Enhancing Hierarchical Vector Quantized Autoencoders for Image Synthesis through Multiple Decoders

Dario Serez[1,3][0009−0003−4006−3575], Marco Cristani[4][0000−0002−0523−6042], Vittorio Murino[1,2,4][0000−0002−8645−2328], Alessio Del Bue[1][0000−0002−2262−4872], and Pietro Morerio[1][0000−0001−5259−1496]

[1] PAVIS Department, Italian Institute of Technology (IIT), Genoa, Italy
{dario.serez,vittorio.murino,alessio.delbue,pietro.morerio}@iit.it
[2] DIBRIS Department, University of Genova (UniGE), Genoa, Italy
[3] DITEN Department, University of Genova (UniGE), Genoa, Italy
[4] Department of Computer Science, University of Verona (UniVR), Verona, Italy
marco.cristani@univr.it

Abstract. Vector Quantized Variational Autoencoders (VQ-VAEs) have gained popularity in recent years due to their ability to represent images as discrete sequences of tokens that index a learned codebook of vectors, enabling efficient image compression. One variant of particular interest is VQ-VAE 2, which extends previous works by representing images as a hierarchy of sequences, resulting in finer-grained representations.
In this study, we further enhance such hierarchical autoencoder approach by introducing multiple decoders, which allow to represent images as a sum of multi-scale contributions in the pixel space. Our proposed model, the Multi Scale (MS) VQ-VAE, not only enables better control over the encoding of each sequence (resulting in improved explainability and codebook usage) but, as a consequence, also shows advantages in image synthesis. Our experiments demonstrate that the MS-VQVAE achieves comparable or superior reconstructions on various datasets and resolutions, as well as greater stability across runs. Moreover, we include a proof-of-concept trial to showcase the potential applications of our model in image synthesis.

Keywords: VQ-VAE · Hierarchical Autoencoder · Image synthesis

1 Introduction

Vector Quantized Variational Autoencoders (VQ-VAE) [18] are popular in Computer Vision for their ability to learn discrete low-dimensional representations of images by indexing a codebook (or dictionary) of learnable vectors. Notably, VQ-VAE and its extensions [6, 28, 14] have been successfully combined with autoregressive models to perform image synthesis [19, 29, 2]. Despite this success, the original algorithm presents limitations in reconstructing the fine-grained information of the encoded images, especially at high-resolutions, where the details

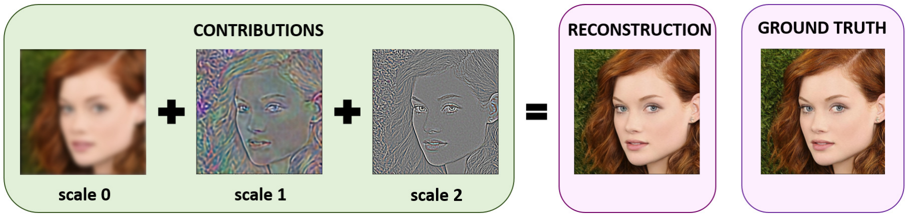

Fig. 1. MS-VQVAE separately reconstructs the hierarchy of quantized sequences in the pixel space, from coarse (scale 0) to fine (scale 2), thus enhancing the explainability of what is encoded at each level. The final sample is given by the sum of these contributions. Example with three scales on the Celeb-A dataset (1024 × 1024 resolution).

are more complex and abundant. As a result, the loss of important information may impact the quality of reconstructions. Therefore, a key challenge in current research is to improve the accuracy of the result while preserving the compressibility of the representation. Later on, VQ-VAE 2 [20] was introduced to address these limitations by representing images as a hierarchy of sequences. This approach breaks down the image into multiple levels of abstraction, capturing its different aspects. By constructing a sequence hierarchy, VQ-VAE 2 can better learn the complex structure of images, leading to more accurate reconstructions. In this study, we further enhance the hierarchical Autoencoder by introducing multiple decoders to represent images as a sum of contributions in the pixel space, as shown in Figure 1. Our approach is inspired by a "*divide-and-conquer*" strategy, where a common CNN-Encoder generates multiple latent sequences at different scales, each responsible for a specific contribution. The sequences are then separately quantized and decoded at the original resolution, and the resulting image at a specific scale i is obtained by adding contributions from all scales $\leq i$. The proposed method, which we name Multi Scale (MS) VQ-VAE, improves explainability over its predecessor (VQ-VAE 2) by allowing direct control and meaningful decoding of each latent-sequence content.

In this work, we also provide a proof-of-concept study where we show the potential applications and benefits of our MS-VQVAE in image synthesis. The whole pipeline (called 2-stage sampling) consists in training an autoregressive model (stage-2) on top of the VQ-VAE (stage-1) discrete representations. This allows the generation of new sequences of indices, which can be decoded with the pretrained Autoencoder. Since it is not necessary to sample all sequences to decode a meaningful image, poor quality samples can be discarded after the generation of the first scale only. Additionally, once the "coarse" part of the image is given, it is possible to modify only its "details" (sequences at scales > 0) multiple times, allowing for the generation of different versions of the same sample.

Further details about the used methodologies are given in Section 3, while in Section 2 differences with respect to VQ-VAE 2 are discussed, as well as improvements introduced by other hierarchical quantized autoencoders. In Section 4 we present the experimental results and evaluate the proposed model on the Imagenet [3] (resolution 256 × 256) and CelebA-HQ [15] (resolution 1024 × 1024) datasets, showing better stability and codebook usage with respect to VQ-VAE 2, as well as general better performances in terms of reconstruction. Furthermore, in Sections 3 to 5, we discuss the potential future applications of our model in image-synthesis tasks.

2 Previous Work

Improving Quantization: The original VQ-VAE work [18] suffers from a known issue called *codebook collapse* (or *index collapse*) [12], where only a subset of available codebook vectors is actually used to encode information. To address this problem, researchers have proposed various improvements such as Expectation-Maximization [21], Decomposed Vector Quantization [12] Continuous Relaxation [16, 10] and Codebook Restart [4]. Although any of these algorithms can be implemented in our method, we employ the basic EMA quantization in order to fairly compare with the VQ-VAE 2 baseline, as described in Section 3. Nevertheless, by introducing multiple decoders, we observe an increased utilization of the codebook in our sequences, as shown in Section 4.
Perceptual Loss and GAN Discriminator: To improve the quality of output images while keeping good compression rates, VQ-GAN [6] replaced the $l2$ reconstruction loss with a combination of Patch-GAN discriminator [9] and perceptual loss [11, 30]. Later on, ViT-VQGAN [28] further improved this result with a Style-GAN discriminator [24] and a Vision Transformer (ViT) [5] based Autoencoder. Since our main concern is to fairly compare with the VQ-VAE 2 baseline, in this work we use CCN-Autoencoders and $l2$ loss on all sequences. An improved implementation with more complex architectures and loss functions is left to future research.
Hierarchical Quantized Autoencoders for Image Modeling: A distinct category of methods proposes the concept of a *hierarchical* Autoencoder and shows its advantages. Specifically, [27] utilizes Mean Squared Error (*MSE*) to compare multiple quantized sequences with their encoded counterpart in order to achieve higher levels of compression, but without exploring possible applications in image modeling. [7] combines the hierarchical structure with autoregressive decoders to enable end-to-end sampling. In Contrast, our MS-VQVAE is designed for 2-stage sampling, which has become common practice for many different methods [19, 29, 2]. A different category of approaches [31, 14, 1] incorporates a stack of codes in the quantization bottleneck, which can be viewed as a form of hierarchy. Among them, [14, 1] are notable for introducing Residual Quantization, where the stack of quantized codes is viewed in a coarse-to-fine manner and summed to obtain the full representation. In contrast, in our MS-VQVAE, the latent residuals exist at different scales (resolutions) and are summed only once decoded in the pixel space. Our work is closely related to VQ-VAE 2 [20], which involves quantizing a stack of latent codes at different resolutions, concatenating them, and decoding the resulting sequence. However, we observed that this technique lacks the ability to control the content of each sequence and often under-exploits codebook dictionaries.
In contrast to all these methods, our proposed MS-VQVAE employs multiple decoders that are explicitly optimized using a residual-based approach to provide

greater control over the content of each sequence. Each Decoder is responsible for a specific subset of the latent codes, and the resulting residuals of these decodings are summed to obtain the full representation. Our architecture also presents advantages for image sampling, where partial generations can naturally be decoded in order to early remove poor quality samples, while multiple and different “details” can be generated on top of the selected “coarse” image.

3 Methodology

3.1 Background

The core of vanilla VQ-VAE [18] is the quantization process, which defines a learnable codebook of vectors $e \in \mathbb{R}^{K \times D}$, where K is the codebook size and D is the dimension of each vector e_i. After encoding the image x, each latent of the output $z_e(x)$ is associated with the nearest embedding in the codebook:

$$z_q(x) = e_k \quad \text{where } k = \text{argmin}_j \| z_e(x) - e_j \| \tag{1}$$

Finally, the Decoder reconstructs the original image from the quantized latent vectors $z_q(x)$. The loss function is composed of three terms:

$$\mathcal{L} = \underbrace{\|x - \hat{x}\|_2^2}_{\text{recons.}} + \underbrace{\|\text{sg}[z_e(x)] - e\|_2^2}_{\text{codebook}} + \beta \underbrace{\|z_e(x) - \text{sg}[e]\|_2^2}_{\text{commitment}} \tag{2}$$

where $\hat{x}$ is the Decoder reconstruction, *sg* denotes the stop gradient operation and β is a constant term usually set to 0.25. The three terms represent the reconstruction, the codebook, and the commitment loss, respectively. A variation is also proposed, where the second term is removed and the embeddings are learned using an Exponential Moving Average (EMA). In detail, each codebook entry e_i is updated at every step t following:

$$e_i^{(t)} := m_i^{(t)} / N_i^{(t)} \tag{3}$$

where $m_i^{(t)}$, $N_i^{(t)}$ represent at each step the mean vector and the usage count with respect to codeword e_i, and they are updated according to the $n_i^{(t)}$ encoder outputs that are closest to the embedding e_i at step t:

$$m_i^{(t)} := m_i^{(t-1)} \cdot \gamma + \sum_j^{n_i^{(t)}} z_{i,j}^{(t)} (1 - \gamma); \quad N_i^{(t)} := N_i^{(t-1)} \cdot \gamma + n_i^{(t)} (1 - \gamma) \tag{4}$$

Here γ is a constant factor set between 0 and 1, usually to 0.99.

3.2 Multi Scale VQ-VAE

The proposed MS-VQVAE architecture is depicted in Figure 2. The input image is fed to a fully convolutional Encoder, which produces M latent images $z_{e,m}$

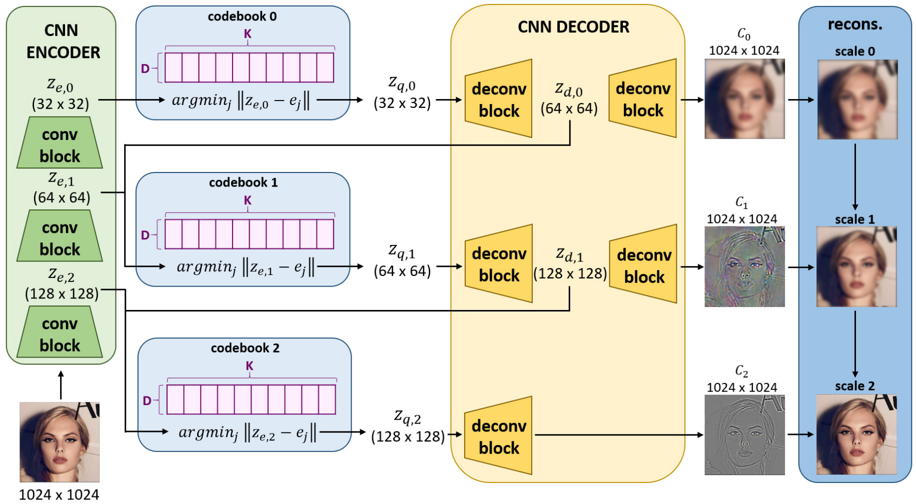

Fig. 2. Full pipeline of our method using three scales ($32 \times 32, 64 \times 64, 128 \times 128$) on an example taken from the CelebA-HQ dataset (1024×1024). The encoding process produces the three sequences $z_{e,0}, z_{e,1}, z_{e,2}$, which are separately quantized to $z_{q,0}, z_{q,1}, z_{q,2}$. The decoding process results in the three contributions C_0, C_1, C_2, which are added to obtain the reconstruction at each scale. During decoding, the intermediate sequences of the first two scales $z_{d,0}, z_{d,1}$ are concatenated to $z_{e,1}, z_{e,2}$, respectively.

at different scales $m \in \{0, 1, \dots, M-1\}$, where $z_{e,0}$ encodes global content and $z_{e,M-1}$ encodes the finest details. Each latent sequence has its own codebook of vectors $e_m \in \mathbb{R}^{K \times D}$, and the quantization process follows Equation (1) to obtain M quantized latents $z_{q,m}$. We use EMA strategy for the learning of each codebook (Equation (4)) and we do not prevent codebook collapse, in order to ensure fair comparisons with previous work. Decoding involves M separate decoders that upsample each sequence to the original size. For $m \in \{0, 1, \dots, M-2\}$ the latents undergo a two-step decoding process, with intermediate sequences $z_{d,m}$ that are concatenated to $z_{e,m+1}$ (before quantization) to allow an information flow between decoders. For scale $m = M-1$, the sequence is directly upsampled to the final resolution. The decoding process produces M residual images C_m. The reconstruction at a given scale m is then the result of the summation of contributions $0, 1, \dots, m-1$:

$$\hat{x}_m = \sum_{i=0}^{m} C_i \tag{5}$$

which implies that $\hat{x}_0 = C_0$.

The overall loss function is a generalization of Equation (2) for multiple scales, without the codebook term due to the EMA algorithm:

$$\mathcal{L} = \underbrace{\sum_{m=0}^{M} \|x_m - \hat{x}_m\|_2^2}_{\text{MS recons.}} + \underbrace{\sum_{m=0}^{M} \beta \|z_{e,m}(x) - \text{sg}[e_m]\|_2^2}_{\text{MS committment}} \tag{6}$$

where $\hat{x}_m$ is defined as in equation Equation (5). In the reconstruction term, the ground-truths images x_m are defined M as:

$$x_m = \begin{cases} x & \text{if m = M - 1} \\ \mathcal{B}(x_{m+1}, \kappa, \sigma) & \text{otherwise} \end{cases} \tag{7}$$

where $\mathcal{B}$ denotes the *Gaussian Blur* operation, κ is the kernel which depends on the image resolution r and is computed as $\sqrt{r} - 1$, and σ is the standard deviation given as a function of the kernel:

$$\sigma = \frac{1}{3}\left(\frac{(\kappa - 1)}{2} - 1\right) + \frac{4}{5} \tag{8}$$

By doing so, the ground truths corresponding to lower scales appear as low-frequency versions of the original sample. This mechanism forces the early sequences to focus only on the global structure of images, ignoring the high-frequency details. The explainability of our method is provided by the multi-scale reconstruction mechanism, since we can asses and show that low scales encode low frequency information (the global content of images), while higher scales encode the high frequencies (details).

3.3 Image Sampling: Intuitions

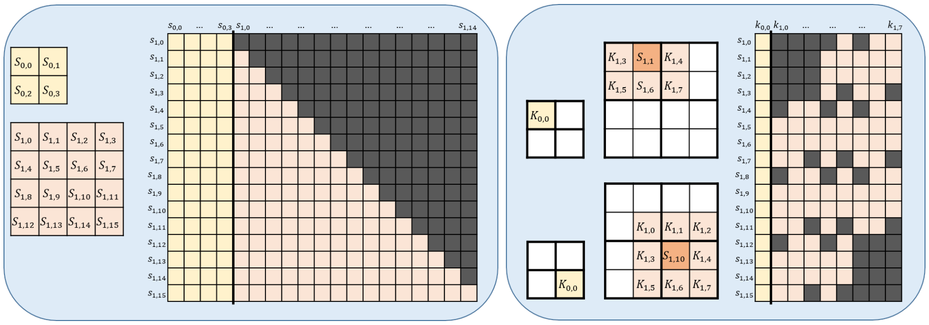
$S_{0,0}$ $S_{0,1}$
$S_{0,2}$ $S_{0,3}$
$S_{1,0}$ $S_{1,1}$ $S_{1,2}$ $S_{1,3}$
$S_{1,4}$ $S_{1,5}$ $S_{1,6}$ $S_{1,7}$
$S_{1,8}$ $S_{1,9}$ $S_{1,10}$ $S_{1,11}$
$S_{1,12}$ $S_{1,13}$ $S_{1,14}$ $S_{1,15}$
$s_{0,0}$ … $s_{0,3}$ $s_{1,0}$ … … … … … … $s_{1,14}$
$s_{1,0}$ $s_{1,1}$ $s_{1,2}$ $s_{1,3}$ $s_{1,4}$ $s_{1,5}$ $s_{1,6}$ $s_{1,7}$ $s_{1,8}$ $s_{1,9}$ $s_{1,10}$ $s_{1,11}$ $s_{1,12}$ $s_{1,13}$ $s_{1,14}$ $s_{1,15}$
$K_{0,0}$
$K_{1,3}$ $S_{1,1}$ $K_{1,4}$
$K_{1,5}$ $K_{1,6}$ $K_{1,7}$
$K_{0,0}$
$K_{1,0}$ $K_{1,1}$ $K_{1,2}$
$K_{1,3}$ $S_{1,10}$ $K_{1,4}$
$K_{1,5}$ $K_{1,6}$ $K_{1,7}$
$k_{0,0}$ $k_{1,0}$ … … … $k_{1,7}$
$s_{1,0}$ $s_{1,1}$ $s_{1,2}$ $s_{1,3}$ $s_{1,4}$ $s_{1,5}$ $s_{1,6}$ $s_{1,7}$ $s_{1,8}$ $s_{1,9}$ $s_{1,10}$ $s_{1,11}$ $s_{1,12}$ $s_{1,13}$ $s_{1,14}$ $s_{1,15}$

Fig. 3. *Left*: Causal Attention matrix as it would be computed in a case with two sequences s_0, s_1 of length 4 and 16, respectively. Each token $s_{1,i}$ can attend to all s_0 and previous $s_{1,j<i}$, according to Equation (9). *Right*: Local attention defined on two sequences of the same length, when considering two kernels 1 × 1 and 3 × 3. The attention matrix reduces to $O(16 \times (1+8))$. The left part shows what indices the tokens $S_{1,1}$ (corner case) and $S_{1,10}$ can attend.

In their paper [20], the authors propose to train M different models (one for each hierarchy), in order to perform autoregressive sampling of the VQ-VAE 2 learned tokens (codebook indices). For sequence s_m, the likelihood of each token i is a function of previous tokens of the same sequence $s_{m,j<i}$ and all tokens of previous sequences $s_{n<m}$:

$$p(s_m) = \prod_i p(s_{m,i} | s_{m,j<i}, s_{n<m}) \tag{9}$$

However, this approach presents a significant challenge as increasing the number of hierarchies and their sequence length requires huge resources in terms of memory storage and sampling time. For instance, the base attention mechanism of Transformers [25] would have a space complexity of $O(m \times N)$ and a sampling time complexity of $O(m)$, where m is the length of the current sequence and N is the sum of the lengths of all sequences up to m (Left part of Figure 3). As a direct consequence, autoregressive models still struggle to synthesise high-resolutions images, and the state-of-the-art in this field is detained by Generative Adversarial Networks [24, 23].
Leveraging the fact that our MS-VQVAE is directly optimized to separate between the coarse and fine-grained details at different scales (due to Equation (7)), we hypothesise that a sampling algorithm based on Transformers [25], would not require all the previous context information for sequences $s_{m>0}$. Instead, one can define m different local kernels $K_{1,\dots,m}$ that would provide all the needed context for the sampling of a token i of sequence s_m, with a significant reduction in terms of memory requirements (see the right part of Figure 3). In Section 4 we provide a proof-of-concept experiment that shows the feasibility of this method.

4 Experiments

MS-VQVAE: We conducted a comparative study of our proposed MS-VQVAE with the existing literature, specifically VQ-VAE 2, on two widely used datasets, namely Imagenet [3] and CelebA-HQ [15], with image resolutions of 256×256 and 1024×1024, respectively. To perform a fair comparison, we replicated and trained the original VQ-VAE 2 model, adopting its original hyperparameters as closely as possible. All our training runs have been trained for 250 epochs employing the Adam optimizer [13] with a learning rate of $3e-4$ and betas 0.9 and 0.5. Moreover, our MS-VQVAE utilized a warmup learning rate for the first five epochs, followed by a cosine decay for the rest of training. For each run, we report several reconstruction metrics evaluated with the *torchmetrics* [17] library, in order to ensure fairness and reproducibility. More in detail, the reconstructions are measured in terms of Mean Squared Error (MSE), Structural Similarity Index [26] (SSIM), Learned Perceptual Image Patch Similarity [30] (LPIPS), Peak Signal-to-Noise Ratio (PSNR) and relative Frechet Inception Distance [8] (rFID). In order to provide a complete comparison, we also report for each experiment the number of training parameters and the codebook usage and perplexity, which indicate the proportion of used codebook vectors and how well they are distributed, respectively.

Table 1 presents the comparison between our MS-VQVAE and the re-implemented VQ-VAE 2 on three experiments which we name as “Large” (L): $8\times8, 16\times16, 32\times32$, “Medium” (M): $16\times16, 32\times32, 64\times64$ and “Small” (S): $32\times32, 64\times64, 128\times128$. In all runs, we utilized a fixed latent vector dimension of 64 and a codebook size of 256, which we determined to be adequate to capture the variances of the dataset.

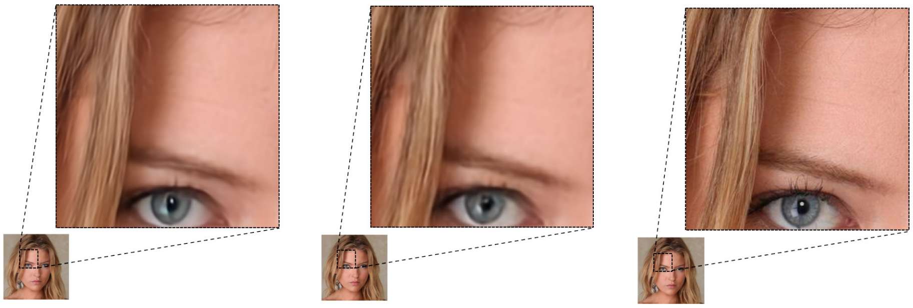

Fig. 4. Qualitative comparison of the "Small" runs on the CelebA-HQ dataset. Our MS-VQVAE (*left*) achieves a very similar reconstruction quality with respect to VQ-VAE 2 (*center*). In both cases, some fine-grained information is lost (e.g. some eye and skin details) when compared to the ground truth (*right*).

Table 1. Experiments on CelebA-HQ (1024 × 1024), grouped by latent sequences length. * reports the mean codebook usage and perplexity over all sequences.

Experiment	# params (M)	CB Usage (%)	Perplexity	MSE ↓ ($1e^{-3}$)	SSIM ↑	LPIPS ↓	PSNR ↑	rFID↓
VQ-VAE 2 (L)	2.2	100-69-18 (62^*)	179-146-35 (120^*)	**2.35**	**0.75**	0.39	**26.28**	**64.47**
Ours (L)	54.3	38-76-100 (**71***)	88-170-163 (**140***)	2.43	**0.75**	**0.38**	26.13	65.05
VQ-VAE 2 (M)	2.2	100-38-7 (48^*)	135-81-11 (75^*)	1.30	0.80	0.28	28.85	38.76
Ours (M)	25.3	100-100-100 (**100***)	222-195-121 (**179***)	**1.15**	**0.81**	**0.25**	**29.36**	**38.38**
VQ-VAE 2 (S)	2.1	100-100-35 (78^*)	110-194-54 (119^*)	0.60	0.88	0.14	32.18	**6.44**
Ours (S)	13.5	67-72-100 (**79***)	142-118-109 (**123***)	**0.52**	**0.90**	**0.13**	**32.79**	6.98

MS-VQVAE achieves comparable or greater performance in all runs, even though it requires more parameters. This gap widens at higher compression rates since more decoding layers are used. Nevertheless, we consider the increased number of parameters to be manageable on most modern hardware, thus not compromising the usability of our model. In general, we have not observed any failures of our method compared to VQ-VAE 2, even at high compression rates. The superior codebook usage in all runs highlights the stability of our method. For the (M) run, VQ-VAE 2 primarily used one sequence, a phenomenon called *codebook collapse*. This does not take place with our approach, as we optimize the reconstruction error directly on each sequence, thereby compelling all the contributions to be present. Avoiding *codebook collapse* is a major challenge in current research as it compromises the stability and usability of the Autoencoder, as discussed in Section 2.

In our study of the Imagenet dataset, we conducted two separate experiments for both methods. The first one ("Small" - (S)) utilized two sequences ($32-64$) and a codebook size of 512, while the second ("Large" - (L)) increased both the compression rate ($16-32$) and codebook size (1024). Table 2 presents a comprehensive comparison between the two methods and shows the superior reconstructions of MS-VQVAE in almost all metrics, especially for higher compression rates. Unlike Table 1, we did not include codebook usage, as it remained at 100 % in all cases. This behavior is likely due to the increased complexity of the dataset (which consists of a wide and diverse range of classes), with respect to CelebA-HQ, which contains only human faces.

Table 2. Experiments on Imagenet (256 × 256), grouped by latent sequences length. * reports the mean perplexity over all sequences.

Experiment	# params (M)	CB Size	Perplexity	MSE ↓ ($1e^{-3}$)	SSIM ↑	LPIPS ↓	PSNR ↑	rFID↓
VQ-VAE 2 (L)	1.4	1024	559-699 (629*)	1.79	0.82	0.195	27.46	22.17
Ours (L)	37.4	1024	815-599 (**707***)	**1.54**	**0.83**	**0.164**	**28.10**	**18.00**
VQ-VAE 2 (S)	1.3	512	273-380 (**327***)	0.59	**0.93**	0.058	32.23	5.39
Ours (S)	9.4	512	371-262 (317*)	**0.50**	**0.93**	**0.046**	**32.95**	**4.49**

For a comprehensive evaluation, we present in Table 3 a comparison of the rFid score for different techniques that employ distinct quantization algorithms and reconstruction losses. A noteworthy observation can be made by comparing the second row, which corresponds to a VQ-GAN model trained on a VQ-VAE 2 pipeline, to the standard VQ-VAE 2 method. The considerable disparity in the rFID score, i.e., 1.45 vs 5.39, highlights the potential of employing the perceptual loss and GAN discriminator for enhancing the quality of reconstructions. This result may suggest the exploration of incorporating these techniques into our MS-VQVAE model in future research.

Table 3. rFid comparison on the validation set of Imagenet for different well-known methods. * indicates a re-trained model. The × in *sequences* column indicates a Residual Quantization, as described in [14]

Method	Sequences	CB size	rFID↓
VQ-GAN [6]	16	1024	7.94
VQ-GAN	32-64	512	1.45
RQ-VAE [14]	8 × 16	16384	1.83
Vit-VQGAN [28]	32	8192	1.28
VQ-VAE 2*	32-64	512	5.39
Ours	32-64	512	4.49

Proof-Of-Concept Image Synthesis: We conduct a proof-of-concept study in order to better enlighten the application of our method in the field of image synthesis. In detail, we train two additional models for 200 epochs (one for VQ-VAE 2 and one for MS-VQVAE) on a 128 px resolution version of the Flickr-Faces-HQ (FFHQ) [24] dataset, and use them as stage-1 models to perform image sampling. In order to reduce the required computational resources, both Autoencoders have highly compressed latent sequences $(8 \times 8, 4 \times 4)$ and a codebook size of 512.

After this first step, we prove the feasibility and advantages of the concepts introduced in Section 3.3 by training different Transformer models [25], each one performing sampling on a specific sequence. For each Autoencoder, we first train a full autoregressive pipeline on both sequences, as depicted in Figure 3 (Left). In the second setup, we implement a MaskGit [2] style transformer to perform image synthesis in a fast way. In particular for the learning of scale one tokens, we designed local-kernel attention in the style of Figure 3 (Right), with kernel dimensions of 3 × 3 and 5 × 5, respectively. In other words, each token of the latent sequence can attend to a maximum of 9 tokens from sequence 0 and 24 tokens from sequence 1, depending on its position. All models are trained for 50 epochs, and have 8 blocks, 8 heads, and a latent dimension of 256.

Table 4. Reported comparison for all the sampling experiments. The Codebook Usage is reported as the mean between sequences.

Method	# params (M)	Autoencoder		Image synthesis	
		CB Usage	rFID ↓	FID ↓	IS ↑
VQ-VAE 2 AR	15.9	61.71%	97.78	137.62	1.83
VQ-VAE 2 MG	14.7			134.97	1.86
MS-VQVAE AR	23.4	75.68 %	97.26	139.29	1.86
MS-VQVAE MG	22.2			133.63	1.86

Table 4 shows the quantitative sampling results in terms of FID and Inception Score (IS) [22] for the Autoregressive (AR) and MaskGit (MG) runs. The cumulative number of parameters (sum Autoencoder and Transformers) is also reported, as well as the mean codebook usage and the reconstruction Fid of the Autoencoders. The results outline how the kernel-based attention can achieve good performances while reducing the space complexity of $O(n^2)$. It is also worth noting that our MS-VQVAE maintains its stability also in this case, with increased codebook usage and comparable performances. Additionally, Figure 5 depicts how our model can be used in order to obtain multiple variations of the same image, by generating the scale 1 samples multiple times.

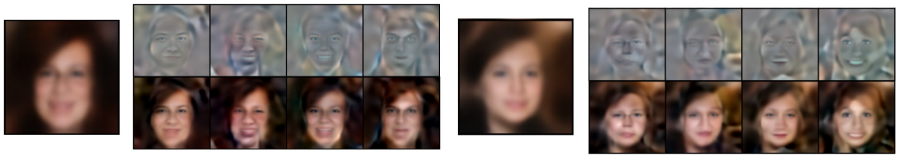

Fig. 5. Samples obtained with the MaskGit-based Transformer and our MS-VQVAE model. For each of the two "coarse" samples (generated from the 4×4 sequence), we sample 4 different contributions (*Top*) from the 8×8 codebook and sum them to obtain the final image (*Bottom*). Note that the low quality is due to the high compression rate of the Autoencoder and the overall small size (number of parameters) of the model.

5 Conclusion and Future Work

In this paper we further developed upon the idea of a hierarchical quantized Autoencoder model, in order to provide more robust and explainable reconstructions. In particular, we showed how the proposed MS-VQVAE can better utilise the information contained in each latent sequence compared to its predecessor [20], while keeping a good reconstruction quality. We also provided a proof-of-concept method that can be useful to perform image synthesis at high resolutions reducing computational costs. We believe that future implementations of this method may be beneficial for upcoming research on autoregressive image synthesis [19, 29, 28, 6, 2], where generating high-resolution images above 256×256 remains a significant challenge.

References

1. Adiban, M., Stefanov, K., Siniscalchi, S.M., Salvi, G.: Hierarchical residual learning based vector quantized variational autoencoder for image reconstruction and

generation. In: British Machine Vision Conference (2022)

2. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11315–11325 (2022)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
4. Dhariwal, P., Jun, H., Payne, C., Kim, J.W., Radford, A., Sutskever, I.: Jukebox: A generative model for music. ArXiv **abs/2005.00341** (2020)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv **abs/2010.11929** (2021)
6. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12868–12878 (2021)
7. Fauw, J.D., Dieleman, S., Simonyan, K.: Hierarchical autoregressive image models with auxiliary decoders. ArXiv **abs/1903.04933** (2019)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS (2017)
9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5967–5976 (2017)
10. Jang, E., Gu, S.S., Poole, B.: Categorical reparameterization with gumbel-softmax. ArXiv **abs/1611.01144** (2017)
11. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016)
12. Łukasz Kaiser, Roy, A., Vaswani, A., Parmar, N., Bengio, S., Uszkoreit, J., Shazeer, N.M.: Fast decoding in sequence models using discrete latent variables. In: International Conference on Machine Learning (2018)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2015)
14. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11513–11522 (2022)
15. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
16. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: A continuous relaxation of discrete random variables. ArXiv **abs/1611.00712** (2017)
17. Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, William Falcon: TorchMetrics - Measuring Reproducibility in PyTorch (2 2022). https://doi.org/10.21105/joss.04101, https://github.com/Lightning-AI/metrics
18. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: NIPS (2017)

19. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
20. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)
21. Roy, A., Vaswani, A., Neelakantan, A., Parmar, N.: Theory and experiments on vector quantized autoencoders. ArXiv **abs/1805.11063** (2018)
22. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
23. Sauer, A., Schwarz, K., Geiger, A.: Stylegan-xl: Scaling stylegan to large diverse datasets. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022)
24. Tero Karras, Samuli Laine, T.A.: A style-based generator architecture for generative adversarial networks. IEEE[Online]. Avaliable: https://ieeexplore.ieee.org/document/8953766 **3** (2019)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
26. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**, 600–612 (2004)
27. Williams, W., Ringer, S., Ash, T., MacLeod, D., Dougherty, J., Hughes, J.: Hierarchical quantized autoencoders. Advances in Neural Information Processing Systems **33**, 4524–4535 (2020)
28. Yu, J., Li, X., Koh, J.Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., Wu, Y.: Vector-quantized image modeling with improved vqgan. ArXiv **abs/2110.04627** (2022)
29. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., Hutchinson, B.C., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., Wu, Y.: Scaling autoregressive models for content-rich text-to-image generation. ArXiv **abs/2206.10789** (2022)
30. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 586–595 (2018)
31. Zheng, C., Vuong, T.L., Cai, J., Phung, D.: Movq: Modulating quantized vectors for high-fidelity image generation. Advances in Neural Information Processing Systems **35**, 23412–23425 (2022)