Channel-Recurrent Autoencoding for Image Modeling

Wenling Shang
University of Amsterdam
wshang@uva.nl

Kihyuk Sohn
NEC Labs
ksohn@nec-labs.com

Yuandong Tian
Facebook AI Research
yuandong@fb.com

Figure 1: Comparison demonstrating our channel-recurrent VAE-GAN's superior ability to model complex bird images. Based on the high-quality generation of Stage1 64×64 images, higher-resolution Stage2 images can be further synthesized unsupervisedly.

Abstract

Despite recent successes in synthesizing faces and bedrooms, existing generative models struggle to capture more complex image types (Figure 1), potentially due to the over-simplification of their latent space constructions. To tackle this issue, building on Variational Autoencoders (VAEs), we integrate recurrent connections across channels *to both inference and generation steps, allowing the high-level features to be captured in global-to-local, coarse-to-fine manners. Combined with adversarial loss, our channel-recurrent VAE-GAN (crVAE-GAN) outperforms VAE-GAN in generating a diverse spectrum of high resolution images while maintaining the same level of computational efficacy. Our model produces interpretable and expressive latent representations to benefit downstream tasks such as image completion. Moreover, we propose two novel regularizations, namely the KL objective weighting scheme over time steps and mutual information maximization between transformed latent variables and the outputs, to enhance the training.*

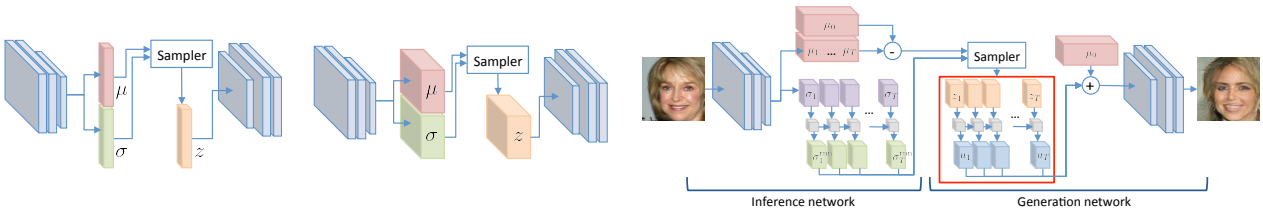
1. Introduction

Tremendous progress has been made in generative image modeling in recent years. Autogressive models, such as PixelRNN [31], describe image densities autoregressively in the pixel level, leading to sharp generations albeit at a high computational cost and without providing latent representations. Generative adversarial networks (GANs) [11] have shown promise, but are limited to modeling high density regions of data distributions [3, 38, 42] and difficult to train [2, 32].

Variational Autoencoder (VAE) [23] is a directed graphical model that approximates a data distribution through a variational lower bound (VLB) of its log likelihood. VAE introduces an approximate posterior parameterized by a deep neural network (DNN) that probabilistically encodes the input to a latent representation. It is directly applicable for a wide range of downstream tasks from photo-editing [16] to policy learning [18], which neither GANs nor autoregressive models are equipped with. Inference in VAE can be done efficiently by a forward pass of the DNN, making it promising for real-time applications. As opposed to GANs, the reconstruction objective of VAE assures a comprehensive input space mode coverage. However, such wide mode coverage, when combined with the KL regularization to the approximated posterior, becomes a downside for modeling complex and high-dimensional images, resulting in blurry image generation [5]. To resolve blurriness while preserving a meaningful latent space, [25] augments a VAE with an auxiliary adversarial loss, obtaining VAE-GAN.

However, recent works for high resolution (64×64 and above) unsupervised image modeling are restricted to images such as faces and bedrooms, whose intrinsic degrees of freedom are low [6, 25, 28]. Once images become spatially and contextually complex, e.g. birds photographed in their natural habitats, aforementioned models struggle to produce sensible outputs (Figure 1), likely due to their latent space constructions lacking the capacity to represent complex input distributions.

In this work, we aim at resolving the limitations of VAE, such as blurry image generation and lack of expressiveness to model complex input spaces, in an unsupervised way. While keeping graphical model unchanged to retain its original efficacy such as efficient probabilistic inference and generation, we propose to augment the architecture of inference and generation networks via recurrent connections across channels of convolutional features, leading to the *channel-*



(a) standard VAE (b) convolutional VAE (c) Proposed channel-recurrent VAE

Figure 2: Illustrations of (a) standard VAE, (b) its convolutional variant cVAE and (c) the proposed crVAE.

recurrent VAE (crVAE). Our approach is motivated by observing a common drawback to VAE and VAE-GAN: the fully-connected (FC) layers between the latent space and convolutional encoder/decoder. Although FC layers can extract abstract information for high-level tasks such as recognition [24], it omits much local descriptions that are essential for detailed image modeling as in our case. Instead, we build latent features on convolutional activation without FC layers. The proposed architecture sequentially feed groups of convolutional features sliced across channels into an LSTM [19], so that for each time step, the associated latent channels are processed based upon accumulated information from previous time steps to ensure temporal coherence while reducing the redundancy and rigidity from FC layers by representing distinguishing information at different time steps. As a result, our model disentangles factors of variation by assigning general outlining to early time steps and refinements to later time steps. Analogously to VAE-GAN, We derive crVAE-GAN by adding an additional adversarial loss, along with two novel regularization methods to further assist training.

We evaluate the performance of our crVAE-GAN in generative image modeling of a variety of objects and scenes, namely birds [4, 40, 41], faces [26], and bedrooms [45]. We demonstrate the superiority of our crVAE-GAN qualitatively via 64×64 image generation, completion, and an analysis of semantic contents for blocks of latent channels, as well as quantitatively via inception scores [35] and human evaluation. Specifically, significant visual enhancement is observed on the more spatially and contextually complex birds dataset. We provide further empirical evidence through higher-resolution (128×128 or 224×224) image synthesis by stacking an extra generation network on top of 64×64 generations from crVAE-GAN and VAE-GAN, similarly to [46]. Unlike [46], the success of generating higher-resolution 2nd stage images without condition variables is heavily dependent on the quality of the 1st stage generations. Our results verify the importance of channel-recurrent architecture in providing a solid 1st stage foundation to achieve high quality 2nd stage generation. Lastly, we remark on the computational virtues of crVAE-GAN.

The merits of our method are summarized as follows:

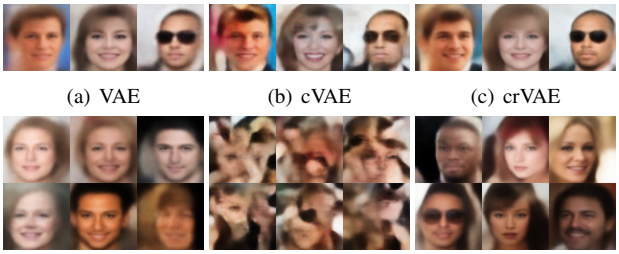
- We integrate temporal structure to the latent space via LSTMs in replacement of the rigid FC layers to recurrently process latent channels, attaining a global-to-local, coarse-to-fine generation.
- Our framework not only preserves the beneficial probabilistic latent space from VAE, allowing wide mode coverage, efficient posterior inference and training, but improves its expressiveness and interpretability.
- Our crVAE-GAN, combined with two novel regularization methods, is capable of modeling complex input spaces when existing models fail. We visually and quantitatively demonstrate significant improvement on high-resolution image generation and related tasks over VAE-GAN.
- Our model, while producing state-of-the-art level image generations, maintains the computational efficacy from VAE.

Code and pretrained models can be found at: https://github.com/WendyShang/crVAE.

2. Related Works

Recent advances in deep generative modeling predominantly come from autoregressive models, Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). Autoregressive models such as PixelRNN and PixelCNN [30, 31, 36] directly characterize the probability density function over the pixel space. Although these models produce sharp images, they have slow inference, demand heavy GPU parallelization for training, and do not explicitly learn a latent space. GAN [11] is another popular method in which a generator competes against a discriminator, producing outputs that imitate the inputs. GANs suffer from several notable issues: limited distribution coverage [3, 38, 42], training instability [2, 27, 32] and lack of probabilistic latent spaces to encode a given input. VAEs [23] consist of a bottom-up inference network and a top-down generation network parameterized by DNNs that are jointly trained to maximize the VLB of the data log-likelihood. Although VAEs are mathematically elegant, easy to train, fast in inference and less GPU demanding than autoregressive models, its KL divergence penalty paired with reconstruction objective hampers realistic image generation since it overly stretches the latent space over the entire training set [5, 38].

Attempts are made to combine the aforementioned methods. PixelVAE [15] integrates PixelCNN into VAE decoder but is still computationally heavy. RealNVP [9] employs an invertible transformation between latent space and pixel



(a) VAE (b) cVAE (c) crVAE

(d) VAE (e) cVAE (f) crVAE

Figure 3: (top) reconstructions with latent variables drawn from the approximated posterior and (bottom) generations from the prior for VAE, convolutional VAE (cVAE), and the proposed crVAE.

space that enables exact log-likelihood computation and inference, but the model is restricted by the invertibility requirement. Adversarial Variational Bayes (AVB) [28] theoretically builds more flexible approximated posterior via adversarial learning but empirically still outputs blurry generations. VAE-GAN [25] stitches VAE with GAN to enhance the generation quality while preserving an expressive latent space without introducing excessive computational overhead. However, VAE-GANs are still not competitive in complex image classes as we note from Figure 1. To tame complex input spaces, recent works [29, 33, 43] leverage side information such as text description, foreground mask and class labels as conditional variables hoping that the consequential conditional distributions are less tangled. But learning a conditional distribution requires additional labeling efforts both at training and downstream applications.

Our approach handles complex input spaces well without the aid of conditional information. It follows the pipeline of VAE-GAN but employs a core channel recurrency to transform convolutional features into and out of the latent space. Such changes are similar in spirit as recent works on improving approximate posterior and prior for VAEs [8, 22, 34], however, we do not change the prior or the posterior to maintain algorithmic simplicity and computational efficiency. Our recurrent module builds lateral connections between latent channels following a similar philosophy as in the deep autoregressive networks (DARN) [14]. But latent variables in DARN are sequentially drawn conditioned on the samples from the previous time steps and thus can be slow in inference. DRAW networks [12, 13] are related to ours as they also recurrently iterate over latent variables for generation. DRAW iterates over the entire latent variables multiple times and incrementally reconstructs pixel-level input at each iteration, whereas we only iterate between blocks of latent channels and reconstruct once, thus it is computationally more efficient and learns interpretable latent subspaces.

3. Channel-Recurrent Autoencoding

This section introduces the proposed channel-recurrent architecture, motivated by observing limitations of the standard VAE and convolutional VAE (cVAE). Then, we extend crVAE-GAN with an adversarial loss to render realistic images. Furthermore, we introduce two latent space regularization techniques specific to our proposed channel-recurrent architecture, namely, the KL objective weighting and the mutual information maximization.

3.1. Latent Space Analysis of VAEs

VAE approximates the intractable posterior of a directed graphical model with DNNs (Figure 2(a)), maximizing a VLB of the data log likelihood:

$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x|z)\big] + D_{\text{KL}}(q_\phi(z|x)\|p(z))\big]$$

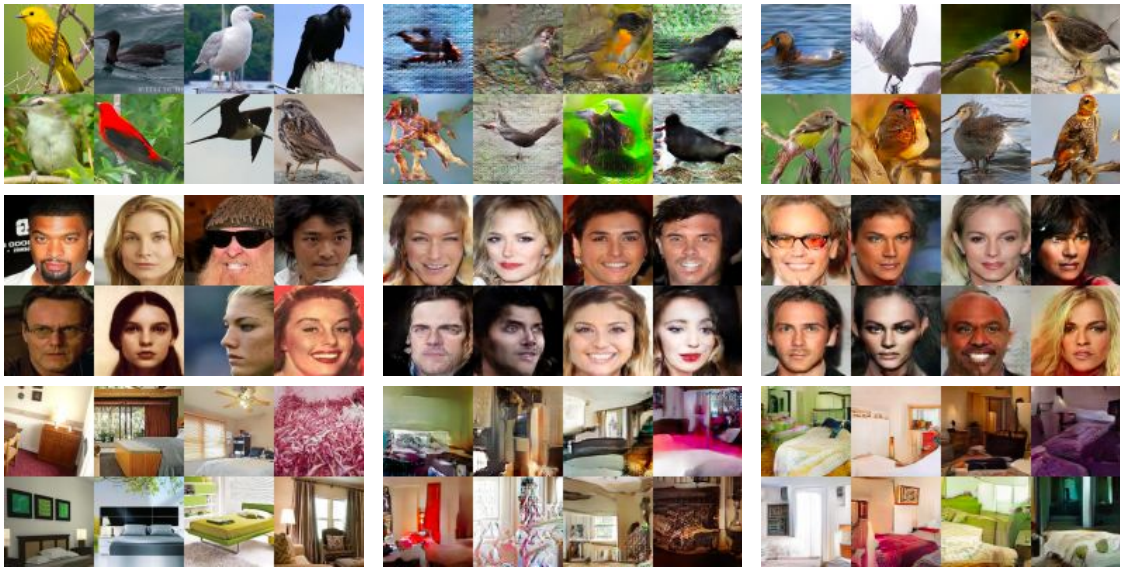
where the approximate posterior $q_\phi(z|x)$ is modeled as a diagonal Gaussian and the prior $p(z)$ as a standard Gaussian. We refer to $q_\phi(z|x)$ as an inference network and $p_\theta(x|z)$ a generation network. The latent space of standard VAE is modeled as 1-dim vector $z\in\mathbb{R}^c$ [23], whereas for cVAE the latent space is modeled as a 3-dim tensor $z\in\mathbb{R}^{w\times h\times c}$ [37].

Overly smoothed reconstructions (Figure 3(a)) and generations (Figure 3(d)) as well as a lack of sample diversity are major downsides of VAEs. A potential cause is that the naive parameterization of the latent space, its associated prior and approximated posterior, may not be able to reflect a complex data distribution [8, 22, 38]. One would be tempted to provide a fix by adding an image-specific prior to the latent space, such as spatial structure, leading to cVAE (Figure 2(b)), whose inference and generation networks, different from standard VAE, are fully convolutional. By making the approximated posterior spatially correlated, cVAE is able to learn with more local details during inference, reflected by higher quality reconstructions (Figure 3(b)). However, the latent variables sampled from spatially independent prior of cVAE ignores the global structure of face shapes and produce chaotic samples (Figure 3(e)).

3.2. Channel-Recurrent Variational Autoencoder

Employing a prior with spatially dependent latent variables, such as a full-covariance Gaussian prior, is one remedy to the problem of chaotic image generation in cVAEs. However, such prior complicates the optimization due to significantly increased number of parameters (e.g., full covariance matrix $\sim O((w\times h\times c)^2)$), especially when the latent space is large [10, 14]. Alternatively, we can structure the covariance to have dense dependencies across the spatial dimension and conditional dependencies along channels, in the hope that such setup can guide each channel to model different aspects of an image.

One possible way to apply the desired structure is to introduce hierarchy to the latent variables, which requires sequential sampling and complicates the training. Thus, tackling from a different direction, we opt to adapt the network architecture. We propose to factorize the convolutional latent space into blocks across channels, flatten the activation to model spatial dependency, and connect the blocks via an LSTM [19] to allow communication and coordination among them for coherency across



(a) ground truth (b) VAE-GAN (c) crVAE-GAN

Figure 4: 64×64 resolution image generation of (top) Birds, (middle) CelebA and (bottom) LSUN using (b) baseline VAE-GAN and (c) our proposed crVAE-GAN along with (a) real examples.

channels. That is, the transformation of the current block of latent channels always takes account of the accumulative information from the preceding time steps, serving as a guidance. Concretely, during generation, $z{=}[z_1, \cdot\cdot, z_T]$ with $z_i{\in}\mathbb{R}^{w\times h\times \frac{c}{T}}$ sampled from standard Gaussian prior is passed through an LSTM to obtain a transformed representation $u = [u_1, \cdot\cdot, u_T]{=}\text{LSTM}(z)$, which is then projected back to the pixel space. Similarly, during inference, the mean path shares the same architecture as in cVAE; the variance path slices latent variables into T blocks of size $w{\times}h{\times}\frac{c}{T}$, each referred to as σ_i and feeds σ_i's into another LSTM to output σ_i^{rnn} as the final variances for the approximate posterior. Our proposed model, referred to as the channel-recurrent VAE (crVAE), can both reconstruct and generate with higher visual quality than VAE and cVAE, as shown in Figure 3(c) and 3(f). More mathematical intuition and details for our design are in the Supplementary Materials.

3.3. Additional Regularization

Inspired by [25], we adopt an adversarial loss to generate realistic images, leading to crVAE-GAN. Additionally, we propose two novel regularizers to enhance the latent space quality of crVAE-GAN for better semantic disentanglement and more stable optimization.

3.3.1 Generating Realistic Images with crVAE-GAN

We extend crVAE to crVAE-GAN with an auxiliary adversarial loss on top of the generation network outputs for realistic image synthesis. The discriminator D maps an image sampled from either the posterior or prior into a binary value:

$$\max_{\phi,\theta} \mathcal{L}_{\text{VAE}} + \beta\mathbb{E}_{z\sim\{q_\phi(z|x),p(z)\}}\left[\log D(p_\theta(x|z))\right]$$

$$\max_{D} \mathbb{E}_{x\sim X}\left[\log D(x)\right] + \mathbb{E}_{z\sim\{q_\phi(z|x),p(z)\}}\left[\log(1{-}D(p_\theta(x|z))\right].$$

Training can be done by min-max optimization as in [25].

Figure 5: (Left) generations from the baseline VAE-GAN without MI regularization has much less artifacts than those from models trained with MI regularization on z (middle) or $\text{FC}_{\text{gen}}(z)$ (right), implying such regularization is not compatible with VAE-GAN.

3.3.2 Weighting the KL Objective

The KL objective of crVAE can be written as follows:

$$\sum_{t=1}^{T}(1-\alpha_t)D_{\text{KL}}(q_\phi(z_t|x)\|p(z_t)), \tag{1}$$

where $\alpha_t{=}0,\ \forall t{\in}\{1,\cdots,T\}$ yields the equivalent expression to standard VAE objective. The channel recurrent architecture additionally enables different weights to regularize the KL objective at each time step. Specifically, noting that the later time steps can be heavily influenced by the earlier ones due to the recurrent connection, we gradually reduce the regularization coefficients of the KL divergence to balance. From an information theoretic perspective [1, 39], the earlier time steps with larger coefficients hold a tighter information bottleneck, meaning that these latent channels shall convey general outlines, whereas the later time steps

with smaller coefficients, i.e., a more flexible information flow, constitute diverse details conditioned on the sketched outlines. Figure 9 demonstrates the resulting effects where earlier time steps output rough profiles and later ones craft the details.

3.3.3 Mutual Information Regularization

To increase training stability, we borrow the idea of mutual information (MI) maximization from [7] as an additional regularization. By recovering the latent variables from the generated image, the generation network encourages the administration of latent information to the output space. The regularization objective is written as:

$$\mathbb{E}_{z\sim\{q_\phi(z|x),p(z)\},\tilde{x}\sim p_\theta(x|z)}\left[q_\psi(z|\tilde{x})\right] \tag{2}$$

where z can be sampled either from approximate posterior $q_\phi(z|x)$ or prior $p(z)$. The CNN encoding path is shared between q_ψ and D as in [7], mapping generated image to reconstruct z and to a binary value, respectively.

Note that the formulation in Equation (2) is not restricted to z, and we empirically found that relating the transformed representation u=$\text{LSTM}_{\text{gen}}(z)$ with the output of q_ψ under crVAE-GAN framework is much more effective. However, similar regularization is found detrimental for VAE-GAN, both with z and the transformed representation $\text{FC}_{\text{gen}}(z)$, as shown in Figure 5. We compare samples generated from the baseline VAE-GAN and those trained with additional regularization to relate z or $\text{FC}_{\text{gen}}(z)$ with the outputs, but the outcomes of the latter are visibly worse. We contemplate that VAE-GAN lacks an equivalent transformation step as in crVAE-GAN via channel-recurrent architecture, which particularly functions to enhance the latent representations. More empirical demonstration are in Section 4.2 and implementation details in the Supplementary Materials.

4. Experiments

For evaluation, we first synthesize 64×64 natural images, followed by a 2nd stage generation to 128×128 or 224×224 on top of the 1st stage generation. To demonstrate the latent space capacity, image completion tasks are performed via optimization based on latent representations. Finally, we explore the semantics of the learned latent channels, particularly with respect to different time steps.

Three datasets, covering a diverse spectrum of contents, are used for evaluation. Birds dataset is composed of three datasets, namely Birdsnap [4], NABirds [40] and Caltech-UCSD Birds-200-2011 [41], containing 106,474 training and 5974 validation images. CelebA [26] contains 163,770 training and 19,867 validation images of face. LSUN bedroom (LSUN) [45] contains 3,033,042 training and 300 validation images. The ROIs are cropped and scaled to 64×64 and 128×128 for Birds and CelebA; the images are scaled and cropped to 64×64 and 224×224 for LSUN. Complete Implementation details are in the Supplementary Materials.

Model (64×64)	Birds	CelebA	LSUN
VAE-GAN	5.81±0.09	21.70±0.15	16.6%
crVAE-GAN	10.62±0.12	24.16±0.33	29.9%
crVAE-GAN + MI	**11.07**±0.12	**26.20**±0.19	**53.5**%
Model (128 or 224)	Birds	CelebA	LSUN
VAE-GAN	14.97±0.11	19.09±0.19	20.5%
VAE-GAN + Perc.	14.61±0.24	27.09±0.26	10.5%
crVAE-GAN	29.14±0.45	**42.66**±0.45	**34.8**%
crVAE-GAN + Perc.	**32.13**±0.37	35.03±0.39	**34.2**%

Table 1: Quantitative evaluation on generating 64×64 and higher-resolution images. For Birds and CelebA, inception scores are reported. For LSUN, the frequency of selection by mechanical turk workers as the most realistc generation among all models is reported.

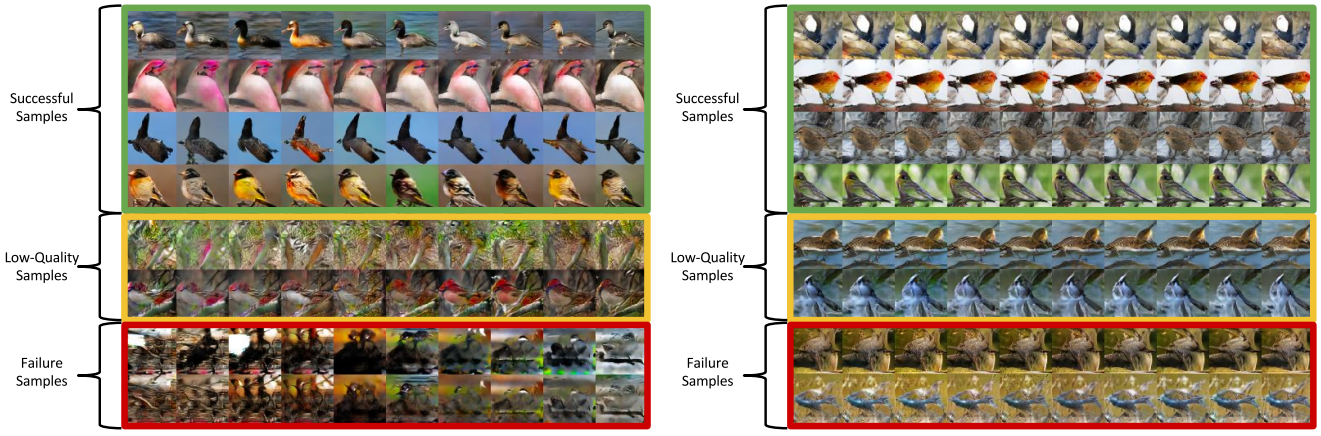
	Occlusion	VAE-GAN	crVAE-GAN
Birds	lower	28.9%	**71.1**%
	upper	35.2%	**64.8**%
CelebA	eye	18.0%	**82.0**%
	mouth	22.7%	**77.3**%
	half	34.4%	**65.6**%
LSUN	center	23.4%	**76.6**%

Table 2: The frequency of selection by mechanical turk workers as the more realistic completion using VAE-GAN and crVAE-GAN.

4.1. Stage 1: 64×64 Image Generation

We compare our crVAE-GAN to the baseline VAE-GAN for 64×64 image generation (Figure 4). This also serves as the first step towards generating higher-resolution images later on. For Birds, VAE-GAN generates colorful images (Figure 4(b)), but details are highly obscured, indicating the latent space conveys mostly low-level information such as color and edges, but not much high-level semantic concepts. By contrast, crVAE-GAN generates significantly more realistic birds with decent diversity in color, background and poses (Figure 4(c)). Generating aligned faces and structured bedrooms are less difficult than birds, but crVAE-GAN still exhibits clear superiority over VAE-GAN.

For quantitative evaluation, we measure inception scores on Birds and CelebA using ImageNet pretrained VGG11 models finetuned on bird and face recognition task [44], respectively. As no bedroom classification dataset is available, we instead conduct a user study for LSUN. We ask a mechanical turk worker to select the most realistic out of 3 generated images (corresponding to VAE-GAN and crVAE-GAN, with and without MI regularization), repeating for 2000 times. We observe improved inception scores in Table 1, which agrees with the visual observation. The frequency of selection of each model by mechanical turk workers on generated LSUN images in Table 1 also verifies that our proposed model outperforms the baseline with a significant margin. More non-curated image samples are in the Supplementary Materials.



(a) crVAE-GAN without MI regularization (b) crVAE-GAN with MI regularization

Figure 6: The same z's are sampled for each row from a standard Gaussian prior and projected back to the pixel space using the snapshot of decoders at last 10 epochs of training. (a) and (b) correspond to crVAE-GAN trained without and with the MI regularization. Top 4 rows are successful samples, 5th and 6th are low-quality ones, and bottom 2 are failure cases. Clear improvement in stability is observed for (b) in terms of color oscillations between consecutive epochs and failure case mode collapsing.

Figure 7: Stage2 generation from 64×64 to 128×128 (Birds and CelebA) and 224×224 (LSUN). Proposed crVAE-GAN provides a solid foundation to assure Stage2 success, clearly contrasting the baseline VAE-GAN.

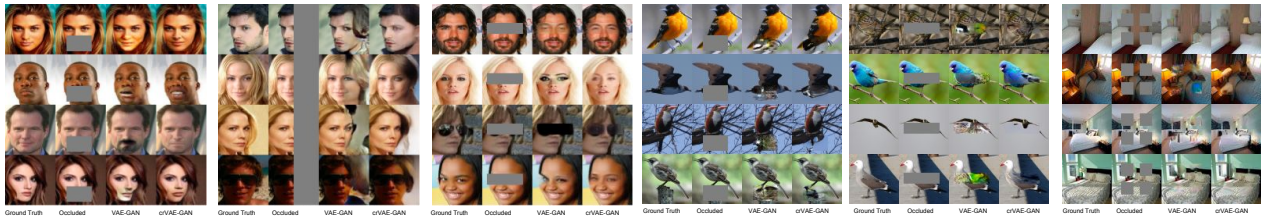
4.2. Effect of Mutual Information Regularization.

We examine the effects of Mutual Information (MI) regularization on training crVAE-GAN by projecting the same z to the output pixel space via generation networks through the last 10 epochs of training. Figure 6(a) and 6(b) show the results without and with MI regularization respectively, where top 4 rows are successful samples, 5th and 6th rows lower quality samples and bottom 2 rows failure cases. First, note that both models supply crisp, coherent and realistic samples when successful and their inception scores are also close (Table 1). Nonetheless, without MI regularization, generated samples between consecutive epochs oscillate and failure cases tend to collapse to the same mode despite originating from different z's. By contrast, with the regularization, the convergence becomes stable and the mode collapsing phenomenon no longer exists even for failed generations. High variance and mode collapsing are two well-known issues of adversarial training [32]. MI maximization overcomes these issues by (1) enforcing the latent messages to be passed to the outputs and (2) regulating the adversarial gradients–recall that MI and adversarial objectives share the same encoding path. Unless specified otherwise, the crVAE-GAN results reported are trained with MI regularization.

4.3. Stage2: Higher Resolution Image Generation

To further assess the quality of Stage1 64×64 generations, we raise the generation resolution to 128×128 for Birds and CelebA,[1] and to 224×224 for LSUN. Our Stage2 generation network is designed similarly to that of StackGAN [46], but generation is done in an unsupervised way without any condition variables. Thanks to the nature of our framework, the Stage1 outputs are composed of both generated and reconstructed samples. We can utilize not only both sources of "fake" images in training but also an additional

[1] We decide to generate higher-resolution images of 128×128 for Birds and CelebA since the ROIs of are approximately of this resolution.



(a) CelebA: mouth, right-half, eyes (b) Birds: lower, upper (c) LSUN: blocks

Figure 8: Image completion with VAE-GAN and crVAE-GAN.

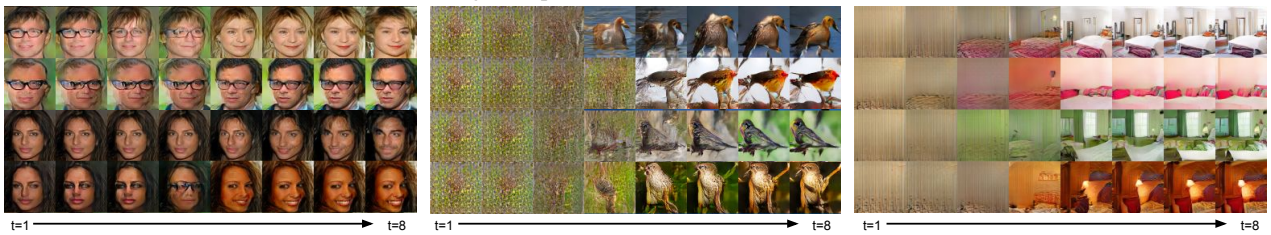


Figure 9: Progressively drawing samples from a standard Gaussian prior over time steps. We observe how image generation evolves by determining global structure at earlier time steps and gradually adding more details later on. Recall the first 3 time steps carry more KL-weight than the rest hence a visual leap occurs around $t = 3$ or 4.

perceptual loss [20] of the reconstructed images to regularize the Stage2 network. Figure 7 presents generated samples from Stage2 networks with and without perceptual loss while taking generation outputs of VAE-GAN and crVAE-GAN Stage1 models as input. Table 1 provides quantitative evaluations following the protocol of Section 4.1. The qualitative and quantitative results imply that a high quality Stage1 generation is essential for Stage2 success, despite that the Stage2 network can correct some of the Stage1 mistakes. Since crVAE-GAN supplies much higher quality Stage1 generations than VAE-GAN, it also produces more visually pleasing Stage2 generations. We also observe that the inception scores in this case can diverge from visual fidelity, e.g. the Stage2 CelebA results with perceptual loss exhibit higher visual quality than without but lower inception scores. Lastly, other mechanisms besides stacking generation networks can be applied to raise image resolution further such as variations of the recently proposed progressive GAN [21], a worthy future direction but out of scope of this paper. More non-curated image samples, details on the Stage2 objectives and model architecture are in the Supplementary Materials.

4.4. Image Completion

To verify how faithfully the latent manifold from crVAE-GAN reflects the semantic meaning of the input space, we conduct image completion task using Stage1 models. We occlude parts of validation images, namely right-half, eye and mouth regions for CelebA, upper and lower part in Birds and blocks for LSUN, and then optimize their latent representations z's to fill in the missing parts [43]:

$$\min_z \ \left[\|\hat{x}\odot m - x\odot m\|_2^2 + \gamma \log \mathcal{N}(z;0,\mathbf{I}) + \tau \log(1-D(\hat{x}))\right],$$

where $\hat{x} = \text{gen}(z)$ refers to the output of the generation network, $m \in \{0,1\}^{3\times 64\times 64}$ is a mask whose entries are 0 if corresponding pixel locations are occluded and 1 otherwise, and $\odot$ represents element-wise multiplication. Qualitative examples are in Figure 8, with more in the Supplementary Materials. VAE-GAN struggles to complete some missing regions, e.g., right half of the faces and sunglasses in Figure 8(a), or generates excessive noise, e.g. the sky in Figure 8(b). By contrast, crVAE-GAN is more competent in retracting off-orbit latent points back to the actual manifold by better embedding the high-level semantic information. Since there may exist multiple solutions in completing an image besides the ground truth, reconstruction error is not an ideal metric for quantitative measurement. Instead, we conduct human evaluation by presenting completed results from VAE-GAN and crVAE-GAN to mechanical turk workers and asking them to select the more realistic one. The selection frequency of each model out of 128 randomly selected pairs is reported in Table 2. Overall our crVAE-GAN again outperforms VAE-GAN with a substantial margin.

4.5. Latent Channel Semantics

Our crVAE-GAN processes the latent variables sequentially, allowing a global-to-local and coarse-to-fine progression. Figure 9 highlights this progression by first initializing all latent variables to be zero, and then gradually sampling blocks of latent variables from the standard Gaussian prior. Due to the weighted KL penalty, the first four time steps tend to operate on the overall tone of an image: defining background color theme, outlining the general shapes, etc. The second half attends the details: the texture of feathers for Birds, facial expressions for CelebA, lighting for LSUN, etc. For a concrete example, the first two rows from CelebA demonstration both starts with an outline of

(a) expression ($t = 8$) (b) azimuth ($t = 5$) (c) background ($t = 5$) (d) belly size ($t = 8$) (e) color ($t = 3$) (f) window ($t = 7$)

Figure 10: Interpolating between z_t and $\tilde{z}_t$, for a selected t, while fixing other z_i's. We observe a gradual shift of an attribute towards the semantic direction encoded by $\tilde{z}_t$ while preserving most of the other factors.

(a) glasses ($t = 4$) (b) hair style ($t = 1$) (c) body size ($t = 6$) (d) color ($t = 6$) (e) linen ($t = 4$) (f) light ($t = 4$)

Figure 11: We manipulate the *style* of a certain attribute by sampling different z_t for a selected t while fixing the rest to demonstrate a multi-dimensional variation as opposed to a mono-dimensional variation such as in β-VAE [17].

men with glasses, then gradually diverge to different hair styles, opposite poses, one taking off while the other solidifying the presence of glasses, and etc. Such progressing phenomenon also suggests that latent channels at different time steps carry their own interpretable semantics.

To investigate this hypothesis, we first draw random samples from the prior across 8 time steps, denoted by $z = [z_1, \cdot\cdot, z_t, \cdot\cdot, z_8]$; we then draw a new sample $\tilde{z} = [z_1, \cdot\cdot, \tilde{z}_t, \cdot\cdot, z_8]$ by only changing a representation at time t that shows semantically meaningful changes from z; finally we generate images by interpolating between z_t and $\tilde{z}_t$ while fixing other z_i's. We note an interesting tendency where images sharing certain characteristics can be manipulated in the same way by interpolating towards the same $\tilde{z}_t$. For example, the CelebA samples in Figure 9 suggest that t=5 decides pose variation; indeed, if two faces both look into the right (Figure 10(b)), traversing z_5's of these faces to the same $\tilde{z}_5$ (selected by visual inspection) will smoothly frontalize their poses while retaining other factors. Similar observations can also be made with Birds and LSUN.

Additionally, we conjecture that our model grants more freedom for semantic variations by associating an explainable factor to a latent subspace rather than a single latent unit [7, 17]. To investigate this hypothesis, we generate several different *styles* of the same factor by sampling different z_t's while fixing the other z_i's. For instance, Figure 11(a) not only takes the glasses on/off but also switches from eyeglasses to sunglasses, from thin frame to thick frame. In comparison, existing works on controlling factors of variation through latent unit manipulation, such as infoGAN [7] and β-VAE [17], can only shift the controlled factor along a single direction, e.g. a latent unit that controls existence of glasses does not allow different glasses styles.

Channel recurrency is not yet perfect in explaining latent semantics. In particular, determining the direction representing a certain factor of variation still requires visual inspection. Nevertheless, the preliminary demonstrations of this intriguing property shed light on future research in learning more semantically meaningful latent spaces.

5. Computations

Another prominent advantage of crVAE-GAN to the baseline as well as other state-of-the-art autoregressive models [22, 31] is found from computational aspects. First, as LSTMs share weights over time, for Stage1 generations, our proposed model has 130M parameters, when the baseline VAE-GAN with the same number of latent variables and the same encoder/decoder architecture has 164M. For the same reason, training our model consumes much less GPU memory. For example, in our implementation, during Stage1 optimization, crVAE-GAN requires around 4.5GB memory with a batch size of 128 while VAE-GAN requires 6.2GB. Finally, the inference and generation complexities for crVAE-GAN is on the same order as those for VAE-GAN. In wall clock time, for a mini-batch of 128 images using a Titan X, crVAE-GAN on average takes 5.8 ms for inference and 4.0 ms for generation; VAE-GAN takes 2.6 ms for inference and 2.2 ms for generation. Meanwhile, autoregressive models are significantly slower in evaluation: even under careful parallelization, it is reported that PixelCNN [31] takes 52K ms and inverse autoregressive flow [22] 50 ms to generate a single 32×32 image on a Titan X.

6. Conclusion

We propose the channel-recurrent autoencoding framework to improve the latent space constructions for image modeling upon the baseline VAE models. We evaluate the performance of our proposed framework via generative image modeling, such as image generation, completion, and latent space manipulation. Future research includes building more interpretable features via channel recurrency and extrapolating our framework to other tasks.

Acknowledgments

We are grateful to Samuel Schulter, Paul Vernaza, Max Welling, Xiang Yu for their valuable comments on the manuscripts. We acknowledge Zeynep Akata for discussions during the preliminary stage of this work. We also thank NVIDIA for the donation of GPUs and Bosch for providing resources.

References

[1] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. *arXiv*, 2016. 4

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv*, 2017. 1, 2

[3] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and equilibrium in generative adversarial nets (gans). *ICML*, 2017. 1, 2

[4] T. Berg, J. Liu, S. Woo Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*, 2014. 2, 5

[5] O. Bousquet, S. Gelly, I. Tolstikhin, C.-J. Simon-Gabriel, and B. Schoelkopf. From optimal transport to generative modeling: the vegan cookbook. *arXiv*, 2017. 1, 2

[6] J. Cha. Implementations of (theoretical) generative adversarial networks and comparison without cherry-picking. https://github.com/khanrc/tf.gans-comparison, 2017. 1

[7] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016. 5, 8

[8] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel. Variational lossy autoencoder. *arXiv*, 2016. 3

[9] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *ICLR*, 2017. 2

[10] J. Duchi. Derivations for linear algebra and optimization. *Berkeley, California*, 2007. 3

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 2

[12] K. Gregor, F. Besse, D. J. Rezende, I. Danihelka, and D. Wierstra. Towards conceptual compression. In *NIPS*, 2016. 3

[13] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. In *ICML*, 2015. 3

[14] K. Gregor, I. Danihelka, A. Mnih, C. Blundell, and D. Wierstra. Deep autoregressive networks. In *ICML*, 2014. 3

[15] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, and A. Courville. Pixelvae: A latent variable model for natural images. *arXiv*, 2016. 2

[16] Q. Guo, C. Zhu, Z. Xia, Z. Wang, and Y. Liu. Attribute-controlled face photo synthesis from simple line drawing. *arXiv*, 2017. 1

[17] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 8

[18] I. Higgins, A. Pal, A. A. Rusu, L. Matthey, C. P. Burgess, A. Pritzel, M. Botvinick, C. Blundell, and A. Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. *ICML*, 2017. 1

[19] S. Hochreiter and J. Schmidhuber. Long short-term memory. In *Neural Computation*, 1997. 2, 3

[20] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 7

[21] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability and variation. *arXiv*, 2017. 7

[22] D. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improving variational inference with inverse autoregressive flow. In *NIPS*, 2016. 3, 8

[23] D. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2013. 1, 2, 3

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2

[25] A. B. L. Larsen, S. K. Sonderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In *ICML*, 2016. 1, 3, 4

[26] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 2, 5

[27] X. Mao, Q. Li, H. Xie, R. Y. Lau, and Z. Wang. Multi-class generative adversarial networks with the l2 loss function. *arXiv*, 2016. 2

[28] L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *ICML*, 2017. 1, 3

[29] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. *ICML*, 2017. 3

[30] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *ICML*, 2016. 2

[31] A. v. d. Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu. Conditional image generation with pixelcnn decoders. *NIPS*, 2016. 1, 2, 8

[32] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016. 1, 2, 6

[33] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *ICML*, 2016. 3

[34] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *ICML*, 2015. 3

[35] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NIPS*, 2016. 2

[36] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv*, 2017. 2

[37] K. Sohn, X. Yan, and H. Lee. Learning structured output representation using deep conditional generative models. In *NIPS*, 2015. 3

[38] L. Theis, A. v. d. Oord, and M. Bethge. A note on the evaluation of generative models. In *ICLR*, 2016. 1, 2, 3

[39] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv*, 2000. 4

[40] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The

fine print in fine-grained dataset collection. In *CVPR*, 2015. 2, 5

[41] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2, 5

[42] Y. Wu, Y. Burda, R. Salakhutdinov, and R. Grosse. On the quantitative analysis of decoder-based generative models. *arXiv*, 2016. 1, 2

[43] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *ECCV*, 2016. 3, 7

[44] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv*, 2014. 5

[45] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv*, 2015. 2, 5

[46] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv*, 2016. 2, 6