When Deep Learning Meets Steganography: Protecting Inference Privacy in the Dark

Qin Liu[a], Jiamin Yang[a], Hongbo Jiang[a], Jie Wu[b],
Tao Peng[c], Tian Wang[d], Guojun Wang[c]

[a] Hunan University

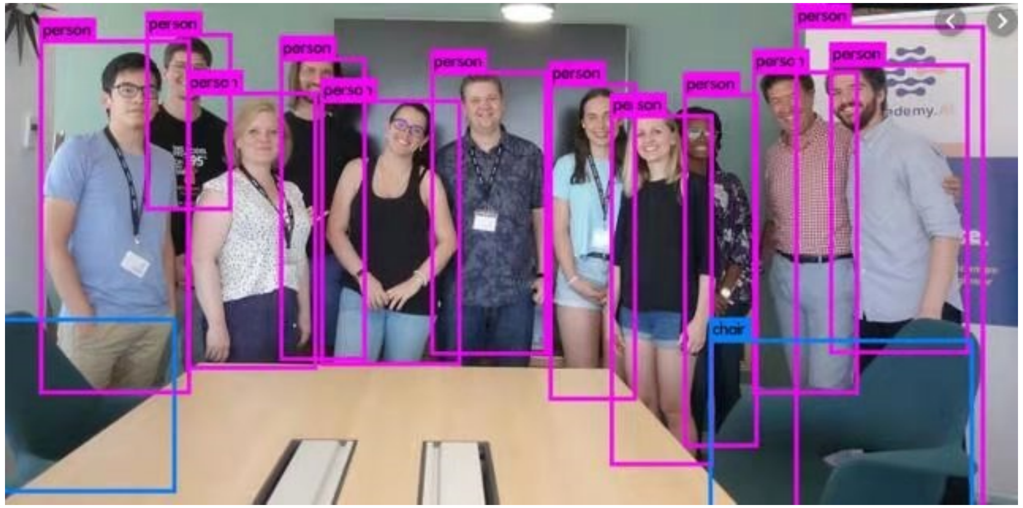
[b] Temple University

[c] Guangzhou University

[d] Beijing Normal University&UIC

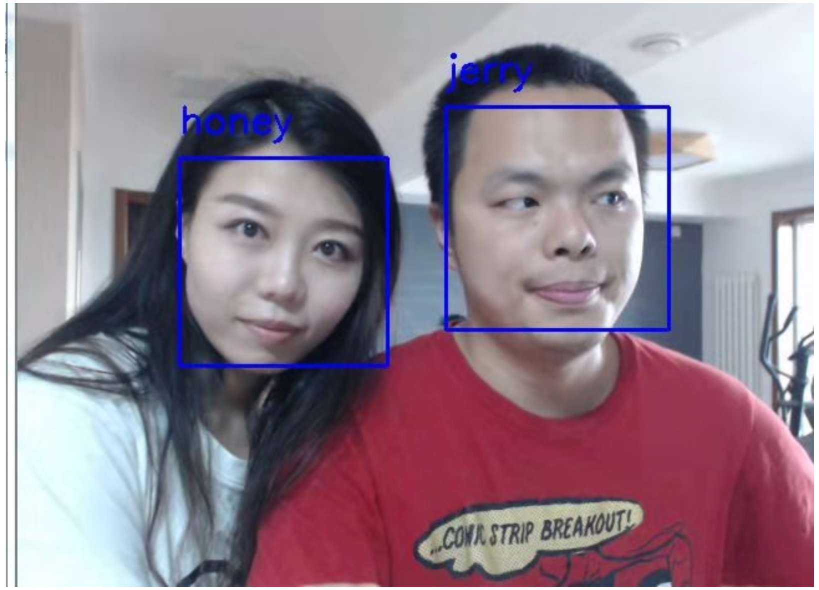
IEEE INFOCOM 2022

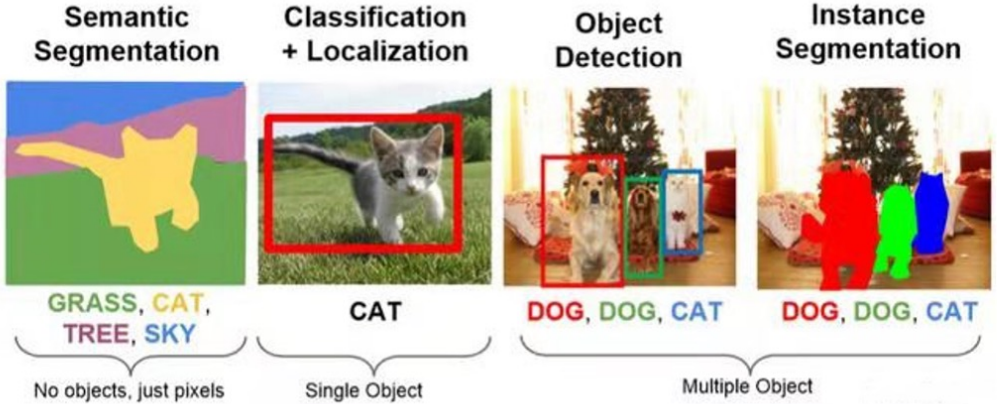
The application of DNN



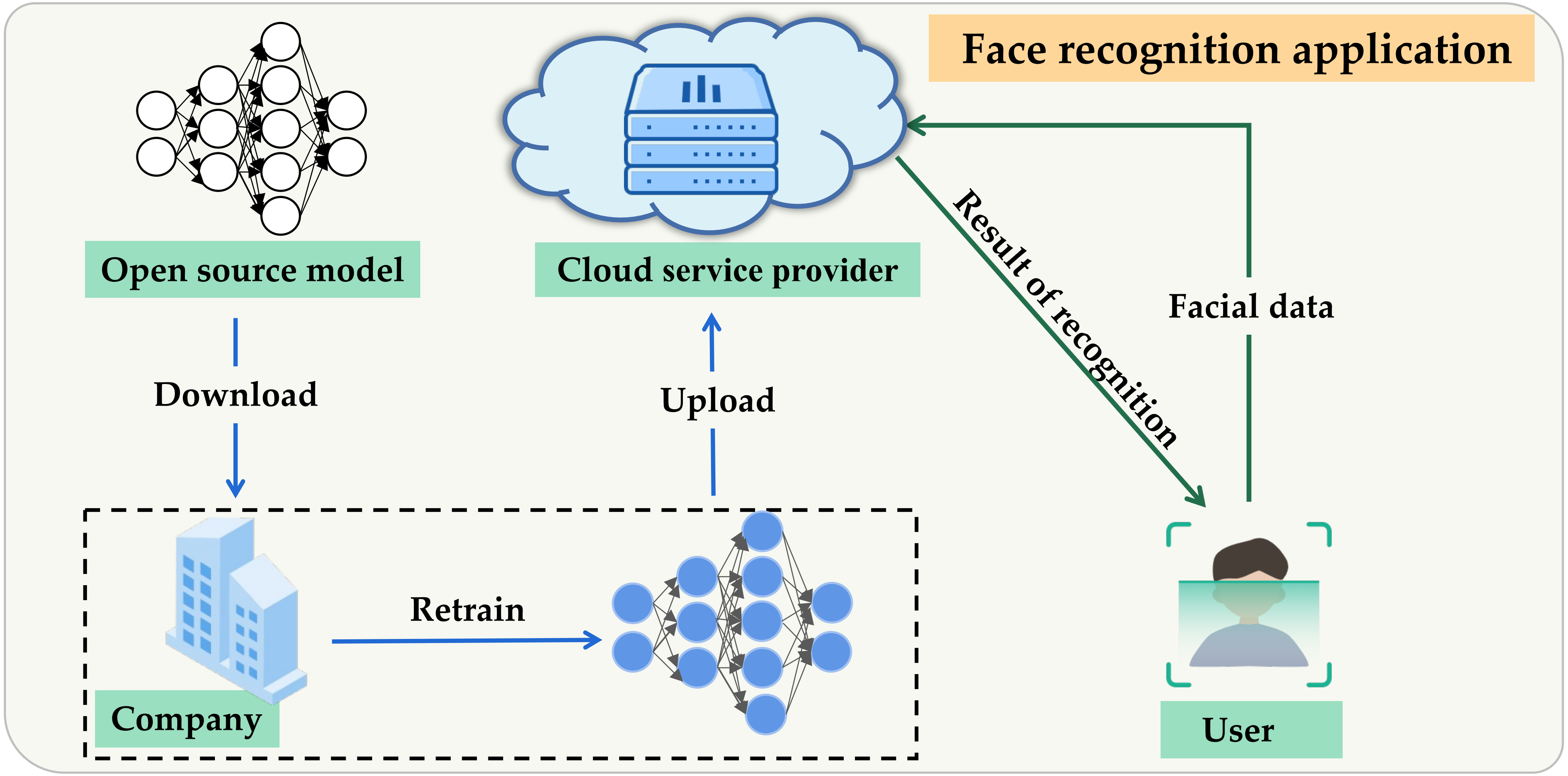
Area used:

- Computer vision
- Image processing
- Face recognition
- Natural language processing
- …



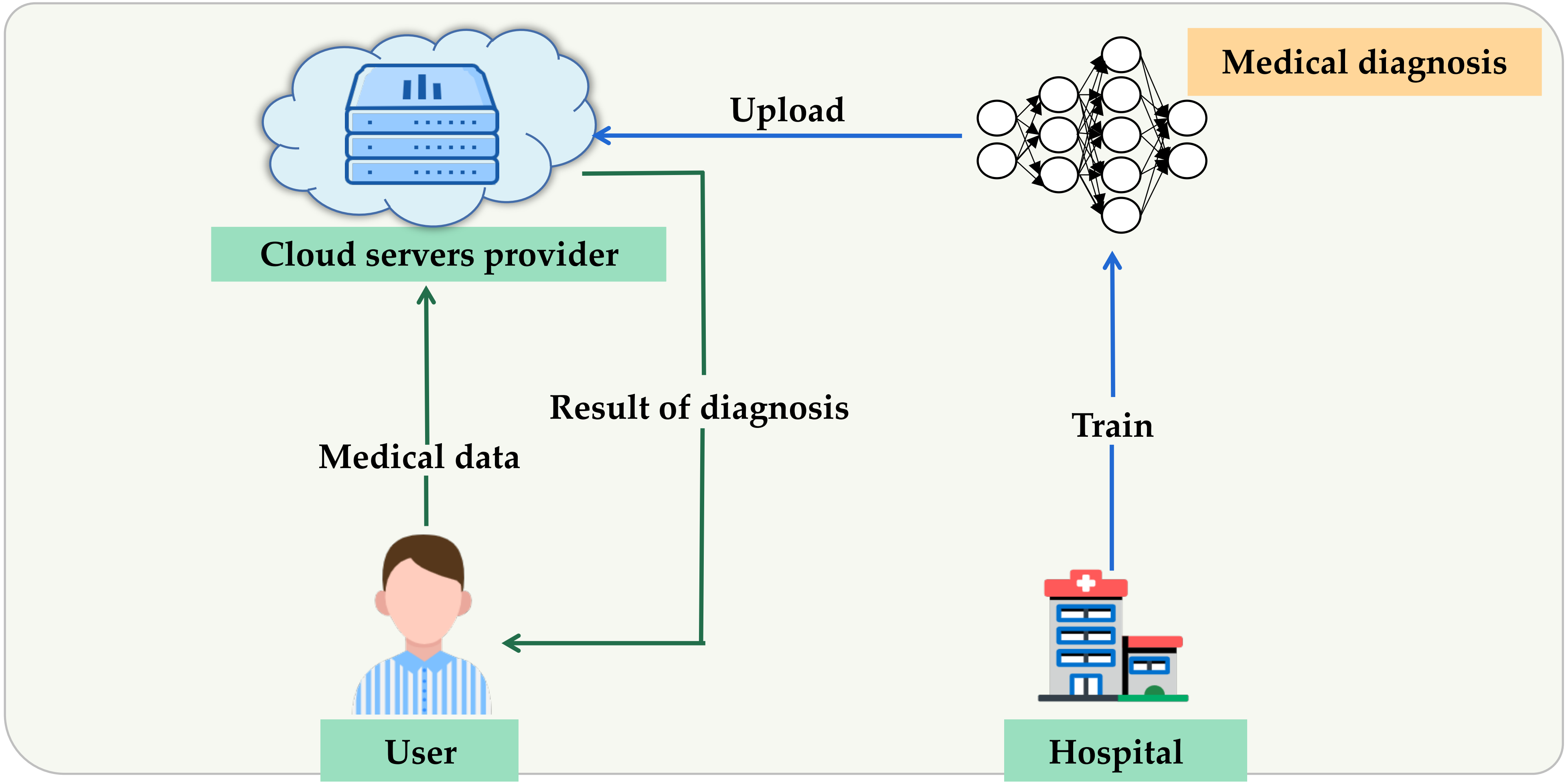


Inference privacy protection in cloud-based deep learning



- Retrain the DNN by training datasets of users.
- How to protect inference privacy by retraining the model?

Inference privacy protection in cloud-based deep learning



- User directly use the predefined DNN provided by the cloud.
- How to protect inference privacy without modifying the model?

Research status

- **Cryptographic techniques**

  MiniONN(CCS’ 17) , ABY^3(CCS’18) , Trident(NDSS’20)

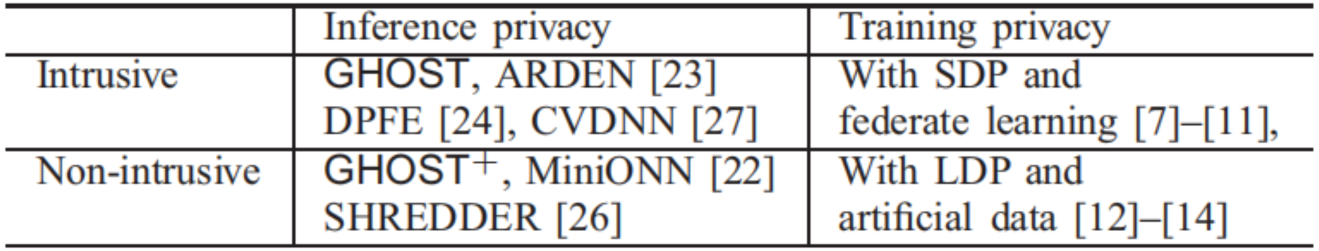
- **Trusted execution environments**

  Slalom(ICLR’18), Chiron(arXiv’18)

- **Noise injection**

  ARDEN(SIGKDD’18) , CVDNN(ICLR’20), SHREDDER(ASPLOS’20)

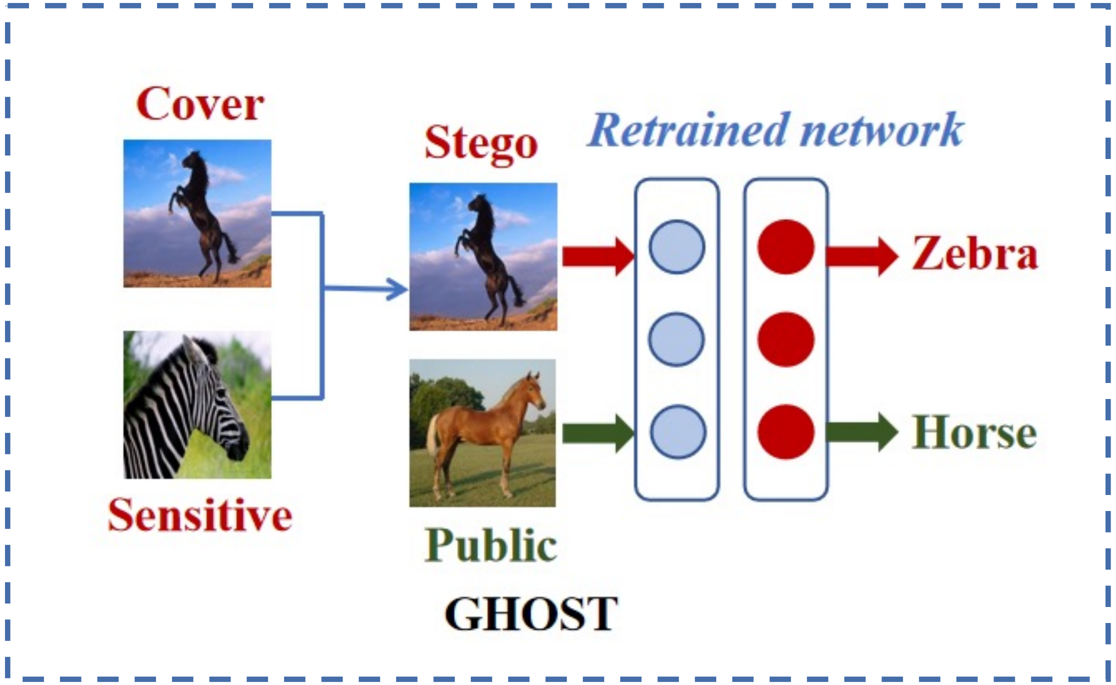
Private deep learning solutions

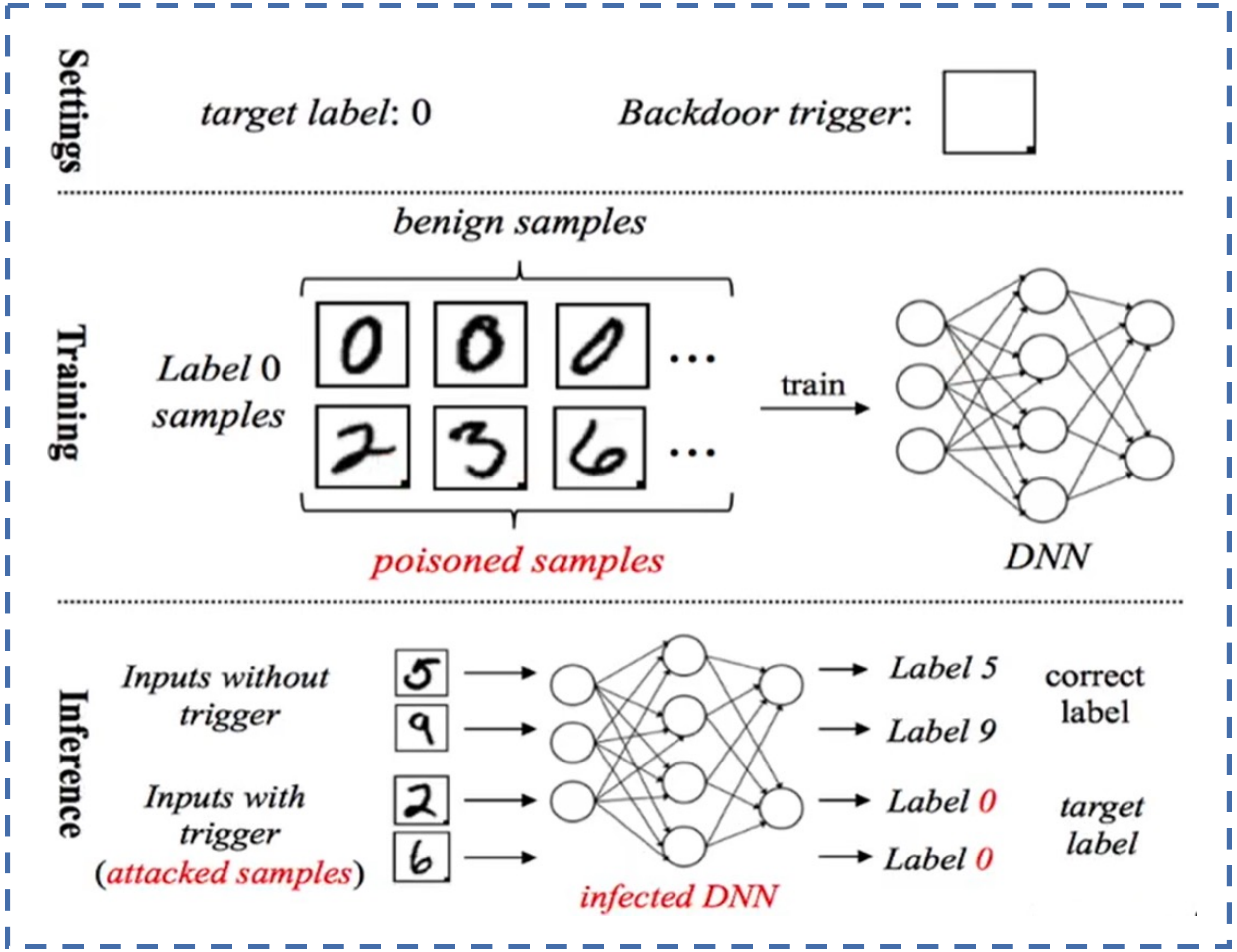


	Inference privacy	Training privacy
Intrusive	GHOST, ARDEN [23] DPFE [24], CVDNN [27]	With SDP and federate learning [7]–[11],
Non-intrusive	GHOST$^+$, MiniONN [22] SHREDDER [26]	With LDP and artificial data [12]–[14]

How to preserve inference privacy while ensuring high scalability and accuracy?

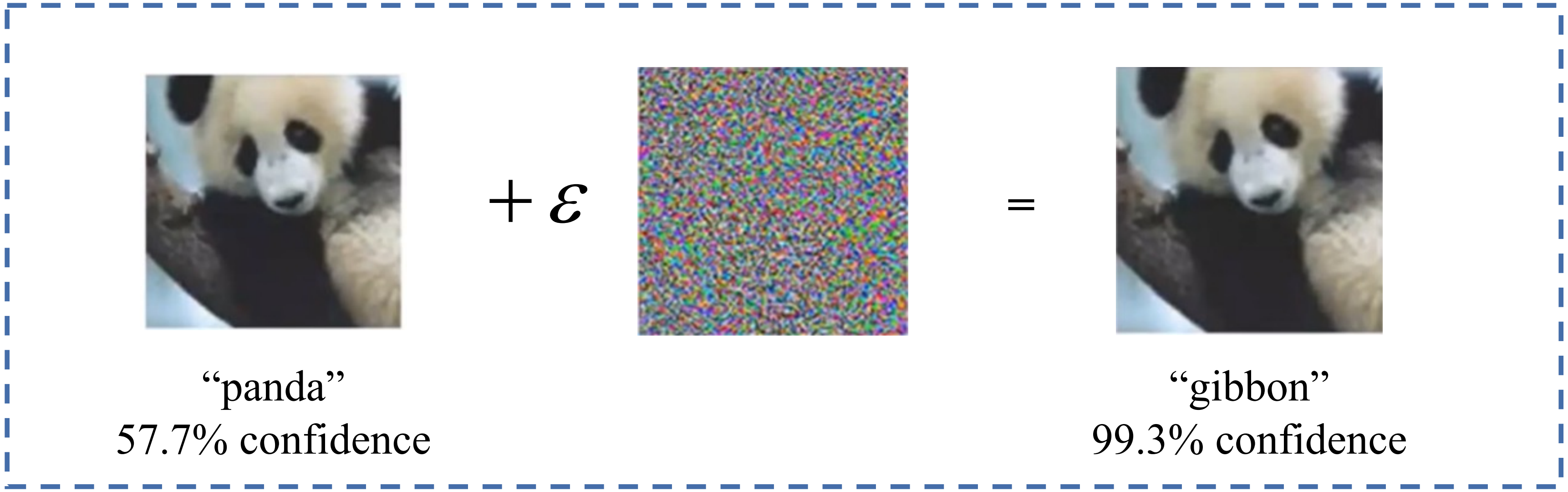
Contributions

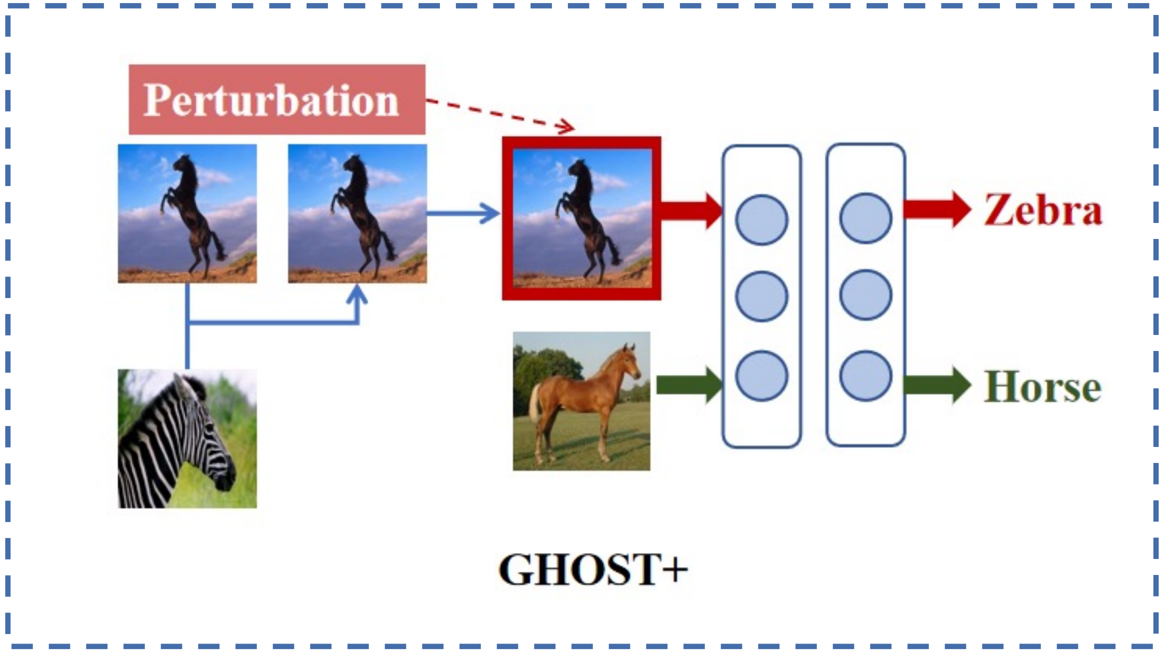




- DNNs are vulnerable to **backdoor attacks.**
- GHOST retrains the DNN into a **poisoned network** .

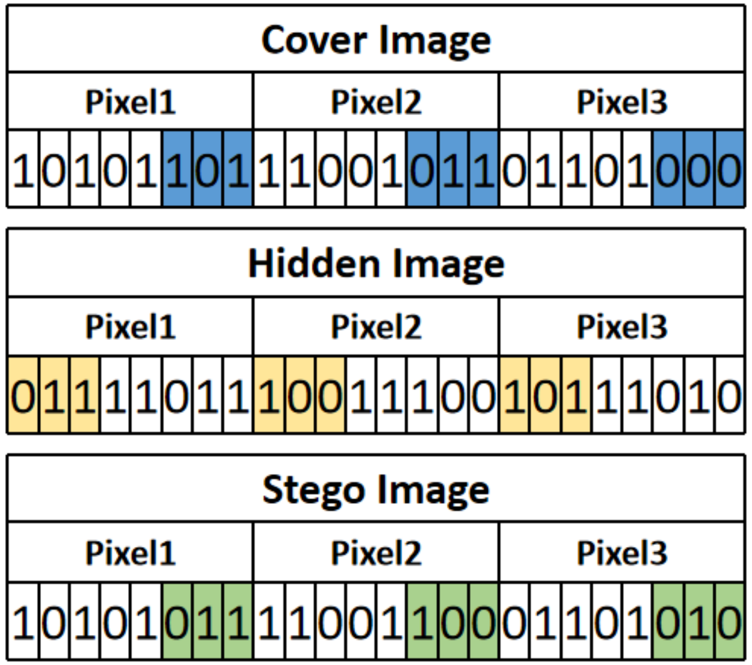
Contributions





- DNNs are vulnerable to adversarial attacks.
- GHOST+ generates adversarial perturbations by GAN.

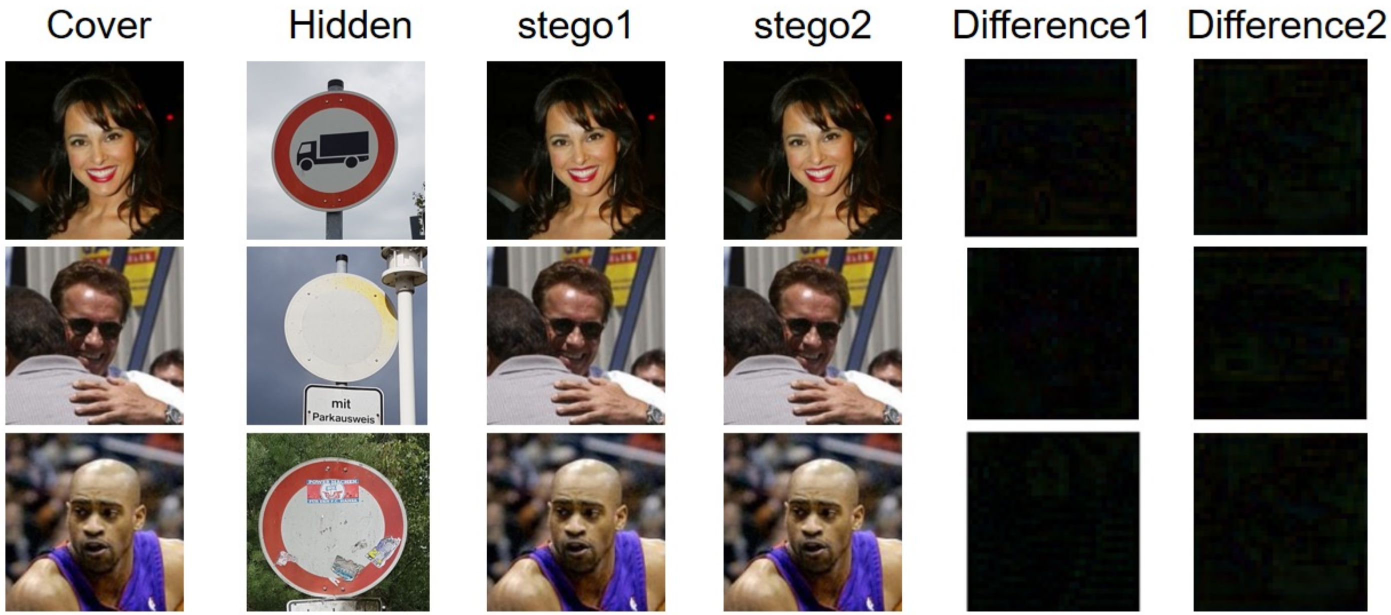
Least significant bit[1]



Cover Image		
Pixel1	Pixel2	Pixel3
10101101	11001011	01101000

Hidden Image		
Pixel1	Pixel2	Pixel3
01111011	10011100	10111010

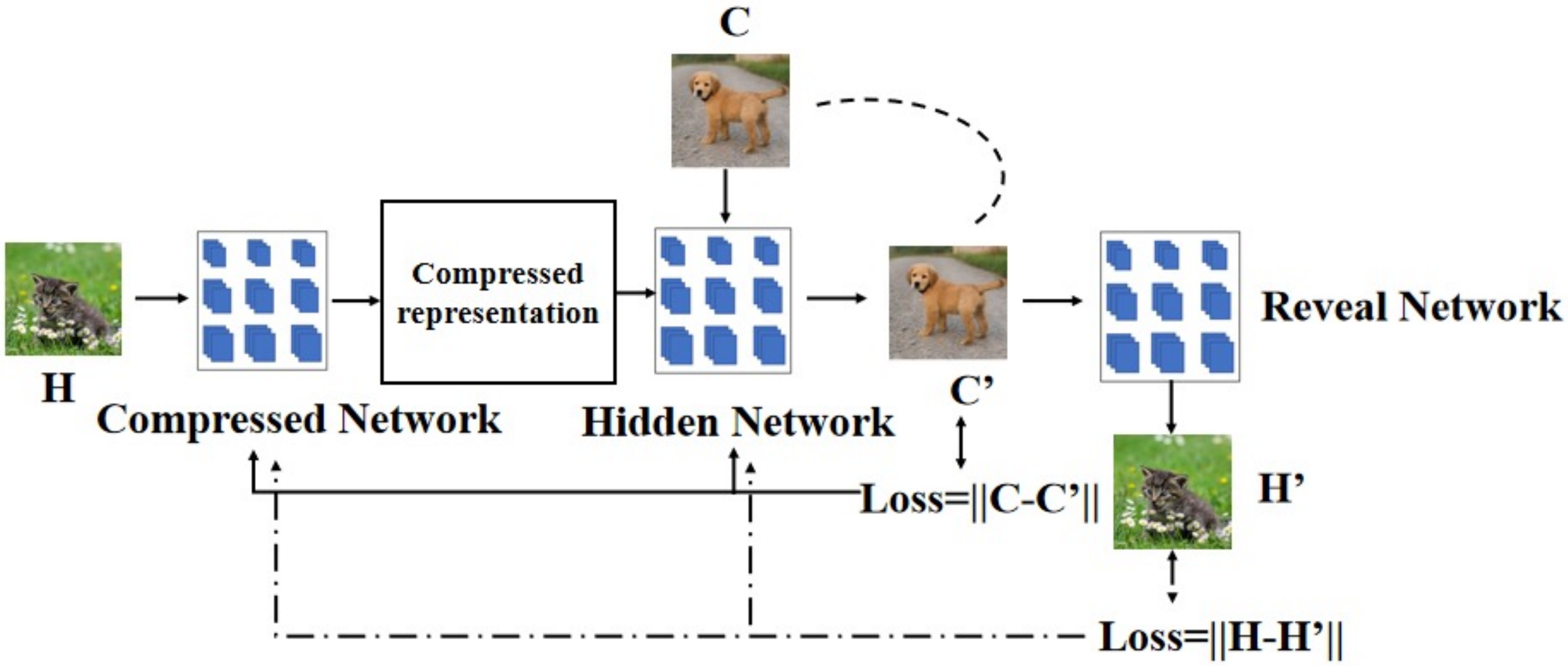
Stego Image		
Pixel1	Pixel2	Pixel3
10101011	11001100	01101010



Replace the lowest three bits of the cover image with the **highest three bits** of the hidden image.

[1] C. K. Chan and L. M. Cheng, "Hiding data in images by simple LSB substitution," Pattern Recognition, 2004.

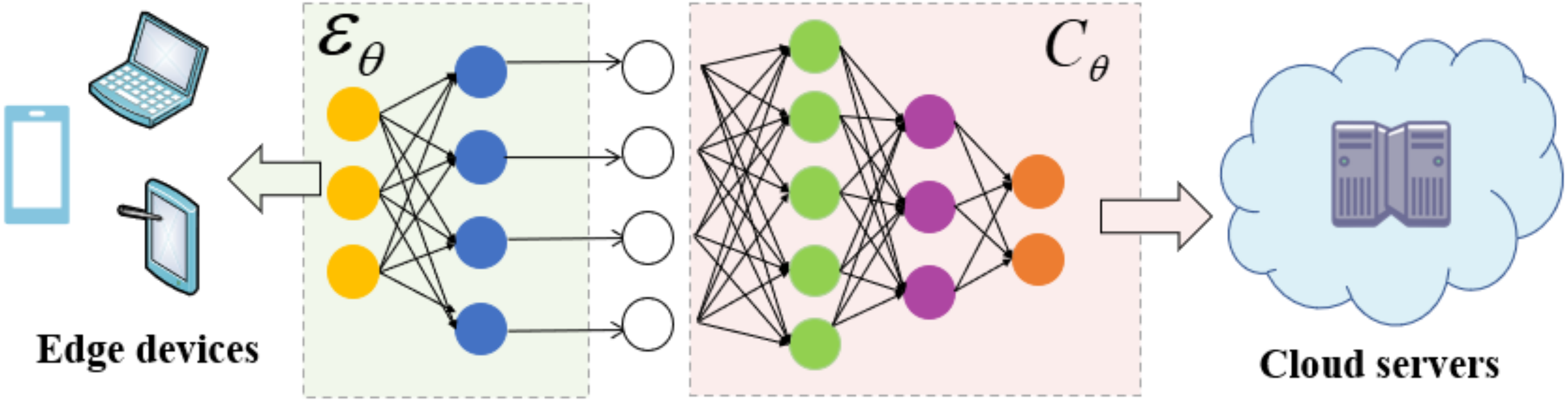
Neural network-based steganography[2]



$$L_{NNS} = E[|C - C'|] + \beta E[|H - H'|]$$

[2] S. Baluja, "Hiding images within images," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

System model

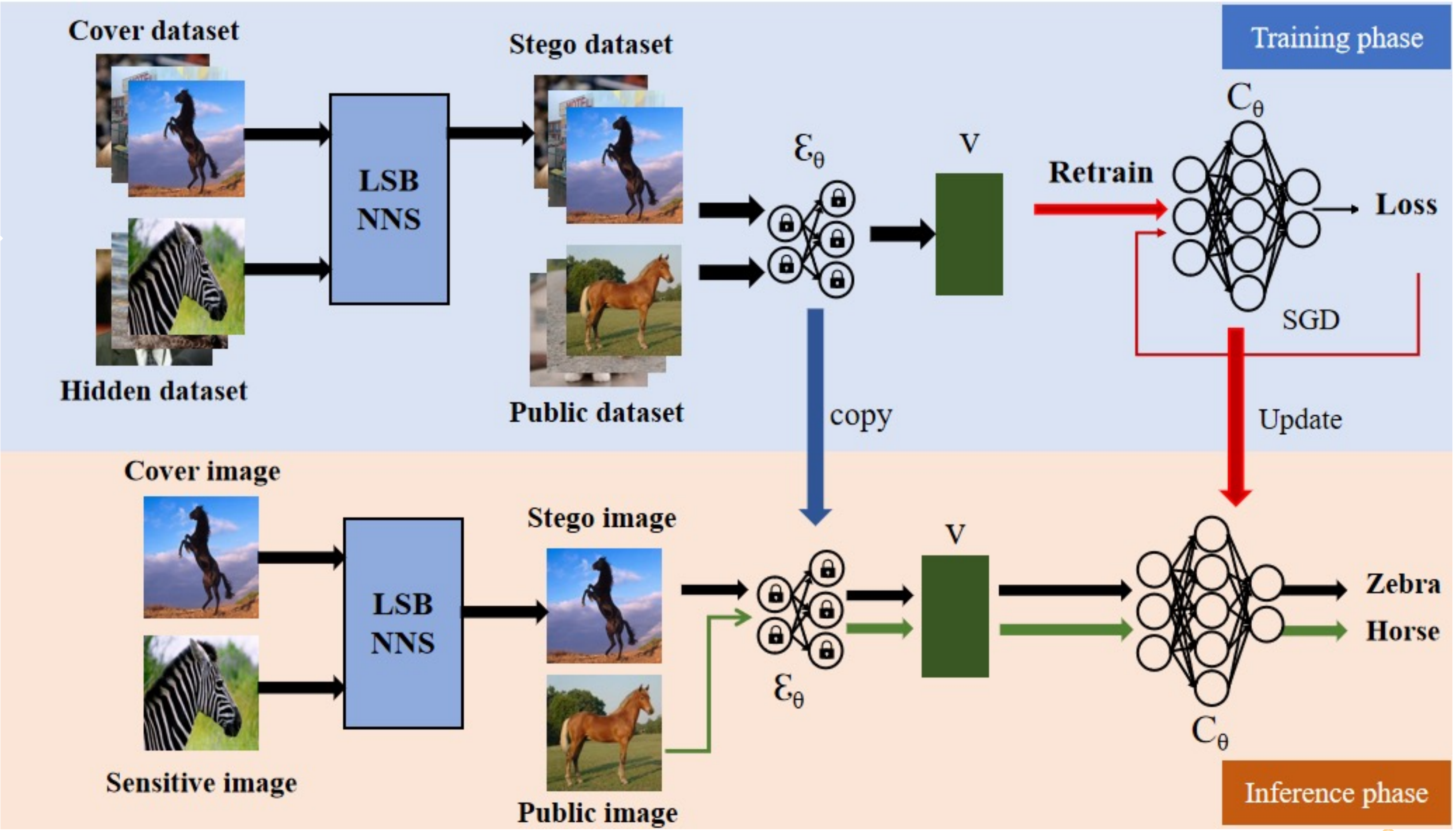


$\mathcal{E}_\theta$: Extracts features embedded and sends the intermediate value v to the Cθ.

C_θ : Calculates the final output and gets it back.

The intrusive solution GHOST

Covert Retraining



$$L(C; x^r, \tilde{c}^r) = L(C; x^r) + \lambda L(C; \tilde{c}^r)$$

Covert Inference

The non-intrusive solution GHOST^{+}

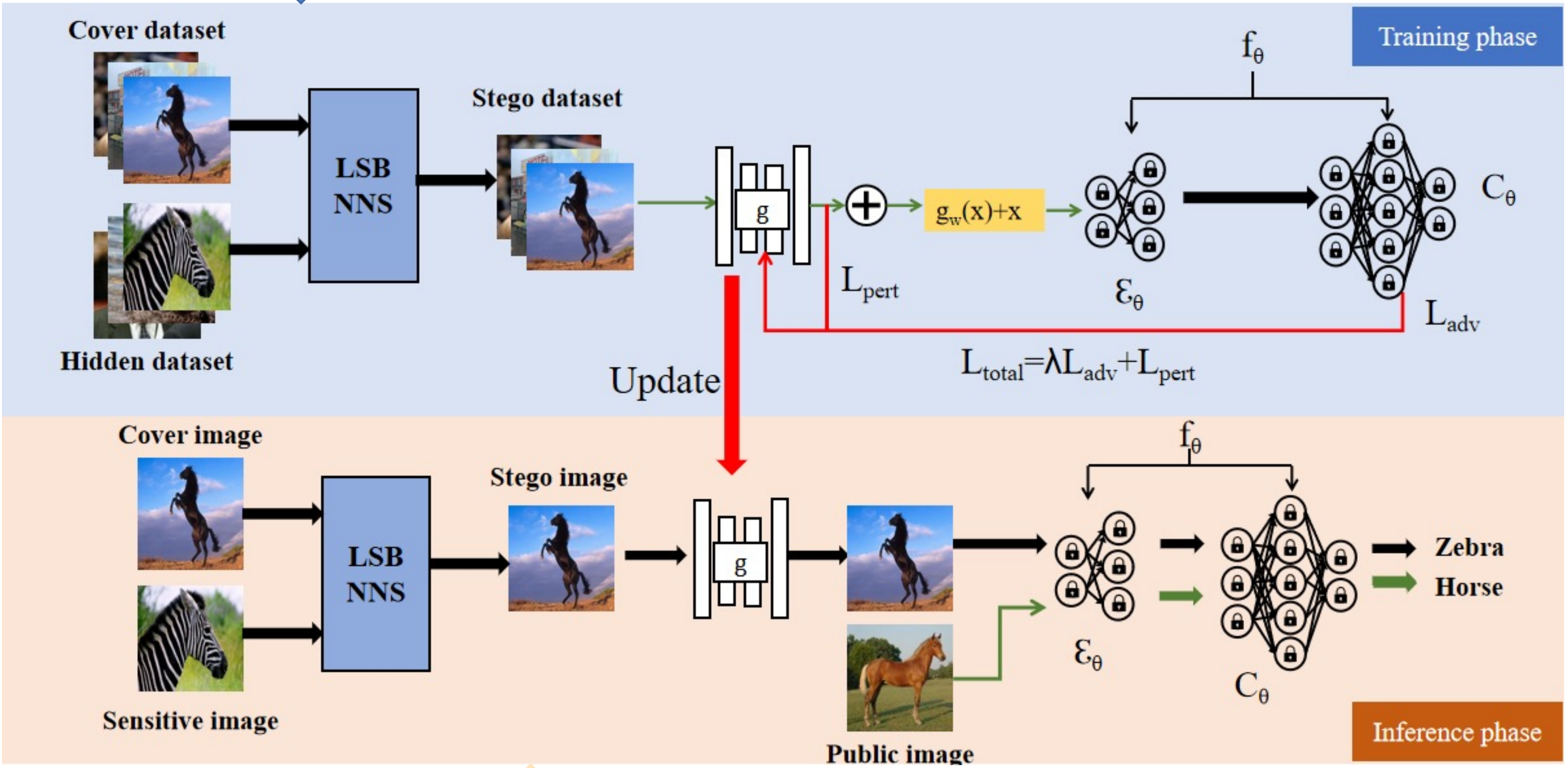
Adversarial PerturbationTraining

$$L_{adv} = E[L(f_{\theta}(g_{w}(\widetilde{C}) + \widetilde{C}), l_{H})]$$

$$L_{pert} = E[\| g_{w}(\widetilde{C}) \|_{2}]$$

$$L_{total} = \lambda L_{adv} + L_{pert}$$

$$\lambda = \left\{ \begin{array}{l} 1/2, \quad if\ L_{pert} > L_{adv} \\ 2, \quad if\ \ L_{pert} \le L_{adv} \end{array} \right\}$$



Adversarial Inference

Experiment setup

- **Edge device specifications:** NVIDIA GeForce MX250 running CUDA V10.2.141.
- **Cloud specifications:** NVIDIA GeForece RTX 3080 GPU running CUDA V11.4.56.

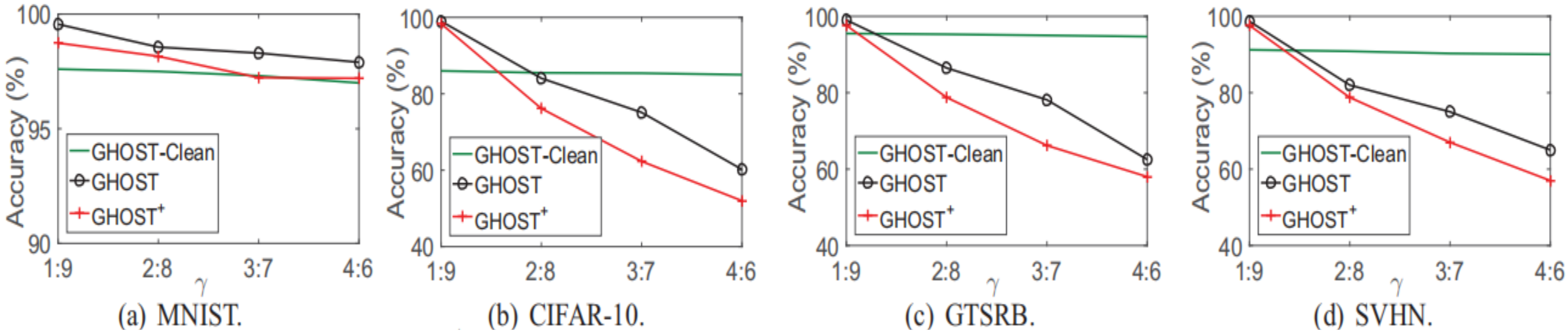
Experimental datasets and model parameters

Dataset	# of images	# of classes	Input size	Model architecture	Accuracy	$\vert\alpha\vert,\vert B\vert,\vert E\vert$
MNIST	70,000	10	28×28×1	2Conv+2Pooling+ 2Dense	98.25	[0.001,256,20]
CIFAR-10	60,000	10	32×32×3	4Conv+2Pooling+4BN+4Dropout+3Dense	87.13	[0.001,128,100]
GTSRB	51,839	43	32×32×3	6Conv+3Pooling+ 4Dropout+3Dense	96.21	[0.001,128,50]
SVHN	99,289	10	32×32×3	AlexNet[57]	91.79	[$5e^{-4}$,128,50]

Performance

- The inference accuracy of our solutions under the setting of **γ** = {1:9, 2:8, 3:7, 4:6}. **(γ=n$_s$/n$_c$)**

***n$_s$* :** *the number of sensitive labels* ***n$_c$* :** *the number of public labels*



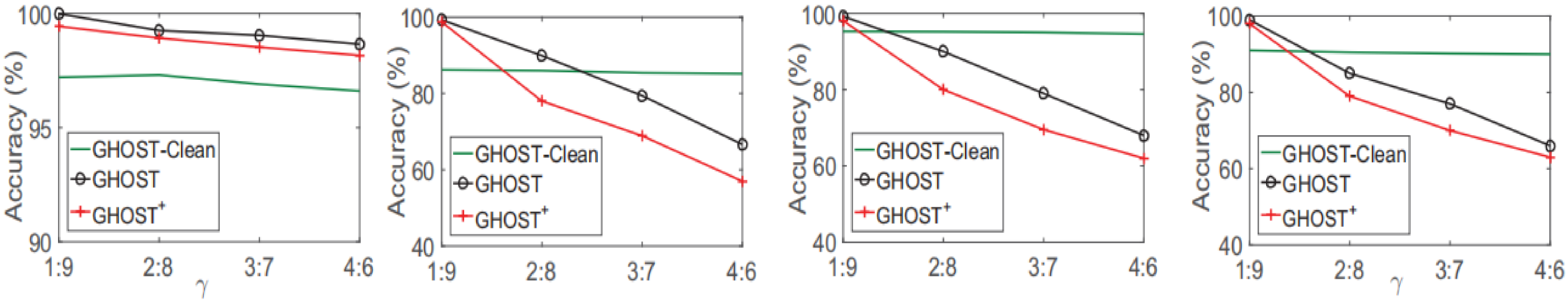
(a) MNIST. (b) CIFAR-10. (c) GTSRB. (d) SVHN.

I. The inference accuracy of sensitive samples **decreases as the ratio γ increases**.

II. Under the same settings, **GHOST performs better than GHOST+**.

Performance

- The inference accuracy of our solutions under the setting of **γ** = {1:9, 2:8, 3:7, 4:6}. **($\gamma = n_s/n_c$)**
- ***n_s* :** *the number of sensitive labels* ***n_c* :** *the number of public labels*

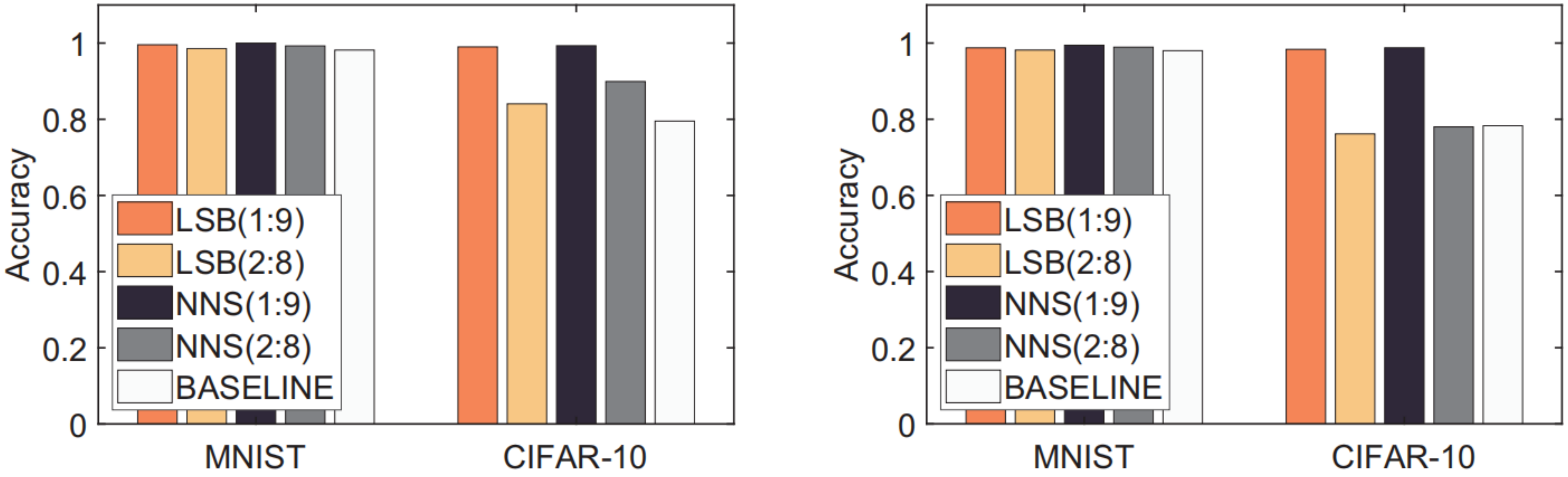


(a) MNIST. (b) CIFAR-10. (c) GTSRB. (d) SVHN.

I. The inference accuracy of sensitive samples **decreases as the ratio γ increases**.

II. NNS performs better than LSB with limited embedding capacity.

Performance comparison



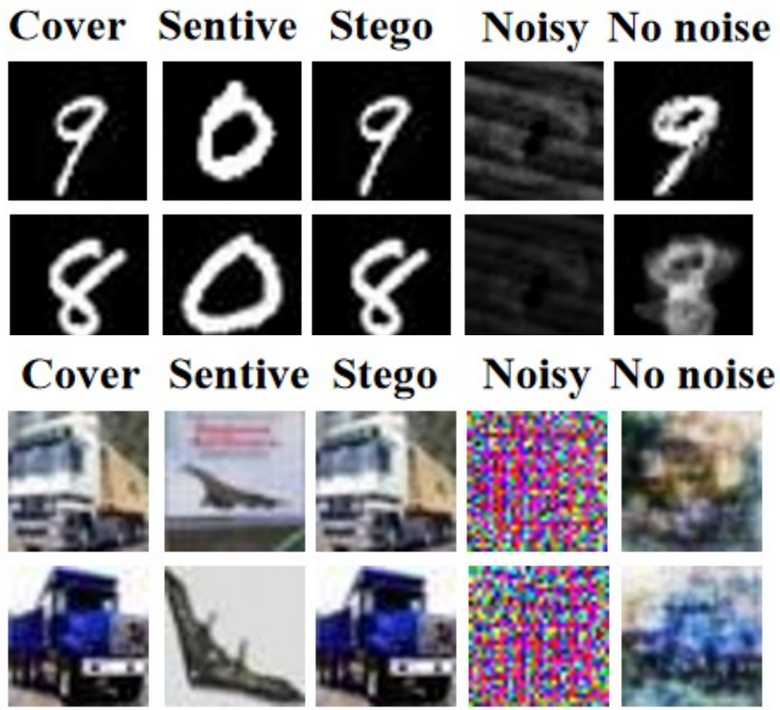
(a) GHOST vs. ARDEN. (b) GHOST$^+$ vs. SHREDDER.

[3]J. Wang, J. Zhang, W. Bao, X. Zhu, and P. S. Yu, “Not just privacy:Improving performance of private deep learning in mobile cloud,” in Proc. of SIGKDD, 2018.

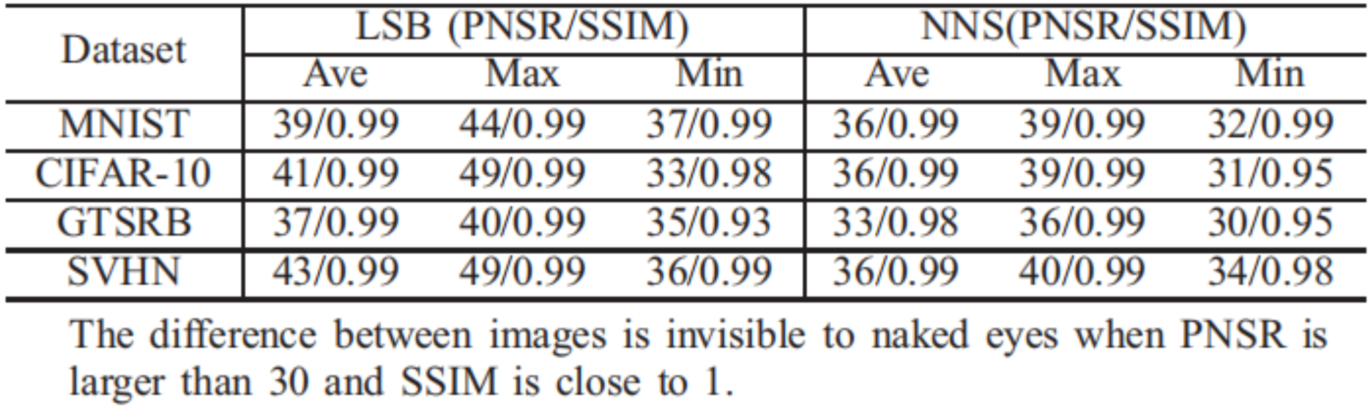
[4] F. Mireshghallah, M. Taram,P. Ramrakhyani, D. Tullsen, and H. Es_x0002_maeilzadeh, “Shredder: Learning noise distributions to protect inference privacy,” in Proc. of ASPLOS, 2020.

Privacy

- **Removing noise using the CDA.**



- **The invisibility of hidden images .**



Dataset	LSB (PNSR/SSIM)			NNS(PNSR/SSIM)		
	Ave	Max	Min	Ave	Max	Min
MNIST	39/0.99	44/0.99	37/0.99	36/0.99	39/0.99	32/0.99
CIFAR-10	41/0.99	49/0.99	33/0.98	36/0.99	39/0.99	31/0.95
GTSRB	37/0.99	40/0.99	35/0.93	33/0.98	36/0.99	30/0.95
SVHN	43/0.99	49/0.99	36/0.99	36/0.99	40/0.99	34/0.98

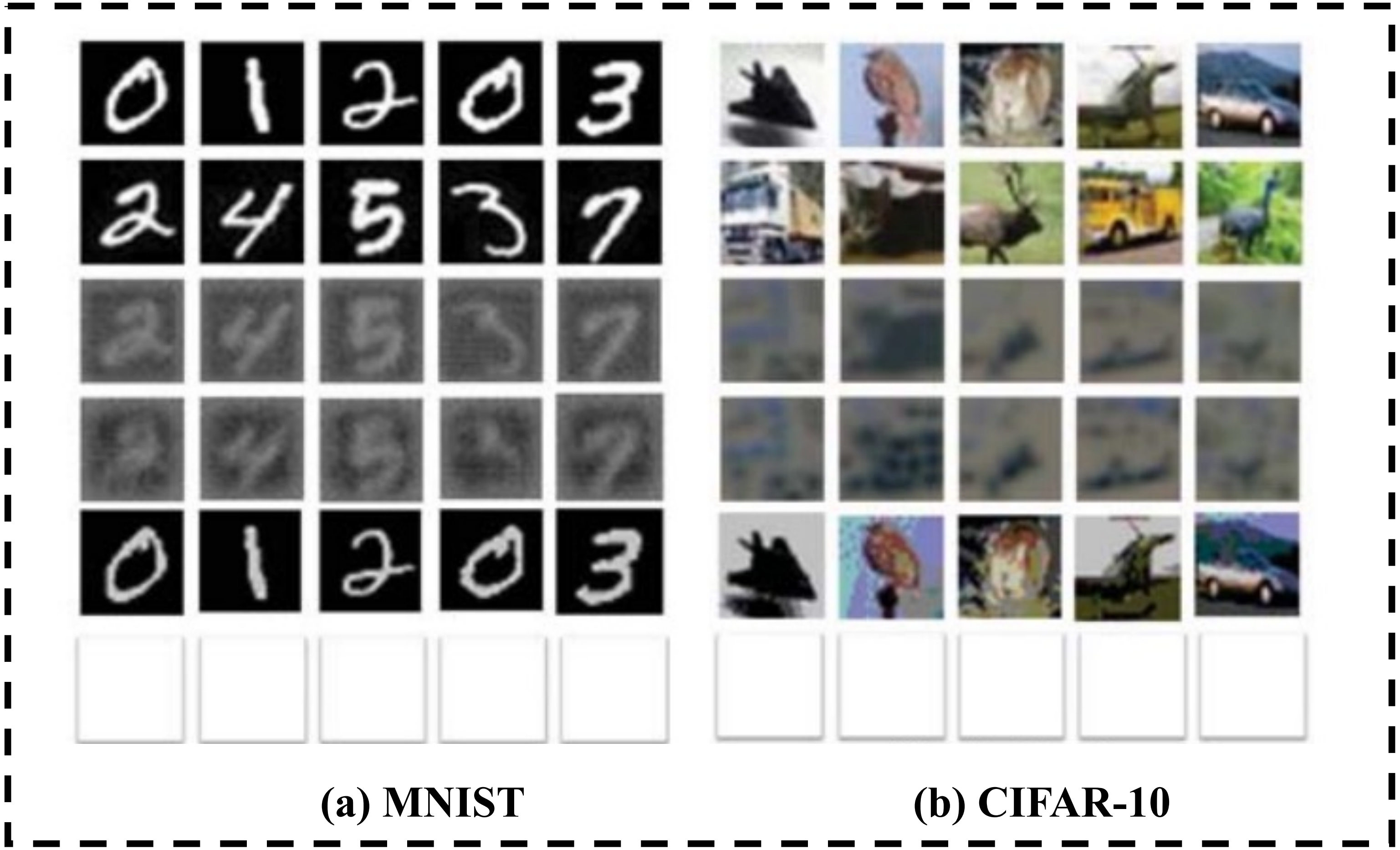
The difference between images is invisible to naked eyes when PNSR is larger than 30 and SSIM is close to 1.

The denoised image could not show the existence of sensitive image.

By calculating PSNR/SSIM, it is hard for the observer to detect the difference.

Privacy

- **Feature inversion attacks**



(a) MNIST **(b) CIFAR-10**

It is hard to extract useful information of sensitives images.

Conclusion

- This is the **first work** that successfully utilizes **image steganography** and **adversarial attacks** to protect inference privacy in the dark.
- We propose two private inference solutions, **GHOST** and **GHOST^{+}**, both of which employ the traditional **LSB** and recent **NNS** techniques to hide sensitive images.
- Experimental results show that our solutions **outperform** the state-of-the-art solutions when the number of sensitive types is **within a given range**.

Future work

- For GHOST, it requires a **larger-scale neural network** of better learning and discrimination power.
- For GHOST+, it requires **training a stronger generator** to generate adversarial perturbations for a variety of sensitive types.
- We try to implement our solutions **in other domains**, such as voice and text.

Thanks for your attention

Questions & Answers?

Qin Liu

Hunan University

gracelq628@126.com