Supplementary Material for "Factorized Diffusion Autoencoder for Unsupervised Disentangled Representation Learning"

Abstract

This supplemental accompanies submission "Factorized Diffusion Autoencoder for Unsupervised Disentangled Representation Learning", including more ablation study results, implementation details and visualization results.

More Ablation Study Results

Ablation study results on Shapes3d (Kim and Mnih 2018) and Cars3d (Reed et al. 2015) are shown in Table 1 and Table 2, respectively. "$E + G$" denotes using encoder E and generator G without masks, which is degraded to DiffAE (Preechakul et al. 2022) (concept number $N = 1$) as our baseline method. "CMFNet" denotes the content-mask factorization network. "$\mathcal{L}_{CD}$" and "$\mathcal{L}_{ME}$" denote content decorrelation loss and mask entropy loss, respectively. The conclusions are the same as those on MPI3D (Gondal et al. 2019) in the main manuscript.

More Implementation Details

Details of Diffusion Model Training and inference strategies of the diffusion model were the same as those used by EDM (Karras et al. 2022; Dhariwal and Nichol 2021). Second order Heun (Ascher and Petzold 1998) was adopted as ODE solver. Time step was $t_{i<N_t} = \sigma_{max}^{1/\rho} + \frac{i}{N_t-1}(\sigma_{min}^{1/\rho} - \sigma_{max}^{1/\rho})^\rho$, where $N_t = 1000$, $\sigma_{min} = 0.002$, $\sigma_{max} = 80$ and $\rho = 7$. Noise scale function was $\sigma_t = t$. Loss weighting function was $\lambda(\sigma_t) = 1/\sigma_t^2 + 1/\sigma_{data}^2$, where $\sigma_{data} = 0.5$.

Network Architecture of Mask Decoder D_M The network architecture of mask decoder D_M is shown in Table 3 The architecture is similar to that of the generator of DCGAN (Radford, Metz, and Chintala 2016), except that different types of normalization layers and activation layers are used. The network takes $N \times d_m$ mask codes ($d_m = 80$) as input and outputs $N \times 1 \times 64 \times 64$ masks.

More Visualization Results

Visualization of Masks

Visualization of masks on MPI3D (Gondal et al. 2019) and Shapes3d (Kim and Mnih 2018) are shown in Figure 1 and Figure 2, respectively. It can be observed that, interpretable concepts are extracted by the masks.

Table 1: Ablation study on Shapes3d (Kim and Mnih 2018).

Components	$E + G$	CMFNet	$\mathcal{L}_{CD}$	$\mathcal{L}_{ME}$	DCI	FVAE	MIG
1 (baseline)	✓				0.114	0.432	0.007
2	✓	✓			0.861	0.960	0.433
3	✓	✓	✓		0.863	0.977	0.435
4	✓	✓		✓	0.886	0.976	0.448
5 (full model)	✓	✓	✓	✓	0.917	0.987	0.473

Table 2: Ablation study on Cars3d (Reed et al. 2015).

Components	$E + G$	CMFNet	$\mathcal{L}_{CD}$	$\mathcal{L}_{ME}$	DCI	FVAE	MIG
1 (baseline)	✓				0.307	0.959	0.023
2	✓	✓			0.355	0.888	0.097
3	✓	✓	✓		0.404	0.898	0.109
4	✓	✓		✓	0.417	0.898	0.125
5 (full model)	✓	✓	✓	✓	0.418	0.918	0.137

Table 3: Architecture of mask decoder D_M.

Layer	Parameter	Input size	Output size
Linear	$d_m \times 6144$	$N \times d_m$	$N \times 6144$
Reshape	-	$N \times 6144$	$N \times 384 \times 4 \times 4$
GroupNorm	group num: 32	$N \times 384 \times 4 \times 4$	$N \times 384 \times 4 \times 4$
Upsample	scale: 2	$N \times 384 \times 4 \times 4$	$N \times 384 \times 8 \times 8$
Conv2d	kernel: 3, stride: 1, pad: 1	$N \times 384 \times 8 \times 8$	$N \times 384 \times 8 \times 8$
GroupNorm	group num: 32	$N \times 384 \times 8 \times 8$	$N \times 384 \times 8 \times 8$
SiLU	-	$N \times 384 \times 8 \times 8$	$N \times 384 \times 8 \times 8$
Upsample	scale: 2	$N \times 384 \times 8 \times 8$	$N \times 384 \times 16 \times 16$
Conv2d	kernel: 3, stride: 1, pad: 1	$N \times 384 \times 16 \times 16$	$N \times 256 \times 16 \times 16$
GroupNorm	group num: 32	$N \times 256 \times 16 \times 16$	$N \times 256 \times 16 \times 16$
SiLU	-	$N \times 256 \times 16 \times 16$	$N \times 256 \times 16 \times 16$
Upsample	scale: 2	$N \times 256 \times 16 \times 16$	$N \times 256 \times 32 \times 32$
Conv2d	kernel: 3, stride: 1, pad: 1	$N \times 256 \times 32 \times 32$	$N \times 128 \times 32 \times 32$
GroupNorm	group num: 32	$N \times 128 \times 32 \times 32$	$N \times 128 \times 32 \times 32$
SiLU	-	$N \times 128 \times 32 \times 32$	$N \times 128 \times 32 \times 32$
Upsample	scale: 2	$N \times 128 \times 32 \times 32$	$N \times 128 \times 64 \times 64$
Conv2d	kernel: 3, stride: 1, pad: 1	$N \times 128 \times 64 \times 64$	$N \times 64 \times 64 \times 64$
GroupNorm	group num: 32	$N \times 64 \times 64 \times 64$	$N \times 64 \times 64 \times 64$
SiLU	-	$N \times 64 \times 64 \times 64$	$N \times 64 \times 64 \times 64$
Conv2d	kernel: 3, stride: 1, pad: 1	$N \times 64 \times 64 \times 64$	$N \times 1 \times 64 \times 64$

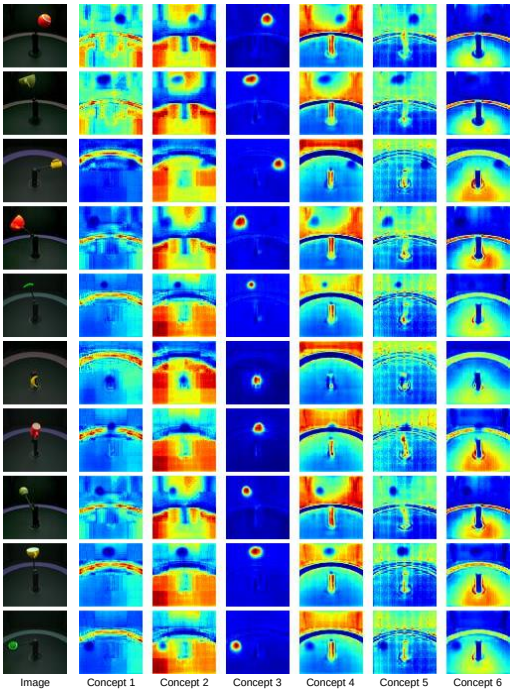

Figure 1: Visualization of masks learned by our FDAE on MPI3D (Gondal et al. 2019).

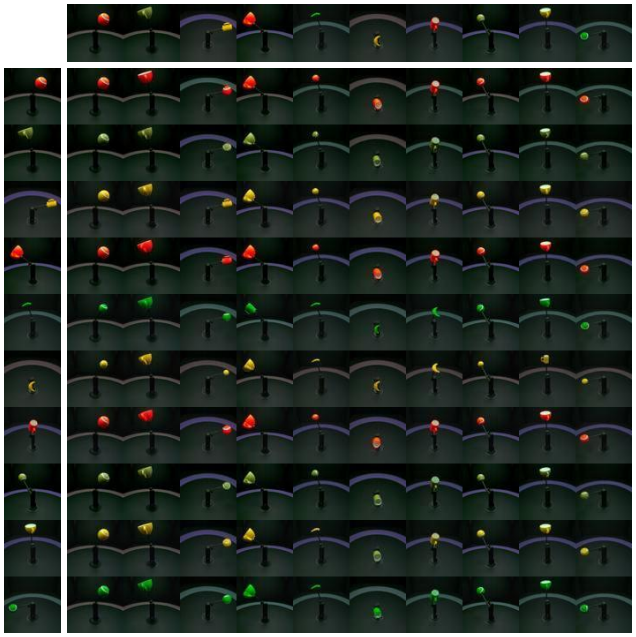

Figure 3: Images generated by swapping content codes and mask codes on MPI3D (Gondal et al. 2019). The image in the i-th row and the j-th column of the box is generated using the content codes from the i-th image on the left and the mask codes from the j-th image on the top.

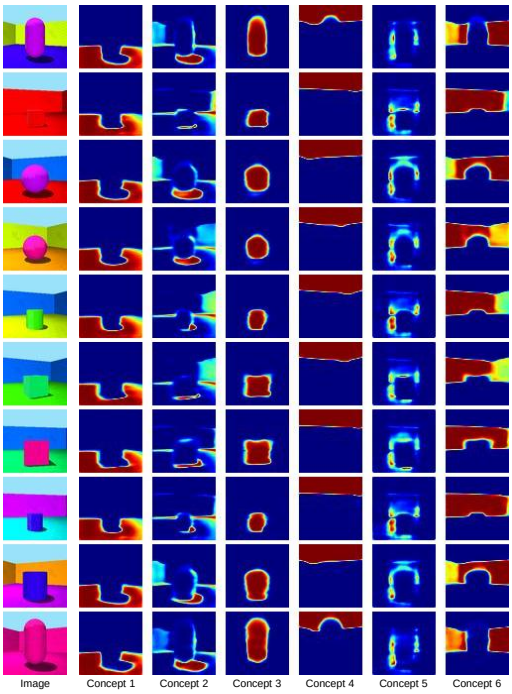

Figure 2: Visualization of masks learned by our FDAE on Shapes3d (Kim and Mnih 2018).

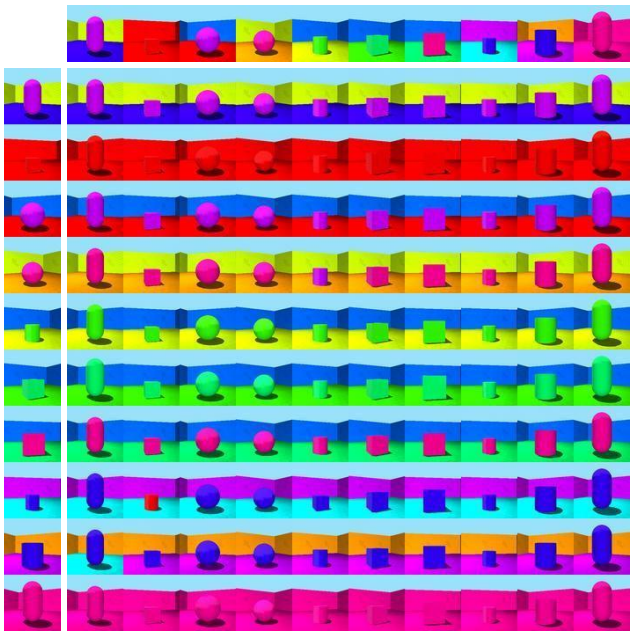

Figure 4: Images generated by swapping content codes and mask codes on Shapes3d (Kim and Mnih 2018). The images are generated as those in Figure 3.

Swapping Content Codes and Mask Codes

In addition to the visualization of swapping content codes and mask codes on Cars3d (Reed et al. 2015) and Market-1501 (Zheng et al. 2015) in the main manuscript, we conducted the same visualization on MPI3D (Gondal et al. 2019) and Shapes3d (Kim and Mnih 2018) in Figure 3 and Figure 4, respectively.

Consistent with the conclusions presented in the main manuscript, our observations indicate that content codes primarily represent appearances, while mask codes predominantly represent shapes and positions. Some failure cases are also shown, such as the croissants in the 5-th and the 6-th columns in Figure 3 and two images in the 8-th and 9-th rows in Figure 4. Although a majority of factors are disentangled as indicated by the quantitative evaluation metrics, there are still some confusions for instances with similar appearances. This demonstrates that more fine-grained representation learning is required in the future work.

References

Ascher, U. M.; and Petzold, L. R. 1998. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. USA: Society for Industrial and Applied Mathematics, 1st edition. ISBN 0898714125.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. *ArXiv*, abs/2105.05233.

Gondal, M. W.; Wuthrich, M.; Miladinovic, D.; Locatello, F.; Breidt, M.; Volchkov, V.; Akpo, J.; Bachem, O.; Schölkopf, B.; and Bauer, S. 2019. On the Transfer of Inductive Bias from Simulation to the Real World: a New Disentanglement Dataset. In *Neural Information Processing Systems (NeuIPS)*, volume 32.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. *ArXiv*, abs/2206.00364.

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

Preechakul, K.; Chatthee, N.; Wizadwongsa, S.; and Suwajanakorn, S. 2022. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10609–10619.

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations (ICLR)*.

Reed, S. E.; Zhang, Y.; Zhang, Y.; and Lee, H. 2015. Deep Visual Analogy-Making. In *Neural Information Processing Systems (NeuIPS)*, volume 28.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision (ICCV)*, 1116–1124.