XRayGAN: Consistency-preserving Generation of X-ray Images from Radiology Reports

Xingyi Yang, Nandiraju Gireesh, Eric Xing, Pengtao Xie

Song Wang, Liyan Tang

https://arxiv.org/abs/2006.10552

To train medical students to become qualified radiologists, a large number of X-ray images collected from patients are needed, but due to privacy concerns, such images are typically difficult to obtain.

Hence, we develop methods to generate view-consistent, high-fidelity, high-resolution X-ray images from radiology reports.

Challenges

1. Radiology reports are long and have complicated structure. How to effectively understand the semantics to generate high-fidelity images that accurately reflect the contents of the report?

2. How to generate high-resolution images?

3. From a single report, images with different views (frontal, lateral, etc) need to be generated. How to ensure the consistency of images (e.g., from the same patient)?

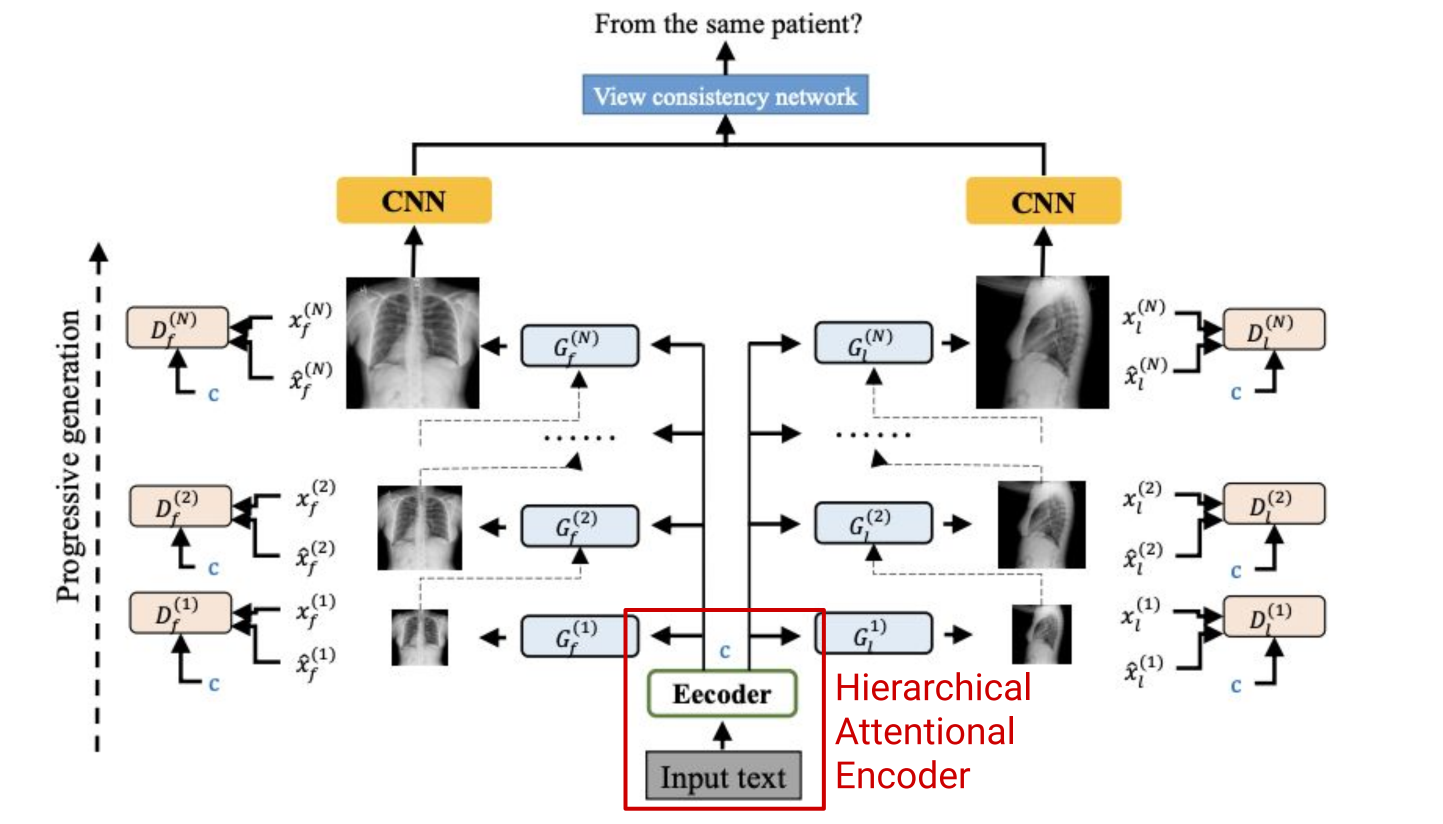
Methods - 1

1. Radiology reports are long and have complicated structure. How to effectively understand the semantics to generate high-fidelity images that accurately reflect the contents of the report?

   -> Hierarchical Attentional Encoder (HAE)

2. How to generate high-resolution images?

3. From a single report, images with different views (frontal, lateral, etc) need to be generated. How to ensure the consistency of images (e.g., from the same patient)?


From the same patient?
View consistency network
CNN
CNN
Progressive generation
Eecoder
Input text
Hierarchical Attentional Encoder

Methods - Hierarchical Attentional Encoder

Takes the radiology reports as inputs, encodes it into an embedding vector that captures the semantics of the report.

Word -> Sequence of words -> Sentence -> Sequence of sentences - > Report

Use a hierarchical LSTM network:

- Word-level LSTM: Takes the sequence of words in one sentence as inputs, learns latent embeddings for each word and the sentence.
- Sentence-level LSTM: Takes the embeddings of a sequence of sentences as inputs, learns an embedding of the report.

Zichao Yang, et al. Hierarchical attention networks for document classification.

Methods - Attentional Word-level LSTM

For a word $w^{(t)}$ at position t, its embedding is $c_{word}^{(t)} = W_e w^{(t)}$

* using one hot encoding

* Learnable embedding matrix

Zichao Yang, et al. Hierarchical attention networks for document classification.

Methods - Attentional Word-level LSTM

For a word $w^{(t)}$ at position t, its embedding is $c_{word}^{(t)} = W_e w^{(t)}$

-> fed into a bi-directional LSTM

Zichao Yang, et al. Hierarchical attention networks for document classification.

Methods - Attentional Word-level LSTM

For a word $w^{(t)}$ at position t, its embedding is $c_{word}^{(t)} = W_e w^{(t)}$

-> fed into a bi-directional LSTM

-> hidden states $\overleftarrow{h}_{word}^{(t)}$, $\overrightarrow{h}_{word}^{(t)}$ that capture contextual information of $w^{(t)}$.

$$\overrightarrow{h}_{it} = \overrightarrow{\mathrm{GRU}}(x_{it}), t \in [1, T],$$
$$\overleftarrow{h}_{it} = \overleftarrow{\mathrm{GRU}}(x_{it}), t \in [T, 1].$$

Zichao Yang, et al. Hierarchical attention networks for document classification.

Within a sentence, some words (e.g., medical terms) are more important than others.

We use an attentional module to calculate the attention score for each word, and use these attention scores to re-weight the word embeddings we obtained from word-level LSTM.

Methods - Attentional Word-level LSTM

-> Importance score of word $w^{(t)}$: $e_{word}^{(t)} = v_w \tanh(W_w h_{word}^{(t)} + b_w)$

* Concatenation of two hidden states

(W_w: learnable weight matrix, v_w and b_w: learnable weight vectors)

- (Feed *h* through an one-layer Multi-layer Perceptron)

Zichao Yang, et al. Hierarchical attention networks for document classification.

Methods - Attentional Word-level LSTM

-> Importance score of word $w^{(t)}$: $e_{word}^{(t)} = v_w \tanh(W_w h_{word}^{(t)} + b_w)$

(W_w: learnable weight matrix, v_w and b_w: learnable weight vectors)

-> Normalize using softmax: $\alpha_{word}^{(t)} = \exp\left(e_{word}^{(t)}\right) / \sum_{j=1}^{T} \exp(e_{word}^{(j)})$

Zichao Yang, et al. Hierarchical attention networks for document classification.

Methods - Attentional Word-level LSTM

-> Importance score of word $w^{(t)}$: $e_{word}^{(t)} = v_w \tanh(W_w h_{word}^{(t)} + b_w)$

(W_w: learnable weight matrix, v_w and b_w: learnable weight vectors)

-> Normalize using softmax: $\alpha_{word}^{(t)} = \exp\left(e_{word}^{(t)}\right) / \sum_{j=1}^{T} \exp(e_{word}^{(j)})$

-> Representation of the sentence: $c_{sent} = \sum_{j=1}^{T} \alpha_{word}^{(j)} h_{word}^{(j)}$

* weighted sum of words' representations

Zichao Yang, et al. Hierarchical attention networks for document classification.

Methods - Attentional Sentence-level LSTM

Given a sequence of sentences in the report

-> word-level LSTM to obtain an embedding of each sentence

-> feed embeddings of sentences into a sentence-level LSTM

-> an embedding of the report

Zichao Yang, et al. Hierarchical attention networks for document classification.

Methods - Attentional Sentence-level LSTM

-> bi-directional LSTM to capture contextual information

-> calculate importance score of each sentence

-> normalize, and calculate the weighted sum of hidden states

-> representation of the report

Zichao Yang, et al. Hierarchical attention networks for document classification.

Methods - 2

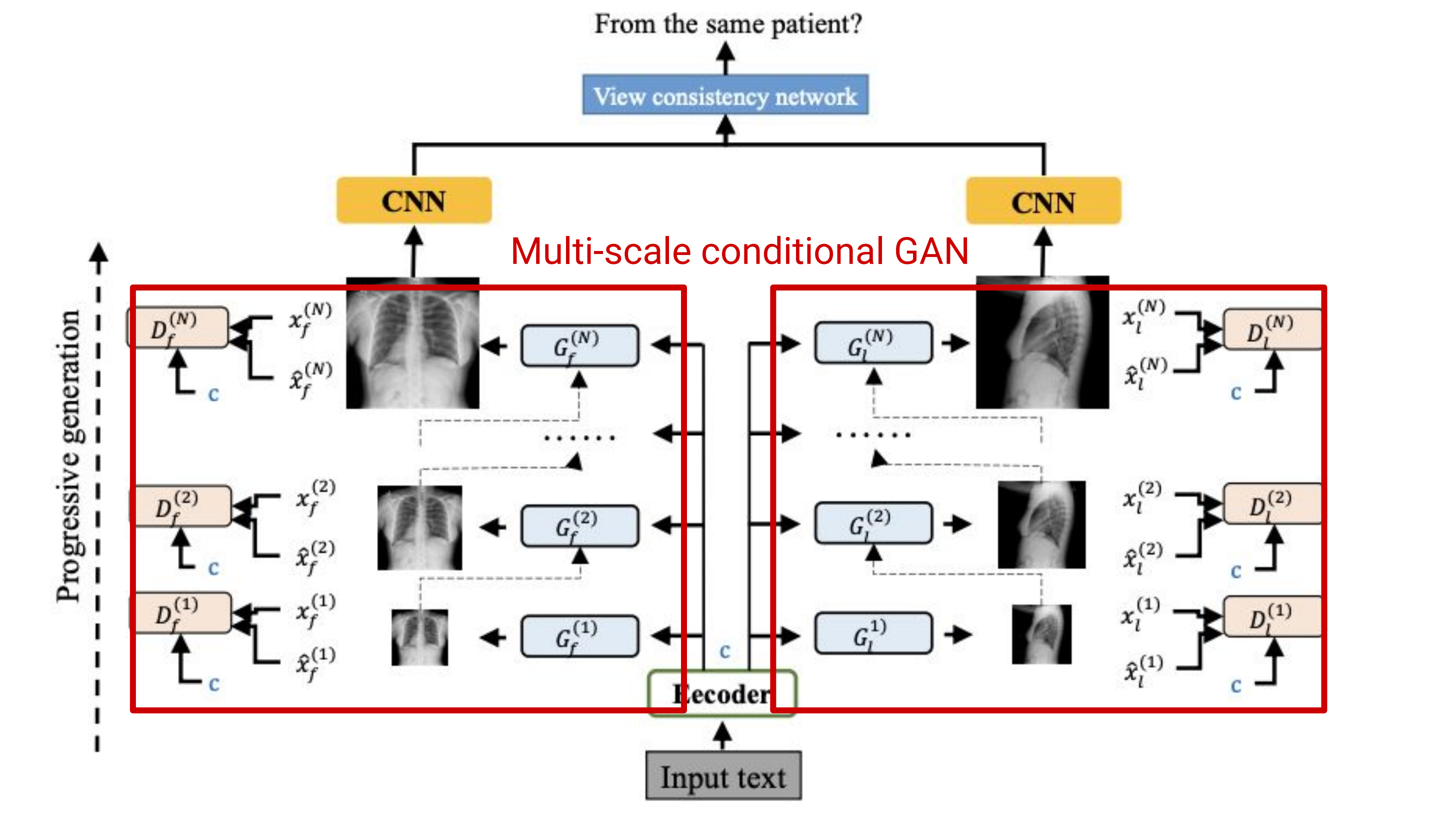
1. Radiology reports are long and have complicated structure. How to effectively understand the semantics to generate high-fidelity images that accurately reflect the contents of the report?

   -> Hierarchical Attentional Encoder (HAE)

2. How to generate high-resolution images?

   -> Multi-Scale Conditional GAN

3. From a single report, images with different views (frontal, lateral, etc) need to be generated. How to ensure the consistency of images (e.g., from the same patient)?


From the same patient?
View consistency network
CNN
CNN
Multi-scale conditional GAN
Progressive generation
Input text

Methods - Multi-scale conditional GAN

Basic genera

Progressive g

lower-resolut ogy report.

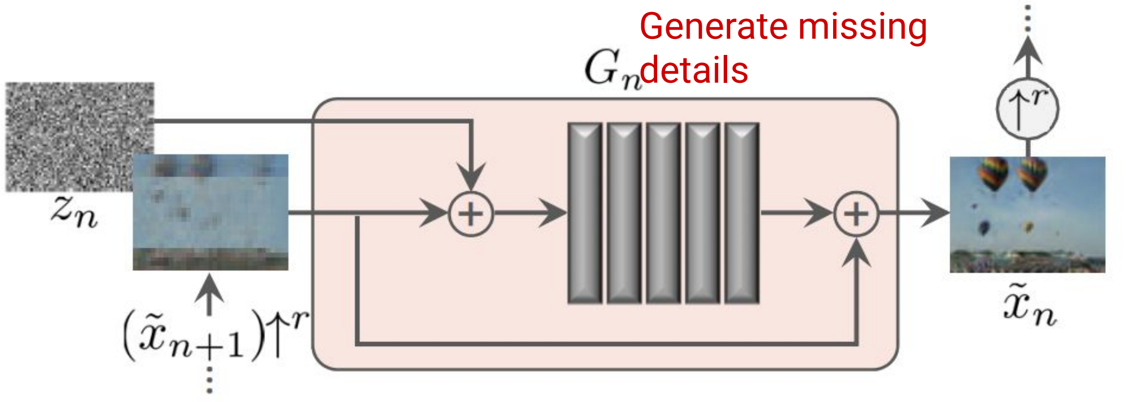
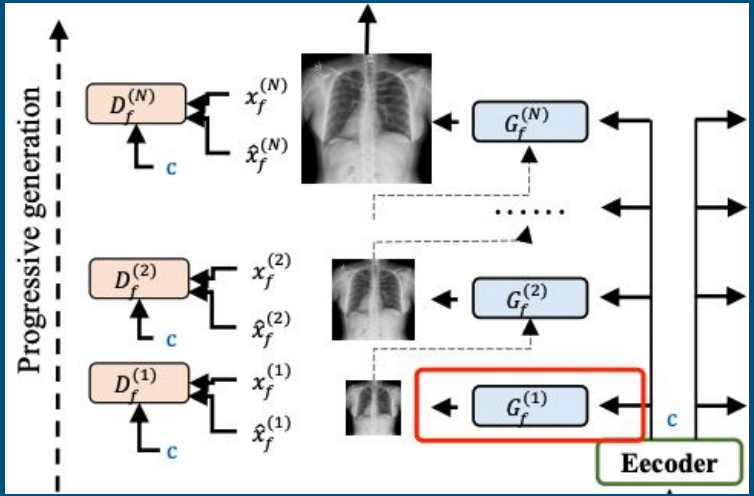


Figure 5: **Single scale generation.** At each scale n, the image from the previous scale, $\tilde{x}_{n+1}$, is upsampled and added to the input noise map, z_n. The result is fed into 5 conv layers, whose output is a residual image that is added back to $(\tilde{x}_{n+1})\uparrow^r$. This is the output $\tilde{x}_n$ of G_n.

Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. https://arxiv.org/pdf/1905.01164.pdf

Methods - Basic Generator

Basic generator $G_f^{(1)}$ takes embedding c of the report as input.

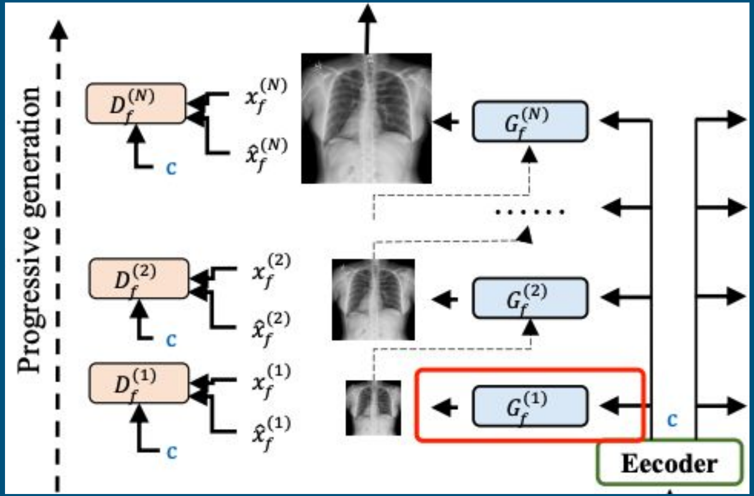


Methods - Basic Generator

Basic generator $G_f^{(1)}$ takes embedding c of the report as input.

-> Frontal-view $x_f^{(1)}$ with the lowest resolution

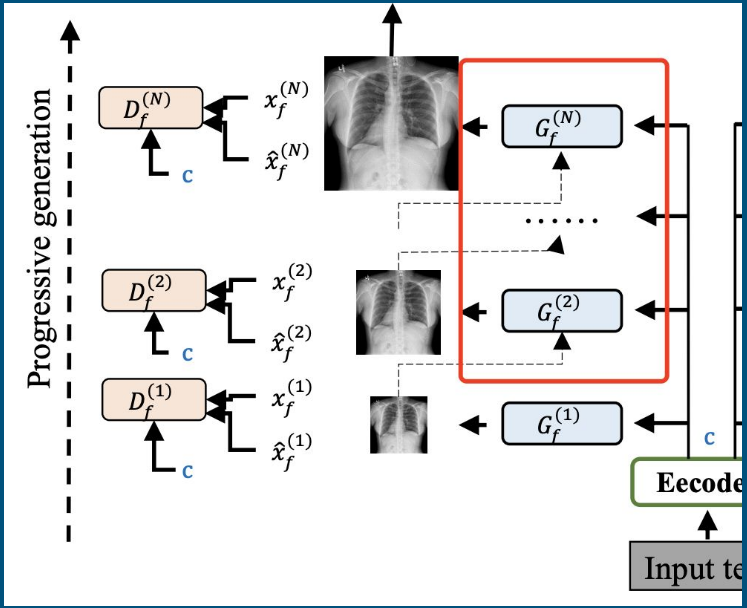
* 32 x 32



Methods – Progressive Generators

Produce images with increasing resolutions, by a factor of 2.

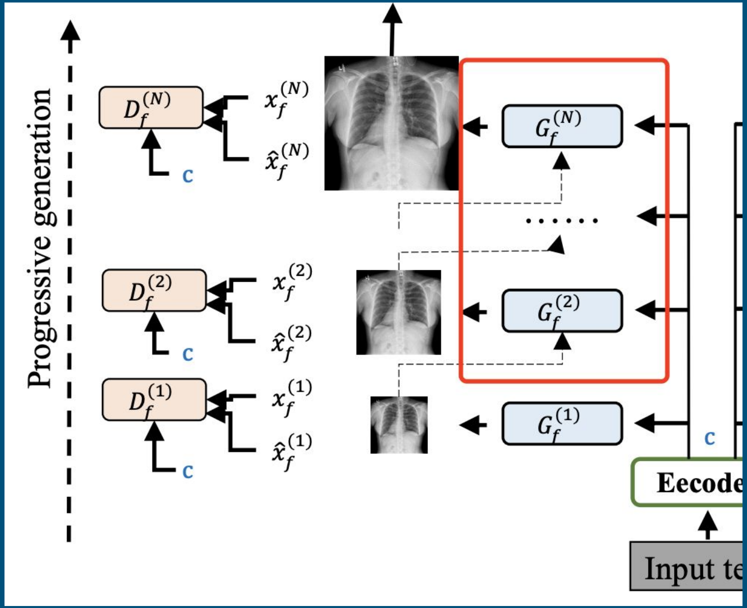
32 x 32 -> 64 x 64 -> 128 x 128 -> 256 x 256



Methods – Progressive Generators

Produce images with increasing resolutions, by a factor of 2.

The outputs at stage n-1 are the inputs at stage n.



Methods – Progressive Generators

Produce images with increasing resolutions, by a factor of 2.

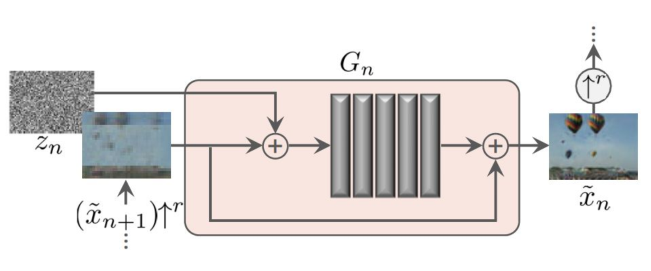
The outputs at stage n-1 are the inputs at stage n.

$$x_f^{(n)} = \alpha G_f^{(n)}\left(x_f^{(n-1)}, c\right) + (1-\alpha)U(x_f^{(n-1)})$$

* Generate higher-resolution image containing more fine-grained visual information

$U(x_f^{(n-1)})$ is to resize $x_f^{(n-1)}$ to the size of $x_f^{(n)}$.

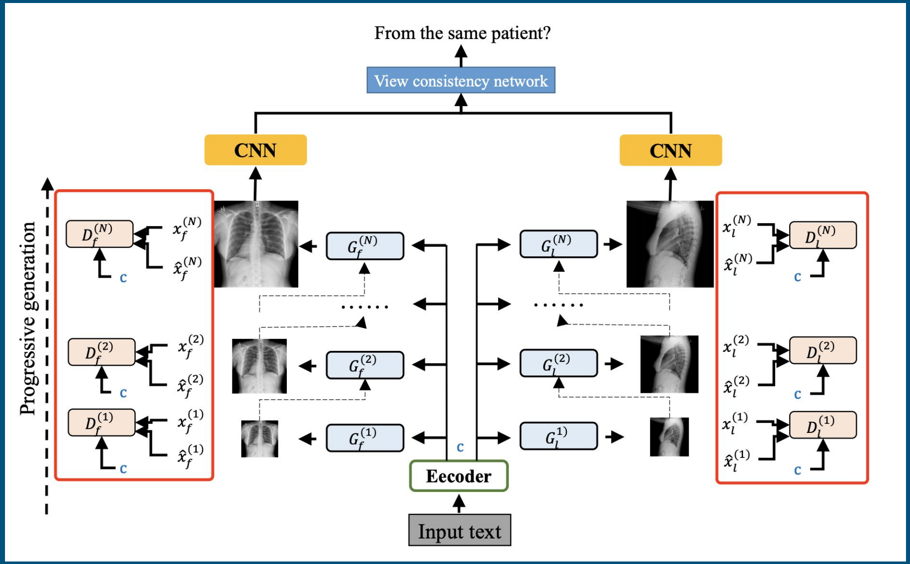
α is the level of blending.



Methods - Discriminators

Similar to GAN*, we use discriminators at each stage to judge whether images are generated or real.

Generators are learned in a way that such a discrimination is difficult to achieve.



*Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. https://arxiv.org/pdf/1905.01164.pdf

Methods - 3

1. Radiology reports are long and have complicated structure. How to effectively understand the semantics to generate high-fidelity images that accurately reflect the contents of the report?

   -> Hierarchical Attentional Encoder (HAE)

2. How to generate high-resolution images?

   -> Multi-Scale Conditional GAN

3. From a single report, images with different views (frontal, lateral, etc) need to be generated. How to ensure the consistency of images (e.g., from the same patient)?

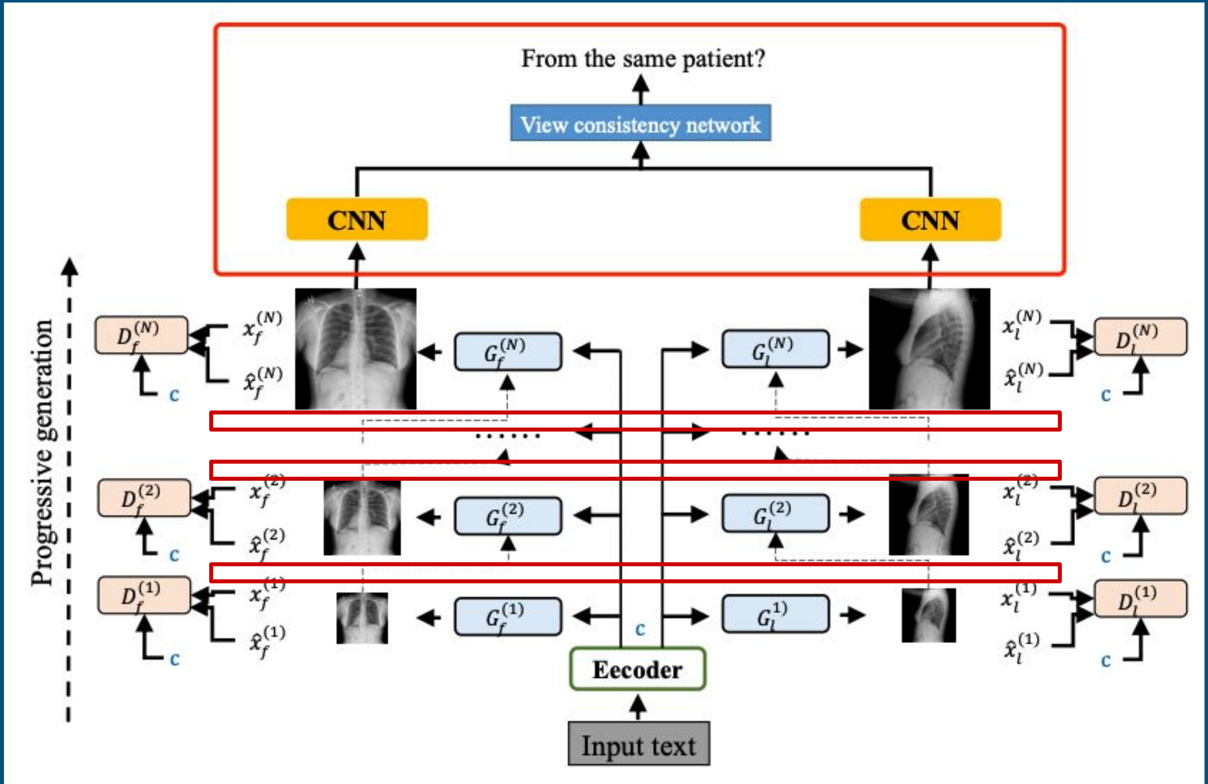
   -> View Consistency Network

Methods - View Consistency Network

Trained off-line using real X-ray images:

- Consistent: Frontal-view and Lateral-view images belonging to the same patient
- Inconsistent: Frontal-view and Lateral-view images belonging to different patient

The view consistency network is trained offline and its weights are frozen during XRayGAN training.


From the same patient?
View consistency network
CNN
CNN
Progressive generation
$D_f^{(N)}$
$x_f^{(N)}$
$\hat{x}_f^{(N)}$
$G_f^{(N)}$
$G_l^{(N)}$
$x_l^{(N)}$
$\hat{x}_l^{(N)}$
$D_l^{(N)}$
$D_f^{(2)}$
$x_f^{(2)}$
$\hat{x}_f^{(2)}$
$G_f^{(2)}$
$G_l^{(2)}$
$x_l^{(2)}$
$\hat{x}_l^{(2)}$
$D_l^{(2)}$
$D_f^{(1)}$
$x_f^{(1)}$
$\hat{x}_f^{(1)}$
$G_f^{(1)}$
$G_l^{1)}$
$x_l^{(1)}$
$\hat{x}_l^{(1)}$
$D_l^{(1)}$
c
Eecoder
Input text

Methods - View Consistency Network

A modified Resnet-18 $f^{(n)}$ is used to compute image embeddings at stage n.

Methods - View Consistency Network

A modified Resnet-18 $f^{(n)}$ is used to compute image embeddings at stage n.

Calculate embeddings of frontal-view and lateral-view images: $f^{(n)}(x_f^{(n)})$ and $f^{(n)}(x_l^{(n)})$

Methods - View Consistency Network

A modified Resnet-18 $f^{(n)}$ is used to compute image embeddings at stage n.

Calculate embeddings of frontal-view and lateral-view images: $f^{(n)}(x_f^{(n)})$ and $f^{(n)}(x_l^{(n)})$

Probability that they are consistent:

$$\sigma(\sum_j w_j \left| f_j^{(n)}\left(x_f^{(n)}\right) - f_j^{(n)}\left(x_l^{(n)}\right) \right|)$$

(σ: sigmoid function, j: dimensions of embeddings)

Objective Function

*Minimize Obj = 1 * adversarial loss +*
*100 * reconstruction loss +*
*(-1) * view-consistency reward*

Adversarial Loss: The generators aim to minimize this loss and the discriminators aim to maximize this loss.

Reconstruction Loss: the pixel-level L2 distance between a generated image and the ground-truth image.

View-consistency Loss: measuring how likely two images are from the same patient

Result - 1

Table 1: Results achieved by different methods on the test sets of Open-i and MIMIC-CXR.

	Open-i				MIMIC-CXR			
Method	IS↑	FID↓	SSIM↑	VC ↑	IS↑	FID↓	SSIM↑	VC ↑
Real	-	-	-	0.669	-	-	-	0.589
GAN-INT-CLS [15]	1.015	272.7	0.201	0.525	1.039	214.3	0.291	0.555
StackGAN [16]	1.043	243.4	0.138	0.487	1.063	245.5	0.212	0.471
AttnGAN [17]	1.055	226.6	0.171	0.508	1.067	232.7	0.231	0.487
XRayGAN	**1.081**	**141.5**	**0.343**	**0.627**	**1.112**	**86.15**	**0.379**	**0.580**

XRayGAN achieves better performance than the three baselines.

Result - 2

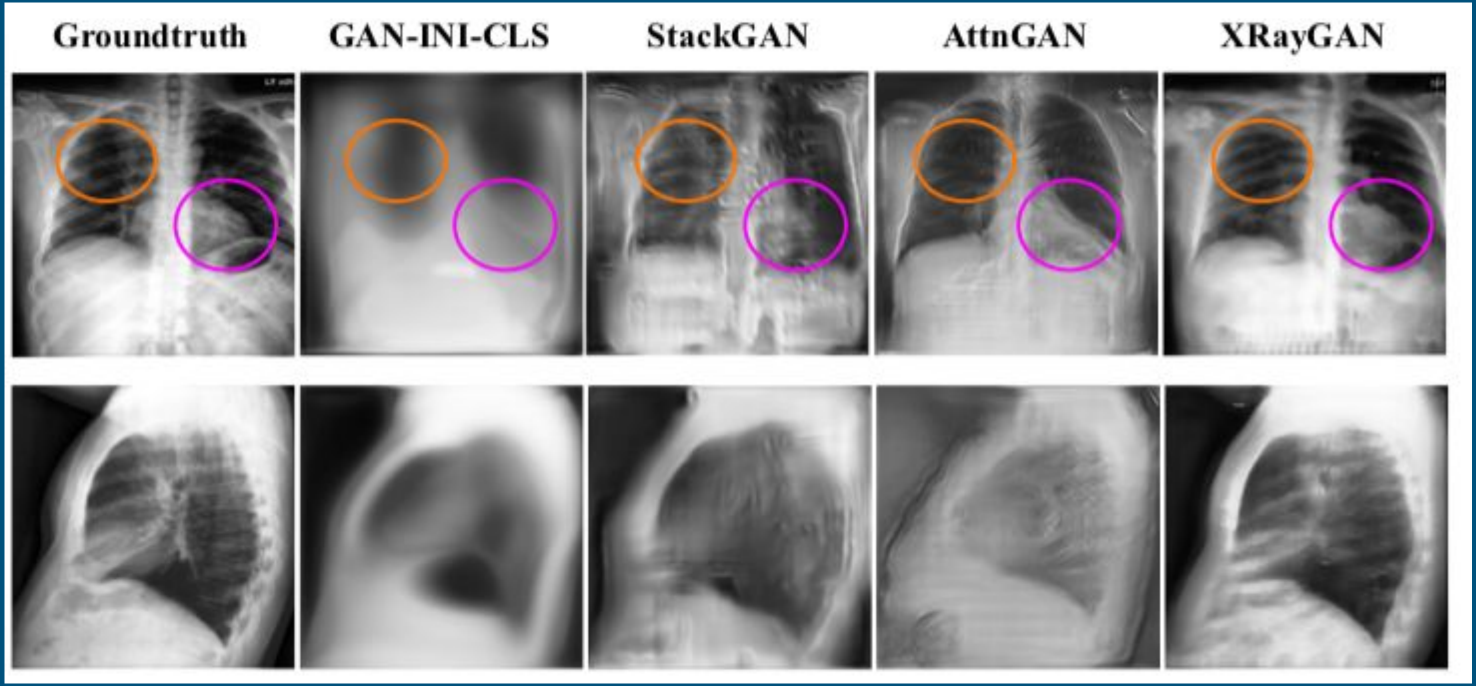
Table 2: Ablation study on the Open-i dataset.

Method	Open-i IS↑	FID↓	SSIM↑	VC↑
Real	-	-	-	0.669
w/o VCN	1.079	147.4	**0.353**	0.623
w/o MSCGAN	1.022	235.5	0.164	0.605
w/o HAE	1.079	153.2	0.300	0.621
Full	**1.081**	**141.5**	0.343	**0.627**

VC: View Consistency

1. VCN explicitly encourages the generators to generate consistent frontal and lateral images.
2. MSCGAN is essential in generating high-resolution and high-fidelity images.
3. Using hierarchical LSTM to capture the linguistic hierarchy in the report can better understand the report, which further makes the generated images have better fidelity.

Result - 3



1. The images generated by XRayGAN are clear and visually very similar to the groundtruth.
    a. VCN encourage the lateral image to be as clear as the frontal image.
    b. It generates better frontal-view image since it uses multi-scale progressive GANs to improve the resolution of generated images.

2. The images generated by XRayGAN correctly reflect the clinical findings in the reports.

Summary and takeaways

1. A hierarchical attentional encoder is used to capture the hierarchical linguistic structure and various clinical importance of words and sentences
2. A multi-scale conditional GAN (MSCGAN) is used to generate high-resolution X-ray images in a progressive manner.
3. Proposed an XRayGAN where a view consistency network (VCN) is used to encourage the generated frontal-view and lateral-view images to be from the same patient.

GitHub Repo: https://github.com/UCSD-AI4H/XRayGAN.

Questions?