bioinformatics [54, 55] for measuring the degree of correlation among independent variables. It appears as a common regularization term in representation learning and deep learning. Mutual information is an information measure commonly used in information theory that represents the degree of randomness reduction of another independent variable when one independent variable is known. However, mutual information can only be leveraged to measure a pair of independent variables.

Total correlation (TC) has played an important role in disentangled representation learning for measuring the overall correlation among multi-dimensional variables [56]. It is a challenging task to calculate the TC value due to its unknown distribution of real data. Cheng *et al* [57] decomposed TC and used mutual information bound to estimate the TC value. This paper proposes a novel decomposition path of total correlation based on mutual information bound, and for the first time, applies this method to the VAE, revealing its crucial significance for unsupervised disentangling.

3. VAE with Flexible Disentanglement

Symbol	Explanation
n	Latend variable z has n dimensions
z_i	The ith dimension's latent variable in the multidimensional latent variable z
e^*	In the case of n is even
o^*	In the case of n is odd
$cp(i,j)$	Mechanism for selecting a pair of dimensions z_i, z_j without replacement
z_r	The final remaining variable r under the $cp(i,j)$ of o^*
$q(z_{-i,j})$	The joint distributions of all dimensions except z_i, z_j
$I(z_i\,;\,z_j)\vert_{q(z)=q(z_i,z_j)\cdot q(z_{-i,j})}$	Calculating the mutual information of z_i and z_j using importance sampling under the target distribution where $q(z_i,z_j)$ and $q(z_{-ij})$ are independent of each other

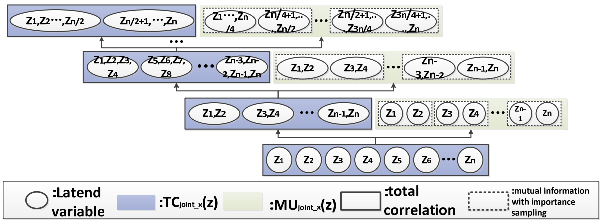
Table 1.Explanation of mathematical symbols

For the convenience of understanding, Table 1 are the mathematical symbols' explanation used in the following chapters.

3.1. New Path to Total Correlation Decomposition

β-TCVAE [2] decomposes $E_{p(x)}[D_{KL}(q(z|x)||p(z)]$ to get the following objective:

$$E_{p(x)}[D_{KL}(q(z|x)||p(z)] = D_{KL}(q(z,x_n)||q(z)p(x_n)) + D_{KL}(q(z)||\prod_i q(z_i)) + \sum_i D_{KL}(q(z_i)||p(z_i)).$$



Bottom-up decomposition of TC

Figure 1.The total correlation is gradually decomposed into the sum of the joint distribution total correlation, denoted as TCjoint_x(z). The importance sampling of mutual information between the marginal distributions of the internal joint distribution, referred to as MUjoint_x(z).

(1)

It is generally believed that the disentanglement of β-TCVAE [2] mainly comes from two aspects in objective (1): 1). The mutual information between latent variables and data variables $I(x;z)$ is represented as the first item; 2). The independence among latent variables $TC(z)$ is represented as the second item. The third term constrains the difference between the latent variable distribution and the prior distribution. This paper mainly analyzes the second item.

Study [57] proffered two distinct pathways for total correlation decomposition, yet it was bereft of practical suggestions regarding the actual applications of total correlation decomposition. Within this paper, we posit a novel decomposition pathway, designated as bottom-up decomposition, which is well-suited for unsupervised disentanglement scenarios in the VAE framework. (Due to limitations in the length of this article, the derivation and comparison of the three distinct pathways for decomposition, as well as the rationale for why our proposed method is better suited for application within the VAE framework, have been relegated to the supplementary materials.)

The following is a brief introduction to the decomposition path we proposed. The complete decomposition process is in the supplementary materials A.3. Refer to Fig 1 for the decomposition process detail.

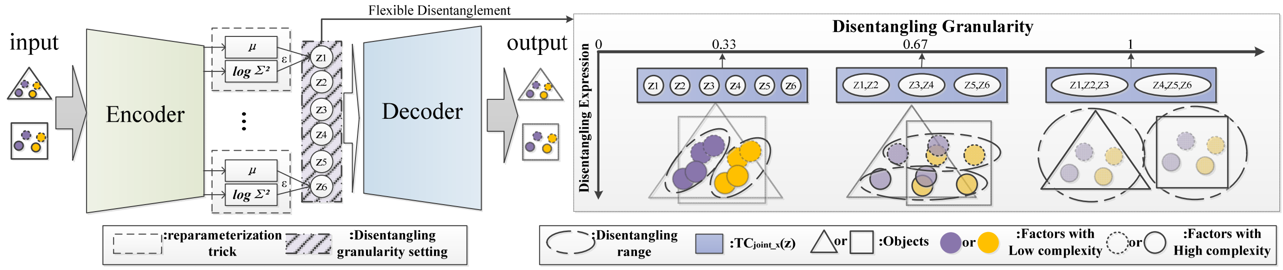
Assuming that z has n dimensions, then:

$$TC(z) = D_{KL}(q(z)||\prod_k q(z_k)) = \begin{cases} \sum_{cp(i,j)} I(z_i\,;\,z_j)|_{q(z)=q(z_i,z_j)\cdot q(z_{-i,j})} + D_{KL}(q(z)||\prod_{cp(i,j)} q(z_i,z_j)) & , e^* \\ \sum_{cp(i,j)} I(z_i\,;\,z_j)|_{q(z)=q(z_i,z_j)\cdot q(z_{-i,j})} + D_{KL}(q(z)||\prod_{cp(i,j)} q(z_i,z_j)\cdot q(z_r)) & , o^* \end{cases} \quad (2)$$

The explanation of each item in objective (2) is as follows:

The first term $\sum_{cp(i,j)} I(z_i\,;\,z_j)|_{q(z)=q(z_i,z_j)\cdot q(z_{-i,j})}$ is called "mutual information summation" term, which represents the accumulation of mutual information combinations of all pairs z_i, z_j of latent variables selected by $cp(i,j)$.

The second term is called "joint distributions total



(a) Model structure (b) Disentangling Granularity and Disentangling Expression

Figure 2. As shown in Fig. 2(a). We can control the disentangled expressions of unsupervised models in different feature domains by changing the disentangling granularity of latent variables. The figure above intuitively shows that as the disentangling granularity becomes smaller, the model focuses more on disentangling subtle features with lower complexity, while as the disentangling granularity becomes larger, the model pays more attention to disentangle features with higher complexity.

correlation" term. During derivation, we discovered that the target distribution of the importance sampling in the first term $I(z_i\,;\,z_j)$ satisfy $q(z_i, z_j)$ and $q(z_{-i,j})$ are independent of each other. The second term can also be regarded as the accuracy constraint for calculating the mutual information summation term. It avoids sampling bias by ensuring that the proposed distribution $q(z_i, z_j)$ is not too far from the target distribution during importance sampling.

Then we set $TC_{joint\_1}(z)$ as:

$$TC_{joint\_1}(z) := \begin{cases} D_{KL}(q(z)\,||\prod_{cp(i,j)} q(z_i, z_j)) & ,e^* \\ D_{KL}(q(z)\,||\prod_{cp(i,j)} q(z_i, z_j)\cdot q(z_r)) & ,o^* \end{cases}. \tag{3}$$

We replace $\sum_{cp(i,j)} I(z_i\,;\,z_j)|_{q(z)=q(z_i z_j)\cdot q(z_{-i,j})}$ with $MU_{joint\_1}(z)$, then $TC_{joint\_1}(z) = TC(z) - MU_{joint\_1}(z)$.

From objective (2), knowing $MU_{joint\_1}(z)$ is always positive, defined as the accumulation of mutual information. Thus, we have $TC_{joint\_1}(z) \le TC(z)$. This inequality brings us two interesting inferences:

Inference 1: The total correlation of joint distributions must be less than or equal to the total correlation of marginal distributions that compose the joint distributions.

Inference 2: We can split and refine the constraint granularity of $TC(z)$. $MU_{joint\_1}(z)$ constrains the independence of marginal distributions within joint distributions, while $TC_{joint\_1}(z)$ constrains the independence among joint distributions.

Furthermore, the $TC(z)$ decomposition has an iterative relationship and we can continue by treating $TC_{joint\_1}(z)$ as a new $TC(z)$. All joint distributions $q(z_i, z_j)$ selected by $cp(i,j)$ in the $TC_{joint\_1}(z)$ are regarded as the elements of new latent variables distributions set $\{q(z_{1\_1}), q(z_{1\_2})..q(z_{1\_n/2})\}$ ($(n+1)/2$ new variables in total when n is odd, and $n/2$ in total when n is even), then we have below iterative objectives:

$$\begin{aligned} TC_{joint\_2}(z) &= TC_{joint\_1}(z) - MU_{joint\_2}(z), \\ TC_{joint\_3}(z) &= TC_{joint\_2}(z) - MU_{joint\_3}(z), \\ &.... \end{aligned} \tag{4}$$

Finally, there is the following objective:

$$TC(z) = MU_{joint\_1}(z) + MU_{joint\_2}(z) + \ldots + I(z_{f\_1}\,;\,z_{f\_2}), \tag{5}$$

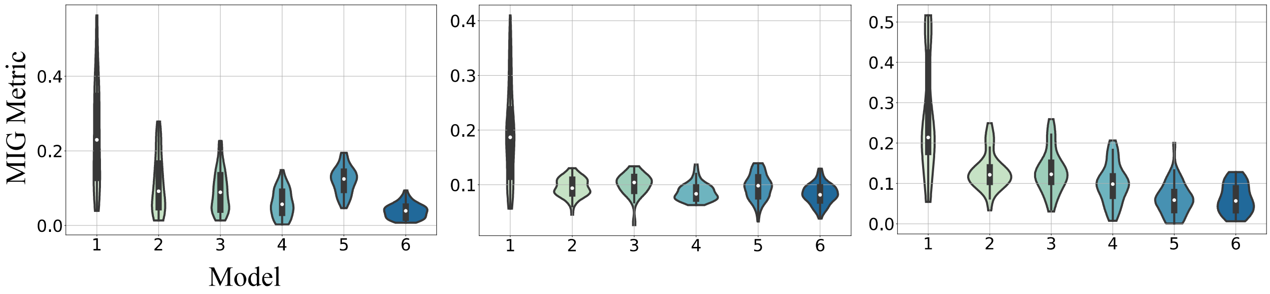
where $f = int(log_2 n) - 2$. If $f < 1$, then $z_{f\_i} = z_i$, and n is the total dimension of z, and $int(c)$ is the rounded-up integer of c.

3.2. β-STCVAE model

3.2.1 Novel Objective Function

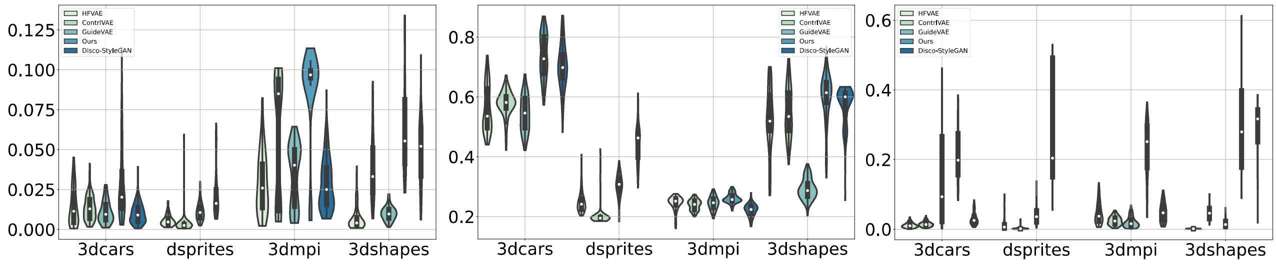
The real data features $x \in R^N$ are assumed to be divided into two parts: conditionally independent features $v \in R^K$ and conditionally dependent features $w \in R^H$ [16]. The conditional independent features satisfy the independence: $logp(v|x) = \sum_K log(v_k|x)$. And the real data distribution can be simulated by a parametric neural network $p(x|v, w) = net(v, w)$, which is used to find an inferred posterior distribution $q(z|x)$. Without knowing the complexity of real data features, the disentangling method matching prior distribution among latent variables is searched during training, which poses a challenge to the correct allocation of parameter capacity for VAEs. Abstractly, the total parameter capacity (Cz) of the network consists of two parts: parameter capacity allocated to independent features (Cv) and parameter capacity allocated to dependent features (Cw), thus $Cz = Cv + Cw$. The parameter capacity can be effectively utilized when it nearly matches v and w, otherwise, the learning ability of the model will be weakened.

We can constrain $TC_{joint\_x}(z)$ and $MU_{joint\_x}(z)$ separately based on inference 2 in section 3.1. Intuitively, it brings the ability to separate regions of independence constraints, $TC_{joint\_x}(z)$ constrains the independence between macro contents, and $MU_{joint\_x}(z)$ constrains the independence of micro features under the same content.



(a) violin on shape3d (b) violin on cars3d (c) violin on dSprites

Figure 3. The Fig. 4(a), Fig. 4(b) and Fig. 7(c) are the MIG-based disentangling metric violin graphs of different VAE models on shape3d, cars3d, and dSprites datasets. Models are abbreviated (1=StcVae_8_2; 2=TcVae_16; 3=TcVae_8; 4=BetaVae_8; 5=FactorVae_8; 6=DIP-I_8). Abbreviated writing method: (Model name)_(latent variables)_(Grouping factor). For example, Stcvae_8_2 represents a Stcvae model with 8 latent variables when the total number of latent dimensions is 16, and the grouping factor is 2. TcVae_8 represents a TcVae model with 8 latent variables.



(a) MIG evaluation (b) Factor evaluation (c) SAP evaluation

Figure 4. On four different datasets, we trained five models, with experimental groups arranged from left to right as follows: HFVAE, ControlVAE, GuidedVAE (unsupervised), stcvae, and DisCo (based on Style-GAN). The DisCo model experiment is missing from the dsprites dataset because the paper does not provide a pre-trained model on Style-GAN for dsprites. We fixed the latent variable dimension of all models to be 8 (the latent variable dimensions of stcvae and HFVAE are 16, the grouping factors of stcvae are 2, and the sub-group size of HFVAE is 2). As can be seen from Fig, stcvae obtains the best disentanglement score on almost all training sets and evaluation scales, except for the 3dshapes dataset under MIG evaluation, where it slightly falls short of DisCo. At the same time, we also note that there are slight differences between the disentanglement evaluation experimental results and those in the baseline, due to the parameter settings of the evaluation method, such as the number of iterations, the size of latent variable dimension, and the size of batch size during iteration, etc. More experimental parameter settings can be found in the supplementary materials.

distributions to obtain a more complex prior distribution, causes the disentangling granularity under the optimal ELBO to rise again (pink line area of Fig. 2(d)). Additionally, we discovered that as the disentangling strength enhances, the optimal ELBO trajectory moves up. This phenomenon implies that the model allocates more parameter capacity to learn dependent features as a balancing mechanism to counteract the high disentangling strength.

Therefore, we suggest the following three points to improve VAEs' representation learning performance:

1) When the total parameter capacity is relatively low, the allocation of parameter capacity should bias toward dependent features;
2) When the total parameter capacity is relatively high, modify the prior distribution that matches the data feature complexity learned from model.
3) When the disentangling strength is greater, the larger the disentangling granularity should be selected.

4.2.2. Disentangling Tendency in Different Regions.

After the generation of numerous samples from various datasets, we can summarize the following laws:

- Dependent feature overfitting on the Upper right of Fig. 2(d): Caused by Cw being too high. Typically when traversing a single latent variable, the generated samples' features change significantly and contain various features that cannot be allocated to a specific attribute or category.
- Dependent feature underfitting on the Lower left of Fig. 2(d): Caused by Cw being too low. Typically when traversing a single latent variable, the generated samples are reconstructions of input samples with low or corrupt quality.
- Independent features overfitting on the Lower right of