Factorized Diffusion Autoencoder for Unsupervised Disentangled Representation Learning

Ancong Wu[1], **Wei-Shi Zheng**[1,2,3*]

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou, China
[3] Guangdong Key Laboratory of Information Security Technology, Sun Yat-sen University, Guangzhou, China
wuanc@mail.sysu.edu.cn, zhwshi@mail.sysu.edu.cn

Abstract

Unsupervised disentangled representation learning aims to recover semantically meaningful factors from real-world data without supervision, which is significant for model generalization and interpretability. Current methods mainly rely on assumptions of independence or informativeness of factors, regardless of interpretability. Intuitively, visually interpretable concepts better align with human-defined factors. However, exploiting visual interpretability as inductive bias is still under-explored. Inspired by the observation that most explanatory image factors can be represented by "content + mask", we propose a content-mask factorization network (CMFNet) to decompose an image into different groups of content codes and masks, which are further combined as content masks to represent different visual concepts. To ensure informativeness of the representations, the CMFNet is jointly learned with a generator conditioned on the content masks for reconstructing the input image. The conditional generator employs a diffusion model to leverage its robust distribution modeling capability. Our model is called the Factorized Diffusion Autoencoder (FDAE). To enhance disentanglement of visual concepts, we propose a content decorrelation loss and a mask entropy loss to decorrelate content masks in latent space and spatial space, respectively. Experiments on Shapes3d, MPI3D and Cars3d show that our method achieves advanced performance and can generate visually interpretable concept-specific masks. Source code and supplementary materials are available at https://github.com/wuancong/FDAE.

Introduction

Learning interpretable disentangled representation (Bengio, Courville, and Vincent 2012; Locatello et al. 2019) is fundamental for improving model generalization and interpretability in downstream tasks that require recognition of manually defined factors. Since annotations of the factors are generally unavailable for real-world data, unsupervised disentangled representation learning (Zhu, Xu, and Tao 2021; Voynov and Babenko 2020; Ren et al. 2021; Yang et al. 2022) is a significant field of computer vision.

Most existing methods operate under the assumption that different factors are independent (Higgins et al. 2016) or under the assumption that representations are informative (Chen et al. 2016). However, they rarely take into account the visually interpretable concepts, which intuitively better align with human-defined factors. To uncover interpretable visual concepts, we leverage visually interpretability to introduce inductive bias for representation disentanglement, since Locatello et al. (Locatello et al. 2019) have proven that unsupervised disentanglement is fundamentally impossible without inductive biases on both model and data. Generally, an image is composed of different visual concepts, and most explanatory factors are related to the content and position of these concepts. Figure 1 shows some examples of visual concept factorization, where the concepts are represented by multiple groups of "content + mask".

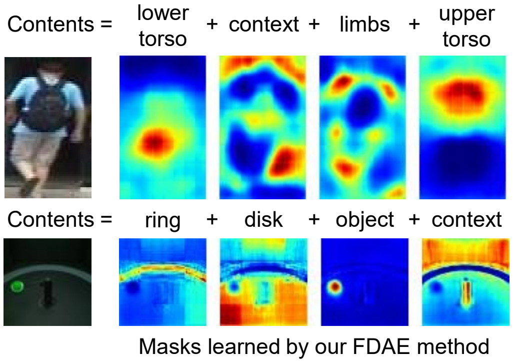

Contents = lower torso + context + limbs + upper torso
Contents = ring + disk + object + context
Masks learned by our FDAE method

Figure 1: Examples of visual concept factorization on Market-1501 (Zheng et al. 2015) and MPI3D (Gondal et al. 2019). Inspired by visual interpretability, we learn disentangled representations by decomposing an image into multiple groups of "content + mask" to uncover visual concepts.

Inspired by this observation, we propose a Content-Mask Factorization Network (CMFNet) that consists of an image encoder and a mask decoder to factorize an image. Given an input image, the image encoder extracts a latent code and splits it into multiple groups of content codes and mask codes to represent different underlying visual concepts. In each group, we aim to learn semantic information in the content code and position information in the mask code. The mask decoder then takes the mask codes as input to generate masks. The content code and mask of each group are aggregated to form a content mask representing a specific

[*] Corresponding author

visual concept. Finally, the content masks of different factors are summed up as a condition map for a conditional image generator, implemented using a diffusion probabilistic model (Karras et al. 2022) to take advantage of its robust distribution modeling ability. To achieve informative and disentangled representation learning, we jointly train the CMFNet and the conditional image generator for image reconstruction, so that the information of the input image is extracted in the condition map and decomposed into different content masks. We call this model the Factorized Diffusion Autoencoder (FDAE). To further enhance disentanglement of different visual concepts, we impose constraints on both the content codes and masks to achieve inter-group content mask disentanglement. A content decorrelation loss and a mask entropy loss are proposed to decorrelate the content masks in latent space and spatial space, respectively.

Our model achieves state-of-the-art results on benchmark datasets Shapes3d (Kim and Mnih 2018), MPI3D (Gondal et al. 2019). Furthermore, our model is capable of generating interpretable masks for understanding the learned concepts.

The main technical contributions of our method are

- We propose the Factorized Diffusion Autoencoder (FDAE), which incorporates the inductive bias of visual interpretability to uncover explanatory visual concepts through content-mask factorization.
- To enhance disentanglement of visual concepts, we introduce constraints of content decorrelation loss and mask entropy loss for multiple groups of content masks in latent space and spatial space, respectively.

Related Work

Disentangled Representation Learning

Disentangled representations (Locatello et al. 2019) should separate the interpretable, independent and informative factors of variations in the data. Most existing approaches concentrate on imposing regularization based on independence (e.g., β-VAE (Higgins et al. 2016)) or informativeness (e.g., InfoGAN (Chen et al. 2016)), while interpretability is still largely ignored. Based on the interpretability assumption, PS-CS model (Zhu, Xu, and Tao 2021) enforces spatial constriction constraint for local feature maps by rectangular masks. Compared with PS-CS model that ignores encoding of the masks, our method separately encodes contents and masks to better facilitate disentanglement and informativeness of content-related factors and position-related factors.

Generally, current approaches explore the latent space of generative models to uncover underlying factors (Voynov and Babenko 2020; Ren et al. 2021). Most methods rely on variational autoencoder (e.g., FactorVAE (Kim and Mnih 2018), β-TCVAE (Chen et al. 2018), DAVA (Estermann and Wattenhofer 2023)) and generative adversarial network (e.g., closed-form latent factorization (Shen and Zhou 2021), GANSpace (Härkönen et al. 2020), DeepSpectral (Khrulkov et al. 2021)). Unsupervised disentanglement method for transformers (e.g., Visual Concepts Tokenization (Yang et al. 2022)) and diffusion models (e.g., DisDiff-VQ (Yang et al. 2023)) are still under-explored. Compared with DisDiff-VQ that relies on independence constraints, our diffusion-based method is inspired by visual interpretability and tends to uncover human-defined factors better.

Diffusion Models for Representation Learning

Diffusion probabilistic model (Ho, Jain, and Abbeel 2020; Rombach et al. 2022) has shown promising image generation ability. Recently, diffusion-based representation learning has attracted increasing attention. PDAE (Zhang, Zhao, and Lin 2022) and Asyrp (Kwon, Jeong, and Uh 2022) discover semantic information in the latent space of pre-trained diffusion models. Diffusion Autoencoders (DiffAE) (Preechakul et al. 2022; Xiang et al. 2023) perform image reconstruction by diffusion model to learn representations.

Based on Diffusion Autoencoders (DiffAE), our method further introduce a content-mask factorization network to uncover disentangled concepts inspired by visual interpretability in an unsupervised manner. This approach is under-explored for diffusion-based representation learning.

Factorized Diffusion Autoencoder

Problem Formulation

For representation disentanglement (Locatello et al. 2019), a real-world image can be assumed to be generated by a two-step process: (1) sampling random variable $\mathbf{z} \in \mathbb{R}^{N_z}$ from a factor distribution $P_{factor}(\mathbf{z})$ and (2) sampling image data $\mathbf{x}$ from conditional data distribution $P_{data}(\mathbf{x}|\mathbf{z})$. Each dimension of $\mathbf{z}$ controls variation of an explanatory factor (e.g., color or position) and is independent of another dimension.

To uncover the underlying factors in unlabeled image $\mathbf{x}$, we expect to learn a concept distribution $Q_{con}(\mathcal{C}|\mathbf{x})$, where $\mathcal{C} = \{\mathbf{C}^1, \mathbf{C}^2, ..., \mathbf{C}^N\}$ is a set of independent visual concepts. On the one hand, $\mathbf{C}^{n_1}$ is informative to predict z_{n_1} and and $\mathbf{C}^{n_1}$ does not contain information of z_{n_2} for $n_1 \neq n_2$. On the other hand, concept set $\mathcal{C}$ should be able to reconstruct $\mathbf{x}$ as factor $\mathbf{z}$, so that we learn $Q_{data}(\mathbf{x}|\mathcal{C})$ to approximate the conditional data distribution $P_{data}(\mathbf{x}|\mathbf{z})$.

To model distributions $Q_{con}(\mathcal{C}|\mathbf{x})$ and $Q_{data}(\mathbf{x}|\mathcal{C})$, we introduce a content-mask factorization network (image encoder E and mask decoder D_M) and a conditional image generator G, respectively. Since joint learning of $Q_{con}(\mathcal{C}|\mathbf{x})$ and $Q_{data}(\mathbf{x}|\mathcal{C})$ can be regarded as image encoding and decoding, our model is called Factorized Diffusion Autoencoder (FDAE), of which the overview is shown in Figure 2.

Content-Mask Factorization Network

To achieve unsupervised disentanglement of interpretable representations, Locatello et al. (Locatello et al. 2019) have theoretically proven that inductive biases are required for both model and data. In this work, we concentrate on learning visual representations and introduce inductive bias for disentanglement inspired by visual interpretability. As shown in Figure 1, when interpreting an image, humans often decompose it into multiple visual concepts, which can typically be represented by their content and position.

Based on such assumption on image data, we take into account the "content + mask" factorization when modeling the concept distribution $Q_{con}(\mathcal{C}|\mathbf{x})$. Each concept $\mathbf{C}^n$ in $\mathcal{C}$ is expected to contain visually interpretable information related

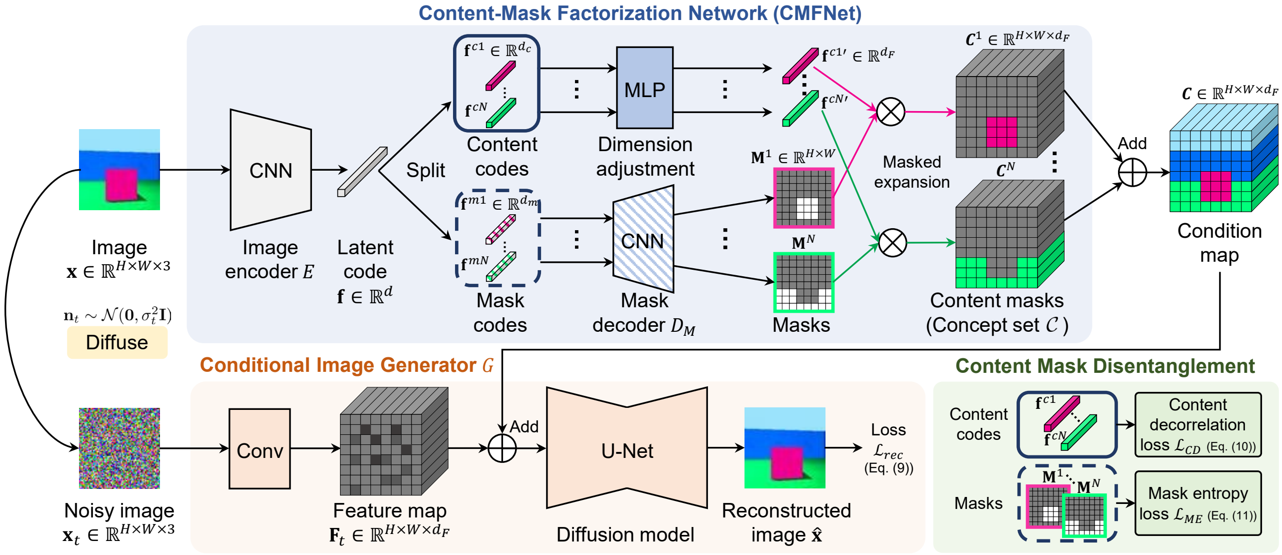

Content-Mask Factorization Network (CMFNet)
Image
$\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$
CNN
Image encoder E
Latent code
$\mathbf{f} \in \mathbb{R}^{d}$
Split
$\mathbf{f}^{c1} \in \mathbb{R}^{d_c}$
$\mathbf{f}^{cN}$
Content codes
$\mathbf{f}^{m1} \in \mathbb{R}^{d_m}$
$\mathbf{f}^{mN}$
Mask codes
MLP
Dimension adjustment
CNN
Mask decoder D_M
$\mathbf{f}^{c1\prime} \in \mathbb{R}^{d_F}$
$\mathbf{f}^{cN\prime}$
$\mathbf{M}^1 \in \mathbb{R}^{H\times W}$
$\mathbf{M}^N$
Masks
Masked expansion
$\boldsymbol{C}^1 \in \mathbb{R}^{H\times W\times d_F}$
$\boldsymbol{C}^N$
Content masks
(Concept set $\mathcal{C}$)
Add
$\boldsymbol{C} \in \mathbb{R}^{H\times W\times d_F}$
Condition map
$\mathbf{n}_t \sim \mathcal{N}(\mathbf{0}, \sigma_t^2\mathbf{I})$
Diffuse
Conditional Image Generator G
Noisy image
$\mathbf{x}_t \in \mathbb{R}^{H\times W\times 3}$
Conv
Feature map
$\mathbf{F}_t \in \mathbb{R}^{H\times W\times d_F}$
Add
U-Net
Diffusion model
Reconstructed image $\hat{\mathbf{x}}$
Loss
$\mathcal{L}_{rec}$
(Eq. (9))
Content Mask Disentanglement
Content codes
$\mathbf{f}^{c1}$
$\mathbf{f}^{cN}$
Content decorrelation loss $\mathcal{L}_{CD}$ (Eq. (10))
Masks
$\mathbf{M}^1$
$\mathbf{M}^N$
Mask entropy loss $\mathcal{L}_{ME}$ (Eq. (11))

Figure 2: The overview of Factorized Diffusion Autoencoder (FDAE). The Content-Mask Factorization Network (CMFNet) encodes the input image $\mathbf{x}$ and factorizes it into N groups of content codes and masks, which correspond to N uncovered visual concepts. Conditioned on the combination of the content masks, the conditional image generator G reconstructs the input image $\mathbf{x}$. Content decorrelation loss $\mathcal{L}_{CD}$ and mask entropy loss $\mathcal{L}_{ME}$ enhance disentanglement of different visual concepts.

to an explanatory factor. To achieve this, we represent a concept $\mathbf{C}^n$ as a combination of a content code and a mask. The content code contains the semantic information of a concept, while the mask contains information about the concept's position and region. To extract content codes and masks, we introduce image encoder E and mask decoder D_M.

Image Encoder E Given input image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$, we employ a convolutional neural network (CNN) as image encoder to extract latent code $\mathbf{f} = E(\mathbf{x})$. To represent multiple visual concepts, the latent code $\mathbf{f}$ is split into N groups of content codes and mask codes by

$$\mathbf{f} = [\overbrace{\mathbf{f}^{c1}, \mathbf{f}^{c2}, ...\mathbf{f}^{cN}}^{N \text{content codes}}, \overbrace{\mathbf{f}^{m1}, \mathbf{f}^{m2}, ...\mathbf{f}^{mN}}^{N \text{mask codes}}], \tag{1}$$

where $\mathbf{f} \in \mathbb{R}^{d}$ is the concatenation of content codes $\mathbf{f}^{c1}, \mathbf{f}^{c2}, ...\mathbf{f}^{cN} \in \mathbb{R}^{d_c}$ and mask codes $\mathbf{f}^{m1}, \mathbf{f}^{m2}, ...\mathbf{f}^{mN} \in \mathbb{R}^{d_m}$. The dimensionality of $\mathbf{f}$ is $d = N(d_c + d_m)$.

Mask Decoder D_M We expect that the position and region information in the mask codes $\mathbf{f}^{mn}$ $(n = 1, 2, ..., N)$ can be explicitly represented by an interpretable mask image. We adopt a CNN with upsampling modules as mask decoder D_M. The mask image $\mathbf{M}^n \in \mathbb{R}^{H\times W}$ is decoded by

$$\begin{aligned} &\mathbf{M}^n_{out} = D_M(\mathbf{f}^{mn}), \\ &\mathbf{M}^n = \text{Softmax}(\mathbf{M}^1_{out}, \mathbf{M}^2_{out}, ..., \mathbf{M}^N_{out})_n, \end{aligned} \tag{2}$$

where the Softmax function is applied to normalize N masks to ensure that the values of mask images are probabilities between 0 and 1.

Content Masks To represent the n-th visual concept in an interpretable manner, we aggregate the content code $\mathbf{f}^{cn}$ and the mask code $\mathbf{M}^n$ to obtain a content mask $\mathbf{C}^n \in \mathbb{R}^{H\times W\times d_F}$, where d_F is the dimensionality that matches the condition input of the conditional image generator G.

To this end, we first map the content code $\mathbf{f}^{cn} \in \mathbb{R}^{d_c}$ to $\mathbf{f}^{cn\prime} \in \mathbb{R}^{d_F}$ by Multi-Layer Perceptron (MLP) to adjust the dimensionality. Next, we perform masked expansion for content code $\mathbf{f}^{cn\prime}$ and mask $\mathbf{M}^n$ to obtain concept $\mathbf{C}^n$ by

$$\mathbf{c}^n_{i,j} = m^n_{i,j} \cdot \mathbf{f}^{cn\prime}, \tag{3}$$

where vector $\mathbf{c}^n_{i,j} \in \mathbb{R}^{d_F}$ and scalar $m^n_{i,j} \in [0, 1]$ are the elements in the i-th row and the j-th column of $\mathbf{C}^n \in \mathbb{R}^{H\times W\times d_F}$ and $\mathbf{M}^n \in \mathbb{R}^{H\times W}$, respectively.

With the cooperation of image encoder E and mask decoder D_M, the input image $\mathbf{x}$ is factorized into a set of N content masks $\mathcal{C} = \{\mathbf{C}^n\}^N_{n=1}$. We call this module the Content-Mask Factorization Network (CMFNet).

Reconstruction by Conditional Diffusion Model

To learn informative representations, the content masks $\mathcal{C}$ extracted by CMFNet should be able to reconstruct $\mathbf{x}$ as the ground-truth factor $\mathbf{z}$ in "Problem Formulation" section. To model data distribution $Q_{data}(\mathbf{x}|\mathcal{C})$, we exploit diffusion probabilistic model (DPM) (Ho, Jain, and Abbeel 2020) as conditional image generator G to leverage its robust distribution modeling ability.

Preliminaries of Diffusion Probabilistic Model (DPM) The denoising diffusion probabilistic model (Ho, Jain, and Abbeel 2020) is a parameterized Markov chain for generating samples by learning to reverse a diffusion process, which gradually adds noise to the data until the signal is destroyed.

Given data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, the forward (diffusion) process gradually adds Gaussian noise to the data according to a variance schedule $\beta_1, ..., \beta_T$ as follows

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{4}$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \tag{5}$$

The backward process is a Markov chain with learned transitions starting at $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ as follows:

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \tag{6}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\sigma}_\theta(\mathbf{x}_t, t)), \tag{7}$$

where θ denotes diffusion model parameters.

Factorized Diffusion Autoencoder (FDAE) Inspired by diffusion autoencoders (Preechakul et al. 2022), joint training of image encoder and DPM conditioned on the latent codes can learn informative representation through image reconstruction. This is because the latent codes are forced to learn information lost during the forward diffusion process (Zhang, Zhao, and Lin 2022).

To represent all uncovered visual concepts by a single variable, we combine the content masks $\mathbf{C}^1, \mathbf{C}^2, ..., \mathbf{C}^N$ to form a condition map $\mathbf{C} = \sum_{n=1}^{N} \mathbf{C}^n$. Then, the condition map $\mathbf{C} \in \mathbb{R}^{H \times W \times d_F}$ is added to intermediate feature map $\mathbf{F}_t \in \mathbb{R}^{H \times W \times d_F}$ of conditional image generator G by

$$\mathbf{F}_t{}' = \mathbf{F}_t + \mathbf{C}, \tag{8}$$

where t is the time step, $\mathbf{F}_t{}'$ is the feature map input to the backbone U-Net (Ronneberger, Fischer, and Brox 2015) of the diffusion probabilistic model (DPM).

For joint learning of CMFNet (E and D_M) and conditional image generator G, we follow the EDM (Karras et al. 2022) approach and adopt the reconstruction loss as follows:

$$\min_{E, D_M, G} \mathcal{L}_{rec} = \mathbb{E}_{\mathbf{x},t} \left[\lambda(\sigma_t) \|G(\mathbf{x}_t, \mathbf{C}, t) - \mathbf{x}\|_2^2\right], \tag{9}$$

where $\mathbf{x} \sim P_{data}(\mathbf{x})$ represents image data. At time step t of the forward process, $\mathbf{x}_t = \mathbf{x} + \mathbf{n}_t$ is the noisy image, where $\mathbf{n}_t \sim \mathcal{N}(\mathbf{0}, \sigma_t^2\mathbf{I})$. σ_t is the noise schedule. $G(\mathbf{x}_t, \mathbf{C}, t)$ is the reconstructed image $\hat{\mathbf{x}}$. $\lambda(\sigma_t)$ is loss weighting function. More details are presented in the supplementary material.

The framework of encoding by content-mask factorization network and decoding by conditional diffusion model is called the Factorized Diffusion Autoencoder (FDAE).

Inter-Group Content Mask Disentanglement

To further enhance disentanglement of the latent code $\mathbf{f}$, we decorrelate content codes $\mathbf{f}^{cn}$ and mask codes $\mathbf{f}^{mn}$ of different groups, since different uncovered visual concepts should be independent to each other as the ground-truth factors $\mathbf{z}$ in "Problem Formulation" section. To accomplish this, we decorrelate the content masks in both the latent space and the spatial space by imposing constraints on content codes $\mathbf{f}^{cn}$ and masks $\mathbf{M}^n$.

Content Decorrelation Loss $\mathcal{L}_{CD}$ In the latent space of content code $\mathbf{f}^{cn}$, we expect that $\mathbf{f}^{cn_1}$ and $\mathbf{f}^{cn_2}$ of different groups are decorrelated for $n_1 \neq n_2$ $(n_1, n_2 \in \{1, ..., N\})$. We compute the inter-group content code covariance $\text{cov}(\mathbf{f}^{cn_1}, \mathbf{f}^{cn_2})$ and make it close to 0. The content decorrelation loss is formulated as

$$\begin{aligned} \min_{E} \mathcal{L}_{CD} &= \sum_{n_1 \neq n_2} |\text{cov}(\mathbf{f}^{cn_1}, \mathbf{f}^{cn_2}) - 0| \\ &= \sum_{n_1 \neq n_2} \left| \frac{1}{N_{tr} - 1} \sum_{s=1}^{N_{tr}} \widetilde{\mathbf{f}_s^{cn_1}}^\top \widetilde{\mathbf{f}_s^{cn_2}} - 0 \right|, \end{aligned} \tag{10}$$

where $\widetilde{\mathbf{f}_s^{cn_1}}, \widetilde{\mathbf{f}_s^{cn_2}}$ are centered content codes normalized by $\ell2$-norm and N_{tr} is the number of training samples.

Mask Entropy Loss $\mathcal{L}_{ME}$ In the mask $\mathbf{M}^n$, the element $m_{i,j}^n$ in the i-th row and the j-th column is the probability that the content of the n-th visual concept appear in pixel (i, j). Due to the Softmax normalization operation in Eq. (2), the elements in mask $\mathbf{M}^n$ satisfy $\sum_{n=1}^{N} m_{i,j}^n = 1$.

We assume that different concepts occupy different regions of the image and should be decorrelated in the spatial space. In the region related to the n-th concept, we expect that $m_{i,j}^n$ is close to 1 and $m_{i,j}^{n'}$ $(n' \neq n)$ is close to 0, which indicates that the uncertainty of $m_{i,j}^1, ..., m_{i,j}^N$ is low. Hence, we introduce the mask entropy loss as follows:

$$\min_{E, D_M} \mathcal{L}_{ME} = \frac{1}{HWN_{tr}} \sum_{i,j,s} \frac{1}{N} \sum_{n=1}^{N} -m_{i,j,s}^n \log(m_{i,j,s}^n), \tag{11}$$

where $m_{i,j,s}^n$ is the value of pixel (i, j) in mask $\mathbf{M}_s^n$ of the s-th sample; H, W are the height and the width of mask $\mathbf{M}_s^n$; N_{tr} is the number of training samples.

Model Training

Loss Function To learn informative and disentangled representations, the total loss function is formulated by

$$\min_{E, D_M, G} \mathcal{L} = \mathcal{L}_{rec} + w_{CD}\mathcal{L}_{CD} + w_{ME}\mathcal{L}_{ME}, \tag{12}$$

where w_{CD} and w_{ME} are trade-off parameters.

Unsupervised Metric for Selecting Concept Number We define an unsupervised metric self-MIG to select the concept number N by mutual information difference between the codes of different concepts in feature $\mathbf{f} = [\mathbf{f}^{c1}, ..., \mathbf{f}^{cN}, \mathbf{f}^{m1}, ..., \mathbf{f}^{mN}]$ (content code $\mathbf{f}^c$ and mask code $\mathbf{f}^m$). Each code $\mathbf{f}^c$, $\mathbf{f}^m$ was quantized by K-means (MacQueen 1967) to form a discrete feature $\mathbf{q} = [q_1, ..., q_{2N}] \in Z^{2N}$. We adapt mutual information gap (MIG) (Chen et al. 2018) to self-MIG and apply it on $\mathbf{q}$ as follows:

$$\text{self-MIG} = \max_{i=1,...,2N} (H(q_i) - \max_{j \neq i} I(q_i; q_j))/H(q_i), \tag{13}$$

where H, I denote discrete entropy and mutual information.

Self-MIG measures disentanglement by the most independent content code or mask code. A larger self-MIG value indicates a greater degree of independence. For implementation details, please see the "Implementation Details" section.

Experiments

We evaluated unsupervised disentanglement representation learning on Shapes3d (Kim and Mnih 2018), MPI3D (Gondal et al. 2019), Cars3d (Reed et al. 2015) and attribute prediction on complex real-world dataset Market-1501 (Zheng et al. 2015). Visualization was performed to understand the learned concepts on the above datasets as well as face dataset FFHQ (Karras, Laine, and Aila 2019).

Experimental Setup

Datasets Evaluations were carried out on three datasets.
(1) **Shapes3d** (Kim and Mnih 2018) is a synthetic dataset of 3D shapes generated from 6 factors (floor color, wall color, object color, object scale, object shape and orientation). There are totally 480,000 samples.
(2) **MPI3D** (Gondal et al. 2019) dataset contains real-world images that capture 3D printed objects with variations of 6 factors (object color, object shape, object size, background color, camera height, horizontal axis and vertical axis). We evaluated on the MPI3D-real-complex version[1], which contains 460,800 samples.
(3) **Cars3d** (Reed et al. 2015) is a synthetic dataset generated by 183 3D car models with variations of 24 rotation angles and 4 camera elevations.

Evaluation Metrics Disentanglement metric (DCI) (Eastwood and Williams 2018), FactorVAE score (FVAE) (Kim and Mnih 2018) and mutual information gap (MIG) (Chen et al. 2018) were applied for evaluation. Disentanglement (D) value of DCI is reported by default. In each experiment, we reported the average performances of 10 models, each trained with a different random seed.

Since our disentangled representations are vector-wise, we followed the approach in COMET (Du et al. 2021) to separately post-process the content code and mask code of each concept using Principal Component Analysis (PCA) (Jolliffe 2002) to preserve d_r main components. This results in a latent code of dimensionality $d_{test} = 2Nd_r$. We fixed $d_{test} = 36$ to determine d_r. As the MIG metric requires the computation of discrete mutual information, we quantized each content code and mask code using K-means (MacQueen 1967), where the number of clusters K was set to 20, following the commonly used evaluation code of Locatello et al. (Locatello et al. 2019).

Implementation Details

To select the concept number N, we varied it from 2 to 10 and trained our FDAE. Then, as illustrated in "Model Training" of the methodology section, self-MIG was applied on discrete features quantized by K-means ($K = 20$) for each trained model. Finally, the N corresponding to the largest self-MIG was selected as default parameter, as shown in Table 1. By default, we used concept number $N = 6$ for Shapes3d (Kim and Mnih 2018), MPI3D (Gondal et al. 2019) and used $N = 2$ for Cars3d (Reed et al. 2015).

Concept number N	2	4	6	8	10
Shapes3d (default 6)	0.86	0.88	**0.92**	0.89	0.78
MPI3D (default 6)	0.85	0.86	**0.88**	0.87	0.56
Cars3d (default 2)	**0.76**	0.62	0.74	0.67	0.71

Table 1: Unsupervised metric self-MIG (↑) for selecting concept number N.

For our FDAE model, we employed ResNet-18 (He et al. 2016) as the backbone model of image encoder E. Input image $\mathbf{x}$ was resized to 64×64. Dimensionalities of the content codes (d_c), mask codes (d_m) and content masks (d_F) were all set to 80. After the final average pooling layer of ResNet, a fully connected layer was applied to obtain a latent code of dimensionality $d = N(d_c + d_m)$. Next, a fully connected layer was applied for dimensionality adjustment of content codes. We constructed the mask decoder D_M by following the architecture of the generator in DCGAN (Radford, Metz, and Chintala 2016). In the image generator G, the noisy image $\mathbf{x}_t$ first passed through a 3×3 2D convolution layer and then through U-Net (Ronneberger, Fischer, and Brox 2015), which was the same as that in EDM (Karras et al. 2022). All modules of FDAE were trained from scratch.

In our loss function, we set $w_{CD} = 2.5 \times 10^{-5}$ for content decorrelation loss $\mathcal{L}_{CD}$ in Eq. (10) and set $w_{ME} = 1.0 \times 10^{-4}$ for mask entropy loss $\mathcal{L}_{ME}$ in Eq. (11). For optimization, we used RAdam (Liu et al. 2020) with learning rate 1.0×10^{-4} for $100,000$ iterations and the batch size was set to 32. Training and inference of the diffusion model were the same as those used by EDM (Karras et al. 2022).

The training process takes 21 hours on 1 NVIDIA RTX 3090. More details are presented in supplementary material.

Compared Methods

(1) **VAE-based methods**: FactorVAE (Kim and Mnih 2018), β-TCVAE (Chen et al. 2018), DisCo-VAE (Ren et al. 2021) and DAVA (Estermann and Wattenhofer 2023);
(2) **GAN-based methods**: InfoGAN-CR (Chen et al. 2016), LatentDiscovery (LatentDisco) (Voynov and Babenko 2020), ClosedForm (Shen and Zhou 2021), GANSpace (Härkönen et al. 2020), DeepSpectral (Khrulkov et al. 2021) and DisCo-GAN (Ren et al. 2021);
(3) **Transformer-based method**: Visual Concepts Tokenization (VCT) (Yang et al. 2022);
(4) **Diffusion-based methods**: Diffusion Autoencoders (DiffAE) (Preechakul et al. 2022) and DisDiff-VQ (Yang et al. 2023).

We used the results reported in the papers of these methods for comparison by default. On MPI3D, we evaluated the best competitor VCT (Yang et al. 2022) using its released code on MPI3D-real-complex for fair comparison and also reported other published results on MPI3D-real. For fair comparison with DiffAE (Preechakul et al. 2022) as our baseline (concept number $N = 1$), we adopted the same network architecture and training strategy as our method.

[1] The MPI3D dataset contains two different versions of real-world data: MPI3D-real and MPI3D-real-complex. While most previous works have been evaluated on the MPI3D-real version, only the MPI3D-real-complex version is currently available on the dataset project page. This version captures more complex objects with variations of the same factors.

Method		Shapes3d			MPI3D (real/real-complex*)			Cars3d		
		DCI	FVAE	MIG	DCI	FVAE	MIG	DCI	FVAE	MIG
VAE	FactorVAE	$0.611^{\pm 0.082}$	$0.840^{\pm 0.066}$	$0.434^{\pm 0.143}$	$0.240^{\pm 0.051}$	$0.152^{\pm 0.025}$	$0.099^{\pm 0.029}$	$0.161^{\pm 0.019}$	$0.906^{\pm 0.052}$	$0.142^{\pm 0.023}$
	β-TCVAE	$0.613^{\pm 0.114}$	$0.873^{\pm 0.074}$	$0.406^{\pm 0.175}$	$0.237^{\pm 0.056}$	$0.179^{\pm 0.017}$	$0.114^{\pm 0.042}$	$0.140^{\pm 0.019}$	$0.855^{\pm 0.082}$	$0.080^{\pm 0.023}$
	DisCo-VAE	$0.844^{\pm 0.033}$	$0.956^{\pm 0.041}$	$0.331^{\pm 0.161}$	$0.288^{\pm 0.021}$	$0.391^{\pm 0.075}$	$0.068^{\pm 0.030}$	$0.211^{\pm 0.041}$	$0.761^{\pm 0.114}$	$0.103^{\pm 0.028}$
	DAVA	$0.780^{\pm 0.030}$	$0.820^{\pm 0.030}$	$\mathbf{0.620}^{\pm 0.050}$	$0.270^{\pm 0.030}$	$0.480^{\pm 0.050}$	$0.110^{\pm 0.050}$	$0.230^{\pm 0.040}$	$0.940^{\pm 0.010}$	$\underline{0.150^{\pm 0.010}}$
GAN	InfoGAN-CR	$0.478^{\pm 0.055}$	$0.587^{\pm 0.058}$	$0.297^{\pm 0.124}$	$0.241^{\pm 0.075}$	$0.439^{\pm 0.061}$	$0.163^{\pm 0.076}$	$0.020^{\pm 0.011}$	$0.411^{\pm 0.013}$	$0.011^{\pm 0.009}$
	LatentDisco	$0.380^{\pm 0.062}$	$0.805^{\pm 0.064}$	$0.168^{\pm 0.056}$	$0.196^{\pm 0.038}$	$0.391^{\pm 0.039}$	$0.097^{\pm 0.057}$	$0.216^{\pm 0.072}$	$0.852^{\pm 0.039}$	$0.086^{\pm 0.029}$
	ClosedForm	$0.525^{\pm 0.078}$	$0.951^{\pm 0.021}$	$0.307^{\pm 0.124}$	$0.318^{\pm 0.014}$	$0.523^{\pm 0.056}$	$0.183^{\pm 0.081}$	$0.243^{\pm 0.048}$	$0.873^{\pm 0.036}$	$0.083^{\pm 0.024}$
	GANSpace	$0.284^{\pm 0.034}$	$0.788^{\pm 0.091}$	$0.121^{\pm 0.048}$	$0.229^{\pm 0.042}$	$0.465^{\pm 0.036}$	$0.163^{\pm 0.065}$	$0.209^{\pm 0.031}$	$0.932^{\pm 0.018}$	$0.136^{\pm 0.006}$
	DeepSpectral	$0.513^{\pm 0.075}$	$0.929^{\pm 0.065}$	$0.356^{\pm 0.090}$	$0.248^{\pm 0.038}$	$0.502^{\pm 0.042}$	$0.093^{\pm 0.035}$	$0.222^{\pm 0.044}$	$0.871^{\pm 0.047}$	$0.118^{\pm 0.044}$
	DisCo-GAN	$0.708^{\pm 0.048}$	$0.877^{\pm 0.031}$	$0.512^{\pm 0.068}$	$0.292^{\pm 0.024}$	$0.371^{\pm 0.030}$	$0.222^{\pm 0.027}$	$0.271^{\pm 0.037}$	$0.855^{\pm 0.074}$	$\mathbf{0.179}^{\pm 0.037}$
Trans	VCT	$\underline{0.884^{\pm 0.013}}$	$\underline{0.957^{\pm 0.043}}$	$\underline{0.525^{\pm 0.028}}$	$0.475^{\pm 0.005}$ $0.467^{\pm 0.064}*$	$0.689^{\pm 0.035}$ $0.779^{\pm 0.058}*$	$0.227^{\pm 0.048}$ $\mathbf{0.204}^{\pm 0.060}*$	$\underline{0.382^{\pm 0.080}}$	$\underline{0.966^{\pm 0.029}}$	$0.117^{\pm 0.045}$
Diff	DisDiff-VQ	$0.723^{\pm 0.013}$	$0.902^{\pm 0.043}$	-	$0.337^{\pm 0.057}$	$0.617^{\pm 0.070}$	-	$0.232^{\pm 0.019}$	$\mathbf{0.976}^{\pm 0.018}$	-
	DiffAE (base)	$0.114^{\pm 0.008}$	$0.432^{\pm 0.019}$	$0.007^{\pm 0.002}$	$\underline{0.506^{\pm 0.016}*}$	$\underline{0.861^{\pm 0.020}*}$	$0.059^{\pm 0.003}*$	$0.307^{\pm 0.015}$	$0.959^{\pm 0.030}$	$0.023^{\pm 0.027}$
	FDAE (ours)	$\mathbf{0.917}^{\pm 0.038}$	$\mathbf{0.987}^{\pm 0.023}$	$0.473^{\pm 0.075}$	$\mathbf{0.644}^{\pm 0.031}*$	$\mathbf{0.903}^{\pm 0.030}*$	$\underline{0.197^{\pm 0.021}*}$	$\mathbf{0.418}^{\pm 0.036}$	$0.918^{\pm 0.027}$	$0.137^{\pm 0.020}$

Table 2: Comparison with the state-of-the-art methods. The results are presented as "mean$^{\pm \text{std}}$", of which the best is in bold type and the second best is marked by underline. "Trans" denotes transformers and "Diff" denotes diffusion models. On MPI3D, results with/without "*" denote evaluations on MPI3D-real/MPI3D-real-complex dataset.

Method	DINO	VCT	DiffAE	FDAE (ours)	APR (sup)
Top color	29.6	56.9	66.8	70.8	74.0
Bottom color	40.8	50.8	51.6	56.2	73.8
Gender	57.9	73.4	68.0	76.6	88.9
Hair	66.5	73.2	71.0	77.6	84.4
Backpack	73.6	74.1	73.6	76.6	84.9
Average	53.7	65.7	66.2	71.6	81.2

Table 3: Attribute prediction accuracy (%) on Market-1501. Linear SVM classifier is applied to the representations. APR (Lin et al. 2019) is supervised deep model as upper bound.

Model Comparison and Analysis

Comparison with the state-of-the-art Methods The performances of comparative evaluations are reported in Table 2. Our method outperforms all other methods in terms of the DCI score. On MPI3D-real-complex, our method also outperforms the best competitor VCT (Yang et al. 2022) on MPI3D-real, which achieves comparable performances on two different versions of MPI3D. On Shapes3d and Cars3d, the MIG score of the vector-wise concept representations (ours, VCT and DisDiff) are not as good as the best results achieved by the scalar-wise representations of the VAE-based or GAN-based methods, because MIG measures representation-factor mutual information difference between the top-2 dimensions.

Pedestrian Attribute Prediction on Market-1501 To evaluate more complex real-world vision task beyond uncovering factors in controlled environments, we applied the commonly used attribute annotations (Lin et al. 2019) of pedestrian appearance on Market-1501 (Zheng et al. 2015) to evaluate attribute prediction. The attributes include top color (8 types), bottom color (9 types), gender, hair (short or long) and backpack (yes or no). Predicting attributes for pedestrians is challenging due to their non-rigid nature and the complex environment they are in, which includes lighting variations and background clutters.

The attribute prediction evaluation involved several steps. First, unsupervised representation learning was performed on the training set. Next, representations were extracted from both the training set and the testing set using the learned model. Finally, for each attribute, linear SVM classifiers (Cristianini and Ricci 2008) were trained on the training representations and tested on the testing representations. The content codes processed by PCA (Jolliffe 2002) as illustrated in "Evaluation metrics" were used as input for the linear SVM classifiers.

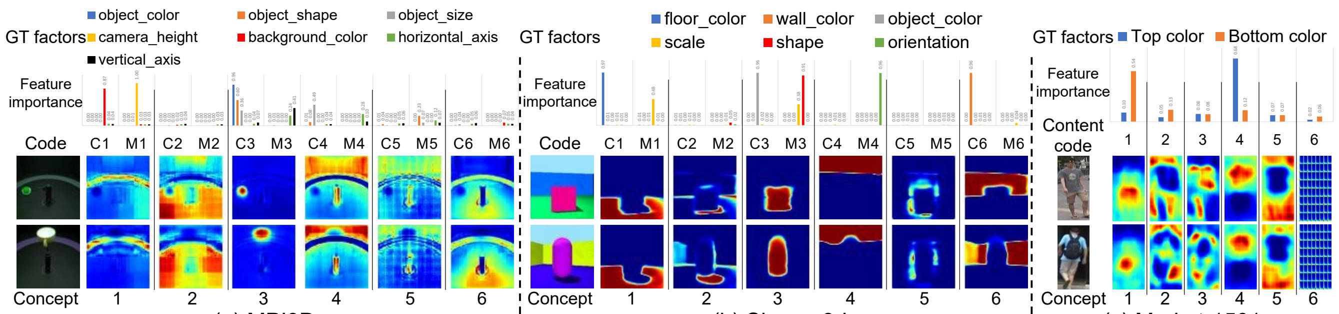
We compared with a representative self-supervised learning method DINO (ResNet-18) (Caron et al. 2021), a competitive unsupervised disentanglement method VCT (Yang et al. 2022) and our baseline method DiffAE (Preechakul et al. 2022). For our method, content codes were extracted for testing. The results of supervised deep attribute recognition method APR (Lin et al. 2019) are reported as upper bound. The prediction accuracies are shown in Table 3.

Our method FDAE achieves the best accuracy among unsupervised representation learning methods, with top color accuracy approaching that of the supervised APR model. This shows that our method is more effective for learning generalizable representations from non-rigid human bodies.

Visualization Results

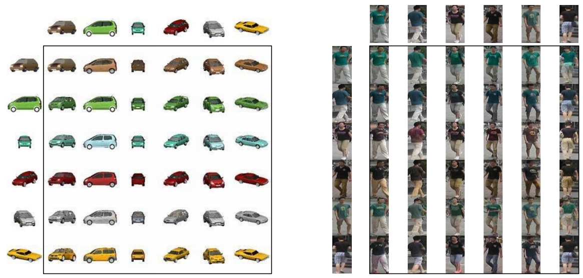
To further understand our model FDAE, we provide visualization results of the uncovered visual concepts and image manipulation by swapping specific concepts and latent traversal. More visualization results are shown in the supplementary material.

Visualization of Uncovered Concepts Our method is capable of generating interpretable masks for understanding the learned visual concepts. Besides the masks, we also



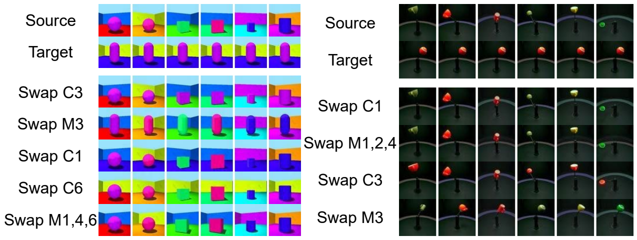
(a) MPI3D (b) Shapes3d (c) Market-1501

Figure 3: Visualization of visual concepts learned by our Factorized Diffusion Autoencoder (FDAE). For each concept, the masks and feature importance histograms of each ground-truth (GT) factors are shown. "C1-C6" denotes content codes and "M1-M6" denotes mask codes.



(a) Cars3d (b) Market-1501

Figure 4: Images generated by swapping content codes and mask codes on Cars3d and Market-1501. 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.



(a) Shapes3d (b) MPI3D

Figure 5: Examples of images generated by swapping specific latent code indicated on the left. The images are generated by using the latent codes extracted from the source image with the indicated latent code swapped to the one extracted from the target image. The notations "C1-C6" and "M1-M6" are corresponding to those in Figure 3.

show the feature importances of the latent codes. Following computation of the DCI metric (Eastwood and Williams 2018), we trained the Gradient Boosted Decision Trees (GBDT) (Friedman 2001) to predict each ground-truth factor by latent code $\mathbf{f}$. The feature importances of each content code $\mathbf{f}^c$ or each mask code $\mathbf{f}^m$ were summed up and displayed by histograms. The masks learned by our FDAE on MPI3D, Shapes3d and Market-1501 are shown in Figure 3.

The learned masks demonstrate that the main instances in the images, such as objects and backgrounds, can be learned as different visual concepts. Although some visual concepts do not learn meaningful representations, such as concepts 5 and 6 on MPI3D and concept 5 on Shapes3d, these concepts have little impact on image reconstruction due to their low probabilities on the masks.

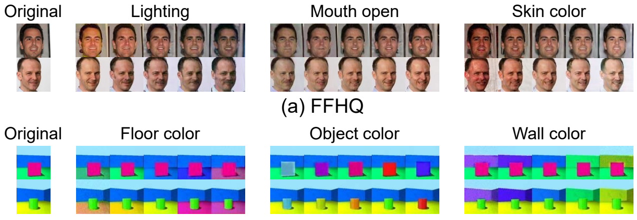
For most factors on Shapes3d, feature importances of a specific latent code are generally over 0.9, which indicate that different content codes (C1-C6) and mask codes (M1-M6) are effectively disentangled. On MPI3D that exhibits more complex variations, factors such as object color, background color, and camera height are successfully disentangled. On Market-1501, attribute prediction of top color and bottom color mainly depends on concept 4 and concept 1.

The interpretable masks make it easy to determine which group of content code and mask code may be useful for downstream tasks without navigating the latent space.

Swapping Content Codes and Mask Codes To show that content codes and mask codes are disentangled, we randomly selected 6 images from Cars3d (Reed et al. 2015) and Market-1501 (Zheng et al. 2015). Then, we performed pairwise latent code swapping for image generation. For a pair of images $\mathbf{x}_1$ and $\mathbf{x}_2$, we used the content codes $\mathbf{f}_1^{c1}, ..., \mathbf{f}_1^{cN}$ extracted from $\mathbf{x}_1$ and the mask codes $\mathbf{f}_2^{m1}, ..., \mathbf{f}_2^{mN}$ extracted from $\mathbf{x}_2$ to generate new images. As shown in Figure 4, content codes mainly represent appearances and the mask codes mainly represent the shapes and viewpoints.

Swapping Specific Latent Codes From the masks in Figure 3, we can gain insight into the information extracted by each content code and mask code. To demonstrate that these latent codes are disentangled, we randomly selected some samples from Shapes3d and MPI3D. We extracted the latent codes from the source images and swapped one of the latent codes with that extracted from a target image to generate new images. Some examples are shown in Figure 5.

On Shapes3d, the factors of object color, object shape, floor color, wall color and orientation can be independently controlled by swapping latent codes C3, M3, C1, C6 and



(a) FFHQ

(b) Shapes3d

Figure 6: Visualization of latent traversal on content codes.

M1+M4+M6, respectively. On more complex MPI3D, the factors of ring color, camera height, object color and orientation can be controlled by latent codes C1, M1+M2+M4, C3 and M3, respectively. Note that, the orientation factor on Shapes3d and camera height factor on MPI3D affect multiple instances, so that controlling this factor requires swapping multiple mask codes simultaneously, which can be easily selected by observing the masks in Figure 3. Failure cases are rare. The first image in the last row of MPI3D mistakenly changes the coffee cup to a beer cup. This may be caused by similar holders for the two types of cups.

Latent Traversal on Principal Component of Concepts To explore the effect of the latent space, we show some qualitative latent traversal results on Shapes3d (Kim and Mnih 2018) as well as a commonly used real-world face dataset FFHQ (Karras, Laine, and Aila 2019). As illustrated in "Evaluation Metrics" in the experiment section, PCA (Jolliffe 2002) was applied to post-process each content code and mask code. Latent traversal on the direction of the most principal component was conducted for some content codes. The latent traversal results are shown in Figure 6, which demonstrate that our method is effective for learning meaningful factors of faces and objects.

Further Evaluations

Ablation Study We evaluated the main components of our FDAE model. The results on MPI3D are reported in Table 4. "$E + G$" 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 in Eq. (10) and mask entropy loss in Eq. (11), respectively.

Compared to the baseline method (experiment 1), CMFNet (experiment 2) significantly improves the DCI and MIG scores, indicating the effectiveness of factorizing concepts into multiple groups of "content + mask". Loss functions $\mathcal{L}_{CD}$ (experiment 3) and $\mathcal{L}_{ME}$ (experiment 4) can further improve disentanglement of the representations and are complementary to each other (experiment 5). This demonstrates the effectiveness of decorrelating visual concepts in both the latent space and the spatial space. Similar conclusions are drawn from the ablation study results on Cars3d and Shapes3d reported in the supplementary material.

Impact of Concept Number N We varied concept number N from 2 to 10 on MPI3D. The case of $N = 1$ is the baseline method without factorization. For each content code and mask code, we fixed the number of principle component d_r as 3 for PCA. The results are shown in Table 5.

Components	$E + G$	CMFNet	$\mathcal{L}_{CD}$	$\mathcal{L}_{ME}$	DCI	FVAE	MIG
1 (baseline)	✓				0.506	0.861	0.059
2	✓	✓			0.631	0.879	0.152
3	✓	✓	✓		0.642	0.893	0.161
4	✓	✓		✓	0.637	0.888	0.180
5 (full)	✓	✓	✓	✓	0.644	0.903	0.197

Table 4: Ablation study on MPI3D.

N	Dimension	DCI	FVAE	MIG
1 (baseline)	80	0.506	0.861	0.059
2	12	0.437	0.770	0.221
4	24	0.629	0.878	0.182
6 (default)	36	0.644	0.903	0.197
8	48	0.646	0.916	0.151
10	60	0.669	0.905	0.074

Table 5: Effect of concept number N on MPI3D.

The performance remains relatively stable when N varies from 4 to 8. When $N = 2$, there are not sufficient content masks to capture the variations of concepts, resulting in a lower DCI score indicating inferior disentanglement. When $N = 10$, the dimensionality of the latent code is larger and the representation is less compact than that in other cases, leading to a decrease in the MIG score. For unsupervised learning, concept number N can be selected by unsupervised metric self-MIG as illustrated in "Model Training" in the methodology part.

Conclusion

We introduce visual interpretability as inductive bias for unsupervised disentangled representation learning, which is still under-explored in existing approaches that focus on imposing independence-based or informativeness-based constraints. To learn visually interpretable representations that align with human-defined factors, we factorize an image into multiple visual concepts represented by "content + mask" and reconstruct the image from the concepts to ensure informativeness of the representations. To achieve this, we propose the Factorized Diffusion Autoencoder (FDAE), which consists of a content-mask factorization network (CMFNet) and a conditional diffusion-based image generator. Extensive experiments on benchmark datasets in controlled environment and real-world pedestrian dataset in uncontrolled environment show the effectiveness of our method. Furthermore, the visually interpretable masks of our method facilitate understanding the uncovered concepts. This work demonstrates the potential of visual interpretability as inductive bias for unsupervised representation disentanglement.

Acknowledgments

This work was supported partially by the National Science Foundation for Young Scientists of China (62106288), the Guangdong Basic and Applied Basic Research Foundation (2023A1515012974), the Guangzhou Basic and

Applied Basic Research Scheme (2024A04J4066), the NSFC (U21A20471) and Guangdong NSF Project (No. 2023B1515040025).

References

Bengio, Y.; Courville, A. C.; and Vincent, P. 2012. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35: 1798–1828.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *IEEE/CVF international conference on computer vision (ICCV)*, 9650–9660.

Chen, R. T.; Li, X.; Grosse, R. B.; and Duvenaud, D. K. 2018. Isolating sources of disentanglement in variational autoencoders. In *Advances in neural information processing systems (NeurIPS)*, volume 31.

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 (NeurIPS)*, volume 29.

Cristianini, N.; and Ricci, E. 2008. *Support Vector Machines*, 928–932. Boston, MA: Springer US. ISBN 978-0-387-30162-4.

Du, Y.; Li, S.; Sharma, Y.; Tenenbaum, J.; and Mordatch, I. 2021. Unsupervised learning of compositional energy concepts. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34.

Eastwood, C.; and Williams, C. K. I. 2018. A Framework for the Quantitative Evaluation of Disentangled Representations. In *International Conference on Learning Representations (ICLR)*.

Estermann, B.; and Wattenhofer, R. 2023. DAVA: Disentangling Adversarial Variational Autoencoder. In *International Conference on Learning Representations (ICLR)*.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29: 1189–1232.

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.

Härkönen, E.; Hertzmann, A.; Lehtinen, J.; and Paris, S. 2020. Ganspace: Discovering interpretable gan controls. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C. P.; Glorot, X.; Botvinick, M. M.; Mohamed, S.; and Lerchner, A. 2016. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations (ICLR)*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *ArXiv*, abs/2006.11239.

Jolliffe, I. T. 2002. Principal Component Analysis. In *International Encyclopedia of Statistical Science*.

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

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 4401–4410.

Khrulkov, V.; Mirvakhabova, L.; Oseledets, I.; and Babenko, A. 2021. Disentangled Representations from Non-Disentangled Models. *ArXiv*, abs/2102.06204.

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

Kwon, M.; Jeong, J.; and Uh, Y. 2022. Diffusion Models already have a Semantic Latent Space. *ArXiv*, abs/2210.10960.

Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; and Yang, Y. 2019. Improving person re-identification by attribute and identity learning. *Pattern recognition*, 95: 151–161.

Liu, L.; Jiang, H.; He, P.; Chen, W.; Liu, X.; Gao, J.; and Han, J. 2020. On the Variance of the Adaptive Learning Rate and Beyond. In *International Conference on Learning Representations (ICLR)*.

Locatello, F.; Bauer, S.; Lucic, M.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *International Conference on Machine Learning (ICML)*, 4114–4124.

MacQueen, J. B. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 1: 281–297.

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.

Ren, X.; Yang, T.; Wang, Y.; and Zeng, W. J. 2021. Learning Disentangled Representation by Exploiting Pretrained Generative Models: A Contrastive Learning View. In *International Conference on Learning Representations (ICLR)*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent

Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241. Springer.

Shen, Y.; and Zhou, B. 2021. Closed-Form Factorization of Latent Semantics in GANs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1532–1540.

Voynov, A.; and Babenko, A. 2020. Unsupervised discovery of interpretable directions in the gan latent space. In *International conference on machine learning (ICML)*, 9786–9796. PMLR.

Xiang, W.; Yang, H.; Huang, D.; and Wang, Y. 2023. Denoising Diffusion Autoencoders are Unified Self-supervised Learners. *ArXiv*, abs/2303.09769.

Yang, T.; Wang, Y.; Lu, Y.; and Zheng, N. 2022. Visual concepts tokenization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35.

Yang, T.; Wang, Y.; Lv, Y.; and Zh, N. 2023. DisDiff: Unsupervised Disentanglement of Diffusion Probabilistic Models. *ArXiv*, abs/2301.13721.

Zhang, Z.; Zhao, Z.; and Lin, Z. 2022. Unsupervised Representation Learning from Pre-trained Diffusion Probabilistic Models. *ArXiv*, abs/2212.12990.

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.

Zhu, X.; Xu, C.; and Tao, D. 2021. Where and What? Examining Interpretable Disentangled Representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5857–5866.