Improving Generative Modelling in VAEs using Multimodal Prior

Vinayak Abrol, *Member, IEEE,* Pulkit Sharma, and Arijit Patra

***Abstract*—In this paper we propose a conditional generative modelling (CGM) approach for unsupervised disentangled representation learning using variational autoencoder (VAE). CGM employs a multimodal/categorical conditional prior distribution in the latent space to learn global uncertainty in data by modelling the variations at local level. Thus, the proposed framework enforces the model to independently estimate the inherent patterns within each category, which improves the interpretability of the latent representations learned by the VAE model. The evidence lower bound objective for training the generative model is maximized using a mutual information criterion between the global latent categorical variable and the encoded inputs. Further, the approach has a built-in mechanism for bounding the information flow between the encoder and the decoder which addresses the problems of posterior collapse in conventional VAE models. Experiments on a variety of datasets demonstrate that our objective can learn disentangled representations and the proposed approach achieves competitive results on various task such as generative modelling, image classification and image denoising.**

***Index Terms*—Generative modelling, autoencoders, matching network, representation learning**

I. Introduction

GENERATIVE models have shown great promise for unsupervised learning via capturing rich distribution of complex data such as natural images, text and speech [1]. In particular, the aim is to extract semantically meaningful low-level (region-oriented) and high-level (object-oriented) hidden attributes of the given data. Recently, neural network based generative models have become successful frameworks for this class of problems e.g., image retrieval [2], super-resolution [3], image recognition [4], video captioning [5], video dialog [6] and music transcription [7]. Popular approaches include, variational autoencoders (VAEs) [8], [9], [10] and generative adversarial networks (GANs) [11]. VAEs are probabilistic graphical models for latent modelling with the ability to approximate (at-least theoretically) a desired distribution [8]. The data generated using VAEs is of good quality and VAEs naturally collapse most dimensions in the latent space, which result in interpretable latent space [12]. Thus autoencoders are more suitable for data compression and generating meaningful semantic features. In contrast, GANs are explicitly set up to optimize for generative tasks and are typically better deep generative models as compared to VAEs [11]. GANs use another network (so-called Discriminator) to compare generated and real data with the aim of achieving an equilibrium between Generator and Discriminator. However, this makes them difficult to work with as they require a lot of training data and tuning [13]. Some efforts have also been taken to learn more flexible generative models by combining adversarial training of GANs with VAEs [14], [15], [16].

In unsupervised representation learning, generative models encode the observed data such as images in an informative latent space. In particular, VAEs are designed to have an isotropic Gaussian prior latent distribution, and are trained by maximizing the likelihood of the observed data by marginalizing over the latent variables [9]. This is achieved via optimizing the evidence lower bound consisting of a data reconstruction error term, and a divergence term between latent inference and prior distribution. In practice, most VAEs are inadequate from the viewpoint of matching between the prior and true distribution, and suffers from their inability to learn latent features that are disentangled [12], [17]. The main reason for this is the original formulation of VAE, which is designed to learn a generative model (via an encoder-decoder pair), and not to produce informative latent features. This results in two major problems 1) uninformative latent representations and 2) variance over-estimation [12]. The former problem occurs when the decoder is too expressive and models the data distribution on its own without the use of latent representation, which drives the divergence term in evidence lower bound to zero [18], [19]. The latter problem occurs due to unbounded information flow from input to latent space [20]. These problems cause VAEs to overfit on the dataset thus driving the inference distribution away from prior distribution on the latent variable [12], [21]. Recently, many studies have proposed solutions for improving the representation learning capabilities of VAEs via a modified objective function [12], [15], [18], [22], [23].

In this work, we propose an alternative yet simple generative modelling framework with multimodal prior distribution targeted towards image applications such image generation, classificaton and denoising. In order to train such models, we define the evidence lower bound objective (ELBO) with categorical conditional prior. We call this conditional generative modelling (CGM) and show that it is capable of extracting interpretable hidden attributes of data. The disentanglement in latent representations is achieved by introduction of global

Manuscript received August 14, 2019; revised January 15, 2020 and March 16, 2020; accepted June 23, 2020. Date of publication 2020; date of current version 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Heng Tao Shen. (*Corresponding author: V. Abrol)*

V. Abrol is with the Mathematical Institute, University of Oxford, Oxford OX26GG, UK. (e-mail: abrol@maths.ox.ac.uk).

P. Sharma, and Arijit Patra are with Department of Engineering Science, University of Oxford, Oxford OX13PJ, U.K. (e-mail: pulkit.sharma@eng.ox.ac.uk,arijit.patra@exeter.ox.ac.uk).

Digital Object Identifier 10.1109/TMM.2020.

latent variables over various categories of available training images. In unsupervised settings, the CGM model induces clustering of images in latent space while supervised setting can be used if label information is available. In addition, a mutual information criterion is used to ensure that the latent variables provide useful information about the input image space, which is finally used by the decoder. Each global latent variable (similar to a clustering regime) represents a learned categorical context conditioned over that category. Such latent variables are aimed to capture the global/local uncertainty of the data arising due to inter/intra class variations. This ensures that the encoder match to at least one of the categorical global latent variable, and the information between latent and input space is maximized.

The variance over-estimation problem is addressed by maximizing the likelihood of the data with a bounded information rate between encoder and decoder. The proposed approach ensures the inference distribution of the encoder for a given input is a Gaussian distribution with a learned mean but fixed pre-determined standard deviation. Further, this work shows that contrary to more complex and flexible models, CGM models are expressive enough without the need of specifically fine-tuning neural network architectures.

The rest of the paper is organized as follows: Section II briefly introduces VAEs followed by the proposed CGM approach and the associated model training procedure. In Section III we discuss and provide a comparison with the relevant related works. Section IV presents experimental results and finally the paper is concluded in Section V.

II. Proposed Approach

We first provide a brief background on generative modelling using variational inference. Next, we describe the proposed CGM approach. The section discusses an approach to maximize mutual information between input and latent space in order to learn categorical prior distributions. Finally, we describe the overall loss function for training the proposed CGM model.

A. Background

We consider the problem of probabilistic generative modelling where the goal is to learn a probability distribution $p(\mathbf{x}|\theta)$ from data points $\mathbf{x}$ parameterized by θ [9]. The generation of $\mathbf{x}$ via 'generator' model is achieved with introduction of latent variables $\mathbf{z}$ such that $p(\mathbf{x}|\theta) = \int p(\mathbf{z}|\theta)p(\mathbf{x}|\mathbf{z},\theta)d\mathbf{z}$ [24]. Here, $p(\mathbf{z}|\theta)$ is the prior distribution commonly restricted to tractable standard Gaussian distribution $\mathcal{N}(0, I)$ [12], [25]. Since, marginal distribution $p(\mathbf{x}|\theta)$ can not be computed analytically, typically variational inference using an 'encoder' model parameterized by ϕ with amortized inference distribution $q(\mathbf{z}|\mathbf{x},\phi)$, is employed to approximate it with a lower bound

$$p(\mathbf{x}|\theta) \geq \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\phi)}[p(\mathbf{x}|\mathbf{z},\theta)] - \lambda D[q(\mathbf{z}|\mathbf{x},\phi)\|p(\mathbf{z}|\theta)]. \quad (1)$$

The first term on right hand side of eq(1) measures how accurate is the generator, the second term denotes the divergence loss (KL divergence being a popular choice) that measures how closely the encoded latent variables match a unit Gaussian and λ is the scaling parameter [8]. In GANs, this divergence term is replaced by an adversarial network which tries to discriminate between original and generated samples [13], [26], [27], [28].

B. Conditional Generative Modelling (CGM)

The proposed CGM aims to model conditional generative distributions of the form

$$p(\mathbf{x}|\mathbf{C},\theta) = \int p(\mathbf{z}|\mathbf{C},\theta)p(\mathbf{x}|\mathbf{z},\mathbf{C},\theta)d\mathbf{z}, \quad (2)$$

where $\mathbf{C}$ is the data-dependent conditional variable(s) used to estimate the conditional prior $p(\mathbf{z}|\mathbf{C},\theta)$ and likelihood $p(\mathbf{x}|\mathbf{z},\mathbf{C},\theta)$. The evidence lower bound in this case will be

$$\mathbb{E}_{q(\mathbf{z}|\mathbf{x},\phi)}[p(\mathbf{x}|\mathbf{z},\theta)] - \lambda D[q(\mathbf{z}|\mathbf{x},\phi)\|p(\mathbf{z}|\mathbf{C},\theta)], \quad (3)$$

where the conditional prior $p(\mathbf{z}|\mathbf{C},\theta)$ being intractable is approximated via the variational posterior $q(\mathbf{z}|\mathbf{C},\phi)$. The CGM model is designed to inherently solve the latent space exploding and variance over-estimation problem of conventional variational models. The first problem is addressed by introduction of global latent variables and maximizing a mutual information criteria to ensure that the latent variables provide useful information about the input to decoder (Section II-C).

The variance over-estimation problem is solved by bounding the information flow from the input to latent space. A given input is first encoded to approximate the mean μ of inference distribution $q(\mathbf{z}|\mathbf{x},\phi)$, followed by addition of noise $z = \mu + \epsilon$, where ϵ is i.i.d Gaussian with variance less than one. This is in contrast to conventional variational models where variance is also estimated. An alternative approach towards bounding the information rate is by using a variational infomax bound [29]. But this requires training a separate auxiliary network and thus, there are no guarantees on tightness of the mutual information bound. Moreover, it increases the model complexity and training time. On similar lines, we came across the work in [22], which also uses a variational inference to construct a lower bound on the information bottleneck objective. Recently, work in [30], [31] showed that bounding the variance prevents posterior collapse or equivalently prevents the divergence term from vanishing.

C. Moment Matching in Local and Global Latent Variables

In previous section, for simplicity the conditional variable is denoted by $\mathbf{C}$ instead of individual categorical variables i.e., $[\mathbf{g}^1 \dots \mathbf{g}^K]$ (see Fig. 1). This results in a K-categorical N-dimensional latent embedding space $\in \mathbb{R}^{K\times N}$ which captures the global uncertainty of the data, while allowing us to still sample the distribution at the local level. As the model trains on more observations the latent distribution encoded by model amounts to the posterior $p(\mathbf{z}|\mathbf{C},\theta)$ instead of the conventional zero-information prior $p(\mathbf{z}|\theta)$. Since, the latent space is discrete due to K-categories, the actual conditional prior distribution $q(\mathbf{z}|\mathbf{x},\mathbf{g}^K,\phi)$ for a category is continuous and deterministic. Further, defining a simple uniform prior over all categories we obtain a constant KL divergence term over the prior $p(\mathbf{g}^K|\theta)$ which can be ignored while training (see Proposition 1).

In the CGM, an example is first locally encoded and then is used to update network parameters such that it

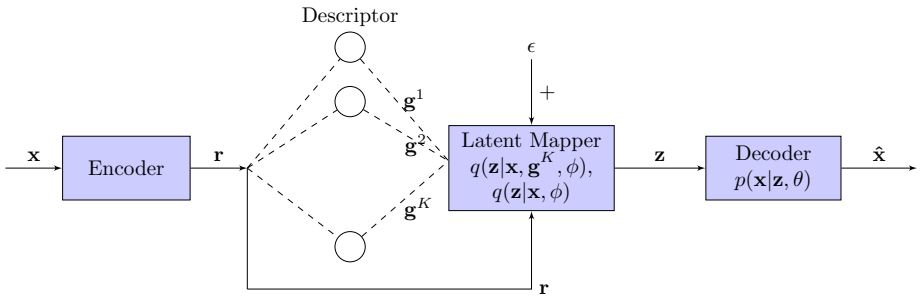


Fig. 1. Block diagram of the proposed CGM approach. Input vector $\mathbf{x}$ is encoded to a local representation vector $\mathbf{r}$ via an Encoder. A Descriptor summarizes representations of similar inputs as one of the global categorical prototypes $\mathbf{g}^K$. With all classes well-separated $\mathbf{g}^K$ corresponds to the means of encoded features in each class. Both $\mathbf{r}$ and $\mathbf{g}^K$ are used by the Latent Mapper to generate samples (with a fixed variance) from the posterior distribution and underlying latent prior distribution. Finally, a Decoder using the sampled latent variable attempts to reconstruct the input. The overall training of CGM model involve a joint loss comprising of individual losses corresponding to Encoder/Decode, Descriptor and Latent Mapper.

match with global conditional encoding of a category. In unsupervised setting example assignment/relevance can be achieved by nearest neighbour search or a similarity metric, while in supervised setting it is done using class labels. The matching between local and global encoding is calculated using Maximum-Mean Discrepancy (MMD) which quantifies the match between examples from two distributions by comparing their moments [32], [33]. The basic idea is that the distances between distributions can be represented as distances between mean embedding, and in its general form MMD is defined as

$$M(p\|q) = \sup_{f \in F} (\mathbb{E}_p[f(a)] - \mathbb{E}_q[f(b)]), \tag{4}$$

where divergence $M(p\|q)$=0 if p=q, only when F=$\{f, \|f\|_{\mathcal{H}} \leq 1\}$ is a unit ball in a Reproducing Kernel Hilbert Space $\mathcal{H}$ [34], [35]. Hence, given samples $a_1 \dots a_S \sim p$ and $b_1 \dots b_T \sim q$, and with a positive semi-definite kernel $\mathcal{K}(\cdot, \cdot)$, under the unit ball assumptions on the evaluation function, we have

$$\begin{aligned} M(p\|q) =& \|\mu_p - \mu_q\|^2_{\mathcal{H}} \\ =& \mathbb{E}_{p(a),p(b)}[\mathcal{K}(a,b)] - 2\,\mathbb{E}_{p(a),q(b)}[\mathcal{K}(a,b)] \\ &+ \mathbb{E}_{q(a),q(b)}[\mathcal{K}(a,b)] \\ =& \frac{1}{S^2}\sum_{i}^{S}\sum_{i'}^{S}\mathcal{K}(a_i, a_{i'}) - \frac{2}{ST}\sum_{i}^{S}\sum_{j}^{T}\mathcal{K}(a_i, b_j) \\ &+ \frac{1}{T^2}\sum_{j}^{T}\sum_{j'}^{T}\mathcal{K}(b_j, b_{j'}). \end{aligned} \tag{5}$$

Its implementation requires the conventional kernel trick where we compute the distance between the means μ_p, μ_q of the two distributions mapped into a reproducing kernel Hilbert space [36]. Note that computing MMD is equivalent to computing an energy distance with respect to some negative-type semimetric [37]. Hence, for a given domain-specific notion of distance (negative type semimetric) we can define an equivalent similarity (kernel) which results in a computationally attractive criterion for training the models. This approach was recently studied in [38] for MMD based GANs.

In CGM, we compute the empirical estimates of MMD metric for samples from the posterior distribution $q(\mathbf{z}_i|\mathbf{x}_i, \phi)$ using encoded inputs depending on sample assignment, and corresponding global categorical distribution $q(\mathbf{z}_i|\mathbf{x}_i, \mathbf{g}^K, \phi)$ using the reparametrization trick. Considering eq(5), μ_p is the sample mean of kernel transformed latent representations of inputs and similarly μ_q is the sample mean corresponding to global prototypes. This is to ensures that the information between latent and input space is maximized, and to encourage disentanglement improving discriminative power of latent features. The MMD metric is also a pseudo measure which maximizes the mutual information between input and latent space. The use of mutual information based objectives is popular in various works such as information maximizing GANs or InfoGANs [13], variational lossy autoencoders [12], [29] and representation learning [1]. The proposed CGM model is not a standard VAE as the conditional prior is categorical i.e., a mixture of Gaussians with hard assignment[1]. However, it can be shown that it still maximizes a variational lower bound of the marginal log-likelihood i.e.,

Proposition 1. $\mathbb{E}_{\tilde{p}(\mathbf{x})}[\log p(\mathbf{x}|\theta)] \geq \mathcal{L}_E + \mathcal{L}_M$.

where $\tilde{p}(\mathbf{x})$ is the empirical data distribution, $\mathcal{L}_E$ is the data error loss and $\mathcal{L}_M$ is the MMD loss.

D. The CGM Model

In our implementation the CGM model has three core components as shown in Fig. 1:

1) An encoder $E(\cdot)$ parameterised as a neural network, that encodes the examples $\mathbf{x}_i$ from input space to a representation $\mathbf{r}_i = E(\mathbf{x}_i)$.
2) A descriptor $A(\cdot)$ that summarizes the encoded inputs to learn K global latent prototype representations. To achieve order invariance in global representation we employ the mean operation $\mathbf{g}^K = \frac{1}{R}\sum_i^R \mathbf{r}_i^K$ which worked well in our initial experiments. We observed marginal performance improvement with a weighted average using attention mechanism, however, further studies are required to make any claims.

[1]During submission process we came across a related article [39] which uses a GMM prior in latent space. The regularization with GMM prior rely upon ad-hoc parameters, namely eta (in NLL) and alpha (in regularizer), to which it is sensitive and are difficult to tune. Full ELBO objective is based on KL divergence which is assumed to ensure an uniform approximate posterior; this is only true for small data, and in practice the NLL term in the ELBO will overpower the regularizer. Also it is unclear if the mean-field approximation actually constraints the Monte Carlo sampling.

3) A Latent mapper $L(\cdot)$ that generate samples from the posterior distribution $q(\mathbf{z}_i|\mathbf{x}_i, \phi)$ and $q(\mathbf{z}_i|\mathbf{x}_i, \mathbf{g}^K, \phi)$ using the reparametrization trick.
4) A decoder $D(\cdot)$ parameterised as a neural network, that takes as input the sampled local latent variable $\mathbf{z}_i$ and estimates the distribution $p(\mathbf{x}_i|\mathbf{z}_i, \theta)$.

E. Model Training

The training of a CGM model is achieved by jointly optimizing the overall loss function:

$$\mathcal{L}_{CGM} = \mathcal{L}_E + \lambda\mathcal{L}_M + \beta\mathcal{L}_D, \tag{6}$$

where $\mathcal{L}_E$ is the data error loss for optimizing the encoder and decoder networks, $\mathcal{L}_M$ is the MMD loss, and $\mathcal{L}_D$ is the descriptor loss for updating global prototypes. If output distribution is chosen to be Gaussian, $\mathcal{L}_E$ corresponds to mean squared error loss, while for the case of Bernoulli (e.g., for binary MNIST images) it corresponds to binary cross entropy loss. λ, β are constants, and in our experiments, we found the model training to be robust and stable w.r.t. these constants. For MMD loss we considered the Gaussian kernel

$$\mathcal{K}(a, b) = \exp^{\frac{-\|a-b\|^2}{2N}}, \tag{7}$$

although other handcrafted kernels can also be explored as long as they are positive semi-definite so as to enable a projection into a reproducing kernel Hilbert space. In order to update the K global prototypes of the descriptor we considered the following loss after computing the example assignments [40]:

$$\mathcal{L}_D = \begin{cases} \|\mathbf{r}_i - \mathbf{g}^K\|_2^2 & \text{if } i \in K \\ \max(0, m - \|\mathbf{r}_i - \mathbf{g}^K\|_2^2) & \text{otherwise.} \end{cases} \tag{8}$$

The network encodes the examples $\mathbf{x}_i$ from input space to a representation $\mathbf{r}_i = E(\mathbf{x}_i)$, which are matched with a global prototype. Ideally, prototypes $\mathbf{g}^K$ should be updated (using back-propagation) along with latent features, on the entire training set. To implement this efficiently, updates are performed on mini-batches. If all classes are well-separated with the margin m, then $\mathbf{g}^K$ will approximately correspond to the means of features in each class [41]. Thus, the loss in eq(8) ensures that each encoded example match to at least one of the global categorical representation, and this encourages disentanglement in latent features.

III. Related Works

In this section we compare the proposed CGM approach with the existing closely related generative modelling and representation learning approaches.

A. Information Maximizing VAE (InfoVAE)

The closest work related to the proposed CGM approach is of the InfoVAE proposed in [12]. The family of InfoVAE models are also based on using a mutual information maximization term with the standard VAE loss term

$$\mathcal{L}_{IVAE} = \mathcal{L}_{VAE} - \gamma\mathcal{I}_{q(\mathbf{x},\mathbf{z},\phi)}(\mathbf{x}, \mathbf{z}), \tag{9}$$

where for $\gamma < 1$, $\mathcal{L}_{IVAE} = \mathcal{L}_E + \mathcal{L}_M$. Hence, InfoVAE is a special case of CGM model with $\beta = 0$ i.e., without considering global latent variables. In other words, for $\beta < 1$, the learned latent space in CGM tends to behave like InfoVAE. However, it is not possible to train an InfoVAE model having $\gamma = 1$ without bounding the mutual information between input and latent space. In other words, the mutual information can be maximized to infinity by making $q(\mathbf{z}|\mathbf{x}, \phi)$ a deterministic mapping with zero variance[2]. Although this problem was highlighted in [12], there was no explicit solution presented to restrict such modelling behaviour. While CGM also maps the latent representations in a deterministic way to one of the K global distributions, each input is encoded with a fixed variance determined by the bound on the information flow designed into the model. Further, the CGM model uses conditional prior instead of standard Gaussian prior used in InfoVAE.

B. Conditional VAE (CVAE)

In CVAE, conditioning on the context is done by adding the dependence both in the inference distribution $q(\mathbf{z}|\mathbf{x}, \mathbf{c}, \phi)$ and the decoder $p(\mathbf{x}|\mathbf{z}, \mathbf{c}, \theta)$, so they can be considered as deterministic functions of the context [42]. Here, context variable $\mathbf{c}$ is usually the class label. This helps in controlling the data generation process as the decoder can generate examples from a specific class. The usual way of implementing this is to concatenate the sampled latent variable $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ with a specified label and pass through the decoder. This again can lead to the problem of uninformative latent representations where the decoder is powerful enough to just focus on the label rather than the latent representation [43]. This issue is more prominent when the latent dimension is small. CGM model eliminates this problem by introduction of global latent variables (in contrast to local latent variables for computing divergence) such that we generate new examples by sampling one of the learned K-conditional prior distributions. Further, the encoder of CGM model doesn't use the class label. As described in Section II-C the model can also be trained in unsupervised settings where no label information is available, and the K-categorical modelling is performed not in the label space but rather in the prototype space which itself is defined by the model and learned during the training.

C. Meta Learning

Meta learning aims at inducing knowledge (e.g., a set of rules) such that an existing learning method can be evolved to perform on different learning problems [44]. As an example, in metric space meta learning such an idea of 'learning to learn' can be attempted by efficient optimization on small subsets of multiple tasks in a metric space and that metric space can be used for evaluating the extent or quality of the learning process. Models that include a conditioning class-variable can be used for classification as well, instead of usual generative tasks. For instance, for few-shot classification problem we usually

[2]The publicly available implementation by the authors doesn't uses the conventional reparametrization trick to sample the latent variable. In other words the posterior $q(\mathbf{z}|\mathbf{x}, \phi)$ is just a point mass. This is not the case with CGM where reparametrization is used as in case of VAEs.

employ some distance metric to compare target examples to the available prototype observations, with such a metric computed between feature representations obtained through a transformation of the image space into a lower dimensional space. Such a behaviour is inherent in CGM due to the matching between local and global prototype latent variables. A similar effort in this direction is of matching networks [33], [45] which at an abstract level can be interpreted as a CGM model where the descriptor is another network whose parameters are optimized using MMD after fixing a baseline VAE [46].

IV. Experimental Results

In this section we present various experimental studies to show the effectiveness of the proposed CGM approach, and compare it with existing state-of-the-art approaches. The performance is evaluated for various tasks namely density estimation, image generation, classification and image completion using standard datasets.

TABLE I
LOG LIKELIHOOD ESTIMATES FOR DIFFERENT MODELS ON THE MNIST DATASET, AND % CLASSIFICATION ERROR ON THE TEST SET OF MNIST DATASET. * AND ** SCORES ARE FOR THE CASE OF DYNAMICALLY AND STATISTICALLY BINARIZED IMAGES, RESPECTIVELY.

Model	NLL*	NLL**	Error
VLAE (PixelCNN) [18]	78.5	79.0	1.55
InfoVAE (DCGAN) [12]	80.7	81.2	1.62
MAE (PixelCNN) [20]	77.9	78.4	1.51
CGM1 (DCGAN)	78.2	78.8	1.49
CGM2 (DCGAN)	77.6	77.5	1.35

A. Training Setup

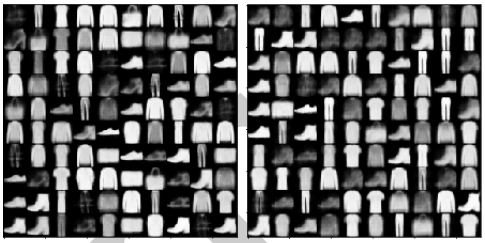
All of the networks for this work were trained using the PyTorch toolkit, where weights were initialized using the Xavier initialization scheme and biases were initialized to zero [47]. The Adam optimizer was used with an initial learning rate of 10^{-4}. All the networks were trained for 200 epochs. A batch size of 64, 64, 100, and 256 was used for MNIST [48], fashion MNIST [49], CIFAR-10 [50], and ImageNet [51] experiments, respectively. Parameter m is set to 1.

B. Behaviour On MNIST

In this section, the performance of the proposed CGM approach is evaluated on the MNIST, and fashion-MNIST datasets. The scaling constants are set as $\lambda = 1000$, $\beta = 1$, and ϵ is sampled with variance 0.02. For a fair comparison, we use the DCGAN architecture[3] for training the model and the choice of λ is kept similar to InfoVAE. We compared two variations of our CGM model, 1) CGM1 trained in unsupervised setting where example relevance is computed using nearest neighbour search; 2) CGM2 trained in supervised setting where example relevance is just the class label. In all experiments the training parameters are set empirically after experimentation.

[3]DCGAN refers to a fully convolutional neural network architecture replacing pooling operations with spatial downsampling convolutions and eliminating fully connected layers [52].

(a) (b)



(c) (d)

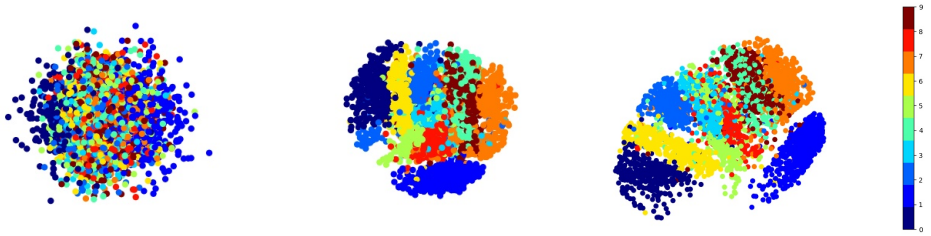
Fig. 2. Reconstructed and generated images for (a,b) MNIST and (c,d) Fashion-MNIST datasets using CGM2.

1) Density estimation: In this experiment we evaluated CGM on density estimation problem with a fixed information rate of 15 bits per image (BPI). The latent dimension N is set to 10. The performance of the proposed approach is compared with existing best performing approaches namely InfoVAE, Variational lossy autoencoder (VLAE) [18], and mutual posterior-divergence autoencoder (MAE) [20]. Both VLAE and MAE employ popular PixelCNN architecture for model training [53]. The results are reported in Table I in terms of marginal negative log-likelihood (NLL) in *nats* evaluated via importance sampling. It can be observed that the proposed approach achieves comparable and/or better performance to existing approaches.

2) Image generation: Next, in Fig. 2 we demonstrate image reconstruction and generation behavior of the proposed CGM approach on MNIST and fashion-MNIST datasets. The latent dimension N is set to 10, with previously described setting of all other parameters. It can be observed that the reconstructed images are of high quality, while the generated images are slightly blurry but sharp enough.

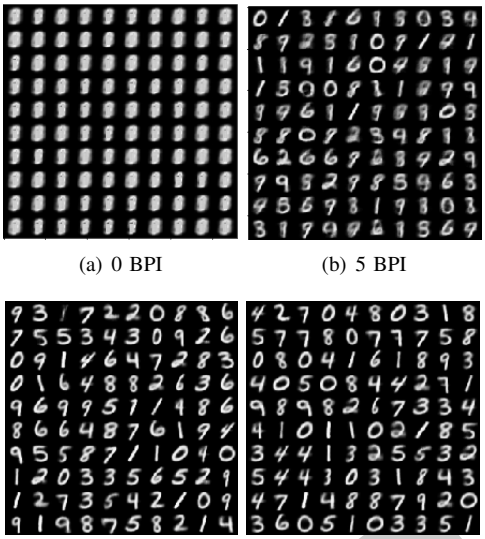
To investigate further, Fig. 3 further plots the latent variables for $N = 2$ on the test set of MNIST. As a comparison we have also plotted the same for InfoVAE and CVAE. In case of CVAE, the latent variables are highly overlapping, while for InfoVAE and CGM one can observe good disentanglement among latent features of different classes. This is because, in CVAE we model $p(\mathbf{z}|\mathbf{c})$, which is inferred variationally and hence the posterior $q(\mathbf{z}|\mathbf{x}, \mathbf{c})$ for any class is roughly $\mathcal{N}(0, \mathbf{I})$, as evident from the plot. CGM inherits the benefits of both conditional modelling so as to be able to generate class specific examples, and informative latent modelling like InfoVAE. Note that the disentanglement is expected to improve as the latent dimension increases.

3) Classification: In order to quantify the clustering behaviour of latent space we calculated the classification error by computing class assignment $\mathcal{S}(q(\mathbf{y}|f(\mathbf{z})))$, where $f(\cdot)$ is



(a) CVAE (b) InfoVAE (c) CGM1

Fig. 3. Visualization of 2D latent representations on MNIST test set.



(a) 0 BPI (b) 5 BPI

(c) 10 BPI (d) 15 BPI

Fig. 4. Reconstructed images for MNIST dataset at variable information-rate using CGM2.

a network that maps the latent variable to the logits and $\mathcal{S}(\cdot)$ is the softmax function. These results for the case of dynamically binarized images are also reported in Table I. Clearly, CGM achieves the best performance in both supervised and unsupervised setting due to its ability to learn disentangled representations.

4) Effect of bounding the information rate: As mentioned earlier the idea of introducing a mutual information constraint between the input and its latent variable is not new and many existing works employ this in some form to overcome the shortcomings of regular VAEs. However, mutual information itself is not a quantity that is easy to comprehend and specify. In practice, mutual information is always data dependent and it is difficult to access how much mutual information is enough to achieve a desired modelling behavior. Hence, for representation learning the usual way is to qualitatively inspect the model training, which makes existing methods less practical. This is not the case with the CGM approach, as it uses a fixed variance while encoding which determines the information rate, and for a fixed rate I it is given as $\sigma^2 = 1/(4^{\frac{I}{N}})$. As shown in Fig. 4, with increase in the information rate from 0 to 15 BPI, the quality of both the generated and reconstructed images increases. In comparison, approaches such as InfoVAE rely on manual inspection of vanishing variance. However, we observed that for a given information rate both CGM and InfoVAE achieve similar MSE, except that CGM results in better disentangled latent representations.

C. Behaviour on ImageNet and CIFAR-10

This section investigates CGM1 for ImageNet [51] and CGM2 for CIFAR-10, two popular benchmark datasets of natural images. For ImageNet, we used the residual-CNN [54] architecture of comparable size to the one in [53], with latent dimension $N = 600$, number of categories $K = 100$, and information rate of 200 BPI. The scaling constants are set as $\lambda = 1200$, $\beta = 1$, and variable ϵ is sampled with variance 0.02. For CIFAR-10, we used the DCGAN architecture with a latent dimension of $N = 300$ and an information rate of 120 BPI. The scaling constants are set as $\lambda = 1100$, $\beta = 1$, and ϵ is sampled with variance 0.02. In all experiments, the training hyperparameters are obtained empirically.

Table II reports the performance of the proposed approach on the density estimation problem. As a comparison we have considered both likelihood-based auto-regressive generative models (e.g. Pixel-RNN) and variationally trained latent models (e.g. δ-VAE). It can be observed that CGM achieves comparable performance to state-of-the-art approaches on challenging datasets with simpler network architecture. We expected to improve upon this by using a more powerful VAE architecture. Our initial experiments with complex architectures in particular, with PixelSNAIL decoder resulted in significant gains. Further, the generated images shown in Fig. 5 on the ImageNet dataset using CGM2 (PixelSNAIL) demonstrate the effective modelling capability of the proposed framework.

D. Image Completion

In this experiment, we evaluate the performance of the proposed CGM2 approach for image completion task. We carried out the experiments using MNIST [48] and CIFAR-10 [50], both containing 10 classes with approximately 60,000 images in the dataset. For each dataset we tested reconstruction abilities of a trained denoising CGM2 network on input images corrupted by additive Gaussian noise. For both datasets, we used the DCGAN architecture, number of categories $K = 10$, and the noise variance was set to 0.5. For MNIST, the latent

TABLE II
LOG LIKELIHOOD ESTIMATES (BITS/DIM) FOR DIFFERENT MODELS ON CIFAR-10 (32×32 IMAGES) AND IMAGENET (64×64 IMAGES).

Model	CIFAR-10 (Test)	ImageNet (Valid)
Pixel-RNN [53]	3.06	3.63
Gated Pixel-RNN [55]	3.02	3.57
PixelSNAIL [56]	2.85	3.56
δ-VAE (PixelSNAIL) [30]	2.83	3.58
MAE (PixelCNN) [20]	2.95	78.4
CGM1 (ResNet)	-	3.60
CGM2 (DCGAN)	3.07	-
CGM1 (PixelSNAIL)	-	3.57
CGM2 (PixelSNAIL)	2.81	-

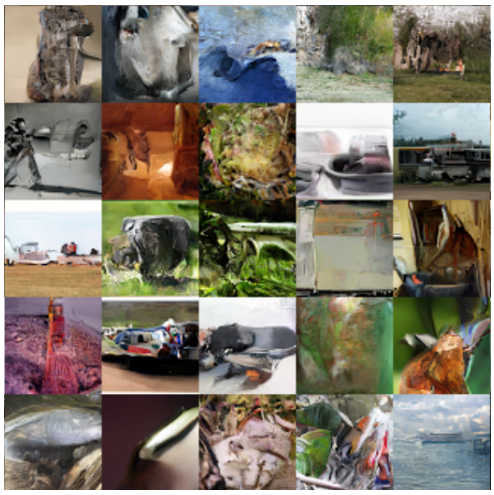
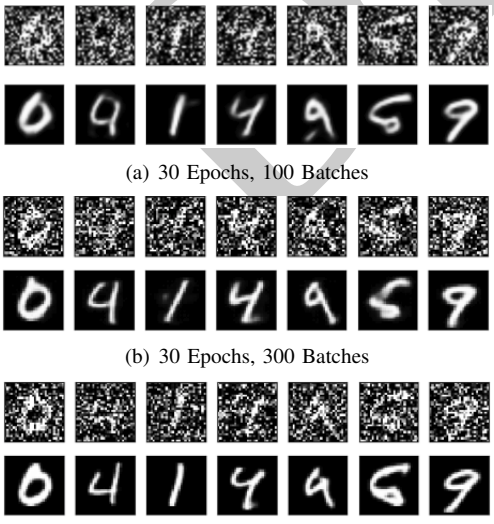


Fig. 5. Generated images for ImageNet dataset using CGM1.

dimension $N = 200$ and the scaling constants are set as $\lambda = 1000$, $\beta = 1$, and ϵ is sampled with variance 0.02. For CIFAR-10, the latent dimension $N = 300$ and the scaling constants are set as $\lambda = 1100$, $\beta = 1$, and ϵ is sampled with variance 0.02.



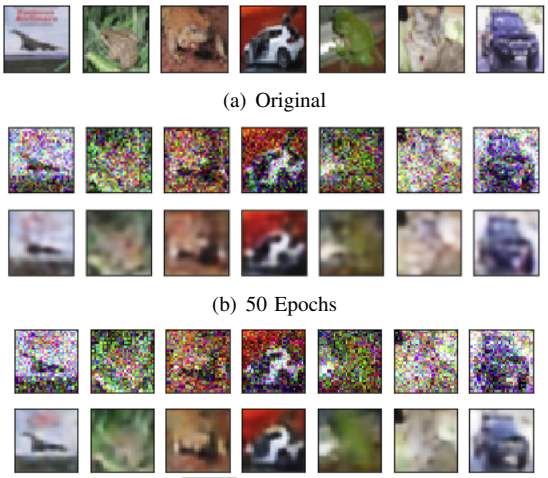
(a) 30 Epochs, 100 Batches

(b) 30 Epochs, 300 Batches

(c) 70 Epochs

Fig. 6. Noisy and Denoised images for MNIST dataset using CGM2.

As shown in Fig. 6 the model learns to make good predictions



(a) Original

(b) 50 Epochs

(c) 100 Epochs

Fig. 7. Noisy and Denoised images for CIFAR-10 dataset using CGM2.

of the underlying images on MNIST dataset. Also, for a given epoch as the number of categorical examples increases, the better the estimate of categorical conditional distribution become and hence, the predictions look more similar to the underlying ground truth.

Similarly, Fig. 7 shows the results on some images from CIFAR-10 dataset. It's worth mentioning that the goal of this experiment is not to surpass a state-of-the-art image recovery approach but just to show the capability of the proposed generative modelling approach. The recovery performance is expected to improve, especially in terms of edge details by using state-of-the-art CNN architectures designed for such tasks with additional image priors and regularization.

E. Discussion

Uninformative latent variable due to behaviour of KL divergence in ELBO and variance overestimation problems are well known, where existing works solves the former by carefully choosing parameters (e.g., β-VAE), label conditioning (e.g, CVAE) or redefining ELBO objective using MMD based ELBO. However, this still requires manually checking the variance, a theoretical proof/justification of which is presented in InfoVAE-Proposition2. CGM improves upon these by mapping each input to a categorical prior and fixed variance to bound the information flow. The main contributions of the paper are multimodal prior and information bounding method which learns an informative latent space with benefits of MMD based ELBO objective e.g., Fig. 4 confirms the effect of information rate controlling behaviour in image generation. Our initial attempt in this work is to show an easy way of training a generative model which eliminated the existing problems without compromising much on overall performance. Again our preliminary results confirm that with better architectures significant performance gains can be achieved.

V. CONCLUSION AND FUTURE WORK

In this paper, we have introduced CGM, a categorically conditional generative modelling approach with fixed multimodal priors. We have demonstrated that this model can learn

disentangled representation of data. The overall generative framework is regularized by restricting the information rate of the encoder network, and by maximizing a mutual information criterion between the global latent categorical variable and the encoded inputs. This ensures that the latent conditional prior distribution doesn't have vanishing variances, and information between latent and input space is maximized efficiently. We evaluated our model on various tasks ranging from unsupervised clustering, classification to image completion using popular datasets and achieved competitive results compared to the current state-of-the-art. In future applications to sequence prediction or data (such as audio or video) that requires modelling of temporal dynamics using a conditional generative approach would be worth considering. Additionally, the proposed CGM method's applications towards categorically defined rehearsal and pseudo-rehearsal approaches for alleviating catastrophic forgetting may be attempted to enable continual learning.

APPENDIX A
PROOF OF PROPOSITION 1:

Assuming K categories with conditional variables $\mathbf{g}^K$, the marginal distribution can be defined in terms of summation over categorical distributions as:

$$\begin{aligned}
\log p(\mathbf{x}|\theta) &= \int \sum_{i}^{K} p(\mathbf{x},\mathbf{z},\mathbf{g}^k[i=k]\ |\theta) d\mathbf{z} \\
&= \log \mathbb{E}_{\mathbf{z},\mathbf{g}^k}\left[\frac{p(\mathbf{x}|\mathbf{z},\mathbf{g}^k,\theta)p(\mathbf{z}|\mathbf{g}^k,\theta)p(\mathbf{g}^k|\theta)}{q(\mathbf{z},\mathbf{g}^k|\mathbf{x})}\right] \\
&\geq \mathbb{E}_{\mathbf{z},\mathbf{g}^k}\left[\log \frac{p(\mathbf{x}|\mathbf{z},\mathbf{g}^k,\theta)p(\mathbf{z}|\mathbf{g}^k,\theta)p(\mathbf{g}^k|\theta)}{q(\mathbf{z},\mathbf{g}^k|\mathbf{x})}\right] \\
&\geq \mathbb{E}_{\mathbf{z},\mathbf{g}^k}\left[\log p(\mathbf{x}|\mathbf{z},\mathbf{g}^k,\theta) + \log \frac{p(\mathbf{z}|\mathbf{g}^k,\theta)}{q(\mathbf{z}|\mathbf{x},\mathbf{g}^k,\phi)} + \right. \\
&\left. \log \frac{p(\mathbf{g}^k|\theta)}{q(\mathbf{g}^k|\mathbf{x},\phi)}\right]
\end{aligned} \tag{10}$$

where the relation follows from Jensen's inequality. The third term is a constant since we have assumed a uniform prior $p(\mathbf{g}|\theta)$ over all categories. The second term is the KL divergence between inference distribution and conditional prior distribution. From eq(9), for $\gamma \leq 1$ and bounded information flow between latent and input space, KL divergence can be replaced by MMD divergence. Taking the expectation over data distribution $\tilde{p}(\mathbf{x})$ proves our result.

REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 8, pp. 1798–1828, August 2013.

[2] I. Gonzlez-Daz, C. E. Baz-Hormigos, and F. Daz-de-Mara, "A generative model for concurrent image retrieval and roi segmentation," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 169–183, January 2014.

[3] Y. Guo, Q. Chen, J. Chen, Q. Wu, Q. Shi, and M. Tan, "Auto-embedding generative adversarial networks for high resolution image synthesis," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2726–2737, November 2019.

[4] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2049–2058, November 2015.

[5] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, "From deterministic to generative: Multimodal stochastic RNNs for video captioning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 10, pp. 3047–3058, October 2019.

[6] W. Jin, Z. Zhao, M. Gu, J. Yu, J. Xiao, and Y. Zhuang, "Video dialog via multi-grained convolutional self-attention context networks," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2019, pp. 465 – 474.

[7] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 927–939, May 2016.

[8] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations (ICLR)*, April 2017, pp. 1–11.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, April 2014, pp. 1–9.

[10] R. Weng, J. Lu, Y. Tan, and J. Zhou, "Learning cascaded deep auto-encoder networks for face alignment," *IEEE Transactions on Multimedia*, vol. 18, no. 10, pp. 2066–2078, October 2016.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *International Conference on Advances in Neural Information Processing Systems (NIPS)*, December 2014, pp. 2672–2680.

[12] S. Zhao, J. Song, and S. Ermon, "Infovae: Information maximizing variational autoencoders," in *International conference of Association for the Advancement of Artificial Intelligence (AAAI)*, January 2019, pp. 5885–5892.

[13] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *International Conference on Advances in Neural Information Processing Systems (NIPS)*, December 2016, pp. 2172–2180.

[14] A. B. L. Larsen, S. K. Snderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International Conference on Machine Learning (ICML)*, June 2016, pp. 1558–1566.

[15] W. Xu, S. Keshmiri, and G. Wang, "Adversarially approximated autoencoder for image generation and manipulation," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2387–2396, September 2019.

[16] T. Miyato, S. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, August 2019.

[17] E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh, "Disentangling disentanglement in variational autoencoders," in *International Conference on Machine Learning (ICML)*, June 2019, pp. 4402–4412.

[18] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel, "Variational lossy autoencoder," in *International Conference on Learning Representations (ICLR)*, April 2017, pp. 1–14.

[19] A. H. Jha, S. Anand, M. Singh, and V. Veeravasarapu, "Disentangling factors of variation with cycle-consistent variational auto-encoders," in *European Conference on Computer Vision (ECCV)*, September 2018, pp. 829–845.

[20] X. Ma, C. Zhou, and E. Hovy, "MAE: mutual posterior-divergence regularization for variational autoencoders," in *International Conference on Learning Representations (ICLR)*, May 2019, pp. 1–12.

[21] Y. Xian, S. Sharma, B. Schiele, and Z. Akata, "f-VAEGAN-D2: A feature generating framework for any-shot learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 10 275–10 284.

[22] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *International Conference on Learning Representations (ICLR)*, April 2017, pp. 1–16.

[23] T. Lucas and J. Verbeek, "Auxiliary guided autoregressive variational autoencoders," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, September 2018, pp. 443–458.

[24] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International Conference on Machine Learning (ICML)*, June 2014, pp. 1278–1286.

[25] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems (NIPS)*, December 2015, pp. 3483–3491.
[26] H. Huang, R. He, Z. Sun, T. Tan *et al.*, "IntroVAE: Introspective variational autoencoders for photographic image synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, December 2018, pp. 52–63.
[27] P. Ge, C. Ren, D. Dai, J. Feng, and S. Yan, "Dual adversarial autoencoders for clustering," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–8, 2019.
[28] X. Guo, R. Nie, J. Cao, D. Zhou, L. Mei, and K. He, "FuseGAN: learning to fuse multi-focus image via conditional generative adversarial network," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 1982–1996, August 2019.
[29] M. Phuong, M. Welling, N. Kushman, R. Tomioka, and S. Nowozin, "The mutual autoencoder: Controlling information in latent code representations," 2018. [Online]. Available: https://openreview.net/forum?id=HkbmWqxCZ
[30] A. Razavi, A. van den Oord, B. Poole, and O. Vinyals, "Preventing posterior collapse with delta-VAEs," in *International Conference on Learning Representations (ICLR)*, May 2019, pp. 1–11.
[31] D. Braithwaite and W. B. Kleijn, "Bounded information rate variational autoencoders," in *Deep Learning day of ACM Conference on Knowledge Discovery And Data Mining (KDD)*, August 2018, pp. 1–10.
[32] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *International Conference on Advances in Neural Information Processing Systems (NIPS)*, December 2007, pp. 513–520.
[33] Y. Li, K. Swersky, and R. Zemel, "Generative moment matching networks," in *International Conference on Machine Learning (ICML)*, July 2015, pp. 1718–1727.
[34] R. Fortet and E. Mourier, "Convergence de la répartition empirique vers la répartition théorique," in *Annales scientifiques de l'École Normale Supérieure*, vol. 70, no. 3, 1953, pp. 267–285.
[35] S. Vishwanathan, N. N. Schraudolph, and A. J. Smola, "Step size adaptation in reproducing kernel hilbert space," *Journal of Machine Learning Research*, vol. 7, pp. 1107–1133, June 2006.
[36] J. W. Green, "Moore-closed spaces, completeness and centered bases," *General Topology and its Applications*, vol. 4, no. 4, pp. 297–313, May 1974.
[37] D. Sejdinovic, B. Sriperumbudur, A. Gretton, K. Fukumizu *et al.*, "Equivalence of distance-based and RKHS-based statistics in hypothesis testing," *The Annals of Statistics*, vol. 41, no. 5, pp. 2263–2291, October 2013.
[38] M. Bikowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *International Conference on Learning Representations (ICLR)*, May 2018, pp. 1–15.
[39] N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. H. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with gaussian mixture variational autoencoders," 2016.
[40] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface: L2 hypersphere embedding for face verification," in *ACM International Conference on Multimedia*, October 2017, pp. 1041–1049.
[41] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision (ECCV)*, October 2016, pp. 499–515.
[42] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *International Conference on Advances in Neural Information Processing Systems (NIPS)*, December 2015, pp. 3483–3491.
[43] J. Engel, M. Hoffman, and A. Roberts, "Latent constraints: Learning to generate conditionally from unconditional generative models," in *International Conference on Learning Representations (ICLR)*, April 2018, pp. 1–12.
[44] S. Ritter, J. X. Wang, Z. Kurth-Nelson, S. M. Jayakumar, C. Blundell, R. Pascanu, and M. Botvinick, "Been there, done that: Meta-learning with episodic recall," in *International Conference on Machine Learning (ICML)*, July 2018, pp. 4354–4363.
[45] S. Bartunov and D. P. Vetrov, "Fast adaptation in generative models with generative matching networks," in *International Conference on Learning Representations (ICLR) Workshop*, April 2017, pp. 1–10.
[46] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. Rezende, and S. M. A. Eslami, "Conditional neural processes," in *International Conference on Machine Learning (ICML)*, July 2018, pp. 1704–1713.
[47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, December 2019, pp. 8024–8035.
[48] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.
[49] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," 2017.
[50] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep. TR-2009, April 2009.
[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009, pp. 248–255.
[52] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *International Conference on Learning Representations (ICLR)*, May 2016, pp. 1–13.
[53] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *International Conference on International Conference on Machine Learning (ICML)*, June 2016, pp. 1747–1756.
[54] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision (ECCV)*, October 2016, pp. 630–645.
[55] A. Van Den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with pixelcnn decoders," in *International Conference on Advances in Neural Information Processing Systems (NIPS)*, December 2016, pp. 4797–4805.
[56] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, "PixelSNAIL: An improved autoregressive generative model," in *International Conference on Machine Learning (ICML)*, July 2018, pp. 864–872.