Self-supervised Enhancement of Latent Discovery in GANs

Silpa Vadakkeeveetil Sreelatha[1*], **Adarsh Kappiyath**[2*], **S Sumitra**[3]

[1] TCS Research, Pune, India
[2] Flytxt Mobile Solutions, Trivandrum, India
[3] Indian Institute of Space Science and Technology, Trivandrum, India
silpavs.43@gmail.com, kadarsh22@gmail.com, sumitra@iist.ac.in

Abstract

Several methods for discovering interpretable directions in the latent space of pre-trained GANs have been proposed. Latent semantics discovered by unsupervised methods are relatively less disentangled than supervised methods since they do not use pre-trained attribute classifiers. We propose Scale Ranking Estimator (SRE), which is trained using self-supervision. SRE enhances the disentanglement in directions obtained by existing unsupervised disentanglement techniques. These directions are updated to preserve the ordering of variation within each direction in latent space. Qualitative and quantitative evaluation of the discovered directions demonstrates that our proposed method significantly improves disentanglement in various datasets. We also show that the learned SRE can be used to perform Attribute-based image retrieval task without further training.

Introduction

Generative Adversarial Networks (GAN) are generative models that have witnessed significant performance improvements in image synthesis over the last decade (Goodfellow et al. 2014). It has many applications, including image, audio, and video generation, image manipulation and editing, image-to-image translation, and many others.

The latent space of GANs is hard to interpret due to its high dimensional and abstract structure. Various architectures such as InfoGAN (Chen et al. 2016), Structured GAN (Deng et al. 2017), and many others learn interpretable and meaningful representations from images by either maximizing the information or promoting independence between the latent variables. The fundamental drawback of these approaches is that they fail in the case of complex datasets since the generation quality degrades as they learn to disentangle. To alleviate this problem, recent works such as (Shen et al. 2020), (Voynov and Babenko 2020) discover interpretable directions directly from the latent space of pre-trained GANs. (Voynov and Babenko 2020) performs unsupervised learning to identify distinguishable directions while (Härkönen et al. 2020), (Spingarn-Eliezer, Banner, and Michaeli 2020) and (Shen and Zhou 2021) obtains directions analytically. These directions need not be completely disentangled.

We propose Scale Ranking Estimator(SRE), a model learned via **self-supervision** strategy to enhance disentanglement in the directions derived by current posthoc disentanglement approaches. Self-supervision is a successful training paradigm for deep learning models that allows them to learn in a label-efficient manner. In essence, SRE enhances disentanglement by enforcing the order of variation within each transformation. Our method is independent of the GAN architecture used. We perform extensive qualitative and quantitative analysis on synthetic and natural datasets to show that the proposed method improves the disentanglement of existing directions. SRE learns to encode the magnitude of variation in each direction. We demonstrate a practical application where these encodings can be directly used for Attribute-based image retrieval task.

Related Work

Generative Adversarial Networks GANs are one of the most popular generative models that shows promising results on image synthesis (Goodfellow et al. 2014). It consists of a Generator and Discriminator that learns in an adversarial setting. Recent variants of GANs such as StyleGAN (Karras, Laine, and Aila 2019), StyleGAN-2 (Karras et al. 2020), Progressive GAN (Karras et al. 2018) and BigGAN (Brock, Donahue, and Simonyan 2019) are shown to be very successful in generating high-resolution images. Progressive GAN, a successor of conventional GAN, attempts to generate high-resolution images by progressively growing the generator and discriminator. StyleGAN and StyleGAN-2 learn a mapping network that maps the z-space to w-space that is more disentangled.

Post-hoc Disentanglement from pre-trained GANs (Higgins et al. 2017), (Dupont 2018), (Lin et al. 2020) e.t.c. disentangle factors of variations very well in synthetic datasets, but they fail to do so in complex natural datasets. These classical disentanglement learning techniques improve disentanglement at the cost of generation quality. To overcome this limitation, extensive research has been conducted in the field of learning interpretable directions from pre-trained models. They can be categorized into three based on the learning paradigm used :

- **Supervised** : (Bau et al. 2019) computes the agreement

[*]These authors contributed equally.

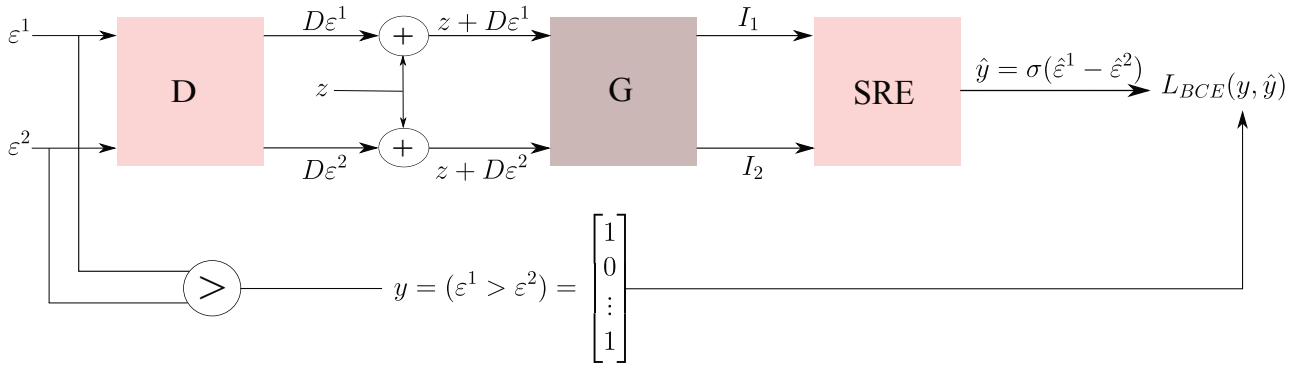

Figure 1: Illustration of the proposed approach; D is initialized with existing post-hoc disentanglement directions. We first compute two linear combinations of directions in D, where the coefficients are values in scale vectors ε^1 and ε^2, respectively. These linear combinations are then added to latent code z, which gives a pair of shifted latent codes. Generator G outputs a pair of images which are passed to the SRE. SRE decodes the scale vectors, $\hat{\varepsilon}^1$ and $\hat{\varepsilon}^2$ from the pair of images. Binary cross-entropy loss is computed based on the difference between the predicted scale vectors and pseudo-ground truth labels. Pseudo-ground truth labels are the original pairwise ordering between the values in ε^1 and ε^2.

between the output of a pre-trained semantic segmentation network and the spatial location of the unit activation map to identify the concept encoded in each unit. (Shen et al. 2020) and (Yang, Shen, and Zhou 2021) use off-the-shelf classifiers to discover interpretable directions in the latent space. A conditional normalizing flow version of (Shen et al. 2020) and (Yang, Shen, and Zhou 2021) is explored in (Abdal et al. 2021). The main limitation of the above approaches is that they require pre-trained networks, which may not be available for complex transformations.

- **Unsupervised** : (Voynov and Babenko 2020) discovers interpretable directions in an unsupervised manner by jointly updating a candidate direction matrix and reconstructor that predicts the perturbed direction. (Peebles et al. 2020) proposes a regularization term that forces the Hessian of a generative model with respect to its input to be diagonal. However, such methods require training. (Härkönen et al. 2020) observed that applying PCA on the latent space of Style-GAN and BigGAN retrieves human-interpretable directions. (Shen and Zhou 2021) and (Spingarn-Eliezer, Banner, and Michaeli 2020) obtained a closed-form solution by extracting the interpretable directions from the weight matrices of pre-trained generators. These methods are computationally inexpensive since they do not require any form of training. (Voynov and Babenko 2020) and (Peebles et al. 2020) attempts to learn directions that are easily distinguishable while (Shen and Zhou 2021), (Spingarn-Eliezer, Banner, and Michaeli 2020) and (Härkönen et al. 2020) finds directions of maximum variance. However, none of these approaches ensure that only a single factor of variation gets captured in a transformation. Our method addresses this problem by defining a self-supervision task that promotes disentanglement on directions captured by these methods.
- **Self-supervised** : (Jahanian, Chai, and Isola 2020) and (Plumerault, Borgne, and Hudelot 2020) make use of user-specified simple transformations as a source of self-supervision to learn corresponding directions. The main drawback of these approaches is that, user-specified edits are hard to obtain for complex transformations. Unlike these methods, our method relies on transformations discovered by unsupervised methods and hence can discover a wide variety of disentangled transformations.

Proposed Method

Firstly, we provide the intuition behind our approach. In an entangled transformation, formulating a task that favors the dominant factor of variation will enhance the dominant factor in it. To achieve this, we propose Scale Ranking Estimator (SRE), a neural network that learns to rank the scale of each transformation in generated images. Imposing a ranking on the magnitude of variation in each direction would hopefully force the SRE to distinguish between the factors of variation in the associated transformation and thus capture the dominant factor of variation. The directions could then be updated based on this knowledge. An illustration of the proposed approach is given in Figure 1.

We formally define all the components involved in our training scheme. Let $G : Z \rightarrow I$ be the pre-trained generator, where Z is the latent space and I represents the pixel space. Interpretable directions are discovered from the latent space of generator G. Let $D \in \mathbb{R}^{k \times d}$ denote the matrix whose columns correspond to interpretable directions in latent space. k and d are the latent space dimensionality and the number of interpretable directions, respectively. We also define a neural network $SRE(i; \theta), i \in I$ that outputs the scale of transformation corresponding to each direction in D. D and SRE are the trainable components in our approach, while the parameters of G are non-trainable.

Training scheme

We initialize D with a set of directions obtained from any post-hoc disentanglement method. A linear walk in the la-

tent space is given by $\hat{z} \rightarrow z + D\varepsilon$, where $D\varepsilon$ is the linear combination of directions in D. $\varepsilon = (\varepsilon_1, \varepsilon_2, .., \varepsilon_d) \in \mathbb{R}^d$, where $\varepsilon_i \sim U(-e, e)$ represents the scale of corresponding direction. We sample $\varepsilon^1, \varepsilon^2 \in \mathbb{R}^d$ to generate images $G(\hat{z_1})$ and $G(\hat{z_2})$, where $\hat{z_1} \rightarrow z + D\varepsilon^1$ and $\hat{z_2} \rightarrow z + D\varepsilon^2$. These images are then passed to SRE which predicts $\hat{\varepsilon}^1$ and $\hat{\varepsilon}^2$ based on the information encoded in the generated images.

The loss function to be minimized is as follows :

$$L = \mathop{\mathbb{E}}_{\substack{z \sim N(0,1) \\ \varepsilon^1, \varepsilon^2 \sim U(-e,e)}} \sum_{i=1}^{d} L_{BCE}(y_i, \hat{y}_i), \tag{1}$$

where,

$$\hat{y}_i = \sigma(\hat{\varepsilon_i}^1 - \hat{\varepsilon_i}^2),$$

$$y_i = \begin{cases} 1, & \text{if } \varepsilon_i^1 > \varepsilon_i^2, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Here, L_{BCE} is the binary cross-entropy loss between the predicted output and the pseudo ground-truth, y_i is determined by comparing the scale of transformation used to generate the images as shown in (2). We provide self-supervision using the knowledge already present in the initialized direction matrix to update it further.

We perform weight updates on D and SRE in an alternate fashion. There are two optimization steps in each training iteration. Firstly, we compute the loss as specified in (1) to update the weights of SRE by freezing the weights of direction matrix D. In the subsequent step, we use the updated SRE to recalculate the loss as in (1). The parameters of SRE are now freezed to update D. Training in this manner helps continually transfer some of the information from SRE to D and vice-versa. This is critical since the initialization of SRE is random, whereas the initialization of D is partially learned directions.

As discussed above, ε is sampled from a multivariate uniform distribution with parameters e and $-e$. If the specified range $(-e, e)$ is relatively small, our method becomes highly constrained, making it hard to capture the variations in a disentangled factor. If the range is set very wide, the model has the freedom to allow a lot of variation in the transformation, which can cause it to stay entangled. As a result, determining the correct values for the hyper-parameter e is crucial for improved disentanglement.

Experimental Details

This section discusses the datasets used, pre-trained generators corresponding to each dataset, choice of initialization for D, and hyperparameters involved.

Datasets

We perform the experiments on following datasets:

- **CelebA-HQ** (Karras et al. 2018) consist of 30,000, 1024×1024 resolution images of Celebrity faces.
- **AnimeFaces dataset** (Jin et al. 2017) consist of 64×64 resolution face images of Anime characters.

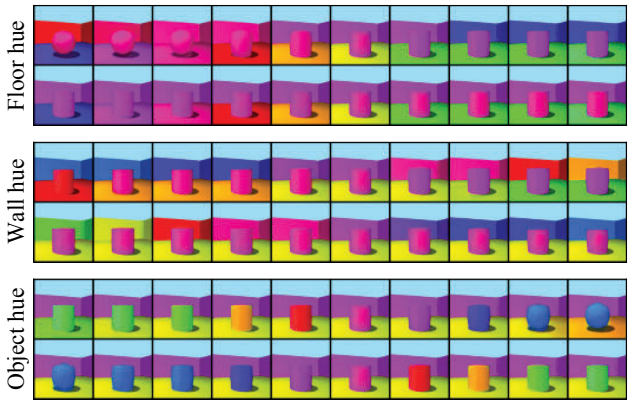

Figure 2: Latent traversal corresponding to Floor hue, Wall hue, Object hue on 3DShapes. For each attribute, the first row corresponds to SeFa and the second row corresponds to SeFa + SRE.

- **LSUN-Cars** (Yu et al. 2015) consist of 512×512 resolution images of cars.
- **3D Shapes** (Burgess and Kim 2018) containing 480,000 images of 64×64 resolution synthetic images with 6 factors of variation.

Pre-trained Generators

We use four different variants of GAN for our experiments to show that our method is independent of the GAN architecture used. **PGGAN** (Karras et al. 2018) is used for generating samples from CelebA-HQ dataset. As a representative of conventional GANs, we use **Spectral Norm GAN** (Miyato et al. 2018) to generate Anime Faces. **StyleGAN**, (Karras, Laine, and Aila 2019) and **StyleGAN-2** (Karras et al. 2020) are used for LSUN-Cars and 3DShapes dataset, respectively. We used the same pre-trained generators that are used in (Voynov and Babenko 2020) and (Shen and Zhou 2021).

Initialization

We mainly use two contrasting post-hoc disentanglement algorithms to obtain initialization for the direction matrix D. As a sampling-based initialisation, we consider **SeFa** (Shen and Zhou 2021) because it does not require any form of training, whereas the other initialization used is based on directions learned by **LD** (Voynov and Babenko 2020), which requires learning to obtain interpretable directions. We show that our method enhances the disentanglement of any set of directions regardless of the paradigm used to generate it, be it sampling or learning. We used the implementation released by the authors of (Voynov and Babenko 2020) and (Shen and Zhou 2021) to derive the directions for initialization on all the datasets.

Hyperparameters

- **Architecture** : For all the four datasets, We utilize ResNet-18 model (He et al. 2016) for SRE while D is

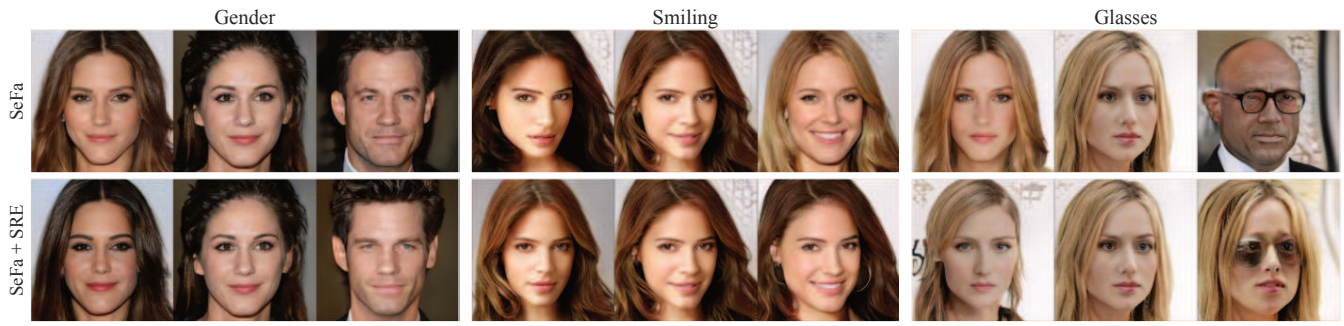

Figure 3: Comparison of latent traversals obtained by SeFa and SeFa + SRE on CelebA-HQ dataset. For each attribute, middle image represents the original image, image to the left and right represents the source image manipulated in positive and negative directions, respectively.

	Pose	Gender	Age	Smile	Glasses
Pose	0.70	0.33	0.22	0.06	0.04
Gender	0.09	0.63	0.76	0.20	0.04
Age	0.09	0.22	0.58	0.27	0.01
Smile	0.17	0.23	0.40	0.43	0.04
Glasses	0.09	0.63	0.76	0.20	0.04

	Pose	Gender	Age	Smile	Glasses
Pose	0.69	0.13	0.20	0.10	0.04
Gender	0.15	0.63	0.06	0.07	0.04
Age	0.15	0.02	0.46	0.16	0.00
Smile	0.05	0.01	0.05	0.43	0.04
Glasses	0.06	0.06	0.07	0.01	0.20

Figure 4: Rescoring matrix obtained for SeFa (Left) and SeFa + SRE (Right) on CelebA-HQ dataset. Each row represents an attribute obtained by moving in the relevant direction, and the column corresponds to attribute predictors used to compute the scores.

a simple linear operator. We discovered that ensuring orthogonality between directions in D in each iteration resulted in better disentanglement.

- **Number of iterations** : We set number of iterations to be 6000 for 3DShapes and 20000 for all other datasets. 3DShapes requires relatively lesser number of iterations since it is a synthetic dataset.
- **Optimization** : We use Adam optimizer to optimize both D and SRE. The learning rate is set to 0.0001. Batch size is 64 for 3DShapes and 8 in the case of CelebA-HQ dataset. For all other datasets, batch size is set to 16.

Results

In this section, we discuss the qualitative and quantitative results for each of the datasets. SeFa + SRE and LD + SRE correspond to our approach where D is initialized with SeFa and LD directions, respectively. We compare the performance of SeFa + SRE with SeFa and LD + SRE with LD directions.

Qualitative Analysis

We conducted a thorough qualitative analysis to evaluate the performance of our proposed approach. Firstly, we analyze the performance of SeFa compared to SeFa + SRE on 3DShapes, CelebA, and LSUN Cars datasets. We plot the latent traversal starting with the original image by traversing in opposite sides along the relevant directions. The range of ε is set from -10 to 10. Figure 2 shows the qualitative results on Shapes3D for three different attributes, floor hue, wall hue, and object hue. It can be observed that floor hue is entangled with object shape and wall hue in the case of SeFa directions while our method improves these directions by disentangling floor hue from the other attributes. A similar trend can be seen in the case of wall hue which is entangled with floor hue in SeFa. More latent traversals on 3DShapes are available in Technical Appendix. Qualitative analysis on CelebA-HQ and LSUN Cars dataset is demonstrated in Figure 3 and Figure 7. Each transformation corresponding to SeFa is entangled with one or more attributes. In Figure 7, car type and color are entangled with zoom in case of SeFa. However, SeFa + SRE disentangles these attributes from zoom. Similarly, SeFa entangles zoom with orientation whereas our method preserves zoom by removing the effect of orientation. Qualitative analysis shows that SRE applied on SeFa initialization disentangles SeFa directions.

We also analyze the directions obtained by LD and SRE applied on LD initialization as shown in Figure 5 and 6. Even though LD seeks distinguishable directions, it can be

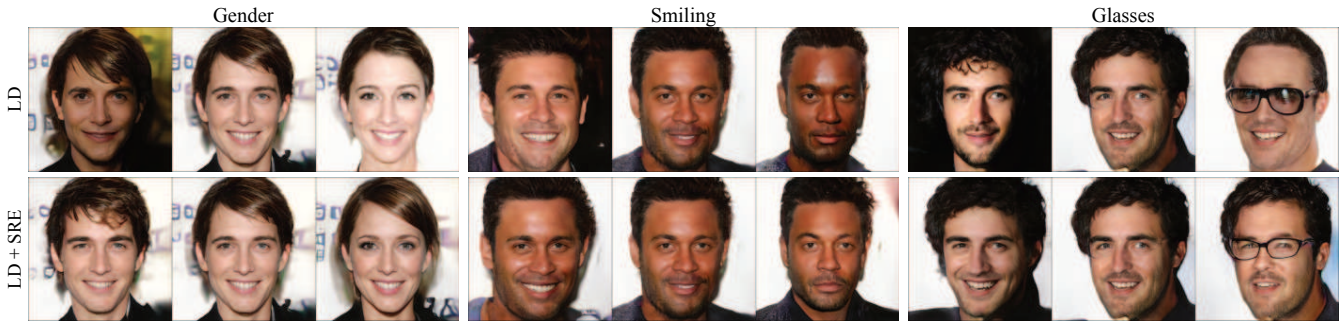

Figure 5: Comparison of latent traversals obtained by LD and LD + SRE on CelebA-HQ dataset. For each attribute, middle image represents the original image, image to the left and right represents the source image manipulated in positive and negative directions, respectively.

Figure 6: Comparison of latent traversals obtained by LD and LD + SRE on AnimeFaces dataset. For each attribute, middle image represents the original image, image to the left and right represents the source image manipulated in positive and negative directions, respectively.

seen that the transformations obtained are quite entangled on both CelebA-HQ and AnimeFaces datasets. We noticed that these transformations are less identity-preserving which is reflected in the Identity preservation accuracies shown in Table 1. Qualitative analysis shows that our approach based on SRE updates the directions so that it results in disentangled and identity-preserving transformations. Additional latent traversal on various datasets and directions are provided in the Technical Appendix.

Quantitative Analysis

We perform quantitative analysis on CelebA-HQ and 3DShapes to evaluate the proposed approach. The two quantitative metrics that we employed to analyze the performance on CelebA-HQ dataset are Rescoring Analysis and Identity Preservation accuracy. We use pre-trained attribute predictors released by the authors of (Yuxuan Han and Fu 2021) to perform rescoring analysis. These attribute predictors are binary classifiers trained on each of the 40 attributes of CelebA dataset (Liu et al. 2015). We perform rescoring analysis similar to that described in (Shen et al. 2020). We take a random sample of 2000 generated images and manipulate them in the direction of the desired attribute. The pre-trained attribute predictors are then used to obtain predictions for the original and altered images. We subsequently compute the absolute value of the difference between the predictions produced for the original image and the manipulated image. The rescoring for the selected direction is computed by taking the mean of this metric across images. Figure 4 shows the rescoring matrix corresponding to SeFa and SeFa + SRE. It can be seen that, when applied on SeFa initialization, SRE better disentangles each of the five attributes compared to SeFa. This analysis supports the qualitative analysis discussed in the previous section. The directions updated by SRE retain the knowledge of individual attributes while reducing the entanglement with other attributes. As observed in the rescoring matrix, SeFa fails to capture Eyeglasses. However, there are directions in SeFa that encodes eyeglasses as the dominant factor. By dominant factor, we mean that it is dominant compared to other factors while the magnitude of the variation is less. SRE disentangles eyeglasses better in one of these directions, which is an interesting observation. This shows that SRE can disentangle factors of finer variation that the initialization struggles to capture. We provide a summary of rescoring to compare the performance of our approach with SeFa and LD. Since the rescoring matrix should be close to the diagonal matrix in case of ideal disentanglement, we compute the ratio of the sum of squares of diagonal elements to that of the off-diagonal elements. Higher the value, better the disentanglement. These values are reported in Table 1.

Identity preservation accuracy is also computed to see

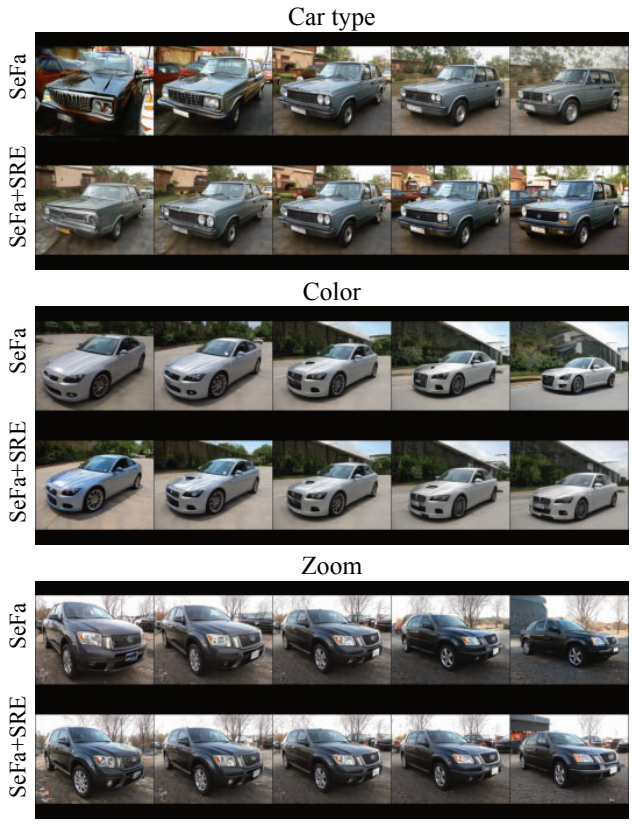

Figure 7: Comparison of latent traversals obtained by SeFa and SeFa + SRE on LSUN Cars dataset. For each attribute, middle image represents the original image.

how effectively SRE retains identity while enhancing disentanglement. We randomly sample 2000 generated images and edit them in the desired direction to obtain the manipulated images. The pair of images are fed into the face recognition model given by (Geitgey 2018), which returns a binary value indicating whether the faces are similar or not. We repeat this procedure in three different directions for all the methods to compute average Identity preservation accuracy. Table 1 summarizes these values. Results suggest that SRE implicitly learns to preserve identity as it learns to disentangle. We believe that our model learns to incorporate smoothness while learning a ranking function on the scale of transformation which helps it to preserve identity.

We also perform quantitative evaluation on 3DShapes since the ground truth factors are readily available. We train SRE and the baselines using seven pre-trained StyleGAN generators for each random seed. Training is done for five different random seeds. We calculate Mutual Information Gap (MIG) (Chen et al. 2018) and Factor-VAE (Kim and Mnih 2018), which are two widely used disentanglement metrics in the literature. This is done by computing the latent space embeddings for real samples by training a GAN inversion network as in (Khrulkov et al. 2021). The evaluation metrics are computed on real samples to analyze the performance of disentangled directions. The mean and standard deviation of the metrics across the models trained for different seeds are reported in Table 2. Both the MIG and Factor-VAE metrics show that SRE outperforms the baselines. Applying SRE on top of SeFa and LD directions increases average MIG by 104.54% and 50% respectively. Results for additional metrics such as DCI (Eastwood and Williams 2018) and β-VAE (Higgins et al. 2017) can be found in the Technical Appendix.

The use of directions derived by post-hoc disentanglement algorithms as initialization is motivated by the fact that some of these (SeFa, GANSpace e.t.c.) are computationally cheap to obtain. However, we also performed extensive experimentation on 3DShapes and CelebA-HQ to evaluate the performance of our approach using random initialization. In comparison to SRE with existing post-hoc disentanglement initialization, we observe that SRE with random initialization requires more iterations to converge and a higher learning rate to optimize the direction matrix D. As shown in Table 3, SRE with random initialization performs reasonably well at the expense of longer training time. We also observe a similar trend in the case of CelebA-HQ (Rescoring value : **6.722**).

We perform real image editing using the directions obtained by SRE. We first approximate the latent vector in the latent space for real images using the GAN inversion paradigm mentioned in (Tov et al. 2021) and then shift the latent vector in the direction of desired attributes to obtain the manipulated images. Qualitative results are provided in Figure 8 which suggests that the SRE directions can also be applied to real images.

Figure 8: Results for Real image editing with respect to multiple attributes in LSUN Cars dataset.

Method	Rescoring(↑)	Identity Preservation Accuracy(↑)
SeFa	0.64	0.61
SeFa + SRE	**7.77**	**0.97**
LD	1.43	0.73
LD + SRE	**3.73**	**0.94**

Table 1: Comparison of Quantitative metrics on CelebA-HQ dataset.

Effect of epsilon (ε)

We devise an ablation to study the effect of ε on training of SRE. We consider three ranges of ε : (-1, 1), (-3, 3), (-10, 10) with all the other parameters fixed. The results are summarized in Table 4. SRE is able to disentangle factors of

Method	MIG(↑)	Factor-VAE Score(↑)
SeFa	0.22±0.01	0.86 ±0.01
SeFa + SRE	**0.45±0.06**	**0.94±0.02**
LD	0.14±0.05	0.78±0.06
LD + SRE	**0.21±0.05**	**0.90±0.05**

Table 2: Comparison of Quantitative metrics on 3DShapes dataset.

Initialization	MIG(↑)	Iterations
Random	0.34 ± 0.04	28000
SeFa	**0.45±0.06**	6000

Table 3: Quantitative metrics of SRE on different initializations averaged across 5 different random seeds.

variation with ε range set to (-1,1). As ε is progressively increased, MIG shows a declining trend. Restricting the range of ε forces the model to accommodate factors that take relatively lesser number of variations, hence forcing the representation to be disentangled. An ε with larger range provides flexibility to accommodate entangled factors, whereas extremely less range of ε will not have sufficient values to properly accommodate variation within a single feature, thus failing to learn any factors of variation properly in both the cases.

Attribute-based Image Retrieval

This section demonstrates an immediate practical application of the learned SRE where it can be directly used for Attribute-based image retrieval task. As we already discussed, our approach updates the initialization and enhances disentanglement in the directions. During the training phase, the Scale Ranking Estimator is updated as well to aid the whole learning process. This section explores the possibility of using the trained SRE for Attribute-based image retrieval without any kind of task specific retraining or fine-tuning.

Given a query image and a specific attribute, our goal is to retrieve images from a pool of real images similar to the query image based on the specified attribute. Attribute-based image retrieval could be of great use in Reverse image search, Person Identification e.t.c. We provide qualitative evidence to show that the SRE that comes as a by-product of our training process can be used for attribute-based image retrieval without any additional training. During the inference, given any image, SRE outputs a vector representation where each value at index i represents the scale or the amount of variation of attribute encoded by the direction at index i in the learned direction matrix D.

We first get the output representations from SRE for all the pool images and the query image. For a given attribute, we obtain the attribute-specific variation for all images by accessing the index corresponding to the direction that encodes the given attribute in the output representations. Pool images are sorted in an ascending order based on the Euclidean distance between their attribute-specific variation to that of the query image. Top K images from the sorted set are K images from the pool most similar to the query image with respect to the specified attribute.

Range of ε	MIG(↑)
(-1, 1)	0.31
(-3, 3)	0.26
(-10, 10)	0.18

Table 4: Effect of ε on performance of SRE.

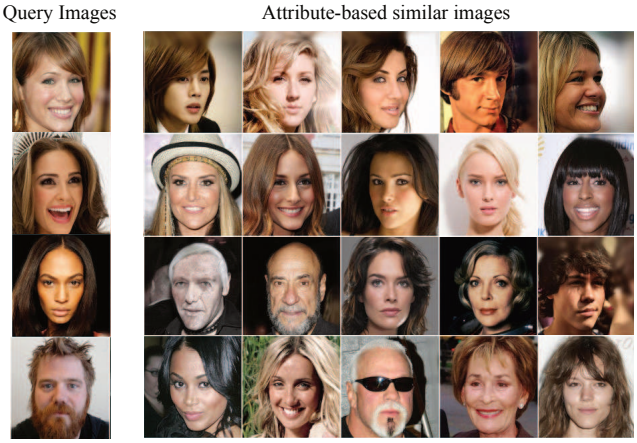

Figure 9: Results for Attribute-based image retrieval task on CelebA-HQ dataset. Each row corresponds to attributes Pose, Gender, Expression and Age respectively.

We empirically demonstrate the performance of Attribute-based image retrieval on CelebA-HQ data set in Figure 9. We considered SRE model trained using SeFa initialization for the analysis. We set $K = 5$ for all the attributes. The attributes considered are Pose, Gender, Expression, and Age. According to qualitative results, SRE performs well on the Attribute-based retrieval task, although it is not explicitly trained to do so.

Conclusion & Future work

We propose a new method for improving disentanglement and interpretability in the directions obtained by existing post-hoc disentanglement methods by learning the Scale Ranking Estimator (SRE). We also provide a thorough quantitative and qualitative analysis of its performance on various real-world and synthetic datasets. Our approach could be used to improve the disentanglement of any set of existing directions regardless of the underlying algorithm used to obtain them. In addition to enhancing disentanglement, trained SRE can also be used for Attribute-based image retrieval without any task-specific training. Computing a closed-form analytical solution to enforce order on the variation in each transformation would also be helpful to enhance the disentanglement by cutting down the training time. Better quantitative metrics need to be proposed to evaluate post-hoc disentanglement methods on natural datasets which consist of complex attributes.

References

Abdal, R.; Zhu, P.; Mitra, N. J.; and Wonka, P. 2021. StyleFlow: Attribute-Conditioned Exploration of StyleGAN-Generated Images Using Conditional Continuous Normalizing Flows. *ACM Transactions on Graphics*.

Bau, D.; Zhu, J.; Strobelt, H.; Zhou, B.; Tenenbaum, J. B.; Freeman, W. T.; and Torralba, A. 2019. GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. In *International Conference on Learning Representations*.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

Burgess, C.; and Kim, H. 2018. 3D Shapes Dataset. https://github.com/deepmind/3dshapes-dataset/.

Chen, R. T. Q.; Li, X.; Grosse, R. B.; and Duvenaud, D. K. 2018. Isolating Sources of Disentanglement in Variational Autoencoders. In *Advances in Neural Information Processing Systems*.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, 2172–2180.

Deng, Z.; Zhang, H.; Liang, X.; Yang, L.; Xu, S.; Zhu, J.; and Xing, E. P. 2017. Structured Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, 3899–3909.

Dupont, E. 2018. Learning Disentangled Joint Continuous and Discrete Representations. In *Advances in Neural Information Processing Systems*.

Eastwood, C.; and Williams, C. K. I. 2018. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*.

Geitgey, A. 2018. Face Recognition. https://github.com/ageitgey/face_recognition.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*.

Härkönen, E.; Hertzmann, A.; Lehtinen, J.; and Paris, S. 2020. GANSpace: Discovering Interpretable GAN Controls. In *Advances in Neural Information Processing Systems*, 9841–9850.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*.

Jahanian, A.; Chai, L.; and Isola, P. 2020. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*.

Jin, Y.; Zhang, J.; Li, M.; Tian, Y.; Zhu, H.; and Fang, Z. 2017. Towards the Automatic Anime Characters Creation with Generative Adversarial Networks. *CoRR*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4396–4405.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Khrulkov, V.; Mirvakhabova, L.; Oseledets, I. V.; and Babenko, A. 2021. Disentangled Representations from Non-Disentangled Models. *CoRR*.

Kim, H.; and Mnih, A. 2018. Disentangling by Factorising. In *International Conference on Machine Learning*, 2649–2658.

Lin, Z.; Thekumparampil, K. K.; Fanti, G. C.; and Oh, S. 2020. InfoGAN-CR and ModelCentrality: Self-supervised Model Training and Selection for Disentangling GANs. In *International Conference on Machine Learning*, 6127–6139.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral Normalization for Generative Adversarial Networks. In *International Conference on Learning Representations*.

Peebles, W.; Peebles, J.; Zhu, J.-Y.; Efros, A. A.; and Torralba, A. 2020. The Hessian Penalty: A Weak Prior for Unsupervised Disentanglement. In *Proceedings of European Conference on Computer Vision*.

Plumerault, A.; Borgne, H. L.; and Hudelot, C. 2020. Controlling generative models with continuous factors of variations. In *International Conference on Learning Representations*.

Shen, Y.; Yang, C.; Tang, X.; and Zhou, B. 2020. InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Shen, Y.; and Zhou, B. 2021. Closed-Form Factorization of Latent Semantics in GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Spingarn-Eliezer, N.; Banner, R.; and Michaeli, T. 2020. GAN Steerability without optimization. *CoRR*.

Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; and Cohen-Or, D. 2021. Designing an Encoder for StyleGAN Image Manipulation. *ACM Transactions on Graphics*.

Voynov, A.; and Babenko, A. 2020. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, 9786–9796.

Yang, C.; Shen, Y.; and Zhou, B. 2021. Semantic Hierarchy Emerges in Deep Generative Representations for Scene Synthesis. *International Journal of Computer Vision*, 1451–1466.

Yu, F.; Zhang, Y.; Song, S.; Seff, A.; and Xiao, J. 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *CoRR*.

Yuxuan Han, J. Y.; and Fu, Y. 2021. Disentangled Face Attribute Editing via Instance-Aware Latent Space Search. In *International Joint Conference on Artificial Intelligence*.