IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 September 2024, accepted 12 September 2024, date of publication 16 September 2024,
date of current version 30 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3461739

RESEARCH ARTICLE

Key Factors Determining the Required Number of Training Images in Person Re-Identification

TAKU SASAKI[1], ADAM S. WALMSLEY[2], KAZUKI ADACHI[1,3], SHOHEI ENOMOTO[1], AND SHIN'YA YAMAGUCHI[1,4]

[1]NTT Laboratories, Tokyo 108-0075, Japan
[2]Department of Mechanical Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[3]Department of Mathematics, Physics, Electrical Engineering and Computer Science, Graduate School of Engineering Science, Yokohama National University, Yokohama, Kanagawa 240-0067, Japan
[4]Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan

Corresponding author: Taku Sasaki (taku.sasaki@ntt.com)

ABSTRACT Focusing on person re-identification datasets, this paper proposes a new method to estimate the test accuracy curve over the training image number in a precise, interpretable, and efficient manner to receive financial and privacy protection benefits. An existing method, neural scaling law, accurately approximates the curve by fitting a regression function to data points of a training image number and the corresponding accuracy. However, fitting such a function does not explain the reason for the estimated curve. Moreover, obtaining a data point updates model parameters with heavy computation. Therefore, this paper investigates the key factors of a person re-identification dataset that determine the regression parameters. By incorporating the found factors, our method becomes interpretable. Simultaneously, the method significantly reduces computation costs since model updates are no longer needed. We experimentally show that our method is as precise as the uninterpretable neural scaling law incurring nearly millions of model updates.

INDEX TERMS Efficiency, interpretability, neural scaling laws.

I. INTRODUCTION

Person Re-Identification (Person Re-ID) [1], [2], [3] is a computer vision task to retrieve images of a specific person and plays an important role in analysis for surveillance and marketing. To have accurate deep-learning-based Person Re-ID models for the test split of a dataset, practitioners often train (or adapt) off-the-shelf pre-trained models with images from the training split of the same dataset. The more training images are used the more accurate the model becomes [4]. At the same time, the computation and data collection fee and the risk of privacy leakage also increase. The financial budget and the privacy risk tolerance in training should depend on the situation of practitioners (e.g., economic and cultural). Some practitioners probably cannot start their Person Re-ID business until the amount they need to pay and the size of the risk to incur in training are clarified. For example, a retail business operator in a small country potentially cannot afford to pay the fee unless they believe it is necessary to meet the required accuracy. In addition, practitioners in privacy-sensitive countries may be hindered by the lack of transparency in the amount of identifiable information uploaded to the cloud for training. These practitioners need to know the required number of training images. However, our preliminary experiments in Figure 1 reveal that the accuracy curve over the training image number is up to the dataset. In the figure, data points of the training image number and the corresponding accuracy form an upward-convex curve for the MSMT17 [5], Market-1501 [6], and PersonX [7] datasets, but MSMT17 requires 3× as many images to achieve the 30% rank-1 accuracy as the others even for the same architecture (i.e., ResNet50 [8]) from the same initial parameters (i.e., pre-trained with ImageNet [9]) by minimizing the same loss function for the same number of epochs with the same learning rate in the same training algorithm (i.e., the mutual-mean teaching [2] unsupervised domain adaptation technique). Therefore, we solve the problem of giving an interpretable guideline to the required training image number

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

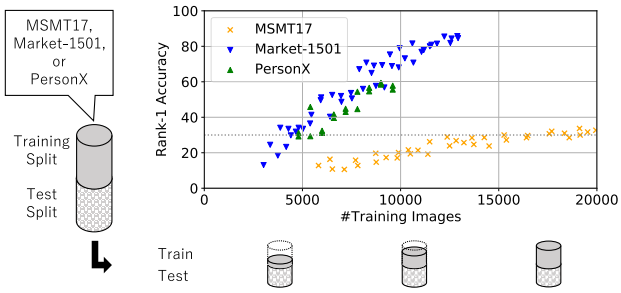

FIGURE 1. Data points of a training image number and the resulting accuracy for MSMT17, Market-1501, and PersonX datasets. The three datasets have respective training and test splits. A data point represents the number of images from the training split used in training and the corresponding test accuracy on all images in the test split. To achieve the 30% rank-1 accuracy (the dotted line), MSMT17 needs its 15K training images while Market-1501 and PersonX only require 5K even for the same Person Re-ID model architecture from the same initial parameters. This paper aims to fit a curve to the data points in an interpretable and efficient manner.

in unseen datasets by estimating the accuracy curve over the training image number within the minimum amount of data collection and computation. This is an entirely new problem even if existing approaches [4], [10] based on Neural Scaling Laws (NSLs) [11], [12], [13], [14], [15], [16] seem to tackle a similar class of problem because such approaches are neither interpretable nor efficient. NSLs start training with a few images and predict the performance gain from additional training images by fitting a regression function to data points of an image number and the corresponding accuracy. NSLs are uninterpretable because they do not explain why the data points form the curve. Moreover, NSLs are inefficient because observing the accuracy of every training image number incurs heavy computation in updating model parameters. In contrast, our method will be interpretable and efficient while keeping the accuracy curve estimation as precise as NSLs.

To introduce interpretability, this paper follows a dataset representation approach [17], [18]. A dataset representation embeds a dataset into a vector to infer a dataset property. Toward the representation for the regression parameter regression, we investigate the underlying relationship between datasets and their accuracy curve over the training image number. As a result, we find that only five key factors of a dataset (i.e., luminance mean, luminance deviation, camera number, person- and camera-wise image volume, and gallery person number) are enough as the dataset representation. The representation explains why datasets require many or few training images and/or how the number of images can be reduced. For example, when the luminance-mean factor has a small value, many training images are required to achieve a certain accuracy; thus, relocating cameras to brighter places can reduce the training image number to achieve a certain accuracy. Furthermore, a small luminance-deviation factor (i.e., most people in the dataset wear clothes of similar colors) requires many training images. Moreover, a large gallery person number is translated into requiring many training images. In other words, dividing the test split into pieces eases training because shorter videos usually have images of fewer gallery persons. The other finding of this paper is that the five key factors, after a simple modification, can be linearly translated into the regression parameters. Thus, we are free from model updates because the linear parameters once *calibrated* (optimized) work for many Person Re-ID datasets.

Relying on the relationship, this paper proposes a new method, namely Interpretable and Efficient accuracy-Curve-over-training-image-number Estimator (IECE), consisting of the dataset representation and representation conversion modules. Given a few labeled images from a dataset, the former module extracts the representation vector of the dataset, which is composed of those five key factors. The latter converts the representation vector into parameters in a new regression function. The parameters reflect significant differences in accuracy curves among datasets.

We assess IECE with Root Mean Squared Error (RMSE) from the estimated curve to ground-truth data points of a training image number and the corresponding accuracy. With the ImageNet-pre-trained ResNet50 Person Re-ID model and the Unreal dataset [19] for linear parameter calibration, IECE marks 3.55, 8.55, and 5.59 RMSE for the MSMT17, Market-1501, and PersonX evaluation, which are less than the double scores achieved by the latest NSL [13] not in an interpretable manner with 470K, 273K, and 67.5K model updates, i.e., 1.80, 4.61, and 4.44 RMSE. The same tendency mostly holds for the TransReID [20] Person Re-ID model and MSMT17 calibration.

Our contributions are listed below.

- We find the five key factors of a Person Re-ID dataset that explain how many training images achieve a certain accuracy.
- We incorporate the five factors for IECE to be not only interpretable but also efficient in estimating the accuracy curve over the training image number.
- Experiments verify that IECE estimates the curve as precisely as the latest NSLs.

II. RELATED WORK

A. PERSON RE-IDENTIFICATION

To find a person of interest (*query* person) in a video stream, Person Re-Identification (Person Re-ID) [2], [3], [21], [22], [23], [24] compares a query person image with every single detected person image from the video stream (*gallery* image) and tells if the two images are of the same person or not. The literature has found difficulties such as occlusion [25], partial bodies [26], viewpoint [7], outfit [27], illumination [28], [29]), and cardinality (e.g., the number of people in the dataset, number of images per person [30], or that of scenes [7], [31]). All these difficulties *could* affect the accuracy curve over the training image number. However, the first four are hard to adopt because they need additional labels. Practitioners are often unable to pay the

labeling fee. Thus, we investigate the relationship between the illumination or cardinality of a dataset and the accuracy curve and find two illumination-related and three cardinality-related key factors. Experiments prove that the five factors are enough for *standard* Person Re-ID datasets to estimate the accuracy curve.

B. NEURAL SCALING LAWS

Even though existing methods [4], [10], [32] using regression functions in Neural Scaling Laws (NSLs) [10], [11], [12], [14], [15], [16] inspired us to propose IECE, the NSL-based methods and IECE have several clear differences. First, existing methods estimate the accuracy of a training image subset obtained by adding a few images to another subset whose accuracy they can know through training and testing. This is an easier problem than ours, with little bias in choosing samples to add, where no bias issues have been discussed. In contrast, we estimate the accuracy of a training image subset without any information about its accuracy. The bias becomes unignorable. Therefore, IECE estimates the expected accuracy curve over all possible training image sampling choices. Second, IECE is interpretable thanks to the key factors translated into the regression parameters, but existing methods are not. Finally, IECE is efficient, but existing methods are not. IECE replaces the heavy computation in model parameter updates with simpler operations in the dataset representation and representation conversion modules.

C. DATASET REPRESENTATION

Previous studies have proposed dataset representation techniques to infer dataset properties. For example, Task2Vec [17] embeds tasks (i.e., the combination of dataset and loss) into the Fischer-information-matrix-based representation vectors and picks a computer vision model pre-trained with the closest task in terms of representation vector as the best pre-trained model for transfer learning. Automatic model Evaluation [18], [33] and its modification [34] are other dataset representation techniques that use the feature mean and covariance or cluster mean to estimate the accuracy degradation due to distribution shifts. In this study, we invent a new dataset representation using the five factors of Person Re-ID datasets that dominate the accuracy curve over the training image number.

III. PROBLEM DEFINITION

This section defines the curve estimation problem on a two-stage basis. First, we define the curve estimation problem in classification datasets, which are easier to understand. Second, we move on to the problem in Person Re-ID datasets.

A. CLASSIFICATION CURVE ESTIMATION

With $\boldsymbol{x}^i \in [0,1]^{c\times h\times w}$ and $y^i \in \{1,\cdots,C^l\}$ representing the i-th image and class label for $(c,h,w,C^l) \in \mathbb{Z}_+^4$ denoting channel number, height, width, and class number, $\mathcal{D}^l = \{(\boldsymbol{x}^i, y^i)\}_{i=1}^{N^l}$ represents the l-th image classification dataset for $l \in \{0,1,\cdots,L\}$. The dataset is randomly divided into training and test splits,[1] i.e., $\mathcal{D}^l_{\rm tr} = \{(\boldsymbol{x}^i_{\rm tr}, y^i_{\rm tr})\}_{i=1}^{N^0_{\rm tr}}$ and $\mathcal{D}^l_{\rm te} = \{(\boldsymbol{x}^i_{\rm te}, y^i_{\rm te})\}_{i=1}^{N^0_{\rm te}}$, where $N^l_{\rm tr} + N^l_{\rm te} = N^l$. We train a classifier f with initial parameters θ_0 using a subset of $\mathcal{D}^l_{\rm tr}$ for $E \in \mathbb{Z}_+$ epochs to minimize the loss function $\mathcal{L}_{\rm cv}$ in the learning rate γ. Note that $(f,\theta_0,E,\mathcal{L}_{\rm cv},\gamma)$ are shared across l. There is an infinite number of ways to sample images to form subsets. By $\mathcal{G}^l_M$, we denote all the ways to generate $M \in \mathbb{Z}_+$ subsets from $\mathcal{D}^l_{\rm tr}$. We let $\mathcal{D}^l_m$ and v^l_m for $m \in \{1,\cdots,M\}$ be the m-th subset of $\mathcal{D}^l_{\rm tr}$ when a certain $g^l \in \mathcal{G}^l_M$ is applied and the corresponding test accuracy on $\mathcal{D}^l_{\rm te}$. Furthermore, we consider a regression function $v(n;\alpha)$ to approximate the accuracy curve over the training image number $n \in \mathbb{Z}_+$, where $\alpha \in \mathbb{R}^A$ for $A \in \mathbb{Z}_+$ denotes the regression parameters. We define the optimal parameters of $\mathcal{D}^l$ by

$$\alpha^l(g^l) = \arg\min_{\alpha} \sum_{m=1}^{M} \left[v^l_m - v(|\mathcal{D}^l_m|;\alpha) \right]^2. \tag{1}$$

The accuracy curve estimation problem estimates

$$\alpha^0 := \mathbb{E}_{g^0\in\mathcal{G}^0}[\alpha^0(g^0)] \tag{2}$$

hinted at by $\left\{(\mathcal{D}^l, \alpha^l(g^l))\right\}_{l=1}^{L}$. We call the $l=0$ and $l \geq 1$ datasets evaluation and calibration datasets, respectively.

B. PERSON RE-ID CURVE ESTIMATION

This paragraph describes the three modifications for Person Re-ID datasets with the superscript l omitted. First, for $i \in \{1,\cdots,N\}$, we assume the camera label of the i-th image $z^i \in \{1,\cdots,Z\}$ can be used, where $Z \in \mathbb{Z}_+$ is the total number of cameras. Second, we use the first sample in the test split as the query sample and the rest as gallery ones. With $i=0$ and $i \geq 1$ corresponding to the labeled query and gallery samples, the test split is represented by $\mathcal{D}_{\rm te} = \{(\boldsymbol{x}^i_{\rm te}, y^i_{\rm te}, z^i_{\rm te})\}_{i=0}^{N_{\rm te}}$, where $y^0_{\rm te} \in \{y^i_{\rm te}\}_{i=1}^{N_{\rm te}}$ and $\{y^i_{\rm tr}\}_{i=1}^{N_{\rm tr}} \cap \{y^i_{\rm te}\}_{i=1}^{N_{\rm te}} = \emptyset$. When $\mathcal{D}_{\rm te}$ has multiple query samples, we average v_m over all samples. Third, we set the constraint that any training subset $\mathcal{D}_m$ has all images such that $y^i_{\rm te} = y$ or no such images for any $y \in \{1,\cdots,C\}$ to match the real-world scenario where persons appear in front of the cameras in turn (i.e., one goes out of the area, then another comes in).

IV. PROPOSED METHOD: IECE

This section proposes Interpretable and Efficient accuracy-Curve-over-training-image-number Estimator (IECE). Section IV-A designs the dataset representation module (i.e., finding the five key factors). Section IV-B implements the representation conversion module. Section IV-C prepares regression functions with parameters that better explain significant differences in accuracy curves among datasets than existing functions.

[1] We assume the training and test samples are independent and identically distributed.

TABLE 1. The rank-10 accuracy of MSMT17-derived datasets. We change the luminance of MSMT17 images. The accuracy is best when the luminance mean and deviation are largest (×1).

	Lum Std ×1	×1/2	×1/4
Lum Mean ×1	62.5%	61.3%	59.1%
×1/2	57.6%	59.0%	59.4%
×1/4	49.9%	53.4%	53.4%

A. DATASET REPRESENTATION MODULE

To find key factors, we analyze Person Re-ID datasets that differ in illumination and cardinality by surveying their test accuracy after the mutual mean teaching [2] domain adaptation over the ImageNet-pre-trained [9] ResNet50 [8] model.

1) ILLUMINATION-RELATED FACTORS

Images from strong lighting are better distinguished than those from weak lighting, where strong or weak lighting is the lighting during day or night [28]. Essentially, we can see the main difference between them in luminance [35]. We argue that high-luminance datasets need fewer training images because the image foreground becomes easy to tell from the background. In addition, we argue that a deviated dataset in terms of luminance needs fewer training images because the image foreground (e.g., clothing color [29]) becomes unique to the person. To support these arguments, we surveyed the MSMT17[2] dataset variants with a modified luminance mean and a deviation by multiplying a coefficient $a^i \in \mathbb{R}$ to the image $\boldsymbol{x}^i$ such that $\mathbb{E}_i[Lum(a^i\boldsymbol{x}^i)]/\mathbb{E}_i[Lum(\boldsymbol{x}^i)]$ and $\mathbb{V}_i[Lum(a^i\boldsymbol{x}^i)]/\mathbb{V}_i[Lum(\boldsymbol{x}^i)]$ fall on $1/2$ and $1/4$, where $Lum : [0,1]^{3\times h\times w} \to [0,1]$ yields the luminance [35] of an image. In Table 1, the dataset with the highest mean and highest deviation achieves the best accuracy. From this insight, we adopt the *Luminance-Mean* (LM) and *Luminance-Deviation* (LD) factors of dataset $\mathcal{D} = \{(\boldsymbol{x}^i, y^i, z^i)\}_{i=1}^N$, i.e.,

$$s_{\mathrm{LM}} = \mathop{\mathbb{E}}_{\boldsymbol{x}\in\mathcal{D}}[Lum(\boldsymbol{x})] \approx \frac{1}{N}\sum_{i=1}^{N} Lum(\boldsymbol{x}^i), \tag{3}$$

$$s_{\mathrm{LD}} = \frac{\sqrt{\mathop{\mathbb{V}}_{\boldsymbol{x}\in\mathcal{D}}[Lum(\boldsymbol{x})]}}{s_{\mathrm{LM}}} \approx \sqrt{\sum_{i=1}^{N}\left(\frac{Lum(\boldsymbol{x}^i)}{s_{\mathrm{LM}}} - 1\right)^2}. \tag{4}$$

Equations (3) and (4) do not need all N samples in the dataset $\mathcal{D}$, as experimentally demonstrated in Section V-C.

2) CARDINALITY-RELATED FACTORS

The number of images or classes dominates model accuracy [30]. We argue that the camera number Z, the shot number (i.e., the number of images per person) $R = N/|\{y^i\}_{i=1}^N|$, and the gallery person number $\{y^i_{\mathrm{te}}\}_{i=1}^{N_{\mathrm{te}}}$ play especially important roles. First, a dataset with many cameras (i.e., larger Z) is likely to have two images of the same person that have similar luminance values but look different due to the difference in image background [36] and thus requires more training images. Second, a dataset with a small shot number (especially $R < Z$) has no two images of the same person from the same camera to guide training and thus requires more training images. Finally, a dataset with many gallery persons (i.e., larger $\{y^i_{\mathrm{te}}\}_{i=1}^{N_{\mathrm{te}}}$) can contain images of persons who are similar to the query person but not are and thus requires more training images.

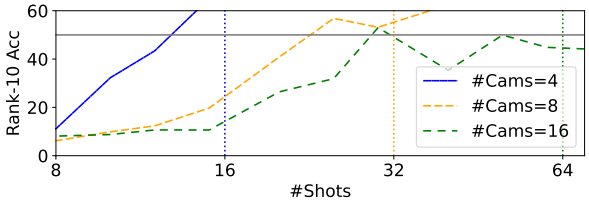

FIGURE 2. The rank-10 accuracy transition over the shot number for different camera numbers in Unreal. The accuracy almost always exceeds 50% (the horizontal gray line) when the shot number reaches four times the camera number (the vertical dashed line).

Fig. 2 illustrates the accuracy over the shot number R and camera number Z with random subsets of the Unreal [19] dataset.[3] First, we can see that a larger Z yields a lower accuracy. Thus, we define the *Camera-Number* (CN) factor as

$$s_{\mathrm{CN}} = \sum_{z=1}^{Z} \min\left(Zp(z), 1\right) \tag{5}$$

$$\approx \sum_{z=1}^{Z} \min\left(Z \cdot \frac{\left|\{i|z^i = z\}_{i=1}^N\right|}{N}, 1\right), \tag{6}$$

with the probability mass function $p(z)$ that $z^i \in \mathcal{D}$ follows. The raw Z is inappropriate for this factor where a camera label z exists such that $p(z) \ll 1/Z$. Second, we can see that accuracy is low especially when $R/Z < 1$, and surpasses 50% when $R/Z \approx 4$. Thus, we design the *Person- and Camera-wise image Volume* (PCV) defined as

$$s_{\mathrm{PCV}} = \frac{R}{s_{\mathrm{CN}}} \approx \frac{N}{\left|\{y^i\}_{i=1}^N\right| \cdot s_{\mathrm{CN}}}. \tag{7}$$

Section V-C again shows the sampling number in (5) and (7) can be reduced.

In Unreal, the rank-10 accuracy of the gallery person number $|\{y^i_{\mathrm{te}}\}_{i=1}^{N_{\mathrm{te}}}|$ was reduced from 86.1% at 375 persons to 81.3% at 750 and 75.9% at 1,500. From this insight, we formulate the *Gallery Person Number* (GPN) factor by

$$s_{\mathrm{GPN}} = \log_{10}\left|\{y^i_{\mathrm{te}}\}_{i=1}^{N_{\mathrm{te}}}\right|. \tag{8}$$

[2]MSMT17 was chosen for having real images of the morning, noon, and night.

[3]Unreal was chosen for having more images of more persons and cameras than MSMT17.

TABLE 2. Existing (left column; $A = 3$) and reduced (right; $A = 2$) regression functions.

$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$		$\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$	
PL3:	$v(n) = \alpha_1 n^{\alpha_2} + \alpha_3$	PL2:	$v(n) = \alpha_1 n^{\alpha_2}$
AC3:	$v(n) = \frac{200}{\pi} \arctan\left(\alpha_1 \frac{\pi}{2} n + \alpha_2\right) + \alpha_3$	AC2:	$v(n) = \frac{200}{\pi} \left(\arctan\left(\alpha_1 \frac{\pi}{2} n + \alpha_2\right) - \arctan \alpha_2\right)$
LN3:	$v(n) = \alpha_1 \ln(n + \alpha_2) + \alpha_3$	LN2:	$v(n) = \alpha_1 \left(\ln(n + \alpha_2) - \ln \alpha_2\right)$
AR3:	$v(n) = 100n \left(1 + \lvert\alpha_1 n\rvert^{\alpha_2}\right)^{-1/\alpha_2} + \alpha_3$	AR2:	$v(n) = 100n \left(1 + \lvert\alpha_1 n\rvert^{\alpha_2}\right)^{-1/\alpha_2}$

B. REPRESENTATION CONVERSION MODULE

The module converts the dataset representation vector $\boldsymbol{s} \in \mathbb{R}^5$ into the regression parameter set $\alpha \in \mathbb{R}^A$, where

$$\boldsymbol{s} = (s_{\mathrm{LM}}, s_{\mathrm{LD}}, s_{\mathrm{CN}}, s_{\mathrm{PCV}}, s_{\mathrm{GPN}})^\top. \tag{9}$$

For simplicity, we assume that every regression parameter is a monotonic and convex function of the factors. Under this assumption, we implement the function as

$$\hat{\alpha}(\boldsymbol{s}; W, b) = W\hat{\boldsymbol{s}} + b, \tag{10}$$

$$\hat{\boldsymbol{s}} = (\boldsymbol{s}/\lambda)^\nu, \tag{11}$$

where division and power operations are performed in every component, $W \in \mathbb{R}^{A\times 5}$, and $b \in \mathbb{R}^A$ are the learnable *slope*, and *intercept* parameters, and $\lambda \in \mathbb{R}^5_+$ and $\nu \in \mathbb{R}^5$ are the hyperparameters called *scaler* and *aligner*. The scaler balances the scale among factors. The aligner brings $\{(\hat{\boldsymbol{s}}, \alpha)\}$ to a flat plane for the slope and intercept parameters well fit. With calibration datasets and their optimal regression parameter values, namely $\{(\mathcal{D}^l, \alpha^l(g^l))\}_{l=1}^L$, where g^l is randomly chosen from $\mathcal{G}^l$, we minimize the following differentiable loss function:

$$\mathcal{L}_{\mathrm{calib}}(W, b) = \sum_{l=1}^{L} \left\| \alpha^l(g^l) - \hat{\alpha}_{\lambda,\nu}(\boldsymbol{s}^l; W, b) \right\|_2^2, \tag{12}$$

where $\boldsymbol{s}^l$ is the representation vector of $\mathcal{D}^l$.

C. SUITABLE REGRESSION FUNCTIONS

The NSL literature has proposed several regression functions. We invent better versions for our problem and even better ones not from NSLs. The left column of Table 2 lists the existing NSL regression functions, namely Power Law (PL) [13], Arctan (AT) [4], Logarithmic (LG) [4], and Algebraic Root (AR) [4], which have $A = 3$ parameters each. The right column lists their better versions by reducing parameters ($A = 2$), which could perform better when $v(0) \approx 0$ as in Fig. 1. Still, the reduced versions do not precisely describe the curve property. Concretely, the PL parameters are $\alpha_1 = v(1)$ and $\alpha_2 = \frac{v'(1)}{v(1)}$, which reflect no information for $n \gg 1$. Even though AC covers $0 < n < \infty$ with $v'(0) = \frac{100}{\pi} \frac{\alpha_1}{\alpha_2^2+1}$ and $v(\infty) = 100 - \frac{200}{\pi} \arctan \alpha_2$, α in AC is nonlinear to $v'(0)$ and $v(\infty)$. LN ignores the $n \gg 1$ information again. In AR, $\alpha_1 = \frac{100}{v(\infty)}$ and $\alpha_2 = \log_{100} 2/(1 - \log_{100} v(1))$ are nonlinear again. Thus, we employ the Terminal Velocity (TV) function,

$$v(n; \alpha) = \alpha_2 \cdot \left(1 - e^{-\frac{\alpha_1}{\alpha_2} n}\right), \tag{13}$$

TABLE 3. Hyperparameter values used throughout experiments.

	LM	LD	CN	PCV	GPN
Scaler ($\boldsymbol{\lambda}$)	+.400	+.500	+4.73	+6.50	+3.50
Aligner ($\boldsymbol{\nu}$)	+1.88	+.500	-3.13	+5.00	-.500

from the physics literature [37]. TV meets $\alpha_1 = v'(0)$ and $\alpha_2 = v(\infty)$. We call α_1 and α_2 the climbing rapidity and terminal accuracy. Since $v(n) < 100\%$ for any $n \in \mathbb{Z}_+$, we also consider another version of TV:

$$v(n; \alpha) = \min(\alpha_2, 100) \cdot \left(1 - e^{-\frac{\alpha_1}{\min(\alpha_2, 100)} n}\right). \tag{14}$$

V. EXPERIMENTS

This section shows our IECE precisely estimates the accuracy curve over the training image number of Person Re-ID datasets. First, we calibrated W and b parameters by minimizing (12) with α^l of $\mathcal{D}^l = \{\mathcal{D}^l_{\mathrm{tr}}, \mathcal{D}^l_{\mathrm{te}}\}$ for $l \in \{1, \cdots, L\}$ and hyperparameters (λ, ν) set to the values[4] listed in Table 3. The optimal regression parameter set α^l of PL, AT, LG, AR, and TV was obtained through (1) using the test accuracy v^l_m on $\mathcal{D}^l_{\mathrm{te}}$ after training using the m-th subset $\mathcal{D}^l_m$ of $\mathcal{D}^l_{\mathrm{tr}}$ for $m \in \{1, \cdots, M\}$ with the Person Re-ID model f_{θ_0}, learning rate γ, epoch size E, and loss function $\mathcal{L}_{\mathrm{cv}}$ defined in Section V-A. Then, we evaluated the Root Mean Squared Error (RMSE),[5] that is,

$$\sqrt{\frac{1}{M} \sum_{m=1}^{M} \left[v^0_m - \hat{v}(|\mathcal{D}^0_m|; \hat{\alpha}_{\lambda,\nu}(\boldsymbol{s}^0; W, b))\right]^2}, \tag{15}$$

of the evaluation dataset $\mathcal{D}^0$, whose value being small means IECE pricisely estimates the curve for $\mathcal{D}^0$. Section V-B introduces Person Re-ID datasets used as $\{\mathcal{D}^l\}_{l=0}^L$ and describes the calculation of the optimal regression parameter α^l. Section V-C reports RMSE results.

A. PERSON RE-ID MODEL AND TRAINING ALGORITHM

ResNet50 [8] and TransReID [20] pre-trained on ImageNet [9] served as the initial Person Re-ID model f_{θ_0}. The cumulative matching characteristic curve at rank-k and mean Average Precision (mAP) [1] were the Person Re-ID accuracy metrics v. The mutual mean teaching [2] was applied in unsupervised and supervised manners over the combination of softmax cross entropy, soft entropy, softmax triplet, and

[4]Hyperparameters are surveyed in the supplementary material.
[5]We chose RMSE because it was used in earlier work [4], [10].

TABLE 4. Dataset summary. The number of cameras (#C), persons, and images are reported along with factor values. For datasets with an explicit split, the person number is presented separately for training and testing while the image number for training (N_{tr}^0), gallery (N_{te}^0), and query. For calibration datasets (Calib), factor values are reported only for the average and standard deviation over the L datasets. 'Eval' represents an evaluation dataset.

Dataset	#C	#Persons	#Images	C/E	s_{LM}	s_{LD}	s_{CN}	s_{PCV}	s_{GPN}
Unreal	34	6,798	1,904,381	Calib	.375	.371	14.8	4.83	2.88
					(.063)	(.068)	(5.5)	(1.66)	(0.25)
MSMT	15	1,040+3,060	32,621+82,161+11,659	Calib	.229	.360	10.1	3.17	2.52
					(.077)	(.006)	(0.9)	(0.27)	(0.32)
				Eval	.297	.365	10.3	2.60	3.49
Market	6	751+750	12,936+13,115+3,368	Eval	.398	.195	5.17	3.38	2.88
PersonX	6	410+750	9,840+27,000+4,500	Eval	.368	.348	6.00	6.00	2.88

TABLE 5. The main results: RMSE achieved by the proposed method, IECE. If RMSE is smaller, the estimated curve is closer to ground-true data points. 'U/S' represents an unsupervised or supervised training. 'Arch' is the Person Re-ID model architecture (i.e., ResNet50 or TransReID). The Unreal, MSMT17, Market-1501, and PersonX datasets are used for calibration (C) and evaluation (E). The median RMSE over evaluation datasets and Person Re-ID metrics (i.e., mAP, rank-1, rank-5, and rank-10) is reported. Numbered regression function names match Table 2. As for TV, the two versions defined in (13) and (14) are tested. The row-wise best RMSE values are bolded. (†: For our environment, TransReID was too large to perform clustering in pseudo label generation [2] for unsupervised learning.)

U/S	Arch	Unreal	MSMT	Market	PersonX	PL3	PL2	AC3	AC2	LN3	LN2	AR3	AR2	TV(13)	(14)
U	Res	C	E	E	E	7.88	12.5	15.1	10.4	12.0	40.2	11.0	13.0	**7.68**	8.29
		-	C	E	E	12.5	21.1	41.8	27.3	18.3	43.6	64.1	21.9	**8.56**	10.9
	Trans	C	E	E	E	(Omitted due to limitation in GPU memory †)									
		-	C	E	E	(Omitted due to limitation in GPU memory †)									
S	Res	C	E	E	E	33.3	25.2	17.4	19.4	15.6	89.0	37.4	44.0	26.5	**5.77**
		-	C	E	E	70.5	55.3	34.9	87.9	42.4	87.6	**10.8**	49.5	641	>1K
	Trans	C	E	E	E	28.8	21.4	21.3	18.7	18.3	37.6	21.2	30.8	17.5	**16.3**
		-	C	E	E	39.9	40.3	82.7	48.7	41.6	32.5	**16.2**	39.9	41.5	57.9

soft softmax triplet as the loss function $\mathcal{L}_{\text{cv}}$ for $E = 50$ epochs in the $\gamma = 3.5\text{e-}4$ learning rate.

B. DATASETS AND OPTIMAL REGRESSION PARAMETERS

We used the Unreal [19], MSMT17 [5], Market-1501 [6], and PersonX [7] datasets. For IECE to perform better, calibration datasets $\{\mathcal{D}^l = (\mathcal{D}^l_{\text{tr}}, \mathcal{D}^l_{\text{te}})\}_{l=1}^L$ are encouraged to support the evaluation dataset $\mathcal{D}^0$ in most factors (i.e., $\{\boldsymbol{s}^l\}_{l=1}^L$ should surround $\boldsymbol{s}^0$ in most components). Yet, assuming that the support stands in all components is not very practical. On such a policy, we fetched $L = 153$ and $L = 162$ calibration datasets from Unreal and MSMT17 that mostly support the others serving as evaluation datasets[6] (for details of fetching and subset generation policy g^l, see Appendices A, B, and C). Table 4 shows $\{\boldsymbol{s}^l\}_{l=1}^L$ is well deviated but not every $\boldsymbol{s}^0$ is in $\mathbb{E}_{l\geq 1}[\boldsymbol{s}^l] \pm \sqrt{\mathbb{V}_{l\geq 1}[\boldsymbol{s}^l]}$ in some components.

C. RESULTS AND DISCUSSION

1) EXISTING REGRESSION FUNCTIONS VS OUR TV FUNCTIONS

Table 5 compares existing regression functions with TV in the median RMSE over all the evaluation datasets and all the Person Re-ID accuracy metrics (for the raw data, see Appendix D). First, we can see that the TV functions (i.e., both (13) and (14)) outperform most existing regression functions that have two or three regression parameters. This superiority owes to the climbing-rapidity and terminal-accuracy parameters directly describing the significant difference in the accuracy curve among datasets. Second, the former and latter TV implementations work better for unsupervised and supervised scenarios. This is because the unsupervised scenario tends to have an accuracy v that reaches the terminal accuracy at n larger than the supervised one. For n such that $v(n)$ has not reached the terminal accuracy, accuracy curves can fit a part of TV functions with $\alpha_2 > 100\%$.

2) STATE-OF-THE-ART NSLS VS IECE

In this discussion and later on, unless otherwise noted, our report focuses on the rank-1 accuracy metric, which is the most commonly used among the four metrics. Table 6 compares the state-of-the-art NSLs [4], [13] with IECE in rank-1 RMSE. Without access to data points, IECE with the best regression function marks RMSE values comparable[7] to that of NSLs in most cases. In Table 6, comparable RMSE was not achieved only in supervised MSMT17

[6]Unreal and MSMT17 are larger than the others, being eligible to generate calibration datasets.

[7]We regard RMSE<17.0 as being "comparable". 17.0 is the largest NSL score in Table 6.

TABLE 6. Rank-1 RMSE for unsupervised ResNet50 Unreal calibration (left) and supervised TransReID Unreal calibration (right) achieved by the State-of-The-Art (SoTA) NSL baselines along with the number of data points accessed (#DP) and the number of model parameter updates (#MU). NSLs' RMSE values are mostly comparable to IECE's ones.

Method			MSMT17 RMSE	MSMT17 #DP	MSMT17 #MU	Market-1501 RMSE	Market-1501 #DP	Market-1501 #MU	PersonX RMSE	PersonX #DP	PersonX #MU
SoTA	NSL	PL3	1.80/1.62	64	470K	4.66/3.92	46	273K	5.97/5.28	18	67.5K
	NSL	AC3	1.84/1.60	64	470K	4.61/3.89	46	273K	4.44/.959	18	67.5K
	NSL	LN3	1.95/1.70	64	470K	17.0/5.36	46	273K	4.55/.950	18	67.5K
	NSL	AR3	1.85/1.65	64	470K	4.69/3.94	46	273K	4.57/.959	18	67.5K
Ours	IECE	Best	3.55/23.6	0	0	8.55/11.8	0	0	5.59/2.37	0	0
Reference	NSL	PL2	4.35/1.63	64	470K	5.89/2.19	46	273K	7.28/3.80	18	67.5K
	NSL	AC2	1.83/1.74	64	470K	4.79/4.17	46	273K	4.39/1.00	18	67.5K
	NSL	LN2	6.88/2.44	64	470K	14.9/4.43	46	273K	9.59/1.13	18	67.5K
	NSL	AR2	5.21/1.76	64	470K	13.2/4.72	46	273K	7.43/4.02	18	67.5K
	NSL	TV(13)	1.83/1.84	64	470K	4.60/1.14	46	273K	4.87/1.52	18	67.5K
	NSL	TV(14)	1.83/1.84	64	470K	7.89/4.34	46	273K	5.13/1.13	18	67.5K

TABLE 7. Ablation study: The smallest RMSE is achieved with all the factors. The column-wise best values are bolded.

U/S	Arch	Calib	Eval	Metric	LM	LD	CN	PCV	GPN	TV(13)	TV(14)
U	Res	Unreal	MSMT	rank-1	✓	✓	✓	✓	✓	**5.90**	11.9
						✓	✓	✓	✓	31.2	21.7
					✓		✓	✓	✓	12.9	**7.10**
					✓	✓		✓	✓	29.3	19.4
					✓	✓	✓		✓	20.0	29.6
					✓	✓	✓	✓		22.7	18.7
				rank-10	✓	✓	✓	✓	✓	**7.62**	**4.12**
						✓	✓	✓	✓	40.9	28.8
					✓		✓	✓	✓	17.4	14.6
					✓	✓		✓	✓	35.3	24.8
					✓	✓	✓		✓	36.6	28.9
					✓	✓	✓	✓		36.9	31.7

TABLE 8. List of the camera-number (CN) and raw-camera-number (RCN) factor values and RMSE results in unsupervised ResNet50 Unreal calibration. Unreal factor values are averaged over all L datasets. The better RMSE value by CN or RCN is bolded.

Dataset	Factor values s_{CN} / s_{RCN}	Rank-1 RMSE TV(13)	Rank-1 RMSE TV(14)	Rank-10 RMSE TV(13)	Rank-10 RMSE TV(14)
Unreal	14.8 / 20.3 = 73%	N/A	N/A	N/A	N/A
MSMT17	10.3 / 15 = 69%	5.90/**5.35**	11.9/**2.37**	7.62/**6.74**	**4.17**/6.63
Market-1501	5.17 / 6 = 86%	**10.5**/14.8	15.4/**7.90**	**5.55**/23.8	8.30/**8.19**
PersonX	6.00 / 6 = 100%	**5.59**/25.2	**7.37**/14.3	**12.3**/40.9	**8.28**/23.0

evaluation, which owed to bad calibration datasets, not to the IECE methodology.

3) ABLATION OF FACTORS

Table 7 lists the RMSE values when one of the key factors is lacking, which suggests that all factors are necessary to keep RMSE small.

4) JUSTIFICATION OF THE CAMERA-NUMBER FACTOR

Table 8 compares the camera-number factor $s_{\mathrm{CN}} = \min(Zp(z), 1)$ to the raw camera number $s_{\mathrm{RCN}} = Z$. As expected, the corrected version had smaller RMSE values than the raw version, especially in PersonX, where the ratio of the two versions (i.e., $s_{\mathrm{CN}}/s_{\mathrm{RCN}} = 100\%$) is significantly different from others (i.e., 73%, 69%, and 86%), suggesting that the camera number correction is required when $p(z)$ is different among datasets.

5) IECE EVALUATION WITH FEWER IMAGES

Fig. 3 illustrates RMSE values when N in (3), (4), (5), (7) is replaced with smaller numbers. We can see that IECE still marks small RMSE values unless we have all N labeled images to feed the dataset representation module. That said, N scales in the data collection or labeling fee.

6) INTERPRETABILITY OF THE ACCURACY CURVE

Table 9 lists the calibrated slope parameter values for the TV function in (14). A positive and negative slope value suggests a positive and negative correlation between scaled and aligned factors $\hat{\boldsymbol{s}}$ and regression parameters α (i.e., the climbing rapidity and terminal accuracy). In reading the correlation coefficient, bear in mind that $\hat{\boldsymbol{s}}$ increases when $\boldsymbol{s}$ decreases if the aligner hyperparameter ν has a negative value in Table 3. As for the luminance-mean and luminance-deviation factors, the aligner and slope are

TABLE 9. Assigned values of the TV(14) slope parameters (W) in the unsupervised ResNet50 Unreal calibration. The left and right values in a cell represent slopes for α_1 and α_2.

Metrics	LM		LD		CN		PCV		GPN	
mAP	+.003	+.003	+.004	+.003	+.004	+.001	-.001	-.004	+.008	+.019
rank-1	+.008	-.000	+.013	-.000	+.006	-.002	-.001	+.000	+.025	-.005
rank-5	+.012	+1.14	+.019	+1.63	+.009	+.694	-.002	+.038	+.048	-2.15
rank-10	+.013	+6.94	+.018	+19.8	+.009	+5.46	-.002	+.442	+.069	-18.6

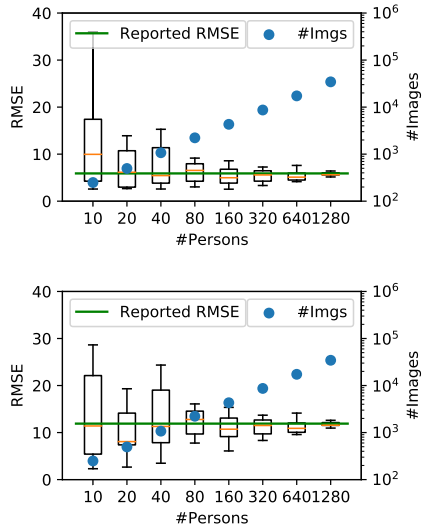

FIGURE 3. RMSE values achieved with fewer labeled images fed to IECE in unsupervised ResNet50 Unreal calibration for rank-1 accuracy on MSMT17 evaluation (left: TV(13), right: TV(14)). Boxplots present median and quartile RMSE values along with the ×1.5 inter-quartile range. The "Reported RMSE" is the one reported in Table 5 with 82K images of 3,060 persons.

almost consistently positive; therefore, a brighter and more deviated dataset can be said to have an accuracy curve with a more rapid climb and higher terminal accuracy. The camera-number factor has negative aligner and slope values, suggesting that datasets with fewer cameras are prone to rapidly climbing and terminating at high accuracy. The person- and camera-wise image volume factor often has positive slope values for α_2 and the aligner is positive, implying datasets with a larger shot number tend to achieve high accuracy with enough training images. The gallery-person-number slope for α_1 is consistently assigned with positive values, but the aligner is negative, so datasets with fewer gallery persons are likely to have rapidly climbing curves. These findings from TV encourage practitioners to relocate cameras and adjust the person number to save the training image number to achieve a certain accuracy.

VI. LIMITATIONS

A. DESIGNED FOR STANDARD PERSON RE-ID DATASETS

We only solved illumination and cardinality as difficulties dominating Person Re-ID accuracy curves. Corner-case datasets have different difficulties. For example, the angle of depression is key in the Bird-View Person Re-ID dataset [38]. Image sharpness dominates Vehicle Re-ID [36], [39], [40], [41], [42], [43], [44] accuracy curves because subtle differences (e.g., tire wheels and emblems) tell two car identities of the same car model in the same color. Yet, new factors for new difficulties would extend the interpretable and computationally efficient accuracy curve estimation toward such corner cases and general computer vision or vision-language datasets beyond Object Re-ID.

TABLE 10. Number of images, unique persons, and cameras in four Unreal scenarios.

Scenario	A	B	C	D
#Images	196,013	369,447	470,004	868,917
#Persons	11,788	6,489	6,599	6,798
#Cameras	6	16	6	6
Cameras located at	Outdoor	Outdoor	Outdoor	Indoor

B. MODEL-WISE CALIBRATION IS INEVITABLE

In IECE, calibration and evaluation datasets must share the model architecture f and initial parameters θ_0 as described in Section III. This can limit the usability of IECE when practitioners estimate the accuracy curve over the training image number of multiple pre-trained models (e.g., practitioners may use Task2Vec [17] to pick the best one from multiple candidates) because they need to prepare calibrated IECEs individually for all candidates. Nevertheless, the Task2Vec dataset representation added to our five factors could lift the constraint on f and θ_0.

C. NO GUARANTEED TOLERANCE TO SAMPLING BIAS

In (2), we formulated the accuracy curve estimation problem to estimate the expected regression parameter averaged over all $g^0 \in \mathcal{G}^0$ as the guideline. However, no theory supports the real regression parameter set of g^0 falling near the guideline (i.e., the variance is small enough). Even though we did not observe a large variance in our experiments, a dataset with diverse images could lead to a large variance. A potential future direction would be estimating $\mathbb{V}_{g^0 \in \mathcal{G}^0}[\alpha^0(g^0)]$ along with $\mathbb{E}_{g^0 \in \mathcal{G}^0}[\alpha^0(g^0)]$.

VII. RESPONSIBILITY TO HUMAN SUBJECTS

In contrast to the Unreal and PersonX datasets consisting of synthetic images, MSMT17 and Market-1501 include identifiable person images. We checked that no ethical issues were reported for the datasets. We used the dataset in a way that met the release agreements, but that unfortunately did not allow us to present actual images as samples to

TABLE 11. Assigned factor values for Unreal scenario combinations. For example, ABC uses images from scenarios A, B, and C, not D. A alone and C alone are excluded for too small LD and CN component values.

Combination	LM	LD	CN	PCV	GPN	# Patterns
A	.530	.175	4.92	N/A	N/A	0
B	.362	.331	12.7	2.86/4.12/5.39/6.72	2.57/2.88/3.18	12
C	.457	.271	5.42	N/A	N/A	0
D	.202	.293	4.19	1.91/2.89/3.41	2.57/2.88/3.18	9
AB	.413	.323	18.6	3.15/4.41/5.80/7.24	2.57/2.88/3.18	12
AC	.468	.262	8.69	2.45/3.48/4.69/5.75	2.57/2.88/3.18	12
AD	.355	.505	9.59	2.17/2.98/4.20/5.02	2.57/2.88/3.18	12
BC	.418	.309	13.3	3.26/4.53/5.82/6.95	2.57/2.88/3.18	12
BD	.287	.424	16.5	3.18/4.83/6.15/7.55	2.57/2.88/3.18	12
CD	.389	.399	8.86	2.46/3.61/4.60/5.88	2.57/2.88/3.18	12
ABC	.435	.307	19.5	3.26/4.76/6.17/7.59	2.57/2.88/3.18	12
ABD	.343	.445	22.2	3.28/4.86/6.15/7.42	2.57/2.88/3.18	12
BCD	.359	.413	18.5	3.64/5.13/6.48/8.11	2.57/2.88/3.18	12
ABCD	.381	.404	23.6	3.52/4.99/6.37/7.90	2.57/2.88/3.18	12
Avg±StdDv	3.75±.063	.371±.068	14.8±5.5	4.83±1.66	2.88±0.25	$L = 153$

TABLE 12. List of MSMT17 training and test person labels in categories A and B.

Category		Person labels	#Persons
Train	A	0-107, 146-219, 381-483, 567-624 676-760, 798-866, 878-920, 944-1021	618
	B	108-145, 220-380, 484-566, 625-675 761-797, 867-877, 921-943, 1022-1040	423
Test	A	0-322, 436-659, 902-1149 1219-1450, 1697-1853, 2005-2241 2353-2556, 2586-2709, 2777-3002	1975
	B	323-435, 660-901, 1150-1218 1451-1696, 1854-2004, 2242-2352 2557-2585, 2710-2776, 3003-3059	1085

TABLE 13. Counting images of persons in categories A and B from every camera. In the training and test splits, a camera often shoots persons in either A or B and seldom ones in the other.

	Train		Test	
z	A	B	A	B
1	4,408	502	13,672	577
2	20	183	23	116
3	360	94	1,062	443
4	236	1,378	585	3,335
5	3,225	1,071	11,231	1,477
6	73	1,605	181	4,502
7	1,821	1,632	6,335	3,065
8	19	776	41	2,496
9	181	1,215	559	3,263
10	0	655	37	2,215
11	145	3,009	544	8,034
12	295	1,069	406	3,111
13	101	3,534	381	9,934
14	3,116	760	10,782	1,039
15	846	292	3,390	432
Total	*14,846*	*17,775*	*49,229*	*44,591*

qualitatively determine whether our key factors reflected image appearance in illumination.

VIII. CONCLUSION

We found the five factors of a Person Re-ID dataset determining the accuracy curve over the training image number. By incorporating the key factors, IECE became interpretable and efficient. Extensive experiments showed

TABLE 14. The number of cameras, persons, and images in training and test splits for every MSMT17 group. Image numbers in test splits are reported separately for gallery and query.

Group	#Cameras	#Persons	#Images
1	15	125+125	5,984+(5,933+792)
2	15	125+125	3,609+(3,609+453)
3	12	125+125	2,922+(2,922+433)
4	14	125+125	2,388+(2,388+426)
5	12	125+125	1,835+(1,835+308)
6	15	125+125	9,460+(9,068+1,108)
7	15	125+125	4,095+(4,095+552)
8	15	125+125	1,780+(1,780+288)

that IECE marked as small RMSE values as uninterpretable NSL baselines that incurred computational costs in nearly millions of Person Re-ID model updates, as long as all the five factors were used even with a reduced number of samples fed to IECE. Also, we found that the TV function from physics can be used as a better regression function with climbing rapidity and terminal accuracy directly explaining the significant differences in accuracy curves among datasets. For the future, we hope subsequent research builds new factors after ours to extend the accuracy-curve-over-training image estimation toward general computer vision and vision-language datasets.

APPENDIX A UNREAL CALIBRATION CONFIGURATION

As shown in Table 10, Unreal consists of four scenarios, which we tag A, B, C, and D. Every scenario has a different image background and a different number of cameras Z; thus the s_{LM}, s_{LD}, and s_{CN} factor values differ in Table 11. We used this difference to maintain the diversity among Unreal-derived calibration datasets. Scenarios B and D could already serve as calibration datasets (A and C did not for the outlier factor values compared to those in our evaluation datasets). To increase the number of calibration datasets, we added the combinations of A, B, C, and D, which introduced more diversity in $(s_{\mathrm{LM}}, s_{\mathrm{LD}}, s_{\mathrm{CN}})$. Still, the combination variation in the other two factors was not large enough. Thus,

TABLE 15. Assigned factor values for MSMT-17 group combinations (e.g., '00111111' represents groups 3-8). '11111001' was removed for CN that was too small and PCV that was too large.

Combination	LM	LD	CN	PCV	GPN	# Patterns
00111111	.301/.154	.363	10.8	2.69	2.01/2.57/2.88	6
01011111	.306/.152	.358	10.7	2.80	2.01/2.57/2.88	6
01101111	.304/.151	.363	10.7	2.87	2.01/2.57/2.88	6
01110111	.307/.153	.363	10.6	2.98	2.01/2.57/2.88	6
01111011	.305/.152	.350	8.49	2.61	2.01/2.57/2.88	6
01111101	.305/.152	.357	9.38	2.93	2.01/2.57/2.88	6
01111110	.306/.152	.357	10.8	3.07	2.01/2.57/2.88	6
10111111	.307/.153	.355	10.7	3.04	2.01/2.57/2.88	6
10101111	.307/.152	.363	10.8	3.32	2.01/2.57/2.88	6
10110111	.308/.153	.360	10.6	3.26	2.01/2.57/2.88	6
10111011	.305/.152	.364	8.88	2.83	2.01/2.57/2.88	6
10111101	.306/.152	.364	10.1	3.20	2.01/2.57/2.88	6
10111110	.305/.152	.363	10.6	3.41	2.01/2.57/2.88	6
11001111	.304/.151	.366	10.8	3.24	2.01/2.57/2.88	6
11010111	.305/.152	.357	10.5	3.37	2.01/2.57/2.88	6
11011011	.300/.149	.352	8.63	3.04	2.01/2.57/2.88	6
11011101	.306/.152	.354	10.0	3.30	2.01/2.57/2.88	6
11011110	.306/.152	.357	10.8	3.37	2.01/2.57/2.88	6
11100111	.307/.153	.356	10.6	3.43	2.01/2.57/2.88	6
11101011	.302/.150	.357	9.05	2.97	2.01/2.57/2.88	6
11101101	.302/.150	.355	10.1	3.38	2.01/2.57/2.88	6
11101110	.304/.151	.370	10.6	3.55	2.01/2.57/2.88	6
11110011	.304/.151	.353	8.22	3.32	2.01/2.57/2.88	6
11110101	.307/.153	.360	9.56	3.39	2.01/2.57/2.88	6
11110110	.306/.152	.360	10.5	3.55	2.01/2.57/2.88	6
11111001	.305	.348	6.29	3.85	N/A	0
11111010	.301/.149	.358	8.91	3.13	2.01/2.57/2.88	6
11111100	.306/.152	.355	9.96	3.44	2.01/2.57/2.88	6
Avg±StdDv	.229±.077	.360±.006	10.1±0.85	3.17±0.27	2.52±0.32	$L = 162$

we sampled $N_{\text{tr}} = 12,000$ training images of $N_{\text{tr}}/(s_{\text{PCV}} \cdot Z)$ persons, $N_{\text{te}} = Z \cdot \exp(s_{\text{GPN}}) \cdot s_{\text{PCV}}$ gallery and $Z \cdot \exp(s_{\text{GPN}})$ query images for $s_{\text{PCV}} \in \{3, 4, 5, 6\}$ and $\exp(s_{\text{GPN}}) \in \{375, 750, 1500\}$. This sampling yielded the final Unreal-derived calibration datasets $\{\mathcal{D}^l = (\mathcal{D}^l_{\text{tr}}, \mathcal{D}^l_{\text{te}})\}_{l=1}^{L}$, counting $L = 153$. The l-th training split $\mathcal{D}^l_{\text{tr}}$ for $l \in \{1, \cdots, L\}$ was further divided into subsets $\{\mathcal{D}^l_m\}_{m=1}^{M}$ and the accuracy v^l_m was tested on $\mathcal{D}^l_{\text{te}}$. In the subset division (i.e., choice of g^l), we paid attention so that the sampling bias would be reduced. To this end, we ensured every image in $\mathcal{D}^l_{\text{tr}}$ appeared in a certain number of subsets using Algorithm 1. The algorithm takes the l-th Unreal dataset $\mathcal{D}^l = (\mathcal{D}^l_{\text{tr}}, \mathcal{D}^l_{\text{te}})$, Person Re-ID model f_{θ_0}, and a parameter $\mathcal{J} \subset \mathbb{Z}_+$ to yield the subsets $\{\mathcal{D}^l_{jr}\}_{jr}$ and the corresponding test accuracy v^l_{jr} on $\mathcal{D}_{\text{te}}$ for $j \in \mathcal{J}$ and $r \in \{0, \cdots, j-1\}$, where (j, r) substitutes for m. Every image in $\mathcal{D}^l_{\text{tr}}$ appeared in $|\mathcal{J}|$ out of $M = \sum_{j \in \mathcal{J}} j$ subsets. Throughout the paper, we used $\mathcal{J} = \{1, 2, 3, 5, 10, 15\}$ and had $M = 36$ subsets from $\mathcal{D}^l_{\text{tr}}$. Using $\{(\mathcal{D}^l_{jr}, v^l_{jr})\}_{jr}$, we calculated the optimal parameter

$$\alpha^l = \arg\min_{\alpha} \sum_{j \in \mathcal{J}} \left(\frac{1}{j} \sum_{r=0}^{j-1} \left[v^l_{jr} - \hat{v}(|\mathcal{D}^l_{jr}|; \alpha) \right]^2 \right). \quad (16)$$

Algorithm 1 Observing Data Points of a Training Image Number and the Corresponding Accuracy. Used for *calibration* Datasets

Require: $\mathcal{D}_{\text{tr}} = \{(\boldsymbol{x}^i_{\text{tr}}, y^i_{\text{tr}}, z^i_{\text{tr}})\}_{i=1}^{N_{\text{tr}}}$
Require: $\mathcal{D}_{\text{te}} = \{(\boldsymbol{x}^i_{\text{te}}, y^i_{\text{te}}, z^i_{\text{te}})\}_{i=0}^{N_{\text{te}}}$
Require: f_{θ_0}
Require: $\mathcal{J} \subset \mathbb{Z}_+$
Ensure: $\{(|\mathcal{D}_{jr}|, v_{jr})\}_{jr}$
for $j \in \mathcal{J}$ **do**
  for $r \in \{0, 1, \cdots, j-1\}$ **do**
    $\mathcal{D}_{jr} = \left\{ (\boldsymbol{x}^i_{\text{tr}}, y^i_{\text{tr}}) \middle| y^i_{\text{tr}} - \frac{Y_r}{j} \in \left\{ 1 \cdots, \frac{Y}{j} \right\} \right\}$
    Train f_{θ_0} with $\mathcal{D}_{jr}$ and observe accuracy v_{jr} on $\mathcal{D}_{\text{te}}$
  end for
end for

APPENDIX B MSMT17 CALIBRATION CONFIGURATION

Even though MSMT17 does not have more than one explicit scenario whereas Unreal had four, we found that MSMT17 persons could be categorized into two categories as described in Table 12, namely A and B. As shown in Table 13, a person from category A and one from B often appear in front of different cameras, which helps us make calibration datasets with variation in CN. To introduce variation in PCV, we sorted persons in categories A and B in the ascending order of the image number and fetched 125 training persons each to make eight groups listed in Table 14, where we chose 125 test persons each that were the most similar to the training persons in the image number.

TABLE 16. The detailed main results.

U/S	Arch	Calib	Eval	Metrics	PL3	PL2	AC3	AC2	LN3	LN2	AR3	AR2	TV(13)	(14)
U	Res	Unreal	MSMT	mAP	**6.97**	11.9	17.4	15.3	8.50	12.8	7.55	8.94	8.77	9.96
				rank-1	**3.55**	17.0	33.5	34.7	18.8	22.6	8.83	8.93	5.90	11.9
				rank-5	4.27	11.5	18.2	17.3	12.0	43.6	**2.48**	4.87	6.61	7.82
				rank-10	10.7	6.05	17.8	17.6	7.80	49.3	6.74	5.45	7.62	**4.17**
			Market	mAP	15.8	22.8	17.5	**11.6**	24.4	30.5	28.8	37.6	14.0	16.1
				rank-1	16.8	25.4	8.87	**8.55**	22.8	36.8	51.0	26.5	10.5	15.4
				rank-5	13.0	23.3	**5.09**	5.18	19.7	75.3	15.3	22.2	5.14	10.8
				rank-10	10.7	20.8	**4.51**	6.69	17.5	80.4	13.1	18.7	5.55	8.30
			PersonX	mAP	7.33	11.4	20.0	**5.76**	12.1	15.8	21.3	20.3	7.74	8.12
				rank-1	8.44	13.1	12.7	6.54	10.7	22.1	36.7	15.5	**5.59**	7.37
				rank-5	**5.05**	9.38	8.50	9.87	7.70	56.8	5.68	10.4	8.38	5.35
				rank-10	6.73	7.36	10.8	10.8	6.62	61.3	**5.23**	7.49	12.3	8.28
		MSMT	Market	mAP	15.1	21.0	46.9	50.1	18.1	26.1	51.6	31.9	**10.5**	10.5
				rank-1	18.9	28.1	41.5	26.6	16.1	37.2	68.5	15.6	**11.0**	11.2
				rank-5	21.5	30.1	43.1	25.2	21.6	43.1	67.0	22.4	**9.39**	14.2
				rank-10	20.8	30.3	42.2	23.2	22.1	45.3	58.8	21.3	**7.72**	10.6
			PersonX	mAP	**4.53**	9.73	276	27.9	16.0	13.9	45.5	17.6	6.62	6.68
				rank-1	9.91	21.3	25.6	25.1	18.5	44.2	90.2	27.6	**6.59**	6.61
				rank-5	**5.01**	17.9	26.6	69.4	22.6	56.3	79.9	6.54	6.77	21.9
				rank-10	**5.44**	7.50	24.9	80.7	18.1	56.2	61.3	65.7	22.9	22.9
S	Res	Unreal	MSMT	mAP	68.5	38.2	65.9	69.2	56.9	**37.3**	42.8	47.7	92.3	50.5
				rank-1	79.9	42.5	43.7	56.7	61.5	65.5	69.7	50.5	58.1	**32.5**
				rank-5	76.4	46.4	31.7	46.8	52.1	78.0	32.4	26.9	44.2	**21.5**
				rank-10	75.1	50.8	27.6	42.2	46.7	82.4	**8.75**	12.5	36.8	17.4
			Market	mAP	27.4	14.1	21.9	19.9	18.2	79.3	252	**10.2**	29.5	15.2
				rank-1	30.2	21.0	18.3	19.0	13.0	91.7	10.0	40.2	12.3	**7.16**
				rank-5	36.3	28.6	16.2	14.9	12.3	97.0	23.6	59.6	9.40	**2.92**
				rank-10	38.7	31.8	16.5	13.5	12.2	98.2	22.3	65.2	8.43	**1.76**
			PersonX	mAP	14.6	21.8	16.6	27.8	18.5	86.2	49.3	34.2	41.7	**4.37**
				rank-1	12.1	10.5	11.6	18.4	6.94	94.6	25.1	37.7	23.5	**3.59**
				rank-5	15.7	8.47	12.2	14.9	5.07	98.5	42.4	52.9	13.8	**1.16**
				rank-10	18.1	10.3	11.8	12.6	4.95	99.3	42.6	57.7	10.0	**.504**
		MSMT	Market	mAP	39.9	9.55	9.29	10.1	22.7	**8.08**	80.3	47.0	13.1	35.7
				rank-1	31.1	24.9	21.1	**2.67**	21.0	8.79	6.23	24.8	4.24	4.24
				rank-5	37.9	27.3	5.47	**3.33**	18.6	97.0	10.1	21.0	3.79	3.81
				rank-10	39.1	33.5	5.34	4.76	19.0	98.2	11.5	52.1	**3.20**	3.24
			PersonX	mAP	105	238	103	167	109	78.2	101	**63.4**	1269	2141
				rank-1	120	77.2	133	165	87.5	58.7	46.9	**37.6**	>10K	>10K
				rank-5	104	91.1	61.0	179	72.0	98.5	**7.85**	71.9	>10K	>10K
				rank-10	101	100	48.7	185	62.1	99.3	**7.52**	75.9	>10K	>10K
	Trans	Unreal	MSMT	mAP	43.7	23.6	47.7	39.8	25.8	9.23	**7.96**	23.7	53.4	47.3
				rank-1	86.3	57.7	73.0	80.5	75.0	**23.6**	93.1	65.1	78.2	68.7
				rank-5	95.5	58.7	72.3	79.6	81.3	**30.8**	94.6	77.0	77.8	62.3
				rank-10	93.4	59.4	67.7	75.3	79.7	**38.2**	94.1	79.8	74.0	57.0
			Market	mAP	17.6	9.55	15.1	19.1	15.6	10.2	**4.54**	92.1	27.4	22.9
				rank-1	28.6	16.7	23.5	24.4	21.0	**11.8**	29.4	21.4	12.4	20.7
				rank-5	29.0	20.0	22.9	18.3	14.3	84.9	**8.70**	12.7	9.19	12.0
				rank-10	30.8	22.7	19.6	15.2	11.8	89.8	**5.46**	28.7	8.37	8.61
			PersonX	mAP	22.0	28.9	**2.92**	9.78	24.9	39.3	50.1	40.0	6.39	9.89
				rank-1	11.5	18.3	7.89	10.8	12.7	37.0	24.3	21.5	18.3	**2.37**
				rank-5	8.01	12.7	8.63	10.6	10.5	92.7	16.3	26.3	16.7	**2.27**
				rank-10	9.06	9.90	7.94	9.15	9.15	95.5	18.2	32.9	12.0	**1.74**
		MSMT	Market	mAP	17.2	21.6	39.1	50.4	21.3	28.0	**10.7**	21.0	13.9	13.9
				rank-1	20.5	15.1	67.9	8.11	13.1	24.2	10.5	**6.48**	18.0	33.2
				rank-5	11.8	14.1	78.5	9.04	16.5	29.4	30.8	11.3	10.5	**4.52**
				rank-10	15.7	16.4	4.88	12.6	17.9	31.9	22.1	10.0	**3.85**	5.72
			PersonX	mAP	80.7	74.4	125	64.0	81.1	**58.7**	63.6	58.8	114	114
				rank-1	59.3	59.1	289	47.0	61.9	33.0	**20.7**	64.0	65.0	82.5
				rank-5	91.5	85.0	122	51.4	72.8	37.1	**11.7**	65.2	170	237
				rank-10	105	87.9	86.9	50.6	75.2	40.4	**6.09**	65.0	78.5	97.7

Since the groups were too small in the image number for a calibration dataset, we considered ${}_8C_6$ combinations of them as well as ones with half luminance mean for the LM variation. After removing a combination with outliers, the resulting $L = 162$ calibration datasets had the factor values listed in Table 15. Algorithm 1 yielded $\{(\mathcal{D}^l_{jr}, v^l_{jr})\}_{jr}$ used in calculating $\{\alpha^l\}_{l=1}^L$ again.

APPENDIX C EVALUATION CONFIGURATION

We used MSMT17, Market-1501, and PersonX datasets for evaluation as they are. Since the sampling bias was not as severe as in calibration, we used a simpler and computational-friendly g^0 algorithm to generate the subsets and observe the corresponding test accuracy $\{(\mathcal{D}^0_m, v^0_m)\}_m$ for the optimal regression parameter calculation, which we refer

Algorithm 2 Making Subsets of a Dataset and Observing the Corresponding Accuracy. Used for *evaluation* Datasets

Require: $\mathcal{D}_{\text{tr}} = \{(\boldsymbol{x}^i_{\text{tr}}, y^i_{\text{tr}}, z^i_{\text{tr}})\}_{i=1}^{N_{\text{tr}}}$
Require: $\mathcal{D}_{\text{te}} = \{(\boldsymbol{x}^i_{\text{te}}, y^i_{\text{te}}, z^i_{\text{te}})\}_{i=0}^{N_{\text{te}}}$
Require: f_{θ_0}
Require: $\mathcal{Y} = \{\eta_m \in \mathbb{Z}\}_{m\in\{1,\cdots,|\mathcal{Y}|\}}$
Ensure: $\{(|\mathcal{D}_m|, v_m)\}_{m\in\{1,\cdots,2|\mathcal{Y}|\}}$
1: **for** $m \in |\{1, \cdots, |\mathcal{Y}|\}|$ **do**
2: $\mathcal{X}_m = \left\{(\boldsymbol{x}^i_{\text{tr}}, y^i_{\text{tr}}) \,\middle|\, y^i_{\text{tr}} \in \{1, \cdots, \eta_m\}\right\}$
3: Train f_{θ_0} with $\mathcal{D}_m$ and observe v_m on $\mathcal{D}_{\text{te}}$
4: $\mathcal{X}_{m+1} = \left\{(\boldsymbol{x}^i_{\text{tr}}, y^i_{\text{tr}}) \,\middle|\, y^i_{\text{tr}} \in \{Y, \cdots, Y - \eta_m + 1\}\right\}$
5: Train f_{θ_0} with $\mathcal{D}_{m+1}$ and observe v_{m+1} on $\mathcal{D}_{\text{te}}$
6: **end for**

TABLE 17. RandPerson (RP) key factor values and the rank-1 RMSE result on the TV(13) function. (†: Since the .885 PCV factor value was too small compared to others in Table 4, we had to use different Unreal-derived calibration datasets.)

U/S	Arch	Calib	Eval	s_{LM}	s_{LD}	s_{CN}	s_{PCV}	s_{GPN}	RMSE
U	Res	Unreal†	RP	.351	.474	17.5	.885	3.85	6.18

to as Algorithm 2. The algorithm takes $\mathcal{D}^0_m = (\mathcal{D}^0_{\text{tr}}, \mathcal{D}^0_{\text{te}})$, the Person Re-ID model f_{θ_0}, and a parameter $\mathcal{Y} \subset \mathbb{Z}_+$ and fetches the images of every $\eta \in \mathcal{Y}$ persons from the head and tail as the training subset $\mathcal{D}^0_m$ and observed the corresponding accuracy v^0_m, for $m \in \{1, \cdots, M = 2|\mathcal{Y}|\}$. We adopted $\mathcal{Y} = \{200, 225, \cdots, 975\}$ for MSMT17, $\mathcal{Y} = \{200, 225, \cdots, 750\}$ for Market-1501, and $\mathcal{Y} = \{200, 225, \cdots, 400\}$ for PersonX.

APPENDIX D DETAILED RESULTS

Table 16 lists the raw data for Table 5.

APPENDIX E EVALUATION IN REAL-WORLD DATASET

We assess the preciseness of IECE with the RandPerson [31] dataset, which is proposed as a more real-world dataset than MSMT17, Market-1501, and PersonX. Table 17 summarizes the key factor values of RandPerson and reports RMSE. We can see that the RandPerson RMSE is small enough to compare with the ones reported earlier in Table 5.

REFERENCES

[1] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3908–3916.
[2] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–15.
[3] Y. Ge, F. Zhu, D. Chen, R. Zhao, and H. Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-ID," in *Proc. Neural Inf. Process. Syst.*, 2020, pp. 11309–11321.
[4] R. Mahmood, J. Lucas, D. Acuna, D. Li, J. Philion, J. M. Alvarez, Z. Yu, S. Fidler, and M. T. Law, "How much more data do I need? Estimating requirements for downstream tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 275–284.
[5] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.
[6] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[7] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 608–617.
[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[10] R. Mahmood, J. Lucas, J. M. Alvarez, S. Fidler, and M. Law, "Optimizing data collection for machine learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 29915–29928.
[11] Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma, "Explaining neural scaling laws," 2021, *arXiv:2102.06701*.
[12] M. Hutter, "Learning curve theory," 2021, *arXiv:2102.04074*.
[13] J. S. Rosenfeld, A. Rosenfeld, Y. Belinkov, and N. Shavit, "A constructive prediction of the generalization error across scales," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–30.
[14] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou, "Deep learning scaling is predictable, empirically," 2017, *arXiv:1712.00409*.
[15] D. Hoiem, T. Gupta, Z. Li, and M. Shlapentokh-Rothman, "Learning curves for analysis of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4287–4296.
[16] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," 2020, *arXiv:2001.08361*.
[17] A. Achille, M. Lam, R. Tewari, A. Ravichandran, S. Maji, C. Fowlkes, S. Soatto, and P. Perona, "Task2Vec: Task embedding for meta-learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6429–6438.
[18] W. Deng and L. Zheng, "Are labels always necessary for classifier accuracy evaluation?" in *Proc. IEEE/CVF Comput. Vis. Pattern Recognit. Conf.*, 2021, pp. 15069–15078.
[19] T. Zhang, L. Xie, L. Wei, Z. Zhuang, Y. Zhang, B. Li, and Q. Tian, "UnrealPerson: An adaptive pipeline towards costless person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11501–11510.
[20] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14993–15002.
[21] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1528–1535.
[22] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *Proc. Assoc. Advancement Artif. Intell.*, vol. 33, Jul. 2019, pp. 8738–8745.
[23] Y. Lin, L. Xie, Y. Wu, C. Yan, and Q. Tian, "Unsupervised person re-identification via softened similarity learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3387–3396.
[24] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical clustering with hard-batch triplet loss for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13654–13662.
[25] P. Chen, W. Liu, P. Dai, J. Liu, Q. Ye, M. Xu, Q. Chen, and R. Ji, "Occlude them all: Occlusion-aware attention network for occluded person re-ID," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11813–11822.
[26] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7073–7082.
[27] Y. Wang, X. Liang, and S. Liao, "Cloning outfits from real-world images to 3D characters for generalizable person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4890–4899.
[28] S. Xiang, G. You, L. Li, M. Guan, T. Liu, D. Qian, and Y. Fu, "Rethinking illumination for person re-identification: A unified view," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 4730–4738.

[29] B. Xu, L. He, X. Liao, W. Liu, Z. Sun, and T. Mei, "Black Re-ID: A head-shoulder descriptor for the challenging problem of person re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 673–681.
[30] J. Vanschoren, "Meta-learning: A survey," 2018, *arXiv:1810.03548*.
[31] Y. Wang, S. Liao, and L. Shao, "Surpassing real-world source training data: Random 3D characters for generalizable person re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3422–3430.
[32] A. Jain, G. Swaminathan, P. Favaro, H. Yang, A. Ravichandran, H. Harutyunyan, A. Achille, O. Dabeer, B. Schiele, A. Swaminathan, and S. Soatto, "A meta-learning approach to predicting performance and data requirements," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3623–3632.
[33] Y. Huang, Z. Zhang, Y. Huang, Q. Wu, H. Huang, Y. Zhong, and L. Wang, "Customized meta-dataset for automatic classifier accuracy evaluation," *Pattern Recognit.*, vol. 146, Feb. 2024, Art. no. 110026.
[34] W. Tu, W. Deng, T. Gedeon, and L. Zheng, "A bag-of-prototypes representation for dataset-level applications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2881–2892.
[35] C. Ridpath and W. Chisholm, "Techniques for accessibility evaluation and repair tools," in *Proc. W3C Work. Draft*, 2000, pp. 1–58.
[36] Y. Yao, L. Zheng, X. Yang, M. Napthade, and T. Gedeon, "Attribute descent: Simulating object-centric datasets on the content level and beyond," 2022, *arXiv:2202.14034*.
[37] A. Garbino, R. S. Blue, J. M. Pattarini, J. Law, and J. B. Clark, "Physiological monitoring and analysis of a manned stratospheric balloon test program," *Aviation, Space, Environ. Med.*, vol. 85, no. 2, pp. 177–182, Feb. 2014.
[38] C. Yan, G. Pang, L. Wang, J. Jiao, X. Feng, C. Shen, and J. Li, "BV-Person: A large-scale dataset for bird-view person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10923–10932.
[39] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2016, pp. 1–6.
[40] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 869–884.
[41] X. Liu, W. Liu, T. Mei, and H. Ma, "PROVID: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 645–658, Mar. 2018.
[42] Y. Zhouy and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6489–6498.
[43] Y. Yao, L. Zheng, X. Yang, M. Naphade, and T. Gedeon, "Simulating content consistent vehicle datasets with attribute descent," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 775–791.
[44] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang, "PAMTRI: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 211–220.

TAKU SASAKI received the B.S. and M.S. degrees in engineering from Tokyo Institute of Technology, Tokyo, Japan, in 2014 and 2016, respectively.

Since 2016, he has been a Research Engineer with NTT Software Innovation Center, Tokyo. His research interests include evolutionary computation, computer vision, and deep learning.

Mr. Sasaki won the Student Best Paper Award at the IEEE Congress on Evolutionary Computation 2015.

ADAM S. WALMSLEY is currently pursuing the bachelor's degree in engineering with The University of British Columbia.

In 2023, he was an Intern with NTT Software Innovation Centre working as a Research Engineer. His research interests include computer vision and deep learning.

KAZUKI ADACHI received the M.E. degree in computer engineering from Yokohama National University, Kanagawa, Japan, in 2019, where he is currently pursuing the Ph.D. degree with the Graduate School of Engineering Science. Since 2019, he has been with NTT Corporation, Tokyo, Japan. He is a Researcher with the Computer and Data Science Laboratories, NTT. His research interests include image recognition and representation learning.

SHOHEI ENOMOTO received the B.S. degree from Tohoku University, in 2014, and the M.S. degree from Tokyo Institute of Technology, in 2016. Since 2016, he has been engaged in researching deep learning and computer vision with NTT Laboratories.

SHIN'YA YAMAGUCHI received the B.E. degree in computer engineering from Yokohama National University, in 2015, and the M.E. degree in computer engineering from the Graduate School of Engineering, Yokohama National University, in 2017. He is currently pursuing the Ph.D. degree with the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, and an Associate Distinguished Researcher with NTT. His research interests include machine learning with synthetic data, generative models, distribution shifts, self-supervised learning, and semi-supervised learning.