Uncertainty-Aware Multi-Shot Knowledge Distillation for Image-Based Object Re-Identification

Xin Jin[1*] **Cuiling Lan**[2†] **Wenjun Zeng**[2] **Zhibo Chen**[1†]
University of Science and Technology of China[1] Microsoft Research Asia[2]
jinxustc@mail.ustc.edu.cn {culan, wezeng}@microsoft.com chenzhibo@ustc.edu.cn

Abstract

Object re-identification (re-id) aims to identify a specific object across times or camera views, with the person re-id and vehicle re-id as the most widely studied applications. Re-id is challenging because of the variations in viewpoints, (human) poses, and occlusions. Multi-shots of the same object can cover diverse viewpoints/poses and thus provide more comprehensive information. In this paper, we propose exploiting the multi-shots of the same identity to guide the feature learning of each individual image. Specifically, we design an Uncertainty-aware Multi-shot Teacher-Student (UMTS) Network. It consists of a teacher network (T-net) that learns the comprehensive features from multiple images of the same object, and a student network (S-net) that takes a single image as input. In particular, we take into account the data dependent *heteroscedastic uncertainty* for effectively transferring the knowledge from the T-net to S-net. To the best of our knowledge, we are the first to make use of multi-shots of an object in a teacher-student learning manner for effectively boosting the single image based re-id. We validate the effectiveness of our approach on the popular vehicle re-id and person re-id datasets. In inference, the S-net alone significantly outperforms the baselines and achieves the state-of-the-art performance.

1 Introduction

Object re-identification (re-id) aims to identify/match a specific object in different places, times, or camera views, from either images or video clips, for the purpose of tracking or retrieval. Because of the high demand in practice, person re-id and vehicle re-id are two dominant research areas for object re-id. In this work, we focus on the popular image-based person and vehicle re-id tasks.

Images to be matched typically have large variations in terms of capturing viewpoints, (human) poses, lighting, and occlusions, making re-id a challenging task (Subramaniam, Chatterjee, and Mittal 2016; Su et al. 2017; Li et al. 2017; Zhao et al. 2017; Ge et al. 2018; Qian et al. 2018; Zhang et al. 2019; Wang et al. 2017; Liu et al. 2018b). These result in inconsistency of visible object regions across images (see Fig. 1(a1)) and lack of comprehensive information from a single image (see Fig. 1(b1)) for matching, and spatial semantics misalignment.

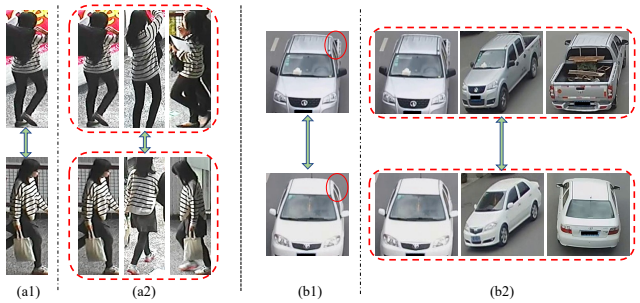

Figure 1: Challenges in image-based re-id: (a1) Inconsistency of visible body regions; and (b1) Lack of comprehensive information from a single image. Observation and Motivation: multi-shot images make it easier to identify whether they are the same person/vehicle as shown in (a2) and (b2).

Generally, for a specific object (*e.g.*, person, vehicle), multiple images captured from different viewpoints or times can provide more comprehensive information, making the identification much easier (see Fig. 1(a2) and (b2)). For example, the difference between the rear of the vehicle is difficult to identify from the two single images in Fig. 1(b1) but the difference becomes very obvious when comparing the two sets of multi-shot images shown in Fig. 1(b2). It is worth noting that for image-based re-id, only a single image is available as a query during inference/testing. The exploration of comprehensive information of multi-shot images is underexplored and remains an open problem.

In this paper, we propose an Uncertainty-aware Multi-shot Teacher-Student (UMTS) Network for exploiting the multi-shot images to enhance the image-based object re-id performance in a teacher-student manner, without increasing the inference complexity or changing the inference setting. We achieve this by distilling knowledge from the multi-shots of the same object and applying it to guide single shot network learning. Fig. 2 shows the flowchart of

[*]This work was done when Xin Jin was an intern at MSRA.
[†]Corresponding Author.

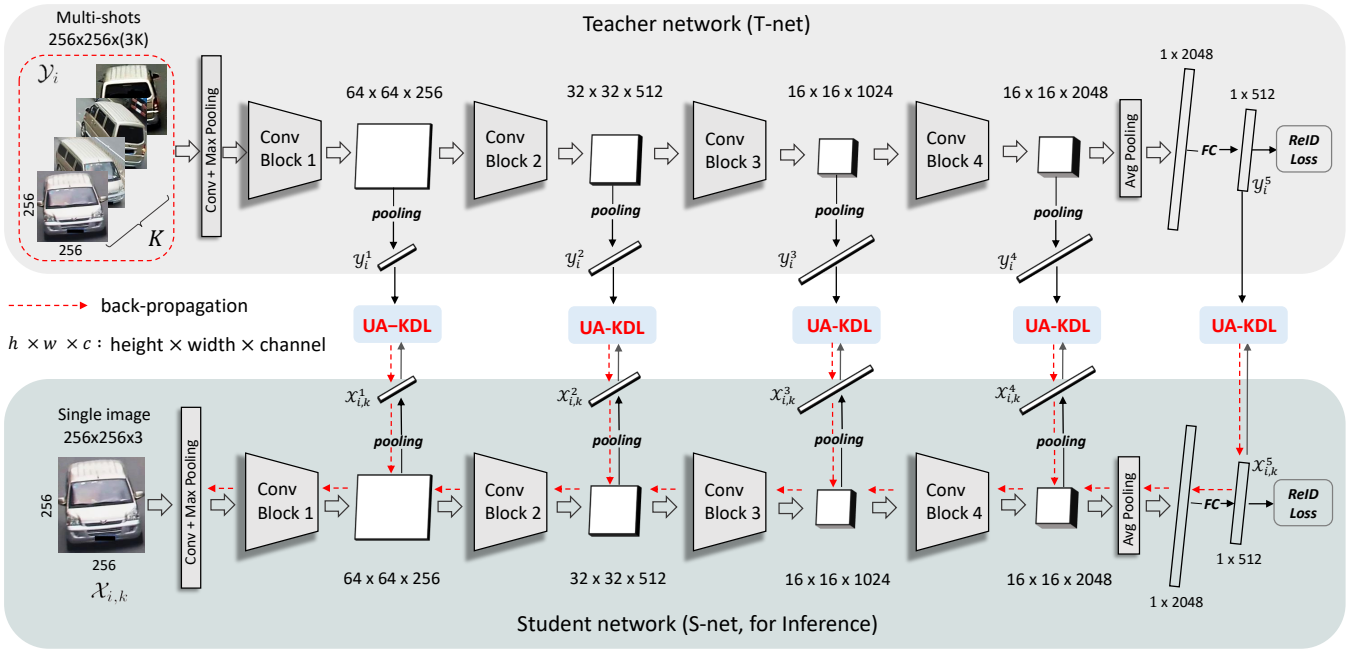

Figure 2: Proposed Uncertainty-aware Multi-shot Teacher-Student (UMTS) Network. It consists of a Teacher-network (T-net) that learns features from the concatenation of multi-shots (*i.e.*, K images, $\mathcal{Y}_i$) of the same identity i, and a Student-network (S-net) that takes a single image $\mathcal{X}_{i,k}$ of the K images as input. To enable efficient feature learning from the T-net, we take into account the data dependent *heteroscedastic uncertainty* and design an Uncertainty-aware Knowledge Distillation Loss (UA-KDL), which we apply at different layers/stages of the teacher-student network. During inference, we use only the S-net.

the proposed Uncertainty-aware Multi-shot Teacher-Student (UMTS) Network. It consists of a Teacher network (T-net) that learns features from the multi-shots (*i.e.*, K images) of the same object, and a Student network (S-net) that takes a single image of the K images as input. In particular, different individual images from the multi-shots have different visible object regions, occlusions, and image quality, and thus different capabilities in approaching the knowledge of the multi-shot images. We take into account the data dependent *heteroscedastic uncertainty* (Kendall and Gal 2017) and design an Uncertainty-aware Knowledge Distillation Loss (UA-KDL) to enable efficient learning of the S-net from the T-net. We conduct extensive ablation studies and demonstrate the effectiveness of the framework and components on both person re-id and vehicle re-id datasets. Our main contributions are summarized as follows:

- We propose a powerful Uncertainty-aware Multi-shot Teacher-Student (UMTS) Network to exploit the comprehensive information of multi-shots of the same object for effective single image re-id, without increasing computational cost in inference.
- We take into account the data dependent *heteroscedastic uncertainty* and design an Uncertainty-aware Knowledge Distillation Loss (UA-KDL), which can efficiently regularize the feature learning at different semantics levels (*i.e.*, layers/stages).
- To the best of our knowledge, we are the first to make use of multi-shots of an object in a teacher-student learning manner for efficient image-based re-id.

2 Related Work

2.1 Image-based Person/Vehicle Re-ID

For image-based person re-id, a lot of efforts are made to address spatial semantics misalignment problem, *i.e.*, across images the same spatial positions usually do not correspond to the same body parts. Many approaches tend to make explicit use of semantic cues such as pose (skeleton), to align the body parts (Kalayeh et al. 2018; Suh et al. 2018; Liu et al. 2018a; Qian et al. 2018; Ge et al. 2018; Zhang et al. 2019). Some approaches leverage attention designs to selectively focus on different body parts of the person (Liu et al. 2017; Zhao, Li, and others 2017). Some other approaches split the feature map to rigid grids for the coarse alignment and jointly consider the global feature and local details (Sun et al. 2018; Wang et al. 2018b). Moreover, several works use GAN to augment the training data with pseudo labels assigned to remedy the insufficiency of training samples (Zheng, Zheng, and Yang 2017; Huang et al. 2018; Zheng et al. 2019b). To address the viewpoint variation problem for vehicle re-id, Zhou *et al.* design a conditional generative network to infer cross-view images and then combine the features of the input and generated views to improve the re-id (Zhou and Shao 2017). In (Zhou and Shao 2018), a complex multi-view feature inference scheme is proposed based on an attention and GAN based model.

Different from the above works, we aim to explore the

comprehensive information of multi-shot images of an object in a teacher-student manner to improve single image based re-id. It is a general re-id framework and we validate its effectiveness for both person re-id and vehicle re-id.

2.2 Teacher-Student Learning

Recent studies show that the knowledge learned by a strong teacher network can improve the performance of a student network (Chen et al. 2017; Zhou et al. 2018; Wang et al. 2019). Hinton *et al*. propose distilling the knowledge in an ensemble of models into a single model (Hinton, Vinyals, and Dean 2015). Romero *et al*. extend this idea to enable the training of a student that is deeper and thinner than the teacher using both the outputs and the intermediate representations learned by the teacher (Romero et al. 2015). Most existing methods focus on learning a light-weight student model from a teacher with the same input data. In contrast, our work aims to distill knowledge from multi-shot images to teach a single shot image feature learning for robust re-id.

2.3 Uncertainty and Heteroscedastic Uncertainty

In Bayesian viewpoint, there are two main types of uncertainty: *epistemic* uncertainty and *aleatoric* uncertainty (Kendall and Gal 2017; Gal 2016). *Epistemic* uncertainty accounts for uncertainty in the model parameters, which is often referred to as *model uncertainty*. *Aleatoric* uncertainty can further be categorized into *homoscedastic uncertainty*, which stays constant for different input data and varies between different tasks, and ***heteroscedastic uncertainty***, which depends on the inputs to the model, with some noisy inputs potentially having poorer predictions than others (*e.g*., due to occlusion or low quality). Under a framework with per-pixel semantic segmentation and depth regression tasks, input-dependent *heteroscedastic uncertainty* together with *epistemic* uncertainty are considered in new loss functions (Kendall and Gal 2017), making the loss more robust to noisy data. In a multi-task setting, Kendall *et al*. show that the task uncertainty captures the relative confidence between tasks, reflecting the uncertainty inherent to the regression/classification task (Kendall, Gal, and Cipolla 2018). They propose using *homoscedastic* uncertainty as a basis for weighting losses in a multi-task learning problem.

In our work, we exploit the *heteroscedastic* uncertainty of the input data (multi-shot images and a corresponding single shot image) to better transfer the knowledge distilled from multi-shot images of an object to each single shot image.

3 Uncertainty-aware Multi-shot Teacher-Student (UMTS) Network

We show the proposed Uncertainty-aware Multi-shot Teacher-Student (UMTS) network in Fig. 2. It consists of a Teacher network (T-net) that learns comprehensive features from the multi-shot images of the same object, and a Student network (S-net) that takes a single image from this multi-shots as input. We aim to exploit the more comprehensive knowledge from the multi-shots of the same identity to regularize/teach the feature learning of a single image for robust *image-based object re-id*. To effectively transfer knowledge from T-net to S-net, we propose an Uncertainty-aware Knowledge Distillation Loss (UA-KDL) and apply them over intermediate layer features and the final matching features, respectively. The entire network can be trained in an end-to-end manner and only the S-net is needed in inference. We discuss the details in the following subsections.

3.1 Motivation: Multi-Shots versus Single-Shot

For an object of the same identity, multiple images captured from different viewpoints/times/places are often available. There are large variations in terms of the visible regions/occlusions, lighting, deformations (*e.g*., poses of person), and the backgrounds. Multiple images can provide more comprehensive information than a single image. For image-based re-id, each identity usually has multiple images in a dataset even though such grouping information cannot be used in inference, where only a single image is used as the query. There are very few works that explicitly explore the multi-shot information to enhance image-based re-id.

We look into whether multi-shot images can lead to better re-id performance and investigate how much benefit it can potentially bring experimentally. As illustrated in Fig. 3, we build three schemes (see (a)(b)(c)) based on the ResNet-50 which is widely used in re-id (He et al. 2016; Wang et al. 2018b; 2018a; Zhang et al. 2019; He et al. 2019). *Scheme A* is a baseline scheme that uses single image for re-id. *Scheme B* and *Scheme C* both assume four-shots of the same identity [1] are used together to obtain the re-id feature vector. *Scheme B* (see (b)) obtains the re-id feature by averaging the feature vectors of the four images while *Scheme C* (see (b)) jointly extracts the re-id feature from the input of four-shot images (input channel number: 3×4).

We show the performance comparisons on the person re-id dataset CUHK03 (labeled setting) (Li et al. 2014) and vehicle re-id dataset VeRi-776 (Liu et al. 2016) in Fig. 3(d). Interestingly, *Scheme B* that simply aggregates the features of four-shots outperforms (*Scheme A*) that uses single image as input by **4.3%** and **7.2%** in mAP accuracy on CUHK03 and VeRi-776 respectively. *Scheme B* ignores the joint feature extraction and interaction among images of the same id and *Scheme C* remedies these by simply concatenating four images together in channel as the input. *Scheme C* outperforms *Scheme A* significantly by **9.0%** and **12.7%** in mAP accuracy on CUHK03 and VeRi-776 respectively. We conclude that there is a huge space for improvement when multi-shot images are available. However, during the inference, in practice, only a single query image is accessible and there is no identity information either for each image in the gallery dataset for image-based re-id. *Scheme B* and *Scheme C* need to take multi-shot images as input and are thus not practical, but somewhat provide performance upper bounds.

To remedy the practical gap, we propose an Uncertainty-aware Multi-shot Teacher-Student (UMTS) Network to transfer the knowledge of multi-shot images to an individual image (see Fig. 2). In inference, as shown in Fig. 3(d), our final scheme UMTS with the S-net alone (Ours), which takes

[1] For each image, based on the groudtruth ids, we randomly select another three images of the same id to have four-shot images.

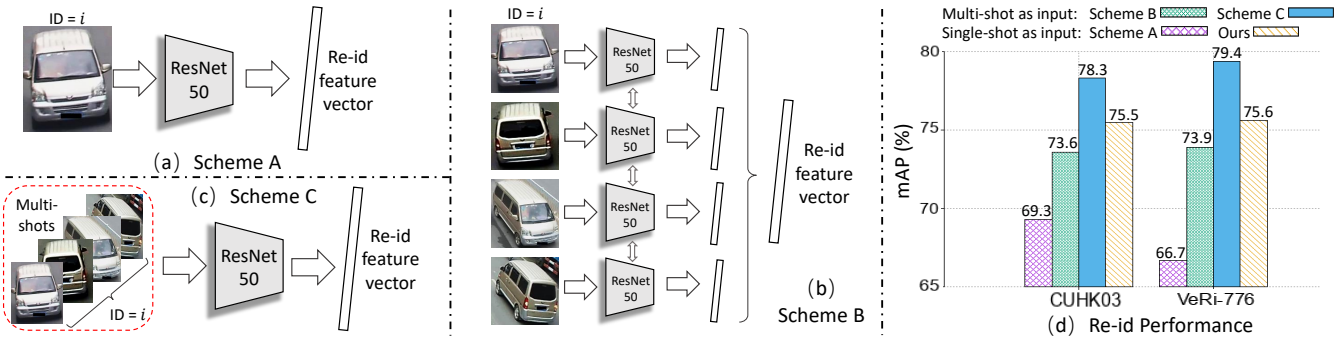

Figure 3: Investigation on whether using multi-shot images can result in better re-id and how much benefit it potentially brings. (a) ***Scheme A*** uses single shot image for re-id. (b) ***Scheme B*** assumes four-shot images are always used and the features of the four images are averaged as the re-id feature. (c) ***Scheme C*** assumes four-shot images are always used and the re-id feature is jointly extracted from the input of four-shot images (we also use this as the Teacher network in our final scheme). (d) Performance comparisons on person re-id dataset CUHK03 (labeled setting) and vehicle re-id dataset VeRi-776.

a single image as input, significantly outperforms *Scheme A* by **6.2%** and **8.9%** in mAP on CUHK03 and VeRi-776 respectively. Note that the model sizes of the three schemes and our final model S-net are almost the same, with *Scheme C* slightly larger (0.11%), which is fair for comparisons.

3.2 Teacher network and Student network

Based on the analysis in subsec. 3.1, we take a simple but effective network as in *Scheme C* (see Fig. 3(c)) as the Teacher network (T-net) (see Fig. 2). More generally, we define the number of shots as K and the input to the T-net is a tensor of size $H \times W \times 3K$. T-net and S-net have the same network structure beside the difference in the number of the input channels. Each network has four stages with each containing multiple convolutional layers, and one fully connected layer, *i.e.*, the fifth stage to obtain the re-id feature vector for matching. For each network, we add the widely-used re-identification loss (ReID Loss) (classification loss (Sun et al. 2018; Fu et al. 2019), and triplet loss with batch hard mining (Hermans, Beyer, and Leibe 2017)) on the re-id feature vectors. Note that our approach is general and any other networks, *e.g.*, PCB (Sun et al. 2018), OSNet (Zhou et al. 2019), can replace the teacher and student networks.

For ease of description, we mathematically formulate the construction of training samples for the T-net and S-net. Given the available images of identity i, we randomly select K images and obtain the set of K-shot images $\mathcal{S}_i = \{\mathcal{X}_{i,1}, \cdots, \mathcal{X}_{i,K}\}$, where $\mathcal{X}_{i,k} \in \mathbb{R}^{H\times W\times 3}$. We obtain an input sample for the T-net by concatenating the K-shots in channel as $\mathcal{Y}_i \in \mathbb{R}^{H\times W\times 3K}$. $\langle \mathcal{Y}_i, \mathcal{X}_{i,k}\rangle$ forms a teacher-student training pair, where $k = 1, \cdots, K$. Different from previous teacher-student networks that share the same input data, an input data to our student network (*i.e.*, $\mathcal{X}_{i,k}$) is only part of the input data to our teacher network (*i.e.*, $\mathcal{Y}_i$).

3.3 Knowledge Transfer with UA-KDL

Re-id aims to learn discriminative feature vectors for matching, *e.g.*, in terms of l_2 distance. We expect the S-net to learn a representation that is predictive of the representation of the T-net, at both the intermediate layers and the final re-id features. At an intermediate feature level, considering that the inputs to the T-net and S-net $\langle \mathcal{Y}_i, \mathcal{X}_{i,k}\rangle$ are different with spatial misalignment in contents, the intermediate feature maps are spatially average pooled to obtain a feature vector before the regularization supervision. We denote $y_i^b = \phi_t^b(\mathcal{Y}_i) \in \mathbb{R}^{c_b}$ as the feature vector at stage b of the T-net with the input $\mathcal{Y}_i$, where $b = 1, \cdots, 5$. Similarly, we denote $x_{i,k}^b = \phi_s^b(\mathcal{X}_{i,k}) \in \mathbb{R}^{c_b}$ as the feature vector at stage b of the S-net with the input $\mathcal{X}_{i,k}$. We encourage the S-net with a single-shot as input to learn/predict the more comprehensive information from the T-net with an input of K-shot images by minimizing the knowledge distillation loss as

$$\mathcal{L}_{KD(i,k)}^b = \|\theta_t^b(y_i^b) - \theta_s^b(x_{i,k}^b)\|^2, \quad (1)$$

where $\theta_t^b(y_i^b)$ and $\theta_s^b(x_{i,k}^b)$ denote projection functions that embed the feature vectors y_i^b and $x_{i,k}^b$ of stage b of the T-net and S-net to the same space/domain. Here $\theta_t^b(y_i^b) = ReLU(BN(W_t^b y_i^b))$ and $\theta_s^b(x_{i,k}^b) = ReLU(BN(W_s^b x_{i,k}^b))$, which is achieved by a fully-connected layer with matrix $W_t^b \in \mathbb{R}^{c_b\times c_b/r_b}$ or $W_s^b \in \mathbb{R}^{c_b\times c_b/r_b}$ followed by a Batch Normalization and ReLU activation function. r_b denotes dimension reduction ratio to reduce the model complexity and aid generalisation. We set r_b=16 for $b = 1, \cdots, 4$, and r_5=4 experimentally. Based on the analysis in subsec. 3.1, we can assume the T-net is always better than the S-net in terms of the feature representations. Thus the projected feature $\theta_t^b(y_i^b)$ of the T-net can be considered as the regression target of the S-net. Besides the updating of projection functions $\theta_t^b(\cdot)$ and $\theta_t^b(\cdot)$, the loss are only back-propagated to the S-net to regularize its feature learning as illustrated in Fig. 2.

Considering that the K samples of the S-net $\mathcal{S}_i = \{\mathcal{X}_{i,1}, \cdots, \mathcal{X}_{i,K}\}$ correspond to the teacher with the same K-shot input $\mathcal{Y}_i$, we can optimize the S-net simultaneously from the K samples with the knowledge distillation loss as

$$\mathcal{L}_{KD(i,:)}^b = \sum_{k=1}^{K} \|\theta_t^b(y_i^b) - \theta_s^b(x_{i,k}^b)\|^2. \quad (2)$$

However, in the above formulation, the *heteroscedastic uncertainty* of each sample to approach the features of the T-net is overlooked, where S-net's samples are equally treated.

Heteroscedastic uncertainty has been studied from the Bayesian viewpoint and applied to per-pixel depth regression and semantic segmentation tasks, respectively (Kendall and Gal 2017; Gal 2016). It captures noise inherent in the observations, which depends on the input data.

For re-id, different individual images have different visible object regions, occlusions, image quality, and thus have different capability/uncertainty in acquiring/approaching the knowledge of the given K-shot images of the same identity. Motivated by the uncertainty analysis in Bayesian deep learning and its application in depth regression (Kendall and Gal 2017), we design an Uncertainty-aware Knowledge Distillation Loss (UA-KDL) as

$$\begin{aligned}\mathcal{L}^b_{UKD(i,:)} = &\sum_{k=1}^{K} \frac{1}{2\sigma_b(y^b_i, x^b_{i,k})^2} \|\theta^b_t(y^b_i) - \theta^b_s(x^b_{i,k})\|^2 \\ &+ \log \sigma_b(y^b_i, x^b_{i,k})^2,\end{aligned} \tag{3}$$

where $\sigma_b(y^b_i, x^b_{i,k})$ denotes the observation noise parameter for capturing *heteroscedastic uncertainty* in regression and is data-dependent. Based on the uncertainty analysis in (Kendall and Gal 2017), minimizing this loss actually is equivalent to maximizing the log likelihood of the regression for the purpose of approaching the feature of T-net by the S-net $p(\theta^b_t(y^b_i)|\theta^b_s(\phi^b_s(\mathcal{X}_{i,k}))$. The introduction of uncertainty factors allows the S-net to adaptively allocate learning efforts on different samples for effectively training the network. For example, for a noisy image with the object being seriously occluded, the uncertainty to approach the feature of multi-shot images is high (*i.e.*, large σ_b) and it is wise to give small weight to the loss to have a smaller effect. The second item can prevent predicting infinite uncertainty (and therefore zero loss for the first item) for all images.

In our framework, the *heteroscedastic uncertainty* for regression depends on *both the feature of the K-shot images (which is the target)* and *the feature of the single image (which intends to approach the target)*. Then we model the log of uncertainty *i.e.*, $\upsilon_b(y^b_i, x^b_{i,k}) := \log(\sigma_b(y^b_i, x^b_{i,k})^2)$,

$$\upsilon_b(y^b_i, x^b_{i,k}) = ReLU(\mathrm{w_b}[\theta^b_t(y^b_i), \theta^b_s(x^b_{i,k})]), \tag{4}$$

where $[\cdot,\cdot]$ denotes the concatenation, $\mathrm{w_b}$ is achieved by a fully connected layer to map $[\theta^b_t(y^b_i), \theta^b_s(x^b_{i,k})]$ to a scalar followed by ReLU. Predicting the log of uncertainty is more numerically stable than predicting σ_b, since this avoids a potential division by zero in (3)(Kendall and Gal 2017).

3.4 Training and Inference

As in Fig. 2, for the K-shot images of the same identity i, the overall optimization loss consists of the widely used re-identification loss $\mathcal{L}_{ReID}$, and the proposed UA-KDLs:

$$\mathcal{L}_{(i,:)} = \mathcal{L}_{ReID(i,:)} + \sum_{b=1}^{5} \lambda_b \mathcal{L}^b_{UKD(i,:)}. \tag{5}$$

Note that we add the UA-KDL at all 5 stages (the first 4 stages and the final re-id feature vector of the last stage) to enable the knowledge transfer on intermediate features and the final re-id features. λ_b is a weight to control the relative importance for the regularization at stage b. In considering the re-id feature of stage 5 is more relevant to the task, we experimentally set $\lambda_5 = 0.5$, and $\lambda_b = 0.1$ for the first 4 stages. We find that training the T-net first to convergence and then fixing the T-net followed by the joint training of S-net and UA-KDL related parameters can produce better performance (about 1.4% gain in mAP on CUHK03(L)) than the end-to-end joint training. This can all along leverage the stable superior performance of the T-net.

In inference, we use only the S-net without any increase in computational or model complexity. The feature vector $x^5_{i,k}$ from stage 5 is the final re-id feature for matching.

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct object re-id experiments on the most commonly-used person re-id dataset, CUHK03 (Li et al. 2014) (including the labeled/detected bounding box settings), and three vehicle re-id datasets of VeRi-776 (Liu et al. 2016), VehicleID (Liu, Tian, and others 2016) and the recent large-scale VERI-Wild (Lou et al. 2019).

We follow common practices and use the cumulative matching characteristics (CMC) at Rank-1, and mean average precision (mAP) to evaluate the performance.

4.2 Implementation Details

We use ResNet-50 (He et al. 2016) to build the T-net, S-net, and baseline respectively. We set K as 4 and add UA-KDLs at all the 5 stages by default. The batch size is set as 64. Following (Hermans, Beyer, and Leibe 2017), a batch is formed by first randomly sampling P identities. For each identity, we then sample K images. Then the batch size is $P \times K$ for the S-net and P for the T-net. For simplicity, we refer to batch size with respect to the S-net hereafter. The input image resolution is set to 256×256 for vehicle re-id and 256×128 for person re-id, respectively.

We use the commonly used data augmentation strategies of random cropping (Zhang et al. 2019), horizontal flipping, label smoothing regularization (Szegedy et al. 2016), and random erasing (Zhong et al. 2017) in both the baseline schemes and our schemes. We use Adam optimizer (Kingma and Ba 2014) for model optimization. All our models are implemented on PyTorch and trained on a single NVIDIA-P40 GPU.

4.3 Ablation Study

We perform comprehensive ablation studies to demonstrate the effectiveness of the designs in our UMTS framework, on both the person re-id dataset CUHK03 (labeled bounding box setting) and the vehicle re-id dataset VeRi-776.

Effectiveness of Our Framework. Table 1 shows the comparisons of our schemes with the baseline. ***Baseline*** denotes the baseline scheme without taking into account multi-shot images. ***MTS*** denotes our **M**ulti-shot **T**eacher-**S**tudent Network with knowledge distillation *without* considering the *heteroscedastic uncertainty* (see formulation (2)). ***UMTS***

Table 1: Performance (%) of our schemes and *Baseline*. *MTS* denotes our Multi-shot Teacher-Student Network. *UMTS* denotes Uncertainty-aware MTS. (bm+bn) denotes the knowledge distillation losses are applied over stage m and n.

Model	CUHK03(L)		VeRi-776	
	Rank-1	mAP	Rank-1	mAP
Baseline	73.5	69.3	91.8	66.7
MTS($b5$)	74.3	71.1	92.9	69.3
UMTS($b5$)	76.6	72.8	93.7	70.9
UMTS($b1$+$b5$)	77.8	73.4	93.9	71.8
UMTS($b2$+$b5$)	78.6	74.3	94.4	73.5
UMTS($b3$+$b5$)	79.3	74.8	94.8	74.1
UMTS($b4$+$b5$)	78.8	74.5	94.4	74.0
MTS(all)	77.7	73.9	94.0	72.8
UMTS(all)	**80.7**	**75.5**	**95.8**	**75.9**

denotes our final **U**ncertainty-aware **M**ulti-shot **T**eacher-**S**tudent Network (see formulation (3)). ***UMTS(all)*** denotes that the UA-KDL is applied at all five stages ($b1$ to $b5$). Similarly, ***UMTS(b5)*** denotes that the UA-KDL is only added at stage 5 while there is no knowledge distillation loss on the other 4 stages. We make the following observations/conclusions.

1) Thanks to the exploration of the knowledge from multi-shot images, and the *heteroscedastic uncertainty*, our final scheme *UMTS(all)* significantly outperforms *Baseline* by **6.2%** and **8.9%** in mAP accuracy on CUHK03(L) and VeRi-776, respectively.
2) By learning the knowledge from multi-shots, our *MTS(all)* outperforms *Baseline* by **4.6%** and **6.1%** in mAP on CUHK03(L) and VeRi-776, respectively.
3) *UMTS(all)*, which introduces the *heteroscedastic uncertainty*, further improves the mAP accuracy by **1.6%** and **2.8%** on CUHK03(L) and VeRi-776, respectively.

4.4 Design Choices of UMTS

Which Stage to Add UA-KDL? To transfer the knowledge from the T-net to the S-net, we add the UA-KDL over the final stage re-id features (which are the most task-relevent) as the scheme *UMTS(b5)*. *UMTS(b5)* outperforms *Baseline* by 3.5% and 4.2% in mAP on CUHK03 and VeRi-776, respectively. We compare the cases of adding an UA-KDL to a different stage (Conv Block), and adding UA-KDLs to all stages (*i.e.*, see Fig. 2). Table 1 shows the results. We observe that on each stage, the adding of UA-KDL leads to obvious improvement and the gains are larger on stages 3, 4 and 2. When UA-KDL are added to all 5 stages, our scheme *UMTS(all)* achieves the best performance.

Influence of the Number of Shots K and Batch Size B. We study the influence of the number of shots K (K=2, 4, 8) on re-id performance under the settings of different batch sizes B (B=32, 64, 96 which is commonly used in re-id) on CUHK03 and VeRi-776 datasets and show the results in Fig. 4. We have the following observations.

1) For batch size B=64, the setting with K=4 shots provides the best performance. That may be because a smaller number of shots (*e.g.*, K=2) for the T-net cannot provide enough comprehensive information. When the shot number is too large, *e.g.*, K=8, the number of samples (equivalent batch size) for the T-net is small, *e.g.*, B/K=8, which is not enough to have precise statistical estimation for the Batch Normalization layers.
2) When the batch size is small, *i.e.*, B=32, large shot number K=8 results in inferior performance because of too small number (B/K=4) of samples in a batch for the T-net. When the batch size is increased to B=96, the performance for K=8 shots further increases on CUHK03 and saturates on VeRi-776. For K=4, the increase of batch size does not bring benefit since using large batch tends to converge to sharp minimums and leads to poorer generalization (Keskar et al. 2017). Besides, too large batch size requires significant GPU memory resources.

We set B=64 and $K = 4$ to trade off the performance and GPU memory requirement.

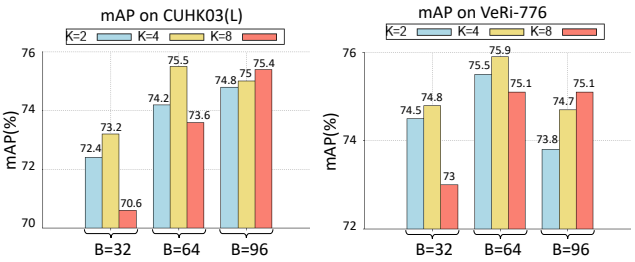

Figure 4: Study on the number of shots K and batch size B.

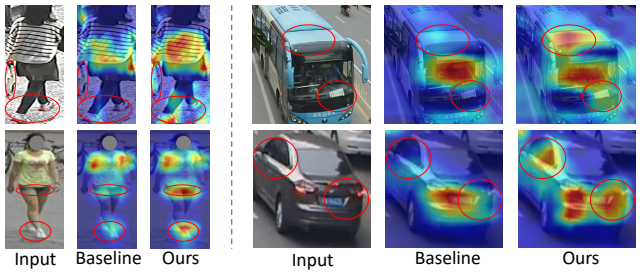

Figure 5: Gradient responses from *Baseline* and our S-net (*Ours*) using the tool Grad-CAM. Best viewed in color.

4.5 Visualization

Visualization of Feature Maps. To understand how the multi-shot images benefit the feature learning of the single image, we use Grad-CAM (Selvaraju et al. 2017) to visualize the gradient responses of *Baseline* and S-net of our *UMTS* in Fig. 5. We observe that *Baseline* tends to pay more attention to some local regions and ignore some potential discriminative regions on an object. In contrast, by exploiting knowledge from multi-shot images which have a more comprehensive perspective, our S-net can pay attention to more regions to capture more discriminative information, such as 'bag', 'shoes' (first row), and 'shorts' (second row) on the persons, and 'inspection sticker' on the bus.

Table 2: Performance (%) comparsions of the proposed UMTS and state-of-the-art methods on the vehicle re-id datasets.

Methods	VeRi-776		VehicleID						VERI-Wild					
			Small=800		Medium=1600		Large=2400		Small		Medium		Large	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
GSTE (TMM) (Bai et al. 2018)	–	–	–	–	–	–	–	–	60.4	31.4	52.1	26.1	45.3	19.5
VAMI (CVPR) (Zhou and Shao 2018)	77.0	50.1	63.1	–	52.8	–	47.3	–	–	–	–	–	–	–
FDA-Net (CVPR) (Lou et al. 2019)	84.2	55.4	–	–	59.8	65.3	55.5	61.8	64.0	35.1	57.8	29.8	49.4	22.7
Part-regularized (CVPR) (He et al. 2019)	94.3	74.3	78.4	–	75	–	74.2	–	–	–	–	–	–	–
MoV1+BS (IJCNN) (Kumar et al. 2019)	90.2	67.6	78.2	86.1	–	81.7	–	78.2	82.9	68.7	77.6	61.1	69.5	49.7
Baseline	91.8	66.7	74.4	80.4	72.4	77.1	69.8	75.2	77.6	65.2	72.7	60.3	63.8	45.0
UMTS	**95.8**	**75.9**	**80.9**	**87.0**	**78.8**	**84.2**	**76.1**	**82.8**	**84.5**	**72.7**	**79.3**	**66.1**	**72.8**	**54.2**

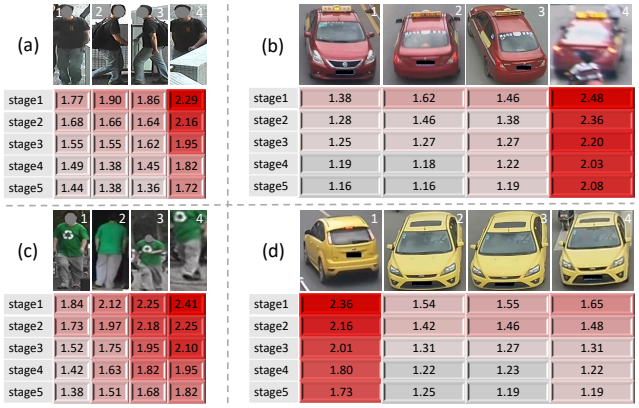

Figure 6: Predicted uncertainty σ_b^2 for the 4-shot images on the 5 stages respectively. Red represents large weight and white means small weight. Best viewed in color.

Visualization of Learned Uncertainty σ_b^2. We visualize the predicted uncertainty factor σ_b^2 for the K-shot images (K=4) on the five stages respectively in Fig. 6. In (a) and (b), the uncertainties of image-4 (I_4) are both relatively large due to occlusion and poor image quality (blur). In (c), the uncertainty values of I_3 and I_4 are larger than that of I_1 and I_2, which may be because of the small scale and incompleteness of the person. Fig. 6 (d) shows that I_1 belongs to the minority and thus has the highest uncertainty while the other similar images have similar low uncertainty.

4.6 Comparison with State-of-the-Arts

We compare the proposed UMTS with other state-of-the-art approaches and show the results in Table 2 and Table 3 for vehicle re-id and person re-id, respectively. With the same network structure in inference, our UMTS significantly outperforms *Baseline(ResNet-50)* on all the datasets, by **9.2%** and **7.2%** in mAP on the large-scale VERI-Wild for vehicle re-id, and CUHK03 (labeled) for person re-id, respectively. Our approach achieves the best performance on *all* the vehicle re-id datasets and most of the person re-id datasets. On the large-scale vehicle dataset VERI-Wild, our approach outperforms the second best approach by 4.0%, 5.0%, 4.5% for small, medium, and large sub-test sets, respectively.

Besides, our proposed UMTS is a general framework, and can be easily applied to other powerful backbone networks to achieve superior performance. Table 3 also shows the comparison when using OSNet (Zhou et al. 2019) as the backbone for person re-id. In comparison with this superior baseline (which outperforms ResNet-50 backbone by 1.8% in mAP), our UMTS achieves 6.3% improvement in mAP, and achieves the best performance.

Table 3: Performance (%) comparsions of UMTS and state-of-the-art methods on the person re-id dataset CUHK03.

Method	CUHK03			
	Labeled		Detected	
	Rank-1	mAP	Rank-1	mAP
HA-CNN (CVPR) (Li, Zhu, and Gong 2018)	44.4	41.0	41.7	38.6
PCB+RPP (ECCV) (Sun et al. 2018)	63.7	57.5	–	–
Mancs (ECCV) (Wang et al. 2018a)	69.0	63.9	65.5	60.5
MGN (ACMMM) (Wang et al. 2018b)	68.0	67.4	66.8	66.0
HPM (AAAI) (Fu et al. 2019)	63.9	57.5	–	–
CAMA (CVPR) (Yang et al. 2019)	70.1	66.5	66.6	64.2
CASN (CVPR) (Zheng et al. 2019a)	73.7	68.0	71.5	64.4
DSA-reID (CVPR) (Zhang et al. 2019)	78.9	75.2	78.2	73.1
Baseline (ResNet-50)	73.5	69.3	70.0	66.0
UMTS (ResNet-50)	**80.7**	**75.5**	77.2	73.4
Baseline (OSNet) (ICCV) (Zhou et al. 2019)	–	–	72.3	67.8
UMTS (OSNet)	–	–	**78.6**	**74.1**

5 Conclusion

In this paper, we propose a simple yet powerful Uncertainty-aware Multi-shot Teacher-Student (UMTS) framework to exploit the comprehensive information of multi-shot images of the same identity for effective single image based re-id. In particular, to efficiently transfer knowledge from the T-net to S-net, we take into account the *heteroscedastic uncertainty* related to the single image input to the S-net and the K-shot images input to the T-net and design an Uncertainty-aware Knowledge Distillation Loss (UA-KDL) which is applied at different semantics levels/stages. Extensive experiments on person re-id and vehicle re-id both demonstrate the effectiveness of the designs. Our UMTS achieves the best performance on all the three vehicle re-id datasets and the person re-id dataset. In inference, we only use the S-net without any increase in computational cost and model complexity.

6 Acknowledgments

This work was supported in part by NSFC under Grant 61571413, 61632001. We would like to thank Yizhou Zhou for the valuable discussion.

References

Bai, Y.; Lou, Y.; Gao, F.; et al. 2018. Group-sensitive triplet embedding for vehicle reidentification. *IEEE TMM* 20(9):2385–2399.

Chen, G.; Choi, W.; Yu, X.; et al. 2017. Learning efficient object detection models with knowledge distillation. In *NeurIPS*.

Fu, Y.; Wei, Y.; Zhou, Y.; et al. 2019. Horizontal pyramid matching for person re-identification. In *AAAI*, volume 33, 8295–8302.

Gal, Y. 2016. *Uncertainty in deep learning*. Ph.D. Dissertation, PhD thesis, University of Cambridge.

Ge, Y.; Li, Z.; Zhao, H.; et al. 2018. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NeurIPS*.

He, K.; Zhang, X.; Ren, S.; et al. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, B.; Li, J.; Zhao, Y.; and Tian, Y. 2019. Part-regularized near-duplicate vehicle re-identification. In *CVPR*, 3997–4005.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huang, Y.; Xu, J.; Wu, Q.; et al. 2018. Multi-pseudo regularized label for generated data in person re-identification. *TIP* 28(3).

Kalayeh, M. M.; Basaran, E.; Gökmen, M.; et al. 2018. Human semantic parsing for person re-identification. In *CVPR*.

Kendall, A., and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 7482–7491.

Keskar, N. S.; Mudigere, D.; Nocedal, J.; et al. 2017. On large-batch training for deep learning: Generalization gap and sharp minima. In *ICLR*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kumar, R.; Weill, E.; Aghdasi, F.; et al. 2019. Vehicle re-identification: an efficient baseline using triplet embedding. *IJCNN*.

Li, W.; Zhao, R.; Tian, L.; et al. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 152–159.

Li, D.; Chen, X.; Zhang, Z.; et al. 2017. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*.

Li, W.; Zhu, X.; and Gong, S. 2018. Harmonious attention network for person re-identification. In *CVPR*.

Liu, X.; Liu, W.; Yang, Y.; et al. 2016. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *ECCV*, 869–884.

Liu, H.; Feng, J.; Qi, M.; et al. 2017. End-to-end comparative attention networks for person re-identification. *TIP* 3492–3506.

Liu, J.; Ni, B.; Zhuang, Y.; et al. 2018a. Pose transferrable person re-identification. In *CVPR*.

Liu, X.; Zhang, S.; Huang, Q.; et al. 2018b. Ram: A region-aware deep model for vehicle re-identification. In *ICME*.

Liu, H.; Tian, Y.; et al. 2016. Deep relative distance learning: Tell the difference between similar vehicles. In *CVPR*, 2167–2175.

Lou, Y.; Bai, Y.; Liu, J.; et al. 2019. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *CVPR*.

Qian, X.; Fu, Y.; Wang, W.; et al. 2018. Pose-normalized image generation for person re-identification. In *ECCV*.

Romero, A.; Ballas, N.; Kahou, S. E.; et al. 2015. Fitnets: Hints for thin deep nets. In *ICLR*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; et al. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

Su, C.; Li, J.; Zhang, S.; et al. 2017. Pose-driven deep convolutional model for person re-identification. In *ICCV*.

Subramaniam, A.; Chatterjee, M.; and Mittal, A. 2016. Deep neural networks with inexact matching for person re-identification. In *NeurIPS*, 2667–2675.

Suh, Y.; Wang, J.; Tang, S.; et al. 2018. Part-aligned bilinear representations for person re-identification. In *ECCV*.

Sun, Y.; Zheng, L.; Yang, Y.; et al. 2018. Beyond part models: Person retrieval with refined part pooling. In *ECCV*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*.

Wang, Z.; Tang, L.; Liu, X.; et al. 2017. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *ICCV*, 379–387.

Wang, C.; Zhang, Q.; Huang, C.; et al. 2018a. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*.

Wang, G.; Yuan, Y.; Chen, X.; et al. 2018b. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 274–282.

Wang, X.; Hu, J.-F.; Lai, J.-H.; et al. 2019. Progressive teacher-student learning for early action prediction. In *CVPR*, 3556–3565.

Yang, W.; Huang, H.; Zhang, Z.; et al. 2019. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *CVPR*, 1389–1398.

Zhang, Z.; Lan, C.; Zeng, W.; et al. 2019. Densely semantically aligned person re-identification. In *CVPR*.

Zhao, H.; Tian, M.; Sun, S.; et al. 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*.

Zhao, L.; Li, X.; et al. 2017. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 3239–3248.

Zheng, M.; Karanam, S.; Wu, Z.; et al. 2019a. Re-identification with consistent attentive siamese networks. In *CVPR*, 5735–5744.

Zheng, Z.; Yang, X.; Yu, Z.; et al. 2019b. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2138–2147.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 3754–3762.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2017. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*.

Zhou, Y., and Shao, L. 2017. Cross-view gan based vehicle generation for re-identification. In *BMVC*, volume 1, 1–12.

Zhou, Y., and Shao, L. 2018. Viewpoint-aware attentive multi-view inference for vehicle re-identification. In *CVPR*, 6489–6498.

Zhou, G.; Fan, Y.; Cui, R.; et al. 2018. Rocket launching: A universal and efficient framework for training well-performing light net. In *AAAI*.

Zhou, K.; Yang, Y.; Cavallaro, A.; et al. 2019. Omni-scale feature learning for person re-identification. *ICCV*.