

Multimodal Adversarially Learned Inference with Factorized Discriminators

Wenxue Chen,¹ Jianke Zhu^{1,2*}

¹ Zhejiang University

² Alibaba-Zhejiang University Joint Institute of Frontier Technologies
{wxchern, jkzhu}@zju.edu.cn

Abstract

Learning from multimodal data is an important research topic in machine learning, which has the potential to obtain better representations. In this work, we propose a novel approach to generative modeling of multimodal data based on generative adversarial networks. To learn a coherent multimodal generative model, we show that it is necessary to align different encoder distributions with the joint decoder distribution simultaneously. To this end, we construct a specific form of the discriminator to enable our model to utilize data efficiently, which can be trained contrastively. By taking advantage of contrastive learning through factorizing the discriminator, we train our model on unimodal data. We have conducted experiments on the benchmark datasets, whose promising results show that our proposed approach outperforms the-state-of-the-art methods on a variety of metrics. The source code will be made publicly available.

1 Introduction

Real-world data usually have several heterogeneous modalities. Co-occurrence of different modalities provides the extra supervision signals, since the different modalities of data are usually associated by the underlying common factors. By exploiting such commonality among different modalities, the representations learned from the multimodal data are promised to be more robust and generalizable across different modalities. Existing works have shown the potential of the representations learned from the multimodal data (Tsai et al. 2019; Shi et al. 2021).

Supervised representation learning requires the labeled data, which are typically scarce comparing to the vast amount of unlabeled data. Recently, some progress on generative modeling has been made in the multimodal setting and demonstrates the power of the representations learned from the multimodal generative models (Suzuki, Nakayama, and Matsuo 2017; Wu and Goodman 2018; Shi et al. 2019; Sutter, Daunhawer, and Vogt 2020). In this work, we focus our attention on generative modeling of the multimodal data.

Generative models offer an elegant way to learn from the data without labels. Variational auto-encoders

(VAEs) (Kingma and Welling 2014) and generative adversarial networks (GANs) (Goodfellow et al. 2014) are two well-known representative deep generative models. In the original formulation, GANs are not able to perform inference like VAEs, the adversarially learned inference (ALI) and bidirectional GAN (BiGAN) (Donahue, Krähenbühl, and Darrell 2017; Dumoulin et al. 2017) proposed an extension by training the inference model and the generative model together using the adversarial process.

Recent approaches (Suzuki, Nakayama, and Matsuo 2017; Shi et al. 2019; Sutter, Daunhawer, and Vogt 2020) have extended the VAEs to the multimodal setting, where learning of the model corresponds to maximize the evidence variational lower bound (ELBO) on the marginal log-likelihood of multimodal data. However, they are not able to utilize the unimodal data efficiently, as combining unimodal ELBOs and multimodal ELBOs naively does not yield a valid lower bound on the marginal log-likelihood of multimodal data anymore, which may even hurt performance. Shi et al. (2021) propose a contrastive version of multimodal ELBO to improve data-efficiency, where the model can make use of the unpaired unimodal data. However, their optimization objective is derived heuristically and empirically weighted to avoid degenerate solutions, where the model only learns to generate the random noise.

To address the above limitations, in this paper, we propose a novel GAN-based multimodal generative model to make use of unimodal data more efficiently. We show that it is necessary to align the different encoder distributions with the joint decoder distribution, simultaneously. To align multiple distributions, we resort to adversarial learning. Specifically, training the discriminators in the D-step of GANs can be regarded as estimating the probability density ratio. This perspective of GANs enables us to factorize the discriminators and compute a wide range of divergences. By factorizing discriminators, we can train discriminators on unimodal data, which improves the data-efficiency. By taking advantages of the flexibility of GANs, we propose a principled way to combine the unimodal and multimodal training under the same framework.

From the above all, the contributions of our work are summarized in threefold as below:

- We show that the different encoder distributions and decoder distribution need to be aligned simultaneously to

*corresponding authors

learn a coherent multimodal generative model.

- We construct a specific form for the discriminator that is able to take advantage of unimodal data and contrastive learning.
- Our proposed approach outperforms the-state-of-the-art methods on a variety of metrics.

The rest of this paper is organized as follows. We summarize the desiderata of multimodal generative modeling in Section 2. We propose the optimization objective to address the desiderata in Section 3.1. We show how to factorize the joint discriminator to enable contrastive learning in Section 3.2. In Section 3.3, we show that it is necessary to factorize the latent space. The experiment results are presented in Section 4. We discuss some related works in Section 5. Finally, Section 6 gives the conclusions.

2 Background

To equip GANs with inference mechanism, Donahue, Krähenbühl, and Darrell (2017); Dumoulin et al. (2017) proposed to align the encoder distribution $q(\mathbf{x}, \mathbf{z}) = q(\mathbf{x})q(\mathbf{z}|\mathbf{x})$ and the decoder distribution $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ using an adversarial process:

$$\min_G \max_D V(D, G) = \mathbb{E}_{q(\mathbf{x})}[\log(D(\mathbf{x}, G_z(\mathbf{x})))] + \mathbb{E}_{p(\mathbf{z})}[\log(1 - D(G_x(\mathbf{z}), \mathbf{z}))] \quad (1)$$

where $q(\mathbf{x})$ is the data distribution, and $p(\mathbf{z})$ is the prior imposed on the latent code. $G_z(\mathbf{x})$ and $G_x(\mathbf{z})$ are the encoder and decoder corresponding to conditional distributions $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{z})$, respectively. The discriminator D takes both \mathbf{x} and \mathbf{z} as input, then determines if the pair comes from the encoder distribution or the decoder distribution.

By training an encoder together, the inference of the generative model can be performed via $p(\mathbf{z}|\mathbf{x}) = q(\mathbf{z}|\mathbf{x})$. When the encoder network $G_z(\mathbf{x})$ and the decoder network $G_x(\mathbf{z})$ are deterministic mappings and the global equilibrium of the game is reached, i.e., $q(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}, \mathbf{z})$, they invert each other almost everywhere (i.e., $\mathbf{x} = G_x(G_z(\mathbf{x}))$ and $\mathbf{z} = G_z(G_x(\mathbf{z}))$) even though the models are trained without the explicit reconstruction loss.

Aligning encoder and decoder distributions does not guarantee faithful reconstruction in the stochastic model due to non-identifiability of the model. Also, the global equilibrium might never be reached in practice. Li et al. (2017) illustrated that we can still achieve autoencoding property through stochastic mappings by adding the conditional entropy regularization terms to the original ALI/BiGAN objective.

In the unimodal setting, only generation $p(\mathbf{x}|\mathbf{z})$ and inference $q(\mathbf{z}|\mathbf{x})$ are considered. However, the situation becomes more complicated in the multimodal setting when we try to model the multimodal data distribution $q(\mathbf{X})$ with the sample $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^M$ having M modalities. As in the unimodal case, we intend to generate multimodal data $p(\mathbf{X}|\mathbf{z})$ as well as perform inference on fully observed multimodal data $q(\mathbf{z}|\mathbf{X})$. Since different modalities may be missed during inference, we want to perform inference given only partial

observation $q(\mathbf{z}|\mathbf{X}_K)$, where \mathbf{X}_K is a subset of K modalities observed. Furthermore we aim to generate other unobserved modalities conditioned on partial observation. Shi et al. (2019) have recently proposed desiderata for multimodal learning. We thus summarize the desiderata here for our multimodal generative model as follows:

- **Desideratum 1. Latent Factorization** The latent code learned by the model can be factorized into the modality-specific code and the code shared across modalities.
- **Desideratum 2. Coherent Joint Generation** The samples generated by the model should follow the joint distribution of the multimodal data.
- **Desideratum 3. Coherent Cross Generation** Given one set of modalities, the model can generate other set of modalities that follows the conditional distribution of the multimodal data.
- **Desideratum 4. Synergy** The model should benefit from observing more modalities.

3 Proposed Method

3.1 Multimodal Adversarially Learned Inference

In the multimodal setting, the observation is the multimodal data, from which we are interested in learning the shared latent representation for all modalities.

Given the multimodal data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^M$ with M modalities and the latent code \mathbf{z} , we consider the following M different unimodal encoder distributions:

$$q_i(\mathbf{X}, \mathbf{z}) = q(\mathbf{x}_1, \dots, \mathbf{x}_M)q(\mathbf{z}|\mathbf{x}_i) \quad (2)$$

where $q(\mathbf{X})$ is the joint distribution of the multimodal data. $q(\mathbf{z}|\mathbf{x}_i)$ is the conditional distribution corresponding to the encoder $G_z(\mathbf{x}_i)$ that maps the modality \mathbf{x}_i to the latent code \mathbf{z} . The samples from distribution $q_i(\mathbf{X}, \mathbf{z})$ are generated by the following steps. Firstly, we take a multimodal data sample \mathbf{X} from the data distribution $q(\mathbf{X})$, and then generate the latent code \mathbf{z} using only one modality $\mathbf{z} = G_z(\mathbf{x}_i)$.

With the assumption that different modalities are conditionally independent given the code \mathbf{z} , we consider the following decoder distribution:

$$p(\mathbf{X}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}_1|\mathbf{z}) \dots p(\mathbf{x}_M|\mathbf{z}) \quad (3)$$

where $p(\mathbf{z})$ is the prior distribution imposed on \mathbf{z} , and $p(\mathbf{x}_i|\mathbf{z})$ is the conditional distribution corresponding to the decoder $G_x(\mathbf{z})$ that maps the latent code \mathbf{z} to modality \mathbf{x}_i .

If we align the distribution $p(\mathbf{X}, \mathbf{z})$ with one of the encoder distributions $q_i(\mathbf{X}, \mathbf{z})$, then we can make sure that: (i) the generative model learns coherent generation, i.e., $p(\mathbf{X}) = q(\mathbf{X})$, which addresses desideratum 2 (coherent joint generation); (ii) we can perform inference on modality \mathbf{x}_i as $p(\mathbf{z}|\mathbf{x}_i) = q(\mathbf{z}|\mathbf{x}_i)$ and cross generation from \mathbf{x}_i to other modalities is coherent as $q(\mathbf{x}_i)q(\mathbf{z}|\mathbf{x}_i) \prod_{j \neq i}^M p(\mathbf{x}_j|\mathbf{z}) = p(\mathbf{z}) \prod_{i=1}^M p(\mathbf{x}_i|\mathbf{z})$. It is necessary to align the decoder distribution $p(\mathbf{X}, \mathbf{z})$ and every unimodal encoder distribution $q_i(\mathbf{X}, \mathbf{z})$ simultaneously, if we want to condition on arbitrary modality. Furthermore, if we want to condition on multiple modalities at the same time, we can introduce multimodal encoder distributions

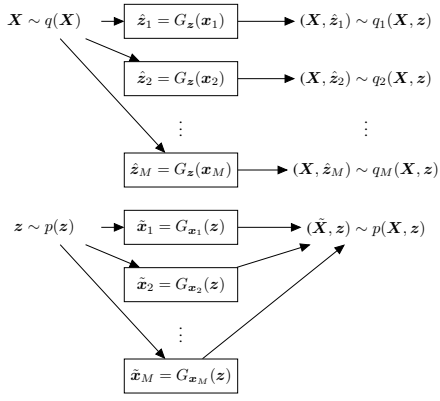


Figure 1: The encoder distributions and the decoder distribution to be aligned. The discriminator $D(\mathbf{X}, \mathbf{z})$ tries to distinguish among tuples $(\mathbf{X}, \hat{\mathbf{z}}_i)$ and $(\tilde{\mathbf{X}}, \mathbf{z})$.

$q(\mathbf{X})q(\mathbf{z}|\mathbf{X}_K)$, where \mathbf{X}_K is a subset of \mathbf{X} with K modalities, and align it with the decoder distribution as well. To fully address the desideratum 3 (coherent cross generation), we need to align all the unimodal and multimodal encoder distributions with the decoder distribution, simultaneously.

Taking consideration of scalability, previous methods (Wu and Goodman 2018; Shi et al. 2019; Sutter, Daunhawer, and Vogt 2020, 2021) proposed to approximate the multimodal encoders using the unimodal encoders instead of introducing the extra multimodal encoders, since it would otherwise require 2^{M-1} encoders in total. There are two common choices to achieve this, the product of experts (PoE) and the mixture of experts (MoE). To align all encoder distributions simultaneously, PoE should therefore be avoided, as the product of experts is sharper than any of its experts. This makes it impossible to align different encoder distributions.

Since our goal is to align all unimodal and multimodal encoders, we only need to align the unimodal encoder distributions and approximate the multimodal encoders via any abstract mean function (Nielsen 2020). The multimodal encoder distributions are aligned automatically by aligning the unimodal encoder distributions alone as $q(\mathbf{z}|\mathbf{X}_K) = \frac{1}{K} \sum_{i=1}^K q(\mathbf{z}|\mathbf{x}_i) = q(\mathbf{z}|\mathbf{x}_1) = \dots = q(\mathbf{z}|\mathbf{x}_M)$, if we take the arithmetic mean (i.e., MoE) as the abstract mean function for example.

As per the discussion above, we only need to align all unimodal encoder distributions and the decoder distribution during training. Afterwards, we can compute the multimodal posterior in the closed-form through the appropriate mean function during testing.

To achieve this goal, we resort to adversarial learning, resulting in **MultiModal Adversarially Learned Inference**:

$$\begin{aligned} \min_G \max_D V(D, G) = & \mathbb{E}_{p(\mathbf{z})} [\log D(G_{\mathbf{X}}(\mathbf{z}), \mathbf{z})[M+1]] \\ & + \sum_{i=1}^M \mathbb{E}_{q(\mathbf{X})} [\log D(\mathbf{X}, G_{\mathbf{z}}(\mathbf{x}_i))[i]] \end{aligned} \quad (4)$$

where D is the discriminator that takes both \mathbf{X} and \mathbf{z} as input and outputs $M+1$ probabilities. G represents the corresponding unimodal encoders and decoders with a slight abuse of notation. Unlike ordinary GANs that only align two distributions, our proposed approach aligns $M+1$ distributions. It can be shown that this objective is equivalent to minimizing the Jensen–Shannon divergence of all distributions (Pu et al. 2018).

3.2 Discriminator Factorization

Given the model specification discussed in Section 3.1, we show that a specific form of the discriminator can be constructed instead of a monolithic one.

It can be shown that the optimal discriminator D^* given the fixed encoders and decoders is:

$$D^*(\mathbf{X}, \mathbf{z})[i] = \begin{cases} \frac{q_i(\mathbf{X}, \mathbf{z})}{p(\mathbf{X}, \mathbf{z}) + \sum_{j=1}^M q_j(\mathbf{X}, \mathbf{z})}, & 1 \leq i \leq M \\ \frac{p(\mathbf{X}, \mathbf{z})}{p(\mathbf{X}, \mathbf{z}) + \sum_{j=1}^M q_j(\mathbf{X}, \mathbf{z})}, & i = M+1 \end{cases} \quad (5)$$

The proof can be found in Pu et al. (2018). This optimal joint discriminator can be expressed with the probability density ratios $r_i(\mathbf{X}, \mathbf{z}) = \frac{q_i(\mathbf{X}, \mathbf{z})}{p(\mathbf{X}, \mathbf{z})}$:

$$D^*(\mathbf{X}, \mathbf{z})[i] = \begin{cases} \frac{r_i(\mathbf{X}, \mathbf{z})}{1 + \sum_{j=1}^M r_j(\mathbf{X}, \mathbf{z})}, & 1 \leq i \leq M \\ \frac{1}{1 + \sum_{j=1}^M r_j(\mathbf{X}, \mathbf{z})}, & i = M+1 \end{cases} \quad (6)$$

According to the model specification discussed before, for each $r_i(\mathbf{X}, \mathbf{z})$, we have:

$$\begin{aligned} r_i(\mathbf{X}, \mathbf{z}) &= \frac{q_i(\mathbf{X}, \mathbf{z})}{p(\mathbf{X}, \mathbf{z})} \\ &= \frac{q(\mathbf{x}_1, \dots, \mathbf{x}_M)q(\mathbf{z}|\mathbf{x}_i)}{p(\mathbf{z}) \prod_{j=1}^M p(\mathbf{x}_j|\mathbf{z})} \\ &= \frac{q(\mathbf{x}_1, \dots, \mathbf{x}_M)}{\prod_k^M q(\mathbf{x}_k)} \frac{q(\mathbf{z}|\mathbf{x}_i)}{p(\mathbf{z})} \prod_j^M \frac{q(\mathbf{x}_j)}{p(\mathbf{x}_j|\mathbf{z})} \\ &= \frac{q(\mathbf{X})}{\prod_k^M q(\mathbf{x}_k)} \frac{q(\mathbf{x}_i)q(\mathbf{z}|\mathbf{x}_i)}{q(\mathbf{x}_i)p(\mathbf{z})} \prod_j^M \frac{q(\mathbf{x}_j)p(\mathbf{z})}{p(\mathbf{x}_j|\mathbf{z})p(\mathbf{z})} \\ &= \frac{q(\mathbf{X})}{\prod_k^M q(\mathbf{x}_k)} \frac{q(\mathbf{x}_i, \mathbf{z})}{q(\mathbf{x}_i)p(\mathbf{z})} \prod_j^M \frac{q(\mathbf{x}_j)p(\mathbf{z})}{p(\mathbf{x}_j, \mathbf{z})} \end{aligned} \quad (7)$$

Now we factorize the joint discriminator into the smaller discriminators that can be trained separately on the unimodal data. For each modality, we train two discriminators. To estimate $\frac{q(\mathbf{x}_i, \mathbf{z})}{q(\mathbf{x}_i)p(\mathbf{z})}$, we take samples from $q(\mathbf{x}_i, \mathbf{z})$ as positive and samples from $q(\mathbf{x}_i)p(\mathbf{z})$ as negative. To estimate $\frac{q(\mathbf{x}_i, \mathbf{z})}{p(\mathbf{x}_i, \mathbf{z})}$, we take samples from $q(\mathbf{x}_i, \mathbf{z})$ as positive and samples from $p(\mathbf{x}_i, \mathbf{z})$ as negative. $\frac{q(\mathbf{x}_i)p(\mathbf{z})}{p(\mathbf{x}_i, \mathbf{z})}$ is then computed by $\frac{q(\mathbf{x}_i, \mathbf{z})}{p(\mathbf{x}_i, \mathbf{z})} / \frac{q(\mathbf{x}_i, \mathbf{z})}{q(\mathbf{x}_i)p(\mathbf{z})}$. As for $\frac{q(\mathbf{x}_1, \dots, \mathbf{x}_M)}{\prod_k^M q(\mathbf{x}_k)}$, we train the separate discriminator by taking samples from $q(\mathbf{X})$ as positive and samples from the product of marginals $\prod_k^M q(\mathbf{x}_k)$

as negative. The advantages of such factorization lies in twofold. Firstly, the discriminator for $\frac{q(\mathbf{X})}{\prod_k^M q(\mathbf{x}_k)}$ can utilize multimodal data efficiently by learning from unpaired samples as well. Secondly, we can update encoders and decoders on unimodal data, since we now have probability density ratios $\frac{q(\mathbf{x}_i, \mathbf{z})}{p(\mathbf{x}_i, \mathbf{z})}$.

A special case for $\frac{q(\mathbf{X})}{\prod_k^M q(\mathbf{x}_k)}$ in the two-modality setting is related to mutual information maximization. Note that each estimated probability density ratio $r_i(\mathbf{X}, \mathbf{z}) \propto \frac{q(\mathbf{x}_1, \mathbf{z})}{q(\mathbf{x}_1)q(\mathbf{x}_2)}$. By updating the decoder distribution to minimize $\mathbb{E}_{p(\mathbf{X}, \mathbf{z})} \log D^*(\mathbf{X}, \mathbf{z})[M + 1] = -\mathbb{E}_{p(\mathbf{X}, \mathbf{z})} \log(1 + \sum_i^M r_i(\mathbf{X}, \mathbf{z}))$, the decoders are updated to maximize the mutual information of two modalities implicitly.

3.3 Latent Space Factorization

As discussed in the previous section, we force different unimodal encoders to have the same conditional distribution $q(\mathbf{X})q(\mathbf{z}|\mathbf{x}_i) = q(\mathbf{X})q(\mathbf{z}|\mathbf{x}_j)$, which may make the modality-specific information lost. Our decoders are the deterministic mappings parameterized by neural networks. Without extra information, the model will struggle to balance between information preserving and distribution matching. We therefore introduce the modality-specific codes to alleviate this problem.

By introducing the modality-specific style codes $\mathbf{S} = \{\mathbf{s}_i\}_{i=1}^M$ and the shared content code \mathbf{c} , with the assumption that modality-specific style codes \mathbf{S} and the shared content code \mathbf{c} are conditionally independent given the observed data \mathbf{X} . The unimodal encoder distributions can be derived as below:

$$q_i(\mathbf{X}, \mathbf{S}, \mathbf{c}) = q(\mathbf{X})q(\mathbf{c}|\mathbf{x}_i) \prod_j^M q(\mathbf{s}_j|\mathbf{x}_j) \quad (8)$$

and the decoder distribution is as follows:

$$p(\mathbf{X}, \mathbf{S}, \mathbf{c}) = p(\mathbf{c}) \prod_i^M p(\mathbf{s}_i)p(\mathbf{x}_i|\mathbf{s}_i, \mathbf{c}) \quad (9)$$

Following the new model specification, the similar result can be obtained for $\frac{q_i(\mathbf{X}, \mathbf{S}, \mathbf{c})}{p(\mathbf{X}, \mathbf{S}, \mathbf{c})}$:

$$\begin{aligned} \frac{q_i(\mathbf{X}, \mathbf{S}, \mathbf{c})}{p(\mathbf{X}, \mathbf{S}, \mathbf{c})} &= \frac{q(\mathbf{x}_1, \dots, \mathbf{x}_M)q(\mathbf{c}|\mathbf{x}_i) \prod_j^M q(\mathbf{s}_j|\mathbf{x}_j)}{p(\mathbf{c}) \prod_k^M p(\mathbf{s}_k)p(\mathbf{x}_k|\mathbf{s}_k, \mathbf{c})} \\ &= \frac{q(\mathbf{X})}{\prod_k^M q(\mathbf{x}_k)} \frac{q(\mathbf{c}|\mathbf{x}_i)}{p(\mathbf{c})} \prod_j^M \frac{q(\mathbf{x}_j)q(\mathbf{s}_j|\mathbf{x}_j)}{p(\mathbf{s}_j)p(\mathbf{x}_j|\mathbf{s}_j, \mathbf{c})} \\ &= \frac{q(\mathbf{X})}{\prod_k^M q(\mathbf{x}_k)} \frac{q(\mathbf{x}_i, \mathbf{s}_i, \mathbf{c})}{q(\mathbf{x}_i, \mathbf{s}_i)p(\mathbf{c})} \prod_j^M \frac{q(\mathbf{x}_j, \mathbf{s}_j)p(\mathbf{c})}{p(\mathbf{x}_j, \mathbf{s}_j, \mathbf{c})} \end{aligned} \quad (10)$$

Now we can train the factorized discriminators as discussed above.

4 Experiments

To examine the effectiveness of our proposed approach, we conduct experiments on three different datasets. Across all the experiments, we use Adam optimizer (Kingma and Ba 2015) with the learning rate 0.0002, in which all models are trained for 250,000 iterations with the batch size 64. We also employ an exponential moving average (Yazici et al. 2019) of the weights for both encoders and decoders with a decay rate 0.9999 and start at the 50,000-th iteration. In all models, the standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is chosen as the prior, and the isotropic Gaussian $\mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ is chosen as the posterior. We make use of the non-saturating loss described in Dandi et al. (2021) to update the encoders and decoders.

4.1 MultiMNIST

In this experiment, we compare the model trained with the factorized discriminator and the joint discriminator to investigate the effectiveness of discriminator factorization.



Figure 2: 5 samples of 4-modality MultiMNIST dataset.

Setup MultiMNIST dataset consists of multiple modalities, where every modality is a rotated version of MNIST (Deng 2012). Every instance in one modality is associated with 30 random instances of the same digit class from other modalities. We perform experiments up to four modalities, in which each modality is rotated counterclockwise 0, 90, 180, 270 degrees, respectively. The performance is assessed by training a linear classifier on top of the frozen encoder of the original MNIST modality (rotated 0 degrees). In our empirical study, the same architecture of the encoders and the decoders is used for each modality across every experiments.

Methods	Modalities (%)		
	2	3	4
Joint	85.96	85.44	85.27
Factorized	97.02	96.94	96.75

Table 1: The classification accuracy of the representations of encoders trained with different discriminators.

Table 1 shows that the performance of the representations of encoders trained with factorized discriminators is improved significantly. This experiment also demonstrates the generalizability on more than two modalities for the proposed method. Additionally, the qualitative results are shown in Figure 3.

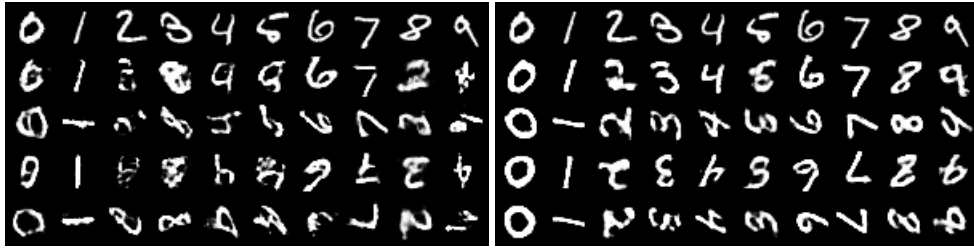


Figure 3: Left: Joint Right: Factorized. From top to bottom: original, reconstruction, translation to 90, translation to 180, translation to 270.

4.2 MNIST-SVHN

We compare the proposed method with the-state-of-the-art methods on MNIST-SVHN.

Setup MNIST-SVHN dataset contains two modalities, MNIST and SVHN (Netzer et al. 2011), where each instance of a digit class is paired with 30 random instances from the other dataset with the same digit class. We use the same architecture of encoders and decoders used by Shi et al. (2021), where MLPs are used for MNIST and CNNs are used for SVHN. Furthermore, we use the same 20-d for the latent space while splitting it into the 10-d content code and 10-d style code. As suggested in Shi et al. (2019, 2021), the following metrics are used to evaluate the performance of the models.

- **Latent accuracy** Latent accuracy is computed by the classification accuracy of a linear classifier trained on the learned representation of encoders.
- **Joint coherence accuracy** Joint coherence accuracy is computed by how often multimodal generations are classified to have the same label. We pre-train the classifiers of each modality and apply them on the multimodal generations $\mathbf{x}_1, \mathbf{x}_2 \sim p(\mathbf{z})p(\mathbf{x}_1, \mathbf{x}_2|\mathbf{z})$ to check how often \mathbf{x}_1 and \mathbf{x}_2 have the same label.
- **Cross coherence accuracy** Cross coherence accuracy is computed similarly to joint coherence accuracy. We apply the pre-trained classifiers on the cross generated samples $\mathbf{x}_1, \hat{\mathbf{x}}_2 \sim q(\mathbf{x}_1)q(\mathbf{z}|\mathbf{x}_1)p(\hat{\mathbf{x}}_2|\mathbf{z})$ and $\hat{\mathbf{x}}_1, \mathbf{x}_2 \sim q(\mathbf{x}_2)q(\mathbf{z}|\mathbf{x}_2)p(\hat{\mathbf{x}}_1|\mathbf{z})$ and compute how often the label remains the same.
- **Synergy coherence accuracy** Synergy coherence accuracy is computed by how often the joint reconstructions have the same label as the original input. The pre-trained classifiers are now applied on $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2 \sim q(\mathbf{x}_1, \mathbf{x}_2)q(\mathbf{z}|\mathbf{x}_1, \mathbf{x}_2)p(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2|\mathbf{z})$

The experimental results are shown in Table 2. Moreover, we report the synergy coherence accuracy for our model by computing the joint posterior of Gaussian experts in the closed-form via geometric mean. It can be seen that our proposed model outperforms others in almost every metrics, especially on the joint coherence accuracy.

Latent space factorization We evaluate the effectiveness of latent space factorization. To this end, we train the linear classifiers on the content code, the style code and the

whole latent code, respectively. Table 3 shows that the classifiers trained on the whole latent code only outperform the classifiers trained on the content code marginally while performing significantly better than the classifiers trained on the style code. Figure 4 depicts the visual effect of latent space factorization. It can be seen that our proposed approach is able to effectively factorize the content from the style.

4.3 CUB Image-Captions

We then evaluate our method on Caltech-UCSD Birds (CUB) dataset (Welinder et al. 2010), where each image of birds is annotated with 10 different captions. Since generating discrete data is challenging for GANs, we circumvent this by modeling it in the continuous feature space. The image features are obtained from the pre-trained ResNet-101 model (He et al. 2016), and the features of caption are extracted from a pre-trained word-level CNN-LSTM autoencoder (Pu et al. 2016).

We adopt the evaluation method employed by Massiceti et al. (2018); Shi et al. (2019) to compute the correlation score as the groundtruth. We perform Canonical Correlation Analysis (CCA) on the feature space of two modalities, which learns two projection $W_1 \in \mathbb{R}^{n_1 \times k}$ and $W_2 \in \mathbb{R}^{n_2 \times k}$ mapping the observations $x_1 \in \mathbb{R}^{n_1}$ and $x_2 \in \mathbb{R}^{n_2}$ to the same dimension $W_1^T x_1$ and $W_2^T x_2$. The correlation score of new observations \tilde{x}_1, \tilde{x}_2 is then computed by:

$$\text{corr}(\tilde{x}_1, \tilde{x}_2) = \frac{\phi(\tilde{x}_1)^T \phi(\tilde{x}_2)}{\|\phi(\tilde{x}_1)\|_2 \|\phi(\tilde{x}_2)\|_2} \quad (11)$$

where $\phi(\tilde{x}_i) = W_i^T \tilde{x}_i - \text{avg}(W_i^T x_i)$

Quantitative results We perform CCA using different features to calculate the correlation score. Although the ground truth of the test set is different from Shi et al. (2019), we include the results of MMVAE for reference. Note that the correlation score of our model is higher than the ground truth score while MMVAE and MVAE are lower than theirs' score.

Qualitative results To visualize the result, we project from the image feature space back to the pixel space by performing a nearest-neighbour lookup using Euclidean distance in the feature space on the training set. Moreover, we employ the pre-trained LSTM decoder to decode the generated caption features back to captions. Figure 5 and Figure 6 show the results of cross-modal generations in both directions.

Data	Methods	Latent (%)		Joint (%)	Cross Generation (%)		Synergy (%)	
		M	S		S \rightarrow M	M \rightarrow S	M	S
100%	JMVAE	84.45 (± 0.87)	57.98 (± 1.27)	42.18 (± 1.50)	49.63 (± 1.78)	54.98 (± 3.02)	85.77 (± 0.66)	68.15 (± 1.38)
	cI-JMVAE	84.58 (± 1.49)	64.42 (± 1.42)	48.95 (± 2.31)	58.16 (± 1.83)	70.61 (± 3.13)	93.45 (± 0.52)	84.00 (± 0.97)
	cC-JMVAE	83.67 (± 3.48)	66.64 (± 2.92)	47.27 (± 4.52)	59.73 (± 3.85)	69.49 (± 2.19)	91.21 (± 6.59)	84.07 (± 3.19)
	MVAE	91.65 (± 0.17)	64.12 (± 4.58)	9.42 (± 7.82)	10.98 (± 0.56)	21.88 (± 2.21)	64.60 (± 9.25)	52.91 (± 8.11)
	cI-MVAE	96.97 (± 0.84)	75.94 (± 6.20)	15.23 (± 10.46)	10.85 (± 1.17)	27.70 (± 2.09)	85.07 (± 7.73)	75.67 (± 4.13)
	cC-MVAE	97.42 (± 0.40)	81.07 (± 2.03)	8.85 (± 3.86)	12.83 (± 2.25)	30.03 (± 2.46)	75.25 (± 5.31)	69.42 (± 3.94)
	MMVAE	92.48 (± 0.37)	79.03 (± 1.17)	42.32 (± 2.97)	70.77 (± 0.35)	85.50 (± 1.05)	—	—
	cI-MMVAE	93.97 (± 0.36)	81.87 (± 0.52)	43.94 (± 0.96)	79.66 (± 0.59)	92.67 (± 0.29)	—	—
	cC-MMVAE	93.10 (± 0.17)	80.88 (± 0.80)	45.46 (± 0.78)	79.34 (± 0.54)	92.35 (± 0.46)	—	—
	MMJSD (MS)	95.76 (± 0.20)	78.12 (± 0.83)	25.33 (± 1.30)	45.81 (± 4.09)	86.33 (± 0.72)	89.66 (± 0.18)	82.62 (± 0.49)
	MMALI (Ours)	96.82 (± 0.59)	87.20 (± 0.75)	79.31 (± 0.01)	83.80 (± 0.01)	95.22 (± 0.02)	96.43 (± 0.01)	96.69 (± 0.01)
20%	JMVAE	77.53 (± 0.13)	52.55 (± 2.18)	26.37 (± 0.54)	42.58 (± 5.32)	41.44 (± 2.26)	85.07 (± 9.74)	51.95 (± 2.28)
	cI-JMVAE	77.57 (± 4.02)	57.91 (± 1.28)	32.58 (± 5.89)	51.85 (± 1.27)	47.92 (± 10.32)	92.54 (± 1.13)	67.01 (± 8.72)
	cC-JMVAE	81.11 (± 2.76)	57.85 (± 2.23)	34.00 (± 7.18)	50.73 (± 0.45)	56.89 (± 6.18)	88.36 (± 4.38)	68.49 (± 8.82)
	MVAE	90.29 (± 0.57)	33.44 (± 0.26)	10.88 (± 9.15)	8.72 (± 0.92)	12.12 (± 3.38)	42.10 (± 5.22)	44.95 (± 5.92)
	cI-MVAE	93.72 (± 1.09)	56.74 (± 7.97)	12.79 (± 6.82)	14.18 (± 2.19)	20.23 (± 4.55)	75.36 (± 5.05)	64.81 (± 4.81)
	cC-MVAE	92.74 (± 2.97)	52.99 (± 8.35)	17.95 (± 12.52)	14.70 (± 1.65)	24.90 (± 5.77)	56.86 (± 18.84)	54.28 (± 9.86)
	MMVAE	88.54 (± 0.37)	68.90 (± 1.79)	37.71 (± 0.60)	59.52 (± 0.28)	76.33 (± 2.23)	—	—
	cI-MMVAE	91.64 (± 0.06)	73.02 (± 0.80)	42.74 (± 0.36)	69.51 (± 1.18)	86.75 (± 0.28)	—	—
	cC-MMVAE	92.10 (± 0.19)	71.29 (± 1.05)	40.77 (± 0.93)	68.43 (± 0.90)	86.24 (± 0.89)	—	—
	MMJSD (MS)	92.78 (± 0.07)	61.17 (± 1.62)	19.29 (± 1.43)	29.17 (± 1.17)	68.05 (± 0.75)	87.97 (± 0.51)	64.85 (± 0.89)
	MMALI (Ours)	94.03 (± 0.64)	80.16 (± 0.74)	75.88 (± 0.01)	72.83 (± 0.02)	86.19 (± 0.03)	84.21 (± 0.02)	85.24 (± 0.01)

Table 2: Baselines JMVAE, MVAE, MMVAE, their contrastive variations, MMJSD with modality-specific code and our method (I stands for IWAE (Burda, Grosse, and Salakhutdinov 2016) and C stands for CUBO (Dieng et al. 2017)). M stands for MNIST, and S represents SVHN.

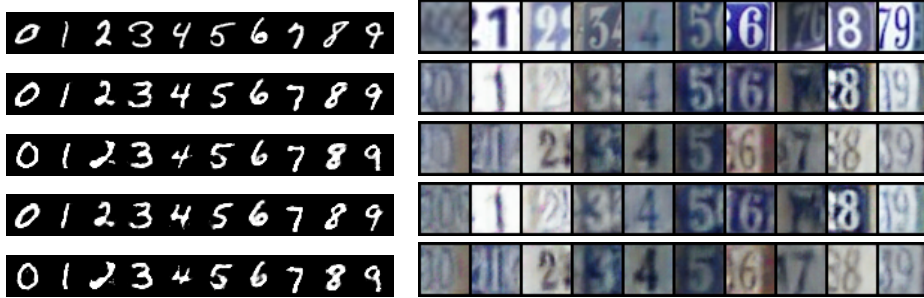


Figure 4: Latent space factorization. From top to bottom: original, reconstruction, reconstruction with random style, cross translation with original style, cross translation with random style.

	Content	Style	Both
MNIST	96.61	23.56	96.88
SVHN	86.40	21.39	86.49

Table 3: Classification accuracy (%) with different latent codes.

5 Related works

Multimodal VAEs In the multimodal setting of VAEs, Suzuki, Nakayama, and Matsuo (2017) introduced JVAE, where an extra multimodal encoder is trained to be aligned with other unimodal encoders. Similarly, Vedantam et al. (2018) utilized an extra multimodal encoder, while their training procedure is split into two stages. Moreover, they proposed to use PoE to handle the missing data during inference. Since introducing the extra multimodal encoders

explicitly is exponential in modalities, Wu and Goodman (2018) proposed MVAE to approximate the multimodal encoder through the product of experts. However, MVAE failed to translate from one modality to another. One of the motivations of PoE was to produce the sharper joint posterior, since it would increase the beliefs when more modalities are observed (Vedantam et al. 2018; Kurle, Günnemann, and van der Smagt 2019). Our analysis suggests if we want to translate among different modalities, the latent code should be aligned. Most recently, Shi et al. (2019) proposed to replace PoE with MoE, which achieved the better cross generation performance. Sutter, Daunhawer, and Vogt (2020) enabled the efficient training by computing the joint posterior in closed-form, which eliminates inefficient sampling to approximate the joint posterior. Furthermore, Shi et al. (2021) introduced a contrastive version of ELBO to improve data-efficiency. By introducing contrastive term, they make the model to utilize the unpaired data. Unfortu-

	Joint	Cross (I \rightarrow C)	Cross (C \rightarrow I)	Ground Truth
MVAE	-0.095	0.011	-0.013	0.273
MMVAE	0.263	0.104	0.135	0.273
MMALI (ours)	0.502	0.488	0.545	0.406

Table 4: Correlation score of Image (I)-Caption (C) pair for joint and cross generation.



a small black bird with a long black beak and a black head



this particular bird has a white belly and breast and a small beak



small yellow and brown bird with a long pointed beak and black and yellow feathers on its breast

Figure 5: From image to caption.



this bird has wings that are brown and has a long neck



a black bird with a long wing span and a small beak



this bird is blue with black and has a very short beak

Figure 6: From caption to image.

nately, their objective is heuristically derived and empirically weighted to avoid the degenerate solutions, where the models only generates random noise. Instead, our model employs a discriminator to achieve contrastive learning.

GANs with specific form of discriminators Miyato and Koyama (2018) proposed a specific form for the discriminator of cGAN. Instead of concatenating the conditional vector to the feature vectors, they constructed the discriminator in the form of the dot product of the conditional vector and the feature vector with certain regularity assumption. They showed that a specific form of discriminator can improve the conditional image generation quality significantly. Our approach is inspired by FactorGAN (Stoller, Ewert, and Dixon 2020), where they proposed to factorize the joint data distribution into a set of lower-dimensional distributions. Such factorization allows the model to be trained on incomplete observation. The factorization in Stoller, Ewert, and Dixon (2020) is quite arbitrary while our factorization naturally follows the model specification.

Factorized latent space Factorizing the latent space into modality-specific and shared codes is not a new idea. In image-to-image translation, MUNIT (Huang, Belongie, and Kautz 2018) introduced content and style codes, which aimed at achieving many-to-many translation. In the multimodal generative modeling, Tsai et al. (2019) proposed to factorize the latent space into multimodal discriminative factors and modality-specific generative factors. Sutter, Daun-

hawer, and Vogt (2020) also used the modality-specific code to improve the performance. Our analysis suggests that latent space factorization is necessary in the multimodal setting when the decoder is limited to the restricted model class.

6 Conclusion

In this work, we proposed a new framework for multimodal generative modeling based on generative adversarial networks. To learn a coherent multimodal generative model, all unimodal encoder distributions were required to be aligned with the joint decoder distribution, simultaneously. The consequence of aligning multiple encoder distributions is that the product of experts should be avoided and the latent space need to be factorized. By factorizing the discriminator, we can utilize the unpaired data more efficiently and embed the contrastive learning mechanism in our framework. We conducted extensive experiments, whose promising results show that our proposed method outperforms the state-of-the-art method on a variety of metrics across different datasets.

Despite of the encouraging result, we discuss the limitations of our approach. We found the performance drop if only more unimodal data are added while keeping the paired multimodal data unchanged. For our future work, we will investigate the semi-supervised learning and extend the method to the sequence generation on more challenging datasets.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grants (61831015).

References

- Burda, Y.; Grosse, R. B.; and Salakhutdinov, R. 2016. Importance Weighted Autoencoders. In Bengio, Y.; and LeCun, Y., eds., *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.
- Dandi, Y.; Bharadhwaj, H.; Kumar, A.; and Rai, P. 2021. Generalized Adversarially Learned Inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8): 7185–7192.
- Deng, L. 2012. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine*, 29(6): 141–142.
- Dieng, A. B.; Tran, D.; Ranganath, R.; Paisley, J.; and Blei, D. 2017. Variational Inference via χ Upper Bound Minimization. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Donahue, J.; Krähenbühl, P.; and Darrell, T. 2017. Adversarial Feature Learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Dumoulin, V.; Belghazi, I.; Poole, B.; Lamb, A.; Arjovsky, M.; Mastropietro, O.; and Courville, A. C. 2017. Adversarially Learned Inference. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.
- Huang, M.-Y., Xunand Liu; Belongie, S.; and Kautz, J. 2018. Multimodal Unsupervised Image-to-Image Translation. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision – ECCV 2018*, 179–196. Cham: Springer International Publishing. ISBN 978-3-030-01219-9.
- Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.
- Kurle, R.; Günnemann, S.; and van der Smagt, P. 2019. Multi-Source Neural Variational Inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 4114–4121.
- Li, C.; Liu, H.; Chen, C.; Pu, Y.; Chen, L.; Henao, R.; and Carin, L. 2017. ALICE: Towards Understanding Adversarial Learning for Joint Distribution Matching. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Massiceti, D.; Dokania, P. K.; Siddharth, N.; and Torr, P. H. S. 2018. Visual Dialogue without Vision or Dialogue. *CoRR*, abs/1812.06417.
- Miyato, T.; and Koyama, M. 2018. cGANs with Projection Discriminator. In *International Conference on Learning Representations*.
- Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.
- Nielsen, F. 2020. On a Generalization of the Jensen-Shannon Divergence and the Jensen-Shannon Centroid. *Entropy*, 22(2): 221.
- Pu, Y.; Dai, S.; Gan, Z.; Wang, W.; Wang, G.; Zhang, Y.; Henao, R.; and Duke, L. C. 2018. JointGAN: Multi-Domain Joint Distribution Learning with Generative Adversarial Nets. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 4151–4160. PMLR.
- Pu, Y.; Gan, Z.; Henao, R.; Yuan, X.; Li, C.; Stevens, A.; and Carin, L. 2016. Variational Autoencoder for Deep Learning of Images, Labels and Captions. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Shi, Y.; Paige, B.; Torr, P.; and N, S. 2021. Relating by Contrasting: A Data-efficient Framework for Multimodal Generative Models. In *International Conference on Learning Representations*.
- Shi, Y.; Siddharth, N.; Paige, B.; and Torr, P. 2019. Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. In *Advances in Neural Information Processing Systems*, 15692–15703.
- Stoller, D.; Ewert, S.; and Dixon, S. 2020. Training Generative Adversarial Networks from Incomplete Observations using Factorised Discriminators. In *International Conference on Learning Representations*.
- Sutter, T.; Daunhawer, I.; and Vogt, J. 2020. Multimodal Generative Learning Utilizing Jensen-Shannon-Divergence.

In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 6100–6110. Curran Associates, Inc.

Sutter, T. M.; Daunhawer, I.; and Vogt, J. E. 2021. Generalized Multimodal ELBO. In *International Conference on Learning Representations*.

Suzuki, M.; Nakayama, K.; and Matsuo, Y. 2017. Joint Multimodal Learning with Deep Generative Models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2019. Learning Factorized Multimodal Representations. In *International Conference on Learning Representations*.

Vedantam, R.; Fischer, I.; Huang, J.; and Murphy, K. 2018. Generative Models of Visually Grounded Imagination. In *International Conference on Learning Representations*.

Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.

Wu, M.; and Goodman, N. 2018. Multimodal Generative Models for Scalable Weakly-Supervised Learning. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montreal, Canada*, 5580–5590.

Yazici, Y.; Foo, C.; Winkler, S.; Yap, K.; Piliouras, G.; and Chandrasekhar, V. 2019. The Unusual Effectiveness of Averaging in GAN Training. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.