ADIS-GAN: AFFINE DISENTANGLED GAN

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper proposes Affine Disentangled GAN (ADIS-GAN), which is a Generative Adversarial Network that can explicitly disentangle affine transformations in a self-supervised and rigorous manner. The objective is inspired by InfoGAN, where an additional affine regularizer acts as the inductive bias. The affine regularizer is rooted in the affine transformation properties of images, changing some properties of the underlying images, while leaving all other properties invariant. We derive the affine regularizer by decomposing the affine matrix into separate transformation matrices and inferring the transformation parameters by maximum likelihood estimation. Unlike the disentangled representations learned by existing approaches, the features learned by ADIS-GAN are axis-aligned and scalable, where transformations such as rotation, horizontal and vertical zoom, horizontal and vertical skew, horizontal and vertical translation can be explicitly selected and learned. ADIS-GAN successfully disentangles these features on the MNIST, CelebA, and dSprites datasets.

1 INTRODUCTION

In a disentangled representation, observations are interpreted in terms of a few explanatory factors. Examples of such factors are the rotation angle(s), scale, or position of an object in an image. Disentangled variables are generally considered as the abstraction of interpretable semantic information and reflection of separatable factors of variation in the data. Many studies have explored the effectiveness of disentangled representations Bengio et al. (2013); N et al. (2017); LeCun et al. (2015); Lake et al. (2017); Tschannen et al. (2018). The information presented in observations is encoded in an interpretable and compact manner, e.g., the texture style and the orientation of the objects. Bengio et al. (2013); LeCun et al. (2015); Lake et al. (2017); Tschannen et al. (2018). The learned representation are more generalizable and can be useful for downstream tasks, such as classification and visualization Bengio et al. (2013); N et al. (2017); Chen et al. (2016).

The concept of disentangled representation has been defined in several ways in the literature Locatello et al. (2019); Higgins et al. (2018); Eastwood & Williams (2018). The necessity of explicit inductive biases both for learning approaches and the datasets is discussed in Locatello et al. (2019). Inductive bias refers to the set of assumptions that the learner uses to predict outputs of given inputs that it has not encountered. For instance, in the dSprites dataset objects are displayed at different angles and positions; such prior knowledge helps to detect and classify the objects. However, the inductive biases in existing deep learning models are mostly implicit. The proposed ADIS-GAN utilizes relative affine transformations (see Section 4) as the explicit inductive bias, leading to axis-aligned and scalable disentangled representations.

The issue of axis-alignment is addressed in Higgins et al. (2018), where each latent dimension should have a pre-defined unique axis-alignment. Without axis-alignment, the features learned by disentangled representation need to be identified with expert knowledge after the training, which could be a cumbersome process when dealing with a large number of features. The axis-alignment property also helps to discover non-dominant attributes (e.g., roll angle of face in CelebA dataset).

The scalability property allows us to make a trade-off between the compactness and expressivity of the disentangled representation. For example, the zoom attribute can be decomposed as horizontal and vertical zoom. A more compact representation encodes the zoom attribute by one latent dimension, while a more expressive representation decomposes the zoom attribute as horizontal and vertical zoom, encoded by two latent dimensions. The scalability property provides a solution to an

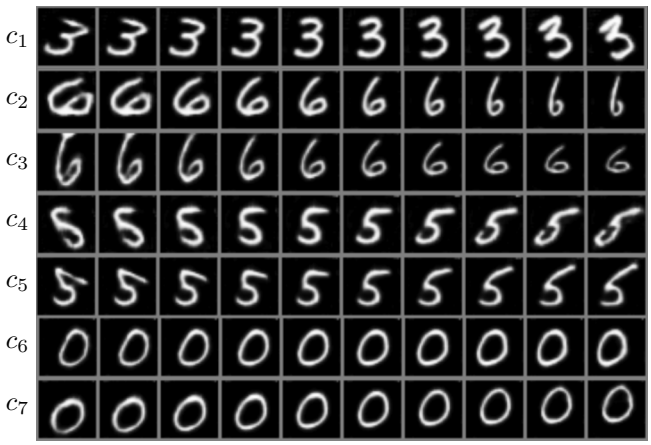

Figure 1: Disentangled representation of affine transformations on the MNIST dataset. c_1: rotation, c_2: horizontal zoom, c_3: vertical zoom, c_4: horizontal skew, c_5: vertical skew, c_6: horizontal translation, c_7: vertical translation. To the best of our knowledge, ADIS-GAN is the first algorithm that can disentangle an entire affine transformation in a self-supervised manner.

open question related to disentangled representations: Should we capture a data generative factor (e.g., zoom) by one or multiple latent dimensions? Many disentanglement metrics Higgins et al. (2017); Chen et al. (2018); Kim & Mnih (2018); Eastwood & Williams (2018) rely on a single latent dimension, while others Denton & Birodkar (2017); Ridgeway & Mozer (2018); Higgins et al. (2018) encode latent factors via multiple latent dimensions.

We motivate the importance of axis-alignment and scalability in particular for affine transformations (see Figure 1), where disentangling object poses from texture and shape is an attractive property of an algorithm in the imaging domain Jaderberg et al. (2015); Bepler et al. (2019); Engstrom et al. (2019). In supervised learning tasks, *Spatial Transformer Network* Jaderberg et al. (2015) can actively spatially transform an image by providing a proper affine transformation matrix. In unsupervised learning tasks, few algorithms have successfully disentangled the affine transformation. In Bepler et al. (2019), an algorithm is introduced that disentangles rotation and translation but not an entire affine transformation.

In this work, we propose ADIS-GAN, which is a Generative Adversarial Network that utilizes the affine regularizer (see Section 4) as an inductive bias to explicitly disentangle the affine transformation. The affine regularizer is rooted in the affine transformation properties of the images, that affect certain properties of the underlying images, while leaving all other properties invariant. We derive the affine regularizer by decomposing the affine matrix into separate transformations and inferring the transformation parameters by maximum likelihood estimation. Unlike the disentangled representations learned by existing approaches, the features learned by ADIS-GAN are axis-aligned and scalable, where transformations such as rotation, horizontal and vertical zoom, horizontal and vertical skew, horizontal and vertical translation can be explicitly selected and learned (see Figure 1).

In the remainder of the paper, we review related work in Section 2. We then compare the difference between GAN, InfoGAN and the proposed method in Section 3. We introduce the ADIS-GAN in Section 4, while in Section 5, we show numerical results showing the axis-aligned and scalable disentangled representation learned by ADIS-GAN. We offer concluding remarks in Section 6.

Our contributions:

1. To the best of our knowledge, ADIS-GAN is the first algorithm that can disentangle an entire affine transformation, including rotation, horizontal and vertical zoom, horizontal and vertical skew, horizontal and vertical translation in a self-supervised manner.

2. The disentangled representations obtained by ADIS-GAN are axis-aligned, where the representations learned by existing algorithms are not. The advantages are two-folds:

(a) The attributes to be learned can be pre-defined, which saves the effort to identify the attributes after the training.
(b) It is possible to extract attributes that are non-dominant, in parallel with the dominant attributes.

3. The disentangled representations obtained by ADIS-GAN are scalable, where the representations learned by existing algorithms are not. The scalability property makes it possible to make a trade-off between the compactness and expressivity of the learned representation.

2 RELATED LITERATURE

Recent approaches learning disentangled representations are largely based on *Variational Autoencoders (VAEs)* Kingma & Welling (2013) and *InfoGAN* Chen et al. (2016). In a standard *VAE*, observation x is encoded to latent representation z from variational distribution $Q(z|x)$ using a deep neural network. In the generative process, the latent variable z is sampled from a prior distribution $P(z)$ and uses another deep neural network to reconstruct the observation x from conditional distribution $P(x|z)$. To achieve the disentanglement, a factorized aggregated posterior $\int_x Q(z|x)P(x)dx$ is encouraged during the training. In a standard *GAN* Goodfellow et al. (2014), latent representation z is sampled from a prior distribution $P(z)$. The fake data x_{fake} is generated by z from distribution $P(x|z)$. A discriminator is introduced to encourage the generated data x_{fake} to be close to the observation x_{real}. To achieve disentanglement, *InfoGAN* Chen et al. (2016) proposes to maximize the mutual information between a subset c of latent representation z and the generated data x_{fake}.

Recently, much attention has been paid to regularizers that promote disentanglement. The β-VAE Higgins et al. (2017) encourages the disentanglement by increasing the weight of the KL regularizer, thus promoting the factorization of the posterior $Q(z|x)$. Both FactorVAE Kim & Mnih (2018) and β-TCVAE Chen et al. (2018) penalize the total correlation, while the former relies on adversarial training and the latter directly calculates the total correlation through the decomposition of the β-VAE objective function. The DIP-VAE Kumar et al. (2018) penalizes the difference between the aggregated posterior and a factorized prior. In the AnnealedVAE Huang et al. (2018), the encoder can concentrate on learning individual factors and variation by gradually increasing the bottleneck capacity. The IB-GAN Jeon et al. (2019) is an extension to InfoGAN rooted in information bottleneck theory, which includes a mutual information upper bound and forms a mutual information bottleneck. The InfoGAN-CR Lin et al. (2019) adds a contrastive regularizer on top on InfoGAN, that compares the changes between the image and latent space.

Even though the aforementioned methods have achieved better disentanglement performance compared to the baseline established by VAE and InfoGAN, none of them have successfully disentangled an entire affine transformations in a scalable and axis-aligned way, which is a desirable property in the imaging domain. Moreover, our affine regularizer is orthogonal to those approaches, thus makes it possible to integrate our affine regularizer with other methods.

In Gidaris et al. (2018); Chen et al. (2019); Wang et al. (2020); Zhang et al. (2019), self-supervised regularization is applied, where they compare the difference of images before and after the affine/projective transformation. The transformation loss is define as: $L = ||M(\theta') - M(\theta)||_2^2$ and a parameterized matrix $M \in \mathbb{R}^{3\times 3}$. We make a more specific assumption about the matrix and further decompose it to achieve disentanglement (see Section 4).

3 BACKGROUND: GAN, INFOGAN AND ADIS-GAN

In a standard *GAN* Goodfellow et al. (2014), latent representation z is sampled from a prior distribution $P(z)$. The fake data x_{fake} is generated from z from conditional distribution $P(x|z)$. A discriminator is introduced to encourage the generated data x_{fake} to be close to the observation x_{real}:

$$\min_G \max_D V(D, G) = L_{\text{adv}}(D, G). \tag{1}$$

To achieve disentanglement, *InfoGAN* Chen et al. (2016) maximizes the mutual information between a subset c of latent representation z and the generated data x_{fake}:

$$\min_G \max_D V(D, G) = L_{\text{adv}}(D, G) - \lambda I(c'; G(c, z)). \tag{2}$$

To explicitly disentangle affine transformation attributes in a scalable and axis-aligned manner, ADIS-GAN has one additional term called affine regularizer on top of InfoGAN:

$$\min_G \max_D V(D,G) = L_{\text{adv}}(D,G) - \lambda I(c'; G(c,z)) - L_{\text{affine}}. \tag{3}$$

4 AFFINE REGULARIZER

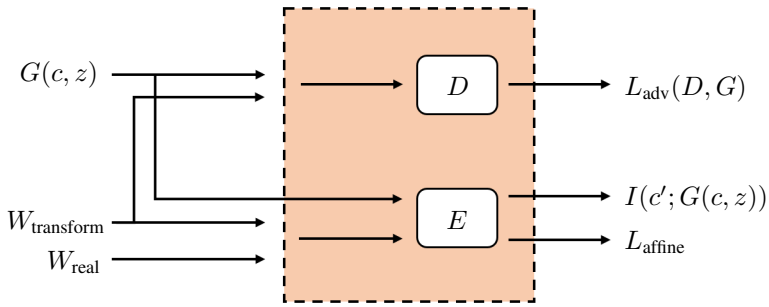

Figure 2: The main framework of ADIS-GAN.

4.1 RELATIVE AFFINE TRANSFORMATION

We assume, without loss of generality, that images can be expressed as the multiplication of the affine transformation $M_{\text{real}}^{\text{basis}} \in \mathbb{R}^{3\times 3}$ and an image basis W_{basis} (see Figure 3):

$$M_{\text{real}}^{\text{basis}} \times W_{\text{basis}} = W_{\text{real}}. \tag{4}$$

The superscript of $M_{\text{real}}^{\text{basis}}$ stands for the initial state and subscript stands for the transformed state of the image W. The image basis W_{basis} is the average manifold of all images within the same category. For instance, the digits "0" , "1", ... "9" in MNIST are different categories.

If we purposely apply a known affine transformation $M_{\text{transform}}^{\text{real}}$ on a sampled image W_{real}, we can obtain an new image $W_{\text{transform}}$:

$$M_{\text{transform}}^{\text{real}} \times W_{\text{real}} = W_{\text{transform}}. \tag{5}$$

According to the definition in equation 4, the transformed image can also be expressed as the multiplication of the $M_{\text{transform}}^{\text{basis}}$ and the image base W_{basis}:

$$M_{\text{transform}}^{\text{basis}} \times W_{\text{basis}} = W_{\text{transform}}. \tag{6}$$

By solving the simultaneous formula from equation 4 to 6:

$$M_{\text{transform}}^{\text{real}} = M_{\text{transform}}^{\text{basis}} \times M_{\text{real}}^{\text{basis}^{-1}}. \tag{7}$$

$M_{\text{transform}}^{\text{real}}$ is the **relative affine transformation** between affine transformation $M_{\text{transform}}^{\text{basis}}$ and $M_{\text{real}}^{\text{basis}}$. The affine regularizer loss is defined as: $L_{\text{affine}} = ||M_{\text{transform}}^{\text{real}} - M_{\text{transform}}^{\text{basis}} \times M_{\text{real}}^{\text{basis}^{-1}}||_2^2$.

4.2 AFFINE TRANSFORMATION DECOMPOSITION

Since M is an affine transformation, it can be decomposed as the multiplication of individual affine transformations M_i . Here we select rotation, horizontal and vertical zoom, horizontal and vertical translation for illustration. Nevertheless, more or fewer attributes can be selected since our disentangled representation is scalable. The affine transformation can be decomposed as:

$$M = M_r \times M_p \times M_q \times M_x \times M_y. \tag{8}$$

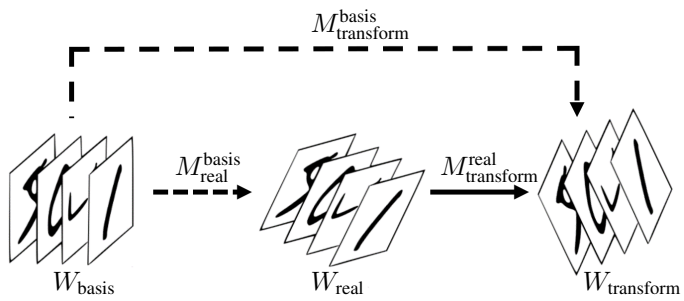

Figure 3: Illustration of relative affine transformation. Solid line stands for the affine transformation between two real images. Dashed line stands for the affine transformation between one real image and the imaginary image basis.

We may take $C = (C_1, C_2, C_3, C_4, C_5)$ as our latent representations according to the affine transformation decomposition (see equation 8). Due to the particular data structure of the affine matrix (see equation 9), we can specifically take $C = (C_r, C_p, C_q, C_x, C_y)$ as our latent representation and derive a relation (see equation 10) between affine matrix M and latent representation C despite the non-commutativity of the multiplication (see Appendix B). Since we can explicitly define the components of the affine matrix (e.g., zoom or decomposed horizontal and vertical zoom), the learned representation is scalable. The axis-alignment property is achieved by assigning the defined attributes to specific latent dimensions. In the context of affine matrix, equation 8 is equivalent to:

$$M = f_M(c(i)) = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} p & 0 & 0 \\ 0 & q & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & x \\ 0 & 1 & y \\ 0 & 0 & 1 \end{bmatrix}. \tag{9}$$

We can derive expressions for each latent representation through the maximum likelihood estimation (see Appendix A) as follows:

$$f_c(M) = \begin{cases} \theta & = \arctan \frac{2(A_{11}A_{21} - A_{12}A_{22})}{A_{11}^2 + A_{22}^2 - A_{12}^2 - A_{21}^2}, \\ p & = A_{11}\cos\theta + A_{21}\sin\theta, \\ q & = -A_{12}\sin\theta + A_{21}\sin\theta, \\ x & = \frac{A_{13}\cos\theta + A_{23}\sin\theta}{p}, \\ y & = \frac{-A_{13}\sin\theta + A_{23}\cos\theta}{q}. \end{cases} \tag{10}$$

The affine regularizer loss is defined as: $L_{\text{affine}} = \sum_i ||C(i)_{\text{transform}}^{\text{real}} - C'(i)_{\text{transform}}^{\text{real}}||_2^2$, $C(i)_{\text{transform}}^{\text{real}} \sim$ Unif(-1,1), $C'(i)_{\text{transform}}^{\text{real}} = \text{Norm}_{\text{AL}}(f(M_{\text{transform}}^{\text{basis}} \times M_{\text{real}}^{\text{basis}^{-1}}))$, $i \in \{\theta, p, q, x, y\}$ (see Algorithm 1).

5 Numerical Results

The goal of the experiments in this section is to investigate, both qualitatively and quantitatively, the disentangled representations obtained by ADIS-GAN. The axis-alignment of the disentangled representations is demonstrated on CelebA dataset Liu et al. (2015), while the scalability of the disentangled representation is illustrated on MNIST dataset. Next quantitative results for ADIS-GAN are presented and compared to benchmarks on the dSprites Matthey et al. (2017) dataset. The parameters of the affine transformation are selected as follows: rotation range: $[-15°, 15°]$, zoom range: [0.8, 1.2], and translation range: [0.8, 1.2].

Algorithm 1: Affine Regularizer

Input: Sampled images: W_{real}, latent vectors: $C(i)^{\text{real}}_{\text{transform}} \sim$ Unif(-1,1), $i \in \{\theta, p, q, x, y\}$
Output: L_{affine}
Latent to affine normalization (LA): $c(i)^{\text{real}}_{\text{transform}}$ = $\text{Norm}_{\text{AL}}(C(i)^{\text{real}}_{\text{transform}})$ [1]
E.g., $\theta \in [-\pi/10, \pi/10] \rightarrow c(\theta)^{\text{real}}_{\text{transform}}$ = $C(\theta)^{\text{real}}_{\text{transform}} \cdot \pi/10$
Affine matrix initialization: $M^{\text{real}}_{\text{transform}}$ = $f_M(c(i)^{\text{real}}_{\text{transform}})$, $i \in \{\theta, p, q, x, y\}$ using equation 9
Affine transformation: $M^{\text{real}}_{\text{transform}} \times W_{\text{real}} = W_{\text{transform}}$
Encoding from images: $M^{\text{basis}}_{\text{transform}}$ = Encoder($W_{\text{transform}}$), $M^{\text{basis}}_{\text{real}}$ = Encoder(W_{real})
Maximum likelihood estimation: $c'(i)^{\text{real}}_{\text{transform}}$ = $f_c(M^{\text{basis}}_{\text{transform}} \times M^{\text{basis}^{-1}}_{\text{real}})$ using equation 10
Affine to latent normalization (AL): $C'(i)^{\text{real}}_{\text{transform}} = \text{Norm}_{\text{LA}}(c'(i)^{\text{real}}_{\text{transform}})$
E.g., $\theta \in [-\pi/10, \pi/10] \rightarrow C'(\theta)^{\text{real}}_{\text{transform}}$ = $c'(\theta)^{\text{real}}_{\text{transform}} \cdot 10/\pi$
Affine reuglarizer loss: $L_{\text{affine}} = \sum_i ||C(i)^{\text{real}}_{\text{transform}} - C'(i)^{\text{real}}_{\text{transform}}||^2_2$, $i \in \{\theta, p, q, x, y\}$

5.1 Qualitative Results

5.1.1 Axis-alignment

In CelebA dataset, typical dominant attributes are azimuth, sunglasses, emotion, etc. Existing methods Chen et al. (2016); Kim & Mnih (2018); Chen et al. (2018) have successfully disentangled those attributes (see Appendix E). However, non-dominant attributes such as the roll, width and length of the face and relative position of face in the frame are rarely tackled. Thanks to the axis-aligned property, ADIS-GAN can explicitly learn those non-dominant attributes (see Figure 4, 5 and 6). It is interesting to note that ADIS-GAN does not rigidly perform affine transformations on the given images but understands the relationship between different facial components. For example, when we try to shorten the length of the face (vertical zoom and translation, see left images on both Figure 5 and 6), ADIS-GAN knows how to complete the bottom of the images with a neck to make the images look realistic. To disentangle facial attributes on the CelebA data, we choose to model the latent space with one categorical code, $c_{\text{cat}} \sim \text{Cat}(K = 10, p = 0.1)$ and 5 continuous codes $c_{\text{cont}} \sim$ Unif(-1,1).

Figure 4: Roll attribute on CelebA dataset.

Figure 5: Vertical (left) and horizontal (right) zoom attributes on CelebA dataset.

[1] $C(i)$: before latent to affine normalization, $c(i)$: after latent to affine normalization.

Figure 6: Vertical (left) and horizontal (right) translation attributes on CelebA dataset.

5.1.2 SCALABILITY

Scalability refers to the flexibility to encode one attribute via one latent vector or, alternatively, to decompose one attribute into sub-attributes and encode them by multiple latent vectors. For example, the zoom attribute can be encoded by one latent vector, or it can be decomposed as horizontal and vertical zoom and encoded by two latent vectors (see Figure 7).

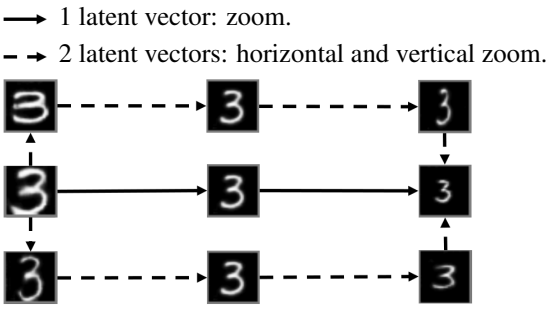

Figure 7: Illustration of scalability. The zoom process indicated in dashed lines is encoded by 2 latent vectors, while the zoom process indicated in solid lines only relies on 1 latent vector. Thanks to the scalable property, we can flexibly switch between the compactness (1 latent vector) and expressivity (2 latent vectors) of the learned representation.

Table 1: Comparison of the axis-alignment and scalability of the disentangled representation learned by different approaches. ADIS-GAN is the only algorithm that generates disentangled representations with both useful properties.

Property	β-VAE	Factor -VAE	β-TCVAE	Annealed -VAE	Info -GAN	Info -GAN-CR	ADIS -GAN
Axis -alignment	no	no	no	no	no	no	yes
Scalability	no	no	no	no	no	no	yes

5.2 QUANTITATIVE RESULTS

To disentangle object style on dSprites, we choose to model the latent space by one categorical code$c_{\text{cat}} \sim \text{Cat}(K = 3, p = 0.33)$, and 4 continuous codes $c_{\text{cont}} \sim \text{Unif}(-1,1)$ that represent rotation, zoom, horizontal and vertical translation. To the best of our knowledge, ADIS-GAN is the first algorithm that can disentangle shape attributes with categorical codes on dSprites datasets. We have observed that shape disentanglement is easier for larger images. We conducted experiments on the "Four Shapes" dataset (see Appendix D). Compared to dSprites, the major difference of the "Four Shapes" dataset is that the images are larger. We found that ADIS-GAN can successfully disentangle shape attribute on "Four Shapes" dataset with categorical codes (see Appendix D). The image enlargement procedures are described in Appendix C.

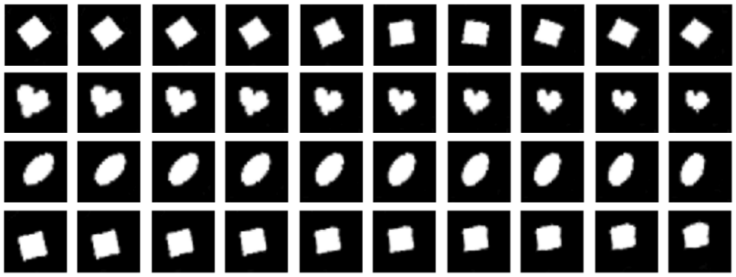

Figure 8: Disentangled representation on the dSprites dataset. The affine transformation attributes (continuous) are axis-aligned from row 1 to 4: rotation, zoom, horizontal translation, and vertical translation. The latent traversal interval is [-1,1] for continuous code. The categorical attributes from row 1 to 4: square, heart, ellipse, and square.

Table 2: Disentanglement scores on the dSprites dataset. The reference values are from Lin et al. (2019) Table 1. Note that we do not apply model centrality here, which is a hyperparameter tuning method, as our focus is on axis-alignment and scalability. For the FactorVAE and BetaVAE scores, we only use continuous codes. For the other scores, we use both continuous and categorical codes. Overall, ADIS-GAN is comparable with the state-of-the-art disentanglement algorithms on the dSprites dataset. We refer to Appendix F for the definations of disentanglement metrics.

	Model	**FactorVAE**	**DCI**	**Explicitness**	**Modularity**	**MIG**	**BetaVAE**
	VAE	$0.63 \pm .06$	$0.30 \pm .10$			0.10	
	β-TCVAE	$0.62 \pm .07$	$0.29 \pm .10$			0.45	
	HFVAE	$0.63 \pm .08$	$0.39 \pm .16$				
	β-VAE	$0.63 \pm .10$	$0.41 \pm .11$		0.21		
	ChyVAE	0.77					
VAE	DIP-VAE		0.53				
	FactorVAE	0.82				0.15	
	FactorVAE (1.0)	$0.79 \pm .01$	$0.67 \pm .03$	$0.78 \pm .01$	$0.79 \pm .01$	$0.27 \pm .03$	$0.79 \pm .02$
	FactorVAE (10.0)	$0.83 \pm .01$	$0.70 \pm .02$	$0.79 \pm .00$	$0.79 \pm .00$	$0.40 \pm .01$	$0.83 \pm .01$
	FactorVAE (20.0)	$0.83 \pm .01$	$0.72 \pm .02$	$0.79 \pm .00$	$0.79 \pm .01$	$0.40 \pm .01$	$0.85 \pm .00$
	FactorVAE (40.0)	$0.82 \pm .01$	$\mathbf{0.74 \pm .01}$	$0.79 \pm .00$	$0.77 \pm .01$	$0.43 \pm .01$	$0.84 \pm .01$
	InfoGAN	$0.59 \pm .70$	$0.41 \pm .05$		0.05		
	IB-GAN	$0.80 \pm .07$	$0.67 \pm .07$				
GAN	InfoGAN(modified)	$0.82 \pm .01$	$0.60 \pm .02$	$0.82 \pm .00$	$0.94 \pm .01$	$0.22 \pm .01$	$0.87 \pm .01$
	InfoGAN-CR	$\mathbf{0.88 \pm .01}$	$0.71 \pm .01$	$\mathbf{0.85 \pm .00}$	$\mathbf{0.96 \pm .00}$	$0.37 \pm .01$	$\mathbf{0.95 \pm .01}$
	ADIS-GAN	$0.86 \pm .02$	$0.71 \pm .02$	$0.78 \pm .00$	$0.95 \pm .01$	$\mathbf{0.46 \pm .01}$	$0.89 \pm .01$

6 Conclusion

This paper introduces the Affine Disentangled GAN (ADIS-GAN) that can explicitly learn the affine transformations via disentangled representations. In contrast to earlier approaches to disentanglement, where inductive biases are not explicit, the disentangled representation obtained by ADIS-GAN is axis-aligned and scalable. The affine regularizer introduced in this paper can be applied to other methods such as the VAE family Kingma & Welling (2013); Higgins et al. (2017); Kim & Mnih (2018); Chen et al. (2018).

Another attractive property of the affine regularizer is that it supports the possibility to utilize expert knowledge as the inductive bias, as it is model based and exploits maximum likelihood estimation. For example, we can disentangle the 2D affine transformations with the decomposed relationships. Task-specific explicit regularizers may provide an alternative pathway for disentanglement compared to existing general implicit regularizers.

REFERENCES

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.50.

Tristan Bepler, Ellen Zhong, Kotaro Kelley, Edward Brignole, and Bonnie Berger. Explicitly disentangling image content from translation and rotation with spatial-vae. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 15435–15445. Curran Associates, Inc., 2019.

Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 2610–2620. Curran Associates, Inc., 2018.

Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 12154–12163. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.01243.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2172–2180. Curran Associates, Inc., 2016.

Emily L Denton and vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4414–4423. Curran Associates, Inc., 2017.

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

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1802–1811, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014.

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

Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loïc Matthey, Danilo J. Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *CoRR*, abs/1812.02230, 2018.

Chin-Wei Huang, Shawn Tan, Alexandre Lacoste, and Aaron C Courville. Improving explorability in variational inference with annealed variational objectives. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 9701–9711. Curran Associates, Inc., 2018.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 2017–2025. Curran Associates, Inc., 2015.

Insu Jeon, Wonkwang Lee, and Gunhee Kim. IB-GAN: Disentangled representation learning with information bottleneck GAN, 2019.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. VARIATIONAL INFERENCE OF DISENTANGLED LATENT CONCEPTS FROM UNLABELED OBSERVATIONS. In *International Conference on Learning Representations*, 2018.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. doi: 10.1017/S0140525X16001837.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

Zinan Lin, Kiran Koshy Thekumparampil, Giulia C. Fanti, and Sewoong Oh. Infogan-cr: Disentangling generative adversarial networks with contrastive regularizers. *CoRR*, abs/1906.06034, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Francesco Locatello, Stefan Bauer, Mario Lučić, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Frederic Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019. Best Paper Award.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Siddharth N, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning disentangled representations with semi-supervised deep generative models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5925–5935. Curran Associates, Inc., 2017.

Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 31*, pp. 185–194. Curran Associates, Inc., 2018.

Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *CoRR*, abs/1812.05069, 2018.

Jiayu Wang, Wengang Zhou, Guo-Jun Qi, Zhongqian Fu, Qi Tian, and Houqiang Li. Transformation gan for unsupervised image synthesis and representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. AET vs. AED: unsupervised representation learning by auto-encoding transformations rather than data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00265.