CONDITIONAL IMAGE GENERATION BY CONDITIONING VARIATIONAL AUTO-ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a conditional variational auto-encoder (VAE) which, to avoid the substantial cost of training from scratch, uses an architecture and training objective capable of leveraging a foundation model in the form of a pretrained unconditional VAE. Training the conditional VAE then involves training an artifact to perform amortized inference over the unconditional VAE's latent variables given a conditioning input. We demonstrate our approach on the image completion task, and show that it outperforms state-of-the-art GAN-based approaches at faithfully representing the inherent uncertainty. We conclude by describing and demonstrating an application that requires an image completion model with the capabilities ours exhibits: the use of Bayesian optimal experimental design to guide a sensor.

1 INTRODUCTION

A major challenge with applying variational auto-encoders (VAEs) to high-dimensional data is the typically slow training times. For example, training a state-of-the-art VAE (Vahdat & Kautz, 2020; Child, 2020) on the 256×256 FFHQ dataset (Karras et al., 2019) takes on the order of 1 GPU-year, but a state-of-the-art generative adversarial network (GAN) (Lin et al., 2021; Karras et al., 2020) can be trained on the same dataset in a matter of GPU-weeks. One hypothesis for the cause of this disparity is that, whereas the "mass-covering" training objective for a VAE forces it to assign probability mass over the entirety of the data distribution, a GAN can "cut corners" by dropping modes (Arora & Zhang, 2017; Arora et al., 2017).

We focus on the problem of *conditional* generative modelling: given an input (e.g. a partially blanked-out image), we wish to map to a distribution over outputs (e.g. plausible completions of the image). Both conditional GANs (Zheng et al., 2019; Zhao et al., 2021) and conditional VAEs (Sohn et al., 2015; Ivanov et al., 2018) are applicable to this problem, with the same disparity in training times that we described for their unconditional counterparts. We present an approach based on the conditional VAE framework but, to mitigate the associated slow training times, we design the architecture so that we can incorporate pretrained unconditional VAEs. We show that re-using publicly available pretrained models in this way can lead to training times competitive with GANs.

Figure 1: **Left column:** Images with most pixels masked out. **Rest:** Completions from our method.

While requiring an existing pretrained model is a limitation, we note that: **(I)** The unconditional VAE need not have been (pre-)trained on the same dataset as the conditional model; we show unconditional models trained on ImageNet are suitable for later use with various photographic image datasets. **(II)** A single unconditional VAE can be used for later training of conditional VAEs on any desired conditional generation tasks (e.g. the same image model may be later used for image completion or image colourisation). **(III)** There is an increasing trend in the machine learning community towards sharing large, expensively trained models (Wolf et al., 2020), sometimes referred to as foundation models (Bommasani et al., 2021). Most of the unconditional VAEs in our experiments use publicly-available pretrained weights released by Child (2020). By presenting a use case for foundation models in image modelling, we hope to encourage even more sharing of pretrained weights in this domain.

We demonstrate our approach on several conditional generation tasks in the image domain but focus in particular on stochastic image completion: the problem of inferring the posterior distribution over images given the observation of a subset of pixel values. The visual quality of image completions produced by our method (see Fig. 1) is close to the state-of-the-art (Zhao et al., 2021), and we show results indicating that our coverage of the "true" posterior over image completions is superior to that of any of our baselines.

For some applications such as photo-editing the implicit distribution defined by GANs is likely to be good enough. We argue that our approach has substantial advantages when image completion is used as part of a larger pipeline. We demonstrate one such situation in Section 5: Bayesian optimal experimental design (BOED) for guiding a sensor or hard attention mechanism (Ma et al., 2018; Harvey et al., 2019; Rangrej & Clark, 2021). In this case, missing modes of the posterior over images is likely to lead to bad decisions. We show that our objective corresponds to the mass-covering KL divergence and therefore covers the posterior well, and outperforms GANs which typically learn a distribution with a fraction of the support of the data distribution (Arora & Zhang, 2017; Arora et al., 2017).

Contributions We develop a method to cheaply convert pretrained unconditional VAEs into conditional VAEs. We demonstrate that this allows conditional VAE training times competitive with GANs, while avoiding the mode-dropping behaviour associated with GANs. Finally, we showcase an application in Bayesian optimal experimental design that benefits from these capabilities.

2 Variational Auto-encoders

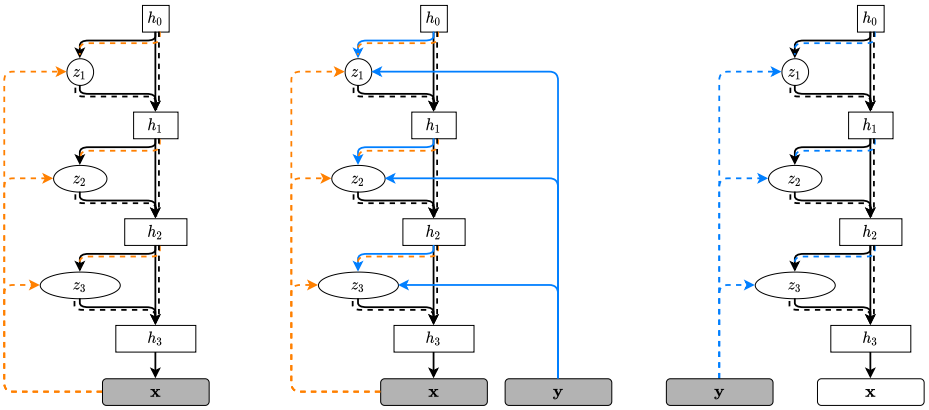
We describe VAEs in terms of three components. **(I)** A decoder with parameters $\theta \in \Theta$ maps from latent variables z to a distribution over data $\mathbf{x}$, which we call $p_{\text{model}}(\mathbf{x}|z;\theta)$. **(II)** There is a prior over latent variables, $p_{\text{model}}(z;\theta)$. This may have learnable parameters, which we consider to be part of θ. Together, the prior and decoder define a joint distribution, $p_{\text{model}}(z,\mathbf{x};\theta)$. Finally, **(III)** an encoder with parameters $\phi \in \Phi$ maps from data to an approximate posterior distribution over latent variables, $q(z|\mathbf{x};\phi) \approx p_{\text{model}}(z|\mathbf{x};\theta)$. Ideally, θ would be learned to maximise the log likelihood $\log p_{\text{model}}(\mathbf{x};\theta) = \log \int p_{\text{model}}(z,\mathbf{x};\theta)\mathrm{d}z$, averaged over training examples. Since this is intractable, θ and ϕ are instead trained jointly to maximise an average of the evidence lower-bound (ELBO) over each training example $\mathbf{x} \sim p_{\text{data}}(\cdot)$:

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\text{ELBO}(\theta,\phi,\mathbf{x})\right] = \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\mathbb{E}_{q(z|\mathbf{x};\phi)}\left[\log \frac{p_{\text{model}}(z;\theta)p_{\text{model}}(\mathbf{x}|z;\theta)}{q(z|\mathbf{x};\phi)}\right] \tag{1}$$

$$= -\mathcal{H}\left[p_{\text{data}}(\mathbf{x})\right] - \text{KL}\big(p_{\text{data}}(\mathbf{x})q(z|\mathbf{x};\phi) \;\big\|\; p_{\text{model}}(z,\mathbf{x};\theta)\big). \tag{2}$$

The data distribution's entropy, $\mathcal{H}\left[p_{\text{data}}(\mathbf{x})\right]$, is typically a finite constant, and this is guaranteed in our experiments where $\mathbf{x}$ is an image with discrete pixel values. Maximising the above objective will therefore drive $p_{\text{model}}(z,\mathbf{x};\theta)$ towards $p_{\text{data}}(\mathbf{x})q(z|\mathbf{x};\phi)$, and so the marginal $p_{\text{model}}(\mathbf{x};\theta)$ towards $p_{\text{data}}(\mathbf{x})$. The KL divergence shown leads to mass-covering behaviour from $p_{\text{model}}(z,\mathbf{x};\theta)$ (Bishop, 2006) so $p_{\text{model}}(\mathbf{x};\theta)$ should assign probability broadly over the data distribution $p_{\text{data}}(\mathbf{x})$. For notational simplicity in the rest of the paper, parameters θ and ϕ are not written when clear from the context.

Hierarchical VAEs (Gregor et al., 2015; Kingma et al., 2016; Sønderby et al., 2016; Klushyn et al., 2019) partition the latent variables z in a way which has been found to improve the fidelity of



(a) Estimating ELBO. (b) Estimating $\mathcal{O}_{\text{for}}$. (c) Sampling $\mathbf{x} \sim p_{\text{cond}}(\cdot|\mathbf{y})$.

Figure 2: A hierarchical VAE architecture with $L = 3$ groups of latent variables. The blocks $h_0, \ldots, h_L$ represent the deterministic hidden state of the decoder. Part (a) shows the computations involved in computing the ELBO for an unconditional VAE. The encoder is shown in orange, and the prior and decoder are shown in black. The dashed lines show the dependencies of the distribution from which z is sampled. Part (b) shows the computations used to compute our training objective $\mathcal{O}_{\text{for}}$. This involves the partial encoder, shown in blue. Part (c) shows the computation graph used when sampling image completions.

the learned $p_{\text{model}}(\mathbf{x})$, especially for the image domain (Vahdat & Kautz, 2020; Child, 2020). In particular, they define z to consist of L disjoint groups, $z_1, \ldots, z_L$. The prior for each z_l can depend on the previous groups through the factorisation

$$p_{\text{model}}(z) = \prod_{l=1}^{L} p_{\text{model}}(z_l|z_{<l}). \tag{3}$$

where $z_{<l}$ is the null set for $l = 1$ and $\{z_1, \ldots, z_{l-1}\}$ otherwise. Fig. 2a shows the hierarchical VAE architecture we base this work on, in which the dependency of the prior for each z_l on $z_{<l}$ is maintained via the decoder's (shown in black) hidden state h_l. The distribution produced by the encoder (shown in orange) for each z_l also depends on the previous hidden state h_{l-1} and therefore factorises as $q(z|\mathbf{x}) = \prod_{l=1}^{L} q(z_l|z_{<l}, \mathbf{x})$. We will parameterise $p_{\text{model}}(z_l|z_{<l})$ and $q(z_l|z_{<l}, \mathbf{x})$ as diagonal Gaussian distributions, as is common for hierarchical VAEs (Sønderby et al., 2016; Vahdat & Kautz, 2020; Child, 2020).

3 Amortized Inference in a Pretrained Artifact

To convert an unconditional VAE architecture to a conditional architecture, we introduce a *partial encoder* with parameters $\hat{\phi} \in \hat{\Phi}$. This is fed a conditioning input $\mathbf{y}$. For example, in the case of image completion, $\mathbf{y}$ could be an image with some pixels masked out. The partial encoder then defines an approximate posterior over the latent variables, $\hat{q}(z|\mathbf{y}; \hat{\phi})$.

The conditional generation task is to approximate $p_{\text{data}}(\mathbf{x}|\mathbf{y})$. Using the partial encoder, we define

$$p_{\text{cond}}(\mathbf{x}|\mathbf{y}; \theta, \hat{\phi}) := \int p_{\text{model}}(\mathbf{x}|z; \theta)\hat{q}(z|\mathbf{y}; \hat{\phi})\mathrm{d}z \tag{4}$$

with learnable parameters θ and $\hat{\phi}$. We can sample from $p_{\text{cond}}(\mathbf{x}|\mathbf{y})$ by sampling $z \sim \hat{q}(\cdot|\mathbf{y})$ and then $\mathbf{x} \sim p_{\text{model}}(\cdot|z)$ as shown in Fig. 2c. This essentially defines a conditional VAE architecture which, unique amongst related work with high-dimensional $\mathbf{x}$ and $\mathbf{y}$ (Sohn et al., 2015; Zheng et al., 2019; Ivanov et al., 2018; Wan et al., 2021), has a decoder $p_{\text{model}}(\mathbf{x}|z; \theta)$ with no dependence on $\mathbf{y}$. This decoder can therefore use an architecture identical to that of an unconditional VAE and also, as we will show later, re-use unconditional VAE weights.

Before describing our method in further detail, we briefly introduce some notation. Let the distribution of paired data be $p_{\text{data}}(\mathbf{x}, \mathbf{y})$. Then recall that training an unconditional VAE matches two joint distributions: the distribution of samples from the generator, $p_{\text{model}}(z, \mathbf{x})$; and the distribution resulting from sampling data $\mathbf{x}$ and encoding it, $p_{\text{data}}(\mathbf{x})q(z|\mathbf{x})$. For notational convenience, we define the following extensions of these joint distributions to include $\mathbf{y}$:

$$p_{\text{model}}(z, \mathbf{x}, \mathbf{y}; \theta) = p_{\text{model}}(z; \theta)p_{\text{model}}(\mathbf{x}|z; \theta)p_{\text{data}}(\mathbf{y}|\mathbf{x}), \tag{5}$$

$$r(z, \mathbf{x}, \mathbf{y}; \phi) = p_{\text{data}}(\mathbf{x}, \mathbf{y})q(z|\mathbf{x}; \phi), \tag{6}$$

where $p_{\text{data}}(\mathbf{y}|\mathbf{x})$ is a (potentially intractable) conditional distribution under $p_{\text{data}}(\mathbf{x}, \mathbf{y})$. Note that $p_{\text{model}}(z, \mathbf{x}, \mathbf{y}; \theta)$ and $r(z, \mathbf{x}, \mathbf{y}; \phi)$ are exactly the two distributions matched by the unconditional VAE objective in Eq. (2) with an additional factor of $p_{\text{data}}(\mathbf{y}|\mathbf{x})$. Therefore, if the unconditional VAE represented by θ and ϕ is well trained, p_{model} and r will be close. From now on, we will use p_{model} and r to refer to any marginals and conditionals of the above joint distributions, with the specific marginal or conditional clear from context.

3.1 Training objective

Our training objective, previously used for training conditional VAEs (Sohn et al., 2015; Ivanov et al., 2018) and neural processes (Garnelo et al., 2018), is

$$\mathcal{O}_{\text{for}}(\theta, \phi, \hat{\phi}) = \mathbb{E}_{p_{\text{data}}(\mathbf{x},\mathbf{y})}\mathbb{E}_{q(z|\mathbf{x})}\left[\log \frac{p_{\text{model}}(\mathbf{x}|z)\hat{q}(z|\mathbf{y})}{q(z|\mathbf{x})}\right] \quad \leq \mathbb{E}_{p_{\text{data}}(\mathbf{x},\mathbf{y})}\left[\log p_{\text{cond}}(\mathbf{x}|\mathbf{y})\right]. \tag{7}$$

This lower-bounds $\log p_{\text{cond}}(\mathbf{x}|\mathbf{y})$ similarly to how the ELBO in an unconditional VAE lower-bounds $\log p_{\text{model}}(\mathbf{x})$. The only difference is that the prior, $p_{\text{model}}(z)$, is replaced by $\hat{q}(z|\mathbf{y})$. This is reflected in Fig. 2b, where each z_l is conditioned on $\mathbf{y}$ via the partial encoder (blue).

We are particularly interested in the properties of the learned partial encoder. Recall the joint distribution $r(z, \mathbf{x}, \mathbf{y}; \phi) = p_{\text{data}}(\mathbf{x}, \mathbf{y})q(z|\mathbf{x}; \phi)$. Then $r(z|\mathbf{y}; \phi)$ is the intractable posterior given by marginalising out $\mathbf{x}$ and conditioning on $\mathbf{y}$. We find that fitting $\hat{\phi}$ to maximise $\mathcal{O}_{\text{for}}(\theta, \phi, \hat{\phi})$ is equivalent to minimising the KL divergence from $r(z|\mathbf{y}; \phi)$ to $\hat{q}(z|\mathbf{y}; \hat{\phi})$. We formalise this statement in the following theorem, which is proven in Appendix B.

Theorem 3.1. *For any set $\hat{\Phi}$ of permissible values of $\hat{\phi}$, and for any $\theta \in \Theta$ and $\phi \in \Phi$,*

$$\arg\max_{\hat{\phi}\in\hat{\Phi}} \mathcal{O}_{\text{for}}(\theta, \phi, \hat{\phi}) = \arg\min_{\hat{\phi}\in\hat{\Phi}} \mathbb{E}_{p_{\text{data}}(\mathbf{y})}\left[KL\big(r(z|\mathbf{y}; \phi) \;\|\; \hat{q}(z|\mathbf{y}; \hat{\phi})\big)\right]. \tag{8}$$

Learning $\hat{q}(z|\mathbf{y}; \hat{\phi})$ to minimise this "forward" KL divergence leads to mass-covering behaviour (Bishop, 2006), and so the learned $\hat{q}(z|\mathbf{y}; \hat{\phi})$ should have good coverage of $r(z|\mathbf{y}; \phi)$. Samples of the latent variables $z \sim \hat{q}(\cdot|\mathbf{y}; \hat{\phi})$, and subsequently samples of $\mathbf{x}$ given z, are therefore likely to be diverse with good coverage of the "true" posterior $p_{\text{data}}(\mathbf{x}|\mathbf{y})$.

3.2 Faster training with a pretrained VAE

To justify using weights trained as part of an unconditional VAE we make the following observation.

Theorem 3.2. *Assume we have a sufficiently expressive encoder and decoder that there exist parameters $\theta^* \in \Theta$ and $\phi^* \in \Phi$ which make the unconditional VAE objective (Eq.* (1)*) equal to its upper bound of $-\mathcal{H}\left[p_{\text{data}}(\mathbf{x})\right]$. Then, given a sufficiently expressive partial encoder,*

$$\max_{\hat{\phi}} \mathcal{O}_{\text{for}}(\theta^*, \phi^*, \hat{\phi}) = \max_{\theta,\phi,\hat{\phi}} \mathcal{O}_{\text{for}}(\theta, \phi, \hat{\phi}).$$

See Appendix B for a proof. This implies that we can use values of θ and ϕ learned using the unconditional VAE objective. Then to train a conditional generative model we need only optimise $\hat{\phi}$. This leads to faster convergence, as well as faster training iterations since we only need to compute gradients for, and perform update steps on, the partial encoder's parameters $\hat{\phi}$. For all of our experiments in Section 4 we use pretrained models released by Child (2020), leveraging between

Table 1: Image completion results. Best performance is shown in **bold**, and second best is underlined. In the last row, t denotes the "temperature" parameter (Child, 2020).

	CIFAR-10			FFHQ-256		
Method	FID↓	P-IDS↑	LPIPS-GT↓	FID↓	P-IDS↑	LPIPS-GT↓
ANP	30.03	5.86	.0447	39.95	0.93	.256
CE	21.92	4.77	.0628	39.02	0.66	.267
RFR	44.35	2.76	.0883	72.50	0.46	.271
PIC	14.73	5.95	.0332	11.60	2.76	.169
CoModGAN	9.65	11.59	.0326	**2.33**	**13.57**	.143
IPA-R	19.21	8.56	.0330	8.82	4.56	.142
IPA (ours)	10.50	13.24	**.0262**	3.93	7.79	.123
IPA ($t = 0.85$, ours)	**8.61**	**14.19**	.0263	3.29	8.50	**.117**

2 GPU-weeks and 1 GPU-year of unconditional VAE training for each dataset. We name our method IPA (Inference in a Pretrained Artifact).

Note that Theorem 3.2 applies only if the unconditional VAE parameters are learned on the same dataset as the conditional VAE is trained on; otherwise there will be a mismatch between the form of p_{data} used in Eq. (1) to fit θ^* and ϕ^*, and the form of p_{data} implicit in the $\mathcal{O}_{\text{for}}$ objective. However we find empirically that we can use unconditional VAE parameters trained on ImageNet (Deng et al., 2009) with IPA on several other photographic image datasets.

4 Experiments

Comparison to image completion baselines We create an IPA image completion model based on the VD-VAE unconditional architecture (Child, 2020), and evaluate it for image completion on two datasets, CIFAR-10 (Krizhevsky et al., 2009) and FFHQ-256 (Karras et al., 2019). We compare against four baselines: Co-Modulated Generative Adversarial Networks (CoModGAN) (Zhao et al., 2021); Pluralistic Image Completion (PIC) (Zheng et al., 2019); Context Encoders (CE) (Pathak et al., 2016); and Attentive Neural Processes (ANP) (Kim et al., 2019). We report some results for another two baselines: we show qualitative results for VQ-VAE (Peng et al., 2021), but not quantitative results because it takes too long (about a minute) to complete each test image. We report results for Recurrent Feature Reasoning for Image Inpainting (RFR) (Li et al., 2020) but with the caveat that it is slow to run on images with many missing pixels and so, although it used a similar computational budget to the other models, its training did not converge.

Given pretrained unconditional VAE parameters, IPA is faster to train than the best-performing baseline, CoModGAN. IPA takes 115 GPU-hours to train on CIFAR-10, and under 7 GPU-weeks on FFHQ-256. The CoModGAN models are trained for 270 GPU-hours and 8 GPU-weeks respectively. We provide more training details in Appendix D.

We report the FID (Heusel et al., 2017) and P-IDS (Zhao et al., 2021) metrics between a set of sampled completions from each method and a reference set. Broadly speaking, these measure the sample quality. To investigate the diversity of samples, and their ability to capture all modes of $p_{\text{data}}(\mathbf{x}|\mathbf{y})$, we also report the LPIPS-GT. We compute this using LPIPS (Zhang et al., 2018), a measure of distance between two images. Specifically, we compute the average over test pairs $(\mathbf{x}, \mathbf{y})$ of $\min_{k=1}^{K}(\text{LPIPS}(\mathbf{x}^{(k)}, \mathbf{x}))$, with each $\mathbf{x}^{(k)} \sim p_{\text{cond}}(\cdot|\mathbf{y})$. As $K \to \infty$, the LPIPS-GT should tend to zero if the ground truth completion is always within the support of $p_{\text{cond}}(\mathbf{x}|\mathbf{y})$. If not, the LPIPS-GT will remain high, penalising methods which miss modes of the posterior. We use $K = 100$.

For the image completion tasks, we sample from $p_{\text{data}}(\mathbf{x}, \mathbf{y})$ by first sampling an image $\mathbf{x}$ from the dataset, and then sampling an image-sized binary mask m from the freeform mask distribution used by Zhao et al. (2021), which is itself based on Yu et al. (2018). We then set $\mathbf{y} = \texttt{concatenate}(\mathbf{x} \odot m, m)$. Here, $\odot$ is a pixel-wise multiplication operation which removes information from the missing pixels. The concatenation is performed along the channel dimension and makes it possible to distinguish between unobserved pixels and zero-valued pixels.

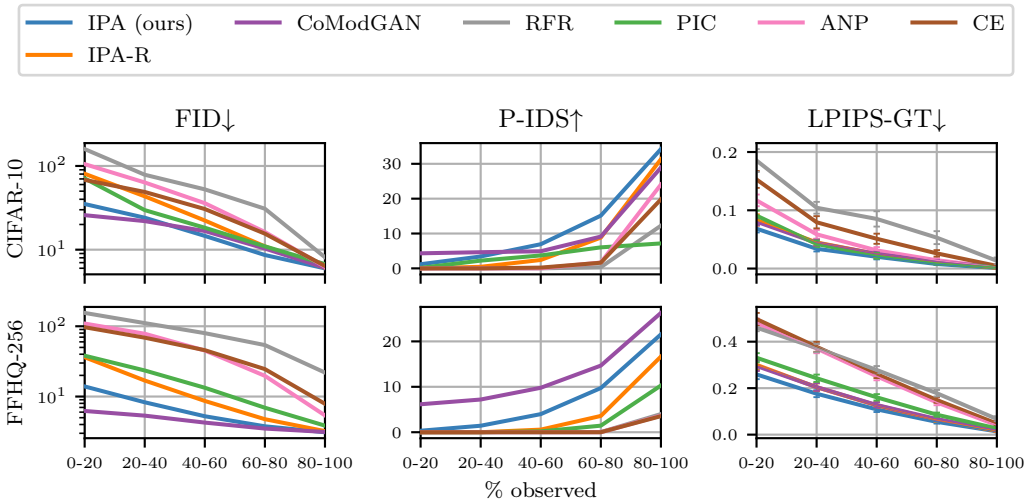


Figure 3: Our three test metrics evaluated for CIFAR-10 (top row) and FFHQ-256 (bottom row), and plotted as a function of the mask distribution. The error bars on LPIPS-GT show the standard error of our estimate for a single trained network.

For evaluation, since the number of observed pixels in freeform masks varies considerably, we follow Zhao et al. (2021) and partition the mask distribution by conditioning the procedure to return a mask with the proportion of pixels observed within some range (0-20%, 20-40%, and so on) and report metrics for each range separately in Fig. 3. To summarise the overall performance in Table 1, we sample masks from a uniformly-weighted mixture distribution over these five partitions.

In terms of the LPIPS-GT scores in Table 1, IPA outperforms the best baselines by roughly 20%. Figure 3 shows that there is an improvement for any proportion of observed pixels. This suggests that IPA produces reliably diverse samples with good coverage of $p_{\text{data}}(\mathbf{x}|\mathbf{y})$. In contrast, we believe that the GAN-based approaches occasionally miss modes of $p_{\text{data}}(\mathbf{x}|\mathbf{y})$ and can therefore fail to capture the ground-truth. This hypothesis is supported by samples from CoModGAN we display in Appendix G. In terms of sample fidelity, as measured by both FID and P-IDS, IPA outperforms all baselines on CIFAR-10 when $> 40\%$ of the image is observed, and comes second to CoModGAN when $< 40\%$ is observed and on FFHQ-256.

Edges-to-photos We provide an additional demonstration of IPA on the Edges2Shoes and Edges2Handbags datasets (Isola et al., 2016), where the task is to generate an image conditioned on the output of an edge detector applied to that image. We downsample the datasets to 64×64 so that we can use unconditional VAEs pretrained on ImageNet (Deng et al., 2009) at this resolution by Child (2020). We show in Fig. 4 that IPA is useful for these tasks, and provide further discussion below. The images generated are diverse and photorealistic, as shown in Appendix H.

Effectiveness of pretraining We now seek to answer the question of how important the pretrained unconditional VAE weights are to IPA. To do so, we compare IPA with conditional VAEs which use the same architecture as IPA but are trained from scratch, and which we will refer to as "from-scratch" baselines. That is, θ and ϕ are randomly initialised and trained to maximise Eq. (7) along with $\hat{\phi}$.

With an infinite training budget, the end-to-end training of the from-scratch baselines is likely to lead them to outperform any IPA models. Nevertheless it is apparent from Fig. 4 that, in the more realistic situation of a finite training budget, using IPA can be beneficial. This is the case even for training budgets of up to a few GPU-weeks on the relatively small CIFAR-10 dataset. In fact, even with only a couple of days of training, IPA on CIFAR-10 (with CIFAR-10 pretraining) achieves better FID and ELBO scores than the from-scratch baseline trained for several weeks.

For Edges2Handbags and Edges2Shoes, training with IPA for 2 days yields performance similar to or better than training with the from-scratch baseline for 1 week, as measured by the ELBO. This

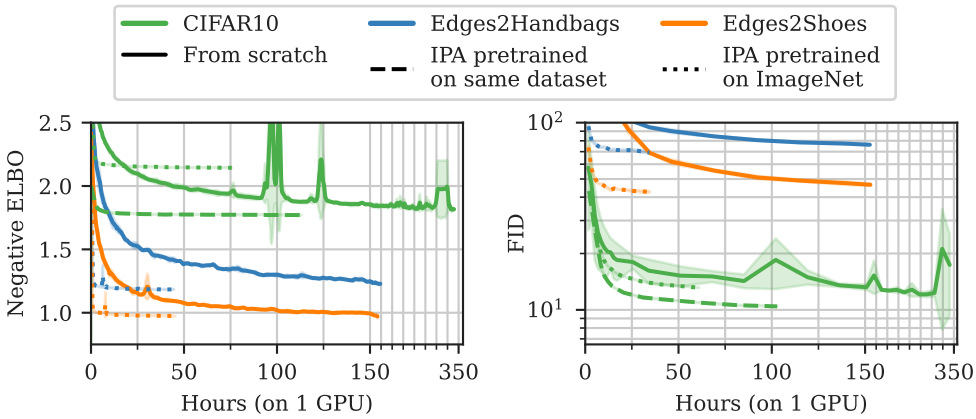


Figure 4: ELBO and FID during training on CIFAR-10 and edges-to-photos using IPA with pretraining on the same dataset, IPA with pretraining on ImageNet, and when trained from scratch. Error bars show standard deviations computed with 3 runs. IPA makes training faster and lower-variance.

is despite IPA on these datasets using a trained ImageNet64 model rather than a model pretrained on those specific datasets, supporting our suggestion that the dataset used for pretraining need not exactly match what IPA is then trained on. When measuring performance with the FID score, IPA looks even more appealing: wherever ELBOs are similar between IPA and the from-scratch baselines, IPA achieves a significantly better FID score. We see that IPA pretrained on ImageNet is less effective for CIFAR-10 than it is for the edges-to-photos datasets, but it nevertheless improves on the from-scratch baseline in terms of ELBO for the first 36 hours of training, and in terms of FID until the from-scratch baseline is trained for at least a week.

An alternative training objective In Table 1 and Fig. 3, we report results for IPA-R, a variation of IPA with a different training objective corresponding to a mode-seeking KL divergence. IPA almost always outperforms IPA-R, but we nonetheless provide a full description of IPA-R in Appendix C.

5 Application: Experimental Design for Medical Imaging

In this section, we demonstrate an application where stochastic image completion, and faithful representation of the posterior $p_{\text{data}}(\mathbf{x}|\mathbf{y})$, is necessary. In particular, we consider whether it is possible to automatically target a chest x-ray at areas most likely to reveal abnormalities. This could avoid the need to scan the entire chest and so bring benefits including reducing the patient's radiation exposure. We do not claim that our system, as it is now, is suitable for use in a clinical setting but believe this is a worthwhile avenue to explore. Specifically, we consider performing a series of x-ray scans, each targeted at only a small portion of the area of interest. We can select the coordinates $c_t = (x_t, y_t)$ of the location to scan at each step t, and this selection can be informed by what was observed in the previous scans. The task we consider is how to select c_t to be maximally informative. In particular, assume we wish to infer a variable v representing, e.g., whether the patient has a particular illness. Bayesian optimal experimental design (BOED) (Chaloner & Verdinelli, 1995) provides a framework to select a value of c_t that is maximally informative about v. It involves taking a Bayesian perspective on the problem of estimating v. We have one posterior distribution over v after taking scans at $c_1, \ldots, c_{t-1}$ and another (typically lower entropy) distribution after conditioning on a scan at c_t as well. The *expected information gain*, or EIG, quantifies the utility of the choice of c_t as the expected difference in entropy between these two distributions. Using BOED involves estimating the EIG and selecting the scan location, c_t, to minimise it.

We use an estimator for the EIG similar to that of Harvey et al. (2019). It requires two components: **(I)** A neural network trained to classify v given a series of scans at locations $c_1, \ldots, c_t$. This outputs a classification distribution which we denote $g(v|f_{c_1,\ldots,c_t}(\mathbf{x}))$, where $f_{c_1,\ldots,c_t}$ is a function mapping from an image to the values of the pixels observed by scans at $c_1, \ldots, c_t$. We use this classification distribution as an approximation of the posterior over v, whose entropy we attempt to minimise by performing BOED. **(II)** A method for sampling image completions conditioned on some observed

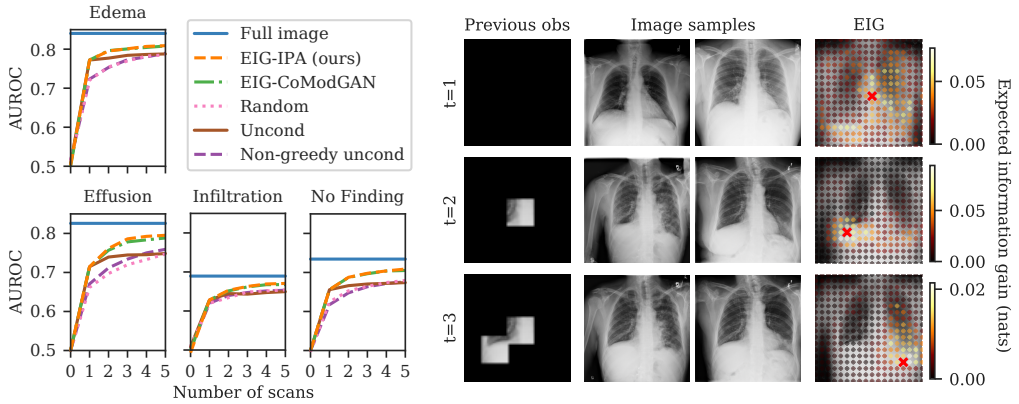


Figure 5: **Left:** Classification AUROC scores after $1, \ldots, 5$ scans chosen with each method. Scores for the "EIG-" methods more quickly approach the upper bound achieved by processing the full image. **Right:** Visualisation of BOED used to select three scan locations for diagnosing 'Effusion'. The left column shows the observations made prior to each time step. We then show samples from IPA (or the dataset when $t = 1$). The rightmost column shows the EIG overlaid on the pixel-space average of sampled images, with the optimal c_t marked by a red cross.

pixel values $f_{c_1,\ldots,c_{t-1}}(\mathbf{x})$. Harvey et al. (2019) used a "stochastic image completion" module which contributed significant complexity to their method. We entirely replace this with IPA.

Let the pixel values observed so far be $\mathbf{y}_{c_1,\ldots,c_{t-1}} = f_{c_1,\ldots,c_{t-1}}(\mathbf{x})$ for a latent image $\mathbf{x}$. Given these, we estimate the EIG of location c_t as

$$\mathrm{EIG}(c_t;\mathbf{y}_{c_1,\ldots,c_{t-1}}) \approx \overbrace{\mathcal{H}\left[\frac{1}{N}\sum_{n=1}^{N} g(\cdot|f_{c_1,\ldots,c_t}(\mathbf{x}^{(n)}))\right]}^{\text{entropy after } t-1 \text{ scans}} - \overbrace{\frac{1}{N}\sum_{n=1}^{N}\mathcal{H}\left[g(\cdot|f_{c_1,\ldots,c_t}(\mathbf{x}^{(n)}))\right]}^{\text{expected entropy after } t \text{ scans}}, \tag{9}$$

where $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$ are sampled image completions from IPA given $\mathbf{y}_{c_1,\ldots,c_{t-1}}$. In Appendix E we report hyperparameters, provide further details of our EIG estimator, and compare it to the estimators used in related work. To select c_t, we simply estimate $\mathrm{EIG}(c_t;\mathbf{y}_{c_1,\ldots,c_{t-1}})$ for many different values of c_t and select the value which maximises it. This process of selecting c_t and then taking a scan is repeated for each $t = 1, \ldots, T$.

We experiment on the NIH Chest X-ray 14 dataset (Wang et al., 2017) at 256×256 resolution. We simulate a scanner which returns a 64×64 pixel patch from this image, and the task is to diagnose the binary presence or absence of an illness. We run separate experiments diagnosing each of edema, effusion, infiltration and "no finding" (an additional label meaning there are no diagnosed illnesses). With appropriate data, this framework could be extended to also infer the severity of a given illness. We envisage BOED being used to select scan locations for an x-ray without necessarily performing an automated diagnosis. However, to quantify the informativeness of the chosen locations, Fig. 5 shows the results of using g to perform a diagnosis, or classification, based on the chosen scan locations. Since the conditional distribution g (used to estimate the EIG) depends on which illness we are classifying, the choice of scan locations is different in each case. We compare against a baseline where the image completion is performed by CoModGAN (our best-performing image completion baseline) rather than IPA, as well as numerous baselines which choose scan locations without image completion; see Appendix E for details.

Our method (denoted EIG-IPA) narrowly but consistently outperforms EIG-CoModGAN. We hypothesise that this is due to the aforementioned tendency of CoModGAN to sometimes collapse to a single mode of the posterior, and exhibit an example of this behaviour on the x-ray dataset in Appendix H. In the BOED context, such "overconfident" image completion could lead to salient scan locations being ignored. Nonetheless, both EIG-IPA and EIG-CoModGAN significantly outperform the other baselines, giving performance much closer to the upper bound of a CNN with access to the entire image. Another benefit of the "EIG-" approaches is that the choice of scan locations is highly interpretable; we can see why a particular location was chosen with visualisations similar to the right

of Fig. 5. This shows the sampled images $\mathbf{x}^{(n)}$ and the estimated EIG for each c_t. In Appendix E, we show that we can further quantify the contribution of each $\mathbf{x}^{(n)}$ to the estimated EIG for each c_t.

6 Related Work

Inference in pretrained VAEs Several prior studies perform conditional generation using a previously trained unconditional VAE. Like us, Rezende et al. (2014); Nguyen et al. (2016); Wu et al. (2018) do so through inference in the VAE's latent space. However, they use non-amortized inference (Gibbs sampling, variational inference, and MCMC respectively), leading to slow sampling times for any new $\mathbf{y}$. Duan et al. (2019) learn variational distributions over z for every possible value of $\mathbf{y}$, but this is not possible when $\mathbf{y}$ is high-dimensional or continuous-valued. Yeh et al. (2017) fit the latent variables of a GAN given observations, but this is neither amortized nor probabilistic.

Conditional VAEs Past research on conditional VAEs (Sohn et al., 2015; Zheng et al., 2019; Ivanov et al., 2018; Wan et al., 2021) has generally been unable to take advantage of pretrained weights as we have due to a difference in architectures: unlike almost all prior work, the IPA decoder does not receive $\mathbf{y}$ as input. The dependence between $\mathbf{y}$ and the decoder's output must therefore be expressed solely through the conditional distribution over the latent variables, $\hat{q}(z|\mathbf{y})$. This is a crucial difference because it means that the decoder can have exactly the same architecture as that of an unconditional VAE. This is key to letting us copy the pretrained weights of an unconditional VAE to speed up training. The exception to the above is Ma et al. (2018) who, like us, use a conditional VAE decoder with no dependence on $\mathbf{y}$. Their use case is very different, however, and they do not consider using pretrained models or use an architecture which can scale to photorealistic images. Leveraging unconditional VAEs lets us drastically reduce the computational budget required to train a conditional VAE. We believe that this paper is the first to demonstrate photorealistic image completion with conditional VAEs at resolutions as high as 256×256.

Image completion Early work on image completion, both before (Bertalmio et al., 2000; 2001; Ballester et al., 2001; Levin et al., 2003; Criminisi et al., 2003) and after (Köhler et al., 2014; Ren et al., 2015) deep learning became the dominant approach, aimed to deterministically fill in missing pixels in images. Even many methods incorporating generative adversarial networks (GANs), which were introduced by Goodfellow et al. (2014) as a tool to learn distributions, have been found to result in little or no diversity in the completions produced for a given input (Song et al., 2018; Yu et al., 2018; 2019; Pathak et al., 2016; Iizuka et al., 2017). However, some recent methods have managed to obtain diverse completions using the GAN framework (Zhao et al., 2020; 2021; Liu et al., 2021). Another approach is to generate a distribution over low-resolution images using VAEs (Zheng et al., 2019; Peng et al., 2021) or transformers (Zheng et al., 2021; Wan et al., 2021), and then use a GAN for upsampling. In contrast, we use a VAE to model image completions at the full resolution. As well as ensuring diverse coverage of the posterior, using such a likelihood-based model enables applications such as out-of-distribution detection for inputs $\mathbf{y}$, which we demonstrate in Appendix F. Another related approach is that of Song et al. (2020), who present a stochastic differential equation-based image model. This can be conditioned on subsets of image pixels to perform image completion, but sampling is slow.

7 Discussion and Conclusion

We have presented IPA, a method to adapt an unconditional VAE into a conditional model. Image completions generated with IPA are close to the state-of-the-art in terms of visual fidelity, and improve on all baselines in terms of their coverage of the posterior as measured by LPIPS-GT. This high-fidelity coverage of the posterior makes IPA ideal for use in Bayesian optimal experimental design, as demonstrated. In addition, IPA has all the benefits of a likelihood-based method, such as the potential to perform out-of-distribution detection. Future work may investigate further improving the image quality by, for example, using a partial encoder with more expressive distributions. Preliminary experiments revealed that normalizing flows could help the partial encoder better match the posterior, but with little impact on the resulting FID scores. Alternative directions include investigating different types of conditional generation, such as conditioning on previous frames to make a video model.

REFERENCES

Sanjeev Arora and Yi Zhang. Do gans actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*, 2017.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, pp. 224–232. PMLR, 2017.

Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8):1200–1211, 2001.

Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 417–424, 2000.

Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pp. I–I. IEEE, 2001.

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pp. 273–304, 1995.

Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

Antonio Criminisi, Patrick Perez, and Kentaro Toyama. Object removal by exemplar-based inpainting. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pp. II–II. IEEE, 2003.

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 conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Yu Duan, Canwen Xu, Jiaxin Pei, Jialong Han, and Chenliang Li. Pre-train and plug-in: Flexible conditional text generation with variational auto-encoders. *arXiv preprint arXiv:1911.03882*, 2019.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning*, pp. 1462–1471. PMLR, 2015.

William Harvey, Michael Teng, and Frank Wood. Near-optimal glimpse sequences for improved hard attention neural network training. *arXiv preprint arXiv:1906.05462*, 2019.

Jakob D Havtorn, Jes Frellsen, Søren Hauberg, and Lars Maaløe. Hierarchical vaes know what they don't know. *arXiv preprint arXiv:2102.08248*, 2021.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. arxiv (2016). *arXiv preprint arXiv:1611.07004*, 2016.

Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning. *arXiv preprint arXiv:1806.02382*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*, 2016.

Alexej Klushyn, Nutan Chen, Richard Kurle, Botond Cseke, and Patrick van der Smagt. Learning hierarchical priors in vaes. *arXiv preprint arXiv:1905.04982*, 2019.

Rolf Köhler, Christian Schuler, Bernhard Schölkopf, and Stefan Harmeling. Mask-specific inpainting with deep neural networks. In *German conference on pattern recognition*, pp. 523–534. Springer, 2014.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics. In *ICCV*, volume 1, pp. 305–312, 2003.

Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7760–7768, 2020.

Ji Lin, Richard Zhang, Frieder Ganz, Song Han, and Jun-Yan Zhu. Anycost gans for interactive image synthesis and editing. *arXiv preprint arXiv:2103.03243*, 2021.

Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting. *arXiv preprint arXiv:2105.02201*, 2021.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. *arXiv preprint arXiv:1809.11142*, 2018.

Tom Minka et al. Divergence measures and message passing. Technical report, Microsoft Research, 2005.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Anh Nguyen, Jason Yosinski, Yoshua Bengio, Alexey Dosovitskiy, and Jeff Clune. Plug & play generative networks: Conditional iterative generation of images in latent space.(2016). *arXiv preprint cs.CV/1612.00005*, 2016.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. *arXiv preprint arXiv:2103.10022*, 2021.

Samrudhdhi Bharatkumar Rangrej and James J. Clark. Achieving explainability in a visual hard attention model through content prediction, 2021. URL https://openreview.net/forum?id=pQq3oLH9UmL.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A DePristo, Joshua V Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *arXiv preprint arXiv:1906.02845*, 2019.

Jimmy SJ Ren, Li Xu, Qiong Yan, and Wenxiu Sun. Shepard convolutional neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pp. 901–909, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Mark Sabini and Gili Rusak. Painting outside the box: Image outpainting with gans. *arXiv preprint arXiv:1808.08483*, 2018.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29*, pp. 2234–2242. Curran Associates, Inc., 2016.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28: 3483–3491, 2015.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *arXiv preprint arXiv:1602.02282*, 2016.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yuhang Song, Chao Yang, Yeji Shen, Peng Wang, Qin Huang, and C-C Jay Kuo. Spg-net: Segmentation prediction and guidance network for image inpainting. *arXiv preprint arXiv:1805.03356*, 2018.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:2007.03898*, 2020.

Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. *arXiv preprint arXiv:2103.14031*, 2021.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Ga Wu, Justin Domke, and Scott Sanner. Conditional inference in pre-trained variational autoencoders via cross-coding. *arXiv preprint arXiv:1805.07785*, 2018.

Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *arXiv preprint arXiv:2003.02977*, 2020.

Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5485–5493, 2017.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5505–5514, 2018.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4471–4480, 2019.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5741–5750, 2020.

Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2021.

Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1438–1447, 2019.

Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Tfill: Image completion via a transformer-based architecture. *arXiv preprint arXiv:2104.00845*, 2021.