2-D latent space models: Layer-wise perceptual training and spatial grounding

Dusan Grujicic*
Department of Electrical Engineering
KU Leuven, Belgium
dusan.grujicic@esat.kuleuven.be

Matthew Blaschko
Department of Electrical Engineering
KU Leuven, Belgium
matthew.blaschko@esat.kuleuven.be

***Abstract*—Deep autoencoder models work on an information bottleneck principle with a lower-dimensional latent space of typically tens or hundreds of dimensions. Two and three-dimensional spaces are key for representing information that can be mapped in physical space, or that can be intuitively navigated and interpreted by a human without extensive background knowledge. Nevertheless, training autoencoder models with extremely low dimensional latent spaces is challenging for multiple reasons: (i) the dimensionality of the latent space is lower than the intrinsic dimensionality of the data manifold, and (ii) optimization of a complicated non-convex objective can lead to convergence to a non-global optimum. In this work, we demonstrate that layer-wise training strategies lead to improved convergence, as well as better perceptual properties when applied to models with extremely low dimensional latent spaces. Experiments on the CelebA dataset show improved performance on the Fréchet Inception Distance over standard autoencoder training. Additionally, we demonstrate the utility of low dimensional, physical latent representations. We map satellite image patches to a low dimensional latent space where we align the representation with the physical attributes of each patch, such as the geographic coordinates. We show that such representations are inherently interpretable and allow for an interactive and physically intuitive approach to generating new images.**

I. Introduction

In some respects, generative models of images via encoder-decoder architectures have become mature technologies. Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and other variants can generate samples from diverse image categories that are visually compelling and show interesting modes of variation corresponding to novel poses and appearance. Such models operate on an information bottleneck principle: original image data are projected into a lower-dimensional latent space capturing the main sources of variation in the data in a compact representation. Often, the dimensionality of this latent space is a parameter determined in a model selection procedure, while optimizing performance metrics of the image generation. The dimensionality of the generator or decoder input reported in the literature is typically in the range of 100 dimensions or more [1]–[7]. However, there are compelling reasons why we might want to develop models with latent dimensions of two or three, with use cases spanning a variety of settings. If a low-enough dimensionality of the latent space can be achieved in a general setting, a small number of known features (e.g. image labels, geo-coordinates etc.) can be used to encode the images in the new space of these features. From a human interaction perspective, these are spaces that we can navigate, visualize, and develop interfaces through which we can interact with such models. From a spatial perspective, we can develop *spatially grounded* models that e.g. align latent space parameters with physical layouts in two or four dimensions [8]–[10]. For example, for geotagged images, we could perform an additional mapping from an low dimensional latent space onto the 2 dimensional geo-coordinates of images, and subsequently use the geo-coordinates for image generation. The work of [11], for example, showcases a model for facial generation conditioned on geographic latitude and longitude, in addition to other conditioning variables such as age, gender, pose and facial landmark points. However, the dimensionality of the encoding is much larger than the one considered in this work. On the other hand, we analyze the problem of training generative models with very low-dimensional latent spaces, with the cardinality of two to four.

Generative models are trained to optimize various criteria depending on the purpose for which they are deployed. Optimization objectives can be designed around: (i) expected image reconstruction error, (ii) perceptual appearance scores, or (iii) distributional approximation of the original data distribution. Expected image reconstruction error is most commonly achieved with a mean squared error penalty and is the most common loss term incorporated in training. Another candidate is the mean absolute error penalty, which alleviates the common issue of excessive blurring that arises when using the mean squared error [3]. However, there exist other differentiable measures of reconstruction quality that are more in line with the human perception, such as the Structural Similarity Index (SSIM) [12] and Multi-Scale Structural Similarity Index (MS-SSIM) [13], which can be used as a measure of reconstruction error and optimized, as in the work of [14]. Additional perceptual appearance scores, such as Fréchet Inception Distance (FID) [15]–[17] are often used during evaluation and provide a measurable quantitative approximation to perceptual quality. Finally, measures of distribution shift,

*Corresponding author

This project was supported by Flanders AI and the MACCHINA project from KU Leuven (grant number C14/18/065).

such as divergence measures or Maximum Mean Discrepancy (MMD) [18], are important when we wish to measure whether a generative model covers the input space. We aim to measure the performance of several training strategies on each of these categories of metrics, in order to gauge their suitability in different settings where generative models with low dimensional latent space are of interest.

The information bottleneck principle, in its instantiation in encoder-decoder models, is built around the idea that the important modes of variation in image data are contained in a low dimensional manifold [19], [20]. In training the encoder-decoder architecture, the data are mapped into the latent space in a manner where it can be reconstructed. Thus, the redundant information common across images is implicitly encoded in the weights of the decoder. The goal of learning in this setting is primarily centered around obtaining a model that achieves minimal reconstruction loss. By contrast, when training a model that projects to two or three dimensions, the latent dimension is determined by extrinsic considerations and in general, the dimensionality reduction will result in a non-trivial reduction in reconstruction accuracy. As such, the distinction between competing objectives of (i) image reconstruction error, (ii) perceptual appearance scores, and (iii) distributional approximation becomes more acute. We find that in this setting, it is important to adapt training to overcome optimization difficulties and emphasize performance on perceptual metrics, rather than just the reconstruction error.

In this work, we explore variants of layer-wise training [21], [22] of generative models, where the model is iteratively trained one layer at a time, after which new layers are added to the network and the process repeated. As per the findings of [17], a two-stage training approach to training a VAE, where the second stage VAE uses the latent representations of the first stage as the observed data, can yield improved results compared to a single-stage training, as long as the latent space is large enough to accommodate the underlying manifold. It is also emphasized that in this regime, the joint training of two stages does not necessarily lead to better performance. We, however, show that in the case of an extremely low latent space dimensionality, layer-wise training improves performance over a standard joint training strategy with respect to the perceptual metric used for the evaluation of deep generative models. In particular, non-greedy layer-wise training, in which the previous layers are further tuned after adding new layers, achieves a performance increase with respect to reconstruction error over the joint training approach. This indicates that even in cases when the dimensionality of the latent space becomes increasingly constrained, non-convexities in the training landscape lead to problems in optimization in a joint framework, making layer-wise training an effective optimization heuristic.

Considering perceptual quality, it appears that an overconstrained latent space leads to excessive blurring as a strategy to minimize the reconstruction error. In this case, greedy layer-wise training, where no further tuning is performed after adding additional layers, can help maintain image sharpness. Moreover, we find that a two-stage t-SNE [23] preimage approach, where we decode the higher dimensional image feature that results in the closest low dimensional latent representation to that of the input image, can yield even better results in distribution approximation at the cost of substantially increased reconstruction error over layer-wise training. Finally, we demonstrate the utility of low dimensional latent spaces in an image colorization setting, in which alignment of the latent space to geographic and color information is performed without substantial loss in reconstruction quality. Source code and models will be made available at the time of publication.

A. Related Work

The work of [6] introduces the Variational AutoEncoder (VAE), which is a neural generative model, consisting of an encoder and decoder, trained by optimizing the variational lower bound on the marginal likelihood of the data, which consists of the reconstruction term and a KL divergence term. The KL divergence term enforces the posterior distribution of the underlying low dimensional representations to match the given prior distribution, which can be seen as a regularization term in the context of training an auto-encoder. Typically, a zero-mean multivariate isotropic Gaussian is used as the prior distribution, and the posterior distribution is assumed to be Gaussian and parametrized by a neural encoder, from which the data can be sampled by using the reparametrization trick [6]. In the work of [24], the authors demonstrate that a modification of the variational lower bound objective, where the KL divergence between the latent factors and the prior is constrained by an upper-bound, can lead to improved disentanglement of underlying generative factors. This constrained optimization problem is transformed into an unconstrained optimization problem with an additional Lagrange multiplier β. Increasing the value of β only allows the most informative latent units to deviate from the Gaussian prior, improving disentanglement. Furthermore, progressively relaxing the constraint on the KL divergence during training can help mitigate the reconstruction penalty, while maintaining disentanglement [25].

The work of [21], [26] demonstrates the advantages of greedy layer-wise training in discriminative models. In a generative setting, the work of [7] proposed a strategy of progressively growing a generative model by gradually increasing the output resolution as additional layers are added. The approach does not, however, progressively decrease the size of the latent space as we do, and does not result in extremely low dimensional latent spaces. Similarly, [27] generate high-resolution images by a multi-resolution Laplacian pyramid, but again do not consider such low dimensional latent spaces.

In the work of [17], a two-stage VAE, with two stages trained separately, is shown to outperform single-stage VAE in the regime where the latent space is large enough compared to the dimensionality the underlying data manifold. Unlike our work, however, it trains a 2 stage network only, while we explore several stages. Additionally, the latent space is large enough and assumed to be larger than the manifold, so the addition of new stages with lower dimensionality of the latent space improves performance. In our case, we explore

the situation where the latent space dimensionality is certainly lower than the presumed dimensionality of the data manifold.

II. Methods

We examine both joint and stage-wise training, while reducing the dimensionality of the latent space in each additional stage, in order to finally reach the desired, extremely low dimensional representation of the data. As per the work on the β-VAE [24], we train the model by optimizing the parameters of the encoder ϕ and those of the decoder θ, with the following objective:

$$\mathcal{L}(\theta, \phi, \beta; \mathbf{x}, \mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \\ -\beta \mathbf{D}_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \quad (1)$$

where $q_\phi(\mathbf{z}|\mathbf{x})$ is the posterior distribution of the latent representation z parameterized by the encoder, and $p_\theta(\mathbf{x}|\mathbf{z})$ the data likelihood parametrized by the decoder. In practice, the likelihood can be viewed as the the reconstruction term [6], while the second term represents the discrepancy between the distribution of the latent representations, which is set to be a multivariate Gaussian whose mean and covariance are the outputs of the encoder, and the zero-mean isotropic unit Gaussian adopted as the prior.

During stage-wise training, after training a VAE in the previous iteration, an additional linear layer that projects the latent embedding to an even lower dimensional latent space is appended to the encoder, and a layer that projects from a low dimensional latent space back to the higher dimensional latent space is prepended to the decoder:

$$\begin{aligned} enc_k(x) &= \sigma\Big(enc_{k-1}(x)W_k^T + b_k\Big), k > 0 \\ dec_k\left(enc_k(x)\right) &= dec_{k-1}\left(\sigma\left(enc_k(x)\tilde{W}_k^T + \tilde{b}_k\right)\right), k > 0 \end{aligned}, \quad (2)$$

where W_k and b_k are the weights and biases of the newly added $k-th$ layer of the encoder, while $\tilde{W}_k$ and $\tilde{b}_k$ represent the added layer of the decoder. For $k = 1$, enc_0 and dec_0 represent the initial, fully convolutional encoder and decoder. At the k-th stage, the optimal encoder weights and biases W^* and optimal decoder weights and biases $\tilde{W}^*$ are obtained as:

$$W^*, \tilde{W}^* = \min_{W,\tilde{W}} \sum_{i=1}^{n} \gamma_k \ell(x_i, dec_k(enc_k(x_i))) + \\ +\beta_k \mathbf{D}_{KL}\left(\mathcal{N}\left(enc_k(x_i)\right)||\mathcal{N}(0, I)\right), \quad (3)$$

where the first term is the reconstruction error and the second term is the KL divergence between the posterior distribution parametrized by the encoder (Gaussian with a mean and covariance matrix, or its diagonal elements, assuming independent dimensions, given by the encoder), and the prior defined as an multivariate zero-mean isotropic Gaussian [6].

In the greedy variant of the training, during the training of the k-th stage, $W = W_k$ and $\tilde{W} = \tilde{W}_k$, and we only optimize the parameters of the newly added layers, while the parameters of the previously added layers in both encoder and decoder are frozen, thus effectively treating the latent representation obtained in the previous training stage as the observed data in the current stage. In the non-greedy variant of the training, $W = \{W_1, ..., W_k\}$ and $\tilde{W} = \{\tilde{W}_1, ..., \tilde{W}_k\}$, i.e. when training the current stage, we also optimize the parameters of the previously trained layers.

III. Experiments

We evaluate the stage-wise training with several different numbers of stages. We also compare the stage-wise approach with joint training of all the layers at once. Additionally, we explore whether or not during stage-wise training, the previous stages should be jointly optimized with the current stage. We evaluate the results in terms of the reconstruction error, perceptual image quality and distributional similarity between the original and reconstructed images.

Additionally, we examine the effect of linear and non-linear mapping to a low dimensional latent space with two baselines, one involving the use of PCA to project the data from a higher dimensional latent space into a low dimensional latent space, and another one involving the use of parametric tSNE. We compare two different variants of each approach.

A. Datasets

We perform experiments on the Celeb-Faces Attributes (CelebA) [28] and the BigEarthNet dataset [29]. The CelebA dataset consists of 202,599 of faces, where each image is annotated with 40 attributes and 5 landmark locations. The training set consists of 160000 images of 8000 people, the valiadation set consists of 20000 images of 1000 people, while the remaining images make up the test set.

The BigEarthNet dataset [29] consists of 590,326 patches from 152 Sentinel-2 tiles, each characterized by a subset of 43 land cover classes, a timestamp that represents the time and date of acquisition, as well as the geographic coordinates of the patch center in Universal Transverse Mercator coordinate system (UTM), which we convert to latitude and longitude. We extract satellite image patches between the latitudes of 48.0082 and 48.4082 and longitudes of 16.1738 and 16.5738, which showcase Vienna and its immediate surroundings, and assign 2312 patches to the train set, 550 patches to the validation set, and the remaining 560 to the test set.

B. Metrics

a) SSIM: We utilize the **S**tructural **Sim**ilarity (SSIM) index [12] to measure the similarity between the original and the reconstructed images. The SSIM index is sensitive to the structural differences between images, of which the human visual system is highly perceptive. The metric compares the local luminance, contrast, and structure across corresponding image regions.

b) FID: We utilize the **F**réchet **I**nception **D**istance (FID) [16] to compute a semantic-based score for appearance. The output of the final pooling layer in a pretrained Inception v3 model is used to obtain image features, which have been found

to be sensitive to the perceptually important cues in images [30]. The mean and covariance of two sets of activations, where one is obtained from the collection of real, and the other from the collection of generated images define two multivariate Gaussian distributions. The Fréchet distance, or the Wasserstein-2 distance is then used to express the discrepancy between these two distributions [31]. It is found that this distance correlates very well with perceptual dissimilarity between images [16].

c) MMD: To evaluate the degree to which the distribution of generated images matches the distribution of the dataset, we compute the **M**aximum **M**ean **D**iscrepancy (MMD) [18], which is a non-parametric two-sample test used to evaluate the difference between distributions. It represents distances between distributions as distances between mean embeddings of features. Using the kernel trick, we avoid having to explicitly compute the embeddings, but evaluate the distances in the original input space using the Gaussian kernel, where the computed MMD is zero only if the distributions are identical. We compute the MMD on images resized to 64x64.

C. Experiment Framework

We use a ResNet-18 model [32] for the convolutional part of the initial encoder (enc_0 in Equation 2), while the initial convolutional part of the decoder (dec_0 in Equation 2) is also based on a ResNet-18 configuration, where we replace the strided convolutions that downsample the feature maps with nearest neighbor upsampling and non-strided convolutions, in addition to reversing the ordering the number of channels in the residual blocks of the decoder. We finally reverse the ordering of residual blocks altogether, resulting in a convolutional decoder that is symmetrical to the encoder. When training the VAE, in the reconstruction term, we optimize the SSIM, as opposed to the standard MSE or BCE minimization, as it is found that it yields images that are perceptually of higher quality [14]. Due to the standardized nature of the CelebA dataset, we do not consider the multi-scale variant of SSIM (MS-SSIM).

Due to its inherent interpretability, interactivity and the potential for a physical spatial interpretation, we opt for the latent dimensionality of 2. For the base model, denoted as Base-128 in Table I, we train a VAE with one 128 dimensional linear layer appended to the encoder, and another 128 dimensional linear layer prepended to the decoder, resulting the latent dimension of 128. We evaluate the reconstruction by computing the Structural SIMilarity index measure (SSIM), Fréchet Inception Distance (FID) and the Maximum Mean Discrepancy (MMD), which represents the upper bound on the model performance with a lower-dimensional latent space. We then evaluate the effect of the number of stages when performing stage-wise training of a VAE that yields a 2-dimensional data representation (results shown in Table I): (i) **2-stage** (G/NG): We train an autoencoder that performs a projection to a 2-dimensional latent space. The model is trained in two stages, where in each stage an additional linear layer is appended to the convolutional encoder and another layer prepended to the convolutional encoder. The layers added in the first and second stage have 128 and 2 neurons, respectivelly. We compare greedy (G) and non-greedy (NG) training, outlined in Section II; (ii) **3-stage** (G/NG): Same as 2-stage, but with 3 training stages and 3 additional layers (128, 32 and 2 neurons); (iii) **4-stage** (G/NG/D): Same as 2-stage, but with 4 training stages and 4 additional layers (128, 32, 8 and 2 neurons). The (D) indicates direct training of all layers at once, and where the number of training epochs is set to 4 times the number used when training an individual stage in a 4-stage configuration. We minimize the mean reconstruction error over the image pixels, and the mean KL divergence over individual latent dimensions, and normalize the value of β (Equation 1) according to the dimensionality of the input data and the latent space, as per [24].

We define two additional baselines built on top of the base VAE with the latent space size of 128. **PCA Baseline:** In the first baseline we utilize PCA [33] to perform a linear projection from the high dimensional latent space to a 2-dimensional space. We project the 128-dimensional codes onto the first two principal axes using the transformation matrix consisting the eigenvectors of the covariance matrix corresponding to the two highest eigenvalues. In the first variant of the PCA baseline, denoted as PCA (P), we reconstruct the latent code by mapping the PCA projections back to the 128-dimensional latent space using the transposed matrix of the chosen eigenvectors, which are then fed the decoder to reconstruct the input image. In the second variant of our approach, denoted as PCA (S), we reconstruct the input images by using a *shotgun* approach, where we compute the preimages for each 2-dimensional PCA representation by randomly sampling 1000 candidates from a zero-mean isotropic Gaussian representing the prior distribution for the 128-dimensional latent codes obtained from the base VAE. We then embed each candidate into a 2-dimensional latent space using the same PCA projection matrix, and decode the candidate whose 2-dimensional embedding has the lowest Euclidean distance to the 2-dimensional representation of the corresponding input sample.

Parametric tSNE: The second baseline is based on the parametric t-distributed stochastic neighbor embedding (Parametric tSNE) [23], [34], where a neural network model is trained to embed the 128-dimensional codes obtained from the VAE into a 2-dimensional space, by minimizing the KL divergence between the Gaussian distance metric in the 128 dimensional latent space and the Student's t-distributed distance metric in the target 2-dimensional space. This yields an embedding of the data in a 2-dimensional latent space, while the inverse of this mapping is found by solving the preimage problem. The preimages are computed via ridge regression from the 2-dimensional latent space back to the 128-dimensional latent space. This version of the tSNE baseline is denoted as tSNE (P). A second variant of the approach for computing the preimages is the similar *shotgun* procedure as described in the case of the *shotgun* variant of the PCA baseline, where the preimage is selected among 1000 candidates sampled from a zero-mean isotropic Gaussian in the

128 dimensional latent space, based on the proximity of its 2-dimensional parametric tSNE embedding to that of the input sample, and use it as an input to the decoder to reconstruct the original image. This version is denoted as tSNE (S).

-	FID	SSIM	MMD
Base - 128	122.1207 ± 0.4202	0.6332 ± 0.0007	0.0070
PCA (P)	196.8630 ± 0.4188	0.3964 ± 0.0006	0.2431
PCA (S)	124.3914 ± 0.4202	0.3023 ± 0.0006	0.0039
tSNE (P)	126.2563 ± 0.8308	0.2767 ± 0.0011	0.0044
tSNE (S)	124.0830 ± 0.4159	0.2752 ± 0.0006	0.0051
2-stage (NG)	208.0426 ± 0.4051	**0.4568 ± 0.0007**	**0.0452**
2-stage (G)	212.4302 ± 0.3575	0.4142 ± 0.0006	0.0724
3-stage (NG)	198.3123 ± 0.4314	**0.4566 ± 0.0007**	0.0493
3-stage (G)	191.5949 ± 0.4257	0.4070 ± 0.0006	0.0748
4-stage (NG)	201.8177 ± 0.4232	**0.4567 ± 0.0007**	0.0462
4-stage (G)	**185.0162 ± 0.4355**	0.4098 ± 0.0006	0.0701
4-stage (D)	217.9415 ± 0.3880	0.4525 ± 0.0007	0.0756

TABLE I

FIRST ROW REPRESENTS THE BASELINE VAE WITH A 128 DIMENSIONAL LATENT SPACE. THE FOLLOWING FOUR ROWS SHOW THE PCA AND TSNE BASED BASELINE, USING THE STANDARD PRE-IMAGE APPROACH (P) AND SHOTGUN APPROACH (S). THE FOLLOWING SIX ROWS SHOW A COMPARISON OF GREEDY (G), AND NONGREEDY (NG) LAYER-WISE TRAINING IN TERMS OF DIFFERENT TRAINING APPROACHES WITH DIFFERENT NUMBER OF STAGES. THE FINAL ROW SHOWS DIRECT (D), NON-LAYER-WISE TRAINING. CONFIDENCE INTERVALS REPRESENT ONE STANDARD ERROR.

As shown in Table I, in terms of the perceptual reconstruction quality measured by FID, the greedy 4-stage progressive training outperforms the stage-wise approach with lower numbers of stages. We therefore adopt the 4-stage approach. The performance of the 4-stage greedy (G) and non-greedy (NG) training is also compared to directly training all the layers at once (D), as well as the base model with a 128 dimensional latent space which represents the performance upper-bound (Base-128). It can be seen that the greedy training, where only the weights of the newly added stage are optimized achieves a better performance in terms of the FID metric, suggesting superior perceptual quality of the reconstructed images. The non-greedy stage-wise training, on the other hand, achieves a higher SSIM. This is expected, as the SSIM represents the reconstruction objective which is directly optimized during training, and the freezing of the layers from previous stages proves to hinder the model from the aspect of the reconstruction objective. The performance of the non-greedy stage-wise training with respect to SSIM is also significantly higher than that of the direct training of all layers at once, which suggests that the stage-wise approach in general leads to convergence to a better optimum. The non-greedy stage-wise also achieves lower MMD than joint training, suggesting that the stage-wise approach yields benefits from the aspect of distributional approximation of the original data.

We also observe in Table I that the shotgun-based approaches achieve a FID score comparable to the Base - 128 model that serves as their backbone. The SSIM on the other hand, is very low, as the chosen candidate preimage does not necessarily correspond to the 128-dimensional latent representation of the input image.

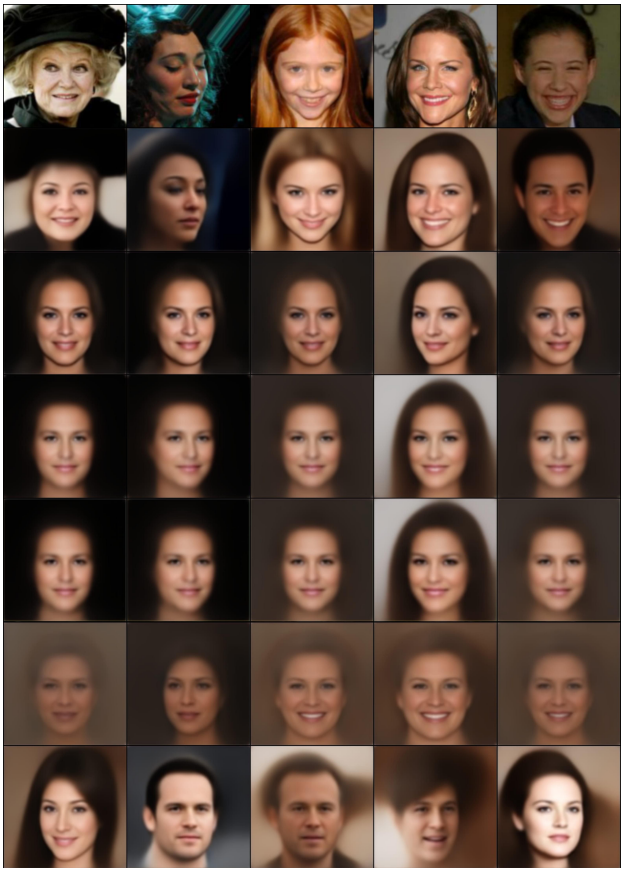

Fig. 1. Comparison of reconstructed images for the base model, 3 different training procedures and 2 baselines, by rows: (i) input images (ii) Base - 128 (iii) 4-stage (G) (iv) 4-stage (NG) (v) 4-stage (D) (vi) PCA (P) (vii) tSNE (S)

Overall, we notice that the performance on image reconstruction does not directly yield a better FID score. The greedy layer-wise training surpasses the performance of non-greedy layer-wise training and joint training of all latent layers in terms of the perceptual quality, despite having a lower SSIM score. Qualitatively, in Figure 1, it can be seen that out of all approaches that encode data into a two dimensional space, the greedy layer-wise training produces images with the highest degree of sharpness, while being able to capture the head pose and rudimentary facial expressions of the original input image. The non-greedy stage-wise training, as well as the non-stage wise training and the PCA baseline, produce exceedingly blurry images. The nonlinear tSNE-based baseline produces sharp images, however, the reconstruction quality is poor.

IV. SPATIALLY GROUNDED AUTOENCODER

The topic of interactive scene generation has received some attention in the literature [35]–[37]. To illustrate the utility of extremely low-dimensional image representations, we learn a low-dimensional representation of satellite patches that allows for the generation of new patches based on the given spatial coordinates. We perform the task of image colorization, where the grayscale satellite images are colorized in the way where

the color of each pixel is predicted based on its 15x15 neighborhood. Predicting the pixel's color, especially based on such a small field-of-view is an challenging problem. We examine the possibility of grounding image patches in a latent space that can be tied to the physical properties of each patch. We leverage two possible sources of information to obtain a physical grounding and allow interactive image generation.

In the first, we leverage the HLS color model [38], which consists of three components: hue, lightness and saturation. The lightness component is directly provided by the grayscale image, while the model is tasked with predicting the hue and saturation color components at each pixel. In the HLS color model, the saturation values lie in the $[0, 1 - |l|]$ range, where l represents the value of the lightness component. The hue component may take any value between 0 and 1. Therefore, for a lightness component value given by the input grayscale image, we minimize the 2-norm of the deviation of the two latent variables from the aforementioned ranges, enforcing the 2-dimensional latent space representations to correspond to valid hue-saturation components under a bi-conical HLS model, as shown in Figure 2.

In the second, we use the coordinates of the satellite image. We calculate the latitude and longitude of each pixel in each satellite image via interpolation, based on the physical size of the patch and the coordinates of its center. For any pixel neighborhood, we enforce that two variables of the latent representation match the latitude and longitude associated with the central pixels by minimizing the L2 distance between them.

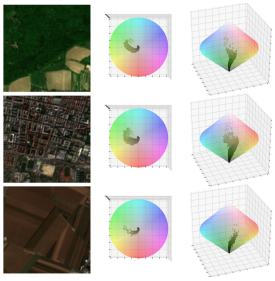

Fig. 2. Visualization of the latent representations corresponding to the hue and saturation color component, constrained to the volume of a bi-conical solid of a HLS color model.

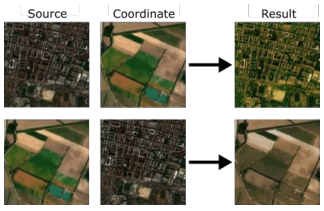

Fig. 3. Colorization based on a given coordinate location.

To efficiently obtain representations of patches around the each pixel in the image in a parallel fashion, we adopt a fully convolutional architecture. The first layer has a 15x15 dimensional kernel, while the subsequent layers perform 1x1 convolutions. Therefore, along each channel dimension at the output of each layer, we obtain a representation of each 15x15 patch in the original image. The 1x1 convolutions after the first layer, as well as the use of local-response normalization instead of batch normalization, make sure that each patch representation is only influenced by the 15x15 neighborhood of its original central pixel. We perform a non-greedy stage-wise training and optimize the mean absolute error between the predicted hue and saturation component and those of the ground truth color image. We evaluate the model performance in terms of the peak signal-to-noise ratio (PSNR). From the results in Table II, we observe that there is no drastic deterioration of the reconstruction error or PSNR with a further projection down to a 4 dimensional latent space. On the other hand, within a 4 dimensional latent space, we obtain a fully physically grounded representation of the data, which is interpretable and interactive. On average, the coordinate loss achieves a slightly better performance than the color loss.

-	MSE	PSNR
VAE 32	0.0861 $\pm$ 0.0018	32.1158 $\pm$ 0.2410
CR 2/VAE 2	0.0875 $\pm$ 0.0018	31.6968 $\pm$ 0.2252
CL 2/VAE 2	0.0895 $\pm$ 0.0019	31.5534 $\pm$ 0.2249
CR 2/CL 2	0.0905 $\pm$ 0.0020	31.2461 $\pm$ 0.2088

TABLE II
COMPARISON OF MSE AND PSNR FOR IMAGE COLORIZATION. CR 2 INDICATES THAT THERE ARE TWO LATENT DIMENSIONS THAT ENCODE A GEOGRAPHIC COORDINATE. CL 2 INDICATES THAT THERE ARE TWO LATENT DIMENSIONS ENCODING COLOR INFORMATION. VAE DENOTES LATENT VARIABLES CONSTRAINED SOLELY BY THE KLD TERM.

Given a model trained to embed the data in a physical, interpretable space of geographic coordinates, we demonstrate the colorization of input images based on a provided set of coordinates. In Figure 3, we demonstrate how, given the coordinates of another image, the model can generate different colored images from the grayscale version of the source image. In the same fashion, the coordinates can be provided interactively, by clicking on the locations on a map.

V. DISCUSSION AND CONCLUSIONS

With the CelebA dataset, the dimensionality of an embedding that can be physically interpretable is significantly lower than the intrinsic dimensionality of the data manifold. Therefore, such an embedding is significantly more challenging, and there is a deterioration of quality, both in terms of reconstruction error and perceptual quality and the distributional approximation, as we go to lower dimensions. Additionally, layer-wise training results in significantly higher perceptual quality, as well as better result with respect to distribution approximation. The colorization of image patches, on the other hand, is a problem where the intrinsic dimensionality of the underlying manifold is low and can be easily tied to color and geographic space. The color space can be aligned naturally to a low dimensional physical space, without significant deterioration of quality, allowing for an interpretable embedding and an interactive approach to image generation.

REFERENCES

[1] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *International Conference on Learning Representations*, Y. Bengio and Y. LeCun, Eds., 2016.

[2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[4] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014.

[6] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *International Conference on Learning Representations*, Y. Bengio and Y. LeCun, Eds., 2014.

[7] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.

[8] X. Deng, Y. Zhu, and S. Newsam, "What is it like down there? generating dense ground-level views and image features from overhead imagery using conditional generative adversarial networks," in *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2018, p. 43–52.

[9] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Oct. 2010, pp. 1277–1287.

[10] D. Grujicic, G. Radevski, T. Tuytelaars, and M. B. Blaschko, "Learning to ground medical text in a 3D human atlas," in *The SIGNLL Conference on Computational Natural Language Learning*, 2020.

[11] Z. Bessinger and N. Jacobs, "A generative model of worldwide facial appearance," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1569–1578.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[13] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[14] J. Snell, K. Ridgeway, R. Liao, B. D. Roads, M. C. Mozer, and R. S. Zemel, "Learning to generate images with perceptual similarity metrics," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 4277–4281.

[15] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training GANs," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, 2016.

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 6629–6640.

[17] B. Dai and D. Wipf, "Diagnosing and enhancing VAE models," in *International Conference on Learning Representations*, 2019.

[18] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[19] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[20] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Icml*, 2011.

[21] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, 2006.

[22] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 2007, pp. 1–8.

[23] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

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

[25] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE," *arXiv preprint arXiv:1804.03599*, 2018.

[26] E. Belilovsky, M. Eickenberg, and E. Oyallon, "Greedy layerwise learning can scale to imagenet," in *International conference on machine learning*. PMLR, 2019, pp. 583–593.

[27] E. L. Denton, S. Chintala, a. szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28, 2015.

[28] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision*, 2015.

[29] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 5901–5904.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[31] D. Dowson and B. Landau, "The Fréchet distance between multivariate normal distributions," *Journal of multivariate analysis*, vol. 12, no. 3, pp. 450–455, 1982.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[34] L. Van Der Maaten, "Learning a parametric embedding by preserving local structure," in *Artificial Intelligence and Statistics*. PMLR, 2009, pp. 384–391.

[35] O. Ashual and L. Wolf, "Specifying object attributes and relations in interactive scene generation," in *International Conference on Computer Vision*, 2019, pp. 4561–4569.

[36] ——, "Interactive scene generation via scene graphs with attributes," in *AAAI Conference on Artificial Intelligence*, vol. 34, no. 09, 2020, pp. 13 651–13 654.

[37] A. Chang, M. Savva, and C. D. Manning, "Interactive learning of spatial knowledge for text to 3d scene generation," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014, pp. 14–21.

[38] N. A. Ibraheem, M. M. Hasan, R. Z. Khan, and P. K. Mishra, "Understanding color models: a review," *ARPN Journal of science and technology*, vol. 2, no. 3, pp. 265–275, 2012.