Maximum Entropy Information Bottleneck for Uncertainty-aware Stochastic Embedding

Sungtae An*
Amazon
ansungt@amazon.com

Nataraj Jammalamadaka
Amazon
jammaln@lab126.com

Eunji Chong
Amazon
chongec@amazon.com

Abstract

*Stochastic embedding has several advantages over deterministic embedding, such as the capability of associating uncertainty with the resulting embedding and robustness to noisy data. This is especially useful when the input data has ambiguity (e.g., blurriness or corruption) which often happens with in-the-wild settings. Many existing methods for stochastic embedding are limited by the assumption that the embedding follows a standard normal distribution under the variational information bottleneck principle. We present a different variational approach to stochastic embedding in which maximum entropy acts as the bottleneck, which we call **Maximum Entropy Information Bottleneck** or **MEIB**. We show that models trained with the MEIB objective outperform existing methods in terms of regularization, perturbation robustness, probabilistic contrastive learning, and risk-controlled recognition performance.*

1. Introduction

Stochastic embedding is a mapping of an input x to a *random variable* $Z \sim p(z|x) \in R^D$ in which the mapped *regions* of similar inputs are placed nearby. Unlike deterministic embedding, where $z = f(x)$ is a point in R^D, stochastic embedding can represent the input uncertainty, such as data corruption or ambiguity, by controlling the spread of probability density over a manifold [33].

Figure 1 depicts a typical stochastic embedding framework with the neural networks parameterized by θ. Input x is mapped to a Gaussian distribution $\mathcal{N}(z; \mu, \Sigma)$ by a stochastic encoder that consists of a backbone feature extractor f_θ^B followed by two separate branches f_θ^μ and f_θ^Σ, each of which predicts the μ and Σ.[1]While the covariance matrix Σ, in prior work as well as in this paper, is assumed to be diagonal where f_θ^Σ outputs a D-dimensional vector, it would be straightforward to extend it to a full covariance matrix, for instance, using a Cholesky decomposition [15]. Embeddings sampled from this Gaussian are then consumed by a decoder f_θ^C for the downstream task, e.g., classification.

Majority of leading methods for stochastic embedding [7, 10, 29, 33, 43] are built upon the variational information bottleneck (VIB) principle [3] where the stochastic encoder $p(z|x)$ is regularized by Kullback–Leibler (KL) divergence, KL$(p(z|x)||r(z))$, where $p(z|x) = \mathcal{N}(z; \mu, \Sigma)$ and $r(z) = \mathcal{N}(z; \mathbf{0}, \mathbf{I})$ in general. This effectively impels the embeddings to be close to a standard normal distribution, which is an explicit assumption that may not always hold.

Furthermore, Bütepage et al. [6] showed that the standard variational autoencoder (VAE) trained with KL divergence from the standard normal prior [27] fails to correlate the latent variance with the input uncertainty; the variance decreases with the distance to the latent means of training data, which is contrary to expectation. Since VAE is a special case of an unsupervised variant of VIB, this phenomenon also holds for VIB; our experiments show VIB assigns smaller variance to more uncertain inputs (see Appendix A). Motivated by this finding, we explicitly use the variance (entropy) as a confidence indicator rather than a measure of input uncertainties and encourage the model to assign larger variance to more certain inputs.

In this paper, we propose *Maximum Entropy Information Bottleneck (MEIB)* to lift such constraints of using a fixed prior and instead use the conditional entropy of the embedding $H(Z|X)$ as the only regularization. Based on the maximum entropy principle [25], we postulate that stochastic uncertainty is best represented by the probability distribution with the largest entropy. By maximizing $H(Z|X)$, the embedding distribution is promoted to be more random, pushing for broader coverage in the embedding space, with a trade-off on the expressiveness of Z about target Y, e.g., the class labels of inputs in classification tasks. The resulting distribution is also the one that makes the fewest assumptions about the true distribution of data [40].

[*] The work was done while the author was an intern at Amazon and a PhD student at Georgia Tech.

[1] We use the terms $f_\theta^\mu(x)$ and $f_\theta^\Sigma(x)$ interchangeably with $f_\theta^\mu(f_\theta^B(x))$ and $f_\theta^\Sigma(f_\theta^B(x))$ respectively.

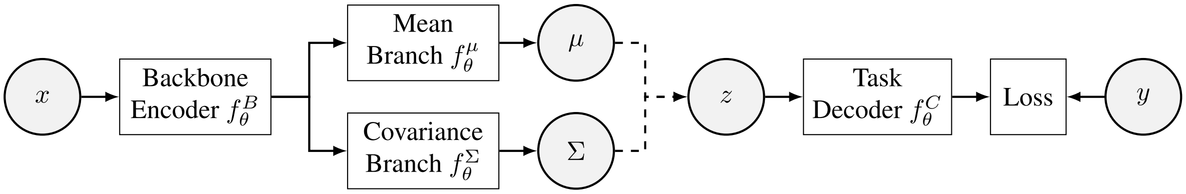
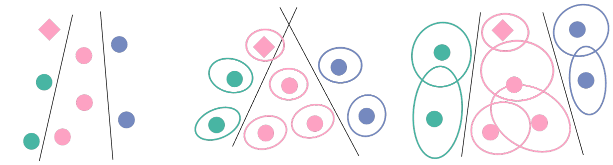

x
Backbone Encoder f_θ^B
Mean Branch f_θ^μ
Covariance Branch f_θ^Σ
μ
Σ
z
Task Decoder f_θ^C
Loss
y

Figure 1. Stochastic embedding framework.



(a) Deterministic (b) Stochastic, KLD (c) Stochastic, ME

Figure 2. Embedding space characteristics. Each color represents a class of data. The color-filled shapes refer to the deterministic or the mean point of stochastic embeddings. The ellipses around the shapes depict the standard deviation of stochastic embeddings. The circles and the diamonds represent training and testing data, respectively. The solid lines are the decision boundaries learned.

Figure 2 depicts our intuition; (a) deterministic encoders would learn embeddings “just enough” to classify the training samples unless any regularization technique, such as a margin loss, is considered. It would be vulnerable to small changes in test inputs. (b) The embedding distribution by typical stochastic encoders (e.g., VIB) trained with the KL divergence regularization will tend to cover a fixed prior. Note that it is generally difficult to pick a true prior distribution. Also, it is unnecessary to restrict the embedding distribution to be within a specific bound. (c) With MEIB, on the other hand, by maximizing the conditional entropy of the stochastic embeddings, we would have a better regularization effect as it makes the area *secured* by the embedding distribution for the given input as broad as possible.

The key contributions of MEIB to the previous stochastic embedding methods are summarized as follows:

- MEIB outperforms existing approaches in the challenging person re-identification task with three popular datasets while it provides a comparable regularization in handwritten digit classification.
- MEIB shows significantly better perturbation robustness compared to VIB in digit classification.
- MEIB performs better than VIB when used in a probabilistic contrastive learning framework.
- By providing reliable uncertainty measurements, MEIB demonstrates outstanding risk-controlled recognition performance in both digit classification and person re-identification tasks.

2. Related Work

Stochastic Embeddings Research on stochastic embeddings has gained popularity in recent years. Oh et al. [33] proposed a similarity-based method based on the pairwise soft contrastive loss and the VIB principle, which has been later applied to human pose embedding [43], cross-modal retrieval [10], and ordinal embedding [29]. Chang et al. [7] proposed stochastic embedding for face recognition using softmax loss and KL divergence regularization without relying on similarity. All of these methods assume embedding distribution to be unit Gaussian. Unlike these approaches, Probabilistic Face Embedding (PFE) [39] turns deterministic embeddings into Gaussians with fixed mean by training a post hoc network to maximize the mutual likelihood of same-class embeddings. Follow-up research on PFE includes its extension to triplets [45] and for spherical space [28]. However, this line of work is limited to fixed embedding mean. The work most closely related to ours is DistributionNet [50] which introduced entropy-based regularization and inspired Yang et al. [48] for their uncertainty-aware loss. However, these methods put a margin to bound the total entropy rather than maximizing it as we do.

Maximum Entropy Maximum entropy is a general principle that has already been widely adopted in designing machine learning models, including supervised and reinforcement learning [2, 51]. Pereyra et al. [35] used the negative entropy of the class prediction distribution, $-H(p_\theta(y|x))$, as a regularization term in the loss function to prevent overconfident predictions. In reinforcement learning, the maximum entropy framework encourages diverse explorations in both on-policy and off-policy settings [22, 23, 31]. However, in most previous work, the entropy regularization has been applied at the decision levels, the distribution of class or action predictions. In this work, on the other hand, we focus on the entropy of the stochastic embedding of inputs.

3. Maximum Entropy Information Bottleneck

MEIB Objective *The maximum entropy principle* [25] states that the current state of knowledge about the given system is best represented by the probability distribution with the largest entropy [40]. By combining this with our hypothesis, the goal is to learn an encoding Z that is maximally expressive about Y while maximizing the expected amount of information (entropy) about $Z|X$:

$$\max_{\theta} I(Z, Y; \theta) \quad \text{s.t. } H(Z|X; \theta) \geq H_c. \tag{1}$$

Introducing a Lagrangian multiplier β, we have the maximization objective:

$$\mathcal{J}_{\text{MEIB}} = I(Z, Y; \theta) + \beta H(Z|X; \theta) \tag{2}$$

where $\beta \geq 0$ controls the trade-off between the predictiveness and the spread of Z given X. Using the lower bound suggested by Alemi et al. [3] for the first term of the objective $I(Z, Y)$, we have

$$\begin{aligned} & I(Z, Y; \theta) + \beta H(Z|X; \theta) \\ & \quad \geq \int dx\, dy\, dz\, p(x) p(y|x) p(z|x) \log q(y|z) \\ & \qquad - \beta \int dx\, dz\, p(x)\, p(z|x) \log p(z|x) = L. \end{aligned} \tag{3}$$

where $q(y|z)$ is the decoder f_{θ}^{C}, which is a variational approximation to $p(y|z)$, that estimates the conditional distribution of the target y given the latent representation z. This lower bound L can be computed by approximating the joint distribution $p(x, y) = p(x)\, p(y|x)$ using the empirical data distribution $p(x, y) = \frac{1}{N} \sum_{n=1}^{N} \delta_{x_n}(x)\, \delta_{y_n}(y)$ [3]

$$\begin{aligned} L \approx \frac{1}{N} \sum_{n=1}^{N} \Bigg[\int & dz\, p(z|x_n) \log q(y_n|z) \\ & - \beta\, p(z|x_n) \log p(z|x_n) \Bigg]. \end{aligned} \tag{4}$$

Consequently, the loss function to be minimized is:

$$\begin{aligned} \mathcal{L}_{\text{MEIB}} = \frac{1}{N} \sum_{n=1}^{N} \Bigg[& \mathbb{E}_{z \sim p(z|x_n)} \Big[- \log q(y_n|z) \Big] \\ & - \beta H(Z|x_n) \Bigg] \end{aligned} \tag{5}$$

where we use the typical reparameterization trick [27] to backpropagate gradients through the sampling of $z \sim p(z|x_n)$. We use a single sample of z by default unless it is specified.

Relationship to VIB The minimization loss function of VIB [3] is:

$$\begin{aligned} \mathcal{L}_{\text{VIB}} &= \frac{1}{N} \sum_{n=1}^{N} \Big[\mathbb{E}_{z \sim p(z|x_n)} \big[- \log q(y_n|z) \big] \\ & \qquad + \beta \text{KL}\big[p(Z|x_n), r(Z) \big] \Big] \quad (6) \\ &= \frac{1}{N} \sum_{n=1}^{N} \Big[\mathbb{E}_{z \sim p(z|x_n)} \big[- \log q(y_n|z) \big] \\ & \qquad - \beta H(Z|x_n) + \beta H(p(Z|x_n), r(Z)) \Big] \quad (7) \\ &\geq \frac{1}{N} \sum_{n=1}^{N} \Big[\mathbb{E}_{z \sim p(z|x_n)} \big[- \log q(y_n|z) \big] \\ & \qquad - \beta H(Z|x_n) \Big] \quad (8) \\ &= \mathcal{L}_{\text{MEIB}} \end{aligned}$$

where $H(p(Z|x_n), r(Z))$ is the cross-entropy of $r(Z)$ relative to the distribution $p(Z|x_n) = \mathcal{N}(Z | f_{\theta}^{\mu}(x_n), f_{\theta}^{\Sigma}(x_n))$, which is given by

$$\begin{aligned} H(p(Z|x_n), r(Z)) &= \frac{1}{2} \Big(D \ln(2\pi) + \sum_{d=1}^{D} \big(\mu_{\theta,d}^2 + \sigma_{\theta,d}^2 \big) \Big) \\ & \geq 0 \end{aligned} \tag{9}$$

where $\mu_{\theta,d}$ and $\sigma_{\theta,d}^2$ are the d-th element of μ_{θ} and σ_{θ}^2, respectively, while $\mu_{\theta} = f_{\theta}^{\mu}(x_n)$ and $\Sigma_{\theta} = \text{diag}(\sigma_{\theta}^2) = f_{\theta}^{\Sigma}(x_n)$; a detailed derivation can be found in Appendix B. Therefore, the VIB loss function is an upper bound of the MEIB loss function with a positive value of β.

Confidence Measure of MEIB MEIB encourages obvious inputs that can be easily classified to take broader embedding areas by assigning larger entropy. On the contrary, the inputs closer to other classes would have smaller entropy to reduce the chance of misclassification according to the loss function. Therefore, we adopt the conditional entropy $H(Z|x)$ as our confidence measure for the input x. Since we assumed the multivariate Gaussian distribution $p(Z|x) = \mathcal{N}(Z | f_{\theta}^{\mu}(x), f_{\theta}^{\Sigma}(x))$, the entropy is given by [11]:

$$\begin{aligned} H(Z|x) &= \frac{1}{2} \ln(2\pi e)^D |\, \Sigma_{\theta} \,| \quad (10) \\ &= \frac{1}{2} \ln |\, \Sigma_{\theta} \,| + \frac{D}{2} (1 + \ln 2\pi). \quad (11) \end{aligned}$$

Specifically, we use the *dimension-wise average* conditional entropy $H(Z|x)/D$ to achieve a dimension-agnostic confidence measure.

4. Experimental Results and Discussion

In this section, we show experimental results for various tasks to demonstrate the effectiveness of MEIB in terms of the regularization, perturbation robustness, and confidence measure. All computational experiments were implemented using PyTorch [34] 1.9 with Python 3.7 on a workstation equipped with an NVIDIA® GeForce® RTX 2080 Ti graphic card. Please refer Appendix H for the implementation and training details for each experimental task.

4.1. Digit Classification

First, we evaluate MEIB on a handwritten digit classification as the simplest form of benchmark task. We use the QMNIST dataset [47] to utilize its larger set of 60,000 testing data compared to 10,000 of those in the original MNIST dataset, while it has almost identical training data. We adopt the same architecture employed in Alemi et al. [3]; the backbone encoder f_θ^B is a multilayer perceptron (MLP) with two fully-connected (FC) layers of 1024 hidden units with ReLU activations. Both f_θ^μ and f_θ^Σ are an FC layer of D hidden units where the exponential (exp) function was applied after f_θ^Σ. We found that applying a batch normalization (BN) layer at the end of f_θ^μ improves the performance of MEIB with a noticeable gap (see Appendix F). The decoder f_θ^C is an FC layer with the softmax function that outputs $p(y|x)$ over ten classes.

Using the same architecture, we compare the following embedding approaches with MEIB: a deterministic baseline, dropout [42], and VIB. The deterministic baseline represents the typical usage of neural network models without stochasticity. By omitting the variance estimator module $f_\theta^\Sigma(x)$, the embedding of input is deterministically given by $z = f_\theta^\mu(x)$. Dropout is one of the most popular regularization methods for neural networks; thus, it is considered the first benchmark regularization method to compare [3, 17]. We use the same deterministic model, but dropout is applied with the probability of 0.5 during the training time. Unlike MEIB and VIB, both deterministic models were trained only with the cross-entropy loss. We set $\beta = \alpha/D$ where α is equal to 0.01 for VIB, 0.1 for MEIB with $D = 2$, and 1 for MEIB with $D = 256$. Please refer to Appendix F for the hyperparameter study. All models were trained with five different random seeds, and we report the mean and standard deviation for each performance metric.

Table 1. QMNIST test set error rate (%)

Method	$D = 2$	$D = 256$
Deterministic	4.64 ± 0.43	1.71 ± 0.04
Dropout	4.08 ± 0.29	1.62 ± 0.02
VIB	3.29 ± 0.32	1.75 ± 0.03
MEIB	3.95 ± 0.41	1.76 ± 0.05
VIB (12 MC samples)	3.21 ± 0.32	1.45 ± 0.02
MEIB (12 MC samples)	3.31 ± 0.35	1.48 ± 0.04

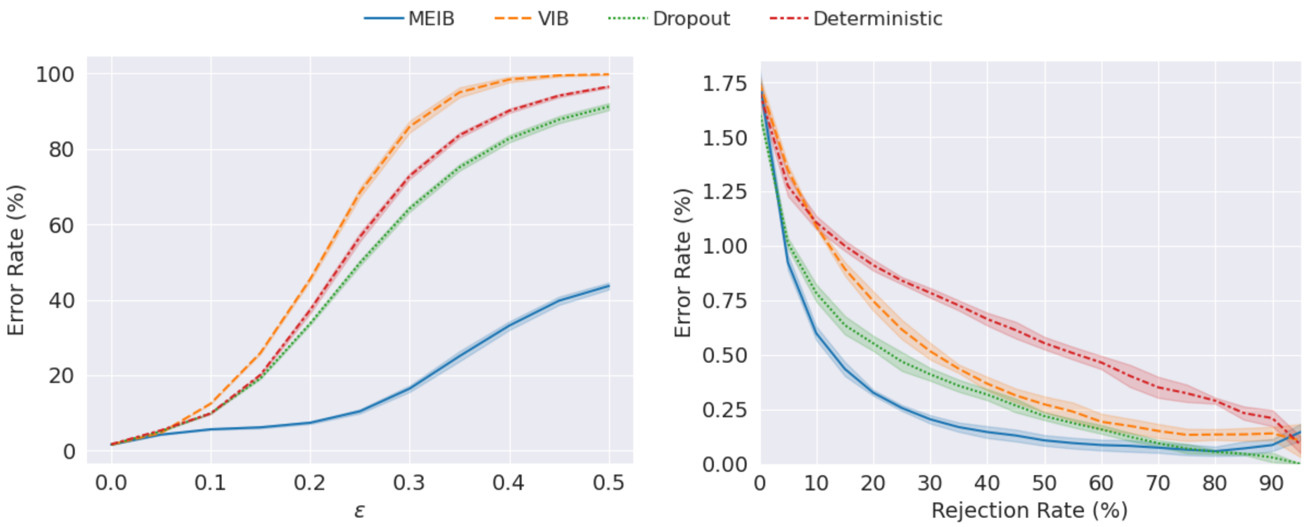
Regularization Effect Table 1 shows the classification results by each method on the QMNIST test set. Specifically, we compared the methods with two different embedding sizes, $D = 2$ and 256. With $D = 2$, the stochastic methods performed better than the deterministic ones regardless of the usage of dropout. Using the larger embedding size of $D = 256$, on the other hand, the deterministic model trained with dropout performed the best while the others have very similar performance considering the standard deviation. However, using 12 Monte Carlo (MC) samples of z, the stochastic methods outperformed the deterministic ones with a larger gap than the single sample case. MEIB provides a reasonable amount of regularization comparable to VIB.

Perturbation Robustness The test time perturbation robustness of a deep neural network model is an important aspect, especially when the model is considered to be deployed for a real-world application. We evaluate the robustness of models toward adversarial examples as an alternative form of perturbation robustness evaluation because models that are weak to adversarial examples might be vulnerable to not only intended attacks but also unexpected noise in test time [18]. Since the primary purpose of MEIB and the other compared methods is not a defense against strong adversarial attacks, we use the Fast Gradient Sign Method (FGSM) [20]:

$$\widetilde{x} = x + \epsilon * \text{sign}(\nabla_x J(\theta, x, y)) \quad (12)$$

where $\widetilde{x}$ is the crafted adversarial example, ϵ is a scale factor of the perturbations, and J is the target loss function, e.g., cross-entropy loss. The FGSM is often regarded as a weak adversary but is still widely used as a first benchmark adversary method due to its simplicity. We report evaluation results with stronger adversaries in Appendix G for interested readers. We crafted the adversarial examples from the QMNIST test set using the FGSM with $\epsilon \in [0.0, 0.5]$ with the step size of 0.5 on all models ($D = 256$) trained with different random seeds for each method. We used 12 MC samples of z for the stochastic methods, MEIB and VIB.

Figure 3a shows the misclassification rate of each method toward the different strengths of the FGSM perturbations. MEIB is more robust than the other methods with significant gaps. Furthermore, the error rate of MEIB increases very slowly with the increasing strength of perturbations until about $\epsilon = 0.3$. On the other hand, VIB is more vulnerable than both deterministic and dropout baselines, typically with more severe perturbations. It might be be-



(a) Perturbation Robustness

(b) Risk-controlled Classification

Figure 3. Performance comparisons on the QMNIST dataset.

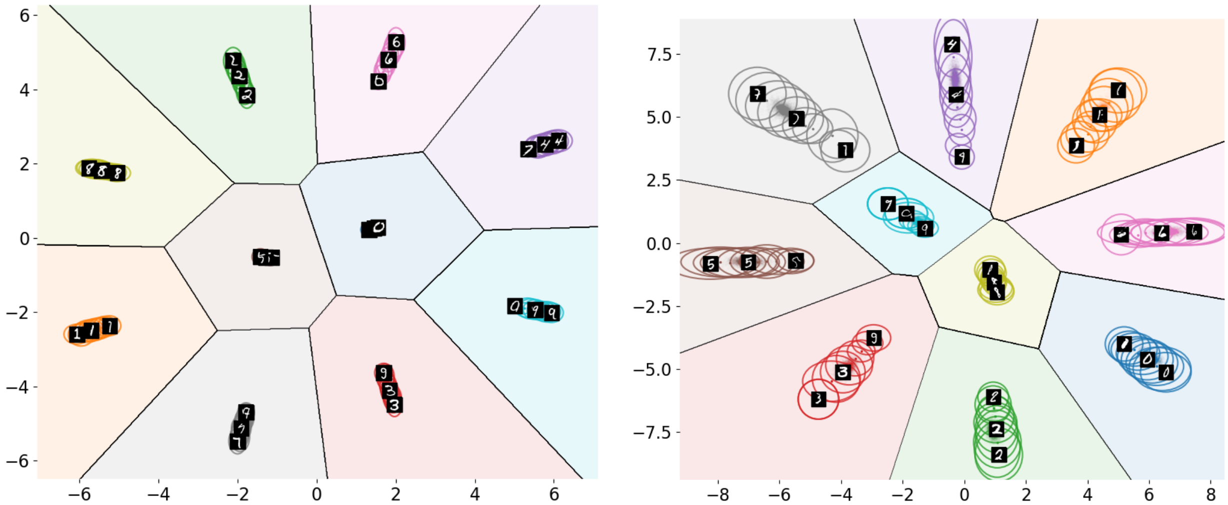
cause we chose the target model with the best performance with the clean dataset, which is a rational choice, for all methods. Thus, a VIB model trained with a different β might yield better robustness, but it still would be difficult to close the gap with MEIB. Please refer to Appendix C for the perturbation robustness with different α values.

Risk-controlled Classification In many real-world application scenarios, it would be favorable to refuse any decision instead of making a false prediction when the model is not confident about the input. For doing this, we need a way to correctly estimate the confidence, or the uncertainty, of inputs to ML models. Using the estimated confidence, we may reject the inputs with insufficient confidence; it is called *risk-controlled recognition* [39]. We evaluated the risk-controlled classification performance on the QMNIST test set by rejecting the inputs by the confidence estimated by each method. Similarly to the confidence measured by MEIB (Section 3), we empirically found that the mean of variance vector from $f_\theta^\Sigma(x)$ of VIB is proportional to the confidence. For the deterministic and dropout baselines, we use the L_2-norm of the embedding vector for each input as a proxy measure of confidence [36]. We set the rejection rate from 0% to 95% with a step size of 5%. A single sample of z was used for MEIB and VIB. Figure 3b shows the risk-controlled classification performance by all methods with $D = 256$. MEIB outperformed all other methods across the most range of the input rejection rate. Specifically, the error rate of MEIB dropped more than half of the initial value after rejecting 10% of uncertain inputs and reached about 0.1% when half of the inputs were rejected.

Embedding Distribution Figure 4 depicts the embedding space learned by VIB and MEIB with $D = 2$. It shows that each embedding distribution by MEIB takes as much area as possible depending on its location from the decision boundaries, which is consistent with our hypothesis. On the other hand, every embedding by VIB has a small standard deviation and thus covers a much smaller area in both dimensions (axes) than those of MEIB. It would be reasonable to consider increasing σ of the prior distribution $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ used in VIB modeling to achieve a similar effect of entropy maximization by MEIB. However, even VIB with very large variance priors still performs much worse than MEIB for both aspects of perturbation robustness and risk-controlled classification (see Appendix D). It implies that MEIB increases σ of inputs in an adaptive way while VIB tries to match the given prior distribution.

4.2. Hedged Instance Embedding

Oh et al. [33] proposed the hedged instance embedding (HIB), a metric learning method that explicitly models the input uncertainty in the stochastic embedding space. HIB utilizes the VIB objective with the probabilistic contrastive learning framework. The authors also proposed a new dataset, called N-digit MNIST, basically images of N adjacent MNIST digits. We examined the HIB framework with the MEIB objective instead of VIB simply by replacing the KL divergence term in the original HIB loss function with the negative conditional entropy of embeddings, keeping all the other aspects of the neural network model same, including the hyperparameters. Please refer to the supplementary material for the architectural details. Table 2 reports the performance of the original HIB with the



(a) VIB

(b) MEIB

Figure 4. 2D embedding space learned for the QMNIST dataset. The ellipses represent the standard deviation of the stochastic embeddings for a subset of training data.

VIB objective and the MEIB variant by the average precision (AP) on the test set. The MEIB variant outperformed the original HIB in every case with the clean test set, except the very high-dimensional embedding of $D = 100$, where both performed well. For the corrupted test data, on the other hand, the original HIB performed slightly better in low-dimensional embedding $D = 2$ scenarios, and both performed comparably with $D = 3$. However, MEIB outperformed with a higher-dimensional embedding $D = 4$ and $D = 100$ in both $N = 2$ and 3 cases. In most applications, it is very uncommon to use such a small size of embeddings with two or three dimensions, except for direct visualization of data relationships [4]. Consequently, it suggests that the MEIB objective would better fit the contrastive learning framework of HIB than the VIB objective upon using a reasonable size of embedding dimensions.

4.3. Person Re-identification

Person re-identification (ReID) is an important computer vision task that is utilized in various applications, including intelligent security and surveillance systems [19]. Unlike typical image classification tasks, the objective of the person ReID task is to find the ranked matches of pedestrian images captured across multiple non-overlapping cameras [8]. Person ReID is challenging due to image-level corruptions and appearance changes, including occlusions [49]. Therefore, it is critical to have an embedding method robust to noises and capable of confidence measuring for potential risk-controllability. For evaluating those aspects, we used three datasets popularly employed in person ReID literature: Market-1501[52], MSMT17[46], and LPW [41]. For LPW, we used a quarter subset of it by selecting every

Table 2. Performance (AP) comparison of HIB with different main framework methods

		$N = 2$		$N = 3$	
D	Test Data	VIB	MEIB	VIB	MEIB
2	Clean	0.955 ± 0.004	$\mathbf{0.959 \pm 0.004}$	0.950 ± 0.003	$\mathbf{0.954 \pm 0.003}$
	Corrupted	$\mathbf{0.840 \pm 0.004}$	0.836 ± 0.010	$\mathbf{0.844 \pm 0.007}$	0.842 ± 0.005
3	Clean	0.980 ± 0.001	$\mathbf{0.982 \pm 0.002}$	0.980 ± 0.003	$\mathbf{0.984 \pm 0.002}$
	Corrupted	0.861 ± 0.003	$\mathbf{0.864 \pm 0.004}$	$\mathbf{0.900 \pm 0.006}$	0.898 ± 0.003
4	Clean	0.991 ± 0.001	$\mathbf{0.993 \pm 0.001}$	0.990 ± 0.002	$\mathbf{0.991 \pm 0.001}$
	Corrupted	0.888 ± 0.003	$\mathbf{0.894 \pm 0.010}$	0.912 ± 0.004	$\mathbf{0.913 \pm 0.003}$
100	Clean	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	Corrupted	0.931 ± 0.005	$\mathbf{0.936 \pm 0.004}$	0.952 ± 0.004	$\mathbf{0.968 \pm 0.002}$

Table 3. Performance on the person ReID datasets

Dataset	Method	mAP (%)	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)
Market-1501	Baseline	75.65 ± 0.21	90.48 ± 0.45	96.47 ± 0.06	97.66 ± 0.15
	DistNet	74.61 ± 0.06	90.10 ± 0.29	96.25 ± 0.23	97.61 ± 0.26
	PFE	76.49 ± 0.23	90.48 ± 0.51	96.53 ± 0.11	97.75 ± 0.20
	DUL	77.12 ± 0.19	90.09 ± 0.17	95.84 ± 0.15	97.29 ± 0.18
	MEIB	$\mathbf{79.67 \pm 0.15}$	$\mathbf{92.14 \pm 0.16}$	$\mathbf{96.79 \pm 0.16}$	$\mathbf{97.83 \pm 0.09}$
MSMT17	Baseline	39.69 ± 0.16	69.31 ± 0.31	82.66 ± 0.17	86.73 ± 0.06
	DistNet	38.16 ± 0.16	68.79 ± 0.21	82.44 ± 0.32	86.67 ± 0.28
	PFE	42.72 ± 0.14	70.62 ± 0.24	83.68 ± 0.17	87.50 ± 0.20
	DUL	38.70 ± 0.18	67.28 ± 0.22	80.98 ± 0.50	85.38 ± 0.25
	MEIB	$\mathbf{44.77 \pm 0.39}$	$\mathbf{73.85 \pm 0.41}$	$\mathbf{85.34 \pm 0.25}$	$\mathbf{88.80 \pm 0.24}$
LPW	Baseline	34.82 ± 0.33	52.76 ± 0.80	67.06 ± 0.94	73.32 ± 0.87
	DistNet	32.37 ± 0.29	50.25 ± 0.20	65.36 ± 0.57	72.02 ± 0.89
	PFE	33.35 ± 0.34	50.36 ± 0.41	64.99 ± 0.67	71.69 ± 0.76
	DUL	33.41 ± 0.65	49.41 ± 0.90	63.90 ± 0.83	70.13 ± 0.83
	MEIB	$\mathbf{37.64 \pm 0.41}$	$\mathbf{54.85 \pm 0.52}$	$\mathbf{69.05 \pm 0.58}$	$\mathbf{75.20 \pm 0.49}$

fourth frame of each identity due to the limited computational resource.

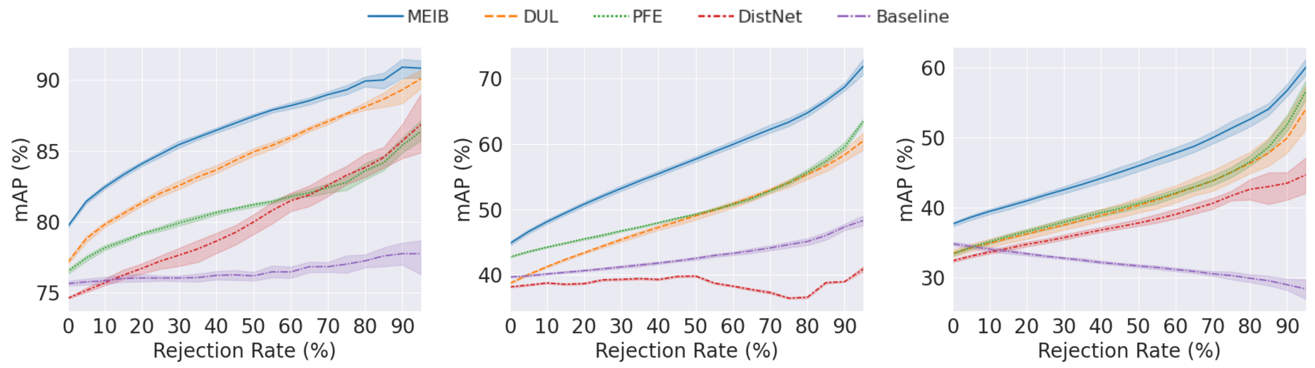
For all methods compared in ReID experiments, we used ResNet50 [24] as the backbone encoder f_θ^B. The mean estimator f_θ^μ of an FC layer with 512 hidden units followed by a BN layer was used for all methods. For the variance estimator f_θ^Σ, we used an MLP of FC512-BN-ReLU-FC512-exp for MEIB and followed the original architecture proposed by the authors for each method, except that we used 512 hidden units for their FC layers. We used $\alpha = 1$ for MEIB and the recommended hyperparameter values by the authors for each method.

We compare the performance of MEIB with the following baseline methods not only from person ReID literature but also from previous work on face recognition, which shares similar task characteristics: DistributionNet (DistNet) [50], Probabilistic Face Embedding (PFE) [39], and Data Uncertainty Learning (DUL, equivalent to VIB) [7]. In addition, we compare the deterministic baseline of the same architecture without the variance estimator f_θ^Σ. We also tried HIB for the ReID task, but we could not achieve meaningful results. All models were trained for the classification task with a softmax classifier of f_θ^C to predict the true identity labels. In testing time, the Euclidean distance between every pair of gallery and query images is calculated using the deterministic embeddings or the mean of stochastic embeddings $f_\theta^\mu(x)$ of them to rank the pairs. For PFE, we use the negative mutual likelihood score (MLS) used in the original work of PFE as the distance metric.

We first evaluated the methods without considering the input confidence and risk control. Table 3 summarizes the results where we report the mean average precision (mAP) [38] and cumulative matching characteristics (CMC) curve [21] at rank-1, rank-5, and rank-10. MEIB outperforms all the others by from 2 up to 6.6 percentage points of mAP throughout the datasets considered. It implies that maximizing the conditional entropy of embeddings in MEIB has better regularization effects than the other methods.

The importance of risk-controlled recognition is more emphasized in the person ReID task due to the cost of misidentifying a person; it can have a serious societal impact, such as falsely tagging someone as a criminal. Considering a real-world application of risk-controlled ReID model deployment, a realistic situation is that we can prepare a well-curated clean set of gallery images in advance and expect that arbitrary query/probe images with potential corruptions will be given after model deployment. Thus, we kept the original test gallery images for each dataset and evaluated the methods with varying amounts of rejected test query images. The query images are sorted using the confidence estimated by each method: MEIB and the deterministic model use the same approaches described in Section 4.1. DistNet and DUL use the mean of variance vector from $f_\theta^\Sigma(x)$, while PFE uses the mean reciprocals of the variance vector elements as their confidence measures. Then, the first $R\%$ of low-confidence images are removed from evaluation. For each remaining query image, the entire gallery images are ranked by the distance, and the evaluation metrics are calculated.

Figure 5 shows the risk-controlled identification performance of the methods for each dataset. While MEIB starts with the best performance among all methods, as shown previously, the amount of performance improvement at lower rejection rates is more significant than the other meth-



(a) Market-1501

(b) MSMT17

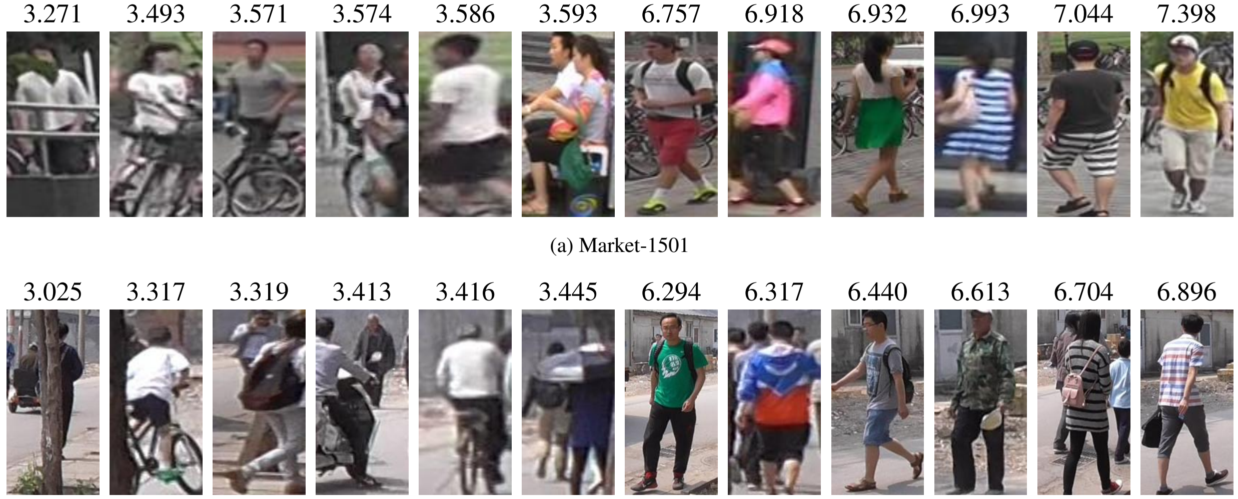
(c) LPW

Figure 5. Risk-controlled person ReID performance.

ods in most cases. Furthermore, the gap between MEIB and the other methods is kept over the most range of the rejection rate. It confirms that MEIB provides an effective confidence measure and risk-control capability. Figure 6 shows the example query images for each dataset (except MSMT17 prohibited by the license) with the associated confidence (conditional entropy) value. There are conspicuous trends: (1) the low-confident images are mostly blurry while the high-confident images are relatively clearer, (2) the persons in the low-confident images are often occluded by another object or person, and (3) the most persons in the low-confident images wear clothes with an achromatic color without patterns while those in the high-confident images wear clothes with vivid colors or patterns such as stripes. This would be another evidence that MEIB provides reasonable confidence measurements.

5. Conclusion

In this work, we presented MEIB, a novel framework that produces stochastic embeddings distributed with the maximum entropy. MEIB provides a regularization effect and perturbation robustness by securing the maximum area in the embedding space, leading to better classification performance. Moreover, the experimental results in digit classification and person ReID tasks showed that MEIB enables an effective risk-controlled recognition by providing reliable uncertainty measurements. While the experiments in this work utilized somewhat elementary neural network backbone encoders, it would be straightforward to combine MEIB with more sophisticated architectures to yield further classification performance improvement.



(a) Market-1501

(b) LPW

Figure 6. Example query images that MEIB is least confident (left six) and most confident (right six) from each ReID dataset. The number above each image is the confidence measured by MEIB as the dimension-wise average entropy.

References

[1] Sravanti Addepalli, Vivek BS, Arya Baburaj, Gaurang Sriramanan, and R Venkatesh Babu. Towards achieving adversarial robustness by enforcing feature consistency across bit planes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1020–1029, 2020. 16

[2] Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, pages 151–160. PMLR, 2019. 2

[3] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. 1, 3, 4

[4] Sungtae An, Shenda Hong, and Jimeng Sun. Viva: semi-supervised visualization via variational autoencoders. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 22–31. IEEE, 2020. 6

[5] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018. 16

[6] Judith Bütepage, Lucas Maystre, and Mounia Lalmas. Gaussian process encoders: Vaes with reliable latent-space uncertainty. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 84–99. Springer, 2021. 1

[7] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5710–5719, 2020. 1, 2, 7

[8] Minghui Chen, Zhiqiang Wang, and Feng Zheng. Benchmarks for corruption invariant person re-identification. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. 6

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 17

[10] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2021. 1, 2

[11] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. 3

[12] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 16

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Ccomputer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 16

[14] 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. 17

[15] Garoe Dorta, Sara Vicente, Lourdes Agapito, Neill DF Campbell, and Ivor Simpson. Structured uncertainty prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5477–5485, 2018. 1

[16] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016. 16

[17] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *Advances in neural information processing systems*, 31, 2018. 4

[18] Justin Gilmer, Nicholas Ford, Nicholas Carlini, and Ekin Cubuk. Adversarial examples are a natural consequence of test error in noise. In *International Conference on Machine Learning*, pages 2280–2289. PMLR, 2019. 4

[19] Shaogang Gong, Chen Change Loy, and Tao Xiang. Security and surveillance. In *Visual analysis of humans*, pages 455–472. Springer, 2011. 6

[20] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 4

[21] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE international workshop on performance evaluation for tracking and surveillance (PETS)*, volume 3, pages 1–7. Citeseer, 2007. 7

[22] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018. 2

[23] Seungyul Han and Youngchul Sung. A max-min entropy framework for reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[24] 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, 2016. 7

[25] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957. 1, 2

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, 2015. 16

[27] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations*, 2014. 1, 3

[28] Shen Li, Jianqing Xu, Xiaqing Xu, Pengcheng Shen, Shaoxin Li, and Bryan Hooi. Spherical confidence learning for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15629–15637, 2021. 2

[29] Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng,

and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13896–13905, 2021. 1, 2

[30] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 16

[31] Jincheng Mei, Chenjun Xiao, Ruitong Huang, Dale Schuurmans, and Martin Müller. On principled entropy exploration in policy optimization. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3130–3136, 2019. 2

[32] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018. 16

[33] Seong Joon Oh, Kevin P. Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew C. Gallagher. Modeling uncertainty with hedged instance embeddings. In *7th International Conference on Learning Representations, (ICLR)*, 2019. 1, 2, 5, 16

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 4

[35] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. Regularizing neural networks by penalizing confident output distributions. In *Workshop Track Proceedings of International Conference on Learning Representations*, 2017. 2

[36] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017. 5, 17

[37] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *6th International Conference on Learning Representations, ICLR, Conference Track Proceedings*, 2018. 16

[38] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008. 7

[39] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6902–6911, 2019. 2, 5, 7

[40] John Shore and Rodney Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on information theory*, 26(1):26–37, 1980. 1, 2

[41] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 6

[42] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 4

[43] Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *European Conference on Computer Vision*, pages 53–70. Springer, 2020. 1, 2

[44] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017. 17

[45] Frederik Warburg, Martin Jørgensen, Javier Civera, and Søren Hauberg. Bayesian triplet loss: Uncertainty quantification in image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12158–12168, 2021. 2

[46] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 6

[47] Chhavi Yadav and Leon Bottou. Cold case: The lost mnist digits. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 4

[48] Gengcong Yang, Jingyi Zhang, Yong Zhang, Baoyuan Wu, and Yujiu Yang. Probabilistic modeling of semantic ambiguity for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12527–12536, 2021. 2

[49] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 6

[50] Tianyuan Yu, Da Li, Yongxin Yang, Timothy M Hospedales, and Tao Xiang. Robust person re-identification by modelling feature uncertainty. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 552–561, 2019. 2, 7

[51] Guanhua Zheng, Jitao Sang, and Changsheng Xu. Understanding deep learning generalization by maximum entropy. *arXiv preprint arXiv:1711.07758*, 2017. 2

[52] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 6

[53] Kaiyang Zhou and Tao Xiang. Torchreid: A library for deep learning person re-identification in pytorch. *arXiv preprint arXiv:1910.10093*, 2019. 16