Tamimg Transformers for High-Resolution Image Synthesis

Patrick Esser, Robin Rombach, Bjorn Ommer

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021

Outline

1. Introduction
2. Related Work
    - iGPT
    - VQVAE-2
3. Approach
    - Learning an Effective Codebook of Image Constituents
    - Learning the Composition of Images
4. Experiments
    - Attention is all you need in the latent space
    - A unified model for image synthesis tasks
    - Building context-rich vocabularies
    - Quantitative comparison to existing models

1. Introduction

- CNN
  - Strong locality bias and spatial invariance
  - Ineffective for holistic understanding
- Transformer
  - Free to learn complex relationships
  - High computational costs

Do we have to re-learn everything we know about the local structure and regularity of images?

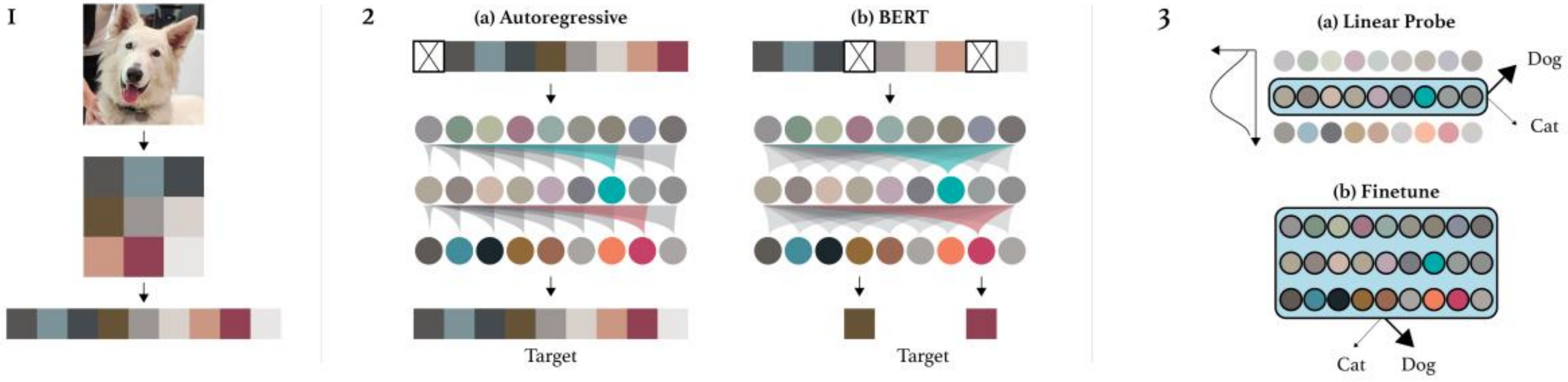
➔ We hypothesize that low-level image structure is well described by a local connectivity, i.e. a convolutional architecture.

1. Introduction

- Use CNNs to learn a context-rich vocabulary of image constituents
- Utilize transformers to efficiently model their composition within high-resolution images

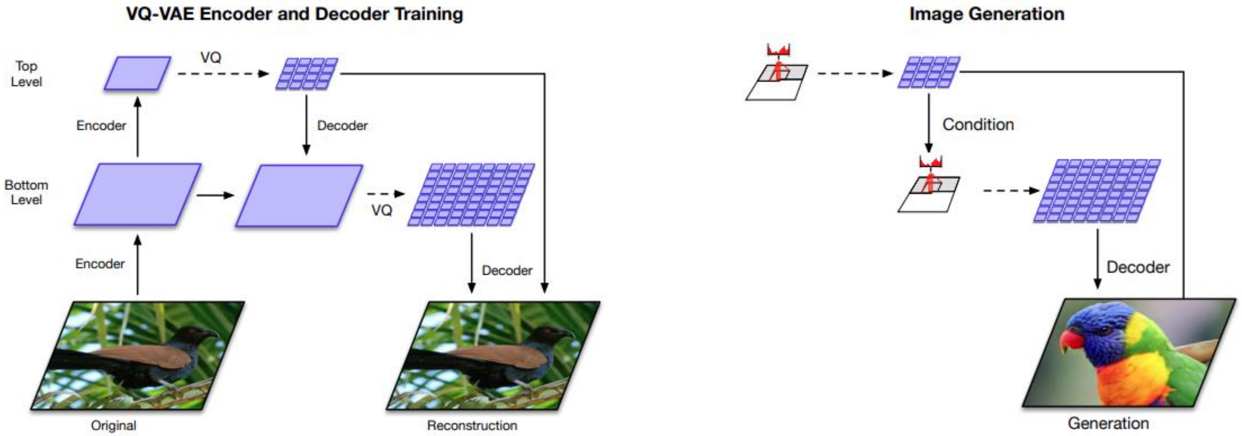
2. Related Work - iGPT

- iGPT
- VQVAE-2



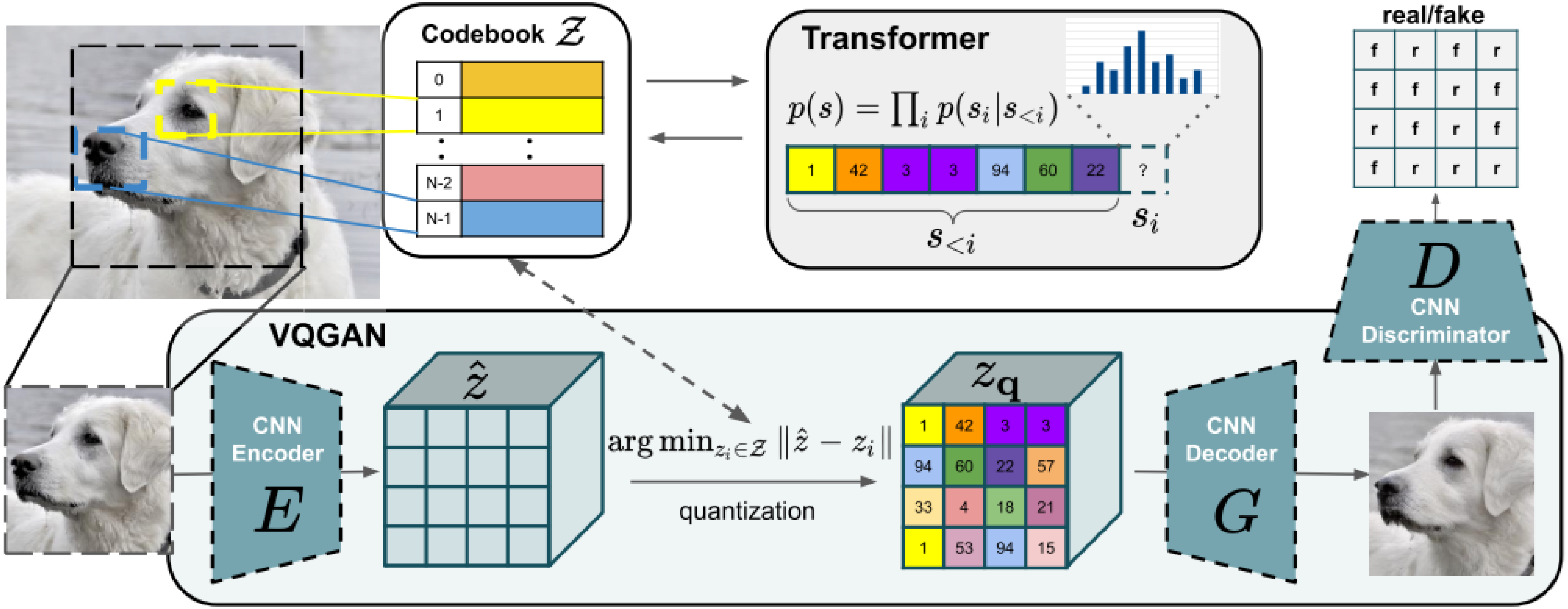
2. Related Work - VQVAE-2

- iGPT
- VQVAE-2



1	1	1	1	1
1	1	1	1	1
1	1	0	0	0
0	0	0	0	0
0	0	0	0	0

3. Approach - Overview

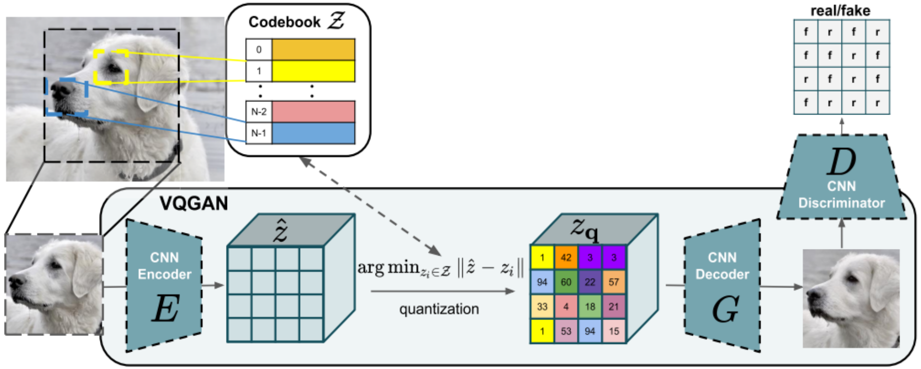


3.1. Learning an Effective Codebook of Image Constituents

- Any image can be represented by a spatial collection of codebook entries

$$x \in \mathbb{R}^{H \times W \times 3} \quad \Rightarrow \quad z_{\mathbf{q}} \in \mathbb{R}^{h \times w \times n_z}$$

- A sequence of **h · w** indices which specify the respective entries in the learned codebook



VQVAE

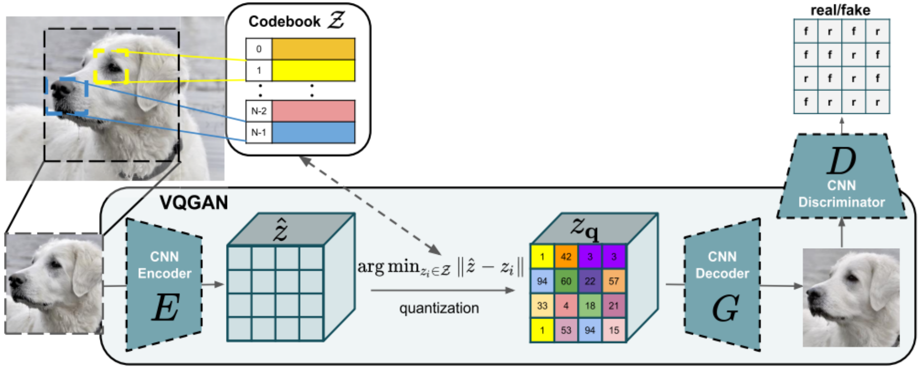
- The encoder **E** and decoder **G** learn to represent images with codes from a discrete codebook

$$z_{\mathbf{q}} = \mathbf{q}(\hat{z}) := \left(\underset{z_k \in \mathcal{Z}}{\arg\min} \|\hat{z}_{ij} - z_k\| \right) \in \mathbb{R}^{h \times w \times n_z}$$

$$\boxed{\hat{z} = E(x) \in \mathbb{R}^{h \times w \times n_z}}$$

$$\boxed{\hat{z}_{ij} \in \mathbb{R}^{n_z}}$$

$$\boxed{\mathcal{Z} = \{z_k\}_{k=1}^{K} \subset \mathbb{R}^{n_z}}$$



VQVAE

- The reconstruction is given by

$$\hat{x} = G(z_{\mathbf{q}}) = G\left(\mathbf{q}(E(x))\right)$$

- The model and codebook can be trained via the loss function

$$\mathcal{L}_{\mathrm{VQ}}(E, G, \mathcal{Z}) = \|x - \hat{x}\|^2 + \|\mathrm{sg}[E(x)] - z_{\mathbf{q}}\|_2^2 + \beta\|\mathrm{sg}[z_{\mathbf{q}}] - E(x)\|_2^2$$

VQGAN

- We propose VQGAN which uses a discriminator and perceptual loss to learn a perceptually rich codebook

$$\mathcal{L}_{\mathrm{GAN}}(\{E, G, \mathcal{Z}\}, D) = [\log D(x) + \log(1 - D(\hat{x}))]$$

- The complete objective for finding the optimal model

$$\mathcal{Q}^* = \underset{E,G,\mathcal{Z}}{\arg\min}\ \max_{D}\ \mathbb{E}_{x\sim p(x)}\Big[\mathcal{L}_{\mathrm{VQ}}(E, G, \mathcal{Z}) + \lambda \mathcal{L}_{\mathrm{GAN}}(\{E, G, \mathcal{Z}\}, D)\Big]$$

$$\boxed{\lambda = \frac{\nabla_{G_L}[\mathcal{L}_{\mathrm{rec}}]}{\nabla_{G_L}[\mathcal{L}_{\mathrm{GAN}}] + \delta}}$$

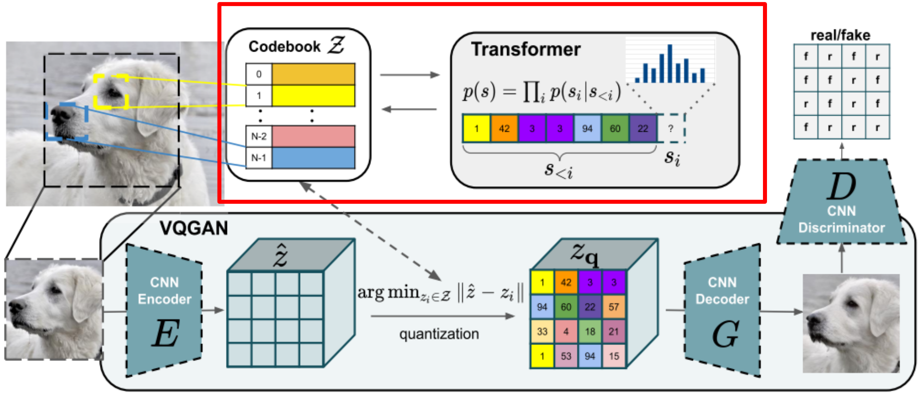
3.2. Learning the Composition of Images

- A sequence **s** of indices from the codebook, which is obtained by replacing each code by its index in the codebook **Z**

$$s_{ij} = k \text{ such that } (z_{\mathbf{q}})_{ij} = z_k$$

$$\boxed{s \in \{0, \ldots, |\mathcal{Z}|-1\}^{h \times w}}$$

$$\boxed{z_{\mathbf{q}} = \mathbf{q}(E(x)) \in \mathbb{R}^{h \times w \times n_z}}$$



3.2. Learning the Composition of Images

- Image-generation can be formulated as autoregressive next-index prediction

$$p(s) = \prod_i p(s_i | s_{<i})$$

- Maximize the log-likelihood

$$\mathcal{L}_{\text{Transformer}} = \mathbb{E}_{x \sim p(x)} \left[-\log p(s) \right]$$

Conditioned Synthesis

- The task is then to learn the likelihood of the sequence given this information **c**

$$p(s|c) = \prod p(s_i|s_{<i}, c)$$

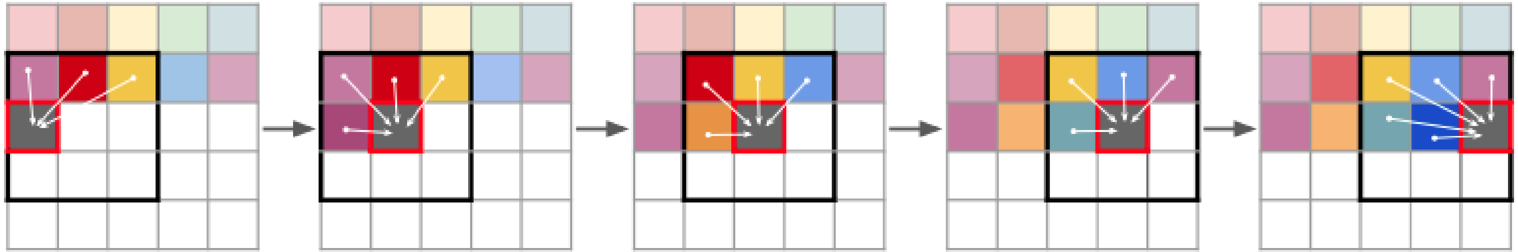
- We first learn another VQGAN to obtain again an index-based representation **r**, then simply prepend **r** to **s**

$$p(s_i|s_{<i}, r)$$

$$\boxed{r \in \{0, \ldots, |\mathcal{Z}_c|-1\}^{h_c \times w_c}}$$

Generating High-Resolution Images

- We have to work patch-wise and crop images to restrict the length of s to a maximally feasible size during training
- Unconditional image synthesis on aligned data, we can simply condition on image coordinates

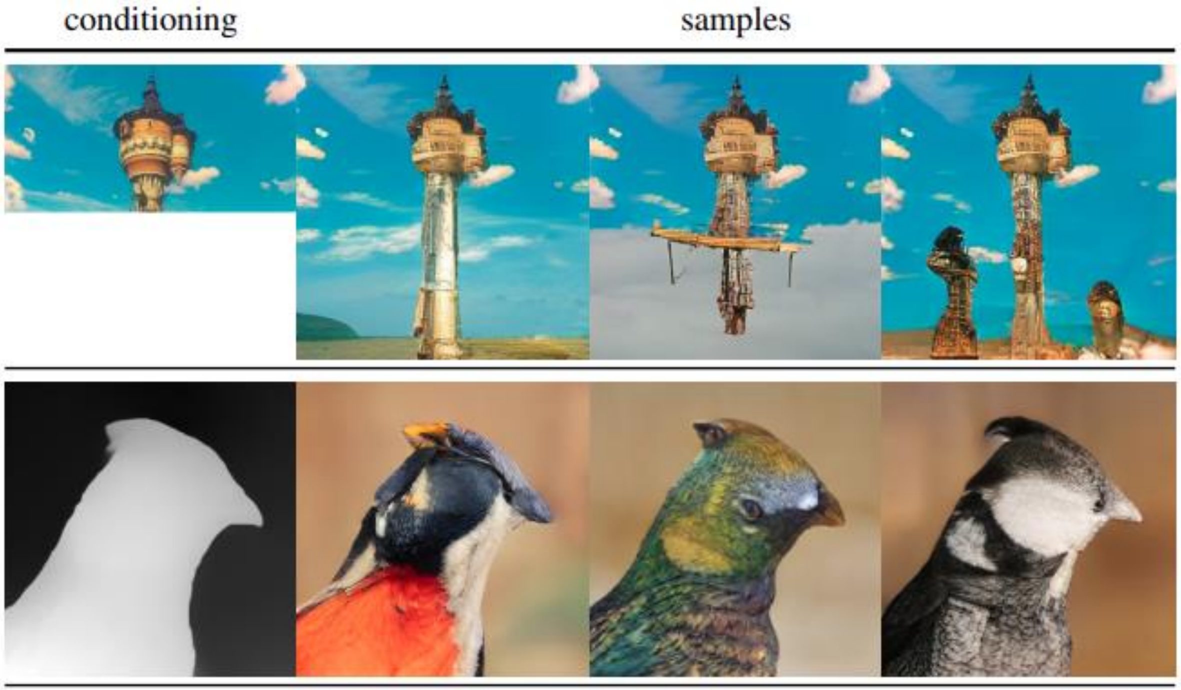


4.1. Attention Is All You Need in the Latent Space

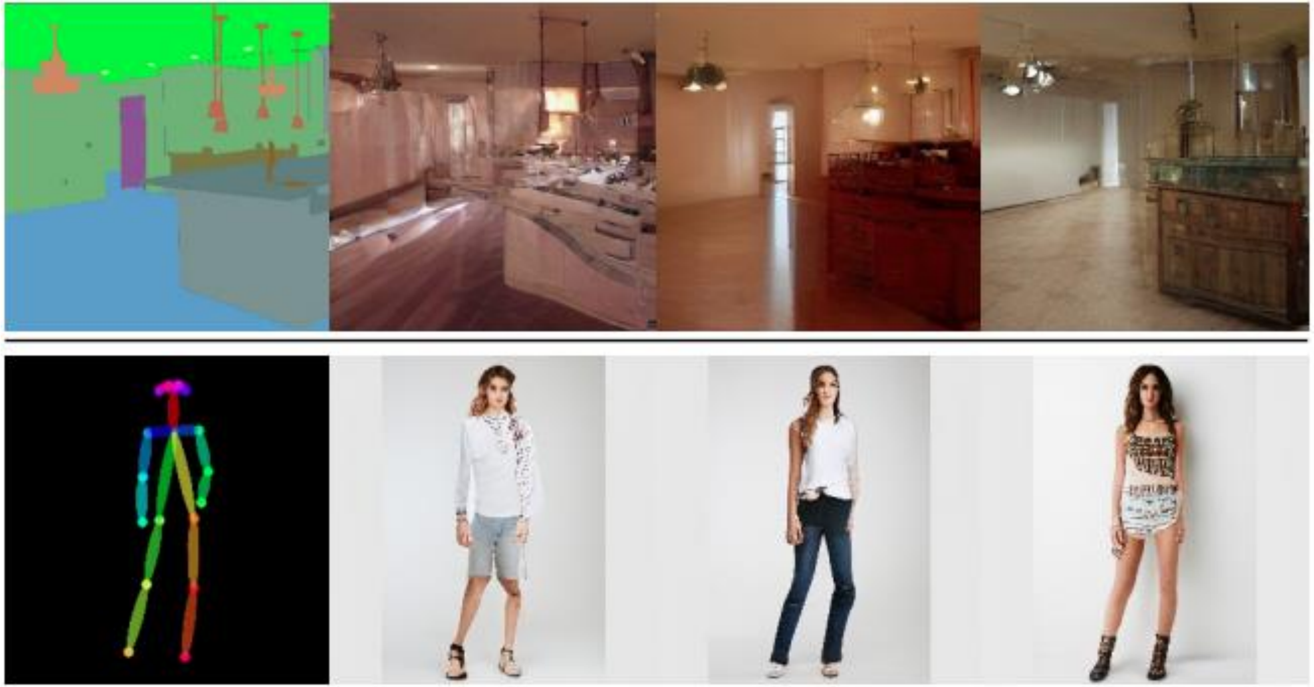
Comparing Transformer and PixelSNAIL architectures across different datasets and model sizes

	Negative Log-Likelihood (NLL)		
Data / # params	Transformer *P-SNAIL steps*	Transformer *P-SNAIL time*	PixelSNAIL *fixed time*
RIN / 85M	**4.78**	4.84	4.96
LSUN-CT / 310M	**4.63**	4.69	4.89
IN / 310M	**4.78**	4.83	4.96
D-RIN / 180 M	**4.70**	4.78	4.88
S-FLCKR / 310 M	**4.49**	4.57	4.64

4.2. A Unified Model for Image Synthesis Tasks



4.2. A Unified Model for Image Synthesis Tasks



4.2. A Unified Model for Image Synthesis Tasks

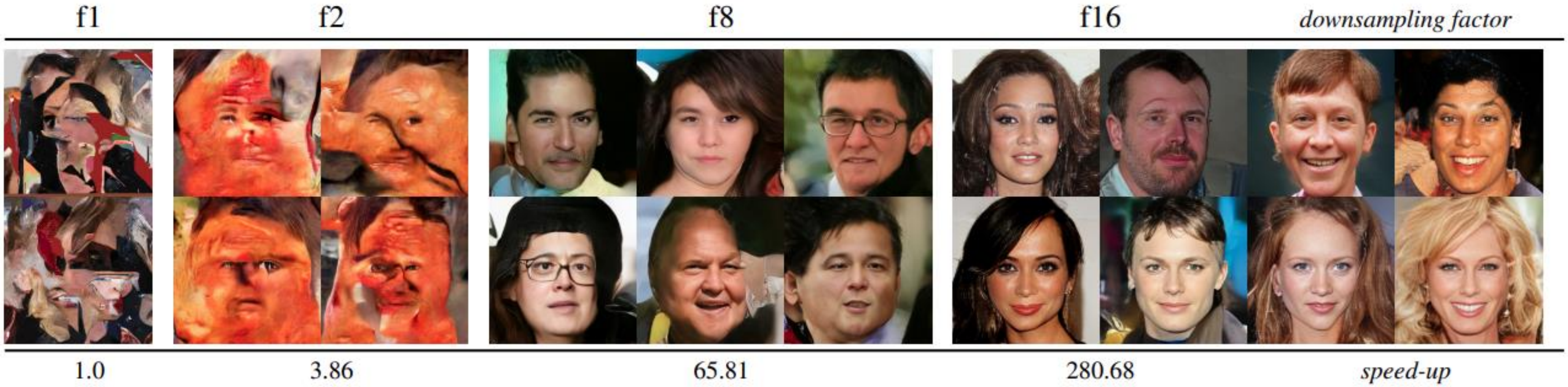
4.2. A Unified Model for Image Synthesis Tasks

4.2. A Unified Model for Image Synthesis Tasks

4.3. Building Context-Rich Vocabularies

How important are context-rich vocabularies?

- Encode images of size H × W into discrete codes of size H/f × W/f



4.4. Quantitative Comparison to Existing Models

Semantic synthesis

Dataset	**ours**	SPADE [46]	Pix2PixHD (+aug) [65]	CRN [9]
COCO-Stuff	22.4	22.6/23.9(*)	111.5 (54.2)	70.4
ADE20K	35.5	33.9/35.7(*)	81.8 (41.5)	73.3

4.4. Quantitative Comparison to Existing Models

Unconditional face synthesis

CelebA-HQ 256×256		FFHQ 256×256	
Method	FID $\downarrow$	Method	FID $\downarrow$
GLOW [33]	69.0	VDVAE ($t = 0.7$) [11]	38.8
NVAE [60]	40.3	VDVAE ($t = 1.0$)	33.5
PIONEER (B.) [21]	39.2 (25.3)	VDVAE ($t = 0.8$)	29.8
NCPVAE [1]	24.8	VDVAE ($t = 0.9$)	28.5
VAEBM [67]	20.4	*VQGAN*+P.SNAIL	21.9
Style ALAE [49]	19.2	BigGAN	12.4
DC-VAE [47]	15.8	**ours**	11.4
ours	10.7	U-Net GAN (+aug) [58]	10.9 (7.6)
PGGAN [27]	8.0	StyleGAN2 (+aug) [30]	3.8 (3.6)