JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Wavelet Diffusion Models are fast and scalable Image Generators

Hao Phung*

Quan Dao*

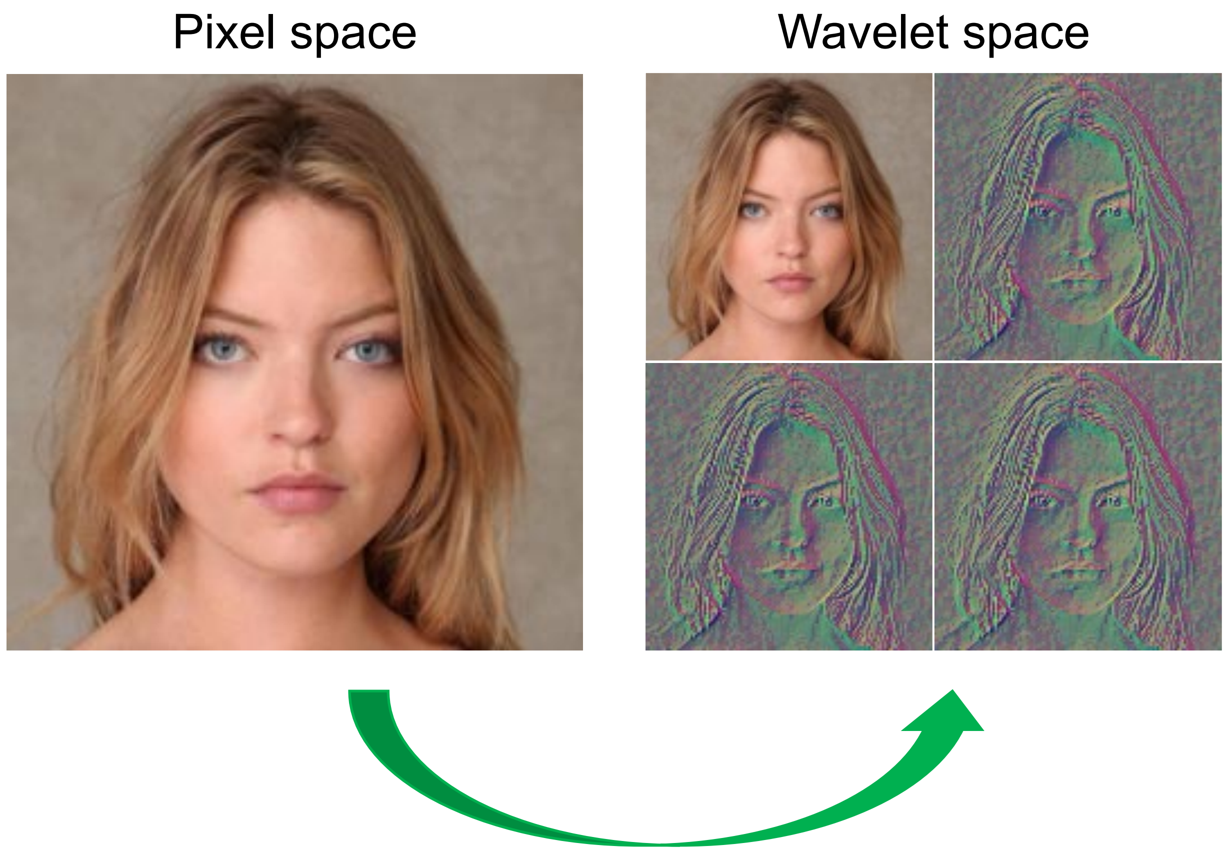
Anh Tran

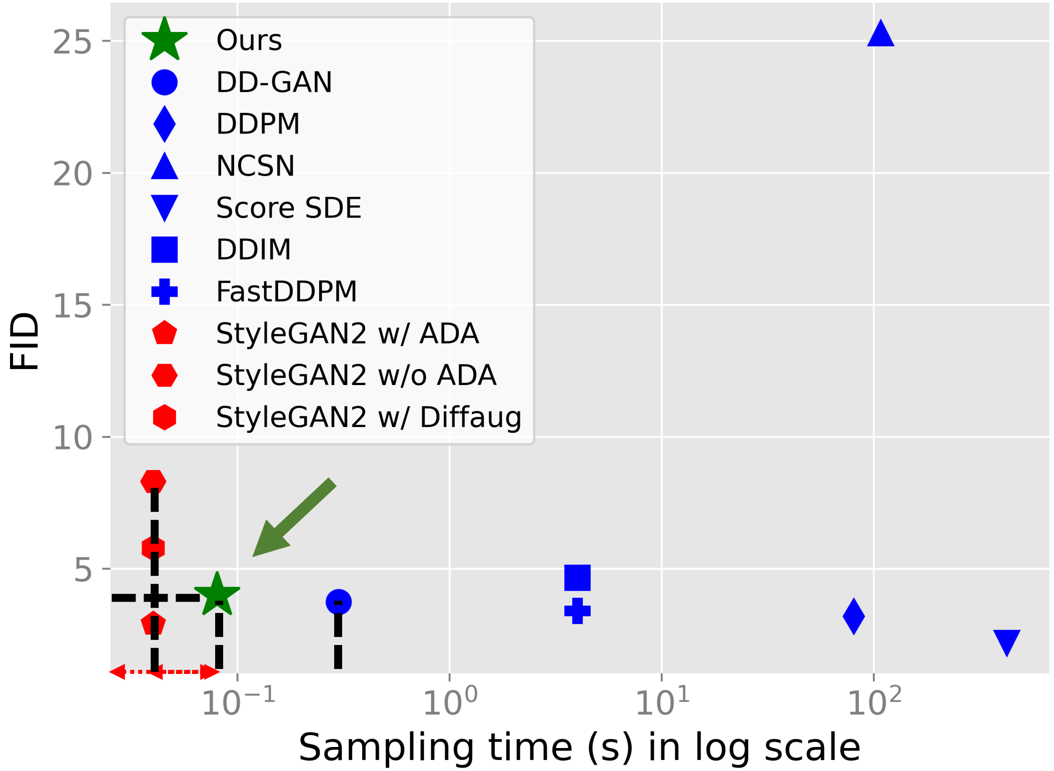
*Equal contribution

Poster: WED-AM-188

Paper

WaveDiff overview





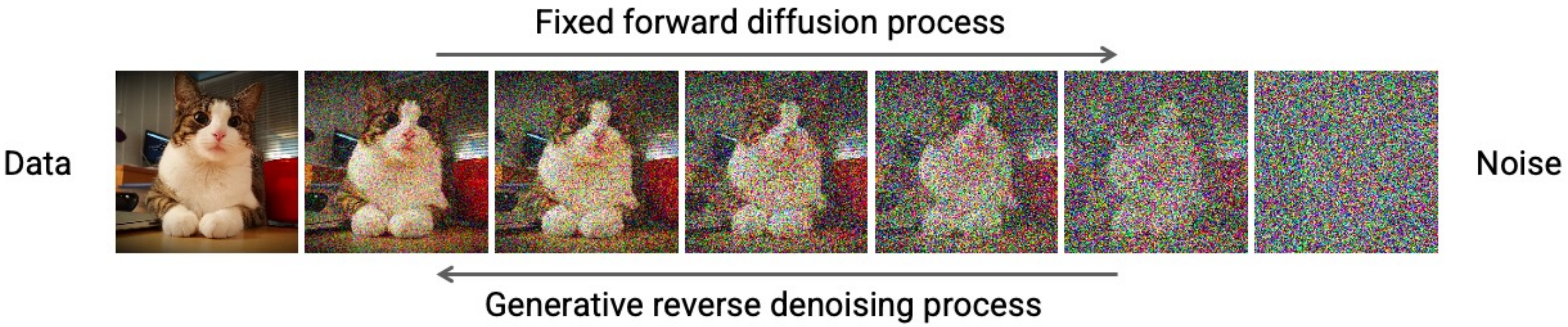
CIFAR-10: Diffusion model that closes the gap speed with GAN methods

Single-image sampling time

	CIFAR-10	STL-10	CelebA(256)	CelebA(512)	CelebA(1024)	Church
Resolution	32	64	256	512	1024	256
#time steps	4	4	2	2	2	4
Time (s)	0.07	0.12	0.08	0.1	0.12	0.16

Produce images up to 1024 × 1024 in a mere 0.1s, which is the first time for a diffusion model to achieve such almost real-time performance.

Diffusion models



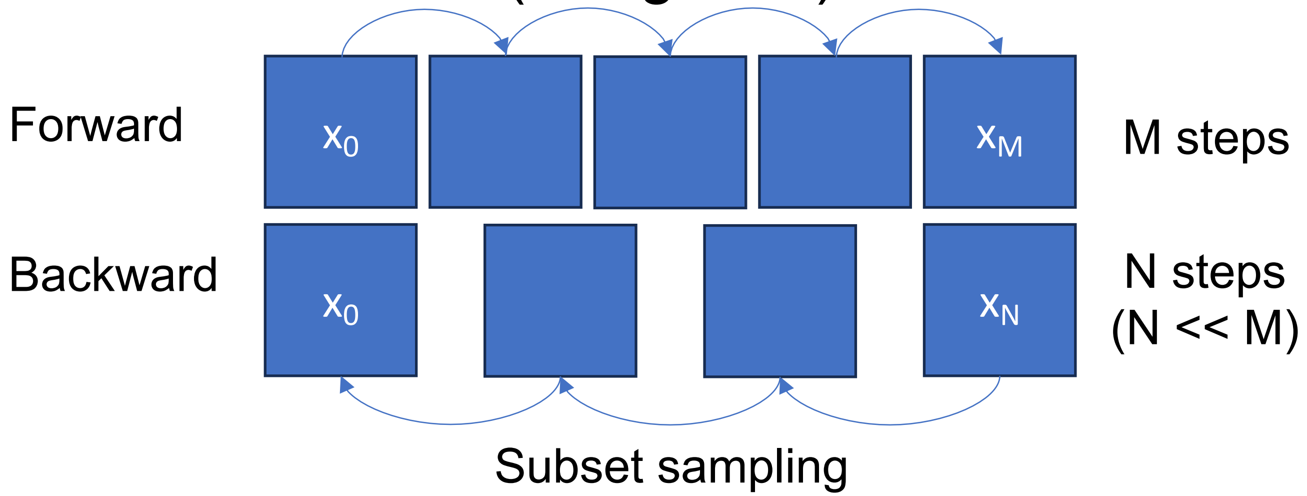
Pros: superior performance on variety of tasks + flexible conditional inputs

Cons: requires thousand-steps traversal to generate a sample. Super low and computational!

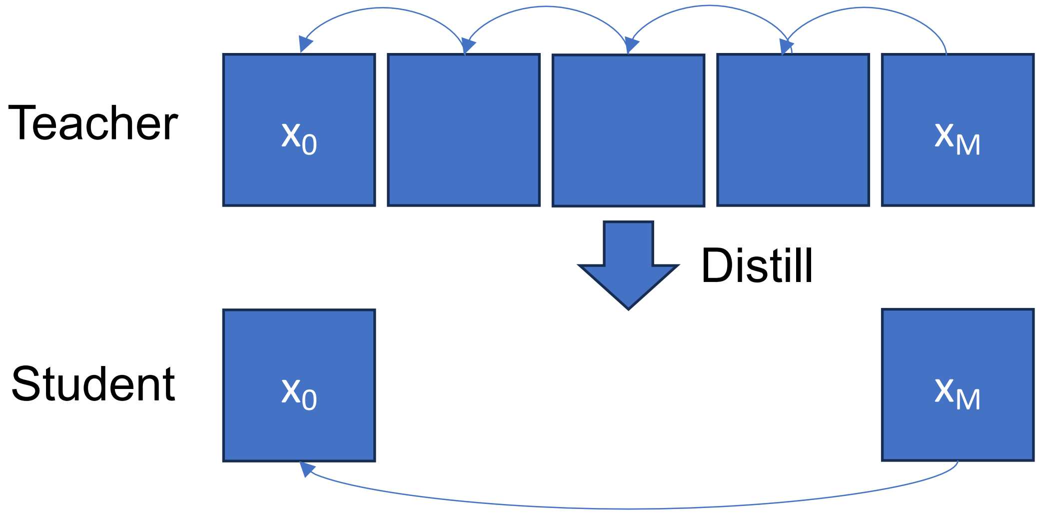
[CVPR 2022 tutorial] Denoising Diffusion-based Generative Modeling: Foundations and Applications

Prior works

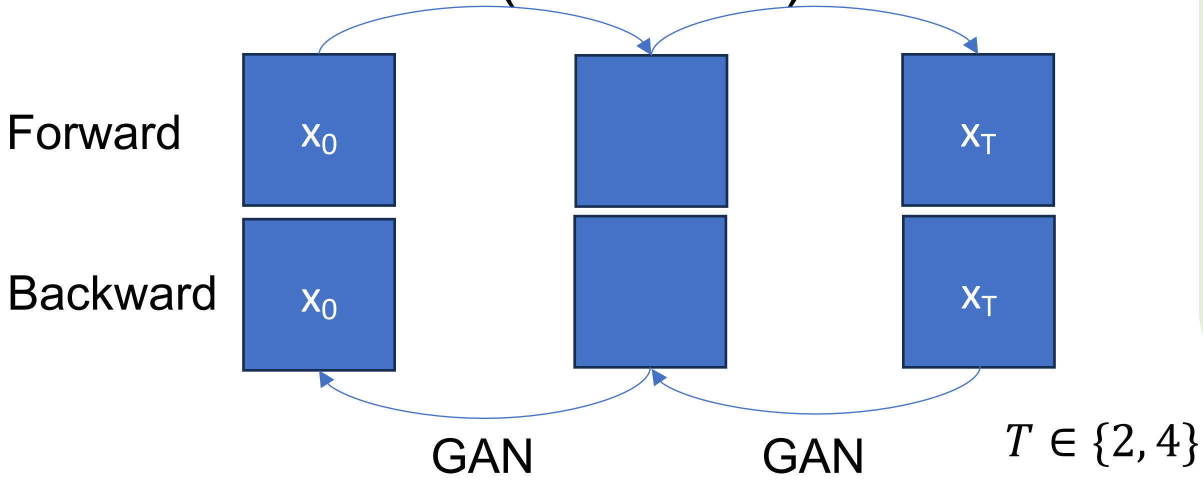
DDIM
(Song et al)



Distillation
(Luhman & Luhman)



DDGAN
(Xiao et al)



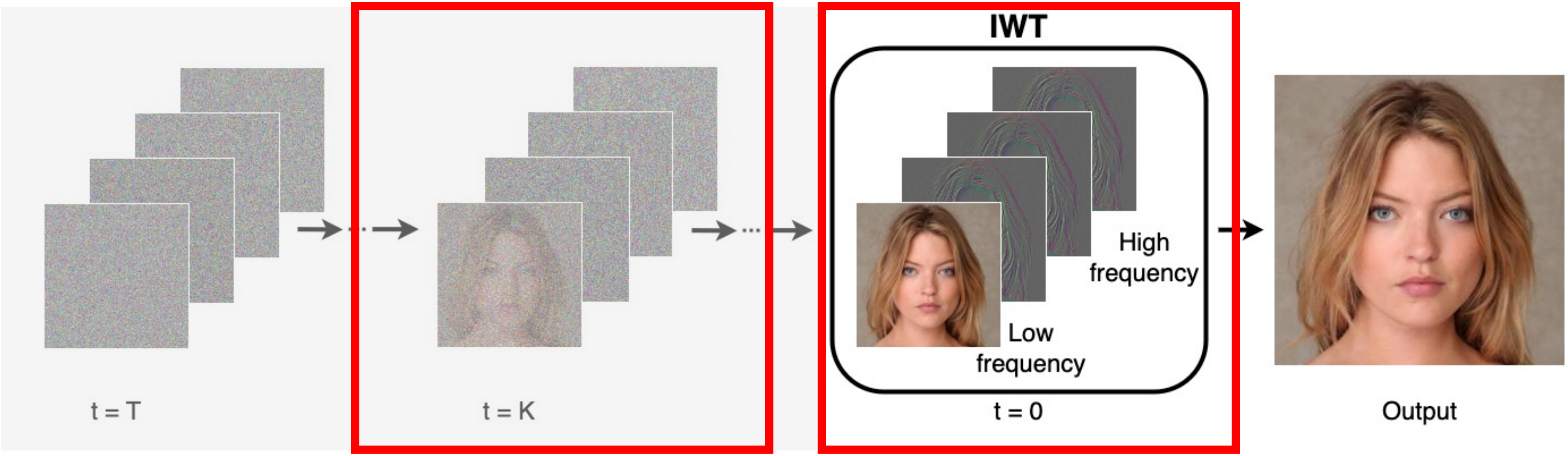
None of them can fully address sampling speed for both training and testing phase.

Our method instead utilizes wavelet transformation to improve sampling efficiency of DDGAN while maintaining competitive visual quality.

Wavelet diffusion scheme

4x smaller dimension. More compute-efficient.

Disentangled representation of low-and-high frequencies. Simpler to Learn.



Training objective

Given noisy sample y_t, latent $z \sim \mathcal{N}(0, \boldsymbol{I})$ and time t, generator outputs clean sample $y'_0 = G(y_t, z, t)$
and then draws the less noisy sample from $y'_{t-1} \sim q(y_{t-1} | y_t, y'_0)$

Original Loss

(1) Adversarial loss:

$$\mathcal{L}_{adv}^{D} = -\log\ (D(y_{t-1}, y_t, t)) + \log\ (D(y'_{t-1}, y_t, t))\ ,$$

$$\mathcal{L}_{adv}^{G} = -\log\ (D(y'_{t-1}, y_t, t))\ ,$$

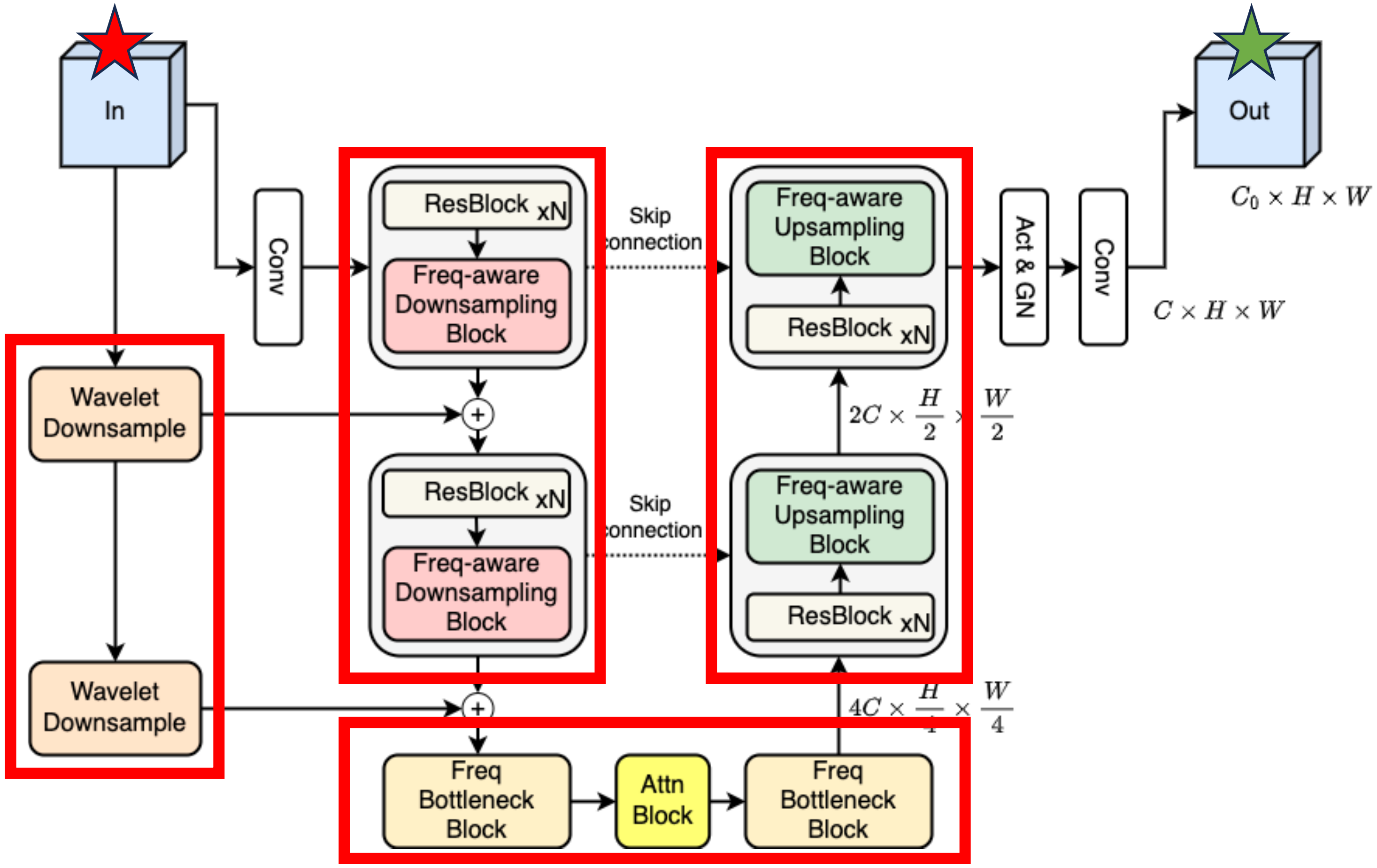
(2) Reconstruction loss: impede the loss of frequency information.

$$\mathcal{L}_{rec} = \ \|y'_0 - y_0\|$$

Our additional loss

Generator loss: $\mathcal{L}^{G} = \mathcal{L}_{adv}^{G} + \lambda\, \mathcal{L}_{rec}$

Wavelet-embedded Generator



1. Wavelet downsample layer
2. Freq-aware up and down block
3. Frequency bottleneck block

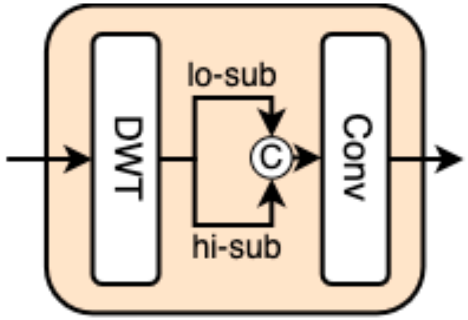
Integrate wavelet information to feature space

Wavelet-embedded Generator

1 Wavelet downsample layer

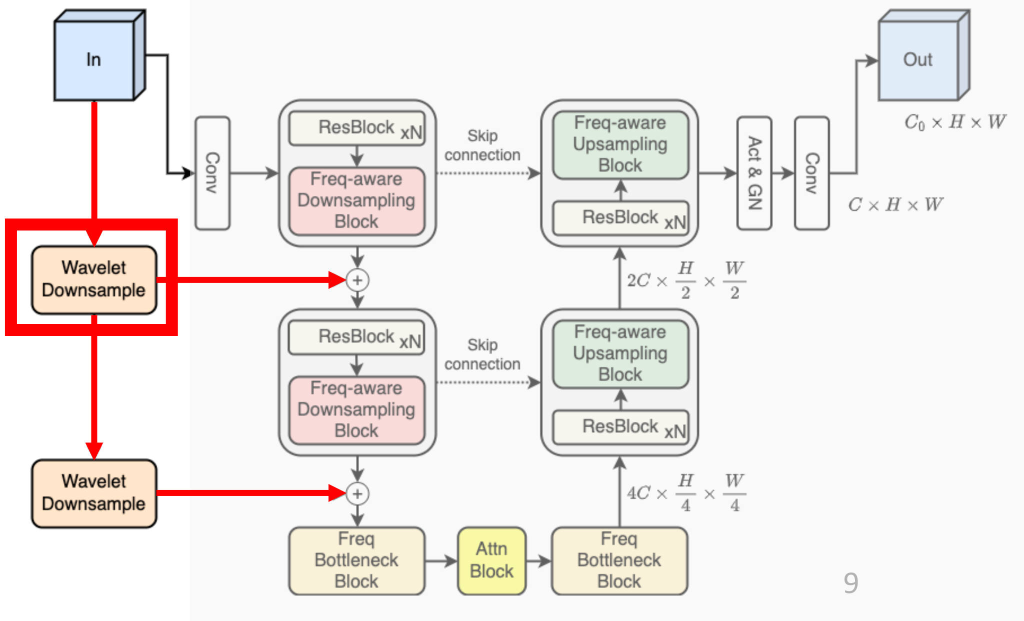
2 Freq-aware up and down block

3 Frequency bottleneck block

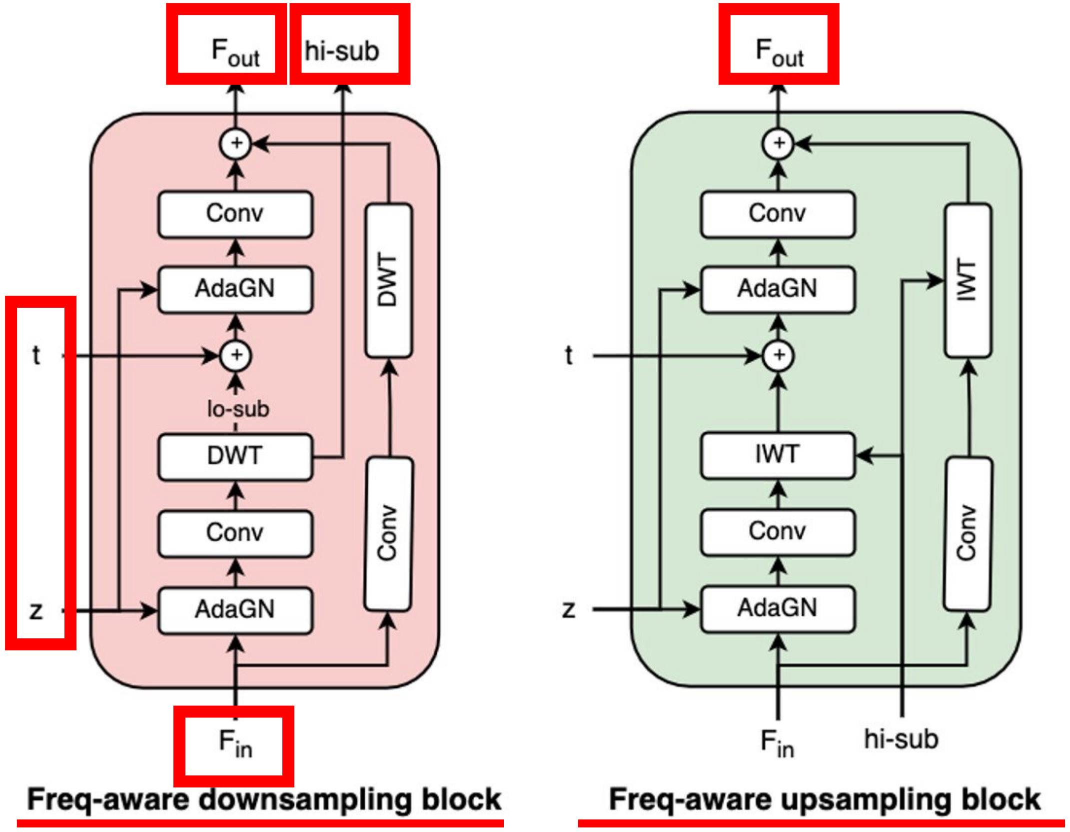


Wavelet Downsample

Inject input signals to feature pyramids.



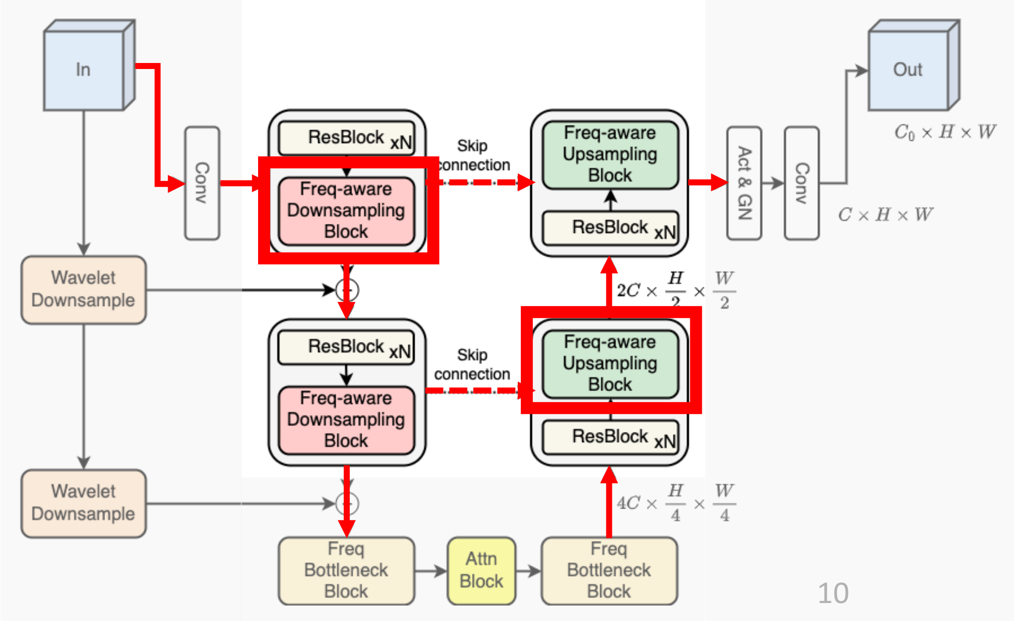
Wavelet-embedded Generator



1 Wavelet downsample layer

2 Freq-aware up and down block

3 Frequency bottleneck block



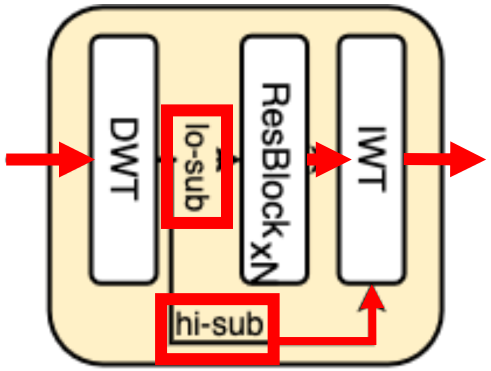
Utilize wavelet transformations for upsampling and downsampling

Wavelet-embedded Generator

1 Wavelet downsample layer

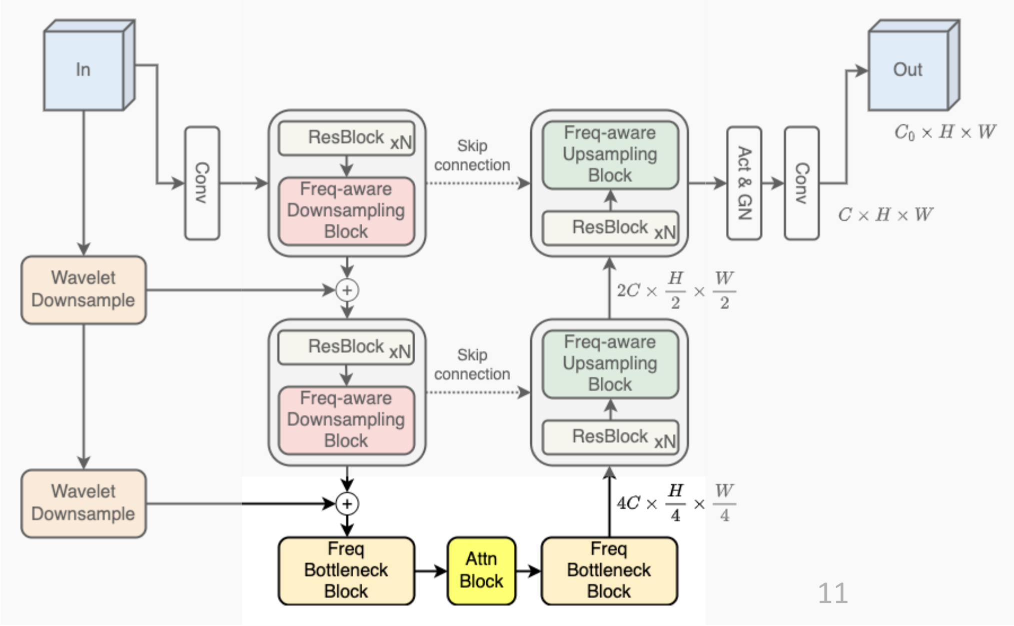
2 Freq-aware up and down block

3 Frequency bottleneck block



Freq Bottleneck Block

Focus on low-frequency subbands while preserving high-frequency details.



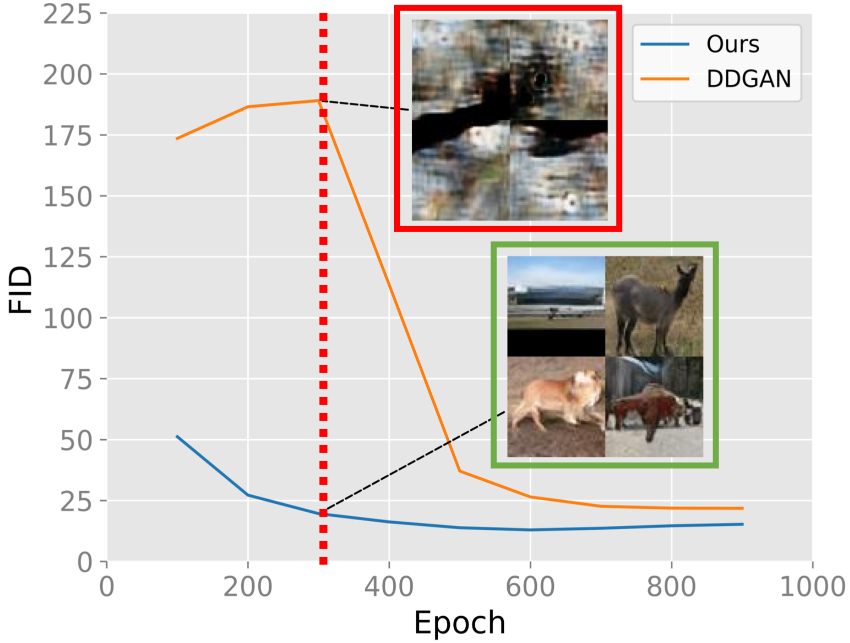
Results – CIFAR10

2.5x faster

Model	FID↓	Recall↑	NFE↓	Time (s)↓
Ours	4.01	0.55	4	**0.08**
DDGAN [50]	**3.75**	0.57	4	0.21 (0.30*)
DDPM [13]	3.21	0.57	1000	80.5
NCSN [42]	25.3	-	1000	107.9
Score SDE (VE) [44]	2.20	0.59	2000	423.2
Score SDE (VP) [44]	2.41	0.59	2000	421.5
DDIM [40]	4.67	0.53	50	4.01
FastDDPM [25]	3.41	0.56	50	4.01
Recovery EBM [8]	9.58	-	180	-
DDPM Distillation [30]	9.36	0.51	1	-
StyleGAN2 w/o ADA [21]	8.32	0.41	1	0.04
StyleGAN2 w/ ADA [19]	2.92	0.49	1	0.04
StyleGAN2 w/ Diffaug [19]	5.79	0.42	1	0.04
Glow [23]	48.9	-	1	-
PixelCNN [33]	65.9	-	1024	-
NVAE [46]	23.5	0.51	1	0.36
VAEBM [49]	12.2	0.53	16	8.79

Results – STL10

Model	FID↓	Recall↑	Time (s)↓
Ours + W-Generator	**12.93**	**0.41**	**0.38**
DDGAN [50]	21.79	0.40	0.58
StyleGAN2 w/o [19]	11.70	0.44	-
StyleGAN2 w/ ADA [57]	13.72	0.36	-
StyleGAN2 + DiffAug [57]	12.97	0.39	-



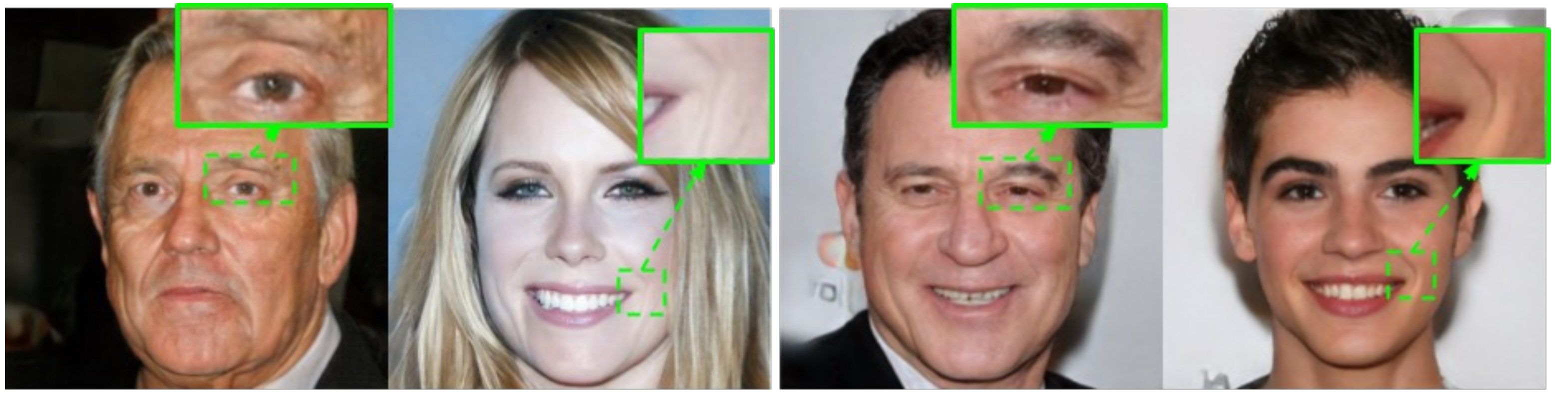
Results – CelebA HQ

Model	FID↓	Recall↑	Time (s)↓
Ours	6.55	0.35	**0.60**
Ours + W-Generator	**5.94**	**0.37**	0.79
DDGAN [50]	7.64	0.36	1.73
Score SDE [44]	7.23	-	-
NVAE [46]	29.7	-	-
VAEBM [49]	20.4	-	-
PGGAN [18]	8.03	-	-
VQ-GAN [6]	10.2	-	-

Model	FID↓	Recall↑	Time (s)↓
CelebA-HQ 512			
Ours + W-Generator	**6.40**	**0.35**	**0.59**
DDGAN [1]	8.43	0.33	1.49
CelebA-HQ 1024			
Ours + W-Generator	**5.98**	**0.39**	**0.59**

256

512



DDGAN Ours

Results - LSUN Church

Model	FID↓	Recall↑	Time (s)↓
Ours + W-Generator	**5.06**	**0.40**	**1.54**
DDGAN [50]	5.25	-	3.42
DDPM [13]	7.89	-	-
ImageBART [5]	7.32	-	-
PGGAN [18]	6.42	-	-
StyleGAN [20]	4.21	-	-
StyleGAN2 [19]	3.86	0.36	-

Ablation on wavelet-embedded generator

Model	FID↓	Time (s)↓
w/o residual	6.25	0.78
w/o up & down	6.23	**0.61**
w/o bottleneck	6.18	0.78
full model	**5.94**	0.79

Alation on CelebA-HQ 256. Each setting is trained for 500 epochs.

Conclusions

Present a novel Wavelet-based Diffusion scheme for efficient sampling.

Integrating wavelet transformations in both pixel and feature space, our method effectively reduces the speed gap with StyleGAN models while delivering competitive benchmarking.

Offer faster training convergence than the baseline.

Facilitate future studies on real-time and high-fidelity diffusion models.

Thank you for your attention!

tienhaophung@gmail.com

Paper