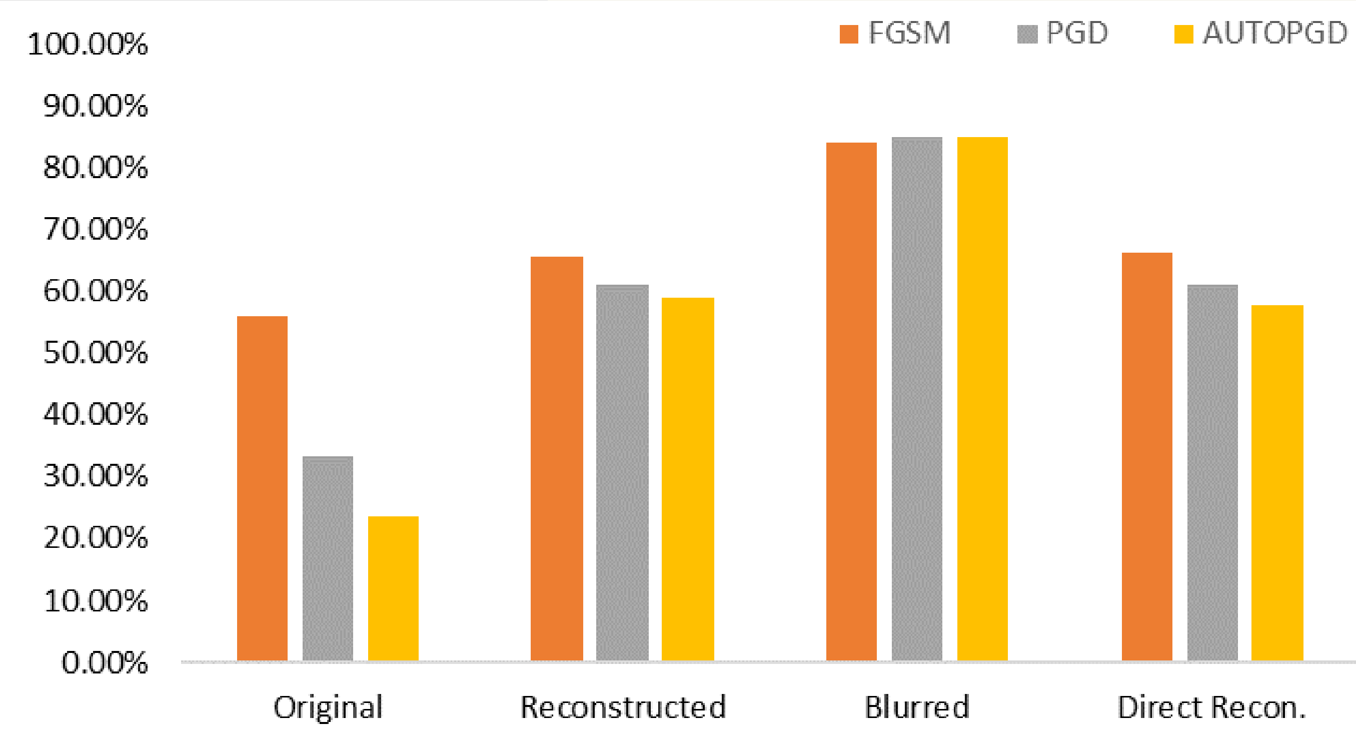
Knowledge distillation



EXPERIMENTS

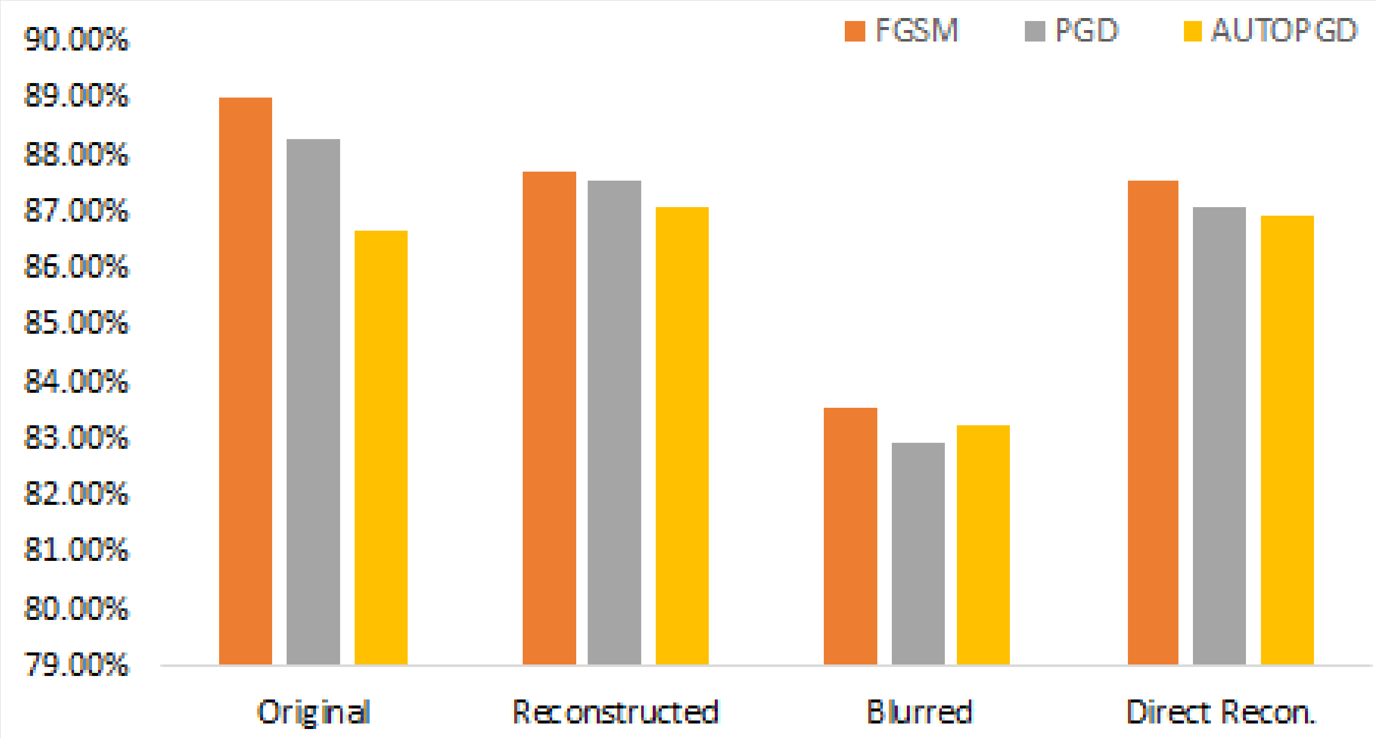
Attack	Original	Reconstructed	Blurred	Direct Recon.
Clean	96.37%	94.34%	93.18%	95.21%
FGSM	55.85%	65.63%	83.85%	66.22%
PGD	33.18%	60.89%	85.03%	60.89%
AUTOPGD	23.56%	58.81%	84.88%	57.63%

Defensive Diffusion on *ViT*



Attack	Original	Reconstructed	Blurred	Direct Recon.
Clean	94.78%	93.04%	91.30%	92.46%
FGSM	89.03%	87.70%	83.55%	87.55%
PGD	88.29%	87.55%	82.96%	87.11%
AUTOPGD	86.67%	87.11%	83.25%	86.96%

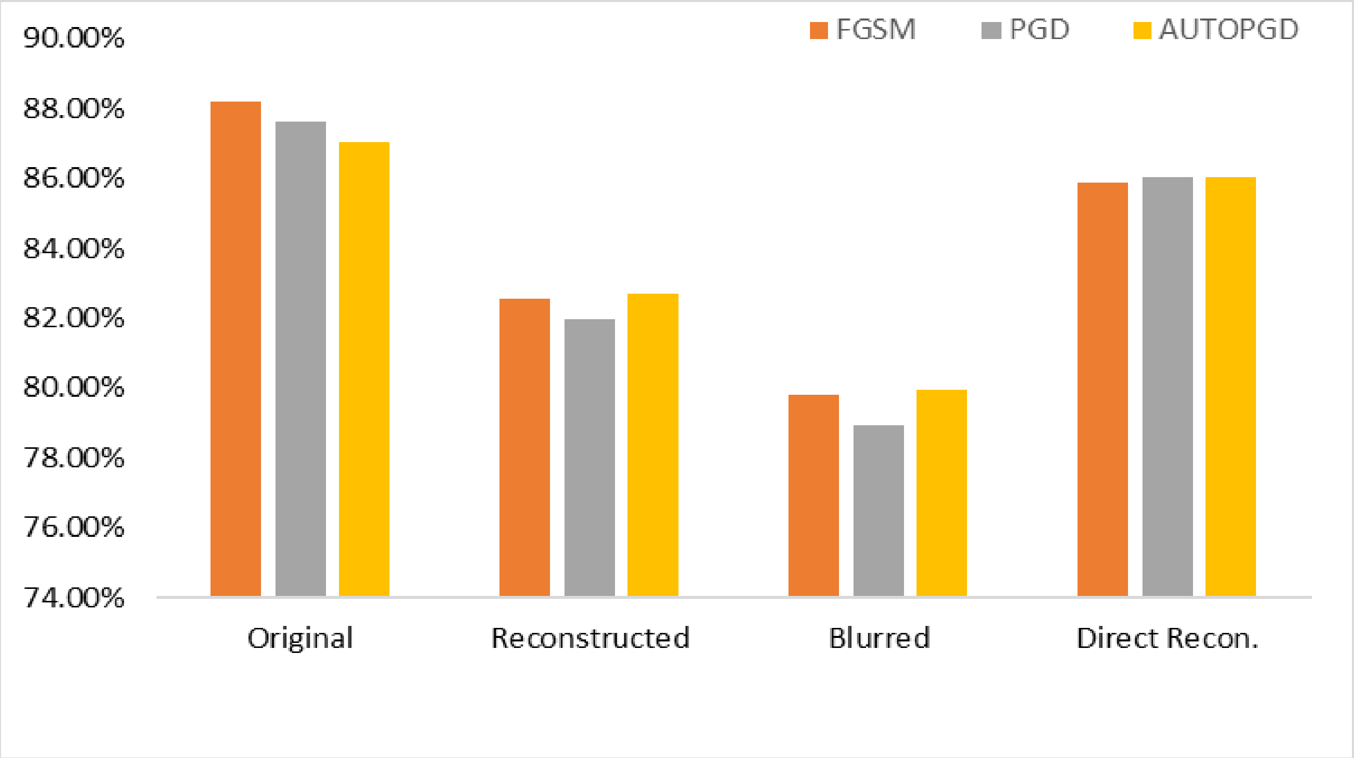
Defensive Diffusion on *SEViT*



Attack	Original	Reconstructed	Blurred	Direct Recon.
Clean	94.27%	91.40%	90.83%	93.27%
FGSM	88.23%	82.56%	79.80%	85.90%
PGD	87.65%	81.98%	78.92%	86.05%
AUTOPGD	87.06%	82.70%	79.94%	86.05%

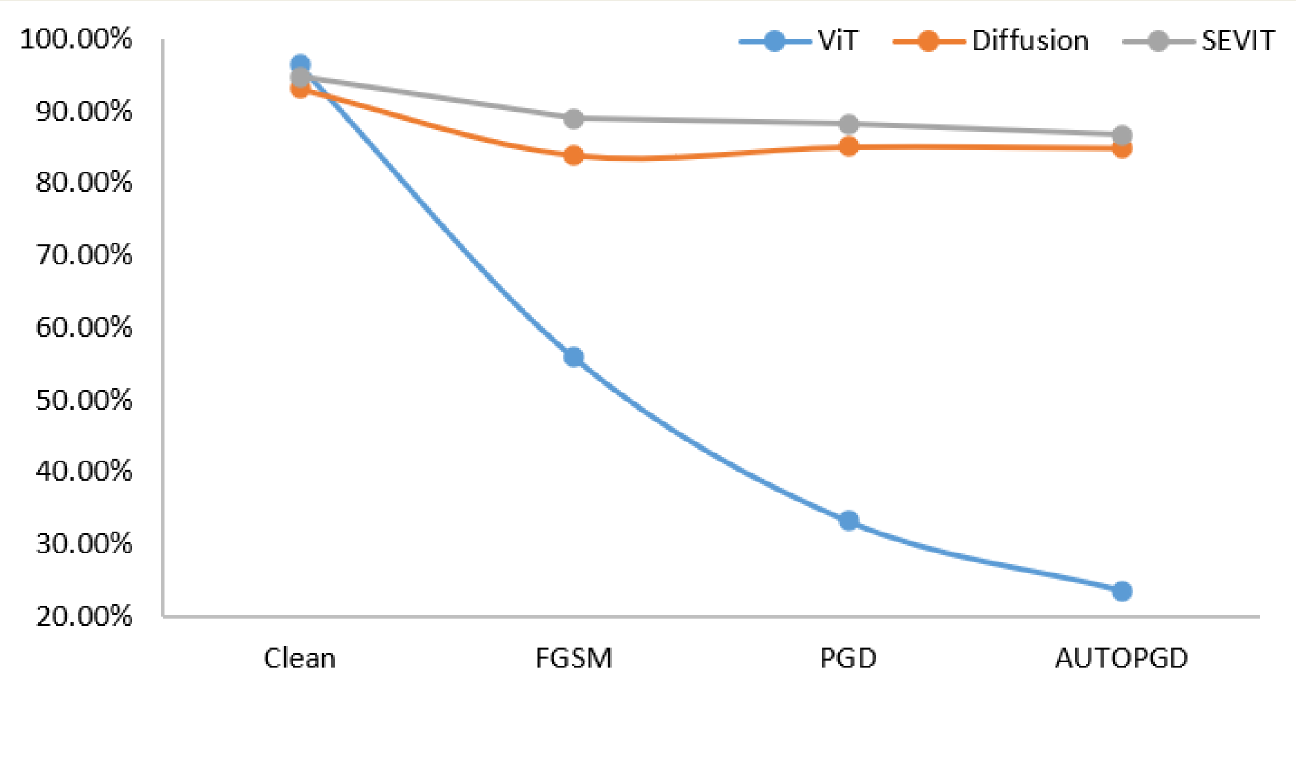
*Student CNN = knowledge transfer from ViT

Defensive Diffusion on Student *CNN*



Results

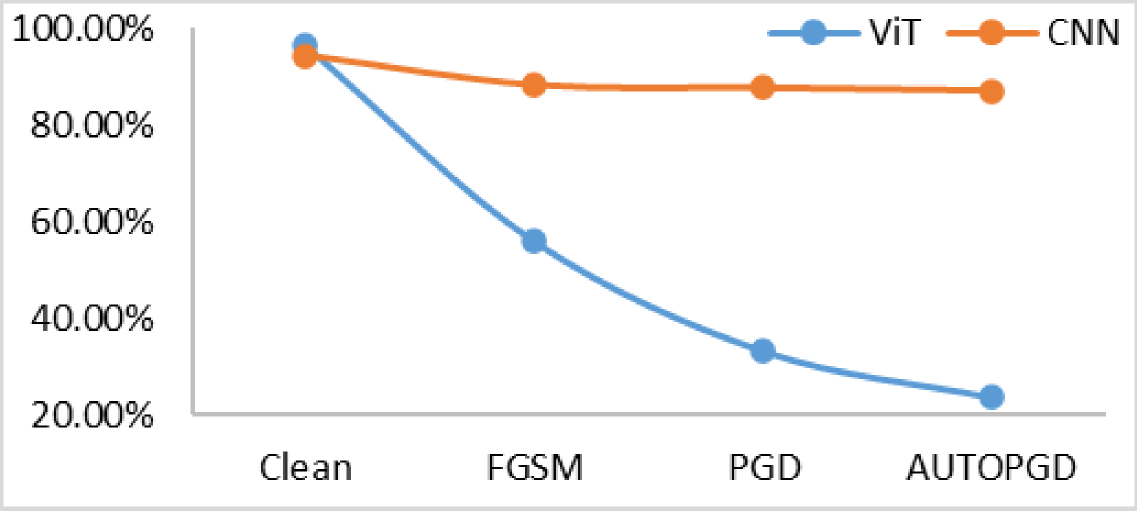
- **ViT vs Diffusion (ViT) vs SEViT**
  - On Attack samples, robust accuracy degrades drastically
  - With blurring diffusion, only a slight decrement (- **8%**) is seen in ViT
  - SEViT (- **6%**) is performing similar to our approach



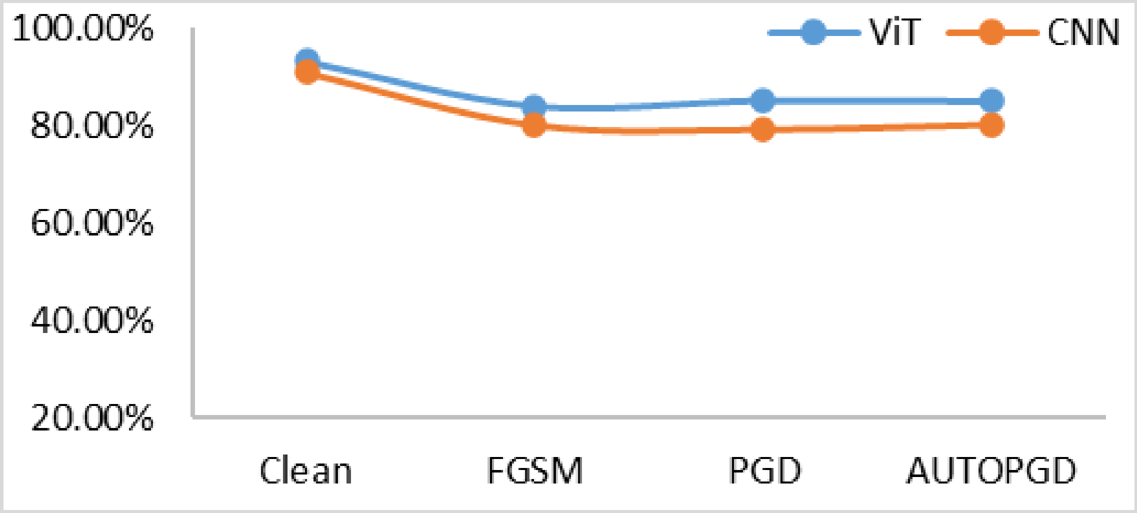
Results Cont.

- **Teacher vs Student**
  - Student CNN is more robust than Teacher ViT even without diffusion
  - Although, in the presence of diffusion, Teacher performed slightly better instead
  - Student CNN is a suitable deployment alternative as well

- **Without Diffusion**



- **With Diffusion (Blurring)**



Observations

- One key observation is that blurred images performed relatively well on both vit and SEViT.
- This encourages us to find a way to maintain the accuracies even after using deblurring diffusion model.
- Potential method could be to increase the blur time steps in the forward diffusion and classify on images that are not fully reconstructed

- The robustness of ViTs can be improved via various diffusion approaches
- Diffusion can act as an adversarial sample purifier to the original model
- Enhance the robustness and attain the same performance as SEViT with much lower computation complexity
- Moreover, knowledge distillation can be used on atop with further enhancement to the model robustness with even lightweight deployment than the original model

Conclusion

Future Work

Combination of proposed method with adversarial training at different level of perturbations.

Extending our approach to natural images will be a more generalized solution.

A major potential to this work is to explore various other popular diffusion approaches as well.

THANK YOU FOR LISTENING !