Taming Transformers for High-Resolution Image Synthesis

Patrick Esser* Robin Rombach* Björn Ommer
Heidelberg Collaboratory for Image Processing, IWR, Heidelberg University, Germany
*Both authors contributed equally to this work

Figure 1. Our approach enables transformers to synthesize high-resolution images like this one, which contains 1280x460 pixels.

Abstract

Designed to learn long-range interactions on sequential data, transformers continue to show state-of-the-art results on a wide variety of tasks. In contrast to CNNs, they contain no inductive bias that prioritizes local interactions. This makes them expressive, but also computationally infeasible for long sequences, such as high-resolution images. We demonstrate how combining the effectiveness of the inductive bias of CNNs with the expressivity of transformers enables them to model and thereby synthesize high-resolution images. We show how to (i) use CNNs to learn a context-rich vocabulary of image constituents, and in turn (ii) utilize transformers to efficiently model their composition within high-resolution images. Our approach is readily applied to conditional synthesis tasks, where both non-spatial information, such as object classes, and spatial information, such as segmentations, can control the generated image. In particular, we present the first results on semantically-guided synthesis of megapixel images with transformers. Project page at https://git.io/JLlvY.

1. Introduction

Transformers are on the rise—they are now the de-facto standard architecture for language tasks [64, 50, 51, 5] and are increasingly adapted in other areas such as audio [12] and vision [8, 15]. In contrast to the predominant vision architecture, convolutional neural networks (CNNs), the transformer architecture contains no built-in inductive prior on the locality of interactions and is therefore free to learn complex relationships among its inputs. However, this generality also implies that it *has to* learn all relationships, whereas CNNs have been designed to exploit prior knowledge about strong local correlations within images. Thus, the increased expressivity of transformers comes with quadratically increasing computational costs, because all pairwise interactions are taken into account. The resulting energy and time requirements of state-of-the-art transformer models thus pose fundamental problems for scaling them to high-resolution images with millions of pixels.

Observations that transformers tend to learn convolutional structures [15] thus beg the question: Do we have to re-learn everything we know about the local structure and regularity of images from scratch each time we train a vision model, or can we efficiently encode inductive image biases while still retaining the flexibility of transformers? We hypothesize that low-level image structure is well described by a local connectivity, i.e. a convolutional architecture, whereas this structural assumption ceases to be effective on higher semantic levels. Moreover, CNNs not only exhibit a strong locality bias, but also a bias towards

spatial invariance through the use of shared weights across all positions. This makes them ineffective if a more holistic understanding of the input is required.

Our key insight to obtain an effective and expressive model is that, *taken together, convolutional and transformer architectures can model the compositional nature of our visual world [44]*: We use a convolutional approach to efficiently learn a codebook of context-rich visual parts and, subsequently, learn a model of their global compositions. The long-range interactions within these compositions require an expressive transformer architecture to model distributions over their consituent visual parts. Furthermore, we utilize an adversarial approach to ensure that the dictionary of local parts captures perceptually important local structure to alleviate the need for modeling low-level statistics with the transformer architecture. Allowing transformers to concentrate on their unique strength — modeling long-range relations — enables them to generate high-resolution images as in Fig. 1, a feat which previously has been out of reach. Our formulation directly gives control over the generated images by means of conditioning information regarding desired object classes or spatial layouts. Finally, experiments demonstrate that our approach retains the advantages of transformers by outperforming previous codebook-based state-of-the-art approaches based on convolutional architectures.

2. Related Work

The Transformer Family The defining characteristic of the transformer architecture [64] is that it models interactions between its inputs solely through attention [2, 32, 45] which enables them to faithfully handle interactions between inputs regardless of their relative position to one another. Originally applied to language tasks, inputs to the transformer were given by tokens, but other signals, such as those obtained from audio [37] or images [8], can be used. Each layer of the transformer then consists of an attention mechanism, which allows for interaction between inputs at different positions, followed by a position-wise fully connected network, which is applied to all positions independently. More specifically, the (self-)attention mechanism can be described by mapping an intermediate representation with three position-wise linear layers into three representations, query $Q \in \mathbb{R}^{N \times d_k}$, key $K \in \mathbb{R}^{N \times d_k}$ and value $V \in \mathbb{R}^{N \times d_v}$, to compute the output as

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^t}{\sqrt{d_k}}\right)V \in \mathbb{R}^{N \times d_v}. \quad (1)$$

When performing autoregressive maximum-likelihood learning, non-causal entries of QK^t, i.e. all entries below its diagonal, are set to $-\infty$ and the final output of the transformer is given after a linear, point-wise transformation to predict logits of the next sequence element. Since the attention mechanism relies on the computation of inner products between all pairs of elements in the sequence, its computational complexity increases quadratically with the sequence length. While the ability to consider interactions between *all* elements is the reason transformers efficiently learn long-range interactions, it is also the reason transformers quickly become infeasible, especially on images, where the sequence length itself scales quadratically with the resolution. Different approaches have been proposed to reduce the computational requirements to make transformers feasible for longer sequences. [48] and [66] restrict the receptive fields of the attention modules, which reduces the expressivity and, especially for high-resolution images, introduces unjustified assumptions on the independence of pixels. [12] and [24] retain the full receptive field but can reduce costs for a sequence of length n only from n^2 to $n\sqrt{n}$, which makes resolutions beyond 64 pixels still prohibitively expensive.

Convolutional Approaches The two-dimensional structure of images suggests that local interactions are particularly important. CNNs exploit this structure by restricting interactions between input variables to a local neighborhood defined by the kernel size of the convolutional kernel. Applying a kernel thus results in costs that scale linearly with the overall sequence length (the number of pixels in the case of images) and quadratically in the kernel size, which, in modern CNN architectures, is often fixed to a small constant such as 3×3. This inductive bias towards local interactions thus leads to efficient computations, but the wide range of specialized layers which are introduced into CNNs to handle different synthesis tasks [46, 70, 59, 74, 73] suggest that this bias is often too restrictive.

Convolutional architectures have been used for autoregressive modeling of images [61, 62, 10] but, for low-resolution images, previous works [48, 12, 24] demonstrated that transformers consistently outperform their convolutional counterparts. Our approach allows us to efficiently model high-resolution images with transformers while retaining their advantages over state-of-the-art convolutional approaches.

Two-Stage Approaches Closest to ours are two-stage approaches which first learn an encoding of data and afterwards learn, in a second stage, a probabilistic model of this encoding. [13] demonstrated both theoretical and empirical evidence on the advantages of first learning a data representation with a Variational Autoencoder (VAE) [34, 54], and then again learning its distribution with a VAE. [17, 68] demonstrate similar gains when using an unconditional normalizing flow for the second stage, and [55, 56] when using a conditional normalizing flow. To improve training efficiency of Generative Adversarial Networks (GANs), [39]

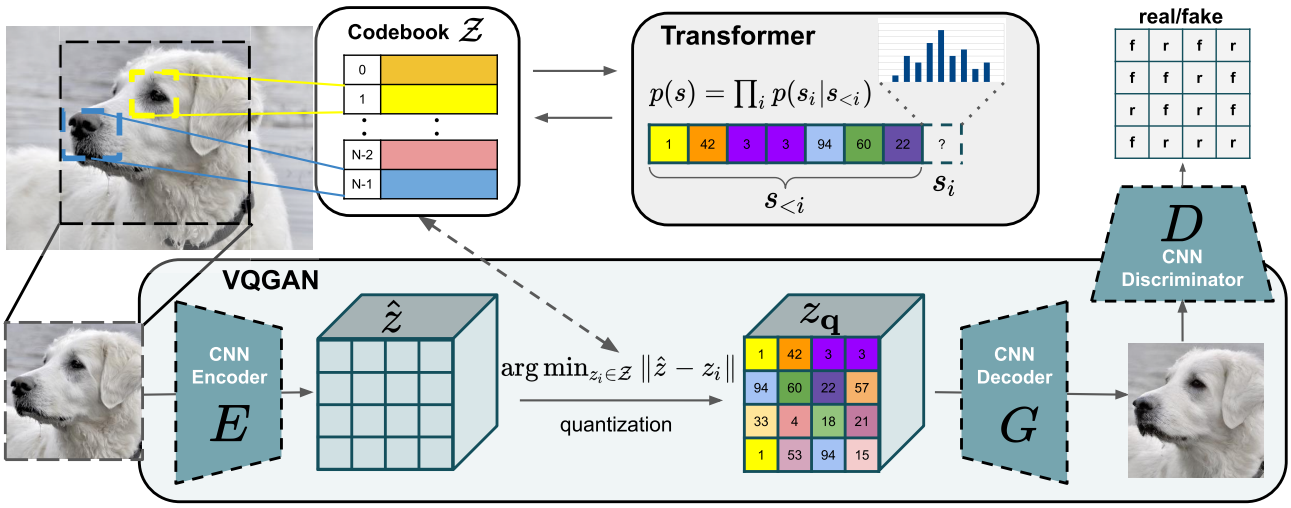

Figure 2. Our approach uses a convolutional *VQGAN* to learn a codebook of context-rich visual parts, whose composition is subsequently modeled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables strong compression while retaining high perceptual quality. This method introduces the efficiency of convolutional approaches to transformer based high resolution image synthesis.

learns a GAN [19] on representations of an autoencoder and [20] on low-resolution wavelet coefficients which are then decoded to images with a learned generator.

[63] presents the Vector Quantised Variational Autoencoder (VQVAE), an approach to learn discrete representations of images, and models their distribution autoregressively with a convolutional architecture. [53] extends this approach to use a hierarchy of learned representations. However, these methods still rely on convolutional density estimation, which makes it difficult to capture long-range interactions in high-resolution images. [8] models images autoregressively with transformers in order to evaluate the suitability of generative pretraining to learn image representations for downstream tasks. Since input resolutions of 32×32 pixels are still quite computationally expensive [8], a VQVAE is used to encode images up to a resolution of 192×192. In an effort to keep the learned discrete representation as spatially invariant as possible with respect to the pixels, a shallow VQVAE with small receptive field is employed. In contrast, we demonstrate that a powerful first stage, which captures as much context as possible in the learned representation, is critical to enable efficient high-resolution image synthesis with transformers.

3. Approach

Our goal is to exploit the highly promising learning capabilities of transformer models [64] and introduce them to high-resolution image synthesis up to the megapixel range. Previous work [48, 8] which applied transformers to image generation demonstrated promising results for images up to a size of 64×64 pixels but, due to the quadratically increasing cost in sequence length, cannot simply be scaled to higher resolutions.

High-resolution image synthesis requires a model that understands the global composition of images, enabling it to generate locally realistic as well as globally consistent patterns. Therefore, instead of representing an image with pixels, we represent it as a composition of perceptually rich image constituents from a codebook. By learning an effective code, as described in Sec. 3.1, we can significantly reduce the description length of compositions, which allows us to efficiently model their global interrelations within images with a transformer architecture as described in Sec. 3.2. This approach, summarized in Fig. 2, is able to generate realistic and consistent high resolution images both in an unconditional and a conditional setting.

3.1. Learning an Effective Codebook of Image Constituents for Use in Transformers

To utilize the highly expressive transformer architecture for image synthesis, we need to express the constituents of an image in the form of a *sequence*. Instead of building on individual pixels, complexity necessitates an approach that uses a discrete codebook of learned representations, such that any image $x \in \mathbb{R}^{H \times W \times 3}$ can be represented by a spatial collection of codebook entries $z_{\mathbf{q}} \in \mathbb{R}^{h \times w \times n_z}$, where n_z is the dimensionality of codes. An equivalent representation is a sequence of $h \cdot w$ indices which specify the respective entries in the learned codebook. To effectively learn such a discrete spatial codebook, we propose to directly incorporate the inductive biases of CNNs and incorporate ideas from neural discrete representation learning [63]. First, we learn a convolutional model consisting of an encoder E and a decoder G, such that taken together, they learn to repre-

sent images with codes from a learned, discrete codebook $\mathcal{Z} = \{z_k\}_{k=1}^K \subset \mathbb{R}^{n_z}$ (see Fig. 2 for an overview). More precisely, we approximate a given image x by $\hat{x} = G(z_{\mathbf{q}})$. We obtain $z_{\mathbf{q}}$ using the encoding $\hat{z} = E(x) \in \mathbb{R}^{h \times w \times n_z}$ and a subsequent element-wise quantization $\mathbf{q}(\cdot)$ of each spatial code $\hat{z}_{ij} \in \mathbb{R}^{n_z}$ onto its closest codebook entry z_k:

$$z_{\mathbf{q}} = \mathbf{q}(\hat{z}) := \left(\underset{z_k \in \mathcal{Z}}{\arg\min} \|\hat{z}_{ij} - z_k\| \right) \in \mathbb{R}^{h \times w \times n_z}. \quad (2)$$

The reconstruction $\hat{x} \approx x$ is then given by

$$\hat{x} = G(z_{\mathbf{q}}) = G\left(\mathbf{q}(E(x))\right). \quad (3)$$

Backpropagation through the non-differentiable quantization operation in Eq. (3) is achieved by a straight-through gradient estimator, which simply copies the gradients from the decoder to the encoder [3], such that the model and codebook can be trained end-to-end via the loss function

$$\mathcal{L}_{\text{VQ}}(E, G, \mathcal{Z}) = \|x - \hat{x}\|^2 + \|\text{sg}[E(x)] - z_{\mathbf{q}}\|_2^2 + \beta \|\text{sg}[z_{\mathbf{q}}] - E(x)\|_2^2. \quad (4)$$

Here, $\mathcal{L}_{\text{rec}} = \|x - \hat{x}\|^2$ is a reconstruction loss, $\text{sg}[\cdot]$ denotes the stop-gradient operation, and $\|\text{sg}[z_{\mathbf{q}}] - E(x)\|_2^2$ is the so-called "commitment loss" with weighting factor β [63].

Learning a Perceptually Rich Codebook Using transformers to represent images as a distribution over latent image constituents requires us to push the limits of compression and learn a rich codebook. To do so, we propose *VQGAN*, a variant of the original VQVAE, and use a discriminator and perceptual loss [36, 26, 35, 16] to keep good perceptual quality at increased compression rate. Note that this is in contrast to previous works which applied pixel-based [62, 53] and transformer-based autoregressive models [8] on top of only a shallow quantization model. More specifically, we replace the L_2 loss used in [63] for $\mathcal{L}_{\text{rec}}$ by a perceptual loss and introduce an adversarial training procedure with a patch-based discriminator D [25] that aims to differentiate between real and reconstructed images:

$$\mathcal{L}_{\text{GAN}}(\{E, G, \mathcal{Z}\}, D) = [\log D(x) + \log(1 - D(\hat{x}))] \quad (5)$$

The complete objective for finding the optimal compression model $\mathcal{Q}^* = \{E^*, G^*, \mathcal{Z}^*\}$ then reads

$$\mathcal{Q}^* = \underset{E,G,\mathcal{Z}}{\arg\min} \max_D \mathbb{E}_{x \sim p(x)} \Big[\mathcal{L}_{\text{VQ}}(E, G, \mathcal{Z}) + \lambda \mathcal{L}_{\text{GAN}}(\{E, G, \mathcal{Z}\}, D) \Big], \quad (6)$$

where we compute the adaptive weight λ according to

$$\lambda = \frac{\nabla_{G_L}[\mathcal{L}_{\text{rec}}]}{\nabla_{G_L}[\mathcal{L}_{\text{GAN}}] + \delta} \quad (7)$$

where $\mathcal{L}_{\text{rec}}$ is the perceptual reconstruction loss [71], $\nabla_{G_L}[\cdot]$ denotes the gradient of its input w.r.t. the last layer L of the decoder, and $\delta = 10^{-6}$ is used for numerical stability. To aggregate context from everywhere, we apply a single attention layer on the lowest resolution. This training procedure significantly reduces the sequence length when unrolling the latent code and thereby enables the application of powerful transformer models.

3.2. Learning the Composition of Images with Transformers

Latent Transformers With E and G available, we can now represent images in terms of the codebook-indices of their encodings. More precisely, the quantized encoding of an image x is given by $z_{\mathbf{q}} = \mathbf{q}(E(x)) \in \mathbb{R}^{h \times w \times n_z}$ and is equivalent to a sequence $s \in \{0, \dots, |\mathcal{Z}|-1\}^{h \times w}$ of indices from the codebook, which is obtained by replacing each code by its index in the codebook $\mathcal{Z}$:

$$s_{ij} = k \text{ such that } (z_{\mathbf{q}})_{ij} = z_k. \quad (8)$$

By mapping indices of a sequence s back to their corresponding codebook entries, $z_{\mathbf{q}} = \left(z_{s_{ij}}\right)$ is readily recovered and decoded to an image $\hat{x} = G(z_{\mathbf{q}})$.

Thus, after choosing some ordering of the indices in s, image-generation can be formulated as autoregressive next-index prediction: Given indices $s_{<i}$, the transformer learns to predict the distribution of possible next indices, i.e. $p(s_i | s_{<i})$ to compute the likelihood of the full representation as $p(s) = \prod_i p(s_i | s_{<i})$. This allows us to directly maximize the log-likelihood of the data representations:

$$\mathcal{L}_{\text{Transformer}} = \mathbb{E}_{x \sim p(x)} \left[-\log p(s) \right]. \quad (9)$$

Conditioned Synthesis In many image synthesis tasks a user demands control over the generation process by providing additional information from which an example shall be synthesized. This information, which we will call c, could be a single label describing the overall image class or even another image itself. The task is then to learn the likelihood of the sequence given this information c:

$$p(s|c) = \prod_i p(s_i | s_{<i}, c). \quad (10)$$

If the conditioning information c has spatial extent, we first learn another *VQGAN* to obtain again an index-based representation $r \in \{0, \dots, |\mathcal{Z}_c|-1\}^{h_c \times w_c}$ with the newly obtained codebook $\mathcal{Z}_c$ Due to the autoregressive structure of the transformer, we can then simply prepend r to s and restrict the computation of the negative log-likelihood to entries $p(s_i | s_{<i}, r)$. This "decoder-only" strategy has also been successfully used for text-summarization tasks [40].

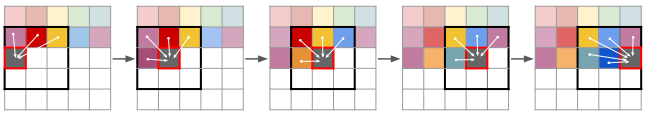

Figure 3. Sliding attention window.

Generating High-Resolution Images The attention mechanism of the transformer puts limits on the sequence length $h \cdot w$ of its inputs s. While we can adapt the number of downsampling blocks m of our *VQGAN* to reduce images of size $H \times W$ to $h = H/2^m \times w = W/2^m$, we observe degradation of the reconstruction quality beyond a critical value of m, which depends on the considered dataset. To generate images in the megapixel regime, we therefore have to work patch-wise and crop images to restrict the length of s to a maximally feasible size during training. To sample images, we then use the transformer in a sliding-window manner as illustrated in Fig. 3. Our *VQGAN* ensures that the available context is still sufficient to faithfully model images, as long as either the statistics of the dataset are approximately spatially invariant or spatial conditioning information is available. In practice, this is not a restrictive requirement, because when it is violated, *i.e*. unconditional image synthesis on aligned data, we can simply condition on image coordinates, similar to [38].

4. Experiments

This section evaluates the ability of our approach to retain the advantages of transformers over their convolutional counterparts (Sec. 4.1) while integrating the effectiveness of convolutional architectures to enable high-resolution image synthesis (Sec. 4.2). Furthermore, in Sec. 4.3, we investigate how codebook quality affects our approach. We close the analysis by providing a quantitative comparison to a wide range of existing approches for generative image synthesis in Sec. 4.4. Based on initial experiments, we usually set $|\mathcal{Z}|= 1024$ and train all subsequent transformer models to predict sequences of length $16 \cdot 16$, as this is the maximum feasible length to train a GPT2-medium architecture (307 M parameters) [51] on a GPU with 12GB VRAM. More details on architectures and hyperparameters can be found in the appendix (Tab. 6 and Tab. 7).

4.1. Attention Is All You Need in the Latent Space

Transformers show state-of-the-art results on a wide variety of tasks, including autoregressive image modeling. However, evaluations of previous works were limited to transformers working directly on (low-resolution) pixels [48, 12, 24], or to deliberately shallow pixel encodings [8]. This raises the question if our approach retains the advantages of transformers over convolutional approaches.

To answer this question, we use a variety of conditional and unconditional tasks and compare the performance between our transformer-based approach and a convolutional approach. For each task, we train a *VQGAN* with $m = 4$ downsampling blocks, and, if needed, another one for the conditioning information, and then train both a transformer and a PixelSNAIL [10] model on the same representations, as the latter has been used in previous state-of-the-art two-stage approaches [53]. For a thorough comparison, we vary the model capacities between 85M and 310M parameters and adjust the number of layers in each model to match one another. We observe that PixelSNAIL trains roughly twice as fast as the transformer and thus, for a fair comparison, report the negative log-likelihood both for the same amount of training time (*P-SNAIL time*) and for the same amount of training steps (*P-SNAIL steps*).

	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

Table 1. Comparing Transformer and PixelSNAIL architectures across different datasets and model sizes. For all settings, transformers outperform the state-of-the-art model from the PixelCNN family, PixelSNAIL in terms of NLL. This holds both when comparing NLL at fixed times (PixelSNAIL trains roughly 2 times faster) and when trained for a fixed number of steps. See Sec. 4.1 for the abbreviations.

Results Tab. 1 reports results for unconditional image modeling on *ImageNet* (IN) [14], *Restricted ImageNet* (RIN) [57], consisting of a subset of animal classes from ImageNet, *LSUN Churches and Towers* (LSUN-CT) [69], and for conditional image modeling of RIN conditioned on depth maps obtained with the approach of [52] (D-RIN) and of landscape images collected from Flickr conditioned on semantic layouts (S-FLCKR) obtained with the approach of [7]. Note that for the semantic layouts, we train the first-stage using a cross-entropy reconstruction loss due to their discrete nature. The results shows that the transformer consistently outperforms PixelSNAIL across all tasks when trained for the same amount of time and the gap increases even further when trained for the same number of steps. These results demonstrate that gains of transformers carry over to our proposed two-stage setting.

4.2. A Unified Model for Image Synthesis Tasks

The versatility and generality of the transformer architecture makes it a promising candidate for image synthesis. In the conditional case, additional information c such as class labels or segmentation maps are used and the goal is to learn the distribution of images as described in Eq. (10). Using the same setting as in Sec. 4.1 (i.e. image size 256×256,

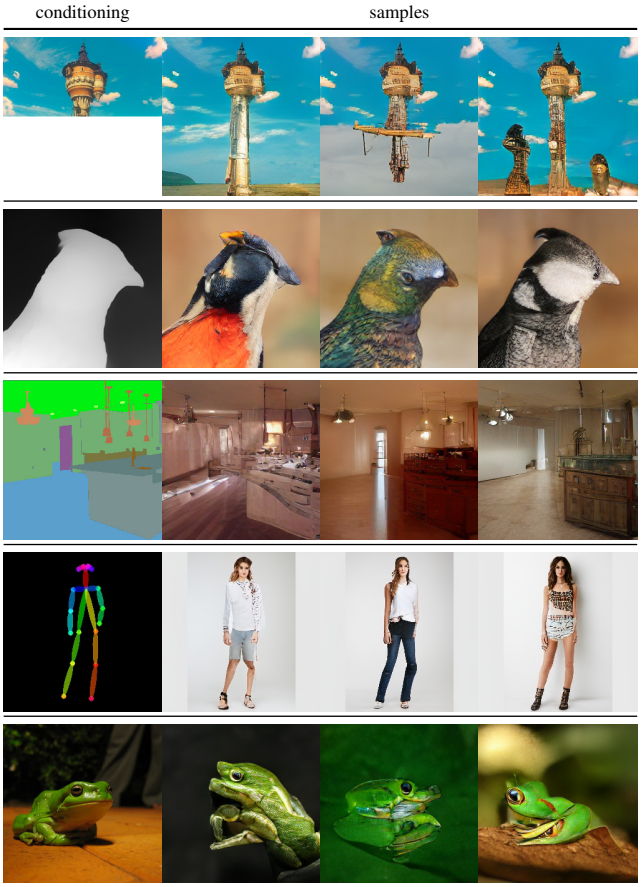

Figure 4. Transformers within our setting unify a wide range of image synthesis tasks. We show 256×256 synthesis results across different conditioning inputs and datasets, all obtained with the same approach to exploit inductive biases of effective CNN based *VQGAN* architectures in combination with the expressivity of transformer architectures. Top row: Completions from unconditional training on ImageNet. 2nd row: Depth-to-Image on RIN. 3rd row: Semantically guided synthesis on ADE20K. 4th row: Pose-guided person generation on DeepFashion. Bottom row: Class-conditional samples on RIN.

latent size 16×16), we perform various conditional image synthesis experiments:

(i): **Semantic image synthesis**, where we condition on semantic segmentation masks of ADE20K [72], a web-scraped landscapes dataset (S-FLCKR) and COCO-Stuff [6]. Results are depicted in Figure 4, 5 and Fig. 6.

(ii): **Structure-to-image**, where we use either depth or edge information to synthesize images from both RIN and IN (see Sec. 4.1). The resulting depth-to-image and edge-to-image translations are visualized in Fig. 4 and Fig. 6.

(iii): **Pose-guided synthesis:** Instead of using the semantically rich information of either segmentation or depth maps, Fig. 4 shows that the same approach as for the previous experiments can be used to build a shape-conditional generative model on the DeepFashion [41] dataset.

(iv): **Stochastic superresolution**, where low-resolution images serve as the conditioning information and are thereby upsampled. We train our model for an upsampling factor of 8 on ImageNet and show results in Fig. 6.

(v): **Class-conditional image synthesis:** Here, the conditioning information c is a single index describing the class label of interest. Results on conditional sampling for the RIN dataset are demonstrated in Fig. 4.

All of these examples make use of the same methodology. Instead of requiring task specific architectures or modules, the flexibility of the transformer allows us to learn appropriate interactions for each task, while the *VQGAN* — which can be *reused* across different tasks — leads to short sequence lengths. In combination, the presented approach can be understood as an efficient, general purpose mechanism for conditional image synthesis. Note that additional results for each experiment can be found in the appendix, Sec. C.

High-Resolution Synthesis The sliding window approach introduced in Sec. 3.2 enables image synthesis beyond a resolution of 256×256 pixels. We evaluate this approach on unconditional image generation on LSUN-CT and FacesHQ (see Sec. 4.3) and conditional synthesis on D-RIN, COCO-Stuff and S-FLCKR, where we show results in Fig. 1, 6 and the supplementary (Fig. 17-27). Note that this approach can in principle be used to generate images of arbitrary ratio and size, given that the image statistics of the dataset of interest are approximately spatially invariant or spatial information is available. Impressive results can be achieved by applying this method to image generation from semantic layouts on S-FLCKR, where a strong *VQGAN* can be learned with $m = 5$, so that its codebook together with the conditioning information provides the transformer with enough context for image generation in the megapixel regime.

4.3. Building Context-Rich Vocabularies

How important are context-rich vocabularies? To investigate this question, we ran experiments where the transformer architecture is kept fixed while the amount of context encoded into the representation of the first stage is varied through the number of downsampling blocks of our *VQGAN*. We specify the amount of context encoded in terms of reduction factor in the side-length between image inputs and the resulting representations, *i.e.* a first stage encoding images of size $H \times W$ into discrete codes of size $H/f \times W/f$ is denoted by a factor f. For $f = 1$, we reproduce the approach of [8] and replace our *VQGAN* by a k-means clustering of RGB values with $k = 512$.

During training, we always crop images to obtain inputs of size 16×16 for the transformer, *i.e.* when modeling images with a factor f in the first stage, we use crops of size $16f \times 16f$. To sample from the models, we always apply them in a sliding window manner as described in Sec. 3.

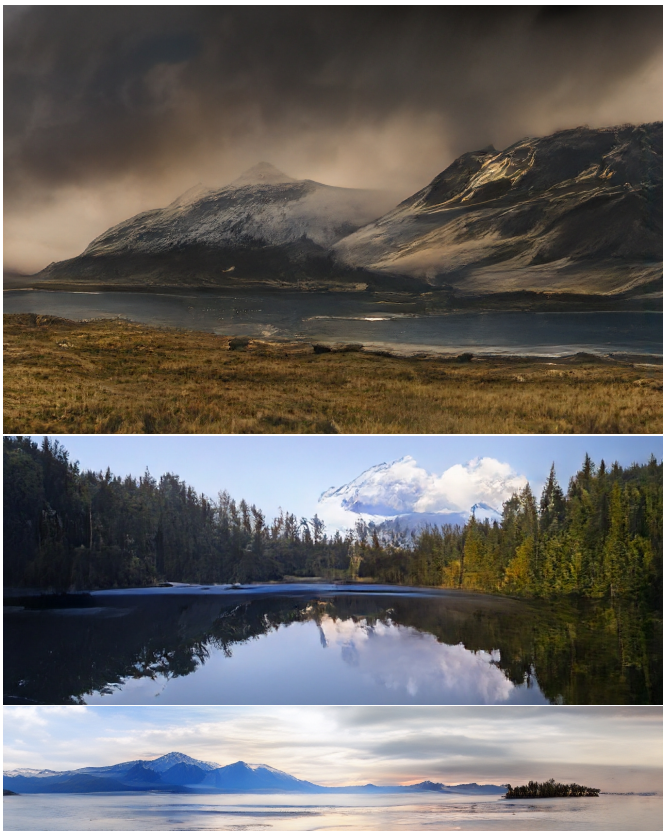

Figure 5. Samples generated from semantic layouts on S-FLCKR. Sizes from top-to-bottom: 1280×832, 1024×416 and 1280×240 pixels. Best viewed zoomed in. A larger visualization can be found in the appendix, see Fig 17.

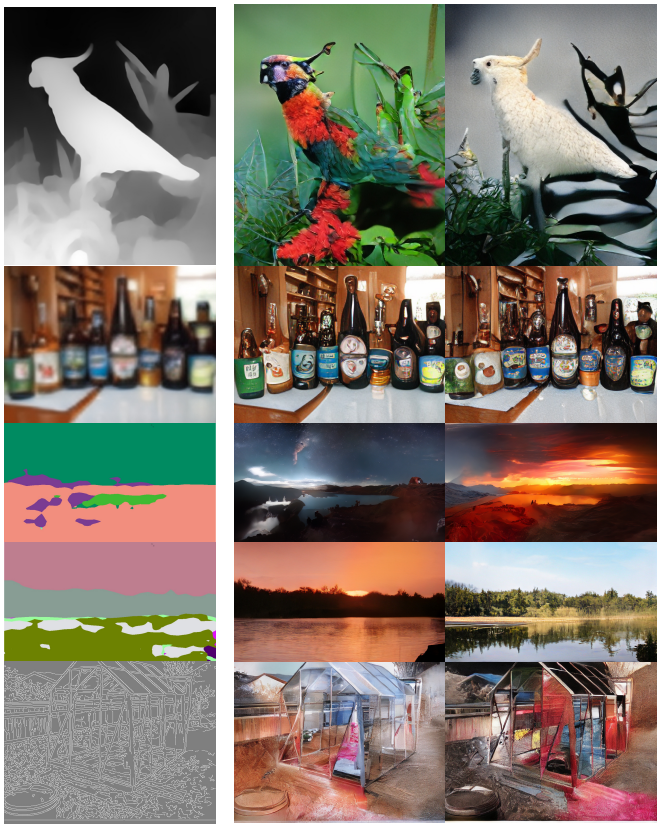

Figure 6. Applying the sliding attention window approach (Fig. 3) to various conditional image synthesis tasks. Top: Depth-to-image on RIN, 2nd row: Stochastic superresolution on IN, 3rd and 4th row: Semantic synthesis on S-FLCKR, bottom: Edge-guided synthesis on IN. The resulting images vary between 368×496 and 1024×576, hence they are best viewed zoomed in.

Results Fig. 7 shows results for unconditional synthesis of faces on *FacesHQ*, the combination of *CelebA-HQ* [27] and *FFHQ* [29]. It clearly demonstrates the benefits of powerful *VQGAN*s by increasing the effective receptive field of the transformer. For small receptive fields, or equivalently small f, the model cannot capture coherent structures. For an intermediate value of $f = 8$, the overall structure of images can be approximated, but inconsistencies of facial features such as a half-bearded face and of viewpoints in different parts of the image arise. Only our full setting of $f = 16$ can synthesize high-fidelity samples. For analogous results in the conditional setting on S-FLCKR, we refer to the appendix (Fig. 10 and Sec. B).

To assess the effectiveness of our approach quantitatively, we compare results between training a transformer directly on pixels, and training it on top of a *VQGAN*'s latent code with $f = 2$, given a fixed computational budget. Again, we follow [8] and learn a dictionary of 512 RGB values on CIFAR10 to operate directly on pixel space and train the same transformer architecture on top of our *VQGAN* with a latent code of size $16 \times 16 = 256$. We observe improvements of 18.63% for FIDs and $14.08\times$ faster sampling of images.

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

Table 2. FID score comparison for semantic image synthesis (256×256 pixels). (*): Recalculated with our evaluation protocol based on [43] on the validation splits of each dataset.

4.4. Quantitative Comparison to Existing Models

In this section we investigate how our approach quantitatively compares to existing models for generative image synthesis. In particular, we assess the performance of our model in terms of FID and compare to a variety of established models (GANs, VAEs, Flows, AR, Hybrid) on (i) *semantic synthesis* in Tab. 2 (where we compare to [46, 65, 31, 9]) and (ii) *unconditional face synthesis* in Tab. 3. Furthermore, to address a direct comparison to the original VQVAE-2 model [53], we train a class conditional ImageNet transformer on 256×256 images, using a *VQGAN* with $\dim \mathcal{Z} = 16384$ and $f = 16$, and additionally compare to BigGAN [4] and MSP [18] in Tab. 4. Note that our model uses $\simeq 10\times$ *less* parameters than VQVAE-2,

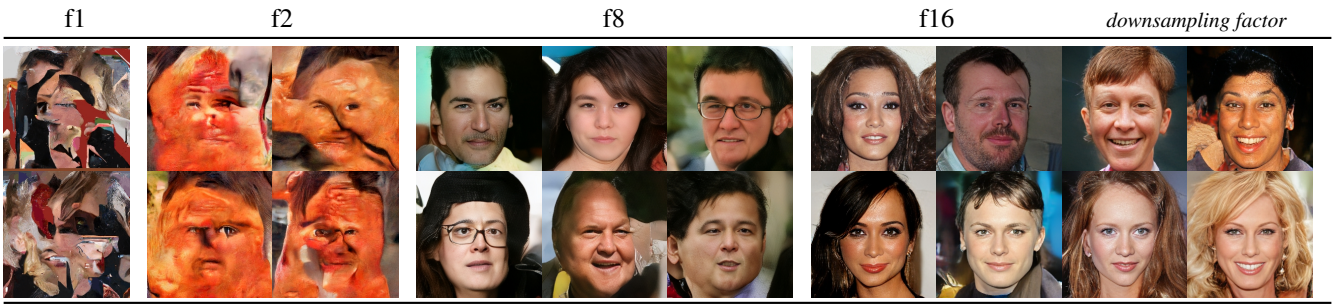

Figure 7. Evaluating the importance of effective codebook for HQ-Faces (CelebA-HQ and FFHQ) for a fixed sequence length $|s| = 16 \cdot 16 = 256$. Globally consistent structures can only be modeled with a context-rich vocabulary (right). All samples are generated with temperature $t = 1.0$ and top-k sampling with $k = 100$. Last row reports the speedup over the f1 baseline which operates directly on pixels and takes 7258 seconds to produce a sample on a NVIDIA GeForce GTX Titan X.

CelebA-HQ 256 × 256		FFHQ 256 × 256	
Method	FID ↓	Method	FID ↓
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)

Table 3. FID score comparison for face image synthesis. CelebA-HQ results reproduced from [1, 47, 67, 22], FFHQ from [58, 28].

Dataset	**ours** (+R)	VQVAE-2 (+R)	BigGAN (-deep)	MSP
IN 256, 50K	19.8 (11.2)	38.1 (∼ 10)	7.1 (7.3)	n.a.
IN 256, 18K	23.5	n.a.	9.6 (9.7)	50.4

Table 4. FID score comparison for class-conditional synthesis. "+R": classifier-based rejection sampling as proposed in VQVAE-2. FID*-values (calculated on reconstructed data, analogous to [53]): ours: 13.5 (8.1), VQVAE-2: 19 (5). BigGAN (-deep) evaluated via `https://tfhub.dev/deepmind` truncated at 1.0.

which has an estimated parameter count of 13.5B (estimation based on `https://github.com/rosinality/vq-vae-2-pytorch`). While some task-specialized GAN models report better FID scores, our approach provides a unified model that works well across a wide range of tasks while retaining the ability to encode and reconstruct images. It thereby bridges the gap between purely adversarial and likelihood-based approaches. Fig. 11, 12, 13 and Fig. 14 contain qualitative samples corresponding to the quantitative analysis in Tab. 4.

How good is the *VQGAN*? Reconstruction FIDs obtained via the codebook provide a lower bound on the achievable FID of the generative model trained on it. To quantify the performance gains of our *VQGAN* over VQVAE-2, we evaluate this metric on ImageNet and report results in Tab. 5. Our *VQGAN* outperforms VQVAE-2 while providing significantly more compression (seq. length of 256 vs. $5120 = 32^2 + 64^2$). As expected, larger versions of

Model	Codebook Size	dim $\mathcal{Z}$	FID ↓
VQVAE-2	64×64 & 32×32	512	∼ 10
VQGAN	16×16	1024	8.0
VQGAN	16×16	16384	4.9
VQGAN	64×64 & 32×32	512	1.7

Table 5. Reconstruction FID on ImageNet (validation split). VQVAE-2 reported their reconstruction FID as "∼ 10".

VQGAN (either in terms of larger codebook sizes or increased code lengths) further improve performance. Using the same hierarchical codebook setting as in VQVAE-2 with our model provides the best reconstruction FID, albeit at the cost of a very long and thus impractical sequence. Furthermore, Fig. 9 qualitatively shows that a standard VQVAE cannot achieve such compressions; the corresponding reconstruction-FIDs read: VQVAE 254.4; *VQGAN* 5.7. Sampling from this VQVAE cannot achieve FIDs below 254.4, whereas our *VQGAN* achieves 21.93 with PixelSNAIL and 11.44 with a transformer (see Tab. 3).

5. Conclusion

This paper adressed the fundamental challenges that previously confined transformers to low-resolution images. We proposed an approach which represents images as a composition of perceptually rich image constituents and thereby overcomes the infeasible quadratic complexity when modeling images directly in pixel space. Modeling constituents with a CNN architecture and their compositions with a transformer architecture taps into the full potential of their complementary strengths and thereby allowed us to represent the first results on high-resolution image synthesis with a transformer-based architecture. In experiments, our approach demonstrates the efficiency of convolutional inductive biases and the expressivity of transformers by synthesizing images in the megapixel range and outperforming state-of-the-art convolutional approaches. Equipped with a general mechanism for conditional synthesis, it offers many opportunities for novel neural rendering approaches.

This work has been supported by the German Research Foundation (DFG) projects 371923335, 421703927 and a hardware donation from NVIDIA corporation.

References

[1] Jyoti Aneja, Alexander G. Schwing, Jan Kautz, and Arash Vahdat. NCP-VAE: variational autoencoders with noise contrastive priors. *CoRR*, abs/2010.02917, 2020. 8

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016. 2

[3] Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR*, abs/1308.3432, 2013. 4

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *7th International Conference on Learning Representations, ICLR*, 2019. 7, 16, 17, 18, 19

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

[6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018. 6

[7] Liang-Chieh Chen, G. Papandreou, I. Kokkinos, Kevin Murphy, and A. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 5

[8] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. 2020. 1, 2, 3, 4, 5, 6, 7, 13, 14, 20, 21

[9] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1520–1529. IEEE Computer Society, 2017. 7

[10] Xi Chen, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel. Pixelsnail: An improved autoregressive generative model. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 863–871. PMLR, 2018. 2, 5

[11] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *CoRR*, abs/2011.10650, 2020. 8

[12] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019. 1, 2, 5

[13] Bin Dai and David P. Wipf. Diagnosing and enhancing VAE models. In *7th International Conference on Learning Representations, ICLR*, 2019. 2

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, 2009. 5

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020. 1

[16] Alexey Dosovitskiy and Thomas Brox. Generating Images with Perceptual Similarity Metrics based on Deep Networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2016. 4

[17] Patrick Esser, Robin Rombach, and Björn Ommer. A Disentangling Invertible Interpretation Network for Explaining Latent Representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2020. 2

[18] Jeffrey De Fauw, Sander Dieleman, and Karen Simonyan. Hierarchical autoregressive image models with auxiliary decoders. *CoRR*, abs/1903.04933, 2019. 7, 16, 17, 18, 19

[19] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2014. 3

[20] Seungwook Han, Akash Srivastava, Cole L. Hurwitz, Prasanna Sattigeri, and David D. Cox. not-so-biggan: Generating high-fidelity images on a small compute budget. *CoRR*, abs/2009.04433, 2020. 3

[21] Ari Heljakka, Arno Solin, and Juho Kannala. Pioneer networks: Progressively growing generative autoencoder. In C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part I*, 2018. 8

[22] Ari Heljakka, Arno Solin, and Juho Kannala. Towards photographic image manipulation with balanced growing of generative autoencoders. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 3109–3118. IEEE, 2020. 8

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 12

[24] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *CoRR*, abs/1912.12180, 2019. 2, 5

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017. 4, 12

[26] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV (2)*, volume 9906 of *Lecture Notes in Computer Science*, pages 694–711. Springer, 2016. 4

[27] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. 7, 8

[28] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 8

[29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR) 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019. 7

[30] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116. IEEE, 2020. 8

[31] Prateek Katiyar and Anna Khoreva. Improving augmentation and evaluation schemes for semantic image synthesis, 2021. 7

[32] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. Structured attention networks, 2017. 2

[33] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS*, 2018. 8

[34] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR*, 2014. 2

[35] Alex Lamb, Vincent Dumoulin, and Aaron C. Courville. Discriminative regularization for generative models. *CoRR*, abs/1602.03220, 2016. 4

[36] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric, 2015. 4

[37] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *AAAI*, pages 6706–6713. AAAI Press, 2019. 2

[38] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. COCO-GAN: generation by parts via conditional coordinating. In *ICCV*, pages 4511–4520. IEEE, 2019. 5

[39] Jinlin Liu, Yuan Yao, and Jianqiang Ren. An acceleration framework for high resolution image synthesis. *CoRR*, abs/1909.03611, 2019. 2

[40] Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. In *ICLR (Poster)*. OpenReview.net, 2018. 4

[41] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 6

[42] Jacob Menick and Nal Kalchbrenner. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 14

[43] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. toshas/torch-fidelity: Version 0.2.0, May 2020. 7

[44] B. Ommer and J. M. Buhmann. Learning the compositional nature of visual objects. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 2

[45] Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference, 2016. 2

[46] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic Image Synthesis with Spatially-Adaptive Normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2019. 2, 7, 33, 34

[47] Gaurav Parmar, Dacheng Li, Kwonjoon Lee, and Zhuowen Tu. Dual contradistinctive generative autoencoder, 2020. 8

[48] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 4052–4061. PMLR, 2018. 2, 3, 5

[49] Stanislav Pidhorskyi, Donald A. Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 14092–14101. IEEE, 2020. 8

[50] A. Radford. Improving language understanding by generative pre-training. 2018. 1

[51] A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 1, 5, 12

[52] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 5

[53] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2, 2019. 3, 4, 5, 7, 8, 16, 17, 18, 19

[54] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning, ICML*, 2014. 2

[55] Robin Rombach, Patrick Esser, and Björn Ommer. Making sense of cnns: Interpreting deep representations and their invariances with inns. In Andrea Vedaldi, Horst Bischof,

Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*, volume 12362 of *Lecture Notes in Computer Science*, pages 647–664. Springer, 2020. 2

[56] Robin Rombach, Patrick Esser, and Bjorn Ommer. Network-to-network translation with conditional invertible neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2784–2797. Curran Associates, Inc., 2020. 2

[57] Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Computer vision with a single (robust) classifier. In *ArXiv preprint arXiv:1906.09453*, 2019. 5

[58] Edgar Schönfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8204–8213. IEEE, 2020. 8

[59] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019. 2

[60] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 8

[61] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1747–1756. JMLR.org, 2016. 2

[62] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders, 2016. 2, 4, 14

[63] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. 3, 4, 13, 15

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2017. 1, 2, 3

[65] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 7

[66] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *ICLR*. OpenReview.net, 2020. 2

[67] Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. Vaebm: A symbiosis between variational autoencoders and energy-based models, 2021. 8

[68] Zhisheng Xiao, Qing Yan, Yi-an Chen, and Yali Amit. Generative latent flow: A framework for non-adversarial image generation. *CoRR*, abs/1905.10485, 2019. 2

[69] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 5

[70] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-Domain Correspondence Learning for Exemplar-Based Image Translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2020. 2

[71] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 2018. 4, 12

[72] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442*, 2016. 6

[73] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A. Efros. View synthesis by appearance flow, 2017. 2

[74] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization, 2019. 2