ELASTIC-INFOGAN: UNSUPERVISED DISENTANGLED REPRESENTATION LEARNING IN IMBALANCED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a novel unsupervised generative model, Elastic-InfoGAN, that learns to disentangle object identity from other low-level aspects in class-imbalanced datasets. We first investigate the issues surrounding the assumptions about uniformity made by InfoGAN (Chen et al. (2016)), and demonstrate its ineffectiveness to properly disentangle object identity in imbalanced data. Our key idea is to make the discovery of the discrete latent factor of variation invariant to identity-preserving transformations in real images, and use that as the signal to learn the latent distribution's parameters. Experiments on both artificial (MNIST) and real-world (YouTube-Faces) datasets demonstrate the effectiveness of our approach in imbalanced data by: (i) better disentanglement of object identity as a latent factor of variation; and (ii) better approximation of class imbalance in the data, as reflected in the learned parameters of the latent distribution.

1 INTRODUCTION

Generative models aim to model the true data distribution, so that *fake* samples that seemingly belong to the modeled distribution can be generated (Ackley et al. (1985); Rabiner (1989); Blei et al. (2003)). Recent deep neural network based models such as Generative Adversarial Networks (Goodfellow et al. (2014); Salimans et al. (2016); Radford et al. (2016)) and Variational Autoencoders (Kingma & Welling (2014); Higgins et al. (2017)) have led to promising results in generating realistic samples for high-dimensional and complex data such as images. More advanced models show how to discover *disentangled* representations (Yan et al. (2016); Chen et al. (2016); Tran et al. (2017); Hu et al. (2018); Singh et al. (2019)), in which different latent dimensions can be made to represent independent factors of variation (e.g., pose, identity) in the data (e.g., human faces).

InfoGAN (Chen et al. (2016)) in particular, tries to learn an unsupervised disentangled representation by maximizing the mutual information between the discrete or continuous latent variables and the corresponding generated samples. For discrete latent factors (e.g., digit identities), it assumes that they are uniformly distributed in the data, and approximates them accordingly using a *fixed uniform* categorical distribution. Although this assumption holds true for many existing benchmark datasets (e.g., MNIST LeCun (1998)), real-word data often follows a long-tailed distribution and rarely exhibits perfect balance between the categories. Indeed, applying InfoGAN on imbalanced data can result in incoherent groupings, since it is forced to discover potentially non-existent factors that are uniformly distributed in the data; see Fig. 1.

In this work, we augment InfoGAN to discover disentangled categorical representations from *imbalanced* data. Our model, Elastic-InfoGAN, makes two modifications to InfoGAN which are simple and intuitive. First, we remodel the way the latent distribution is used to fetch the latent variables; we lift the assumption of any knowledge about class imbalance, where instead of deciding and fixing them beforehand, we treat the class probabilities as *learnable parameters* of the optimization process. To enable the flow of gradients back to the class probabilities, we employ the Gumbel-Softmax distribution (Jang et al. (2017); Maddison et al. (2017)), which acts as a proxy for the categorical distribution, generating *differentiable* samples having properties similar to that of categorical samples. Second, we enforce our network to assign the same latent category for an image I and its transformed image I', which induces the discovered latent factors to be invariant to *identity-preserving* transformations like illumination, translation, rotation, and scale changes. Although there are multiple meaningful ways to partition unlabeled data—e.g., with digits, one partitioning could be based

Figure 1: (**Left** & **Center**): Samples generated with an InfoGAN model learned with a fixed uniform categorical distribution $Cat(K = 10, p = 0.1)$ on balanced and imbalanced data, respectively. Each row corresponds to a different learned latent category. (**Right**): Samples generated with Elastic-InfoGAN using its automatically learned latent categorical distribution. Although InfoGAN discovers digit identities in the balanced data, it produces redundant/incoherent groupings in the imbalanced data. In contrast, our model is able to discover digit identities in the imbalanced data.

on identity, whereas another could be based on stroke width—we aim to discover the partitioning that groups objects according to a high-level factor like identity while being invariant to low-level "nuisance" factors like lighting, pose, and scale changes. Such partitionings focusing on object identity are more likely to be useful for downstream visual recognition applications (e.g., semi-supervised object recognition). In sum, our modifications to InfoGAN lead to better disentanglement and categorical grouping of the data (Fig. 1), while at the same time enabling the discovery of the original imbalance through the learned probability parameters of the Gumbel softmax distribution. Importantly, these modifications do not impede InfoGAN's ability to jointly model both continuous and discrete factors in either balanced or imbalanced data scenarios.

Our contributions can be summarized as follows: (1) To our knowledge, our work is the first to tackle the problem of unsupervised generative modeling of categorical disentangled representations in *imbalanced* data. We show qualitatively and quantitatively our superiority in comparison to InfoGAN and other relevant baselines. (2) Our work takes a step forward in the direction of modeling *real data* distributions, by not only explaining what modes of a factor of variation are present in the data, but also discovering their *respective proportions*.

2 Related Work

Disentangled representation learning Learning disentangled representations of the data has a vast literature (Hinton et al. (2011); Bengio et al. (2013); Yan et al. (2016); Chen et al. (2016); Mathieu et al. (2016); Tran et al. (2017); Denton & Birodkar (2017); Hu et al. (2018); Singh et al. (2019)). InfoGAN (Chen et al. (2016)) is one of the most popular unsupervised GAN based disentanglement methods, which learns disentanglement by maximizing the mutual information between the latent codes and generated images. It has shown promising results for discovering meaningful latent factors in *balanced* datasets like MNIST (LeCun (1998)), CelebA (Liu et al. (2015)), and SVHN (Netzer et al. (2011)). The recent method of JointVAE (Dupont (2018)) extends beta-VAE (Higgins et al. (2017)) by jointly modeling both continuous and discrete factors, using Gumbel-Softmax sampling. However, both InfoGAN and JointVAE assume uniformly distributed data, and hence fail to be equally effective in imbalanced data, evident by Fig. 1 and our experiments. Our work proposes modifications to InfoGAN to enable it to discover meaningful latent factors in *imbalanced* data.

Learning from imbalanced data Real world data have a long-tailed distribution (Guo et al. (2016); Van Horn et al. (2018)), which can impede learning, since the model can get biased towards the dominant categories. To alleviate this issue, researchers have proposed re-sampling (Chawla et al. (2002); He et al. (2008); Shen et al. (2016); Buda et al. (2018); Zou et al. (2018)) and class re-weighting techniques (Ting (2000); Huang et al. (2016); Dong et al. (2017); Mahajan et al. (2018)) to oversample rare classes and down-weight dominant classes. These methods have shown to be effective for the *supervised* setting, in which the class distributions are known a priori. There are

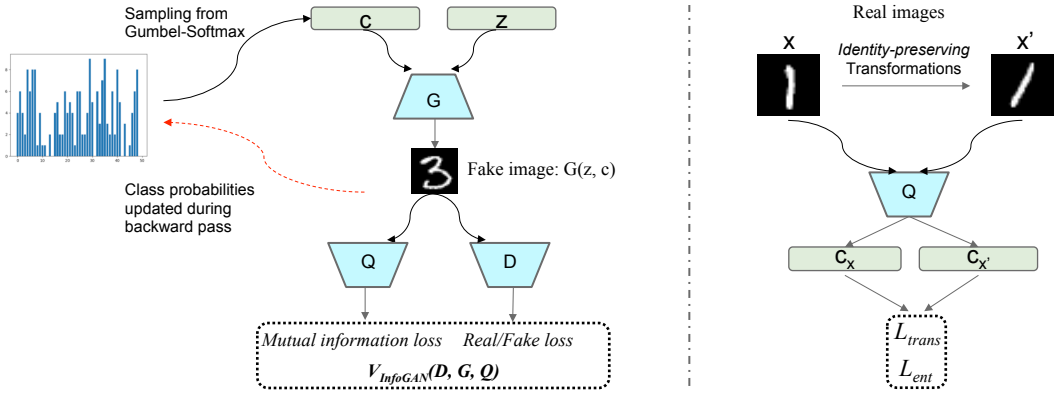

Figure 2: Elastic-InfoGAN takes a sampled categorical code from a Gumbel-Softmax distribution and a noise vector to generate fake samples. Apart from the original InfoGAN (Chen et al. (2016)) loss functions, we have two additional constraints: (1) We take real images x and create a transformed version x' using identity-preserving operations (e.g., small rotation), and force their inferred latent code distributions to be close; (2) We also constrain their entropy to be low. The use of differentiable latent variables from the Gumbel-Softmax enables gradients to flow back to the class probabilities to update them.

also unsupervised clustering methods that deal with imbalanced data in unknown class distributions (e.g., Nguwi & Cho (2010); You et al. (2018)). Our model works in the same *unsupervised* setting; however, unlike these methods, we propose an unsupervised *generative* model method that learns to disentangle latent categorical factors in imbalanced data.

Leveraging data augmentation for unsupervised image grouping Some works (Hui (2013); Dosovitskiy et al. (2015); Hu et al. (2017); Ji et al. (2019)) use data augmentation for image transformation invariant unsupervised clustering or representation learning. The main idea is to maximize the mutual information or similarity between the features of an image and its corresponding transformed image. However, unlike our approach, these methods do not target imbalanced data and do not perform generative modeling.

3 Approach

Let $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ be a dataset of N unlabeled images from k different classes. No knowledge about the nature of class imbalance is known beforehand. Our goal is twofold: (i) learn a generative model G which can learn to disentangle *object category* from other aspects (e.g., digits in MNIST (LeCun (1998)), face identity in YouTube-Faces (Wolf et al. (2011))); (ii) recover the unknown true class imbalance distribution via the generative modeling process. In the following, we first briefly discuss InfoGAN (Chen et al. (2016)), which addressed this problem for the balanced setting. We then explain how InfoGAN can be extended to the scenario of imbalanced data.

3.1 Background: InfoGAN

Learning disentangled representations using the GAN (Goodfellow et al. (2014)) framework was introduced in InfoGAN (Chen et al. (2016)). The intuition is for generated samples to retain the information about latent variables, and consequently for latent variables to gain control over certain aspects of the generated image. In this way, different types of latent variables (e.g., discrete categorical vs. continuous) can control properties like discrete (e.g., digit identity) or continuous (e.g., digit rotation) variations in the generated images.

Formally, InfoGAN does this by maximizing the mutual information between the latent code c and the generated samples $G(z, c)$, where $z \sim P_{noise}(z)$ and G is the generator network. The mutual information $I(c, G(c, z))$ can then be used as a regularizer in the standard GAN training objective. Computing $I(c, G(c, z))$ however, requires $P(c|x)$, which is intractable and hard to compute. The authors circumvent this by using a lower bound of $I(c, G(c, z))$, which can approximate $P(c|x)$ via a neural network based auxiliary distribution $Q(c|x)$. The training objective hence becomes:

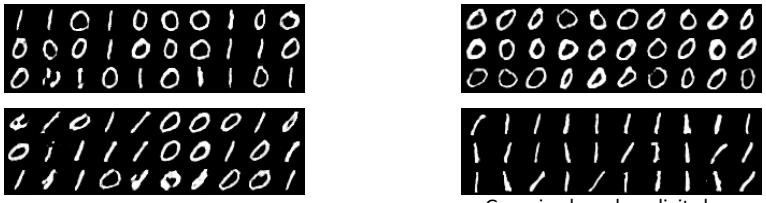

Figure 3: Different ways for unsupervised learning based methods to group unlabeled data; based on rotation (left) vs. digit identity (right). Here, we show two different groups for each grouping.

$$\min_{G,Q} \max_{D} V_{InfoGAN}(D, G, Q) = V_{GAN}(D, G) - \lambda_1 L_1(G, Q), \tag{1}$$

$$L_1(G, Q) = E_{c \sim P(c), x \sim G(z,c)}[\log Q(c|x)] + H(c), \tag{2}$$

where D is the discriminator network, and $H(c)$ is the entropy of the latent code distribution. Training with this objective results in latent codes c having control over the different factors of variation in the generated images $G(z, c)$. To model discrete variations in the data, InfoGAN employs non-differentiable samples from a uniform categorical distribution with fixed class probabilities; i.e., $c \sim Cat(K = k, p = 1/k)$ where k is the number of discrete categories to be discovered.

3.2 ELASTIC-INFOGAN

As shown in Fig. 1, applying InfoGAN to an imbalanced dataset results in suboptimal disentanglement, since the uniform prior assumption does not match the actual ground-truth data distribution of the discrete factor (e.g., digit identity). To address this, we propose two augmentations to InfoGAN. The first is to enable *learning* of the latent distribution's parameters (class probabilities), which requires gradients to be backpropagated through latent code samples c, and the second is to enforce *identity-preserving transformation* invariance in the learned latent variables so that the resulting disentanglement favors groups that coincide with object identities.

Learning the prior distribution To learn the prior distribution, we replace the fixed categorical distribution in InfoGAN with the Gumbel-Softmax distribution (Jang et al. (2017); Maddison et al. (2017)), which enables sampling of differentiable samples. The continuous Gumbel-Softmax distribution can be smoothly annealed into a categorical distribution. Specifically, if $p_1, p_2..., p_k$ are the class probabilities, then sampling of a k-dimensional vector c can be done in a differentiable way:

$$c_i = \frac{\exp((\log(p_i) + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\log(p_j) + g_j)/\tau)} \qquad \text{for } i = 1, ..., k. \tag{3}$$

Here g_i, g_j are samples drawn from $Gumbel(0, 1)$, and τ (softmax temperature) controls the degree to which samples from Gumbel-Softmax resemble the categorical distribution. Low values of τ make the samples possess properties close to that of a one-hot sample.

In theory, InfoGAN's behavior in the class balanced setting (Fig. 1 left) can be replicated in the imbalanced case (where grouping becomes incoherent, Fig. 1 center), by simply replacing the fixed uniform categorical distribution with Gumbel-Softmax with *learnable* class probabilities p_i's; i.e. gradients can flow back to update the class probabilities (which are uniformly initialized) to match the true class imbalance. And once the true imbalance gets reflected in the class probabilities, the *possibility* of proper categorical disentanglement (Fig. 1 right) becomes feasible.

Empirically, however, this ideal behavior is not observed in a consistent manner. As shown in Fig. 3 (left), unsupervised grouping can focus on non-categorical attributes such as rotation of the digit. Although this is one valid way to group unlabeled data, our goal in this work is to prefer groupings that correspond to *class identity* as in Fig. 3 (right).

Learning object identities To capture object identity as the factor of variation, we make another modification to InfoGAN. Specifically, to make the model focus on high level object identity and be invariant to low level factors like rotation, thickness, illumination, etc., we explicitly create these identity-preserving transformations on real images, and enforce the latent prediction $Q(c|x)$ to be invariant to these transformations. Note that such transformations (aka data augmentations) are standard for learning invariant representations for visual recognition tasks.

Formally, for any real image $x \sim P_{data}(x)$, we apply a set of transformations δ to obtain a transformed image $x' = \delta(x)$. It is important to point out that these transformations are not learned over the optimization process. Instead we use fixed simple transformations which guarantee that the human defined object identity label for the original image x and the transformed image x' image remain the same. For example, the digit identity of a 'one' from MNIST will remain the same if a transformation of rotation ($\pm$10 degree) is applied. Similarly, a face identity will remain the same upon horizontal flipping. We hence formulate our transformation constraint loss function:

$$L_{trans}(Q) = \mathsf{d}(Q(c_x|x), Q(c_{x'}|x')) \tag{4}$$

where $\mathsf{d}(\cdot)$ is a distance metric (e.g., cosine distance), and $Q(c_x|x)$, $Q(c_{x'}|x')$, are the latent code predictions for real image x and transformed image x', respectively. Note that ideally $Q(c|x)$, for either $x \sim P_{data}(x)$ or $x \sim P_g(G)$, should have low entropy (peaky class distribution) for proper inference about the latent object category. Eq. 2 automatically enforces a peaky class distribution for $Q(c|x)$ for $x \sim P_g(G)$, because the sampled input latent code c from Gumbel-Softmax is peaky. For $x \sim P_{data}(x)$ though, Eq. 4 alone isn't sufficient as it can be optimized in a sub-optimal manner (e.g., if $c_x \approx c_{x'}$, but both have high entropy). We hence add an additional entropy loss which forces c_x and $c_{x'}$ to have low entropy (s) class distributions:

$$L_{ent}(Q) = \mathsf{s}(Q(c_x|x)) + \mathsf{s}(Q(c_{x'}|x')). \tag{5}$$

The losses L_{trans} and L_{ent}, along with Gumble-Softmax, constitute our overall training objective:

$$\min_{G,Q} \max_{D} L_{final} = V_{InfoGAN}(D, G, Q) + \lambda_2 L_{trans}(Q) + \lambda_3 L_{ent}(Q). \tag{6}$$

$V_{InfoGAN}$ plays the role of generating realistic images and associating the latent variables to correspond to *some* factor of variation in the data, while the addition of L_{trans} will push the discovered factor of variation to be close to object identity. Finally, L_{ent}'s objective is to ensure Q behaves similarly for real and fake image distributions. The latent codes sampled from Gumbel-softmax, generated fake images, and losses operating on fake images are all functions of class probabilities p_i's too. Thus, during the minimization phase of Eqn. 6, the gradients are used to optimize the class probabilities along with G and Q in the backward pass.

4 EXPERIMENTS

In this section, we perform quantitative and qualitative analyses to demonstrate the advantage of Elastic-InfoGAN in discovering categorical disentanglement for imbalanced datasets.

4.1 DATASETS

We use: (1) MNIST (LeCun (1998)) and (2) YouTube-Faces (Wolf et al. (2011)). MNIST is by default a balanced dataset with 70k images, with a similar number of training samples for each of 10 classes. We artificially introduce imbalance over 50 random splits (max imbalance ratio 10:1 between the largest and smallest class). YouTube-Faces is a real world imbalanced video dataset with varying number of training samples (frames) for the 40 face identity classes (as used in Shah & Koltun (2018)). The smallest/largest class has 53/695 images, with a total of 10,066 tightly-cropped face images. All results are reported over the average of: (i) 50 runs (over 50 random imbalances) for MNIST, (ii) 5 runs over the same imbalanced dataset for YouTube-Faces.[1]

We use MNIST to provide a proof-of-concept of our approach. For example, one of the ways in which different 'ones' in MNIST vary is rotation, which can be used as a factor (as opposed to object identity) to group data in imbalanced cases (recall Fig. 3 left). Thus, using rotation as a transformation in L_{trans} should alleviate this problem. We ultimately care most about the YouTube-Faces results since it is more representative of real world data, both in terms of challenging visual variations (e.g., facial pose, scale, expression, and lighting changes) a well as inherent class imbalance. For this reason, the effect of augmentations in L_{trans} will be more reflective of how well our model can work in real world data.

[1]The imbalance statistics for all datasets are provided in the appendix.

4.2 BASELINES AND EVALUATION METRICS

We design different baselines to show the importance of having learnable priors for different latent variables and applying our transformation constraints.

- *Uniform InfoGAN* (Chen et al. (2016)): This is the original InfoGAN with fixed and uniform categorical distribution.
- *Ground-truth InfoGAN*: This is InfoGAN with a fixed, but imbalanced categorical distribution where the class probabilities reflect the ground-truth class imbalance.
- *Ground-truth InfoGAN + Transformation constraint*: Similar to the previous baseline but with our data transformation constraint (L_{trans}).
- *Gumbel-softmax*: In this case, InfoGAN does not have a fixed prior for the latent variables. Instead, the priors are learned using the Gumbel-softmax technique (Jang et al. (2017)).
- *Gumbel-softmax + Transformation constraint*: Apart from having a learnable prior we also apply our transformation constraint (L_{trans}). This is a variant of our final approach that does not have the entropy loss (L_{ent}).
- *Gumbel-softmax + Transformation constraint + Entropy Loss (Elastic-InfoGAN)*: This is our final model with all the losses, L_{trans} and L_{ent}, in addition to $V_{InfoGAN}(D, G, Q)$.
- *JointVAE* (Dupont (2018)): We also include this VAE based baseline, which performs joint modeling of disentangled discrete and continuous factors.

Our evaluation should capture: (1) how well we learn class-specific disentanglement for the imbalanced dataset, and (2) recover the ground-truth class distribution of the imbalanced dataset. To capture these aspects, we apply three evaluation metrics:

- *Average Entropy (ENT)*: Evaluates two properties: (i) whether the images generated for a given categorical code belong to the same ground-truth class i.e., whether the ground-truth class histogram for images generated for each categorical code has a low entropy; (ii) whether each ground-truth class is associated with a single unique categorical code. We generate 1000 images for each of the k latent categorical codes, compute class histograms using a pre-trained classifier[2] to get a $k \times k$ matrix (where rows index latent categories and columns index ground-truth categories). We report the average entropy across the rows (tests (i)) and columns (tests (ii)).
- *Normalized Mutual Information (NMI)* (Xu et al. (2003)): We treat our latent category assignments of the fake images (we generate 1000 fake images for each categorical code) as one clustering, and the category assignments of the fake images by the pre-trained classifier as another clustering. NMI measures the correlation between the two clusterings. The value of NMI will vary between 0 to 1; higher the NMI, stronger the correlation.
- *Root Mean Square Error (RMSE)* between predicted and actual class distributions: measures the accuracy of approximating the true class distribution of the imbalanced dataset. Since the learned latent distribution may not be aligned to the ground-truth distribution (e.g., the first dimension for the learned distribution might capture 9's in MNIST whereas the first dimension for the ground-truth distribution may be for 0's), we need a way to align the two. For this, we use the pre-trained classifier to classify the generated images for a latent variable and assign the variable to the most frequent class. If more than one latent variable is assigned to the same class, then their priors are added before computing its distance with the known prior of the ground-truth class.

4.3 IMPLEMENTATION DETAILS

Transformations (δ) used: (i) MNIST: Rotation (± 10 deg) + Zoom ($\pm 0.1\times$); (ii) YouTube-Faces: Random flipping + Random cropping (scale image by $1.1\times$ and crop 64×64 patch) + Gamma contrast (gamma $\sim U(0.3, 4.0)$). Additional details are in Appendix.

4.4 QUANTITATIVE EVALUATION

We first evaluate disentanglement quality as measured by NMI and average entropy (ENT); see Table 1. Elastic-InfoGAN consistently outperforms InfoGAN, JointVAE, and other baselines. In

[2]We train the classifier by creating a split of training/validation (80/20) on a per class basis. Classification accuracies: (i) MNIST - 98%, (ii) YoutTube-Faces - 96%. See appendix for details.

	MNIST		YouTube-Faces	
	NMI	ENT	NMI	ENT
JointVAE	0.6801	0.7006	0.4384	1.7203
Uniform InfoGAN	0.7765	0.4569	0.6729	1.0299
Ground-truth InfoGAN	0.7827	0.4196	0.6832	0.9577
Ground-truth InfoGAN + Transformation constraint	0.7926	0.3965	0.7349	0.8392
Gumbel-softmax	0.8360	0.3260	0.7704	0.7561
Gumbel-softmax + Transformation constraint	0.8678	0.2585	0.7572	**0.7229**
Elastic-InfoGAN (**Ours**)	**0.8778**	**0.2348**	**0.7768**	0.7240

Table 1: Distentanglement quality, measured by NMI (higher is better) and ENT (lower is better). Elastic-InfoGAN outperforms the baselines for both datasets. This shows that it learns a better disentangled representation which aligns with the ground-truth categories. Learning the prior with the transformation constraint (Ours) results in the best performance, showing their complementarity.

	MNIST	YouTube-Faces
Gumbel-softmax	0.03207	0.02118
Gumbel-softmax + Transformation constraint	0.03283	0.01732
Elastic-InfoGAN (**Ours**)	**0.02699**	**0.01552**

Table 2: Root Mean Square Error (RMSE) between the learned class distribution and ground-truth class distribution. Lower is better. See text for details.

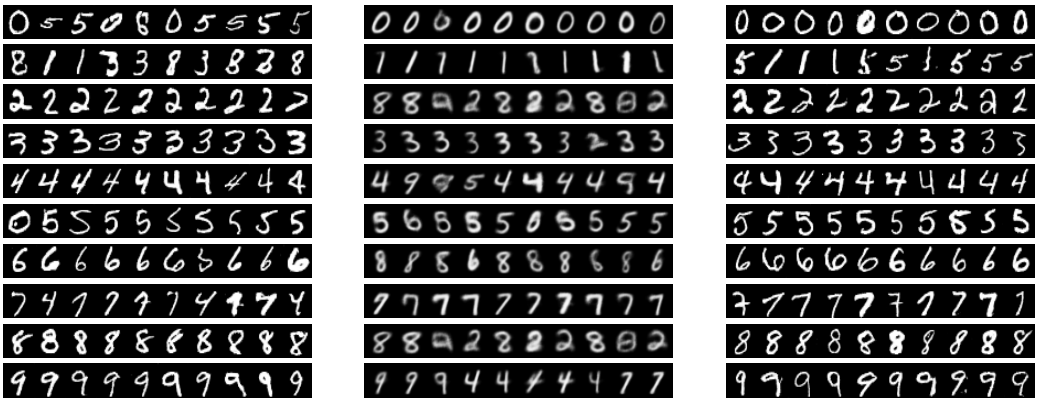

Figure 4: Representative image generations on a random imbalanced MNIST split. Each row corresponds to a learned latent variable. Our approach generates inconsistent images in only row 2 whereas Uniform InfoGAN does so in rows 1,2,6,8 and JointVAE does so in rows 3,5,6,7,9,10.

particular, our full model obtains significant boosts of 0.101 and 0.104 in NMI, and -0.222 and -0.305 in ENT compared to the Uniform InfoGAN baseline for MNIST and YouTube-Faces, respectively. The boost is even more significant when compared to JointVAE: 0.1977, 0.3380 in NMI, and -0.4658, -0.9963 in ENT for MNIST and YouTube-Faces, respectively. This again is a result of the assumption of a uniform categorical prior by JointVAE, along with poorer quality generations. We see that our transformation constraint generally improves the performance for both when the ground-truth prior is known (Ground-truth InfoGAN vs. Ground-truth InfoGAN + Transformation constraint) as well as when the prior is learned (Gumbel-softmax vs. Gumbel-softmax + Transformation constraint). This shows that enforcing the network to learn groupings that are invariant to identity-preserving transformations helps it to learn a disentangled representation in which the latent dimensions correspond more closely to identity-based classes.

Also, learning the prior using the Gumbel-softmax leads to better categorical disentanglement than fixed uniform priors, which demonstrates the importance of learning the prior distribution in imbalanced data. Overall, our approach using Gumbel-softmax to learn the latent prior distribution together with our transformation constraint works better than applying them individually, which demonstrates their complementarity. Interestingly, using a fixed ground-truth prior (Ground-truth InfoGAN) does not result in better disentanglement than learning the prior (Gumbel-softmax). This requires further investigation, but we hypothesis that having a rigid prior makes optimization more difficult compared to allowing the network to converge to a distribution on its own, as there are multiple losses that need to be simultaneously optimized.

Finally, in Table 2, we evaluate how well the Gumbel-softmax can recover the ground-truth prior distribution. For this, we compute the RMSE between the learned prior distribution and ground-

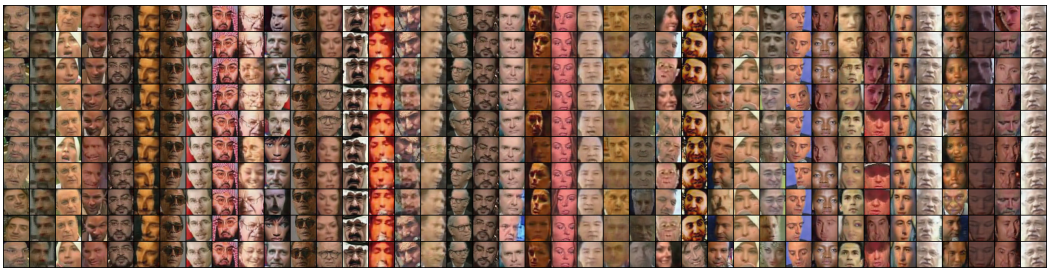

Figure 5: Elastic-InfoGAN image generations on YouTube-Faces. Each column corresponds to a latent variable. Although there are a few redundant latent variables (e.g., last and 5th last columns) or latent variables with multiple identities (e.g., 13th column), in general each latent variable corresponds to a unique identity with diverse variations in pose, translation, and scale. See Fig. 7 in Appendix for Uniform InfoGAN and JointVAE results.

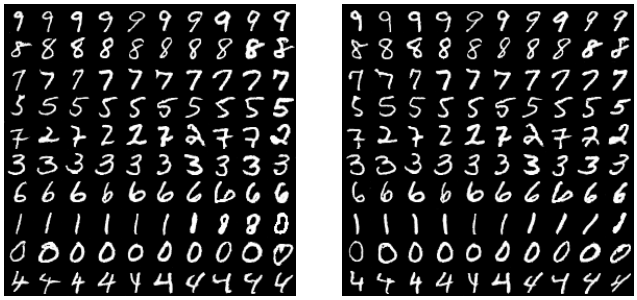

Figure 6: Uniform interpolation of two continuous latent codes between [-1, 1]: r1 varies in the left, while r2 varies in the right. The captured factors appear to be *stroke width* by r1, and *rotation* by r2.

truth prior distribution. Our full model (transformation constraint + entropy loss) produces the best estimate of the true class imbalance for both datasets, as evident through lowest RMSE. Our improvement over the Gumbel-Softmax baseline indicates the importance of our tranformation L_{trans} and entropy L_{ent} losses in approximating the class imbalance.

4.5 Qualitative Evaluation

We next qualitatively evaluate the disentanglement achieved by our approach. Figs. 4, 5, and 7 show results for MNIST and YouTube-Faces. Overall, Elastic-InfoGAN generates more consistent images for each latent code compared to Uniform InfoGAN and JointVAE. For example, in Fig. 4, Elastic-InfoGAN only generates inconsistent images in the second row whereas the baseline approaches generate inconsistent images in several rows. Similarly, in Fig. 7, Elastic-InfoGAN generates faces of the same person corresponding to a latent variable more consistently than the baselines. Both Uniform InfoGAN and JointVAE on the other hand tend to mix up identities within the same categorical code because they incorrectly assume a prior uniform distribution.

4.6 Modeling continuous factors

Finally, we demonstrate that Elastic-InfoGAN does not impede modeling of continuous factors in the imbalanced setting. Specifically, one can augment the input with continuous latent codes (e.g. r1, r2 $\sim$ Unif(-1, 1)) along with the existing categorical and noise vectors. In Fig. 6, we show the results of continuous code interpolation; we can see that each of the two continuous codes largely captures a particular continuous factor (stroke width on left, and digit rotation on the right).

5 Conclusion

In this work, we proposed a new unsupervised generative model that learns categorical disentanglement in imbalanced data. Our model learns the class distribution of the imbalanced data and enforces invariance to be learned in the discrete latent variables. Our results demonstrate superior performance over alternative baselines. We hope this work will motivate other researchers to pursue this interesting research direction in generative modeling of imbalanced data.

REFERENCES

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 1985.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *TPAMI*, 2013.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *JMLR*, 2003.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 2018.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *JAIR*, 2002.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*, 2016.

Emily L Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *NeurIPS*, 2017.

Qi Dong, Shaogang Gong, and Xiatian Zhu. Class rectification hard mining for imbalanced deep learning. In *ICCV*, 2017.

Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. In *TPAMI*, 2015.

Emilien Dupont. Learning disentangled joint continuous and discrete representations. In *NeurIPS*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *IJCNN*, 2008.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. Transforming auto-encoders. In *ICANN*, 2011.

Qiyang Hu, Attila Szab, Tiziano Portenier, Paolo Favaro, and Matthias Zwicker. Disentangling factors of variation by mixing them. In *CVPR*, 2018.

Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *ICML*, 2017.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, 2016.

Kayu Hui. Direct modeling of complex invariances for visual object features. In *ICML*, 2013.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *ICLR*, 2017.

Xu Ji, Joo F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2014.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*, 2017.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018.

Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *NeurIPS*, 2016.

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

Yok-Yen Nguwi and Siu-Yeung Cho. An unsupervised self-organizing learning with support vector ranking for imbalanced datasets. *Expert Systems with Applications*, 2010.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016.

Sohil Atul Shah and Vladlen Koltun. Deep continuous clustering. In *arXiv*, 2018.

Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *ECCV*, 2016.

Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *CVPR*, 2019.

Kai Ming Ting. A comparative study of cost-sensitive boosting algorithms. In *ICML*, 2000.

Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018.

Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011.

Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *SIGIR*, 2003.

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *ECCV*, 2016.

Chong You, Chi Li, Daniel P Robinson, and Rene Vidal. Scalable exemplar-based subspace clustering on class-imbalanced data. In *ECCV*, 2018.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *TPAMI*, 2018.

Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Domain adaptation for semantic segmentation via class-balanced self-training. *ECCV*, 2018.

A APPENDIX

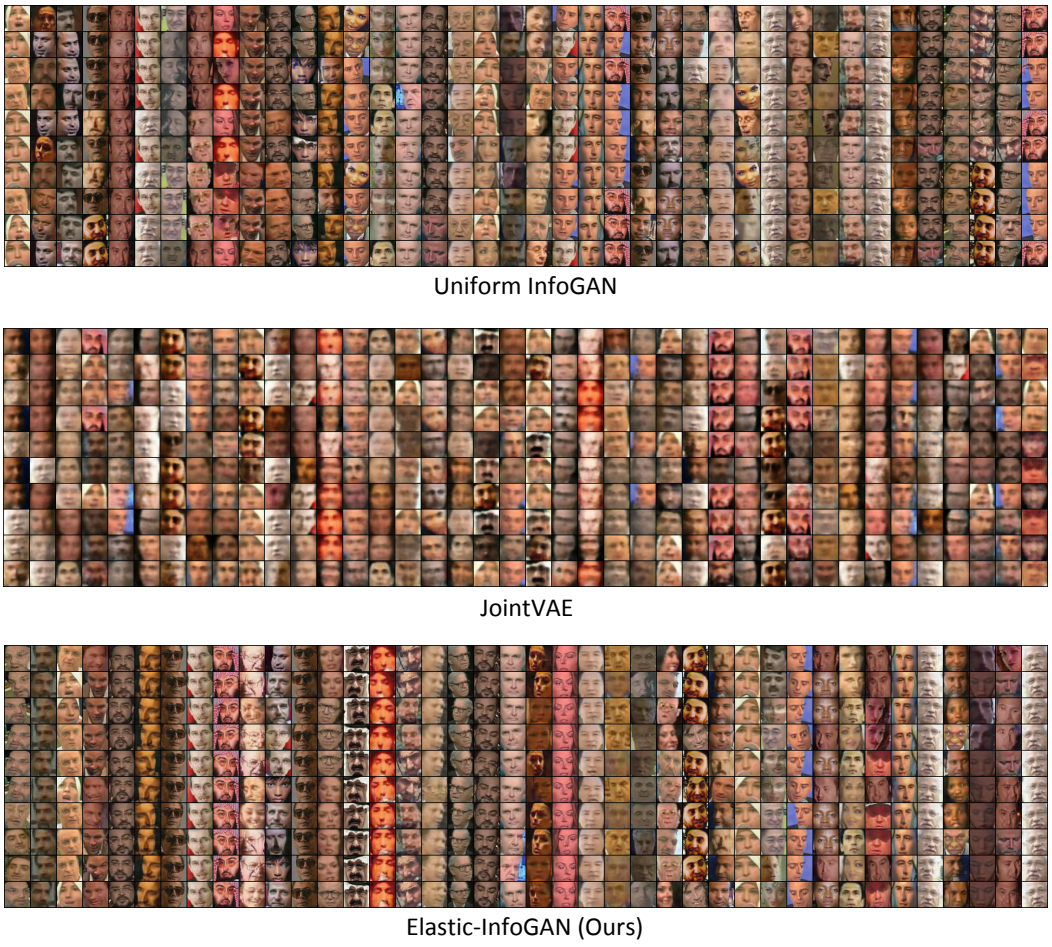

Figure 7: Image generations on YouTube-Faces. Each column corresponds to a latent variable. Overall, our approach generates images belonging to the same person more consistently compared to the baselines.

A.1 IMPLEMENTATION DETAILS (CONTINUED)

For MNIST, we operate on the original 28x28 image size, with 10-dimensional categorical code to represent 10 digit categories. For YouTube-Faces, we crop the faces using bounding box annotations provided, and then resize them to 64x64 resolution, and use a 40-dimensional categorical code to represent 40 face identities (first 40 categories sorted in alphabetical manner), as done in Shah & Koltun (2018). Pre-trained classification architecture used for evaluation for MNIST: 2 Conv + 2 FC layers, with max pool and ReLU after every convolutional layer. For YouTube-Faces classification, we fine-tune a ResNet-50 network pretrained on VGGFace2, for face recognition. We set $\lambda_1 = 1$ (for L_1), $\lambda_2 = 10$ (for L_{trans}), and $\lambda_3 = 1$ (for L_{ent}). These hyperparameters were chosen to balance the magnitude of the different loss terms. Finally, one behavior we observe is that if the random initialization of class probabilities is too skewed (only few classes have high probability values), then it becomes very difficult for them to get optimized to the ideal state. We hence initialize them with the uniform distribution, which makes training much more stable.

Elastic-InfoGAN architecture for MNIST: We follow the exact architecture as described in InfoGAN (Chen et al. (2016)): The generator network G takes as input a 64 dimensional noise vector $z \sim \mathcal{N}(0, 1)$ and 10 dimensional samples from Gumbel-Softmax distribution. The discriminator D

and the latent code prediction network Q share most of the layers except the final fully connected layers.

Elastic-InfoGAN architecture for YouTube Faces We operate on cropped face images resized to 64x64 resolution. Our architecture is based on the one proposed in StackGANv2 (Zhang et al. (2018)), where we use its 2-stage version for generating 64x64 resolution images. The input is a 100 dimensional noise vector $z \sim \mathcal{N}(0, 1)$ and 40 dimensional samples (c) from the Gumbel-Softmax distribution. There is an initial fully connected layer which maps the input (concatenation of z and c) to an intermediate feature representation. A series of a combination of upsampling + convolutional (interleaved with batch normalization and Gated Linear Units) increase the spatial resolution of the feature representation, starting from 1024 (feature size: 4 x 4 x 1024) channels to 64 (feature size: 64 x 64 x 64) channels. For the first stage, a convolutional network transforms the feature representation into a 3 channel output, while maintaining the spatial resolution; this serves as the fake image from the first stage. The next stage uses the 64 x 64 x 64 resolution features, forwards it through a network containing residual blocks and convolutional layers, while again maintaining the spatial resolution of 64 x 64. For the second stage, again a convolutional layer maps the resulting feature into a 64 x 64 resolution fake image, which is the one used by the model for evaluation purposes. The discriminator networks are identical at both stages. It consists of 4 convolutional layers interleaved with batch normalization and leaky ReLU layers, which serve as the common layers for both the D and Q networks. After that, D has one non-shared convolutional layer which maps the feature representation into a scalar value reflecting the real/fake score. For Q, we have a pair of non-shared convolutional layers which map the feature representation into a 40 dimensional latent code prediction.

Training of Elastic-InfoGAN We employ a similar way of training the generative and discriminative modules as described in Chen et al. (2016). We first update the discriminator based on the real/fake adversarial loss. In the next step, after computing the remaining losses (mutual information + L_{trans} + L_{ent}), we update the generator (G) + latent code predictor (Q) + latent distribution parameters at once. Our optimization process alternates between these two phases. For MNIST, we train all baselines for 200 epochs, with a batch size of 64. For YouTube-Faces, we train until convergence, as measured via qualitative realism of the generated images. We use a batch size of 50. $\tau = 0.1$ when used for sampling from Gumbel-Softmax, which results in samples having very low entropy (very close to one hot vectors from a categorical distribution).

A.2 Ground truth class imbalance

Here we describe the exact class imbalance used in our experiments. For MNIST, we include below the 50 random imbalances created. For YouTube-Faces, we include the true ground truth class imbalance in the first 40 categories. The imbalances reflect the class frequency.

A.2.1 MNIST

- 0.147, 0.037, 0.033, 0.143, 0.136, 0.114, 0.057, 0.112, 0.143, 0.078
- 0.061, 0.152, 0.025, 0.19, 0.12, 0.036, 0.092, 0.185, 0.075, 0.064
- 0.173, 0.09, 0.109, 0.145, 0.056, 0.114, 0.075, 0.03, 0.093, 0.116
- 0.079, 0.061, 0.033, 0.139, 0.145, 0.135, 0.057, 0.062, 0.169, 0.121
- 0.053, 0.028, 0.111, 0.142, 0.13, 0.121, 0.107, 0.066, 0.125, 0.118
- 0.072, 0.148, 0.092, 0.081, 0.119, 0.172, 0.05, 0.109, 0.085, 0.073
- 0.084, 0.143, 0.07, 0.082, 0.059, 0.163, 0.156, 0.063, 0.074, 0.105
- 0.062, 0.073, 0.065, 0.183, 0.099, 0.08, 0.05, 0.16, 0.052, 0.177
- 0.139, 0.113, 0.074, 0.06, 0.068, 0.133, 0.142, 0.13, 0.112, 0.03
- 0.046, 0.128, 0.059, 0.112, 0.135, 0.164, 0.142, 0.125, 0.051, 0.037
- 0.107, 0.057, 0.154, 0.122, 0.05, 0.111, 0.032, 0.044, 0.136, 0.187
- 0.129, 0.1, 0.039, 0.112, 0.119, 0.095, 0.047, 0.14, 0.156, 0.064
- 0.146, 0.08, 0.06, 0.072, 0.051, 0.119, 0.176, 0.11, 0.158, 0.028

- 0.035, 0.051, 0.112, 0.143, 0.033, 0.165, 0.082, 0.165, 0.054, 0.161
- 0.041, 0.1, 0.073, 0.054, 0.155, 0.117, 0.091, 0.124, 0.142, 0.104
- 0.052, 0.139, 0.128, 0.133, 0.104, 0.107, 0.058, 0.137, 0.036, 0.107
- 0.055, 0.138, 0.059, 0.074, 0.08, 0.135, 0.085, 0.064, 0.172, 0.139
- 0.141, 0.156, 0.119, 0.062, 0.08, 0.022, 0.043, 0.159, 0.101, 0.118
- 0.11, 0.088, 0.033, 0.062, 0.089, 0.176, 0.161, 0.105, 0.144, 0.032
- 0.157, 0.111, 0.125, 0.099, 0.036, 0.119, 0.036, 0.05, 0.147, 0.121
- 0.119, 0.121, 0.117, 0.152, 0.026, 0.174, 0.027, 0.065, 0.151, 0.049
- 0.057, 0.07, 0.134, 0.118, 0.058, 0.185, 0.07, 0.13, 0.116, 0.063
- 0.102, 0.082, 0.135, 0.046, 0.128, 0.106, 0.116, 0.085, 0.133, 0.066
- 0.057, 0.193, 0.2, 0.123, 0.022, 0.154, 0.115, 0.025, 0.065, 0.047
- 0.056, 0.196, 0.168, 0.052, 0.116, 0.062, 0.099, 0.133, 0.065, 0.053
- 0.04, 0.022, 0.2, 0.194, 0.038, 0.033, 0.161, 0.097, 0.159, 0.056
- 0.04, 0.036, 0.119, 0.204, 0.16, 0.103, 0.089, 0.061, 0.136, 0.052
- 0.112, 0.189, 0.145, 0.163, 0.113, 0.031, 0.028, 0.062, 0.045, 0.112
- 0.071, 0.099, 0.113, 0.175, 0.082, 0.068, 0.03, 0.066, 0.133, 0.164
- 0.134, 0.074, 0.111, 0.091, 0.051, 0.119, 0.044, 0.085, 0.144, 0.148
- 0.103, 0.126, 0.084, 0.117, 0.084, 0.127, 0.131, 0.092, 0.117, 0.019
- 0.096, 0.121, 0.026, 0.046, 0.043, 0.124, 0.165, 0.04, 0.127, 0.213
- 0.117, 0.115, 0.125, 0.128, 0.081, 0.103, 0.073, 0.044, 0.137, 0.077
- 0.037, 0.021, 0.143, 0.165, 0.075, 0.111, 0.028, 0.132, 0.134, 0.154
- 0.154, 0.049, 0.128, 0.089, 0.082, 0.072, 0.034, 0.138, 0.108, 0.146
- 0.078, 0.141, 0.084, 0.139, 0.085, 0.062, 0.035, 0.174, 0.15, 0.053
- 0.112, 0.112, 0.128, 0.112, 0.107, 0.142, 0.032, 0.142, 0.063, 0.049
- 0.084, 0.091, 0.128, 0.129, 0.045, 0.105, 0.05, 0.091, 0.089, 0.188
- 0.062, 0.136, 0.112, 0.153, 0.091, 0.046, 0.089, 0.03, 0.161, 0.12
- 0.143, 0.1, 0.046, 0.166, 0.107, 0.191, 0.026, 0.078, 0.097, 0.047
- 0.077, 0.174, 0.05, 0.098, 0.028, 0.173, 0.067, 0.106, 0.096, 0.13
- 0.105, 0.022, 0.183, 0.056, 0.045, 0.103, 0.081, 0.135, 0.119, 0.149
- 0.083, 0.127, 0.126, 0.028, 0.209, 0.03, 0.066, 0.125, 0.1, 0.107
- 0.138, 0.142, 0.074, 0.091, 0.103, 0.067, 0.12, 0.04, 0.1, 0.124
- 0.058, 0.039, 0.088, 0.113, 0.093, 0.055, 0.162, 0.069, 0.168, 0.155
- 0.02, 0.162, 0.133, 0.138, 0.137, 0.051, 0.069, 0.032, 0.118, 0.14
- 0.071, 0.046, 0.134, 0.119, 0.159, 0.057, 0.039, 0.135, 0.057, 0.184

A.2.2 YOUTUBE-FACES

- 0.0189, 0.0131, 0.0242, 0.0201, 0.0284, 0.0225, 0.0526, 0.0103, 0.062, 0.0306, 0.0365, 0.0053, 0.0106, 0.027, 0.0339, 0.0333, 0.0091, 0.0063, 0.0115, 0.0162, 0.0236, 0.0466, 0.028, 0.069, 0.0119, 0.0063, 0.0241, 0.0053, 0.0064, 0.0241, 0.0053, 0.0375, 0.0277, 0.0562, 0.0594, 0.0258, 0.0082, 0.006, 0.0281, 0.0281

A.3 DISCUSSION ABOUT EVALUATING PREDICTED CLASS IMBALANCE IN SEC. 4.2

To measure the ability of a generative model to approximate the class imbalance present in the data, we derive a metric in Section 4.2 of the main paper, the results of which are presented in Table 2. Even though we do get better results as measured by RMSE between the approximated and the original imbalance distribution, we would like to discuss certain flaws associated with this metric.

In its current form, we compute the class histogram (using the pre-trained classifier, which classifies each fake image into one of the ground-truth categories) for a latent code and associate the latent code to the *most frequent* class. If multiple latent codes get associated to the same ground-truth class, there will be ground-truth classes for which the predicted class probability will be zero. This is rarely an issue for MNIST, as it only has 10 ground-truth classes, and thus in most cases both our method and the baselines assign each latent code to a unique ground-truth class. However, for YouTube-Faces, after associating latent codes to the ground truth categories in this manner, roughly 10-13 ground-truth classes (out of 40) get associated with 0 probability for both our approach and the baselines (due to multiple latent codes being associated to the same majority ground-truth class). Our metric therefore may be too strict, especially for difficult settings with many confusing ground-truth categories.

The tricky part about evaluating how well the model is approximating the class imbalance is that there are two key aspects that need to be simultaneously measured. Specifically, not only should (i) the raw probability values discovered match the ground-truth class imbalance distribution, but (ii) the class probabilities approximated by the latent codes must correspond to the correct ground-truth classes. For example, if the original data had 80% samples from class A and 20% from class B, the generative model should not only estimate the imbalance as 80%-20%, but the model must associate 80% to class A and 20% to class B (instead of 80% to class B and 20% to class A). Another way to evaluate whether a model is capturing the ground-truth class imbalance could be the FID score, but it's worth noting that a method can still have a good FID score without disentangling the different factors of variations.

Given the limitation with our metric on YouTube-Faces, we have also measured the min/max of predicted prior values. For YouTube-Faces, the min/max of predicted and ground-truth priors are: **Gumbel-Softmax**: Min 2.76748415e-05, Max: 0.0819286481; **Ours without** L_{ent}: Min 0.00211485, Max: 0.06152404; **Ours complete**: Min 0.00336615, Max: 0.06798439; and **Ground-Truth**: Min 0.005265, Max: 0.069044. Our full method's min/max more closely matches that of the ground-truth, and the overall ordering of the methods follows that of Table 2 using our RMSE based metric.

In sum, we have made an effort to evaluate accurate class imbalance prediction in multiple ways, but it is important to note that this is an area which calls for better metrics to evaluate the model's ability to approximate the class imbalance distribution.