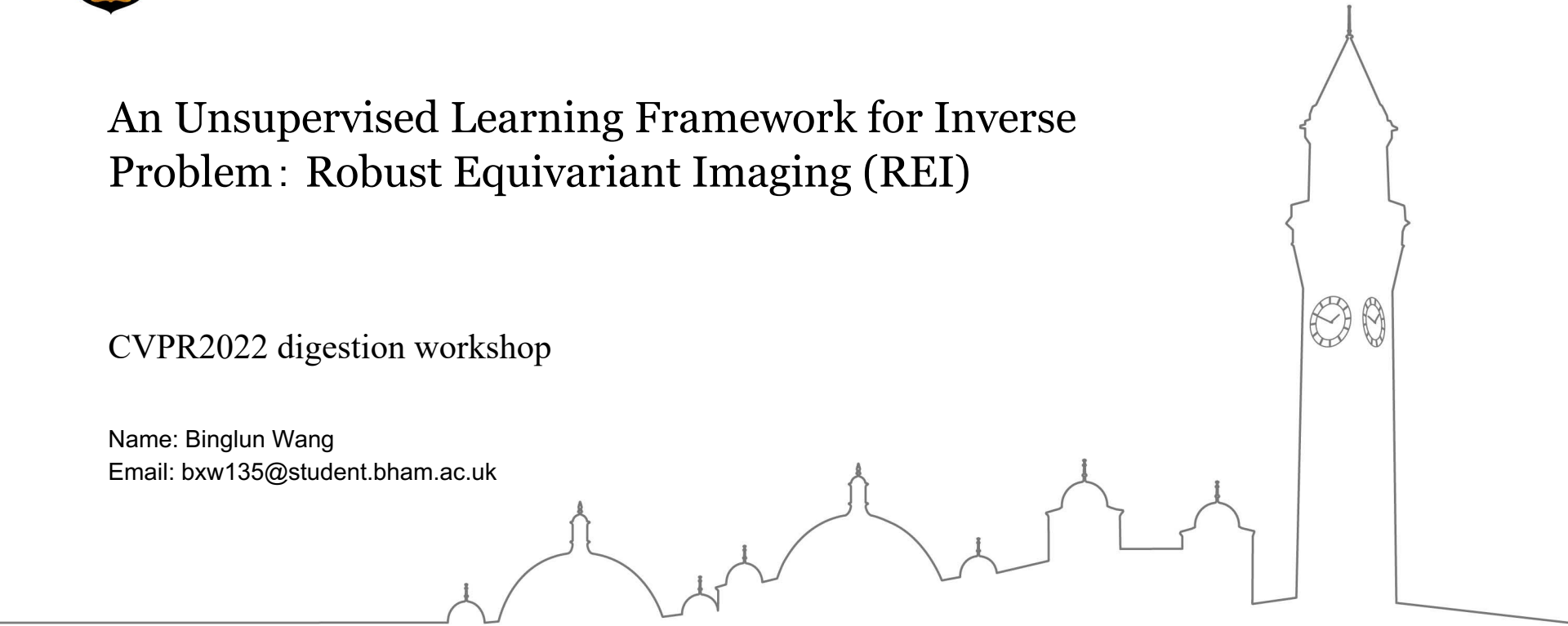
UNIVERSITY OF BIRMINGHAM

An Unsupervised Learning Framework for Inverse Problem: Robust Equivariant Imaging (REI)

CVPR2022 digestion workshop

Name: Binglun Wang
Email: bxw135@student.bham.ac.uk



Inverse Problem

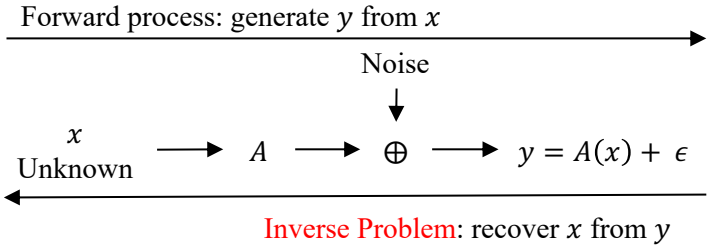


Fig 1. Definition of Inverse Problem [1, 2]

Example of Inverse Problem

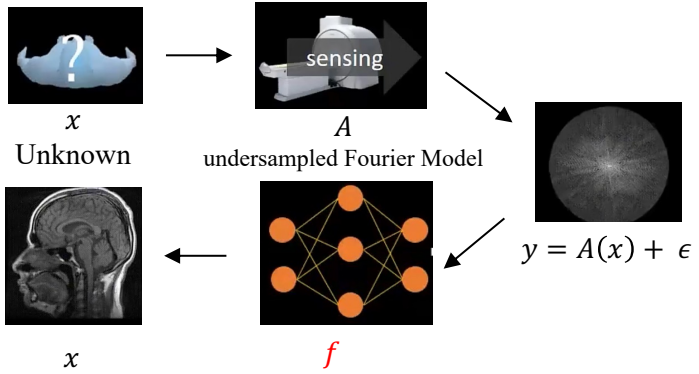


Fig 2. Example of MRI [3]

Supervised Learning

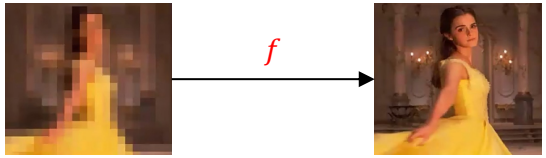


Fig 3. Example of super-resolution [3]

Highlight

SOTA performance
Based on {y, x} pairs

However, in lots of tasks.
Obtaining targets {x} is
expensive or impossible

Unsupervised Learning

In some unsupervised methods [4, 5, 6], the setting is not often met in practice. However, Equivariant Imaging (EI):

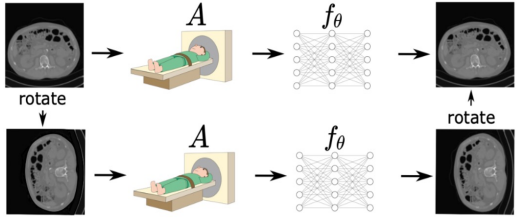


Fig 4. Idea of Equivariant Imaging [1, 2]

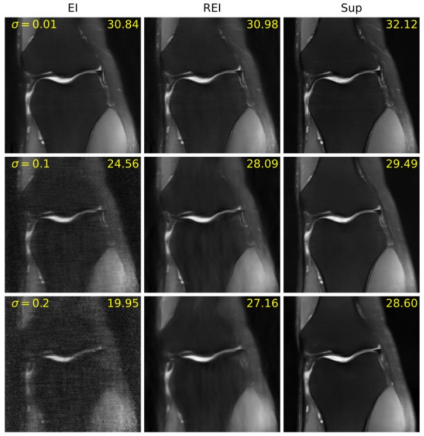
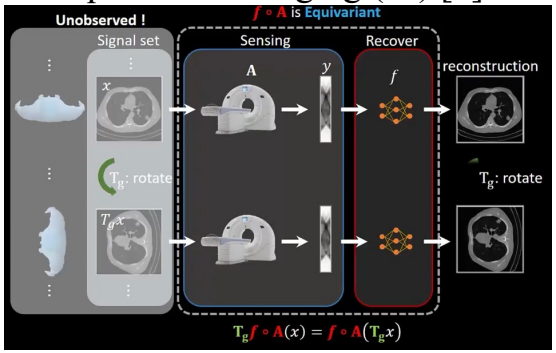


Fig 5. EI problem and REI performance [1]

[1] "Robust Equivariant Imaging: a fully unsupervised framework for learning to image from noisy and partial measurements.", [Chen et al., CVPR2022 Oral]
[2] "Equivariant imaging: Learning beyond the range space.", [Chen et al., ICCV2021]
[3] "Robust Equivariant Imaging: A Fully Unsupervised Framework for Learning To Image From" [Wang, YouTube]
[4] "AmbientGAN: Generative models from lossy measurements." [Bora et al., ICLR2018]
[5] "Deep image reconstruction using unregistered measurements without groundtruth" [Gan et al., ISBI 2021]
[6] "Unsupervised adversarial image reconstruction." [Pajot et al., ICLR2021]

- Equivariant Imaging (EI) [1]



$$\mathcal{L}_{\text{EI}}(\theta) = \mathcal{L}_{\text{MC}}(\theta) + \alpha \mathcal{L}_{\text{EQ}}(\theta) \qquad y \subseteq R^m, x \subseteq R^n$$

With: $$\mathcal{L}_{\text{MC}}(\theta) = \sum_{i=1}^{N} \frac{1}{m} \|y_i - A\left(f_\theta(y_i)\right)\|^2,$$

$$\mathcal{L}_{\text{EQ}}(\theta) = \sum_{i=1}^{N} \sum_{g=1}^{|\mathcal{G}|} \frac{1}{|\mathcal{G}|n} \|T_g f_\theta(y_i) - f_\theta\left(A(T_g f_\theta(y_i))\right)\|^2$$

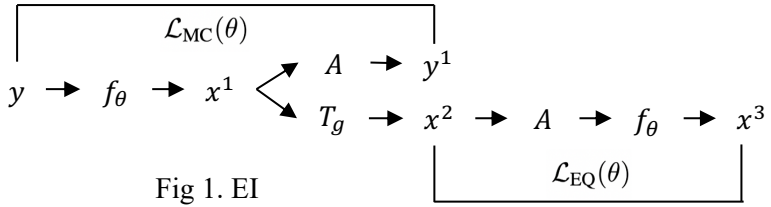


Fig 1. EI

- Self-supervised denoising via Stein's Unbiased Risk Estimator (SURE) [2]

If measurements contain Gaussian noise:

$y|u \sim \mathcal{N}(u, I\sigma^2)$ h_θ that maps $y \mapsto u$

$$\mathcal{L}_{\text{SURE}}(\theta) = \sum_{i=1}^{N} \frac{1}{m} \|y_i - h_\theta(y_i)\|^2 - \sigma^2 + \frac{2\sigma^2}{m} \nabla \cdot h_\theta(y_i)$$

Theorem1: SURE loss is an unbiased estimator of the supervised mean squared loss

$$\mathbb{E}_y\{\mathcal{L}_{SURE}(\theta)\} = \mathbb{E}_{y,u}\{\mathcal{L}_{MSE}(\theta)\}$$

Theorem2 [3]: The divergence term can be approximated by sampling a single i.i.d.

$$\nabla \cdot h_\theta(y) \approx \frac{1}{\tau} b^\top \left(h_\theta(y + \tau b) - h_\theta(y)\right)$$

Where τ is small fixed number, $b \sim \mathcal{N}(0, 1)$

The SURE framework [4, 5, 6] beyond simple additive Gaussian noise, including Poisson and mixed Poisson Gaussian noise, resulting in different loss expressions depending on the noise type.

[1] "Equivariant imaging: Learning beyond the range space.", [Chen et al., ICCV2021]
[2] "Estimation of the mean of a multivariate normal distribution." [Stein, The annals of Statistics1981]
[3] "MonteCarlo SURE: A black-box optimization of regularization parameters for general denoising algorithms." [Ramani et al., ToIP2008]
[4] "Generalized SURE for exponential families: Applications to regularization." [Eldar et al., ToSP2008]
[5] "Image denoising in mixed Poisson-Gaussian noise." [Luisier et al., ToIP2011]
[6] "Least squares estimation without priors or supervision." [Raphan et al., Neural computation2011]

- REI

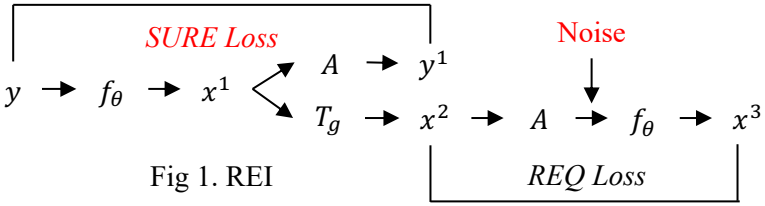


Fig 1. REI

$$\mathcal{L}_{\text{REI}}(\theta) = \mathcal{L}_{\text{SURE}}(\theta) + \alpha \mathcal{L}_{\text{REQ}}(\theta)$$

REQ Loss:

$$\mathcal{L}_{\text{REQ}}(\theta) = \sum_{i=1}^{N} \sum_{g=1}^{|\mathcal{G}|} \frac{1}{|\mathcal{G}|n} \|T_g f_\theta(y_i) - f_\theta(\tilde{y}_i))\|^2$$

SURE Loss:

For Gaussian noise:

$$\mathcal{L}_{\text{SURE}}(\theta) = \sum_{i=1}^{N} \frac{1}{m} \|y_i - A(f_\theta(y_i))\|_2^2 - \sigma^2$$

$$+ \frac{2\sigma^2}{m\tau} b_i^\top \left(A(f_\theta(y_i + \tau b_i)) - A(f_\theta(y_i))\right)$$

where $b_i \sim \mathcal{N}(0, I)$ and τ is a small positive number.

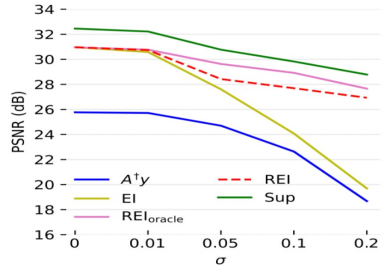
For Poisson noise & Mixed Gaussian-Poisson (MGP): See Paper

- Experiments

Setup: f_θ are residual U-Nets in all networks different loss in EI, REI, Sup.

In Accelerated MRI experiment:

1. A† is masked inverse Fourier transform.
2. Gaussian Noise based SURE loss
3. T_G are set of rotations. $|G|$=360 (degree)



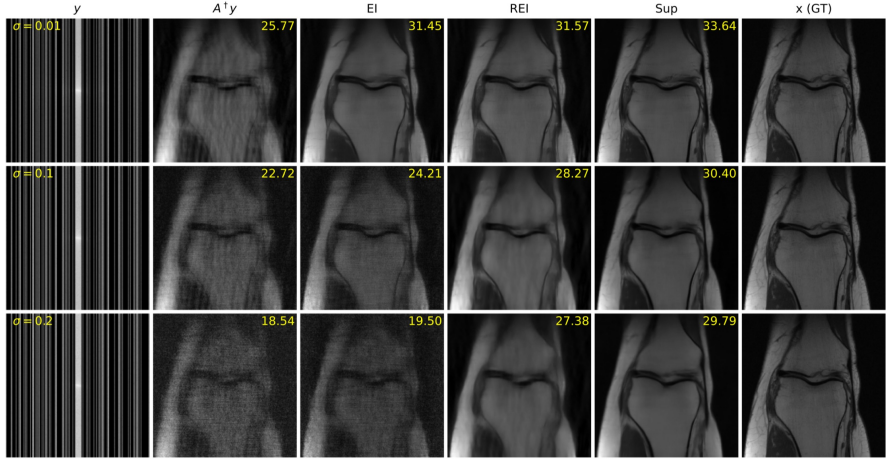
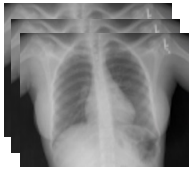


Fig 2. Reconstruction performance (PSNR) of different methods

Inpainting (Poisson noise) & Sparse-view CT (MGP) experiments: See Paper

[1] "Robust Equivariant Imaging: a fully unsupervised framework for learning to image from noisy and partial measurements.", [Chen et al., CVPR2022 Oral]
[2] "Equivariant imaging: Learning beyond the range space.", [Chen et al., ICCV2021]

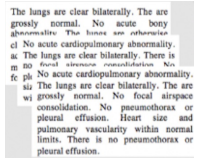
- Problem



Chest X-ray (CXR)

Radiologists

Heavy Workload
Timely Reporting



Report

Fig 1. Radiologists' work [1]

- Anomaly Detection (AD)

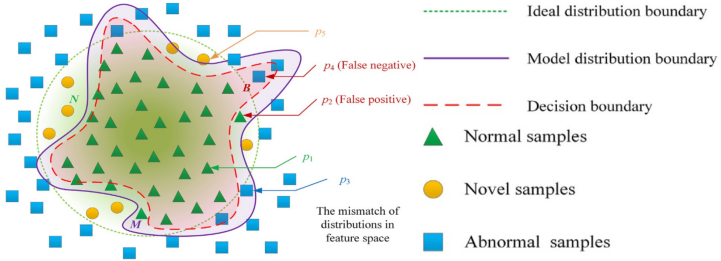


Fig 2. Definition of AD [2]

- Dataset

NIH Clinical Center Chest X-ray [3];
Montgomery datasets;
Shenzhen datasets;
etc.

- Evaluation Metrics

Receiver Operating Characteristics (ROC) curve

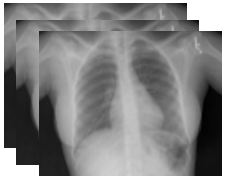
$$TPR = \frac{Number\ of\ correctly\ classified\ positive\ samples}{Number\ of\ positive\ samples}$$

$$FPR = \frac{Number\ of\ misclassified\ negative\ samples}{Number\ of\ negative\ samples}$$

AUC-ROC, Sensitivity, Specificity etc. [4][5]

My Recent work

- Improve work efficiency

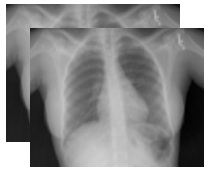


1000 CXRs

Classifier

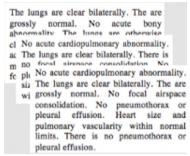
100 Certainly Normal CXRs

Report: No founding



900 CXRs

Radiologists



Report

Auxiliary Diagnosis:
Segmentation
Anomaly Score

Fig 3. Example of Auxiliary Diagnosis [3]

- My Project:

Investigate and summarise several significant modern anomaly detection methods based on GANs in CXR

Create a GUI system to visually compare results from different networks and classify CXR in real time by trained models.

[1] "A Self-boosting Framework for Automated Radiographic Report Generation.", [Wang et al., CVPR2021]
[2] "GAN-based anomaly detection: A review.", [Xia et al., Neurocomputing2022]
[3] "Chestx-ray8: Hospital-scale chest x-ray database and…" [Wang et al., CVPR2017]
[4] "One-class classification: A survey." [Perera et al., arXiv2021]
[5] "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks." [Schlegl et al., MIA2019]

- CXR AD based on GANs

Deep Learning
Good performance
Data hungry [1]

CXR Dataset [2]:
Easily Accessible
Already have open dataset

Supervised Learning → Unsupervised Learning → GANs Method

limited due to:
Abnormal samples
Unpredictability [3]

GAN can make abnormal inferences using adversarial learning of the representation of samples [3]

- GAN [6]:

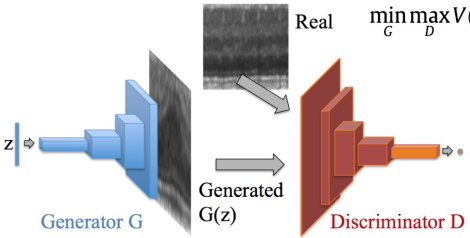


Fig 1. DCGAN Structure [4] [5]

$$\min_G \max_D V(D,G) = \mathbf{E}_{\mathbf{x}\sim P_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbf{E}_{\mathbf{z}\sim P_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

For G fixed, the optimal D:

$$D_G^*(\boldsymbol{x}) = \frac{p_{data}(\boldsymbol{x})}{p_{data}(\boldsymbol{x}) + p_g(\boldsymbol{x})}$$

$$C(G) = \max_D V(G,D) = -\log(4) + 2 \cdot JSD\left(p_{\text{data}} \| p_g\right)$$

$p_g = p_{data}$

- AD based on GAN

Theoretical Support [7][8]:
once the sample distribution and latent variable distribution were correlated through strong constraints, the changes of latent variables could be used to predict whether the sample was abnormal

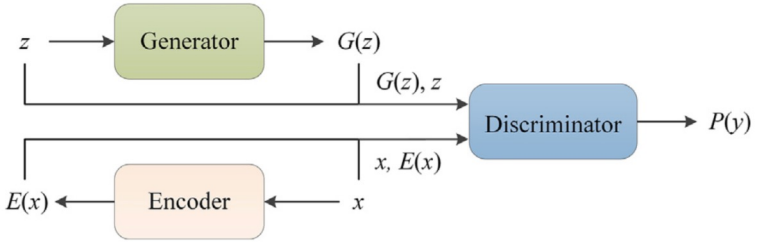
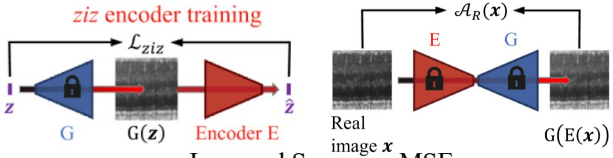
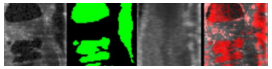


Fig 2. BiGAN and ALI Structure [7] [8]



Loss and Score are MSE



x, GT, $G(E(x))$, $\boldsymbol{x_r}$

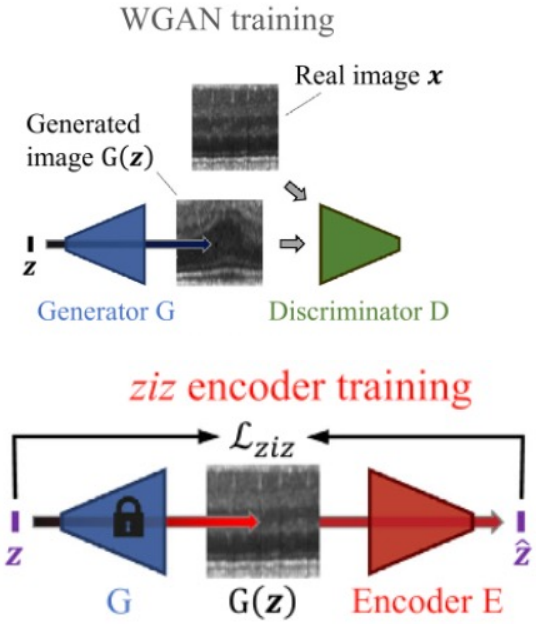
GT: Ground Truth

$\boldsymbol{x_r} = |x - G(E(x))|$

Fig 3. Example of f-AnoGAN [9]

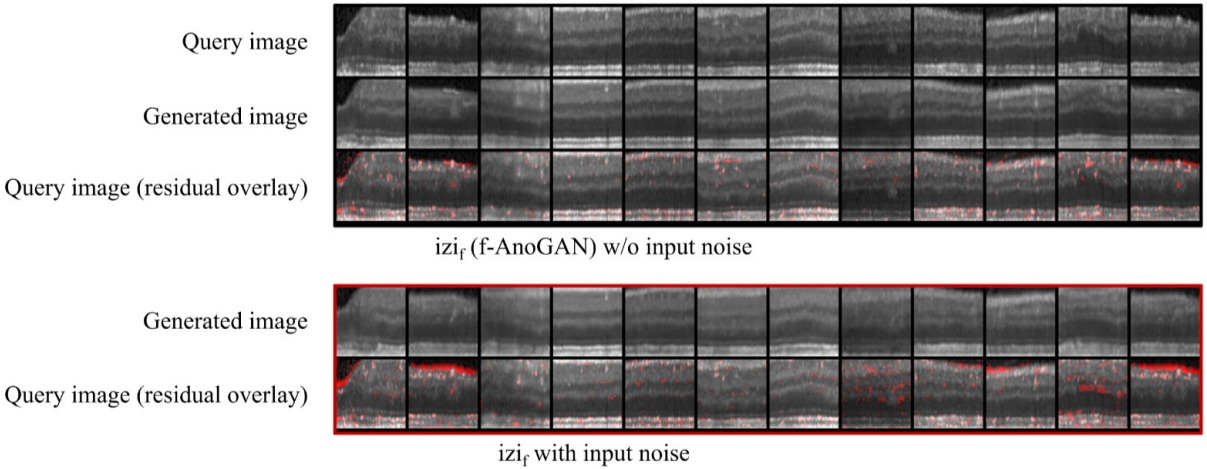
[1] "The AI index 2021 annual report.", [Zhang et al., Stanford University2021] [2] "Chestx-ray8: Hospital-scale chest x-ray database and…" [Wang et al., CVPR2017]
[3] "GAN-based anomaly detection: A review.", [Xia et al., Neurocomputing2022]
[4] "Unsupervised representation learning with deep convolutional generative adversarial networks." [Radford et al., arXiv2015]
[5] "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery." [Schlegl et al., IPMI2017]
[6] "Generative adversarial nets." [Goodfellow et al., NIPS2014]
[7] "Adversarial feature learning." [Donahue et al., ICLR2017]
[8] "Adversarially learned inference" [Dumoulin et al., ICLR2017]
[9] "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks." [Schlegl et al., MIA2019]

- Inverse Problem



x maps to $E(x)$ is an Inverse Problem

- SURE Loss



The previous method is sensitive to noise.
May be SURE can improve the robustness of methods

[1] “f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks.” [Schlegl et al., MIA2019]

Thank you!!!!

- SURE Gaussian noise proof

$$\mathrm{SURE}(h) = d\sigma^2 + \|g(x)\|^2 + 2\sigma^2 \sum_{i=1}^{d} \frac{\partial}{\partial x_i} g_i(x) = -d\sigma^2 + \|g(x)\|^2 + 2\sigma^2 \sum_{i=1}^{d} \frac{\partial}{\partial x_i} h_i(x),$$

We wish to show that

$$\mathrm{E}_\mu \|h(x) - \mu\|^2 = \mathrm{E}_\mu\{\mathrm{SURE}(h)\}.$$

We start by expanding the MSE as

$$\begin{aligned}
\mathrm{E}_\mu \|h(x) - \mu\|^2 &= \mathrm{E}_\mu \|g(x) + x - \mu\|^2 \\
&= \mathrm{E}_\mu \|g(x)\|^2 + \mathrm{E}_\mu \|x - \mu\|^2 + 2\,\mathrm{E}_\mu\, g(x)^T (x - \mu) \\
&= \mathrm{E}_\mu \|g(x)\|^2 + d\sigma^2 + 2\,\mathrm{E}_\mu\, g(x)^T (x - \mu).
\end{aligned}$$

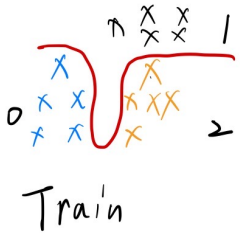
Now we use integration by parts to rewrite the last term:

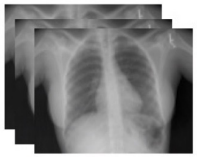
$$\begin{aligned}
\mathrm{E}_\mu\, g(x)^T (x - \mu) &= \int_{\mathbb{R}^d} \frac{1}{\sqrt{2\pi\sigma^{2d}}} \exp\left(-\frac{\|x - \mu\|^2}{2\sigma^2}\right) \sum_{i=1}^{d} g_i(x)(x_i - \mu_i) d^d x \\
&= \sigma^2 \sum_{i=1}^{d} \int_{\mathbb{R}^d} \frac{1}{\sqrt{2\pi\sigma^{2d}}} \exp\left(-\frac{\|x - \mu\|^2}{2\sigma^2}\right) \frac{dg_i}{dx_i} d^d x \\
&= \sigma^2 \sum_{i=1}^{d} \mathrm{E}_\mu \frac{dg_i}{dx_i}.
\end{aligned}$$

Substituting this into the expression for the MSE, we arrive at

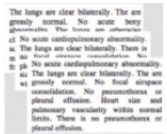
$$\mathrm{E}_\mu \|h(x) - \mu\|^2 = \mathrm{E}_\mu \left(d\sigma^2 + \|g(x)\|^2 + 2\sigma^2 \sum_{i=1}^{d} \frac{dg_i}{dx_i} \right).$$

- Supervised Learning Problem



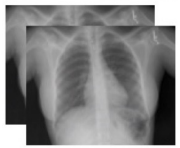


1000 CXRs



Report

Classifier → 100 Certainly Normal CXRs → Report: No founding



900 CXRs

Radiologists

Auxiliary Diagnosis:
Segmentation
Anomaly Score

May lead to delayed treatment

Inability to control the spread of the disease in time

Supplement

- Theory & Technology

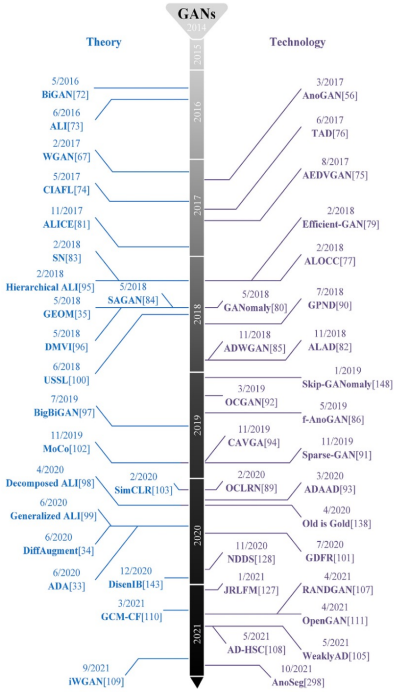


Fig 1. Time series diagram of the theory and application of GAN-based anomaly detection [2]

- CXR AD based on GAN

Model	Dataset	AUC
AnoGAN	COVIDx (balanced test set)	0.54
AnoGAN	Segmented COVIDx (balanced test set)	0.71
RANDGAN	Segmented COVIDx (balanced test set)	**0.77**
RANDGAN	Segmented COVIDx (imbalanced test set)	0.76

Fig 2. RANDGAN (proposed in 2021) [1]
Only compared AnoGAN (proposed in 2017)

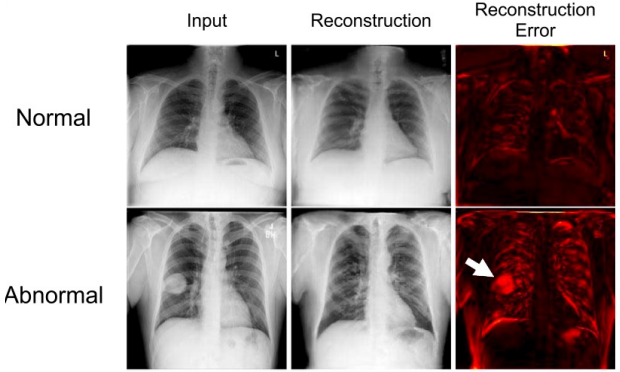


Fig 3. αGAN (proposed in 2021) [3]
No compare & similar with f-AnoGAN (proposed in 2019)

- Research Gap

Not much research & Lack of comparisons in modern anomaly detection methods based on GANs.

Reason:

1. Cross field. At present, many methods are still based on traditional CNN.
2. Many experts work on general anomaly detection methods.
3. Current anomaly detection methods lack explainable research.

But these are no reasons not to research, and my project can make a small contribution to this research gap

SOTA performance is not high
Adjusting parameters, modifying models, designing for losses
New Technologies, like: Classifier free guidance

- Significance

1. Research significance:
Research Gap
2. Application significance:
GUI

[1] “RANDGAN: randomized generative adversarial network for detection of COVID-19 in chest X-ray.” [Motamed et al., Scientific Reports 2021]
[2] “GAN-based anomaly detection: A review.”, [Xia et al., Neurocomputing2022]
[3] “Unsupervised deep anomaly detection in chest radiographs.” [Nakao et al., Journal of Digital Imaging 2021]

Challenge

1. Instability of Training

(Proof details in supplement slides)

[1][2][3][4]

If D is optimal, loss is:

$$C(G) = \max_D V(G, D)$$

$$= -\log(4) + 2 \cdot JSD\left(p_{\text{data}} \| p_g\right)$$

When p_g *and* p_{data} do not overlap:

$JSD = 0$ (No Gradient)

No overlapping:

a. Image are 2D manifold of 3D

b. $p_g \sim G(z)$ It is a mapping from low dimension to high dimension

c. Each training is a sampling of the data

WGAN-GP: There are still problems

$$\min_G \max_D V(D,G) = \mathbf{E}_{\mathbf{z}\sim P_{\mathbf{z}}(\mathbf{z})}[D(G(\mathbf{z}))] - \mathbf{E}_{\mathbf{x}\sim P_{\text{data}}(\mathbf{x})}[D(\mathbf{x})] + \lambda_{gp}\mathbf{E}_{\hat{\mathbf{x}}\sim P(\hat{\mathbf{x}})}\left[\left(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1\right)^2\right]$$

Challenge (Cont.)

2. Dataset

CXR data are high-resolution images, but training and detection on high-resolution images can take up a lot of computational resources

3. Evaluation Metrics

ROC-curve metrics can only measure the degree of anomaly in a way that is uninterpretable. [4]

Previous works had [5][6][7][8]:

Turing test, Residual image, Distribution

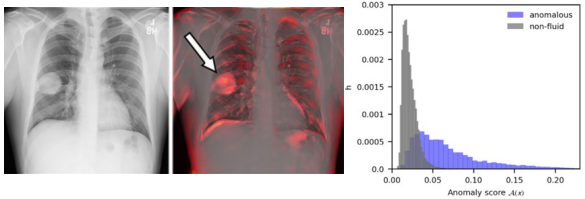


Fig 1. *Left*: Residual image, *Right*: Distribution

[1] “Wasserstein generative adversarial networks.”, [Arjovsky et al., ICLR2017]

[2] “Towards principled methods for training generative adversarial networks.”, [Arjovsky et al., arXiv2017]

[3] “Improved training of wasserstein gans.”, [Gulrajani et al., NIPS2017]

[4] “GAN-based anomaly detection: A review.”, [Xia et al., Neurocomputing2022]

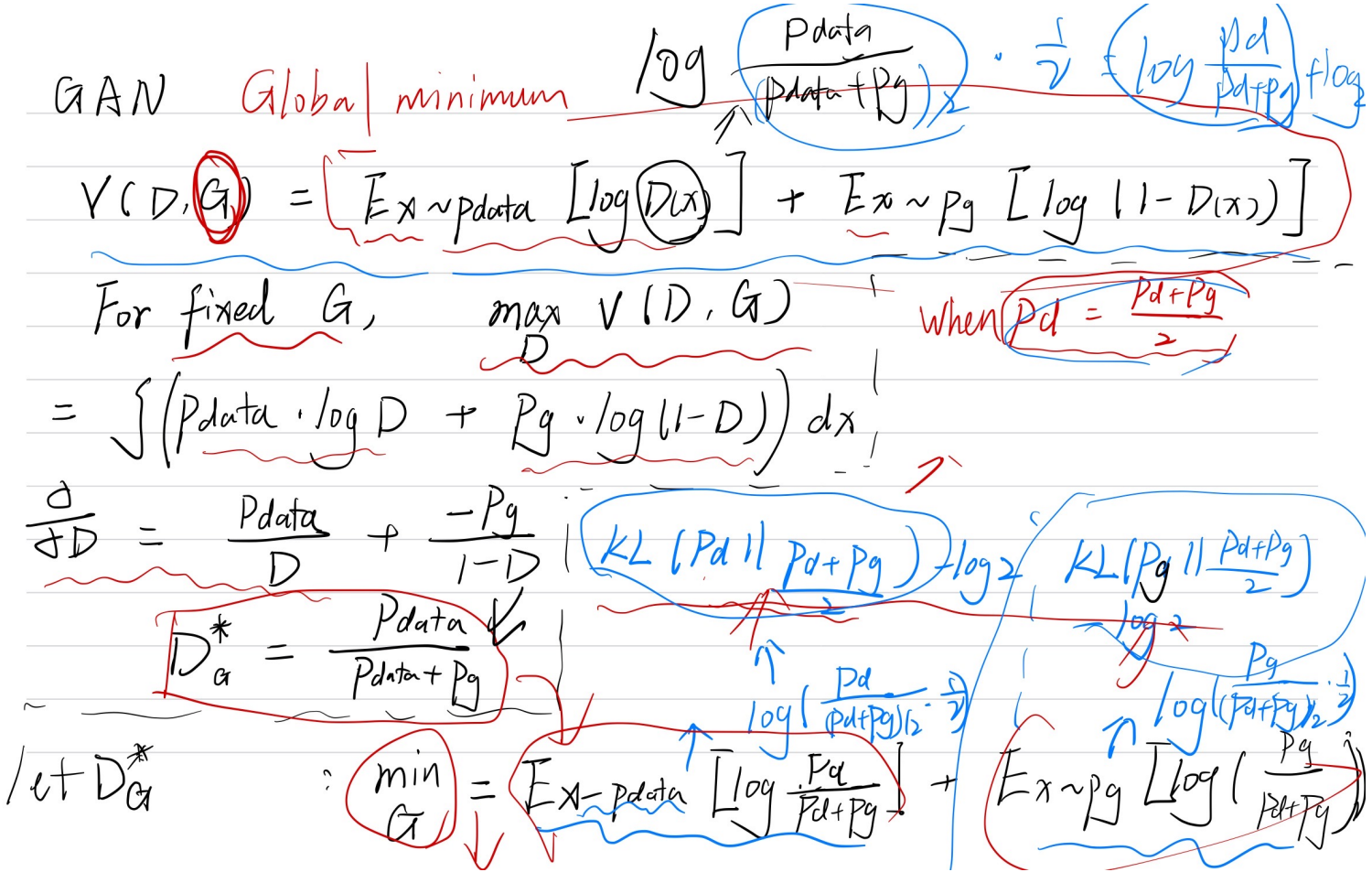
[5] “f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks.” [Schlegl et al., MIA2019]

[6] “Unsupervised representation learning with deep convolutional generative adversarial networks.” [Radford et al., arXiv2015]

[7] “RANDGAN: randomized generative adversarial network for detection of COVID-19 in chest X-ray.” [Motamed et al., Scientific Reports 2021]

[8] “Unsupervised deep anomaly detection in chest radiographs.” [Nakao et al., Journal of Digital Imaging 2021]

- GAN Global Minimum



- GAN Training Problem

In D^*, $\min_G = 2JS(P_d \| P_g) - 2\log 2$

$KL(P_1 \| P_2) = E_{x \sim P_1} \log \frac{P_1}{P_2}$ ($P_1 \in [0,1]$, $P_2 \in [0,1]$)

$JS(P_1 \| P_2) = \frac{1}{2} KL\left(P_1 \| \frac{P_1+P_2}{2}\right) + \frac{1}{2} KL\left(P_2 \| \frac{P_1+P_2}{2}\right)$

$= \frac{1}{2}\left[\int P_1 \log \frac{2P_1}{P_1+P_2} dx + \int P_2 \log \frac{2P_2}{P_1+P_2} dx\right]$

$= \log 2 + \frac{1}{2}\int P_1 \log \frac{P_1}{P_1+P_2} dx + \frac{1}{2}\int P_2 \log \frac{P_2}{P_1+P_2} dx$

When P_1 P_2 are not overlapping, $JS = \log 2$

$P_d \| P_g$

$\log 2$

$0 \leq \log 2$

WGAN

Wasserstein distance Earth - Mover distance

$$W(P_d, P_g) = \inf_{\gamma \sim \Pi(P_d, P_g)} E_{(x,y)\sim\gamma}[\|x-y\|]$$

$$= \frac{1}{K} \sup_{\|f\|_L \le K} E_{x\sim P_r}[f(x)] - E_{x\sim P_g}[f(x)]$$

Lipschitz $|f(x_1) - f(x_2)| \le K|x_1 - x_2|$

Any x_1, x_2 ,

- WGAN

$K \cdot W(P_r, P_g) \approx \max_{w \cdot \|f_w\|_L \le K} E_{x \sim p_r}[f_w(x)] - E_{x \sim p_g}[f_w(x)]$

Trick: Clipping

$[-c, c]$ w

$\min_G \max_D E_{x \sim p_d} D(x) - E_{z \sim p_g} D(G(z))$

G Loss $-E_{z \sim p_g} D(G(z))$

D Loss $E_{z \sim p_g} D(G(z)) - E_{x \sim p_d} D(x)$

- WGAN

Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

Require: : α, the learning rate. c, the clipping parameter. m, the batch size. n_{critic}, the number of iterations of the critic per generator iteration.

Require: : w_0, initial critic parameters. θ_0, initial generator's parameters.

while θ has not converged **do**
  for $t = 0, ..., n_{\text{critic}}$ **do**
    Sample $\{x^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_r$ a batch from the real data.
    Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
    $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))\right]$
    $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$
    $w \leftarrow \text{clip}(w, -c, c)$
  end for
  Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
  $g_\theta \leftarrow -\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))$
  $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$
end while

Adam

① remove sigmoid

② No log

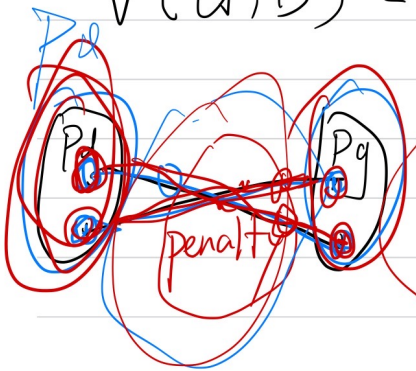
③ clip

④ momentum
Not good

- WGAN-GP

WGAN-GP

$$V(G,D) = \max_{D \text{ 1-Lipschitz}} \{ E_{x\sim p_{data}}[D(x)] - E_{x\sim p_G}[D(x)] \}$$



$$\|\nabla_x D(x)\| \le 1$$

$$V(G,D) \approx \max_D \{ E_{x\sim p_{data}}[Dx] - E_{x\sim p_g}[Dx] - 1 \}^2 + E_{x\sim penalty}(\|\nabla_x Dx\| - 1)^2$$

$$\hat{x}_1 = x_{d_1}\cdot\beta + x_{g_1}\cdot(1-\beta)$$

$\beta \in [0,1]$

$$\max(0, \|\nabla_x Dx\| - 1)\,dx$$