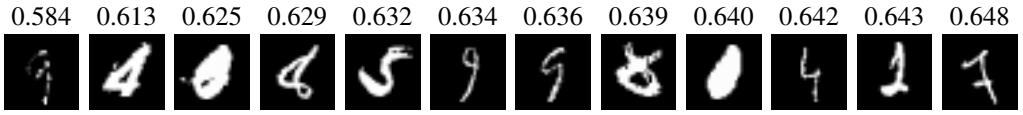
A. Uncertainty Estimation of VIB

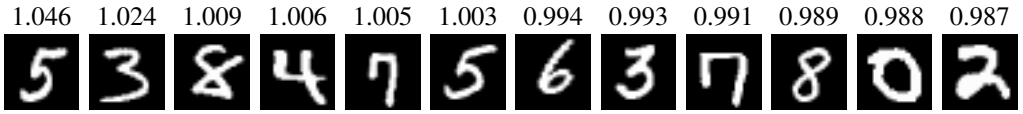
Fig. S1 shows the QMNIST test set samples with the lowest and highest variance averaged over feature dimension, i.e., $\frac{1}{D}\sum_{d=1}^{D}\sigma_{\theta,d}^2$, estimated by VIB. It shows that VIB assigns relatively smaller variances for uncertain inputs and higher variances for clear inputs. Moreover, Fig. S2 shows the risk-controlled classification (see Sec. 4.1 in the main paper) performance of VIB with the ascending and descending order of the average variance. In the ascending order, the samples with small variances are rejected first. Similarly, the high variance samples are rejected first in the descending order case. If we assume that the variance proportionally represents the 'uncertainty' associated with the inputs, we will reject the samples with the higher variance first. In this case of the descending order rejection, the risk-controlled classification did not work as expected; the more samples were rejected, the worse performance was achieved. On the contrary, the risk-controlled classification worked correctly using the ascending order, regardless of its performance. Therefore, we can infer from these results that it is more rational to utilize the variance σ_θ^2 in VIB to represent a 'confident' area/interval rather than the 'uncertainty'.

B. Cross-entropy $H(p_\theta(z|x_n), r(z))$

Let two D-dimensional multivariate Gaussian distributions $p_\theta(z|x_n) = \mathcal{N}(z|\,\mu_\theta, \Sigma_\theta)$ and $r(z) = \mathcal{N}(z; \mu_\mathbf{0}, \Sigma_\mathbf{I})$ where $\mu_\theta = f_\theta^\mu(x_n)$, $\Sigma_\theta = \mathrm{diag}(\sigma_\theta^2) = f_\theta^\Sigma(x_n)$, a diagonal covariance, $\mu_\mathbf{0} = \mathbf{0}$, a zero mean vector, and $\Sigma_\mathbf{I} = \mathbf{I}$, an identity



(a) Low average variance



(b) High average variance

Figure S1. Sample QMNIST images with the uncertainty estimated by VIB.

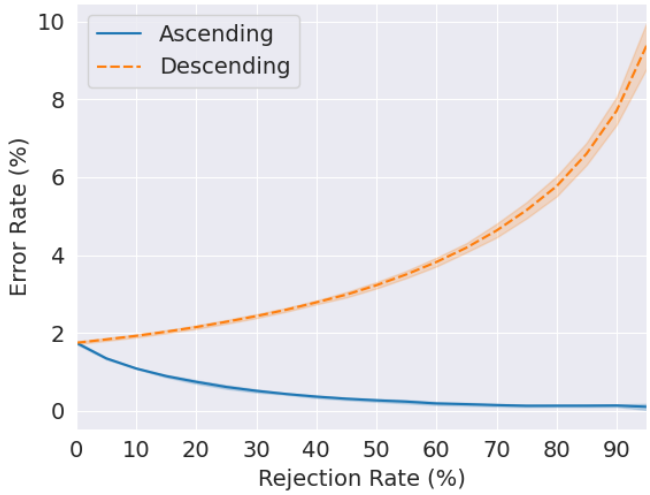


Figure S2. Risk-controlled classification performance of VIB with the ascending and descending order of the average variance.

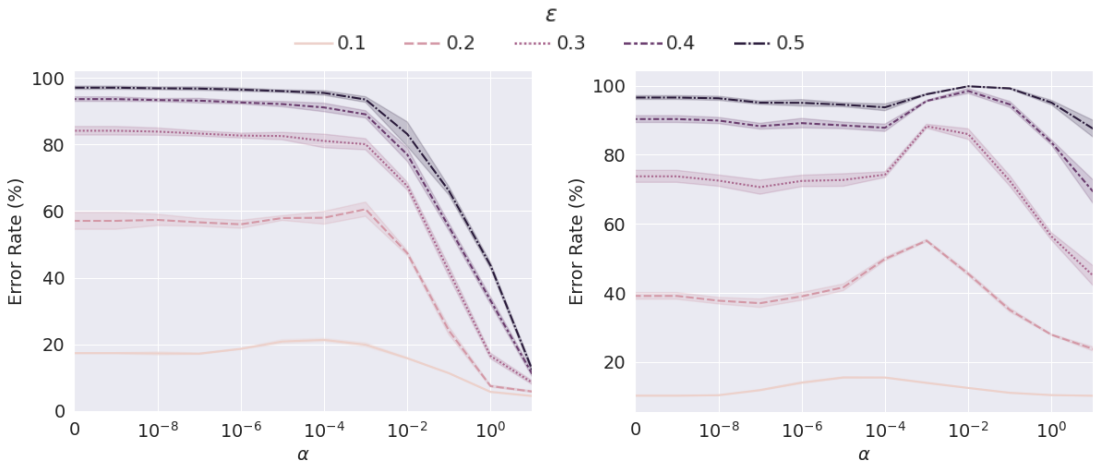
covariance matrix, respectively. Then, $H(p(z|x_n), r(z))$, the cross-entropy of $r(z)$ relative to $p(z|x_n)$, is given by

$$
\begin{aligned}
H(p(z|x_n), r(z)) &= -\int p(z|x_n) \ln r(z)\, dz && (13)\\
&= \frac{1}{2}\int \mathcal{N}(z|\,\mu_\theta, \Sigma_\theta)\Big(D\ln(2\pi) + \ln|\,\Sigma_{\mathbf{I}}\,| + (z-\mu_{\mathbf{0}})^\top \Sigma_{\mathbf{I}}^{-1}(z-\mu_{\mathbf{0}})\Big)\, dz && (14)\\
&= \frac{1}{2}\int \mathcal{N}(z|\,\mu_\theta, \Sigma_\theta)\Big(D\ln(2\pi) + \ln|\mathbf{I}| + (z-\mathbf{0})^\top \mathbf{I}^{-1}(z-\mathbf{0})\Big)\, dz && (15)\\
&= \frac{1}{2}\int \mathcal{N}(z|\,\mu_\theta, \Sigma_\theta)\Big(D\ln(2\pi) + z^\top z\Big)\, dz && (16)\\
&= \frac{1}{2}\left(D\ln(2\pi)\int \mathcal{N}(z|\,\mu_\theta, \Sigma_\theta)\, dz + \int \mathcal{N}(z|\,\mu_\theta, \Sigma_\theta)(z^\top z)\, dz\right) && (17)\\
&= \frac{1}{2}\left(D\ln(2\pi) + \mathbb{E}_{z\sim p_\theta(z|x_n)}\left[z^\top z\right]\right) && (18)\\
&= \frac{1}{2}\left(D\ln(2\pi) + \mathbb{E}_{z\sim p_\theta(z|x_n)}\left[\sum_{d=1}^{D} z_d^2\right]\right) && (19)\\
&= \frac{1}{2}\left(D\ln(2\pi) + \sum_{d=1}^{D}\mathbb{E}_{z\sim p_\theta(z|x_n)}\left[z_d^2\right]\right) && (20)\\
&= \frac{1}{2}\left(D\ln(2\pi) + \sum_{d=1}^{D}\left(\mathbb{E}_{z\sim p_\theta(z|x_n)}\left[z_d\right]^2 + \mathrm{Var}_{z\sim p_\theta(z|x_n)}\left[z_d\right]\right)\right) && (21)\\
&= \frac{1}{2}\left(D\ln(2\pi) + \sum_{d=1}^{D}\left(\mu_{\theta,d}^2 + \sigma_{\theta,d}^2\right)\right) \geq 0 && (22)
\end{aligned}
$$

where each of z_d, $\mu_{\theta,d}$, and $\sigma_{\theta,d}^2$ is the d-th element of the corresponding vector. Note again that $\Sigma_\theta = \mathrm{diag}(\sigma_\theta^2) = f_\theta^\Sigma(x_n)$, a diagonal covariance, i.e., each dimension of z is independent to each other.

C. Perturbation Robustness by Different α

Fig. S3a shows the misclassification rates against the FGSM perturbations for MEIB trained with different α values. It shows a more consistent trend that a larger α provides more robustness for the given perturbation strength, unlike the VIB



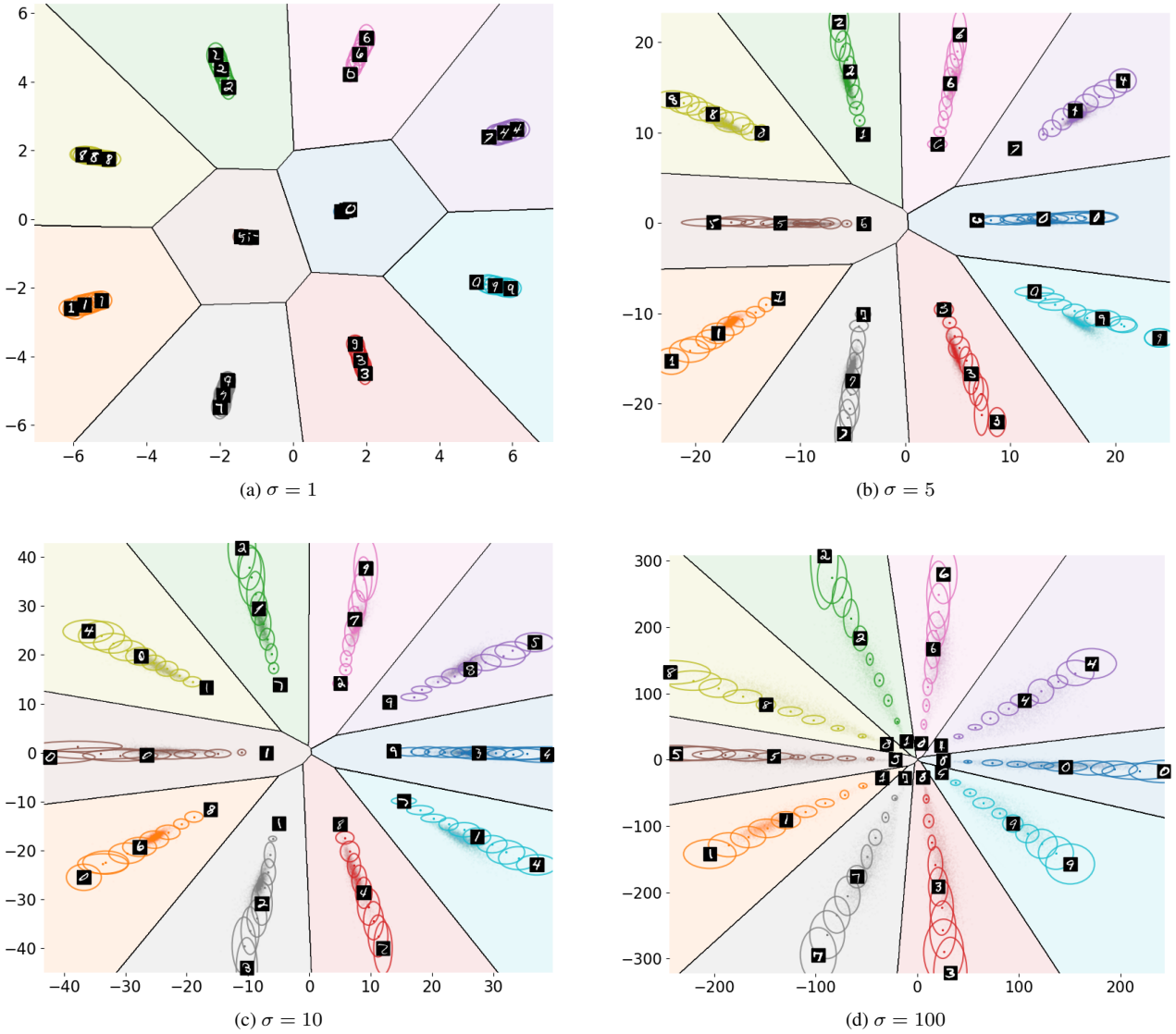
(a) MEIB (b) VIB

Figure S3. Adversarial robustness with different α values.

case that shows saddle-shaped curves in Fig. S3b. It is reasonable because a larger α (thus, a larger β) value encourages the model to assign larger entropy and thus secure a larger area for the distribution of z given input x, which leads to larger margins from the decision boundary.

D. VIB with Large-variance Priors

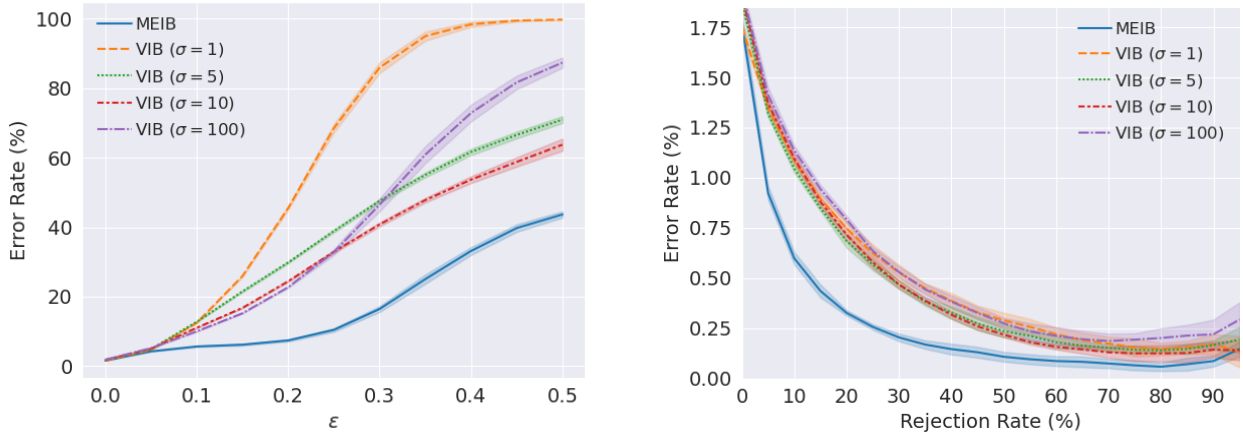
We trained and evaluated VIB models with different values of σ in their prior distribution. Fig. S4 shows the embeddings of VIBs trained with 2-dimensional bottleneck and different values of σ. By visual inspection, it may seem getting similar to those of MEIB shown in Fig. 4b as σ increases. On the other hand, Fig. S5 presents the perturbation robustness (toward FGSM attacks) and risk-controlled classification performance by VIB trained with the different values of σ while keeping the other configurations same as in Sec. 4, including the 256-dimensional bottleneck. It shows that the adversarial robustness is improved with increasing σ, until $\sigma = 10$, but becomes worse again with much larger $\sigma = 100$, while all of them were still worse than MEIB with a significant gap. Moreover, the risk-controlled classification performance has little difference with different σ values. On the other hand, the magnitude of σs estimated by MEIB for QMNIST dataset are in a range below than those by VIB with $\sigma = 5$, as shown in Tab. S1. It provides an insight that MEIB increase σ of inputs in a more effective way.



(a) $\sigma = 1$ (b) $\sigma = 5$

(c) $\sigma = 10$ (d) $\sigma = 100$

Figure S4. 2D embedding space learned for the QMNIST dataset by VIB with different σ of the prior distributions. The ellipses represent the standard deviation of the stochastic embeddings for a subset of training data.



(a) Perturbation Robustness

(b) Risk-controlled Classification

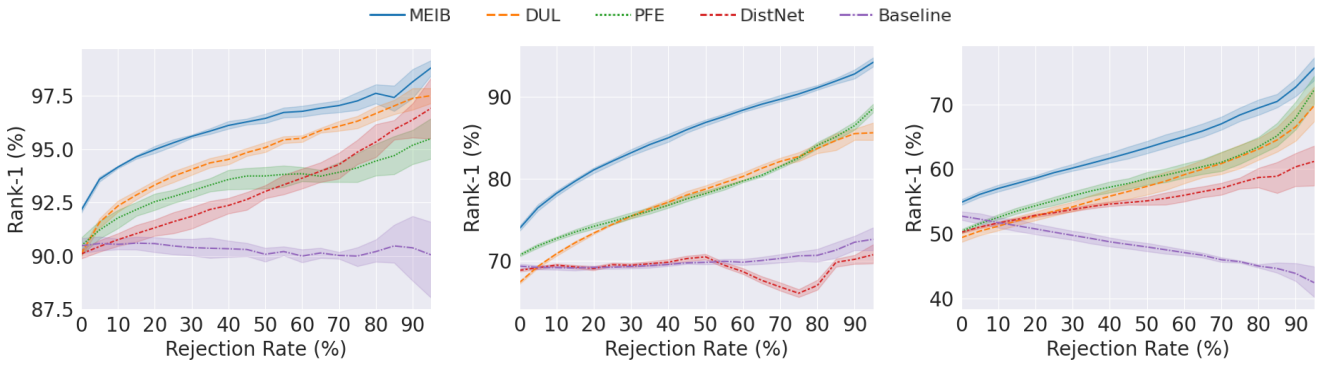
Figure S5. Performance comparisons on the QMNIST dataset.

Table S1. $\| \sigma \|_2$ estimated by each method

Method	Mean	SD	Min	Max
VIB ($\sigma = 1$)	0.38	0.08	0.18	0.53
VIB ($\sigma = 5$)	1.80	0.39	0.56	4.03
VIB ($\sigma = 10$)	3.22	0.84	0.70	9.33
VIB ($\sigma = 100$)	18.77	8.01	2.04	81.42
MEIB	1.10	0.28	0.45	1.95

E. Rank-1 Accuracy for Risk-controlled Person ReID

Fig. S6 shows the risk-controlled CMC rank-1 accuracy for each method on each ReID dataset. The results were obtained from the same experiments described in Sec. 4.3, in addition to the risk-controlled mAP in Fig. 5 of the main paper. It confirms again that MEIB provides the most effective confidence measure and risk-control capability compared to the other methods over the most range of the rejection rate.



(a) Market-1501

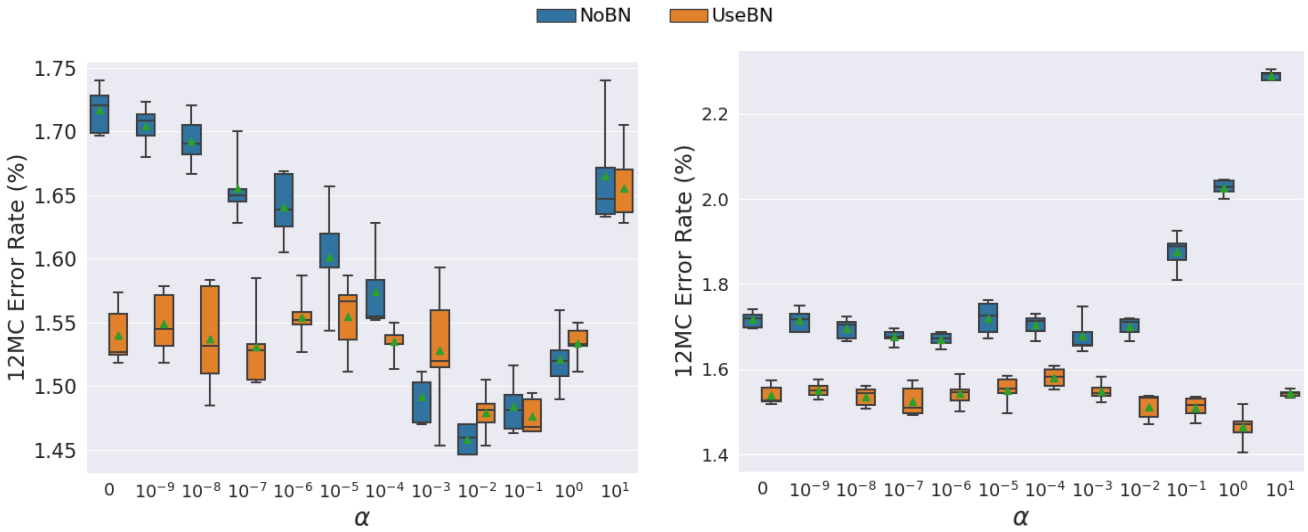
(b) MSMT17

(c) LPW

Figure S6. CMC rank-1 accuracy for risk-controlled person ReID.

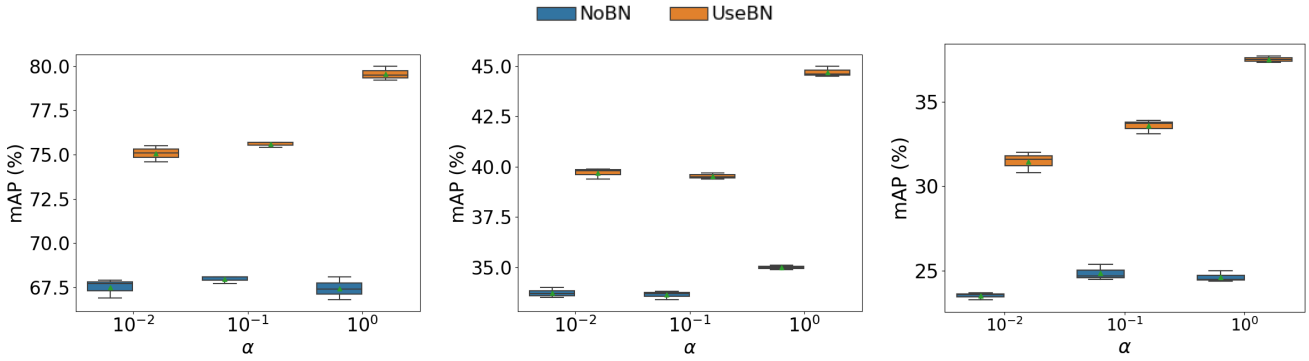
F. Hyperparameter Study

Fig. S7 shows the performance of VIB and MEIB for different values of α and the usage of the BN layer at the end of the f_θ^μ branch in the QMNIST digit classification task with $D = 256$. For VIB, the BN layer is beneficial only for the small values of α ($0 \sim 10^{-4}$), and the best result was achieved without using the BN layer at $\alpha = 0.01$. For MEIB, on the other hand, it is a common phenomenon that using the BN layer at the end of f_θ^μ noticeably improves the performance across all α values, while the best performance was obtained with $\alpha = 1$ using the BN layer. Furthermore, Fig. S8 shows that using the BN layer at the end of the f_θ^μ branch also significantly improves the performance of MEIB on all datasets considered in the person ReID task, commonly with the best performance at $\alpha = 1$. While we need a more concrete study about the effect of BN in MEIB, which we leave as our future work, it can be assumed that $\alpha = 1$ with a BN layer at the end of f_θ^μ is a reasonable default configuration for these tasks.



(a) VIB (b) MEIB

Figure S7. Hyperparameter comparison for QMNIST (lower is better).



(a) Market-1501 (b) MSMT17 (c) LPW

Figure S8. MEIB hyperparameter comparison for person ReID (higher is better).

G. Stronger Adversarial Attacks Experiments

While the FGSM evaluated in Sec. 4 is a popular first-choice baseline method, it is often ineffective compared to multi-step attack methods. Although the main contribution of this paper is not an adversarial defensive method, we evaluated MEIB

and other baseline methods on the following stronger adversaries for a more concrete adversarial robustness analysis: the multi-step Projected Gradient Descent (PGD) [30], two variants of Auto-PGD (APGD), one based on the cross-entropy loss (APGD_{CE}) and one with the difference of logits ratio (DLR) loss (APGD_{DLR}), and AutoAttack (AA), which is an ensemble of those two APGDs [12]. We used the projection on the L_∞-ball of radius ϵ for all attacks where we set $\epsilon = 0.3$. For PGD, we used 1000 steps (PGD-1000) with the step size $\epsilon_{\text{step}} = 0.01$ and 100 random restarts. For APGDs and AA, Expectation over Transformation (EOT) [5] was used with an average over 20 times of the gradient computations for more effective attacks on the stochastic models, MEIB and VIB. We used Adversarial Robustness Toolbox (ART) [32] to leverage its PGD implementation and the AA implementation provided by the authors[2], including APGDs.

Tab. S2 shows the evaluation results on class-balanced $10,000$ samples of the QMNIST test set. MEIB outperformed all the other baseline methods across all different types of adversarial attacks. Although MEIB was also effectively fooled by these multi-step and adaptive PGD attacks, it performs better than other non-adversarially trained models such as Mixup evaluated with the PGD attacks on the original MNIST dataset[1]. It would be possible to achieve better robustness once MEIB is trained with adversarial training methods.

Table S2. Classification Accuracy (%) with different adversarial attacks on the QMNIST test samples

Method	Clean	PGD-1000	APGD_{CE}	APGD_{DLR}	AA
Deterministic	98.39 ± 0.11	1.13 ± 0.04	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Dropout	$\mathbf{98.45 \pm 0.08}$	1.09 ± 0.04	0.01 ± 0.01	0.00 ± 0.00	0.00 ± 0.00
VIB	98.34 ± 0.14	1.18 ± 0.07	0.89 ± 0.05	1.20 ± 0.10	0.06 ± 0.02
MEIB	98.38 ± 0.15	$\mathbf{2.42 \pm 0.44}$	$\mathbf{6.79 \pm 1.40}$	$\mathbf{1.60 \pm 0.28}$	$\mathbf{0.12 \pm 0.06}$

H. Implementation and Training Details

QMNIST Digit Classification All models were trained for 200 epochs with a batch size of 100. Adam optimizer [26] was used with $\beta_1 = 0.5, \beta_2 = 0.999$, and the learning rate of 0.0001, which is decayed by the factor of 0.97 every other epoch. The exponential moving average of the model parameters with a decay factor of 0.999 was tracked during the training, and the final averaged parameters were used at test time. We trained and evaluated all methods with five different random seed values. All pixel values of input digit images were rescaled to $[-1, 1]$. The embedding size of $D = 256$ was used by default for all experiments unless it was specified. For the stochastic methods, MEIB and VIB, we used 12 MC samples of z for the adversarial robustness experiment and a single sample for the risk-controlled classification experiment.

Hedged Instance Embedding We used the same architecture in the original HIB work [33], except for some parts not mentioned by the authors. The backbone encoder f_θ^B consists of two convolutional layers, 32 and 64 filters of 5×5 kernels, each followed by ReLU activation and a max-pooling with 2×2 kernels. Each f_θ^μ and f_θ^Σ is an FC layer of D hidden units, and a BN layer is attached at the end of f_θ^μ. Both MEIB and VIB variants of the HIB models were trained over 500,000 iterations with a batch size of 128. We followed the same batching strategy used in [33] to ensure we had enough number of both positive and negative pairs of images. Adam optimizer [26] was used with $\beta_1 = 0.5, \beta_2 = 0.999$ and the learning rate of 0.0001.

Person Re-identification We utilized the Torchreid framework [53] to implement the methods and conduct the experiments. Except for PFE and DistNet, all other models were initialized with the ResNet50 parameters pre-trained on the ImageNet dataset [13] and trained for 60 epochs with a batch size of 32. Specifically, we adopted the two-stepped transfer learning strategy [16], where the backbone encoders f_θ^B were frozen for the first 5 epochs. AMSGrad optimizer [37] was used with the learning rate of 0.0003, which was reduced by the factor of 10 every 20 epochs. PFE and DistNet were initialized from the fully-trained deterministic baseline model and fine-tuned over another 60 epochs only for the parts specified in their original work. While the same optimizer was used, the initial learning rate was set as 0.0001, which was also reduced by the factor of 10 every 20 epochs. For all methods, each batch consists of 4 random identities and 8 random images for each identity. All input images were rescaled to 128×256, and random horizontal flipping with the probability of 0.5 was applied. We reported the mean and standard deviation of the performance for all models trained with five different random seeds.

[2] https://github.com/fra31/auto-attack

I. Misclassified QMNIST Samples

Fig. S9 shows a subset of QMNIST test images misclassified by MEIB after filter out 80% of the test set in the risk-controlled classification experiment in Sec. 4.1. Although MEIB was overconfident, i.e., assigned inappropriately high entropy values, for these uncertain images, each image has a plausible shape for the predicted class also.

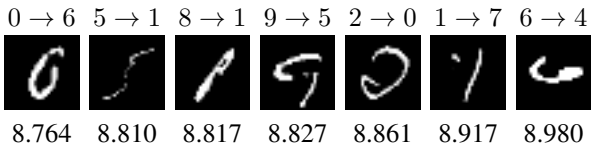
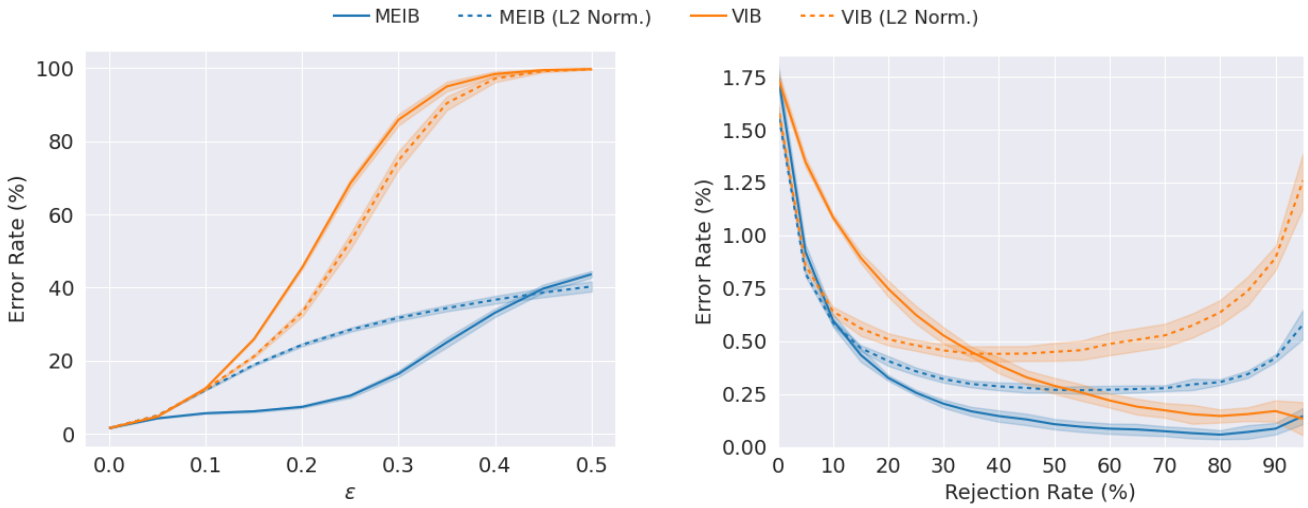


Figure S9. Sample QMNIST images misclassified by MEIB. Two digits above each image represent `Label` →`Prediction` and the number below each image is the conditional entropy estimated.

J. L2 Normalization for MEIB and VIB

L_2 normalization of feature embedding [36] is widely used with the softmax loss [44], angular margin-based loss [14], and contrastive loss [9]. We also trained and tested MEIB and VIB with L_2-normalized z samples to see its benefits on stochastic embeddings. Fig. S10 shows the perturbation robustness and risk-controlled classification performance, similar to Fig. 3, of the MEIB and VIB models with and without L_2 normalization of embeddings. MEIB performs much worse with the L_2 normalization in terms of both perturbation robustness and risk-controlled classification. While VIB shows a slightly improved robustness with L_2 normalization, its risk-controlled classification performance become worse than without the normalization shortly after about 30% of rejection rate. It could be difficult for stochastic embeddings, especially MEIB, to utilize their advantage in angular space since their magnitudes are not considered anymore. For example, the main intuition of MEIB is spreading out each input embedding as wide as possible to take more space within a decision region. However, after applying L_2 normalization, only the angle of each feature vector contributes to the decision regardless of their magnitude in the embedding space.



(a) Perturbation Robustness

(b) Risk-controlled Classification

Figure S10. Comparison of L_2-normalized MEIB and VIB on the QMNIST dataset.