# Cell Variational Information Bottleneck Network

**Zhonghua Zhai** ZHAIZHONGHUA@HIKVISION.COM *Hikvision Research Institute*

**Editors:** Berrin Yanıkoğlu and Wray Buntine

## Abstract

In this work, we propose "Cell Variational Information Bottleneck Network (cellVIB)", a convolutional neural network using information bottleneck mechanism, which can be combined with the latest feedforward network architecture in an end-to-end training method. Our Cell Variational Information Bottleneck Network is constructed by stacking VIB cells, which generate feature maps with uncertainty. As layers going deeper, the regularization effect will gradually increase, instead of directly adding excessive regular constraints to the output layer of the model as in Deep VIB. In each VIB cell, the feedforward process learns an independent mean term and a standard deviation term, and predicts the Gaussian distribution based on them. The feedback process is based on reparameterization trick for effective training.

This work performs an extensive analysis on MNIST dataset to verify the effectiveness of each VIB cells mentioned above, and provides an insightful analysis on how the VIB cells affect mutual information. Experiments conducted on CIFAR-10 also prove that our network is robust against noisy labels during training and against corrupted images during testing. Then, we validate our method on PACS dataset, whose results show that the VIB cells can significantly improve the generalization performance of the basic model. Finally, in a more complex representation learning task, face recognition, our network structure has also achieved very competitive results.

**Keywords:** Variational Information Bottleneck, feature maps with uncertainty, regularization effect

## 1. Introduction

It is meaningful to explain deep neural networks from the perspective of information theory. The information bottleneck was first proposed in Tishby et al. (2000). It is attractive because it makes a balance between the concise representation and the representation with good predictive ability Tishby and Zaslavsky (2015). The main disadvantage of the information bottleneck principle is that it is usually very difficult to calculate mutual information, which severely limits the types of learnable models and makes the information bottleneck only available under some extreme assumptions. To solve this problem, Alemi et al. (2016) uses variational inference to construct the lower bound of the information bottleneck optimization objective, and uses reparameterization trick Kingma and Welling (2013) so that deep neural networks can parameterize the distribution of high-dimensional continuous data and avoiding previous restrictions on discrete or Gaussian cases. However, the experiments in Tsai et al. (2021) show that such information bottleneck methods are still not thorough enough.

To understand information theory from the perspective of regularization, the latent code $Z$ of the input $X$ is defined by the parameter encoder $P(Z|X;\theta)$. The target is to

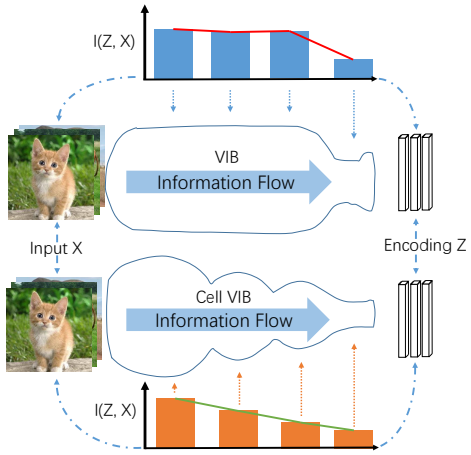


Figure 1: **Concepts** of Cell Variational Information Bottleneck. **Best viewed in color**.

learn a code that can understand the target $Y$ to the greatest extent, measured by the mutual information $I(Z, Y; \theta)$ between the code $Z$ and the target $Y$, and at the same time, to minimize the complexity of $Z$, measured by the mutual information $I(Z, X; \theta)$ between code $Z$ and input $X$. In order to improve the regularity effect, we penalize the redundant information theory measurement between adjacent network layers, that is, use variational approximation to penalize the mutual information between the layers, instead of only penalizing mutual information between $Z$ and $X$. When looking at information theory from another perspective, uncertainty comes from two parts Alemi et al. (2018). First, $Z$ is no longer a certain point, but a random distribution; second, we penalize the KL divergence between the conditional distribution of a given input and a low-dimensional, low-information code space. In this case, for the upper bound of the variational inference of uncertainty, the layer-by-layer information interference is higher than just applying interference to a certain layer Tsai et al. (2021). This is why we establish an information bottleneck in each cell as shown in Figure 1.

The cellVIB network can be generated by simply stacking VIB cells, and VIB cells can also be used as a direct replacement for primitive blocks of any depth in the architecture. Although the template of the cells is universal, there will still be differences in cells of different depths. In the shallow network, the VIB cell plays a weaker regularization role and introduces weaker uncertainty. As the network continues to deepen, the regularization and uncertainty will continue to strengthen. Therefore, the regularization and uncertainty of cellVIB are gradual. The development of a new CNN architecture is a hard engineering task, usually involving many new layer configurations. In contrast, the design of the VIB cell outlined above is very convenient and can be used with the existing most advanced architectures directly. The modules of these architectures can be enhanced by directly replacing with the VIB cells.

In Chapter Experiments, we verify the superiority of cellVIB from different perspective. In order to further illustrate its general applicability, we have adopted different backbone structures and different datasets in different tasks to show that our proposed method is not limited to specific datasets or tasks. First, we analyze the correspondence between $I(Z; Y)$

and $I(Z;X)$ when $\beta$ changes on MNIST dataset, and analyze how $I(Z;Y)$ and $I(Z;X)$ changes with the network depth. The experimental results demonstrate the deficiencies of Deep VIB in removing information redundancy. Subsequently, we also prove through a series of experiments that cellVIB method is robust to input noise and label noise, because the representation $Z$ learned by cellVIB removes more redundant information about $X$. Besides, compared with the deterministic model fitted by maximum likelihood estimation or VIB, cellVIB has a stronger generalization ability. Intuitively speaking, each input image is mapped to a feature map distribution instead of a unique feature map in each layer. The introduction of this uncertainty strengthens the generalization ability of the model. Finally, our model also shows very competitive results in face recognition, a much more complex representation learning task.

The major contributions of this paper are summarized as follows.

- We discovered the existence of mutual information redundancy between layers, and analyzed the influence of mutual information redundancy between layers on the generalization and robustness of the model, which has never been paid attention to before.
- We proposed a new network structure, cellVIB, by gradually increasing the regularization effect and uncertainty as the layer deepens, instead of directly adding too many regular constraints to the output layer. VIB cells can also be used as a direct replacement for primitive blocks of any depth in the architecture.
- We verified the superiority of our model through detailed experiments, and explained from the perspective of mutual information why the results displayed on a smaller dataset can be extended to larger ones.

## 2. Related Work

### 2.1. Deep Architectures

VGGNets Simonyan and Zisserman (2014) and Inception models Szegedy et al. (2015) explains the effect of network depth on performance. Batch normalization (BN) Ioffe and Szegedy (2015) improved the gradient propagation by inserting units for adjusting layer input, thereby stabilizing the learning process. ResNets He et al. (2016a,b) demonstrated the effectiveness of learning deeper networks by using identity-based skip connections. Highway network Srivastava et al. (2015) used a gating mechanism to adjust shortcut connections.

Another research direction is to explore ways to adjust the functional form of network modular components. Grouped convolution can be used to increase the cardinality Ioannou et al. (2017); Xie et al. (2017). Multi-branch convolutions can be interpreted as a generalization of the concept, which makes the combination of operators more flexible Ioffe and Szegedy (2015); Szegedy et al. (2017, 2015, 2016). Recently, compositions learned in an automated manner Liu et al. (2017); Zoph and Le (2016); Zoph et al. (2018) have shown competitive performance. Cross-channel correlation is usually independent of the spatial structure Chollet (2017); Jaderberg et al. (2014) or by using a standard convolution filter with $1 \times 1$ convolution Lin et al. (2013) to jointly map a new combination of features. In this work, we construct Cell Variational Information Bottleneck Network by stacking VIB cells, which generate feature maps with uncertainty.

### 2.2. Information Theory Regularization Methods

Tishby & Zaslavsky Tishby and Zaslavsky (2015) pointed out the idea of applying information theoretic objectives to deep neural networks. However, their work only contained theoretical hypotheses, and did not design experiments to verify their hypotheses. Although not combined with the information bottleneck objective, variational bounds on mutual information has been discussed previously in Agakov (2004). Mohamed & Rezende Mohamed and Rezende (2015) also explored variational bounds on mutual information and applied it to deep neural networks for reinforcement learning. Variational autoencoder Kingma and Welling (2013) is a special case in the unsupervised learning literature, in which the $\beta$ parameter is fixed at 1.0 while Higgins et al. Higgins et al. (2016) explored VAE targets with different $\beta$ values.

Variational information bottleneck Alemi et al. (2016) suggests using variational inference to construct a lower bound of the information bottleneck objective. The reparameterization trick Kingma and Welling (2013) allows the use of deep neural networks to parameterize the distribution to process high-dimensional continuous data. In this work, we have revealed the crux of VIB from the perspective of mutual information redundancy and put forward a feasible solution, cellVIB, by gradually increasing the regularization effect as the layer deepens, instead of directly adding too many regular constraints to the output layer.

## 3. Cell Variational Information Bottleneck

We propose VIB cell which includes distributional representation and KL-divergence regularization in a repeatable small cell.

### 3.1. Distributional Representation

Specifically, we define a hidden layer $\tilde{\mathbf{x}}$ of input $\mathbf{x}$ as a Gaussian distribution,

$$p(\tilde{\mathbf{x}}|\mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}; \tilde{\mathbf{x}}_\mu, \tilde{\mathbf{x}}_\sigma^2 \mathrm{I}) \tag{1}$$

The two parameters of the Gaussian distribution (mean and variance) are both predicted by CNN and related to the input:

$$\tilde{\mathbf{x}}_\mu = f_{\theta_\mu}(\mathbf{x}); \tilde{\mathbf{x}}_\sigma = f_{\theta_\sigma}(\mathbf{x}) \tag{2}$$

where $\theta_\mu$ and $\theta_\sigma$ are the model parameters with respect to output $\tilde{\mathbf{x}}_\mu$ and $\tilde{\mathbf{x}}_\sigma$ respectively. The predicted Gaussian distribution is a diagonal multivariate normal distribution and feature maps of each sample is no longer deterministic, but randomly sampled from $\mathcal{N}(\tilde{\mathbf{x}}; \tilde{\mathbf{x}}_\mu, \tilde{\mathbf{x}}_\sigma^2 \mathrm{I})$. However, the sampling operation is not differentiable, thereby preventing the backpropagation of the gradient flow during training. We use the reparameterization technique Kingma and Welling (2013) to make the model still use gradients as usual. Specifically, we first sample a random noise $\epsilon$ independent of the model parameters from $\mathcal{N}(\mathbf{0}, \mathrm{I})$, and then treat $\tilde{\mathbf{x}}$ as an equivalent sampling feature map,

$$\tilde{\mathbf{x}} = \tilde{\mathbf{x}}_\mu + \epsilon \tilde{\mathbf{x}}_\sigma, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathrm{I}) \tag{3}$$

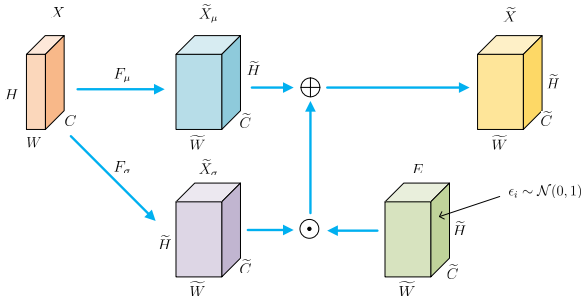


Figure 2: A variational information bottleneck cell. **Best viewed in color**.

The basic structure of cellVIB net is shown in Figure 2. For any given transformation $\mathbf{F}_\mu : \mathbf{X} \to \tilde{\mathbf{X}}_\mu, \mathbf{X} \in \mathbb{R}^{H\times W\times C}, \tilde{\mathbf{X}}_\mu \in \mathbb{R}^{\tilde{H}\times\tilde{W}\times\tilde{C}}$ (for example, a convolutional layer or a convolutional block composed of a series of convolutions), we can construct a corresponding VIB cell as shown below. First, introduce a new transformation $\mathbf{F}_\sigma : \mathbf{X} \to \tilde{\mathbf{X}}_\sigma, \mathbf{X} \in \mathbb{R}^{H\times W\times C}, \tilde{\mathbf{X}}_\sigma \in \mathbb{R}^{\tilde{H}\times\tilde{W}\times\tilde{C}}$ , the complexity of $\mathbf{F}_\sigma$ can be lower than $\mathbf{F}_\mu$ to obtain another feature map equal in size to $\mathbf{X}_\mu$ as the standard deviation term. Second, we sample $E$ on a normal distribution. Finally, we input feature maps $\tilde{\mathbf{X}} = \tilde{\mathbf{X}}_\mu + E \cdot \tilde{\mathbf{X}}_\sigma$ into the subsequent layers.

### 3.2. KL-Divergence Regularization

Eq.(3) shows that during the training process, feature map $\tilde{\mathbf{x}}_\mu$ is affected by $\tilde{\mathbf{x}}_\sigma$, which will prompt the model to predict a small $\tilde{\mathbf{x}}_\sigma$ for all samples to suppress the unstable component in $\tilde{\mathbf{x}}$. In this case, the random feature map will degrade to $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}_\mu$, which is actually the original deterministic feature map. Inspired by the variational information bottleneck Alemi et al. (2016), we explicitly constrain $\mathcal{N}(\tilde{\mathbf{x}}_\mu, \tilde{\mathbf{x}}_\sigma^2)$ in the optimization process to make it close to the normal distribution $\mathcal{N}(\mathbf{0}, \mathrm{I})$, measuring by Kullback-Leibler divergence between them. The Kullback-Leibler divergence term is,

$$\mathcal{L}_{kl} = KL[\mathcal{N}(\tilde{\mathbf{x}}|\tilde{\mathbf{x}}_\mu, \tilde{\mathbf{x}}_\sigma^2)||\mathcal{N}(\epsilon|\mathbf{0}, \mathrm{I})] \tag{4}$$

In Eq.(4), $\mathcal{L}_{kl}$ is a good balancer. In particular, $\tilde{\mathbf{x}}$ is not encouraged to predict the large variance of all samples, which may lead to extreme deterioration of $\tilde{\mathbf{x}}_\mu$, making it difficult for the model to converge. At the same time, $\tilde{\mathbf{x}}$ is not encouraged to predict low variance over all samples, which may in turn lead to a larger $\mathcal{L}_{kl}$ penalty.

### 3.3. Variational Information Bottleneck in Cell

The input variable of a neural network with $I$ layers is represented as $\mathbf{x}$, and the related target output is represented as $\mathbf{y}$. We denote the activation of the hidden layer of the network as $\{\tilde{\mathbf{x}}_i\}_{i=1}^I$.

Now the feedforward network layer can be interpreted as a Markov chain of successive representation Tishby and Zaslavsky (2015), that is,

$$\mathbf{y} \to \mathbf{x} \to \tilde{\mathbf{x}}_1 \to ... \to \tilde{\mathbf{x}}_I \to \hat{\mathbf{y}} \tag{5}$$

Every hidden layer in the network defines a conditional probability $p(\tilde{\mathbf{x}}_i|\tilde{\mathbf{x}}_{i-1})$. For convenience, we use $\mathbf{x} = \tilde{\mathbf{x}}_0$ and $\mathbf{z} = \tilde{\mathbf{x}}_I$.

For the deterministic network model, $p(\tilde{\mathbf{x}}_i|\tilde{\mathbf{x}}_{i-1})$ is absolutely certain. In this situation, the function of the hidden layer is to extract information from the previous layer, while the output layer tries to approximate the real distribution $p(\mathbf{y}|\tilde{\mathbf{x}}_I)$.

Deep VIB introduces a variational information bottleneck in the output layer. On the one hand, it can reduce $I(\tilde{\mathbf{x}}_I, \mathbf{x})$ to remove the redundant content. On the other hand, it can incorporate uncertainty during training. However, only add information bottleneck in the last layer is not sufficient, regardless of to remove redundant information or to add uncertainty.

Our starting point for achieving this goal is to clearly punish the information theory measure of redundancy between each adjacent layer. More specifically, for each hidden layer $\tilde{\mathbf{x}}_i$, we want to minimize the mutual information $I(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i-1})$ between $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_{i-1}$, eliminating the redundancy between layers and maximizing the mutual information $I(\tilde{\mathbf{x}}_i, \mathbf{y})$. Therefore, the objective of layer $\mathcal{L}_i$ becomes

$$\mathcal{L}_i = I(\tilde{\mathbf{x}}_i, \mathbf{y}) - \beta_i I(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i-1}) \tag{6}$$

Among them, $\beta_i \geq 0$ is a coefficient, which determines the strength of the bottleneck. To sum each layer, the goal is to maximize $\sum_i \mathcal{L}_i$.

### 3.4. Cell VIB Applied in Modern Architectures

Applying VIB cells to AlexNet Krizhevsky et al. (2012) or VGGNet Simonyan and Zisserman (2014) is simple. The flexibility of VIB cells means that they can be directly applied to transformations other than standard convolution. To illustrate this point, we develop cellVIB networks by integrating VIB cells into modern architectures with complex design. Figure 3 describes the architecture of the cell-VIB-ResNet module. An equivalent sampling feature map of VIB cells is treated as a non-identity branch of the residual module. The reparameterization trick works before summing with the identity branch.

## 4. Experiments

### 4.1. Experiments of Mutual Information

#### 4.1.1. Dataset and Architecture

We start with experiments on unmodified MNIST LeCun et al. (1998) and use the same architecture as Pereyra et al. (2017), namely, an MLP with fully connected layer in the form of $784 - 1024 - 1024 - 10$ and activated by ReLU.

In our method, each cell in the random encoder has the following form

$$p(\tilde{x}|x) = \mathcal{N}(\tilde{x}|f_\mu^i(x), f_\sigma^i(x)), i = 1, 2, 3 \tag{7}$$

where $i$ is the depth of the network. Among them, $f_\mu^i(x)$ is the MLP cell with the format of $784 - 1024$, $1024 - 1024$, and $1024 - K$ respectively, and the output of $f_\mu^3(x)$ is $\tilde{\mathbf{x}}_\mu$, where $K$ is the size of the bottleneck. $f_\sigma^i(x)$ has the same structure as $f_\mu^i(x)$ and outputs encode $\tilde{\mathbf{x}}_\sigma$ after a softplus transform.

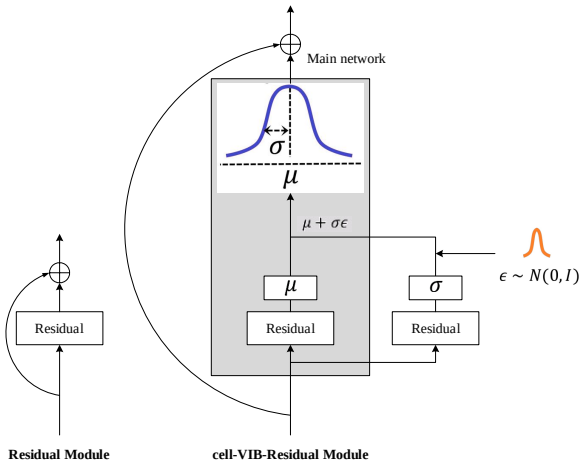


Figure 3: **Left**: The architecture of Residual module. **Right**: our cell-VIB-Residual module. **Best viewed in color**.

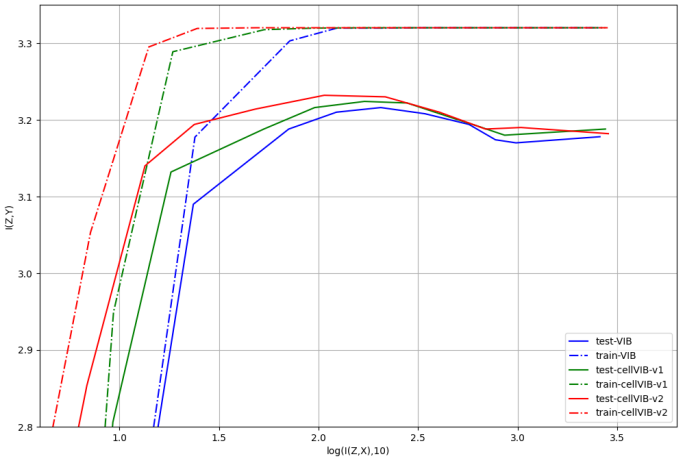


Figure 4: $I(Z, Y)$ vs $I(Z, X)$ of VIB and cellVIB as we vary $\beta$ for $K = 256$. **Best viewed in color**.

The classifier $f_{cls}(z)$ maps the K-dimensional latent code to the $C = 10$ classes, followed by a simple logistic regression model in the form of $q(y|z) = Softmax(y|f_{cls}(z))$. In the latter part, we will consider more complex architecture in other experiments, but here, we want to show the advantages of cellVIB in a simple setting.

#### 4.1.2. Results and Discussion

In Figure 4, we draw the mutual information curve of original VIB (blue curve), cellVIB-v1 (the second and third fully connected blocks replaced by VIB cells, green curve) and

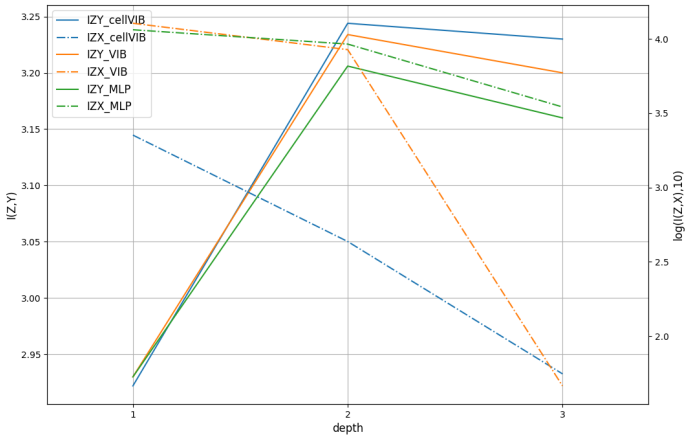


Figure 5: $I(Z,X)$ and $I(Z,Y)$ vs *depth* of MLP,VIB and cellVIB for $K = 256$. **Best viewed in color**.

cellVIB-v2 (all fully connected blocks replaced by VIB cells, red curve), i.e., we plot $I(Z,Y)$ vs $I(Z,X)$ as we vary $\beta$. Notice that the $x$-axis is a logarithimc one.

We can observe that the three different structures have some common characteristics. When we allow more information from the input to the bottleneck (by reducing $\beta$), the mutual information between the embedding and the labels increases on the training set, but not necessarily on the testing set. In other words, increasing $I(Z,X)$ helps the performance of the training set, but may lead to overfitting, which can be clearly seen from Figure 4.

On the other hand, we can observe that as the original fully connected blocks are replaced by VIB cells gradually, the regularization effect of the network becomes stronger. In three different structures, when $I(Z,X)$, the upper bound on the mutual information between the images $X$ and the stochastic encoding $Z$, is equal, the $I(Z,Y)$ value of cellVIB-v2 is greater, that is, the mutual information between encoding Z and label Y is greater, and the $I(Z,Y)$ value of the original VIB is the smallest. It can be seen from Eq.(6) that cellVIB is obviously easier to achieve a more ideal optimization state compared to the original VIB.

We compare our method with the baseline MLP and the baseline variation information bottleneck model Alemi et al. (2016). We also consider the following deterministic limitations of the model. When $\beta_i \to 0 (i = 1,2,3)$, we observe that the cellVIB optimization process tends to make $f_\sigma^i(x) \to 0 (i = 1,2,3)$, so the network becomes nearly deterministic. When $\beta_i \to 0 (i = 1,2)$, we observe that cellVIB degenerates to the original VIB form, at this time, $f_\sigma^i(x) \to 0 (i = 1,2)$.

Based on the above approximation effect, we draw the changes of $I(Z,Y)$ and $I(Z,X)$ in the test phase with the network depth of MLP, VIB and cellVIB respectively, as shown in Figure 5. For the MLP network, since there is no explicit regularization constraint, $I(Z,X)$ only slightly decreases with depth, while $I(Z,Y)$ increases significantly after passing through the second layer of the network and reaches the maximum, and then decreases after passing through the third layer of the network. For the VIB network, since the network only has

regularization constraint in the final output layer, $I(Z, X)$ has a similar trend to MLP when passing the first two blocks, and it drops significantly after passing the third layer. At the same time, $I(Z, Y)$ of VIB has the same trend compared to MLP, but larger than MLP due to the influence of regularization constraint. For the cellVIB network, due to the global regularization constraints of the network, $I(Z, X)$ declines in a more gentle trend and reaches a value close to VIB in the output layer, and $I(Z, Y)$ of cellVIB is the largest of the three models.

### 4.2. Experiments of Robustness

In this section, we evaluate the robustness of the cellVIB model. In image classification tasks, the robustness of the model during training and testing is equally important. For the robustness of the model in the training phase, we consider label flipping, that is, the situation where images are mislabeled in the training set; for the robustness of the model in the testing phase, we apply different levels and types of corruption to the test images. Without loss of generality, we conduct these experiments in the CIFAR-10 dataset.

#### 4.2.1. Datasets and Architecture

CIAFR-10 Krizhevsky et al. (2009) is a widely used benchmark in the image classification task. It contains 60000 $32 \times 32$ color images of 10 classes, with 50000 for training and 10000 for testing. CIFAR-10-C Hendrycks and Dietterich (2019) is a corrupted version of CIFAR-10 test set, in which the images are processed by 19 corruption types falling into four main categories: weather, noise, blur and digital. Each corruption contains 5 severity levels, ranging from the lowest level "1" to the highest level "5".

We follow the settings in the Wide Residual Networks(WRN) Zagoruyko and Komodakis (2016) with 16 layers and the width is 4. The training images are zero-padded by 4 pixels and randomly cropped to $32 \times 32$. All the input data subtracts a mean value and is divided by standard deviation per channel for normalization.

#### 4.2.2. Testing Image Corruption

We train the network on CIFAR-10 training set and report the averaged top-1 accuracy under 5 severity levels on the CIFAR-10-C. WRN can be viewed as the baseline model, VIB and our proposed cellVIB are based on the WRN architecture, the results are summarized in the Table 1.

As shown in Table 1, both VIB and cellVIB outperform the baseline method, indicating that VIB method can enhance the robustness in the testing phase. Moreover, the proposed cellVIB is superior to VIB by a large margin, which suggests that cellVIB has natural advantages in strengthening model's robustness against test image corruption.

#### 4.2.3. Training Label Flips

In real world classification scenario, labeling huge dataset for training is laborious and costly. An alternative approach is collecting some noisy dataset with keyword searching, allowing some mislabeled example. We follow the settings of paper Sukhbaatar et al. (2014) on CIFAR-10 dataset, and set the confusion matrix $Q$ as follows:

| Method | Lv 1 | Lv 2 | Lv 3 | Lv 4 | Lv 5 |
|---|---|---|---|---|---|
| WRN | 85.9 | 79.0 | 73.1 | 65.9 | 53.8 |
| Deep VIB | 86.4 | 79.6 | 73.7 | 66.8 | 54.7 |
| cellVIB | **88.5** | **81.2** | **75.5** | **68.7** | **56.8** |
| ↑ WRN | 2.6 | 2.2 | 2.4 | 2.8 | 3.0 |
| ↑ Deep VIB | 2.1 | 1.6 | 1.8 | 1.9 | 2.1 |

Table 1: Averaged top-1 accuracy (%) under 5 severity levels, Lv means Level and ↑ means performance boost. As the severity increases, the gain is more significant.

$$Q = \begin{pmatrix} 1-r & \frac{r}{9} & \cdots & \frac{r}{9} \\ \frac{r}{9} & 1-r & \cdots & \frac{r}{9} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{r}{9} & \frac{r}{9} & \cdots & 1-r \end{pmatrix}_{10\times 10} \tag{8}$$

where $r$ is the mislabeled ratio in the training set. We compare cellVIB with WRN and VIB under different noisy level $r$, demonstrating the noise resistant property of our proposed method. Top-1 accuracy is reported in Table 2.

| Noise Level | 0% | 10% | 30% | 50% | 70% |
|---|---|---|---|---|---|
| WRN | 95.00 | 90.41 | 87.81 | 82.52 | 69.62 |
| Deep VIB | 95.08 | 91.02 | 87.98 | 83.24 | 70.72 |
| cellVIB | **95.12** | **92.03** | **89.54** | **85.17** | **72.55** |
| ↑ WRN | 0.12 | 1.62 | 1.73 | 2.65 | 2.93 |
| ↑ Deep VIB | 0.04 | 1.01 | 1.56 | 1.93 | 1.83 |

Table 2: Top-1 accuracy (%) on CIFAR-10 test set trained with different noise level.

The results suggest that both VIB and cellVIB can resist the noisy label in the training set while baseline model collapses rapidly as the noise level increases. CellVIB is more powerful in resisting training label noise.

#### 4.2.4. How CellVIB Enhance Model's Robustness?

As stated in Alemi et al. (2018), VIB introduces uncertainty in the feature layer, which has two implications for robustness. First, for the noise of the image, regardless of the type of noise, the errors all increase with the norm of the weight matrix through error propagationTsai et al. (2021), and there is a definite upper bound. Second, for label noise, the uncertainty introduced by VIB in the feature layer is reflected in the logits. When calculating the loss function (such as cross entropy, etc.) with the label, the effect is actually equivalent to the soft label, which has a positive effect on the robustness of label noise.

As demonstrated in Tsai et al. (2021), in joint perturbation, the upper bound of the predicted label probability difference is larger for full-layer perturbation compared to single-layer perturbation, which shows that cellVIB has stronger robustness to both image noise and label noise and is consistent with our experimental conclusions.

### 4.3. Experiments of Generalization

#### 4.3.1. Datasets and Architecture

We also conduct experiments on the PACS dataset Li et al. (2017) to demonstrate that cell VIB performs well in generalization. As a commonly used domain generalization benchmark, PACS has dramatic inter-domain shift across four domains: Art Painting, Cartoon, Photo, and Sketch. It consists of 9991 images in 7 common categories: "dog", "elephant", "giraffe", "guitar", "horse", "house" and "person". Generalization of algorithms is measured by training on any three of the four domains and testing on the rest one. Following Li et al. (2017), we randomly choose 20% images from the training domains as validation set. We take the ImageNet pre-trained ResNet-18 as a base model and finetune the whole network on training domains.

#### 4.3.2. Results and Discussion

In Table 3, we compare our method with Deep VIB. In most domains, except for "photo", our method has better performance than VIB, which shows that cellVIB does help to further improve the generalization performance of the model.

| Domain | Art painting | Cartoon | Photo | Sketch | Ave |
|---|---|---|---|---|---|
| baseline | 83.61 | 78.18 | 94.94 | 78.22 | 83.74 |
| Deep VIB | 85.30 | 78.02 | **95.68** | 77.22 | 84.05 |
| cellVIB | **85.47** | **78.42** | 95.60 | **78.46** | **84.49** |
| ↑ baseline | 1.86 | 0.24 | 0.66 | 0.24 | 0.75 |
| ↑ Deep VIB | 0.17 | 0.40 | -0.08 | 1.24 | 0.44 |

Table 3: Top-1 accuracy (%) on PACAS test set.

The better generalization performance of cellVIB can be explained from two perspectives. First, the regularization introduced between feature maps further reduces the information redundancy between layers, making the mutual information between the final representation $Z$ and the input $X$ smaller. Second, the uncertainty introduced between feature maps makes it difficult for the model to overfit on the training set, thereby forcing the model to learn a more general representation.

### 4.4. Representation Learning

#### 4.4.1. Datasets and Architecture

We use a cleaned version of the MS-Celeb-1M datasets Guo et al. (2016) as our training set, which contains 3,648,176 images in 79,891 identities. Note that we follow the lists

Wang et al. (2019a,b) to remove the overlapped identities between the employed training datasets and the test datasets. 2 benchmarks including LFW Huang et al. (2008) and Agedb-30 Moschoglou et al. (2017), 3 unconstrained benchmarks: CFP Sengupta et al. (2016)[1], IJB-B Whitelam et al. (2017) and IJB-C Maze et al. (2018), and 2 renovations of LFW: CALFW Zheng et al. (2017) and CPLFW Zheng and Deng (2018), are used to evaluate the performance of cellVIB following the standard evaluation protocols.

We construct a face image ($112 \times 112$) by warping a face region using three facial points: the two eyes and the midpoint of the two corners of the mouth. We employ the modified 100-layer ResNet He et al. (2016a) as the backbone network. The head of the baseline model is: BackBone-Flatten-FC-BN with embedding dimensions of 512 and dropout probability of 0.4 to output the embedding feature.

#### 4.4.2. Compared Methods

The original ArcFace Deng et al. (2019) is used as the baseline. In addition, we compared our method with the original Deep VIB. In the training process, both the embedding features and the weights in the classifier are L2-normalized, and cosine similarity is used for evaluation. We re-implement these methods in accordance with every detail in the original literature, and make fair comparisons under the same experimental settings.

| Method | LFW | Agedb-30 | CFP-FP | CALFW | CPLFW | IJB-B(TPR@FPR) | | | IJB-C(TPR@FPR) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 0.001% | 0.01% | 0.1% | 0.001% | 0.01% | 0.1% |
| ArcFace | 99.77 | 97.90 | 98.14 | 96.10 | 92.77 | 88.97 | 94.72 | 96.59 | 93.54 | 96.11 | **97.51** |
| Deep VIB | 99.78 | **98.08** | 98.07 | **96.18** | 92.58 | 87.69 | 94.74 | 96.47 | 93.75 | 96.08 | 97.41 |
| cellVIB | **99.83** | **98.08** | **98.27** | 96.14 | **92.85** | **89.65** | **94.82** | **96.63** | **93.94** | **96.15** | 97.46 |

Table 4: Results of models (ResNet100) trained on MS-Celeb-1M. The ArcFace model outputs a deterministic embeddings. The better performance among each base model are shown in bold numbers.

#### 4.4.3. Evaluation on General Datasets

We compare our method with the baseline and Deep VIB on the general face recognition test set (that is, the test set with limited changes within the group). Table 4 summarizes the results of these evaluations. It is worth noting that on these test sets, the performance of the baseline model has almost reached saturation, and the advantages of cellVIB are not obvious, but cellVIB still slightly improve the accuracy of some of the test sets.

#### 4.4.4. Evaluation on Mixed-quality Datasets

When evaluating challenging datasets, such as IJB-B and IJB-C, these datasets have a large domain gap with high-quality training datasets. At this time, the cellVIB model has better performance compared with the baseline model and Deep VIB, especially when the false acceptance rate is low, as shown in the Table 4.

1. Noted that we only use "frontal-profile" protocol of CFP

#### 4.4.5. Discussion

In most benchmark tests, our proposed method outperforms the benchmark deterministic model. These results show that, compared with the point embedding estimated by the baseline model and the normal distribution of a single sample estimated by Deep VIB, embedding estimated by feature maps with uncertainty has better intra-class compactness and inter-class separability, especially in unconstrained benchmarks: such as CFP with front/side photos, and IJB-C which contains blurred photos collected from YouTube videos. CellVIB has made the most significant progress in the IJB benchmark verification protocol, which is also the most challenging one. This shows that the model with uncertainty is more suitable for unconstrained face recognition scenarios than the deterministic model.

It is worth mentioning that, in the current face recognition task, because the number of hyperparameters in cellVIB is very large, and the hyperparameters are all artificially set, cellVIB may still not achieve the best performance in the above-mentioned backbone. We will try to use reinforcement learning to explore the automatic setting of $\beta_i$ in our future work. We believe this will further improve the performance of cellVIB in complex models.

## 5. Conclusion

We propose Cell Variational Information Bottleneck (cellVIB), which has a novel cell structure, being able to directly replace the intermediate structure of the existing networks. We use a distribution to replace the original feature maps during training, so that the information redundancy between layers is less, and the uncertainty during training is more. Extensive experiments have proved the effectiveness of our cellVIB, which achieve competitive performance on multiple different tasks and datasets. In addition, cellVIB also has better robustness and generalization, and has strong adaptability to noisy data or labels and unknown domains. In future work, we will further explore ways to reduce the number of hyperparameters or use reinforcement learning to automatically learn hyperparameters to further improve the practicality of our model.

## References

David Barber Felix Agakov. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16:201, 2004.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Alexander A Alemi, Ian Fischer, and Joshua V Dillon. Uncertainty in the variational information bottleneck. *arXiv preprint arXiv:1807.00906*, 2018.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016b.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Irina Higgins, Loic 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. 2016.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008.

Yani Ioannou, Duncan Robertson, Roberto Cipolla, and Antonio Criminisi. Deep roots: Improving cnn efficiency with hierarchical filter groups. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1231–1240, 2017.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Speeding up convolutional neural networks with low rank expansions. *arXiv preprint arXiv:1405.3866*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

Hanxiao Liu, Karen Simonyan, Oriol Vinyals, Chrisantha Fernando, and Koray Kavukcuoglu. Hierarchical representations for efficient architecture search. *arXiv preprint arXiv:1711.00436*, 2017.

Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. *arXiv preprint arXiv:1509.08731*, 2015.

Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 51–59, 2017.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. *arXiv preprint arXiv:1507.06228*, 2015.

Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Yu-Lin Tsai, Chia-Yi Hsu, Chia-Mu Yu, and Pin-Yu Chen. Non-singular adversarial robustness of neural networks. *arXiv preprint arXiv:2102.11935*, 2021.

Xiaobo Wang, Shuo Wang, Jun Wang, Hailin Shi, and Tao Mei. Co-mining: Deep face recognition with noisy labels. In *Proceedings of the IEEE international conference on computer vision*, pages 9358–9367, 2019a.

Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. Mis-classified vector guided softmax loss for face recognition. *arXiv preprint arXiv:1912.00833*, 2019b.

Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus benchmark-b face dataset. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 90–98, 2017.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5, 2018.

Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.