Encouraging Disentangled and Convex Representation with Controllable Interpolation Regularization

Yunhao Ge, Zhi Xu, Yao Xiao, Gan Xin, Yunkui Pang, Laurent Itti
University of Southern California, Los Angeles, CA, USA
yunhaoge@usc.edu, itti@usc.edu

Abstract

We focus on controllable disentangled representation learning (C-Dis-RL), where users can control the partition of the disentangled latent space to factorize dataset attributes (concepts) for downstream tasks. Two general problems remain under-explored in current methods: (1) They lack comprehensive disentanglement constraints, especially missing the minimization of mutual information between different attributes across latent and observation domains. (2) They lack convexity constraints, which is important for meaningfully manipulating specific attributes for downstream tasks. To encourage both comprehensive C-Dis-RL and convexity simultaneously, we propose a simple yet efficient method: Controllable Interpolation Regularization (CIR), which creates a positive loop where disentanglement and convexity can help each other. Specifically, we conduct controlled interpolation in latent space during training, and we reuse the encoder to help form a 'perfect disentanglement' regularization. In that case, (a) disentanglement loss implicitly enlarges the potential understandable distribution to encourage convexity; (b) convexity can in turn improve robust and precise disentanglement. CIR is a general module and we merge CIR with three different algorithms: ELEGANT, I2I-Dis, and GZS-Net to show the compatibility and effectiveness. Qualitative and quantitative experiments show improvement in C-Dis-RL and latent convexity by CIR. This further improves downstream tasks: controllable image synthesis, cross-modality image translation and zero-shot synthesis.

1. Introduction

Disentangled representation learning empowers models to learn an orderly latent representation, in which each separate set of dimensions is responsible for one semantic attribute [10, 5, 22]. If we categorize different disentangled representation methods by whether they could *control* the partition of the obtained disentangled latent representation (e.g., explicitly assign first 10 dimensions to be responsible for face attribute), there are two main threads:

(1) **Uncontrollable** disentangled methods, such as Variational Autoencoders (VAEs) [13, 11, 18], add prior constraints (e.g., Gaussian distribution) in latent space to implicitly infer a disentangled latent code. Most are unsupervised methods that can easily generalize to different datasets and extract latent semantic factors. Yet, they struggle to obtain controllable disentanglement because the unsupervised latent encoding does not map onto user-controllable attributes. (2) **Controllable** disentangled methods, which explicitly control the partition of the disentangled latent space and the corresponding mapping to semantic attributes by utilizing dataset attribute labels or task domain knowledge. Because users can precisely control and design their task-driven disentangled latent representation, they are widely used in various downstream tasks: in cross-modality image-to-image translation, I2I-Dis [14] disentangle content and attribute to improve image translation quality (Fig. 1(b)); In controllable image synthesis, ELEGANT [21] and DNA-GAN [20] disentangle

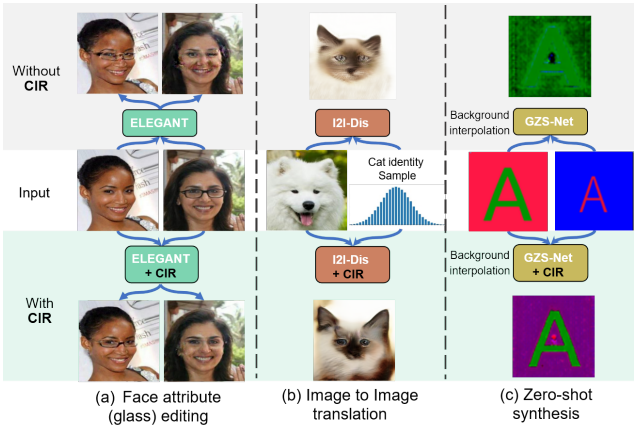

Figure 1. Our proposed approach CIR improves the result quality of 3 tasks by encouraging both disentanglement and convexity in the latent space: (a) Face attribute editing with ELEGANT (add/remove glasses on face); CIR is better able to transfer glasses with less disturbance on other face parts. (b) Image to image translation transfer from a dog image to a cat image with same pose (content); CIR better matches the desired pose with fewer artifacts. (c) Zero-shot synthesis with GZS-Net to synthesize an image with a new background by interpolating in the corresponding latent space; CIR better interpolates the background only without changing letter size, color or font style. See Suppl. fig. 1 for a larger version.

different face attributes to achieves face attribute transfer by exchanging certain part of their latent encoding across images ((Fig. 1(a))). In group supervised learning, GZS-Net [8] uses disentangled representation learning to simulate human imagination and achieve zero-shot synthesis (Fig. 1(c)).

However, controllable disentangled methods suffer from 2 general problems: 1) The constraints on disentanglement are partial and incomplete, they lack *comprehensive* disentanglement constraints. For example, while ELEGANT enforces that modifying the part of the latent code assigned to an attribute (e.g., hair color) will affect that attribute, it does not explicitly enforce that a given attribute will *not* be affected when the latent dimensions for other attributes are changed (Fig. 1(a)). 2) Most of the above-mentioned downstream tasks require manipulating specific attribute-related dimensions in the obtained disentangled representation; for instance, changing only the style while preserving the content in an image-to-image translation task. For such manipulation, the convexity of each disentangled attribute representation (i.e., interpolation within that attribute should give rise to meaningful outputs) is not guaranteed by current methods (Fig. 1, Fig. 3(a) and Fig. 7(a)). Further, convexity demonstrates an ability to generalize, which implies that the autoencoder structure has not simply memorized the representation of a small collection of data points. Instead, the model uncovered some structure about the data and has captured it in the latent space [3]. How to achieve both comprehensive disentanglement, and convexity in the latent space, is under-explored.

To solve the above problems, we first provide a definition of controllable disentanglement with the final goals of *perfect* controllable disentanglement and of convexity in latent space. Then, we use information theory and interpolation to analyze different ways to achieve disentangled (Sec. 3.1) and convex (Sec. 3.2) representation learning. To optimize them together, based on the definition and analysis, we use approximations to create a positive loop where disentanglement and convexity can help each other. We propose Controllable Interpolation Regularization (CIR), a simple yet effective general method that compatible with different algorithms to encourage both controllable disentanglement and convexity in the latent space (Sec. 3.3). Specifically, CIR first conducts controllable interpolation, i.e., controls which attribute to interpolate and how in the disentangled latent space, then reuses the encoder to 're-obtain' the latent code and add regularization to explicitly encourage *perfect* controllable disentanglement and implicitly boost convexity. We show that this iterative approximation approach converges towards perfect disentanglement and convexity in the limit of infinite interpolated samples.

Our contributions are: (i) Describe a new abstract framework for *perfect* controllable disentanglement and convexity in the latent space, and use information theory to summarize potential optimization methods (Sec. 3.1, Sec. 3.2). (ii) Propose Controllable Interpolation Regularization (CIR), a general module compatible with different algorithms, to encourage both controllable disentanglement and convex in latent representation by creating a positive loop to make them help each other. CIR is shown to converge towards perfect disentanglement and convexity for infinite interpolated samples (Sec. 3.3). (iii) Demonstrate that better disentanglement and convexity are achieved with CIR on various tasks: controllable image synthesis, cross-domain image-to-image translation and group supervised learning (Sec. 4, Sec. 5).

2. Related Work

Controllable Disentangled Representation Learning (C-Dis-RL) is different from Uncontrollable Dis-RL (such as VAEs [13, 11, 4]), which implicitly achieves disentanglement by incorporating a distance measure into the objective, encouraging the latent factors to be statistically independent. However, these methods and not able to freely control the relationship between attribute and latent dimensions. C-Dis-RL learns a partition control of the disentanglement from semantic attribute labels in the latent representation and boosts the performance of various tasks: ELEGANT [21] and DNA-GAN [20] for face attribute transfer; I2I-Dis [14] for diverse image-to-image translation; DGNet [22] and IS-GAN [7] for person re-identification; GZS-Net [8] for controllable zero-shot synthesis with group-supervised learning. However, their constraints on disentanglement are implicit and surrogate by image quality loss, which also misses the constraint between different attributes across latent and observation. As a general module, CIR is compatible and complementary with different C-Dis-RL algorithms by directly constraining disentanglement while focusing on minimizing the mutual information between different attributes across latent and observation.

Convexity of Latent Space is defined as a set in which the line segment connecting any pair of points will fall within the rest of the set [17]. Linear interpolations in a low-dimensional latent space often produce comprehensible representations when projected back into high-dimensional space [6, 9]. However, linear interpolations are not necessarily justified in many controllable disentanglement models because latent-space projections are not trained explicitly to form a convex set. VAEs overcome non-convexity by forcing the latent representation into a pre-defined distribution, which may be a suboptimal representation of the high-dimensional data. GAIN [17] adds interpolation in the generator in the middle latent space and uses a discriminative loss to help optimize convexity. Our method controls the interpolation in a subspace of the disentangled latent space and uses disentanglement regularization to encourage a convex latent space for each semantic attribute.

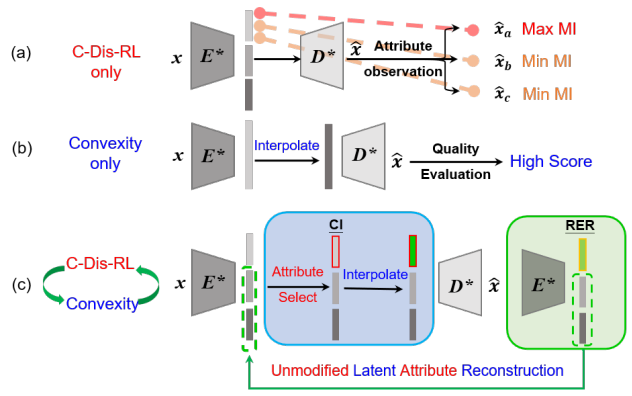

Figure 2. Intuitive understanding of Controllable Interpolation Regularization (**CIR**). (a) Only encourage controllable disentangled representation (C-Dis) with general Mutual Information (MI) constrain method: maximize the MI between the same attribute across latent and observation domains while minimizing the MI between the different attribute across latent and observation domains.(b) Only encourage convexity with interpolation and image quality evaluation. (c) A simple yet efficient method, CIR, encourages both C-Dis and convexity in latent representation. CIR consists of a Controllable Interpolation (CI) module and a Reuse Encoder Regularization (RER) module.

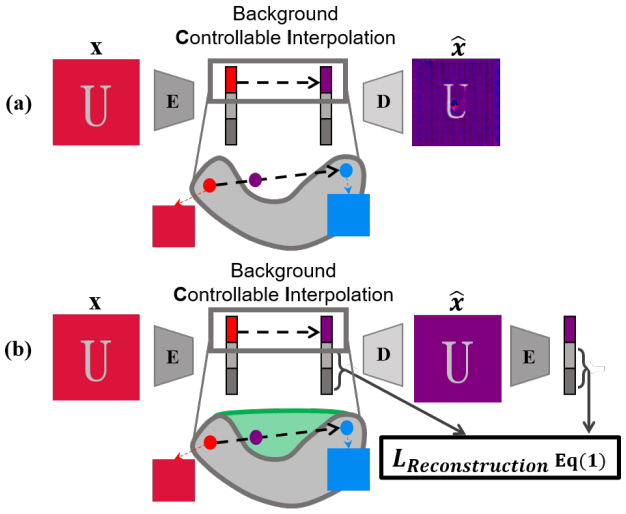

Figure 3. Interpolation in the disentangled latent space for background color of GZS-Net (a) without CIR, the latent space is not convex (purple point out of understandable gray region) and the synthesized image shows some contamination over unmodified attributes (size and foreground letter) (b) Architecture of GZS-Net + **CIR**, which encourages a more disentangled and convex latent space.

3. Controllable Interpolation Regularization

3.1. Mutual Information for Perfect Controllable disentanglement

A general autoencoder structure $(D \circ E)$: $\mathcal{X} \rightarrow \mathcal{X}$ is composed of an encoder network $E : \mathcal{X} \rightarrow \mathbb{R}^d$, and a decoder network $D : \mathbb{R}^d \rightarrow \mathcal{X}$. $\mathbb{R}^d$ is a latent space, compared with the original input space $\mathcal{X}$ (e.g., image space). The disentanglement is a property of latent space $\mathbb{R}^d$ where each separate set of dimensions is responsible for one semantic attribute of given dataset. Formally, a dataset (e.g., face dataset) contains n samples $\mathcal{D} = \{x^{(i)}\}_{i=1}^n$, each accompanied by m attributes $\mathcal{D}_a = \{(a_1^{(i)}, a_2^{(i)}, \dots a_m^{(i)})\}_{i=1}^n$. Each attribute $a_j \in \mathcal{A}_j$ can be either binary (two attribute values, e.g., $\mathcal{A}_1$ may denote wearing glass or not; $\mathcal{A}_1 = \{\text{wear glass}, \text{not wear glass}\}$), or a multi-class attribute, which contains a countable set of attribute values (e.g., $\mathcal{A}_2$ may denote hair-colors $\mathcal{A}_2 = \{\text{black}, \text{gold}, \text{red}, \dots\}$). Controllable disentangled representation learning (C-Dis-RL) methods have two properties: (1) Users can explicitly control the partition of the disentangled latent space $\mathbb{R}^d$ and (2) Users can control the semantic attributes mapping between $\mathbb{R}^d$ to input space $\mathcal{X}$. To describe the ideal goal for all C-Dis-RL, we define a *perfect* controllable disentanglement property in latent space $\mathbb{R}^d$ and the autoencoder.

Definition 1 *perfect* CONTROLLABLE DISENTANGLEMENT (*perfect*-C-D)$(E, D, \mathcal{D})$*: Given a general encoder* $E : \mathcal{X} \rightarrow \mathbb{R}^d$*, a decoder* $D : \mathbb{R}^d \rightarrow \mathcal{X}$*, and a dataset* $\mathcal{D}$ *with* m *independent semantic attributes* $\mathcal{A}$*, we say the general autoencoder achieve perfect controllable disentanglement for dataset* $\mathcal{D}$ *if the following property is satisfied: (1) For encoder* E*, if one attribute* $\mathcal{A}_i$ *of input* x *was specifically modified, transforming* x *into* $\hat{x}$*, after computing latent codes* $z = E(x)$ *and* $\hat{z} = E(\hat{x})$*, the difference between* z *and* $\hat{z}$ *should be zero for all latent dimensions except those that represent the modified attribute. (2) Similarly, for decoder* D*, the latent space change should only influence the corresponding attribute expression in the output (e.g., image) space.*

To encourage a general autoencoder structure model to obtain *perfect* controllable disentanglement property, we propose an information-theoretic regularization with two perspectives (Fig. 2(a)): (1) Maximize the mutual information $(I())$ between the *same* attribute across latent space $\mathbb{R}^d$ and observation input space $\mathcal{X}$; and (2) Minimize the mutual information between the *different* attributes across latent $\mathbb{R}^d$ and observation input space $\mathcal{X}$. Formally:

$$\begin{aligned} &\max_{E,D} \Big[I(x_{\mathcal{A}_i}, E(x)_{\mathcal{A}_i}) + I(E(x)_{\mathcal{A}_i}, D(E(x))_{\mathcal{A}_i}) \Big]; \\ &\min_{E,D} \Big[I(x_{\mathcal{A}_i}, E(x)_{\mathcal{A}_j}) + I(E(x)_{\mathcal{A}_i}, D(E(x))_{\mathcal{A}_j}) \Big]; \end{aligned} \tag{1}$$

where $x_{\mathcal{A}_i}$ and $D(E(x))_{\mathcal{A}_i}$ represent the observation of attribute $\mathcal{A}_i$ in $\mathcal{X}$ domain (e.g., hair color in human image); $E(x)_{\mathcal{A}_i}$ represents the dimensions in $\mathbb{R}^d$ that represent attribute $\mathcal{A}_i$; $i, j \in [1..m]$ and $i \neq j$ (Fig. 2(a)).

3.2. Convexity Constraint with Interpolation

A convex latent space has the property that the line segment connecting any pair of points will fall within the rest

of the space [17]. As shown in Fig. 3(a), the gray region represents the 2D projection of the latent representation of one attribute (e.g., background color) for a dataset. This distribution would be non-convex, because the purple point, though between two points in the distribution (the red and blue points, represent two background color), falls in the space that does not correspond to the data distribution. This non-convexity may cause that the projection back into the image space does not correspond to a proper semantically meaningful realistic image ($\hat{x}$ in Fig. 3(a) influence other unmodified attributes, i.e, size and foreground letter). This limitation makes disentanglement vulnerable and hinders potential latent manipulation in downstream tasks. The result of Fig. 4 and 5in experiments illustrate this problem.

To encourage a convex data manifold, the usefulness of interpolation has been explored in the context of representation learning [2] and regularization [19]. As is shown in Fig. 1(b), we summarize the constraint of convexity in the latent space: we use a dataset-related quality evaluation function $Q()$ to evaluate the "semantic meaningfulness" of input domain samples; a higher value means high quality and more semantic meaning. After interpolation in latent space $\mathbb{R}^d$, we want the projection back into the original space to have a high $Q()$ score. Formally:

$$\max_{E,D}\left\{\mathbb{E}_{x_1,x_2\in\mathcal{D}}\left[Q(D(\alpha E(x_1)+(1-\alpha)E(x_2)))\right]\right\} \quad (2)$$

where x_1 and x_2 are two data samples and $\alpha \in [0..1]$ controls the latent code interpolation in $\mathbb{R}^d$.

The dataset-related quality evaluation function $Q()$ also has different implementations: [17] utilizes additional discriminator and training adversarially on latent interpolations; [3] uses a critic network as a surrogate which tries to recover the mixing coefficient from interpolated data.

3.3. CIR: encourage both C-Dis-RL and Convexity

Our goal is to encourage a controllable disentangled representation, and, for each semantic attribute-related latent dimension, the created space should be as convex as possible. Specifically, we want to optimize both controllable disentanglement (Eq. 1) and convexity (Eq. 2) for each semantic attribute. In practice, each mutual information term in Eq. 1 is hard to optimize directly as it requires access to the posterior. Most of the current methods use approximation to obtain the lower bound for optimizing the maximum [5, 1] or upper bound for optimizing minimum [13]. However, it is hard to approximate so many $(2m(m-1)+2m)$ different mutual information terms in Eq. 1) simultaneously, not to mention considering the convexity of m latent space (Eq. 2) as well. To optimize them together, we propose to use a controllable disentanglement constraint to help the optimization of convexity and in turn, use convexity constraint to help a more robust optimization of the controllable disentanglement. In other words, we create a positive loop between controllable disentanglement and convexity, to help each other. Specifically, as shown in Fig. 1(c), we propose a simple yet efficient regularization method, Controllable Interpolation Regularization (CIR), which consists of two main modules: a Controllable Interpolation (CI) module and a Reuse Encoder Regularization (RER) module. It works as follows: an input sample x goes through E to obtain latent code $z = E(x)$. Because our goal is controllable disentanglement, on each iteration we only focus on one attribute. CI module first selects one attribute $\mathcal{A}_i$ among all m attributes, and then interpolates along the $\mathcal{A}_i$ related latent space in z while preserving the other unselected attributes, yielding $z_{\mathcal{A}_i}$. After D translates the interpolated latent $z_{\mathcal{A}_i}$ back to image space, the RER module takes $D(z_{\mathcal{A}_i})$ as input and reuses the encoder to get the latent representation $z^{re}_{\mathcal{A}_i} = E(D(z_{\mathcal{A}_i}))$. RER then adds a reconstruction loss on the *unmodified latent space* as a regularization:

$$L_{\text{reg}} = ||z_{-\mathcal{A}_i} - z^{re}_{-\mathcal{A}_i}||_{l1} \quad (3)$$

where $z_{-\mathcal{A}_i}$ and $z^{re}_{-\mathcal{A}_i}$ denote the all latent dimensions of $z_{\mathcal{A}_i}$ and $z^{re}_{\mathcal{A}_i}$ respectively, except those that represent the modified attribute $\mathcal{A}_i$. Eq. 3 explicitly optimizes Eq. 1: in each iteration, if the modified latent region $z_{\mathcal{A}_i}$ only influences the expression of $x_{\mathcal{A}_i}$, then, after reusing E, the unmodified region in $E(D(z_{\mathcal{A}_i}))$ should remain as is (min E, D in Eq. 1). On the one hand, for those unselected attributes, their information should be preserved in the whole process (max E, D in Eq. 1). Eq. 3 also implicitly optimizes Eq. 2: if the interpolated latent code is not 'understandable' by E and D, the RER module does not work and the L_{reg} would be large. Fig. 2 (a) and (b) abstractly demonstrate the latent space convexity difference before and after adding CIR to GZS-Net [8]. Convexity and disentanglement are dual tasks in the sense that one can help enhance the other's performance. On the other hand, the reconstruction loss towards *perfect* controllable disentanglement implicitly encourages a convex attribute latent space; The more convex the latent space, the more semantically meaningful samples synthesized by interpolation will help the optimization of controllable disentanglement, which encourages a more robust C-Dis-RL. From the perspectives of loss function and optimization, if the reconstruction loss could decrease to zero for a given dataset augmented by many interpolated samples, then perfect disentanglement and convexification are achieved. That is, CIR forces, in the limit of infinite interpolated samples, the disentangled latent representation of every attribute to be $convex$, where every interpolation along every attribute is guaranteed to be meaningful.

4. Qualitative Experiments

We qualitatively evaluated our CIR as a general module and merged it into three baseline models on three different tasks (Fig. 5): multiple face attributes transfer with ELEGANT [21] (Sec. 4.1), cross modality image translation with

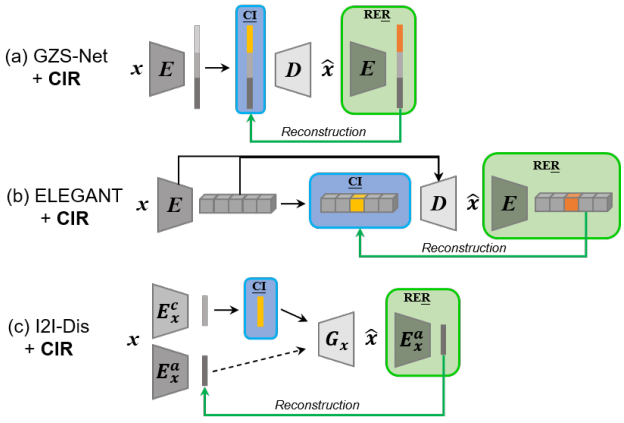

Figure 4. CIR consists of a Controllable Interpolation (CI; shown in blue) module and a Reuse Encoder Regularization (RER; green) module. (a-c) **CIR** compatible to different models. (a) GZS-Net [8] + **CIR** (b) ELEGANT [21] + **CIR**. (c) I2I-Dis [14] + **CIR**. Grey components are the baseline methods.

I2I-Dis [14] (Sec. 4.2) and zero-shot synthesis with GZS-Net [8] (Sec. 4.3). CIR encourages a better disentanglement and convexity in their latent space to further improve their performance.

4.1. CIR boosts multiple face attributes transfer

We conduct the same face attribute transfer tasks as in ELEGANT [21] paper with *CelebA* [15]. **Task 1**: taking two face images with the opposite attribute as input and generate new face images which exactly transfer the opposite attribute between each other (Fig. 5). **Task 2**: generate different face images with the same style of the attribute in the reference images (Fig. 6). Both of the two tasks require a robust controllable disentangled latent space to swap attributes of interest to synthesize new images and the convexity of latent space influences image quality.

Fig. 4(b) shows the high-level structure about how CIR (blue and green block) compatible to ELEGANT (grey). ELEGANT adopts a U-Net [16] structure (autoencoder) to generate high-resolution images with exemplars. In this way, the output of the encoder is the latent code of disentangled attributes and the context information is contained in the output of the intermediary layer of the encoder. ELEGANT adopts an iterative training strategy: training the model with respect to a particular attribute each time. We use the same training strategy but adding our regularization loss term. As shown in Fig. 4 (b), to encourage the disentanglement and convexity of attribute $\mathcal{A}_i$, CIR interpolates $\mathcal{A}_i$-related dimensions in latent code (yellow) and constrains the other latent dimensions to remain unchanged after D and reused E. Specifically, when training ELEGANT about the $\mathcal{A}_i$ attribute **Eyeglasses** at a given iteration, we obtain the latent code $zA = E(A)$ and $zB = E(B)$ with E for each pair of images A and B with opposite $\mathcal{A}_i$ attribute value. The disentangled latent code is partitioned into $z_{+\mathcal{A}_i}$ for latent dimensions related to $\mathcal{A}_i$, and $z_{-\mathcal{A}_i}$ for unrelated dimensions.

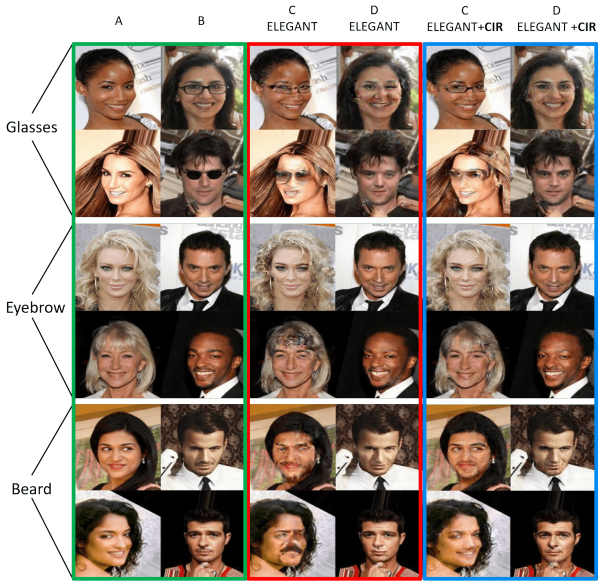

Figure 5. ELEGANT + **CIR** performance (task 1) for two images face attribute transfer (inputs: A,B ; outputs: C,D).

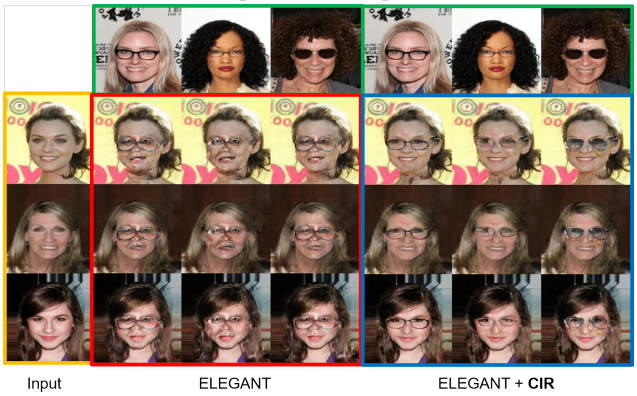

Figure 6. ELEGANT + **CIR** performance (task 2) for face generation by exemplars: Input image (orange) should be modified as different face images with the same style of the Eyeglasses attribute in the reference images (green).

We interpolate in $z_{+\mathcal{A}_i}$ with zA and zB while keeping the other dimensions $z_{-\mathcal{A}_i}$ as is to obtain interpolated latent code $zA_{\mathcal{A}_i}$ and $zB_{\mathcal{A}_i}$. After D and reuse E, we get the reconstructed latent representation $zA^{re}_{\mathcal{A}_i} = E(D(zA_{\mathcal{A}_i}, zA))$ and $zB^{re}_{\mathcal{A}_i} = E(D(zB_{\mathcal{A}_i}, zB))$. The reconstruction loss as a regularization is (an instantiation of Eq. 3):

$$L_{\text{reg}} = ||zA_{-\mathcal{A}_i} - zA^{re}_{-\mathcal{A}_i}||_{l2} + ||zB_{-\mathcal{A}_i} - zB^{re}_{-\mathcal{A}_i}||_{l2} \quad (4)$$

The overall generative loss of ELEGANT + CIR is:

$$\mathcal{L}(G) = L_{\text{reconstruction}} + L_{\text{adv}} + \boldsymbol{\lambda}_{\text{CIR}} \boldsymbol{L}_{\text{reg}} \quad (5)$$

where $L_{\text{reconstruction}}$ and L_{adv} are ELEGANT original loss terms, $\lambda_{\text{CIR}} > 0$ control the relative importance of the loss terms. we keep the discriminative loss. (More network architecture and training details are in Supplementary)

Fig. 5 shows the task 1 performance on two images face attribute transfer. Take Eyeglasses as an example attribute

to swap: A,B are input, the output C and D should keep all other attributes unmodified except for swapping the Eyeglasses. ELEGANT generated C and D have artifacts in Eyeglasses-unrelated regions, which means ELEGANT cannot disentangle well in latent space. After adding CIR, the generated C and D better preserve the irrelevant regions during face attribute transfer, which demonstrates that CIR helps encourage a more convex and disentangled latent space. The Eyebrow and Beard attribute results also show the improvement from CIR. Fig. 6 shows the task 2 performance on face image generation by exemplars. Input image (orange) should be modified as different face images with the same style of the Eyeglasses attribute in the reference images (green). ELEGANT generated new images with artifacts in Eyeglasses-unrelated regions that cannot disentangle well. Synthesis is also inferior in the glasses region, which we posit is due to non-convexity in the eyeglass-related latent space. With the help of CIR, the generated images improve both Eyeglass quality and irrelevant region preservation.

4.2. CIR boosts cross modality image translation

We conduct the same image-to-image translation task as in I2I-Dis [14] paper with *cat2dog* dataset [14]. Fig. 4(c) shows the high-level structure about how CIR (blue and green block) compatible to I2I-Dis (grey). There are two image domains $\mathcal{X}$ (cat) and $\mathcal{Y}$ (dog), I2I-Dis embeds input images onto a shared content space $\mathcal{C}$ with specific encoders ($E^c_{\mathcal{X}}$ and $E^c_{\mathcal{Y}}$), and domain-specific attribute spaces $\mathcal{A}_{\mathcal{X}}$ and $\mathcal{A}_{\mathcal{Y}}$ with specific encoders ($E^a_{\mathcal{X}}$ and $E^a_{\mathcal{Y}}$) respectively. After that, new images can be synthesized by transferring the shared content attribute cross-domain (between cat and dog), such as generating unseen dogs with the same content attribute value (pose and outline) as the reference cat (Fig. 7). Domain-specific attribute $\mathcal{A}_{\mathcal{X}}$ and $\mathcal{A}_{\mathcal{Y}}$ already been constraint by adding a KL-Divergence loss with Gaussian distribution; thus, we can freely sample in Gaussian for synthesis. The shared content space $\mathcal{C}$ could be encouraged as a more convex and disentangled space by CIR.

We use the same network architecture and training strategy as I2I-Dis except for adding our regularization loss term. As shown in Fig. 4 (c), during each training iteration, a cat image x and a dog image y go through corresponding encoders and each of them produce latent codes of domain ($zx_a = E^a_{\mathcal{X}}(x)$, $zy_a = E^a_{\mathcal{Y}}(y)$) and content ($zx_c = E^c_{\mathcal{X}}(x)$, $zy_c = E^c_{\mathcal{Y}}(y)$). Then an interpolated content attribute latent code (yellow) zxy_c (between zx_c and zy_c) concatenates with the domain attribute latent code of cat image zx_a and dog image zy_a respectively and forms two new latent codes, and decoders turns them into new images $u = G_{\mathcal{X}}(zx_a, zxy_c)$, $v = G_{\mathcal{Y}}(zy_a, zxy_c)$. To encourage the disentanglement and convexity of the content attribute, we reuse $E^a_{\mathcal{X}}$ and $E^a_{\mathcal{Y}}$ to get the reconstructed domain attribute latent representations $zx^{re}_a = E^a_{\mathcal{X}}(u)$, $zy^{re}_a = E^a_{\mathcal{Y}}(v)$ and add the reconstruction loss as a regularization (an instantiation of Eq. 3):

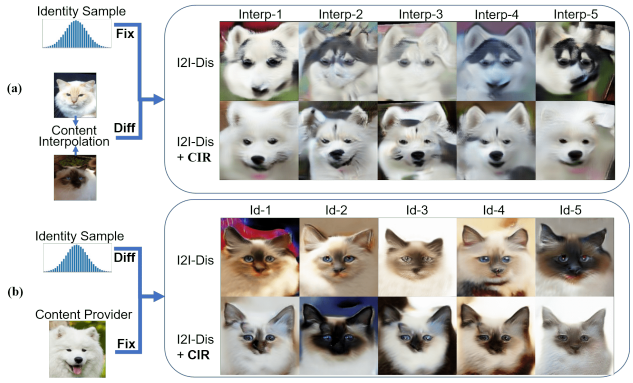

Figure 7. I2I-Dis + **CIR** performance of diverse image-to-image translation. (a) For any dog image sample, create several interpolated images with content (here, pose, ear orientation, etc) in between that of the two reference cat images. (b) For several cat identity samples, synthesize images with that cat's identity but the content (pose, etc) of the reference dog image.

$$L_{\text{reg}} = ||zx^{re}_a - zx_a||_{l1} + ||zy^{re}_a - zy_a||_{l1} \tag{6}$$

The overall loss of I2I-Dis + CIR is

$$\begin{aligned}\mathcal{L} =& \lambda^{content}_{adv} L^c_{adv} + \lambda^{cc}_1 L^{cc}_1 + \lambda^{domain}_{adv} L^{domain}_{adv} + \\ & \lambda^{recon}_1 L^{recon}_1 + \lambda^{latent}_1 L^{latent}_1 + \lambda_{KL} L_{KL} + \boldsymbol{\lambda}_{\text{CIR}} \boldsymbol{L}_{\text{reg}}\end{aligned} \tag{7}$$

where content and domain adversarial loss L^c_{adv} L^{domain}_{adv}, cross-cycle consistency loss L^{cc}_1, self-reconstruction loss L^{recon}_1, latent regression loss L^{latent}_1 and KL loss L_{KL} are I2I-Dis original loss terms, $\lambda > 0$ control the relative importance of the loss terms. (More details in Supplementary).

Fig. 7 shows the image-to-image translation performance. (a) We fix the identity (domain) latent code and change the content latent code by interpolation; generated images should keep the domain attribute (belong to the same dog). I2I-Dis generated dog images have artifacts, which means the non-convex latent space cannot 'understand' the interpolated content code. After adding our CIR, the generated images have both better image quality and consistency of the same identity. (b) We fix the content latent code and change the identity by sampling; generated images should keep the same content attribute (pose and outline). Cat images generated by I2I-Dis have large pose variance (contain both left and right pose), and large face outline variance (ear positions and sizes). After adding our CIR, the generated images have smaller pose and outline variance. (More results in Supplementary)

4.3. CIR boosts zero-shot synthesis

We use the same architecture of autoencoders as GZS-Net [8] and *Fonts* dataset [8]. Fig. 4(a) shows the high-level structure about how CIR (blue and green block) compatible to GZS-Net (grey). The latent feature after encoder E is a

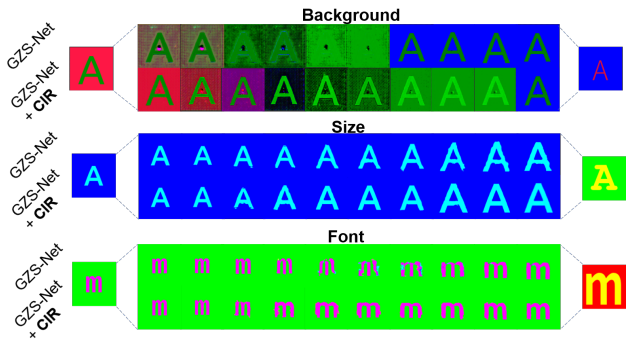

Figure 8. GZS-Net + **CIR** performance of interpolation-based attribute controllable synthesis. Top: Interpolation in the latent space of background color. Middle: interpolation of letter size. Bottom: Interpolation of font style. In all three cases, CIR provides both better disentanglement (attributes other than the interpolated one do not change as much) and higher interpolation quality (interpolated attributes show fewer artifacts).

100-dim vector, and each of the five *Fonts* attributes (content, size, font color, background color, font) covers 20-dim. The decoder D, symmetric to E, takes the 100-dim vector as input and outputs a synthesized sample. We use the same Group-Supervised learning training strategy as GZS-Net except for adding our regularization loss term Eq. 1, which is exactly the same as the one described in Sec. 3.3 and Fig. 3 (b). Besides the reconstruction loss L_{r}, swap reconstruction loss L_{sr} and cycle swap reconstruction loss L_{csr} which are same as GSL, we add a regularization reconstruction loss L_{reg}. The total loss function is:

$$\mathcal{L}(E, D) = L_{\text{r}} + \lambda_{\text{sr}} L_{\text{sr}} + \lambda_{\text{csr}} L_{\text{csr}} + \boldsymbol{\lambda}_{\text{CIR}} \boldsymbol{L}_{\text{reg}} \qquad (8)$$

where $\lambda_{\text{sr}}, \lambda_{\text{csr}}, \lambda_{\text{CIR}} > 0$ control the relative importance of the loss terms.

Fig. 8 shows the interpolation-based controllable synthesis performance on background, size, and font attributes. Take background interpolation synthesis as an example: we obtain background latent codes by interpolating between the left and right images, and each of them concatenates with the unselected 80-dim latent code from the left image. Generated images should keep all other attributes unmodified except for the background. GZS-Net generated images have artifacts in background-unrelated regions, i.e., GZS-Net cannot disentangle well in latent space. After adding our CIR, the generated images better preserve the irrelevant areas during synthesis. The size and font attribute results also show improvement from CIR. (More results in Supplementary).

5. Quantitative Experiments

We conduct five quantitative experiments to evaluate the performance of CIR on latent disentanglement and convexity.

(a) GZS-Net (b) GZS-Net + **CIR**

Figure 9. Disentangled representation analysis. (a) Baseline GZS-Net. (b) With CIR, off-diagonal elements (entanglement across attributes) are reduced.

5.1. Controllable Disentanglement Evaluation by Attribute Co-prediction.

Can latent features of one attribute predict the attribute value? Can they also predict values for other attributes? Under *perfect* controllable disentanglement, we should answer *always* for the first and *never* for the second. We quantitatively assess disentanglement by calculating a model-based confusion matrix between attributes. We evaluate GZS-Net [8] + **CIR** with the *Fonts* [8] dataset (latent of ELEGANT and I2I-Dis are not suitable). Each image in *Fonts* contains an alphabet letter rendered using 5 independent attributes: content (52 classes), size (3), font color (10), background color (10), and font (100). We take the test examples and split them 80:20 for train_{DR}:test_{DR}. For each attribute pair $j, r \in [1..m] \times [1..m]$, we train a classifier (3 layer MLP) from g_j of train_{DR} to the attribute values of r, then obtain the accuracy of each attribute by testing with g_j of test_{DR}. Fig. 9 compares how well features of each attribute (row) can predict an attribute value (column): perfect should be as close as possible to Identity matrix, with off-diagonal entries close to random (i.e., 1 / $|\mathcal{A}_r|$). The off-diagonal values of GZS-Net show the limitation of disentanglement performance; with CIR's help, the co-prediction value shows a better disentanglement.

5.2. Controllable Disentanglement Evaluation by Correlation Coefficient.

For each method, we collect 10,000 images from the corresponding dataset (ELEGANT [21] with *CelebA* [15], GZS-Net with *Fonts* [8]) and obtain 10,000 latent codes by Es. We calculate the correlation coefficient matrix between dimensions in latent space. A near-perfect disentanglement should yield high intra-attribute correlation but low inter-attribute correlation. ELEGANT disentangles two attributes: eyeglasses and mustache, each of which covers 256-dimensions. GZS-Net disentangles five attributes: content, size, font color, background color, and font; each covers 20-dimensions. Fig. 10 shows that CIR improves the disentanglement in latent space, as demonstrated by higher intra-attribute and lower inter-attribute correlations (More details in Suppl.).

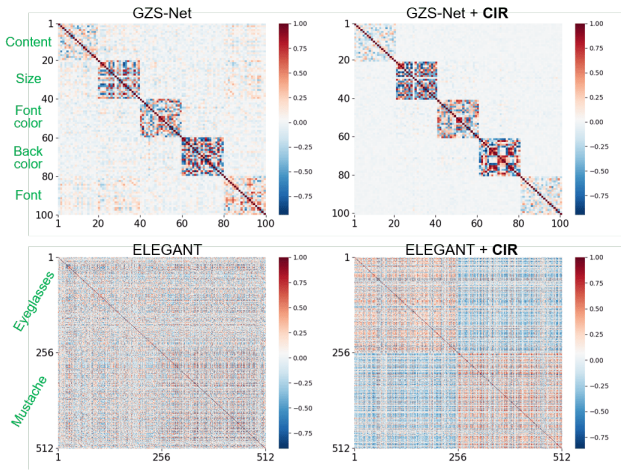

Figure 10. Disentanglement Evaluation by Correlation Coefficient. Intra-attribute correlation increases with CIR (GZS-Net (top): 7.2%, ELEGANT (bottom): 3.2%) while inter-attribute decreases (GZS-Net: 60.9%, ELEGANT: 3.1%).

Table 1. Convexity Evaluation with Image Quality Score

Algorithms	Train images	Test images	High quality probability
ElEGANT	6000	1500	12%
ElEGANT + **CIR**			**60%**
I2I-Dis	1500	1500	18%
I2I-Dis + **CIR**			**33%**
GZS-Net	6000	1000	13%
GZS-Net + **CIR**			**40%**

5.3. Convexity Evaluate with Image Quality Score.

To evaluate the overall convexity in latent space, we use an image quality classifier to evaluate the quality of images generated by interpolating in latent space. We train a specific image quality classifier for each baseline algorithm and corresponding dataset. Take ELEGANT as an instance: To train a classifier for ELEGANT and ELEGANT + CIR, we use 3000 *CelebA* original images as positive, high-quality images. To collect negative images, we first randomly interpolate the latent space of both ELEGANT and ELEGANT + CIR and generate interpolated images for negative low-quality images; then, we manually select 3000 low-quality images (artifact, non-sense, fuzzy ...) and form a 6000 images training set. After training an image quality classifier, we test it on 1500 images generated by interpolation-based attribute controllable synthesis as Exp. 4.1. Table 1 shows the average probability of high-quality images (higher is better). The training and testing for I2I-Dis (+ CIR) and GZS-Net (+ CIR) are similar.

5.4. *Perfect* Disentanglement Property Evaluation.

As we defined in Sec. 3.1, *Perfect* disentanglement property can be evaluated by the difference of the unmodified attribute related dimensions in $\mathbb{R}^d$ after modifying a specific attribute $\mathcal{A}_i$ in image space. For the two methods in each column (Table 2) and corresponding datasets, we modify one attribute value $\mathcal{A}_i$ of each input x and get $\hat{x}$, then obtain latent codes ($z = E(x)$, $\hat{z} = E(\hat{x})$) with two methods' encoders

Table 2. *Perfect* Disentanglement Property Evaluation

Algorithms	ElEGANT	I2I-Dis	GZS-Net
MSE	1.9	1.8	3.42
Algorithms	ElEGANT + **CIR**	I2I-Dis + **CIR**	GZS-Net + **CIR**
MSE	**0.38**	**0.1**	**0.27**

Table 3. Disentanglement Evaluation with StyleGAN Perceptual Path Length Metric. Lower difference is better.

I2I-Dis	I2I-Dis + **CIR**	ElEGANT	ElEGANT + **CIR**
29	21	1.23	0.68

respectively. After we normalized the latent codes from two methods into the same scale, we calculate the Mean Square Error (MSE) of the unmodified region $MSE(z_{-\mathcal{A}_i}, \hat{z}_{-\mathcal{A}_i})$ between z and $\hat{z}$ (lower is better). Table 2 shows that after adding CIR, we obtain a lower MSE, which means CIR encourages a better disentangled latent space.

5.5. C-Dis Evaluation with Perceptual Path Length

We use a method similar to the perceptual path length metric in StyleGAN [12], which measure the difference between consecutive images (their VGG16 embeddings) when interpolating between two random inputs. We subdivide a latent space interpolation path into linear segments. In our experiment, we use a small subdivision epsilon $\epsilon = 10^{-4}$ and linear interpolation (lerp). Thus, the average perceptual path length in latent space $\mathcal{Z}$ is $l_{\mathcal{Z}} = \mathbb{E}\left[\frac{1}{\epsilon^2} d\left(G\left(\text{lerp}\left(\mathbf{z}_1, \mathbf{z}_2; t\right)\right), G\left(\text{lerp}\left(\mathbf{z}_1, \mathbf{z}_2; t+\epsilon\right)\right)\right)\right]$ $\mathbf{z}_1, \mathbf{z}_2$ is the start point and the end point. G can be a decoder in Auto-encoder or generator in a GAN-based model. $t \sim U(0,1)$. d is the distance in VGG16 embeddings. Our results can be seen in Table. 3 where CIR improves the latent disentanglement.

6. Conclusion

We proposed a general disentanglement module, Controllable Interpolation Regularization (CIR), compatible with different algorithms to encourage more convex and robust disentangled representation learning. We show the performance of CIR with three baseline methods ELEGANT, I2I-Dis, and GZE-Net. CIR first conducts controllable interpolation in latent space and then 'reuses' the encoder to form an explicit disentanglement constraint. Qualitative and quantitative experiments show that CIR improves baseline methods performance on different controllable synthesis tasks: face attribute transfer, diverse image-to-image transfer, and zero-shot image synthesis with different datasets: *CelebA*, *cat2dog* and *Fonts* respectively.

Acknowledgment This work was supported by C-BRIC (one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA), DARPA (HR00112190134), the Army Research Office (W911NF2020053), and the Intel and CISCO Corporations. The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

References

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540. PMLR, 2018.

[2] Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai. Better mixing via deep representations. In *International conference on machine learning*, pages 552–560. PMLR, 2013.

[3] David Berthelot, Colin Raffel, Aurko Roy, and Ian Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer. *arXiv preprint arXiv:1807.07543*, 2018.

[4] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016.

[6] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pages 1068–1077. PMLR, 2017.

[7] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. *arXiv preprint arXiv:1910.12003*, 2019.

[8] Yunhao Ge, Sami Abu-El-Haija, Gan Xin, and Laurent Itti. Zero-shot synthesis with group-supervised learning. *arXiv preprint arXiv:2009.06586*, 2020.

[9] Yunhao Ge, Jiaping Zhao, and Laurent Itti. Pose augmentation: Class-agnostic object pose transformation for object recognition. In *European Conference on Computer Vision*, pages 138–155. Springer, 2020.

[10] Irina Higgins, Loic 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. 2016.

[11] I. Higgins, Loïc Matthey, A. Pal, C. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

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

[13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

[14] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018.

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[17] Tim Sainburg, Marvin Thielk, Brad Theilman, Benjamin Migliori, and Timothy Gentner. Generative adversarial interpolative autoencoding: adversarial training on latent space interpolations encourage convex latent distributions. *arXiv preprint arXiv:1807.06650*, 2018.

[18] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1415–1424, 2017.

[19] Vikas Verma, Alex Lamb, Christopher Beckham, Aaron Courville, Ioannis Mitliagkis, and Yoshua Bengio. Manifold mixup: Encouraging meaningful on-manifold interpolation as a regularizer. *arXiv preprint arXiv:1806.05236*, 7, 2018.

[20] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Dna-gan: Learning disentangled representations from multi-attribute images. *arXiv preprint arXiv:1711.05415*, 2017.

[21] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–187, September 2018.

[22] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2138–2147, 2019.