Catch Missing Details: Image Reconstruction with Frequency Augmented Variational Autoencoder

Xinmiao Lin[1], Yikang Li[2], Jenhao Hsiao[2], Chiuman Ho[2], Yu Kong[3]
[1]Rochester Institute of Technology, [2]OPPO US Research, [3]Michigan State University

TUE-AM-165

Summary

Challenges:

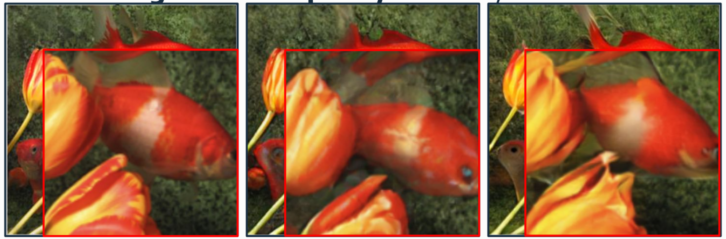
- Reconstruction **deteriorates** with higher compression.
- Features of the middle and higher frequency spectrum are **least recoverable**.

Contributions:

- New model **F**requency **A**ugmented **VAE** (**FA-VAE**) for more accurate details reconstruction.
- New losses **Spectrum Loss (SL)** and **Dynamic Spectrum Loss (DSL)** for learning features of different low/high frequency mixtures.
- New **C**ross-attention **A**utoregressive **T**ransformer (**CAT**) for text-to-image generation with **enhanced attention** mechanism.

Results:

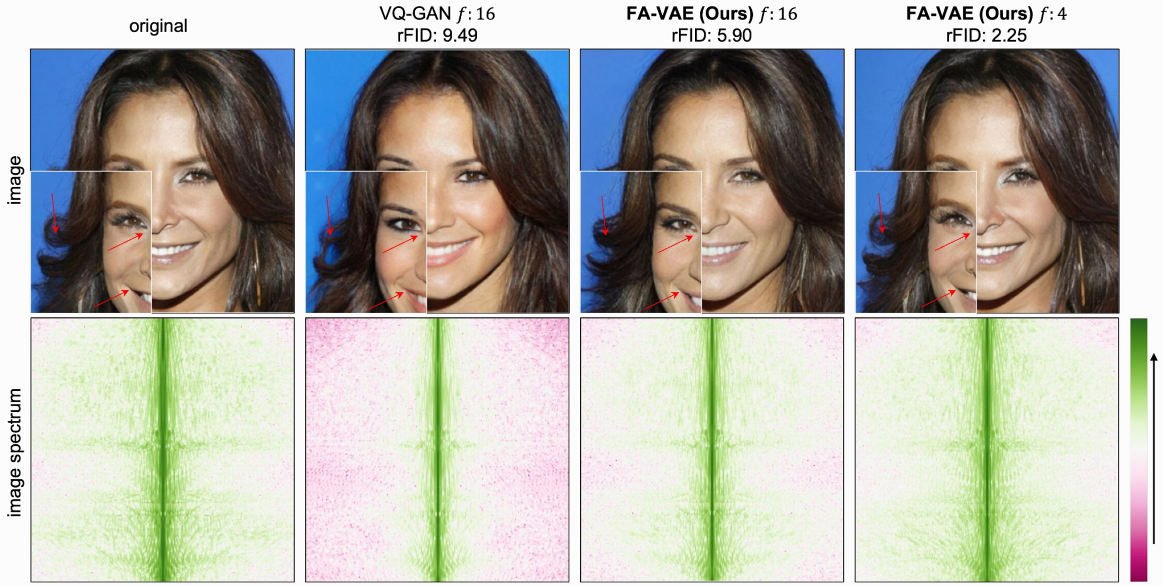
- **FA-VAE improves reconstruction** for various compression rates on several benchmarks.
  - CelebA-HQ, FFHQ, ImageNet
- **CAT yields better generation quality** for T2I synthesis.



original baseline **ours**

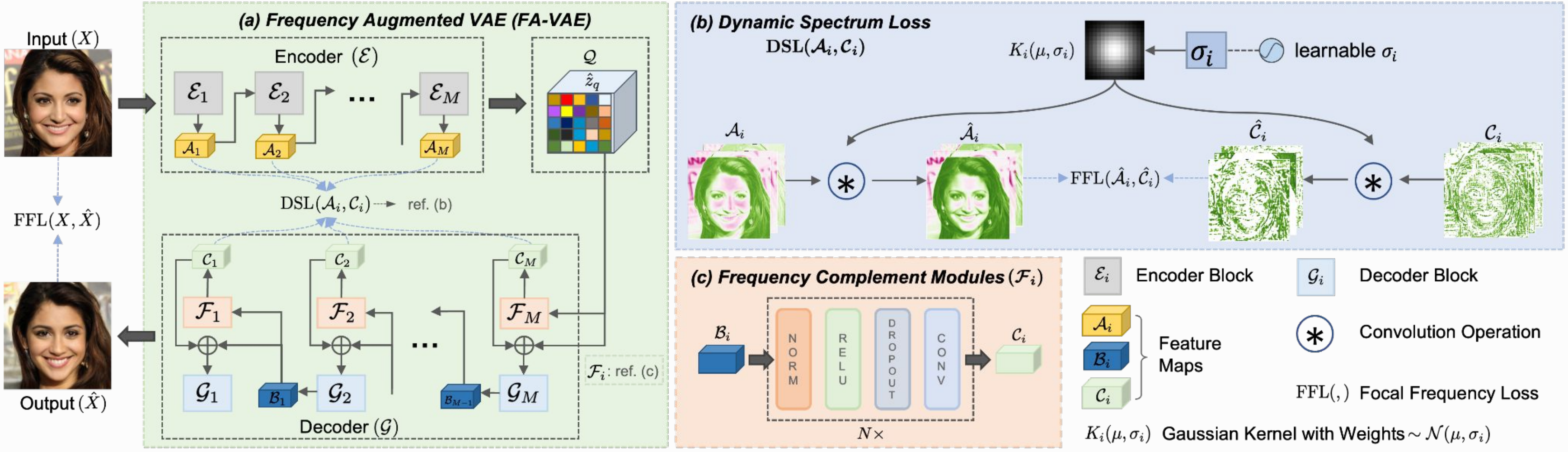
Motivation

- With higher compression rate, **harder to reconstruct** accurately images.
- Features towards middle and higher frequency spectrum are **least recoverable**.
- Existing reconstruction models tend to **ignore alignment** on the frequency spectrum.

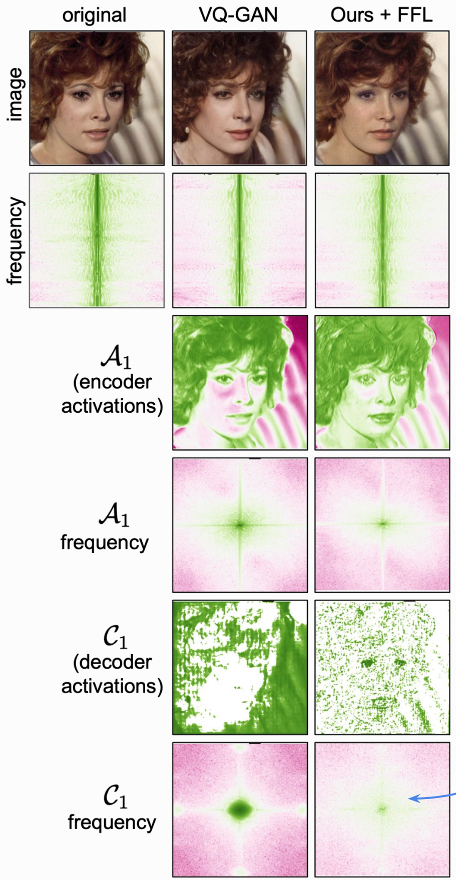


FA-VAE

- **F**requency **A**ugmented **VAE** (**FA-VAE**) learns to complement the reconstructed images with missing features of important frequencies.



Focal Frequency Loss (FFL)



- Focal Frequency Loss (FFL) penalizes the hard frequencies.

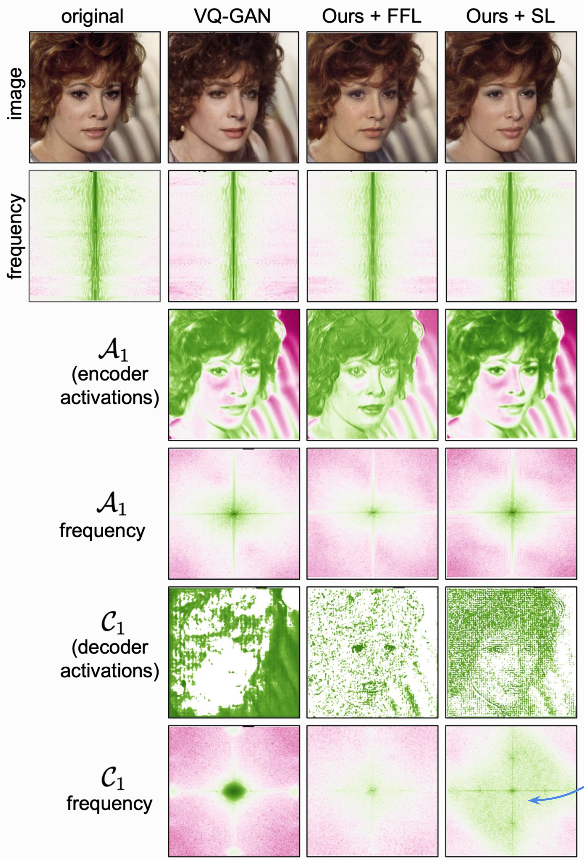
$$\mathrm{FFL}(\mathcal{A}_i, \mathcal{C}_i) = \frac{1}{MN|\mathcal{C}_i|} \sum_{c=0}^{|\mathcal{C}_i|-1} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} w(u,v) J(u,v)$$

encoder activations　　decoder activations　　weights　　frequency distance

  - weights: $w(u,v) = |F_{\mathcal{A}_i}(u,v) - F_{\mathcal{C}_i}(u,v)|$
  - frequency distance: $J(u,v) = |F_{\mathcal{A}_i}(u,v) - F_{\mathcal{C}_i}(u,v)|^2$
  - Discrete Fourier Transform (DFT): $F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cdot e^{-i2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)}$

Noise due to overemphasis on the higher frequency spectrum

Spectrum Loss (SL)



- Penalizes more mismatch in the lower frequency spectrum
  - Because they define the image content
- Diminish the weights towards higher frequency spectrum
  - Details they contain the details
- Apply Gaussian kernels on the activations

$$(\hat{\mathcal{A}}_i, \hat{\mathcal{C}}_i) = (K_i(\mu, \sigma_i) \star \mathcal{A}_i, K_i(\mu, \sigma_i) \star \mathcal{C}_i)$$

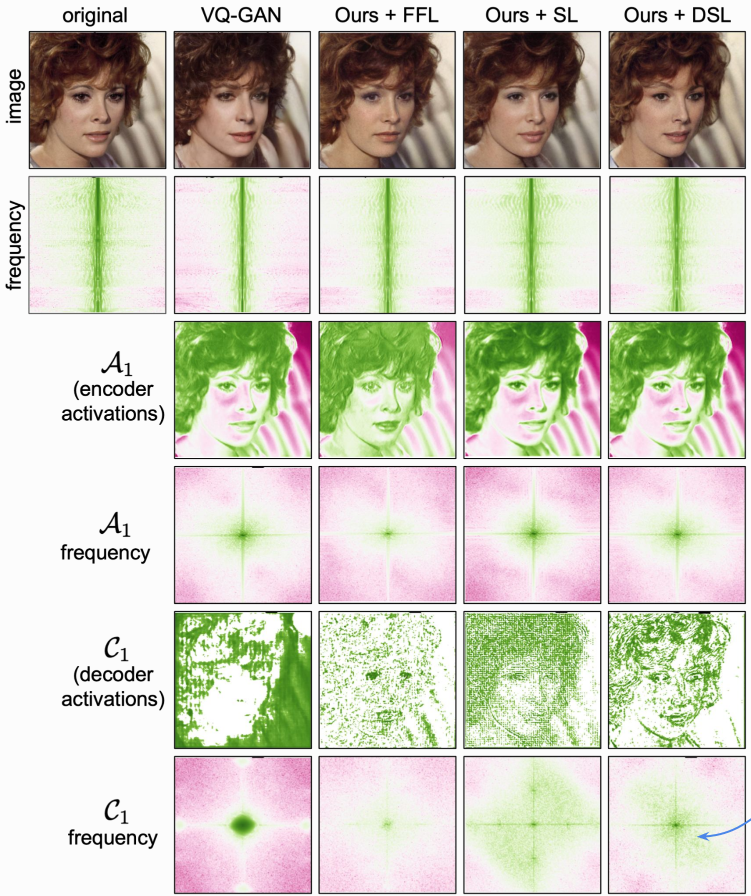
Gaussian Kernels

- **Spectrum Loss (SL)** is defined as:

$$\mathrm{SL}(\mathcal{A}_i, \mathcal{C}_i) = \mathrm{FFL}(\hat{\mathcal{A}}_i, \hat{\mathcal{C}}_i)$$

Better reconstruction on the lower spectrum, checkerboard artifacts due to fixed σ_i

Dynamic Spectrum Loss (DSL)



- Optimize the variances σ_i instead static.
  - Dynamically adjust to different amounts of frequencies needed.
- σ_i are model parameters and optimized as:

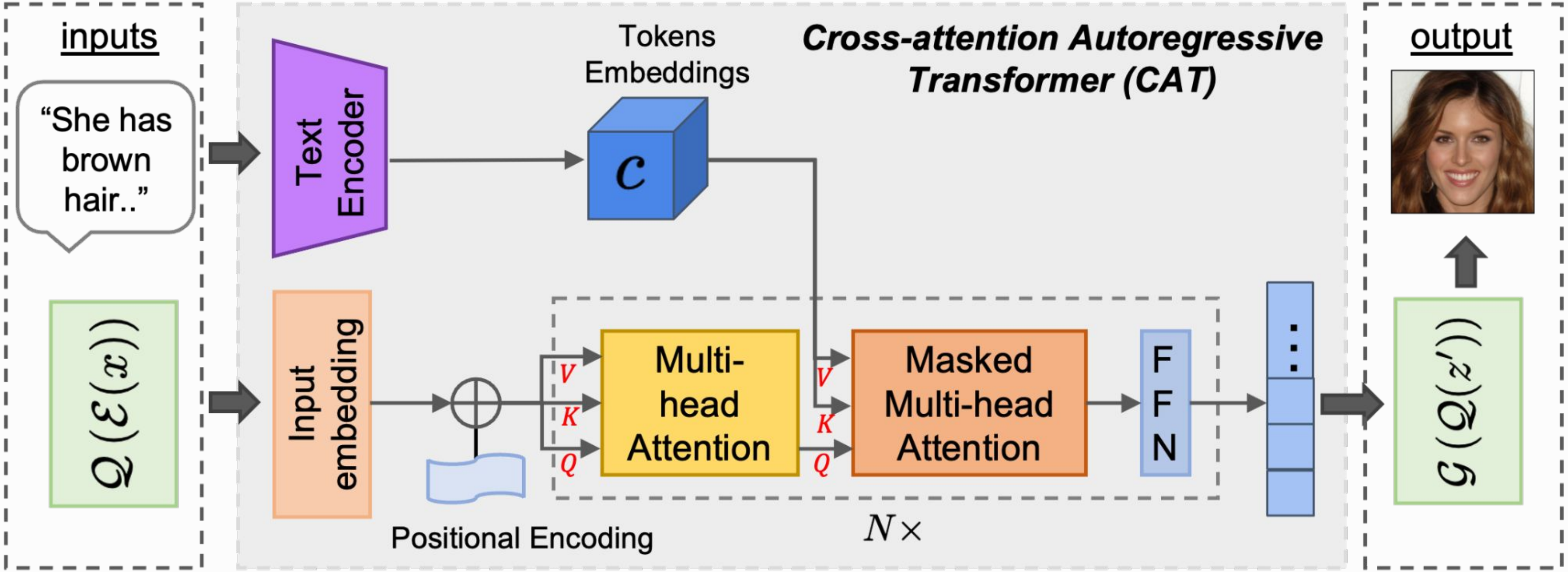
$$\sigma_i^*, \mathcal{E}^*, \mathcal{G}^*, \mathcal{C}^* = \underset{\sigma_i, \mathcal{E}, \mathcal{G}, \mathcal{C}}{\operatorname{argmin}}(\mathcal{L}_{rec} + \mathcal{L}_Q)$$

  - $\mathcal{L}_{rec}$ is the reconstruction loss
  - $\mathcal{L}_Q$ is the quantization loss

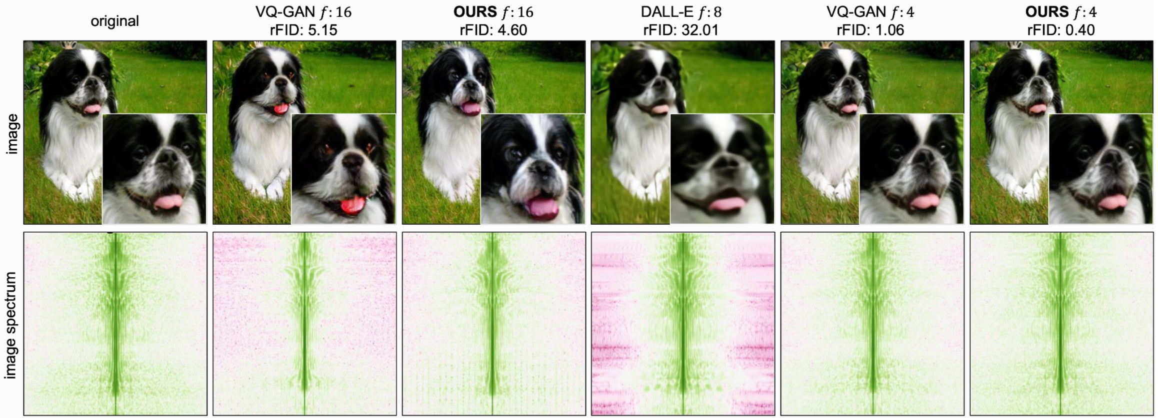
Good balance between low and high frequencies,
No checkerboard artifacts

CAT for T2I

- **C**ross-attention **A**utoregressive **T**ransformer (**CAT**) for text-to-image (T2I) generation task.
  - Uses all token embeddings of a text description for more fine-grained guidance.
  - Embeds cross-attention mechanism to guide generation at each step.



Experiments - Reconstruction



Model	Dataset	Codebook Size	$(h \times w)$	rFID $\downarrow$
RQ-VAE [25]	FFHQ	2048	(8×8)	5.33
FA-VAE (Ours)	FFHQ	2048	(16×16)	4.98
VQ-VAE-2 [39]	ImageNet	512	(64×64) & (32×32)	~ 10 (train)
VQ-GAN [40]	ImageNet	8192	(64×64)	1.06
FA-VAE (Ours)	ImageNet	8192	(64×64)	**0.40**
DALL-E [38]	ImageNet	8192	(32×32)	32.01
VQ-GAN [11]	ImageNet	16384	(16×16)	5.15
VQ-GAN [11]	ImageNet	1024	(16×16)	7.94
VQ-GAN [25]	ImageNet	16384	(8×8)	17.95
RQ-VAE† [46]	ImageNet	16384	(8×8)	10.77
RQ-VAE* [25]	ImageNet	16384	(8×8)	4.73
FA-VAE (Ours)	ImageNet	16384	(16×16)	**4.60**

- FA-VAE gives better reconstruction on different compression rates.
- FA-VAE improves the reconstruction on the frequency spectrum.
- More results in the paper.

Experiments - Generation

Model	FID ↓
AttnGAN [52]	125.98
ControlGAN [26]	116.32
DM-GAN [55]	131.05
DF-GAN [44]	137.60
TediGAN [50]	106.37
LAFITE [54]	12.54
CAT (Ours)	**10.23**

- CAT generates better images for text inputs on CelebA-HQ-MM dataset.
- Images look more realistic.
- More results in the paper.

Thanks

Paper: https://arxiv.org/abs/2305.02541

Code: https://xinmiaolin.github.io/

References

- VQ-GAN: Esser, Patrick et al. “Taming Transformers for High-Resolution Image Synthesis.” *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020): 12868-12878.
- RQ-VAE: Lee, Doyup & Kim, Chiheon & Kim, Saehoon & Cho, Minsu & Han, Wook-Shin. (2022). Autoregressive Image Generation using Residual Quantization.
- VQ-VAE-2: Razavi, Ali et al. “Generating Diverse High-Fidelity Images with VQ-VAE-2.” *Neural Information Processing Systems* (2019).
- DALL-E: Ramesh, Aditya et al. “Zero-Shot Text-to-Image Generation.” *International Conference on Machine Learning* (2021).
- AttnGAN: Xu, Tao & Zhang, Pengchuan & Huang, Qiuyuan & Zhang, Han & Gan, Zhe & Huang, Xiaolei & He, Xiaodong. (2018). AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. 1316-1324. 10.1109/CVPR.2018.00143.
- ControlGAN: Li, Bowen et al. “Controllable Text-to-Image Generation.” *ArXiv* abs/1909.07083 (2019): n. pag.
- DM-GAN: Zhu, Minfeng et al. “DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-To-Image Synthesis.” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019): 5795-5803.
- DF-GAN: Tao, Ming et al. “DF-GAN: Deep Fusion Generative Adversarial Networks for Text-to-Image Synthesis.” *ArXiv* abs/2008.05865 (2020): n. pag.
- TediGAN: Xia, Weihao et al. “TediGAN: Text-Guided Diverse Face Image Generation and Manipulation.” *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021): 2256-2265.
- LAFITE: Zhou, Yufan et al. “Towards Language-Free Training for Text-to-Image Generation.” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021): 17886-17896.