Orthogonal Jacobian Regularization for Unsupervised Disentanglement in Image Generation

Yuxiang Wei[1], Yupeng Shi[1], Xiao Liu[2], Zhilong Ji[2], Yuan Gao[2], Zhongqin Wu[2], Wangmeng Zuo[1,3]
[1] Harbin Institute of Technology, [2] Tomorrow Advancing Life, [3] Pazhou Lab, Guangzhou

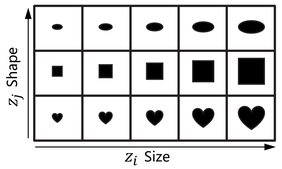
Paper: https://arxiv.org/abs/2108.07668 Project: https://github.com/csyxwei/OroJaR

2021 ICCV OCTOBER 11-17 VIRTUAL

Motivation

Disentanglement:

Each dimension corresponds to the change in one factor of variation (FoV), while being independent to changes in other factors.



Intuition:

When perturbing a single dimension of the network input, the change in the output should be independent (and also uncorrelated) with those caused by the other input dimensions.

Our Solution:

- We propose an **Ortho**gonal **Ja**cobian **R**egularization (**OroJaR**) to encourage generative model to learn disentangled representations.
- Jacobian matrix is used to represent the change caused by the latent input.
- Jacobian vector of each dimension is constrained to be orthogonal to encourage disentanglement.

Method

Orthogonal Jacobian Regularization:

- To encourage disentanglement, we constrain the Jacobian vectors of different latent dimensions to be orthogonal,

$$\left[\frac{\partial G_d}{\partial z_i}\right]^T \frac{\partial G_d}{\partial z_j} = 0. \tag{1}$$

- Taking all latent dimensions into account, we present the Orthogonal Jacobian Regularization (OroJaR):

$$\mathcal{L}_J(G) = \sum_{d=1}^{D} \|\mathbf{J}_d^T \mathbf{J}_d \circ (\mathbf{1} - \mathbf{I})\| = \sum_{d=1}^{D} \sum_{i=1}^{m} \sum_{j \neq i} \left| \left[\frac{\partial G_d}{\partial z_i}\right]^T \frac{\partial G_d}{\partial z_j} \right|^2. \tag{2}$$

Approximation for Accelerated Training:

- According to Hutchinson's estimator, Eqn. (2) can be rewrite as:

$$\mathcal{L}_J(G) = \sum_{d=1}^{D} \mathrm{Var}_{\mathbf{v}} \left[\mathbf{v}^T (\mathbf{j}_d^T \mathbf{j}_d) \mathbf{v}\right] = \sum_{d=1}^{D} \mathrm{Var}_{\mathbf{v}} \left[(\mathbf{j}_d \mathbf{v})^T \mathbf{j}_d \mathbf{v}\right]. \tag{3}$$

- $\mathbf{j}_d \mathbf{v}$ can be further efficiently computed by a first-order finite difference approximation:

$$\mathbf{j}_d \mathbf{v} = \frac{1}{\epsilon}[G(\mathbf{z} + \epsilon \mathbf{v}) - G(\mathbf{z})]. \tag{4}$$

Applications to GANs

Training from scratch:

$$\mathcal{L}_G^{oro} = \underbrace{\mathbb{E}_{\mathbf{z}}[f(1 - D(G(\mathbf{z})))]}_{\text{Standard Adversarial Loss}} + \underbrace{\lambda \mathbb{E}_{\mathbf{z}}[\mathcal{L}_J(G(\mathbf{z}))]}_{\text{OroJaR}}, \tag{5}$$

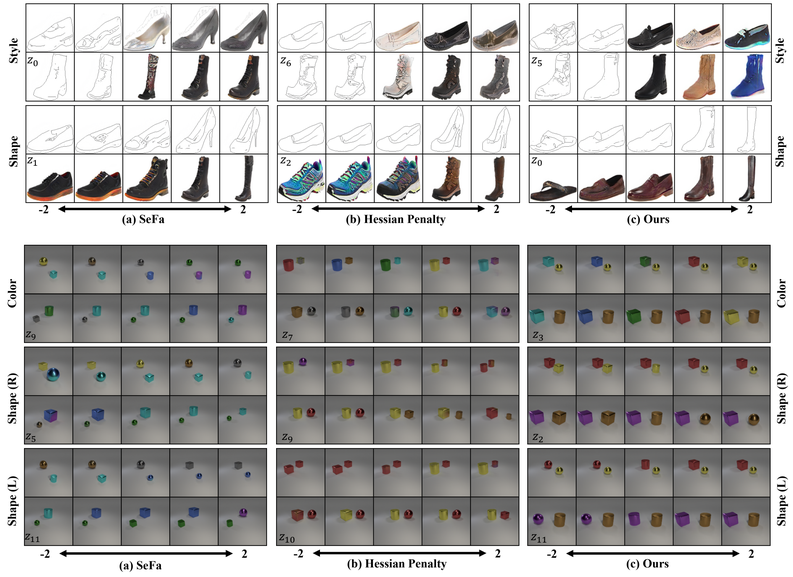
Apply to pretrained GAN:

$$\mathbf{A}^* = \arg\min_{\mathbf{A}} \mathbb{E}_{\mathbf{z},\omega_i} \mathcal{L}_J(G(\mathbf{z} + \eta \mathbf{A} \omega_i)), \tag{6}$$

$\mathbf{A} \in \mathbb{R}^{m \times N}$ is a learnable orthonormal matrix and G is frozen. **OroJaR is now taken *w.r.t* ω_i instead of $\mathbf{z}$.**

Experiments

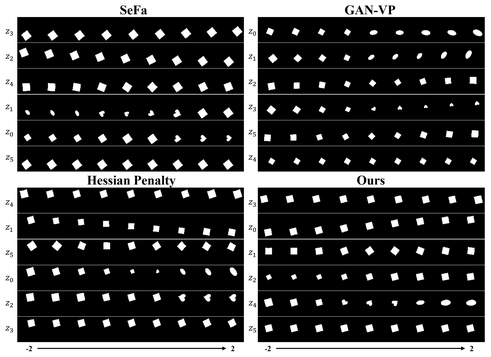
Comparison on Edges+Shoes and CLEVR:



Method	Edges+Shoes			CLEVR-Simple			CLEVR-Complex			CLEVR-U			CLEVR-1FOV		
	PPL (↓)	FID (↓)	VP (↑)	PPL	FID	VP	PPL	FID	VP	PPL	FID	VP	PPL	FID	VP
InfoGAN	2952.2	**10.4**	15.6	56.2	**2.9**	28.7	83.9	**4.2**	27.9	766.7	3.6	40.1	22.1	6.2	-
ProGAN	3154.1	10.8	15.5	64.5	3.8	27.2	84.4	5.5	25.5	697.7	**3.4**	40.2	30.3	9.0	-
SeFa	3154.1	10.8	24.1	64.5	3.8	58.4	84.4	5.5	30.9	697.7	**3.4**	42.0	30.3	9.0	-
Hessian Penalty	554.1	17.3	28.6	39.7	6.1	71.3	74.7	7.1	42.9	61.6	26.8	79.2	20.8	2.3	-
Ours	**236.7**	16.1	**32.3**	**6.7**	4.9	**76.9**	**10.4**	10.7	**48.8**	**40.9**	4.6	**90.7**	**2.8**	**2.1**	-

Experiments

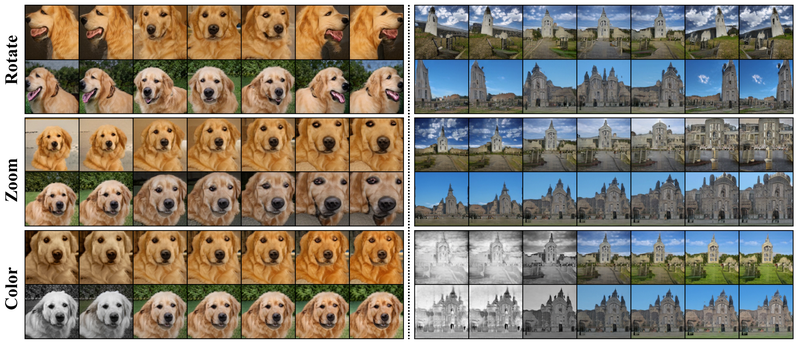
Comparison on Dsprites:



Method	VP(%, ↑)
GAN	30.9 (0.84)
SeFa	48.6 (0.70)
GAN-VP	39.1 (0.48)
Hessian Penalty	48.5 (0.56)
Ours	**54.7 (0.27)**

- Our OroJaR achieves better disentanglement performance.

Discovered directions in pretrained BigGAN:



Ablation Study:

Method	GAN	L0	L1	L2	L3	L4	L0∼1	L0∼2	L0∼3	L0∼4	SeFa	Ours(L0∼3)
VP(%, ↑)	30.9	47.6	48.1	50.2	36.4	35.0	48.8	53.5	**54.7**	52.3	48.6	**54.7**

- The earlier layers are more effective in deactivating the redundant dimensions, resulting in better disentanglement.
- Applying OroJaR to the first multiple layers encourages to learn a better disentangled model.

Conclusion

Experimental results demonstrate that our OroJaR is effective in disentangled and controllable image generation, and performs favorably against the state-of-the-art methods.