Improved Transformer for High-Resolution GANs

Long Zhao[1] Zizhao Zhang[2] Ting Chen[3] Dimitris Metaxas[1] Han Zhang[3]

[1] Rutgers University [2] Google Cloud AI [3] Google Research

Introduction

- In this paper, we explore how to apply the Transformer to high-resolution image generation based on Generative Adversarial Networks (GANs).

- Challenges:

  - The quadratic scaling problem brought by the self-attention operation becomes even worse when generating pixel-level details for high-resolution images.

  - Generating images from noise inputs poses a higher demand for spatial coherency in structure, color, and texture than discriminative tasks, and hence a more powerful yet efficient self-attention mechanism is desired for decoding feature representations from inputs.

Contributions

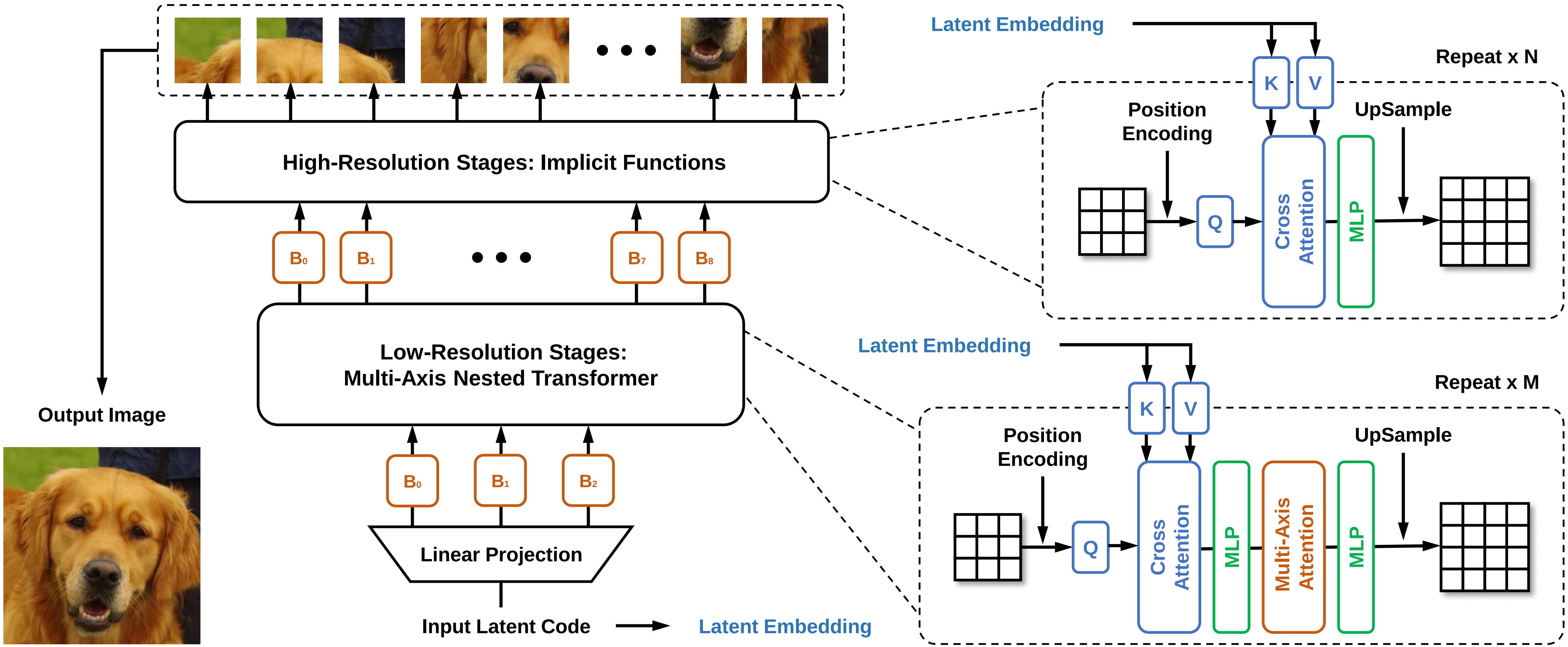
- We propose HiT, a Transformer-based generator for high-fidelity image generation. The resulting architecture easily scales to high-definition image synthesis (with the resolution of 1024 x 1024) and has a comparable throughput to StyleGAN2.

- We present a new form of sparse self-attention operation, namely multi-axis blocked self-attention. It captures local and global dependencies within nonoverlapping image blocks in parallel, each of which uses a half of attention heads.

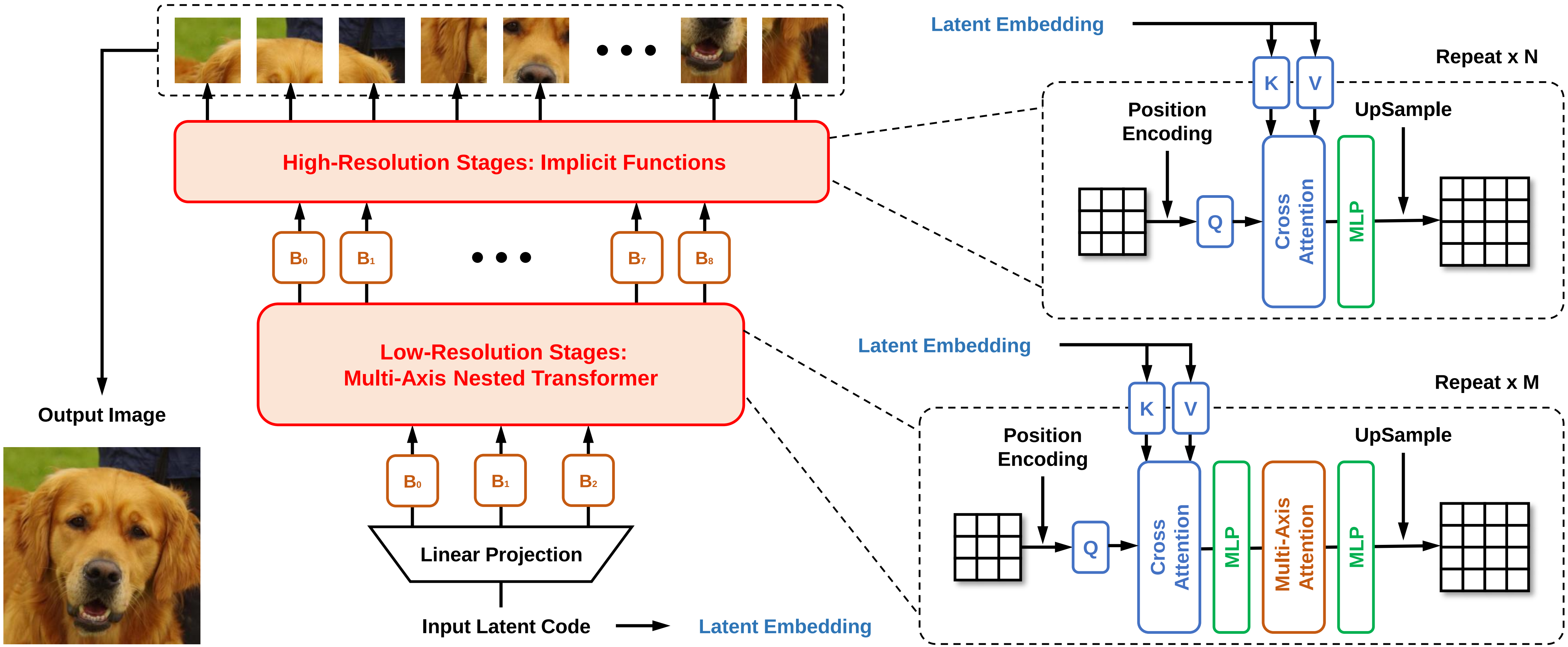
- We introduce a cross-attention module performing attention between the input and intermediate features. This module provides important global information to high-resolution stages where self-attention operations are absent.

- The proposed HiT obtains competitive FID scores of 31.87 and 2.95 on unconditional ImageNet 128 x 128 and FFHQ 256 x 256, respectively, highly reducing the gap between ConvNet-based GANs and Transformer-based ones.

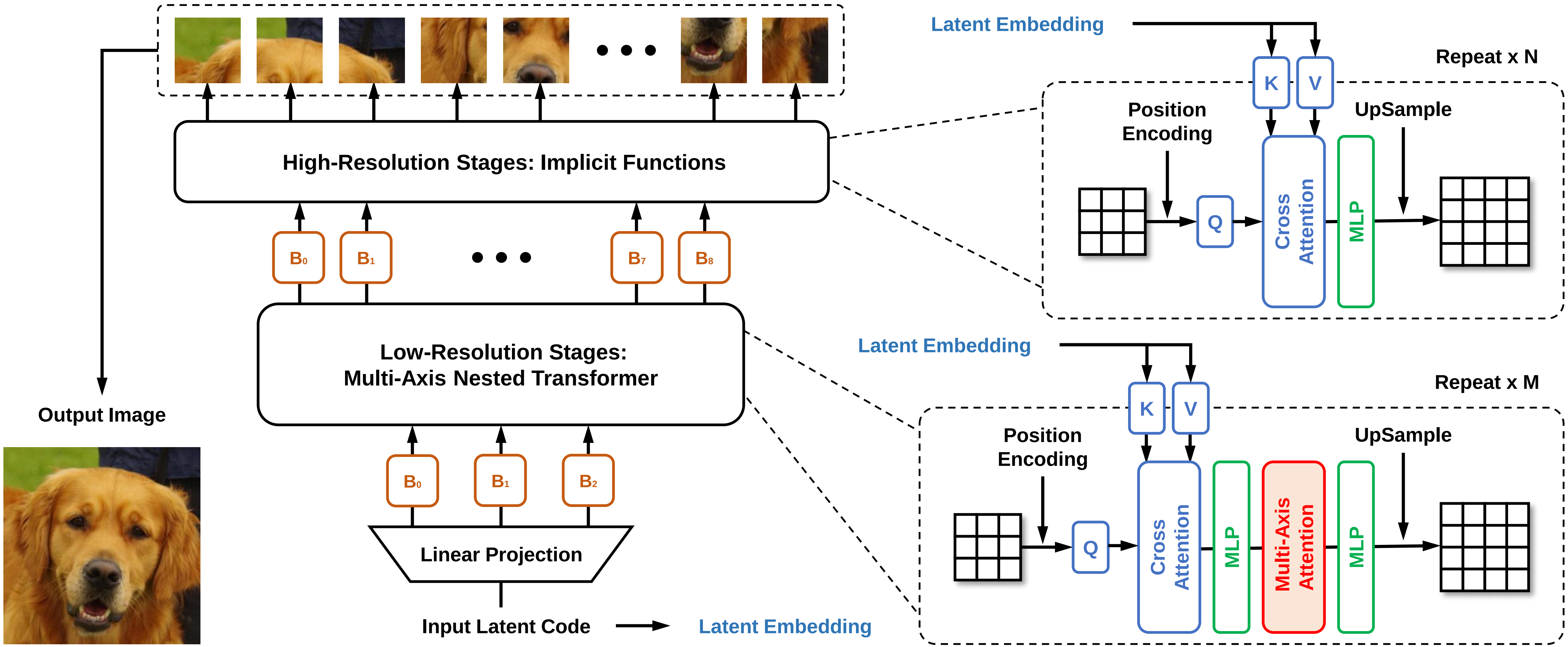
Approach: Main Architecture



Approach: Two-Stage Framework

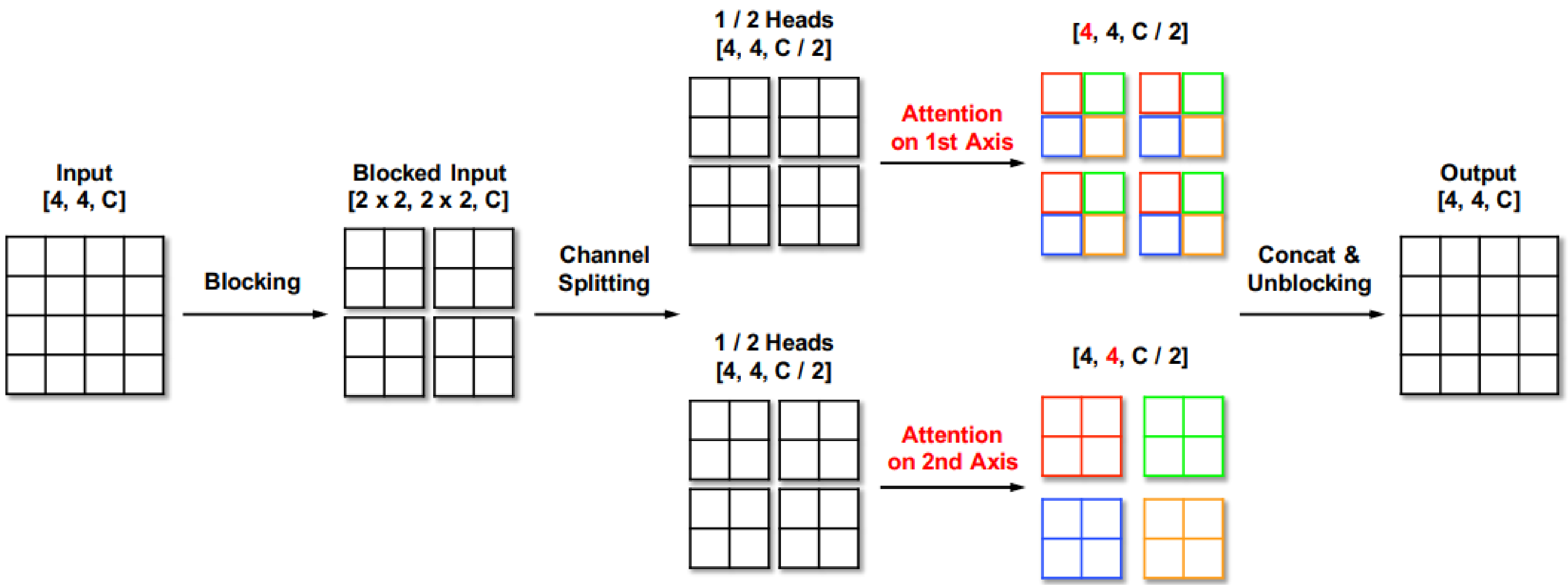


Approach: Multi-Axis Blocked Self-Attention

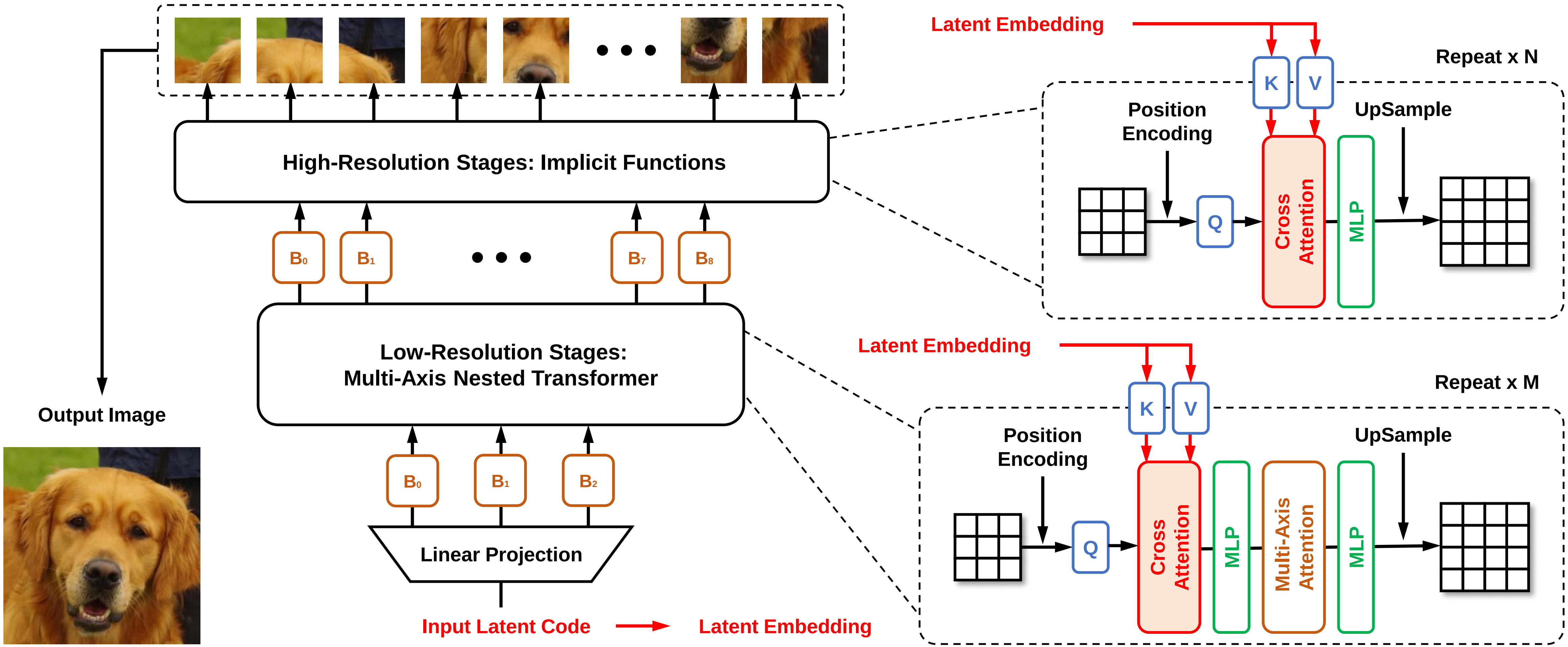


Approach: Multi-Axis Blocked Self-Attention

- The different stages of multi-axis self-attention for a [4, 4, C] input with the block size of b = 2. The input is first blocked into 2 x 2 non-overlapping [2, 2, C] patches. Then regional and dilated self-attention operations are computed along two different axes, respectively, each of which uses a half of attention heads. The attention operations run in parallel for each of the tokens and their corresponding attention regions, illustrated with different colors.



Approach: Cross-Attention for Self-Modulation



Approach: Cross-Attention for Self-Modulation

- Two benefits:

  - Self-modulation stabilizes the generator towards favorable conditioning values and also appears to improve mode coverage.

  - When self-attention modules are absent in high-resolution stages, attending to the input latent code provides an alternative way to capture global information when generating pixel-level details.

Results: ImageNet

- **Left:** Comparison with the state-of-the-art methods on the ImageNet 128 × 128 dataset. † is based on a supervised pre-trained ImageNet classifier.

Method	FID ↓	IS ↑
Vanilla GAN [12]	54.17	14.01
PacGAN2 [30]	57.51	13.50
MGAN [15]	58.88	13.22
Logo-GAN-AE [44]	50.90	14.44
Logo-GAN-RC [44]†	38.41	18.86
SS-GAN (sBN) [7]	43.87	-
Self-Conditioned GAN [31]	40.30	15.82
ConvNet-R_1	39.71	18.61
HiT (Ours)	**31.87**	**21.32**

Results: ImageNet

- **Left:** Comparison with the state-of-the-art methods on the ImageNet 128 × 128 dataset. † is based on a supervised pre-trained ImageNet classifier. **Right:** Reconstruction FID on the ImageNet 256 × 256 dataset. We note that VQVAE-2 utilizes a hierarchical organization of VQ-VAE and thus has two codebooks Z.

Method	FID ↓	IS ↑
Vanilla GAN [12]	54.17	14.01
PacGAN2 [30]	57.51	13.50
MGAN [15]	58.88	13.22
Logo-GAN-AE [44]	50.90	14.44
Logo-GAN-RC [44]†	38.41	18.86
SS-GAN (sBN) [7]	43.87	-
Self-Conditioned GAN [31]	40.30	15.82
ConvNet-R_1	39.71	18.61
HiT (Ours)	**31.87**	**21.32**

Method	Embedding size and $\lvert\mathcal{Z}\rvert$	FID ↓
VQ-VAE [56]	32, 1024	75.19
DALL-E [41]	32, 8192	34.30
VQ-VAE-2 [42]	64, 512 32, 512	10.00
VQGAN [11]	16, 1024	8.00
VQ-HiT (Ours)	16, 1024	**6.37**

Results: Ablation Study

- We start with the INR-based generator [5, 26] conditioned on the input latent code and gradually improve it with the proposed attention components and their variations. O/M denotes “out-of-memory” error: the model cannot be trained for the batch size of one.

	Model configuration	#params (million)	Throughput (images / sec)	FID ↓	IS ↑
	Latent-code conditioned INR decoder [5, 26]	42.68	110.39	56.33	16.19
+	Cross-attention for self-modulation	61.55	82.67	35.94	19.42
+ one of	All-to-all self-attention [58]	67.60	-	O/M	O/M
	Axial attention [14, 60]	67.60	74.21	35.15	19.79
	Blocked local attention [57, 67]	67.60	75.54	33.70	19.96
	Interleaving blocked regional and dilated attention			32.96	20.75
	Multi-axis blocked self-attention (Ours)			32.23	20.96
+	Balancing attention between axes (Full model)	67.60	75.33	**31.87**	**21.32**

References

[5] Bepler et al. “Explicitly disentangling image content from translation and rotation with spatial-VAE”. NeurIPS, 2019.

[26] Kleineberg et al. “Adversarial generation of continuous implicit shape representations”. Eurographics, 2020.

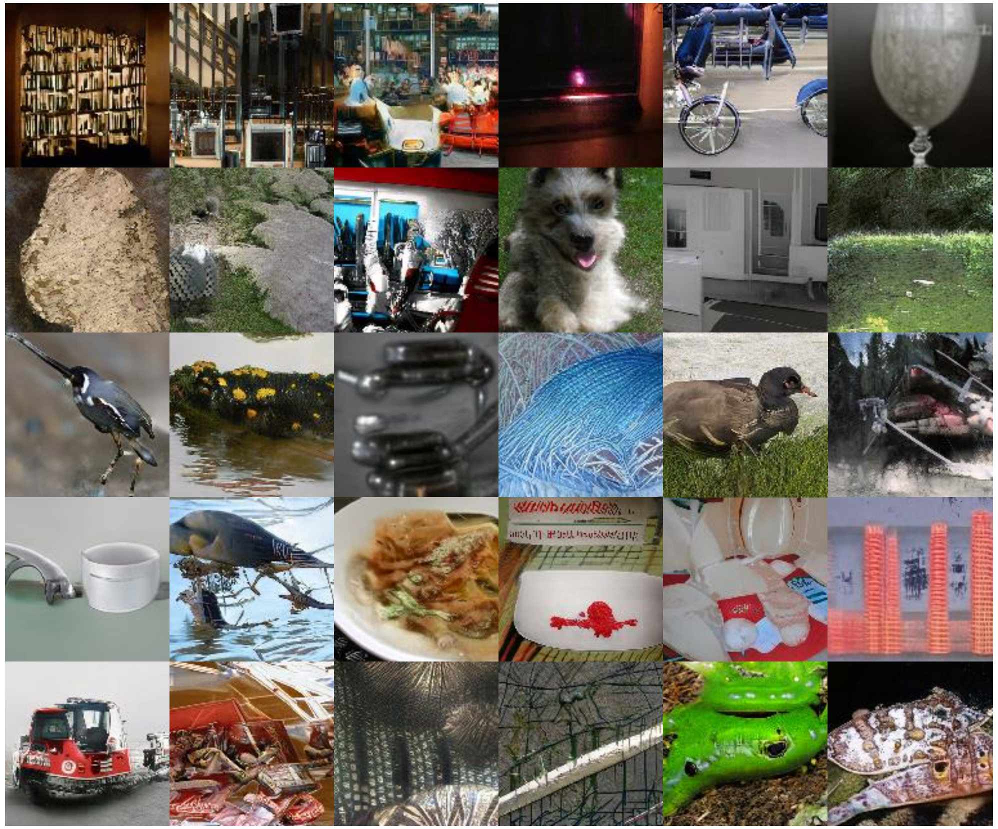
Results: Ablation Study

- Performance as a function of the number of self-attention stages on ImageNet 128 x 128. The attention configuration is defined using the protocol [a, b], where a and b refer to the number of stages in the low-resolution and high-resolution stages of the model, respectively.

Attention configuration	$[0, 5]$	$[1, 4]$	$[2, 3]$	$[3, 2]$	$[4, 1]$
#params (million)	61.55	66.01	67.19	67.52	67.60
Throughput (images / sec)	82.67	80.88	80.22	78.06	75.33
FID $\downarrow$	35.94	34.16	33.69	32.72	31.87

Results: ImageNet 128 x 128

- Uncurated ImageNet 128 × 128 samples from ConvNet-R1 (left, FID: 39.71, IS: 18.61) and the proposed HiT (right, FID: 31.87, IS: 21.32).



Results: Higher Resolution Generation

- Comparison with the state-of-the-art methods on CelebA-HQ (left) and FFHQ (right) with the resolutions of 256 x 256 and 1024 x 1024. bCR [70] is not applied at the 1024 x 1024 resolution.

Method	FID ↓ (CelebA-HQ) $\times 256$	FID ↓ (CelebA-HQ) $\times 1024$
VAEBM [62]	20.38	-
StyleALAE [39]	19.21	-
PG-GAN [21]	8.03	-
COCO-GAN [28]	-	9.49
VQGAN [11]	10.70	-
StyleGAN [23]	-	**5.17**
HiT-B (Ours)	**3.39**	8.83*

Method	FID ↓ (FFHQ) $\times 256$	FID ↓ (FFHQ) $\times 1024$
U-Net GAN [46]	7.63	-
StyleALAE [39]	-	13.09
VQGAN [11]	11.40	-
INR-GAN [50]	9.57	16.32
CIPS [1]	4.38	10.07
StyleGAN2 [24]	3.83	**4.41**
HiT-B (Ours)	**2.95**	6.37*

References

[70] Zhao et al. “Improved consistency regularization for GANs”. AAAI, 2020.

Results: Higher Resolution Generation

- Comparison with the main competing methods in terms of number of network parameters, throughput, and FID on FFHQ 256 x 256. The throughput is measured on a single Tesla V100 GPU.

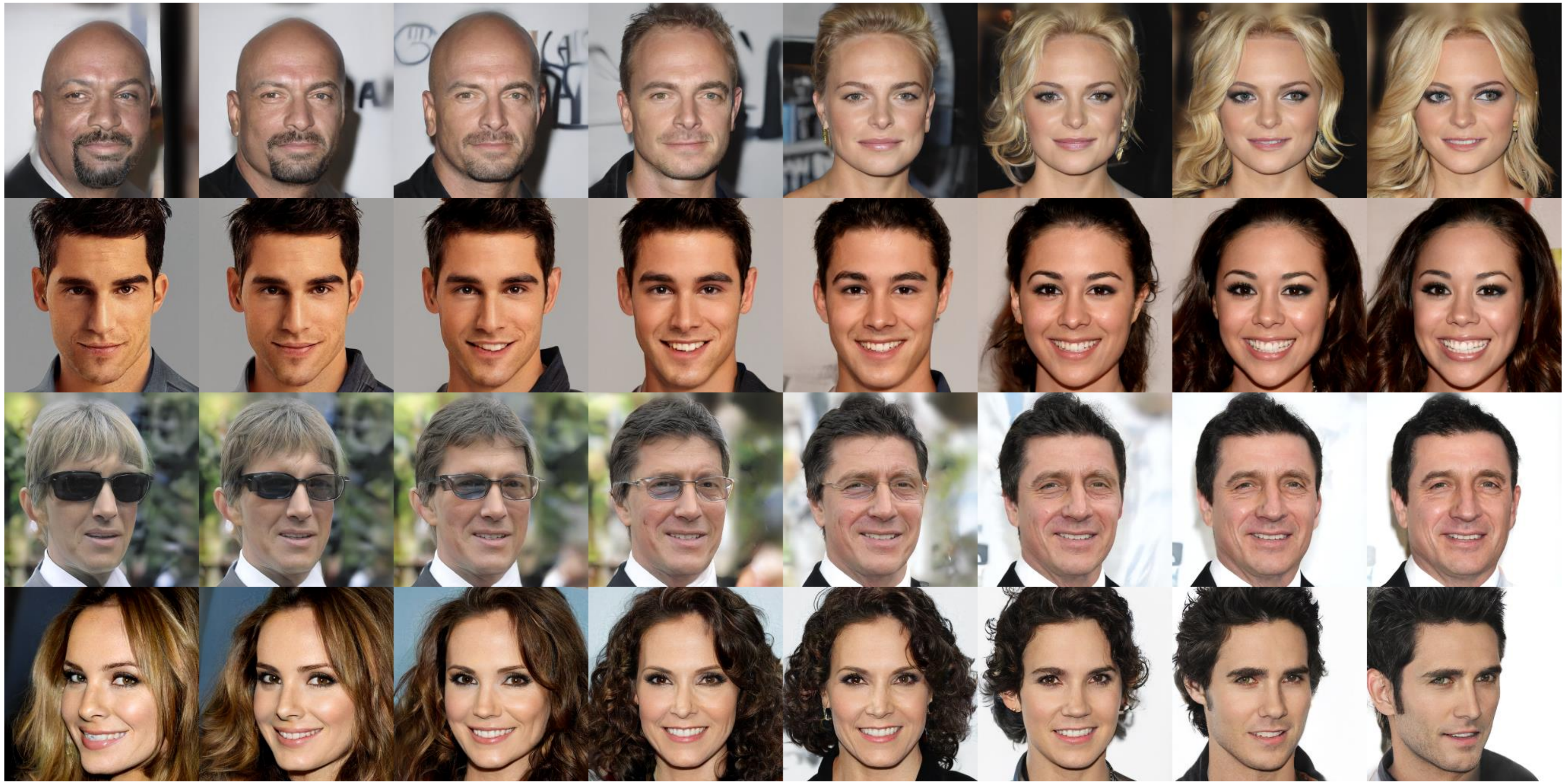
Architecture	Model	#params (million)	Throughput (images / sec)	FID $\downarrow$ (FFHQ $\times$256)
ConvNet	StyleGAN2 [24]	30.03	95.79	3.83
INR	CIPS [1]	45.90	27.27	4.38
	INR-GAN [50]	107.03	266.45	9.57
Transformer	HiT-S	38.01	86.64	3.06
	HiT-B	46.22	52.09	2.95
	HiT-L	97.46	20.67	2.58

Results: CelebA-HQ

- Synthetic face images by HiT-B on CelebA-HQ 1024 x 1024 and 256 x 256.

Results: Latent Interpolation

- Latent linear morphing on the CelebA-HQ 256 x 256 dataset between two synthetic face images – the left-most and right-most ones.



Results: Effectiveness of Regularization

- The effectiveness of bCR [70] on both StyleGAN2 and HiT. † indicates the results of StyleGAN2 are obtained from [22] which uses a lighter-weight configuration of [24].

+ bCR [70]	StyleGAN2 [24]†	HiT-S	HiT-B	HiT-L
✗	5.28	6.07	5.30	5.13
✓	3.91	3.06	2.95	2.58
Δ FID	1.37	3.01	2.35	2.55

References

[22] Karras et al. “Training generative adversarial networks with limited data”. NeurIPS, 2020.

[24] Karras et al. “Analyzing and improving the image quality of StyleGAN”. CVPR, 2020.

[70] Zhao et al. “Improved consistency regularization for GANs”. AAAI, 2020.

Thanks!