Towards Efficient Front-End Visual Sensing for Digital Retina: A Model-Centric Paradigm

Yihang Lou, Ling-Yu Duan, *Member, IEEE*, Yong Luo, Ziqian Chen, Tongliang Liu, Shiqi Wang, *Member, IEEE*, and Wen Gao, *Fellow, IEEE*

***Abstract*—The digital retina excels at providing enhanced visual sensing and analysis capability for city brain in smart cities, and can feasibly convert the visual data from visual sensors into semantic features. With the deployment of deep learning or handcrafted models, these features are extracted on front-end devices, then delivered to back-end servers for advanced analysis. In this scenario, we propose a model generation, utilization and communication paradigm, aiming at strong front-end sensing capabilities for establishing better artificial visual systems in smart cities. In particular, we propose an integrated multiple deep learning models reuse and prediction strategy, which dramatically increases the feasibility of the digital retina in large-scale visual data analysis in smart cities. The proposed multi-model reuse scheme aims to reuse the knowledge from models cached and transmitted in digital retina to obtain more discriminative capability. To efficiently deliver these newly generated models, a model prediction scheme is further proposed by encoding and reconstructing model differences. Extensive experiments have been conducted to demonstrate the effectiveness of proposed model-centric paradigm.**

***Index Terms*—Digital retina, model reuse, model communication, visual sensing.**

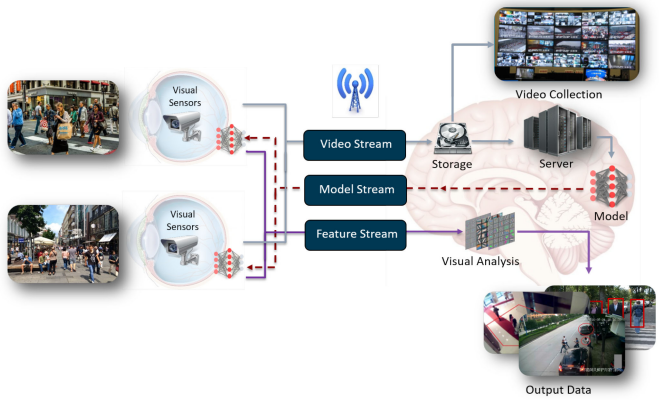

Fig. 1. Illustration of digital retina framework in smart cities. The digital retina includes front-end smart visual sensing and back-end intelligent analysis. The interaction between front-end and back-end includes video stream, model stream, and feature stream.

I. Introduction

In THE era of big data, massive surveillance cameras deployed in urban areas form the visual systems of cities. However, the video streams collected from front-end cameras impose great challenges to the bandwidth and computational resources. Due to the insufficient sensing capabilities of the front-end, the city brain is unable to monitor the current dynamics in real-time. Attempting to address such computational and transmission issues, the emerging digital retina is revolutionizing the artificial vision system of smart cities.

Manuscript received November 14, 2019; accepted January 4, 2020. Date of publication January 15, 2020; date of current version October 23, 2020. This work was supported in part by the National Natural Science Foundation of China under Grants U1611461 and 61425025, and in part by the Australian Research Council Project under Grant DE-1901014738, and in part by the Hong Kong RGC Early Career Scheme under Grants 9048122 and CityU 21211018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Marta Mrak *(Corresponding author: Lingyu Duan.)*

Y. Lou, L.-Y. Duan, Y. Luo, Z. Chen, and W. Gao are with the Institute of Digital Media, Peking University, Beijing 100871, China, and also with the China and the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: yihanglou@pku.edu.cn; lingyu@pku.edu.cn; yongluo@pku.edu.cn; wzziqian@pku.edu.cn; wgao@pku.edu.cn).

T. Liu is with the University of Sydney, Sydney, NSW 2006, Australia (e-mail: tongliang.liu@sydney.edu.au).

S. Wang is with the Department of Computer Science, City University of Hong Kong, Kowloon Tong 999077, Hong Kong (e-mail: shiqwang@cityu.edu.hk).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.2966885

In human visual system, the retina is a crucial and indispensable component. In the retina, the rods and cones cells are responsible for perceiving the low and high light levels, as well as the color vision. Then, light signals are converted into neuronal representations by retina. Besides, retina acts as a filter which conveys specifically required and meaningful visual information to the brain [1], [2]. To complete the perception process, the photosensitive retinal ganglion cells extract the complex features. Therefore, retina not only perceives the visual information, but also works as a highly efficient visual data processing engine in the central nervous system. Motivated by the concept of retina, as illustrated in Fig. 1, the digital retina [3] includes front-end smart visual sensing and back-end intelligent analysis in the city brain.

Regarding the digital retina, the interactions between front-end and back-end are performed in terms of video stream, feature stream and model stream. In smart cities, the main element of visual system is video stream from massive surveillance cameras. The computing system transmits highly compressed video streams from front-ends to back-ends where visual analysis can be intensively performed. Video coding standards such as HEVC [4] and AVS [5] have been developed towards achieving highest quality representation of videos given limited bandwidth. With respective to the feature steam, to improve the data exchange and collaboration efficiency, the MPEG organization launched standardization work of CDVS [6] and CDVA [7], [8],

which generate the compact feature of visual information to achieve interoperability of different ends. In digital retina, the front-end visual sensors directly extract and compress the features, such that the compact features can be efficiently sent to central server. For a typical surveillance visual system, a series of analysis tasks can be performed based on visual features, *e.g.*, person re-identification [9], [10], vehicle retrieval [11]–[14], face recognition [15], [16]. Besides video stream and feature stream, we propose that the interactions of model stream between front-back ends are also necessary in the digital retina, and the entire digital retina framework is illustrated in Fig. 1.

The central process of digital retina is visual representation, transmission and analysis. In this context, the models are the core components in the city brain. The models are usually learned at the central server by leveraging the visual data, and the learned models are subsequently delivered to the front-end devices for feature extraction and compression. As such, the model generation, utilization and communication are essential in establishing the digital retina, especially in the sense that the collected visual data are featured by high variations in terms of locations, time and ambient environments. In particular, the recent advances of artificial intelligence technologies [17] are driving many applications for understanding the current city dynamics and ensuring the securities [18], [19]. However, how the models-of-interest can be feasibly generated by leveraging the existing massive models in different domains have not been fully exploited.

In this paper, we focus on an integrated model-centric solution for efficient front-end sensing in digital retina. The proposed scheme relies on effective model generation and efficient transmission by exploiting the cross-domain and inter-model relationships for the construction of digital retina. The transfer learning, which has been studied for decades from the psychological [20] and educational [21] perspectives, motivates us to design a novel multi-model reuse method based on the widely acknowledged evidence that learning from prior experience or knowledge transfer is beneficial to the model learning. As such, the existing models can be reused for model generation, even when the source and target models are in different domains. Moreover, as knowledge correlations between different source models and target models are different, a novel adaptive model weighting scheme is also incorporated for better model reusing.

It is commonly believed that the energy consumption for the information transmission between neurons in human brain is usually very light due to the selective and adaptive regulation mechanism [22]. This inspires us to propose a low transmission cost strategy that enables incremental and adaptive delivery of the deep learning models. To promote the application of deep learning models, the model compression approaches have been widely investigated to produce light-weight models. However, the model communication, instead of single model compression, has been largely ignored. In smart cities, the models may undergo generation, updating, and distribution process, such that existing models and the to-be-transmitted models usually have high correlation in weighting parameters. In view of this, a model prediction method based on Difference of Models (DoM) has been incorporated for model efficient delivering. In order to efficiently deploy the DoM to front-ends, we also further investigate the lossy compression scheme.

In this work, towards establishing extremely economic, efficient and effective digital retina in smart cities, we aim to explore the model-centric approach to enhance the smart sensing capabilities at the front-end based on the model generation, utilization, and distribution methodologies. The main contributions of this paper are summarized as follows,

- We propose a novel model generation, utilization and communication paradigm towards digital retina to better construct the artificial vision system in smart cities.
- We explore a multi-model reuse method with adaptive model weighting scheme to learn more discriminative models in the desired target domain.
- We design a model prediction scheme from the perspective of transmitting difference of model to efficiently deliver the newly generated models.

This work extends our previous conference version [23] from the following perspectives. First, we propose a novel adaptive model weighting scheme to attentively reuse the knowledge from the source models. Second, we detail the theoretical proof and offer more analysis for the proposed multi-model reuse scheme. Third, we explore more quantization methods for DoM based model prediction. Finally,we conduct extensive experiments in terms of hyper parameters sensitivity, model convergence, ablation study and visualization analysis.

The rest of the paper is organized as follows. In section II, we review the related work including model reuse and deep network compression. We further introduce the model-centric paradigm with generation, utilization and communication in section III. The proposed multiple model reuse and prediction scheme is presented in section IV. Experimental results are demonstrated and discussed in section V. Finally, section VI concludes this paper.

II. Related Work

Model Reuse: Many existing models that contain different knowledge are off the shelf. Recently, these existing models are reused to promote the training for a target model with limited labeled data and computational resources. In [24], Ajakan *et al.* trained hidden layers for fitting multiple domains. Ghifary *et al.* [25] used maximum mean discrepancy measure to reduce distribution mismatch in the latent space. However, these approaches only reuse the structure and weights directly from the source networks and can hardly utilize the existing pre-provided model/features.

Recently, in [26], Yang *et al.* proposed a fixed model reuse (FMR) method that uses a set of additional features for each training sample. In FMR, the method only uses features and more information in the models are not considered. Moreover, only one type of feature can be reused is also a limitation of FMR. In [27], Wu *et al.* proposed a heterogeneous model reuse method via optimizing multiparty multiclass margin in the context that several remote local models are available, such that

multiple local models can be reused to approximate a global model. In [28], Xiang *et al.* proposed a multi-model reuse strategy, but it assumes there are several modalities in each domain and cannot be applied to one modality. Moreover, it only uses the label information and requires several source models involved to be presented in testing phase. In [29], Jha *et al.* proposed a bag of experts method that directly reuses the output features or labels of several expert source models, but the same limitations as in [26] or [28] are observed. Besides, Lou *et al.* [23] developed a multi-model reuse strategy based on multi-view subspace learning, where abundant unlabeled data are utilized.

In this work, we propose a novel and generalized framework that can well alleviate these problems and make full use of the knowledge contained in existing source models towards a better target model, which greatly promotes the model generation and knowledge centric networking in city brain.

Network Compression and Transmission: Despite the promising performance of deep learning models in the current artificial vision system of smart cities, most deep learning models suffer from the enormous cost in terms of storage and computation due to huge amount of parameters. For instance, the VGG16 [30] has more than 60 M parameters, and it is widely acknowledged that there exists abundant redundancy within the model [31]. To compress the deep learning models without sacrificing the performance, numerous approaches have been proposed, which can be categorized into the following categories: (1) Parameters pruning: pruning the neurons with weak response or unnecessary neurons in the deep network [32], [33]. (2) Matrix factorization: reconstructing neuron weights based on the low-rank methods [34], [35]. (3) Filter selection: removing unnecessary filters in network [36], [37]. (4) Quantization: representing the weight compactly in scalar quantization or vector quantization [38], [39]. The above methods all aim to economize the model storage space. However, the transmission and communication of the deep neural network have been largely ignored.

The deep network transmission and communication is a typical type of applications in Knowledge Centric Networking (KCN)[40] which reforms traditional Content Centric Networking (CCN)[41] with machine learning based knowledge generation, utilization and distribution. In digital retina, a model serves as a modality of knowledge for intelligent analysis. The deep neural network transmission aims to utilize and deliver the knowledge concentrated in the network model to facilitate different intelligent applications. In [42], Chen *et al.* formulate the model compression from the perspective of model transmission. In incremental model updating, the redundancy among updating models of different versions can be further exploited to promote numerous applications in front-end visual sensors. Such scheme can be elegantly integrated into compression and communication framework with the existing compression methods.

III. Model-Centric Paradigm With Generation, Utilization, and Communication

In this section, we demonstrate the model generation, utilization and transmission paradigm in the digital retina system of smart cities. The digital retina can be decomposed to the front-end and back-end. Such architecture enjoys the data generation and sensing capability in front-ends and computation capability in back-ends. The front-end devices constantly generate data, which are transmitted to the back-end for knowledge learning. In particular, the knowledge learning from a large quantity of raw data requires high computation and storage cost. The back-end fills the large gap between the low capabilities of the front-end and high demand in terms of the computational cost. Therefore, the generated data on the front-end are constantly conveyed to the back-end server, enabling the progressive training of analysis models. Such front-back end architecture can be elegantly applied to the visual sensing system in smart cities. With this scheme, the front-end cameras constantly capture and transmit visual data to the back-end. As such, the models can be progressively trained or fine-tuned. The temporal and scenario different data distribution can be covered with this manner, consistently providing new knowledge and promoting the performance of the existing models.

For efficient knowledge creation, utilization and distribution, we propose a model-centric paradigm including model generation, utilization and communication in front-back end based digital retina system. As illustrated in Fig. 2, the proposed paradigm involves edge end, which serves as an intermediate layer between the front-end visual sensor and back-end central cloud. By offloading the computation from back-end and caching the data from front-end at the edge side, the whole efficiency of the digital retina system can be significantly improved. Accordingly, there arise multiple requirements for model generation and communication from different perspectives, which can be summarized as follows,

1) Model Generation: The initial models are generated on the back-end with a large quantity of training data. Therefore, the edge nodes need to collect the generated data on the front-end and transmit them to the center server, enabling the progressive training of analysis models. Since the analysis tasks are performed on the front-end visual sensors, the trained models are then distributed to the front-end devices to facilitate the analysis tasks with more discriminative feature representation and analysis.

2) Model Utilization: The generated models on back-end continuously flow through edge-end to the front-end. Therefore, there are many models cached on the edge-end, which have great potential to be used for generating a better target model. In real-world applications, the deployed front-end visual sensors in different locations may deal with the data in different domains. For a specific task, the models trained in different domains contain different knowledge, which can be utilized to generate a better domain adaptive model. For example, the trained models at several edge nodes can be used to promote a target model training, leading to better performance. Again, such paradigm allows the knowledge to be more efficiently utilized, and the models themselves can be customized to better accommodate to the target scenarios.

3) Model Communication: The front-end devices often receive the models from edge-end or back-end for deployment. For a particular task, a series of updating models will be continuously deployed to the front-end from time to time when more

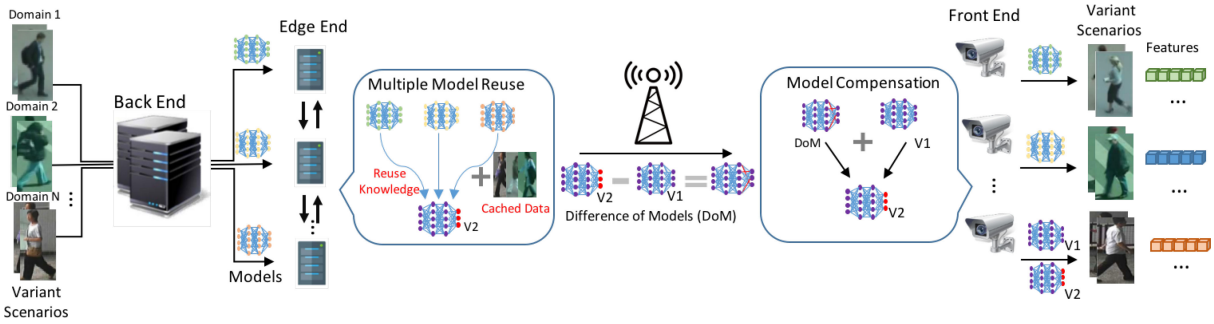

Fig. 2. Illustration of the proposed model generation, utilization and communication paradigm in digital retina. From left to right, the models are generated at back end, reused at edge end, and deployed to front end. The DoM-based model transmission is also performed in the deployment.

training data or better model optimization strategies are available. In such scenario, there exist high redundancies among a series of updating models. Therefore, transmitting entire updated models consumes large unnecessary transmission cost. An alternative solution is to explore the inter-model redundancy. Specifically, transmitting the updated weights can greatly reduce the transmission cost of the to-be-deployed models.

IV. Multiple Model Reuse and Prediction

In various application scenarios, the data distributions show severe domain bias, which is mainly caused by different acquisition conditions. Due to such a domain gap, the generalization capability of the model has been greatly affected. For example, the model trained in one specific domain cannot well generalize to other domains. In the visual system of smart cities, there often exist many similar models for the same analysis task. Collecting data in numerous front-ends for model training is infeasible in practice, due to the unacceptable transmission and annotation cost. However, the knowledge contained in models actually is the abstract representation of data domain, which motivates us to leverage knowledge in existing multiple models to obtain a domain adaptive one. In view of a large quantity of cached models and cached data on edge-end, the edge-end is an ideal place to perform model reuse before actual deployment.

In this section, we detail how the models can be effectively learned based on multiple model reuse, and how the generated models can be efficiently transmitted and delivered to facilitate the intelligent sensing at the front-ends in smart cities.

A. Multi-Model Reuse

In model reuse, the knowledge is reused from source models to target models. We assume there are M source domains and one target domain. Moreover, in the m-th source domain, a deep learning model $f_m(\Theta_m)$ is already trained using abundant data. Note that, the tasks of models in source domains could be different but similar to that in the target domain. By reusing the knowledge in source domains, our goal is to learn a model $f_T(\Theta_T)$ in the target domain with limited labeled data and a large quantity of unlabeled data.

We show the architecture of the proposed multi-model reuse framework in Fig. 3. A mild assumption is made that the pretrained source models and target model are Convolutional

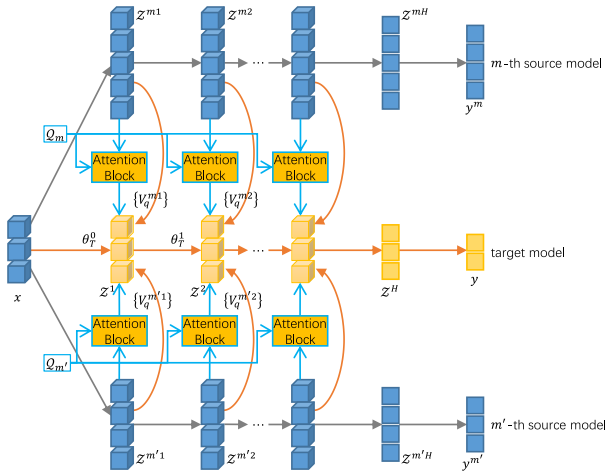

Fig. 3. Architecture of the multi-model reuse framework. The hidden layer representation from different source models are used for target model training. In particular, the attention block is designed for adaptive model weighting.

Neural Networks (CNNs) that are widely used in various visual analysis applications. In our proposed reuse method, since we transfer the knowledge in feature domain, the reuse stage is not constrained to the labeled data. Given input x, we map x as $\mathcal{Z}^h$ and $\mathcal{Z}^{mh}$, which are hidden layer representation in target model and m_{th} source model, respectively. For $\mathcal{Z}^{mh}, m = 1, \ldots, M$, h is the layer index and M is the number of source models. Here, each representation $\mathcal{Z}$ is a Q-order tensor of size $d_1 \times d_2 \times \ldots \times d_Q$. In order to reuse the multiple source models, different $\mathcal{Z}^{mh}$ are mapped into a common subspace. Then, we constrain $\mathcal{Z}^h$ to be close to the representation in common subspace. We use multi-view learning strategy [43] for knowledge reuse, *i.e.*, using $\mathcal{Z}^h$ to reconstruct each $\mathcal{Z}^{mh}$. As such, $\mathcal{Z}^h$ can be considered as a meta-embedding of different source layer representations. In this way, the hidden layer representation can be improved by all features in the source domains. Compared to using only the limited labeled information in the target domain, a target model with better domain adaptive capability can be achieved. The formulation of model reuse $R(\cdot)$ is given by,

$$R(\mathcal{Z}_n^h; \{\mathcal{Z}_n^{mh}\}) = \sum_{m=1}^{M} \alpha_m \left\| \mathcal{Z}_n^{mh} - \mathcal{Z}_n^h \times_1 V_1^{mh} \ldots \times_q V_q^{mh} \right\|_F^2, \quad (1)$$

where each V_q^{mh} is a transformation matrix of size $d_q^{mh} \times d_q^h$, $\times_q$ is the q-mode tensor-matrix product. $\{\alpha_m\}$ are the weights that reflect the importance of different source models and satisfy $\sum_m \alpha_m = 1$. In a typical CNN structure, the representation of the shallow layers is of 2D shape. To ease the difficulty of optimizing $R(\cdot)$, we reduce the parameters in V_Q^{mh} by pooling the hidden representations as vectors. Thus, when $\mathcal{Z}$ is an one-order tensor ($Q = 1$), the formulation $R(\cdot)$ becomes,

$$R(\mathbf{z}_n^h; \{\mathbf{z}_n^{mh}\}) = \sum_{m=1}^{M} \alpha_m \left\| \mathbf{z}_n^{mh} - V^{mh} \mathbf{z}_n^h \right\|_2^2. \quad (2)$$

It is worth mentioning that the risk of over-fitting may be caused by additional parameters $\{V^{mh}\}$. This issue can be alleviated since large amount of unlabeled data are available to train $\{V^{mh}\}$.

B. Adaptive Model Weighting

The α_m in the Eqns. (1) and (2) is the weight that reflects the importance of the m-th source model. In essence, assigning a proper value for hyper-parameter α_m is a tough task because the values of knowledge contained in the different source models are unknown. Besides, with the increase of the number of reuse models, the prediction of hyper-parameters α_m will become more difficult. Attention model which allows the networks to selectively focus on specific information has been employed in image caption, machine translation and action recognition, *etc*. Similarly, selectively reusing knowledge from different models is also crucial to obtain a better target model and ensure the stability of optimization procedure. Therefore, we design an attention block for adaptive model weighting as follows,

$$e_m = q_m^T z_n^{mh}, \quad (3)$$

Here, q_m is the attention vector to be learned for the m-th source model. Then the model weighting parameter e_m is normalized as follows,

$$\alpha_m = \frac{exp(e_m)}{\sum_{m=1}^{M} exp(e_m)}. \quad (4)$$

During the reuse stage, the parameters q_m in attention block will be fine-tuned for adaptively weighting the reuse feature vector from different source models.

C. Joint Optimization

The model reuse is applied on the target model training stage as an extra regularization term. We assume that there are N images $\{x_n\}_{n=1}^N$ in the target domain for training, including limited labeled samples $\{x_n, y_n\}_{n=1}^{N_l}$ and large amounts of unlabeled data $\{x_n\}_{n=N_l+1}^N$. Fig. 4 shows the joint optimization pipeline. To achieve reliable model reuse, the objective function with the joint optimization fashion is given by,

$$\epsilon(\Theta_T; x_n, y_n) = \frac{1}{N_l} \sum_{n=1}^{N_l} L(f_T(\Theta_T; x_n), y_n) + \gamma \sum_{n=1}^{N} \sum_{h=1}^{H'<H} R(\mathcal{Z}_n^h; \{\mathcal{Z}_n^{mh}\}), \quad (5)$$

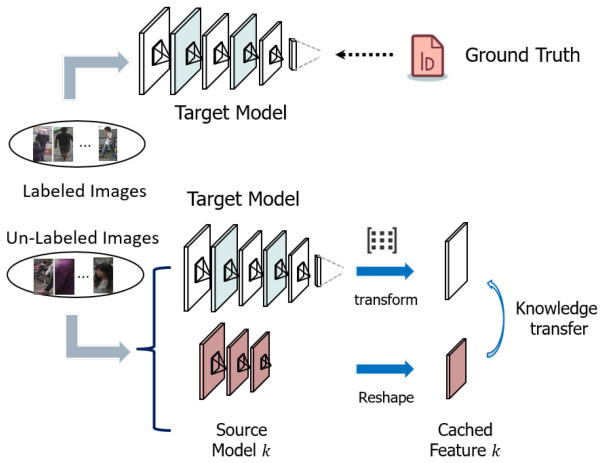

Fig. 4. Architecture of the joint optimization framework. In our learning stage, we use labeled images to optimize the target model, and use unlabeled images for model reuse. The intermediate features of target model are regularized by the features from source models.

where $\Theta_T = \{\theta_T^h\}_{h=0}^H$ is the set of all parameters of the target learning task. h is the layer index and the choice of layers is arbitrary. In order to obtain the common task-independent knowledge, we choose to reuse the lower layers of source models. $\mathcal{Z}_n^h = \varphi(\theta_T^{h-1}; \mathcal{Z}_n^{h-1})$ and $\varphi(\cdot)$ is an activation function. γ is the scale parameter to balance the $L(\cdot)$ and $R(\cdot)$; $L(\cdot)$ is loss of a specific task and $R(\cdot)$ is proposed multi-model reuse regularization term.

D. Theoretical Analysis

Here, we detail the theoretical analysis for the proposed multi-model reuse scheme. For simplicity, we treat the deep learning models as end-to-end learning algorithms, which are adopted in most deep learning methods. More specifically, we denote the prediction function for the target model as f_T, while the prediction functions for the source models are denoted as $f_m, m = 1, \ldots, M$. To measure the accuracy of a model f, we introduce the expected risk $\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim D}[L(f(x), y)]$. Then we have the following theorem, which shows that the proposed model can benefit from the source models with a theoretical guarantee on the expected risk. Here, the least square loss is simply adopted.

Theorem 1: Assume the linearity of their input-output map, *i.e.*, $(f_1 \pm f_2)(x) = f_1(x) \pm f_2(x)$ and $f(x) \leq \|f\|\|x\|$, and the feature space is bounded with an upper boundary r, *i.e.*, $\|x\| \leq r$, then when we adopt the least square loss, we can have

$$\mathcal{R}(f_{T,N_l}) \leq 2\,M \sum_{m=1}^{M} \alpha_m^2 \mathcal{R}(V^m f_m)(r^2/\gamma + 1). \quad (6)$$

Here, $\mathcal{R}(f_{T,N_l})$ is the expected risk of the target model trained using the N_l labeled samples, and $\mathcal{R}(V^m f_m)$ is the risk of the transformed source model *w.r.t.* the data in the target domain. Since the source models are well-trained using abundant labeled data, we believe that there exists (or we can learn) a V^m such that $\mathcal{R}(V^m f_m)$ is small if the source tasks are related to the target task. When $\{\alpha_m\}$ and $\{V^m\}$ are determined appropriately,

the term $M\sum_{m=1}^{M}\alpha_m^2\mathcal{R}(V^m f_m)$ can be small. As such, the expected risk $\mathcal{R}(f_{T,N_l})$ of the target model is guaranteed to be low. We do not have such a guarantee when the designed regularization term does not exist, since the expected risk of the target model would be high, in which the model is prone to over-fit for limited labeled data in the target domain.

Proof of Theorem: From Eqns. (1) and (2), we can see that, for each input data x, the newly proposed regularization pushes the representations on different layers to be similar among different models via learning some transformations. Because the input x for each model is the same, we interpret the regularization as to push the end-to-end functions to be similar via learning some transformations. We rewrite the regularization as $R(f_T) = \gamma\sum_{m=1}^{M}\alpha_m\|f_T - V^m f_m\|_2^2$, where $\sum_{m=1}^{M}\alpha_m = 1$ and V^m is a transformation matrix to be learned by matching f_T and f_m. When minimizing the regularization, it is to push f_T to be close to $\sum_{m=1}^{M}\alpha_m V^m f_m$ (this can be obtained by letting the derivative $R'(f_T) = 0$). To provide some insights, we exploit the least square loss and further rewrite the objective function (5),

$$O(f_T; x_n, y_n) = \frac{1}{N_l}\sum_{n=1}^{N_l}(f_T(x_n) - y_n)^2 + \gamma\|f_T - \sum_{m=1}^{M}\alpha_m V^m f_m\|_2^2. \quad (7)$$

Let $g_T = f_T - \sum_{m=1}^{M}\alpha_m V^m f_m$, we have

$$O(g_T; x_n, y_n) = \frac{1}{N_l}\sum_{n=1}^{N_l}(g_T(x_n) - y_n + \sum_{m=1}^{M}\alpha_m V^m f_m(x))^2 + \gamma\|g_T\|_2^2. \quad (8)$$

Let $g_{T,N_l} = \arg\min_{g_T} O(g_T; x_n, y_n) = f_{T,N_l} - \sum_{m=1}^{M}\alpha_m V^m f_m$, where the subscript N_l indicates that g_{T,N_l} and f_{T,N_l} are learned from the N_l labeled examples. Now we show that the expected risk $\mathcal{R}(f_{T,N_l})$ of the target model can be bounded using the risk $\{\mathcal{R}(f_m)\}_{m=1}^{M}$ of source models.

The expected risk for the target model is as,

$$\begin{aligned}\mathcal{R}(f_{T,N_l}) &= \mathcal{R}\left(g_{T,N_l} + \sum_{m=1}^{M}\alpha_m V^m f_m\right)\\ &= \mathbb{E}_{(x,y)\sim D}\left[\left(\left(g_{T,N_l} + \sum_{m=1}^{M}\alpha_m V^m f_m\right)(x) - y\right)^2\right]\\ &= \mathbb{E}_{(x,y)\sim D}\left[\left(g_{T,N_l}(x) + \sum_{m=1}^{M}\alpha_m V^m f_m(x) - y\right)^2\right]\\ &\le 2\mathbb{E}_{(x,y)\sim D}[(g_{T,N_l}(x))^2]\\ &\quad + 2\mathbb{E}_{(x,y)\sim D}\left[\left(\sum_{m=1}^{M}\alpha_m V^m f_m(x) - y\right)^2\right] \quad (9)\end{aligned}$$

Then,

$$\begin{aligned}\mathcal{R}(f_{T,N_l}) &= 2\mathbb{E}_{(x,y)\sim D}[(g_{T,N_l}(x))^2] + 2\\ &\quad \mathbb{E}_{(x,y)\sim D}\left[\left(\sum_{m=1}^{M}\alpha_m V^m f_m(x) - \sum_{m=1}^{M}\alpha_m y\right)^2\right]\\ &\le 2r^2\mathbb{E}_{(x,y)\sim D}[\|g_{T,N_l}\|_2^2] + 2M\sum_{m=1}^{M}\alpha_m^2\mathcal{R}(V^m f_m), \quad (10)\end{aligned}$$

where,

$$\begin{aligned}&\mathbb{E}_{(x,y)\sim D}\left[\left(\sum_{m=1}^{M}\alpha_m V^m f_m(x) - \sum_{m=1}^{M}\alpha_m y\right)^2\right]\\ &\le M\sum_{m=1}^{M}\alpha_m^2\mathbb{E}_{(x,y)\sim D}[(V^m f_m(x) - y)^2]\\ &= M\sum_{m=1}^{M}\alpha_m^2\mathcal{R}(V^m f_m). \quad (11)\end{aligned}$$

holds because of the Hölder's inequality.

Now, we are going to upper bound the term $\mathbb{E}_{(x,y)\sim D}[\|g_{T,N_l}\|_2^2]$. We have $O(g_{T,N_l}; x_n, y_n) \le O(0; x_n, y_n)$, i.e.,

$$\begin{aligned}&\frac{1}{N_l}\sum_{n=1}^{N_l}\left(g_{T,N_l}(x_n) - y_n + \sum_{m=1}^{M}\alpha_m V^m f_m(x)\right)^2 + \gamma\|g_{T,N_l}\|_2^2\\ &\le \frac{1}{N_l}\sum_{n=1}^{N_l}\left(y_n - \sum_{m=1}^{M}\alpha_m V^m f_m(x)\right)^2. \quad (12)\end{aligned}$$

Then, we have

$$\begin{aligned}\gamma\|g_{T,N_l}\|_2^2 &\le \frac{1}{N_l}\sum_{n=1}^{N_l}\left(y_n - \sum_{m=1}^{M}\alpha_m V^m f_m(x)\right)^2\\ &\le \frac{M}{N_l}\sum_{n=1}^{N_l}\sum_{m=1}^{M}\alpha_m^2(V^m f_m(x) - y_n)^2. \quad (13)\end{aligned}$$

Taking expectation on both side, we have

$$\mathbb{E}_{(x,y)\sim D}[\gamma\|g_{T,N_l}\|_2^2] \le M\sum_{m=1}^{M}\alpha_m^2\mathcal{R}(V^m f_m). \quad (14)$$

Thus,

$$\mathcal{R}(f_{T,N_l}) \le 2M\sum_{m=1}^{M}\alpha_m^2\mathcal{R}(V^m f_m)(r^2/\gamma + 1). \quad (15)$$

This completes the proof.

E. Model Prediction

After generating a better target model, it is necessary to efficiently and economically deploy it to the front-end visual sensors. Model communication is an essential component in visual system. In particular, the new models will be frequently updated with the adoption of model reuse, such that efficient

communication of these models is highly expected. Hence, we investigate difference of models (DoM) [42] between the existing model (*e.g.*, existing in both sender and receiver) and the to-be-transmitted model to explore an economic model communication philosophy. Specifically, the DoM between the prediction and the to-be-compressed models is calculated by the difference for each corresponding weight in each layer. Basically, we target at providing a conceptually meaningful way for deep learning model communication by removing inter model redundancy. Denote the weights of prediction model as W_p and the weights of to-be compressed model as W_c, then the DoM between W_p and W_c can be formulated as follows,

$$W_{DoM}(h,i) = W_c(h,i) - W_p(h,i), \tag{16}$$

where h and i represent the layer index and weight index in each layer, respectively. Then model compensation is performed to recover the model that is desired to be transmitted, *i.e.*,

$$\hat{W}_c(h,i) = \hat{W}_{DoM}(h,i) + W_p(h,i). \tag{17}$$

By means of transmitting DoM, the new model can be recovered at receiver side. In addition, the recovered source models from model prediction are also allowed to facilitate the model training in the target domain through the multi-model reuse method. Moreover, in some cases, only a small portion of the parameters or layers in deep models is updated and the others remain unchanged. In view of this, the DoM computation is based on the comparison between the partially updated weights.

In this work, we further investigate the use of lossy compression method to compress the DoM. We use k bits memory cost to represent each weight parameter in the network. The general structure of the quantization function can be formulated as follows,

$$Q(v) = sc^{-1}(\hat{Q}(sc(v))), \tag{18}$$

where sc^{-1} indicates the inverse of the scaling function, and $\hat{Q}$ is the actual quantization function. The input v could be multi-dimensional weights. In our investigation, we explore 4 different quantization methods on DoM, *i.e.*, Linear Quantization, Log MinMax Quantization, MinMax Quantization, and Vector Quantization [44].

V. Experimental Results

A. Experimental Setup

We conduct experiments on person ReID [9] which is a typical task in smart city applications aiming to find person images of the same identity as the query person image in the database. Four different person ReID datasets, Duke [45], Market1501 [46], MSMT17 [47], CUHK03 [48] are used in the experiments, and the details are shown in Table I. The reason for choosing this task is that variant capture conditions lead to severe domain bias between datasets. As shown in Fig. 5, the domain bias could be reflected in terms of backgrounds, distinct lightings, seasons, resolutions, human races, etc. The model trained on CUHK03, Market1501, MSMT17 only achieves the 6.62%, 17.11% and 30.71% mAP when tested on Duke, which shows the severe domain bias.

TABLE I
Details of the Datasets in Our Experiments

Dataset	Images/IDs	Train	Test
Duke [45]	36,411/1,812	16,522/702	19,919/1,110
Market1501 [46]	32,688/1,501	12,936/751	19,752/750
MSMT17 [47]	126,441/4,101	32,621/1,041	93,820/3,060
CUHK03 [48]	28,192/1,467	26,264/1,367	1,928/100

Fig. 5. Illustration of the domain gaps between person ReID datasets.

In all experiments, the source models are trained on MSMT17, CUHK03 and Market1501 datasets, and the target model is trained and tested on Duke dataset. During model reuse stage, the parameters of source models are fixed and serve as feature extractors to generate feature representations of the training data.

Implementation Details: In this work, we adopt ResNet50 network [49] as our base network, both in the source and target models. It is widely acknowledged that in deep neural networks features in the lower layers are more general and in higher layers are more task-specific [50]. Therefore, we use the lower hidden layer representation Conv3 in source models for reusing. Both in source and target models, classification layer is implemented by fully connected layer and softmax loss (softmax function + cross-entropy loss). For the labeled data, the training follows the standard classification paradigm where each person identity is regarded as a unique class. The models are fine-tuned from ImageNet pretrained weights. The images are resized to $256 * 128$. Batch size and weight decay are set to 32 and $5e-4$, respectively. The total training for model reuse lasts for 300 epochs. Initial learning rate starts from $2 * 10^{-4}$, and is divided by 10 after 150 epochs.

Mean Average Precision: The mAP metric evaluates the overall performance for ReID. Average precision is calculated for each query image as follows,

$$AP = \frac{\sum_{k=1}^{n} P(k) \times gt(k)}{N_{gt}}, \tag{19}$$

where k is the rank of retrieved persons in the sequence, n is the number of retrieved persons, and N_{gt} is the number of relevant persons. $P(k)$ is the precision at cut-off k in the recall list and

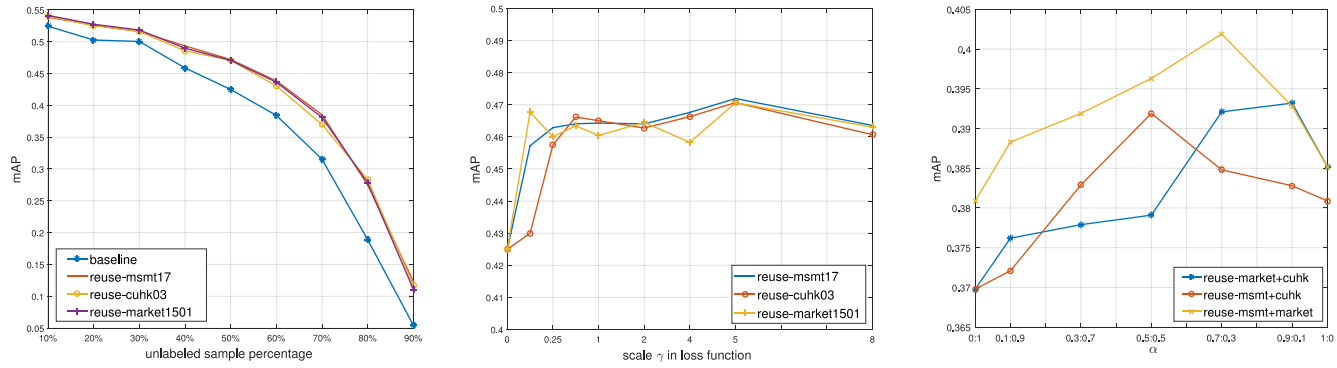

Fig. 6. (a) The performances of reusing different single models on Duke test set by varying the percentage of unlabeled data. (b) The performances by setting different scale γ in the loss function. (c) The performances by setting different importance factors α_m between reused models.

$gt(k)$ indicates whether the k-th recall image is correct or not. Therefore, the mAP is defined as follows,

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}, \tag{20}$$

where Q is the number of total query images. Moreover, Top 1 match rate is also reported in the experiments.

Comparison methods: We compare with the recent person ReID methods as follows,

PCB [9] is a part based convolutional feature representation scheme with refined part pooling method for person re-identification.

DefenseTriplet [10] is an improved triplet embedding mining scheme with classic triplet loss for person re-identification.

AlignedReID [51] uses a novel local feature alignment method to benefit global feature learning.

BagofTricks [52] is a competitive baseline in person ReID. For fair comparison, we do not adopt the reranking and random erasing augmentation mechanism.

Unlabeled Data Usage: The unlabeled data are involved in multi-model reuse, and we split the training set (Duke) into two parts, *i.e.*, the labeled and unlabeled. Regarding the unlabeled data, the label information of a part of training samples are not used in the experiments. To denote the different proportion of labeled and unlabeled data, we use 30% Duke to indicate only 30% samples are used as labeled samples and the rest serve as unlabeled samples during training. As such, the mixed training of labeled and unlabeled data in real-world scenarios can be well simulated in our experiments.

B. Results of Model Reuse

Single Model Reuse: In Fig. 6(a), the results of single model reuse are demonstrated. We provide the results of baseline model trained with the splitted labeled Duke training set. For model reuse, we use three different models trained on MSMT17, CUHK03 and Market1501 as source models. Compared with baseline models, it is clear that the introduction of single model reuse significantly boosts the performance of ReID models. In addition, we can find that with different settings on the percentage of unlabeled data in model reuse, the target model can consistently achieve better performance than baseline model. For example, with the 40% labeled data, reusing MSMT17 model can bring about 5.42% mAP (38.42% $\rightarrow$ 43.84%) performance gains.

TABLE II
THE mAP PERFORMANCE BY SETTING DIFFERENT SCALE γ WHEN REUSING TWO MODELS ON THE DUKE TEST SET. (70% SAMPLES IN DUKE TRAIN SET ARE USED AS UNLABELED SAMPLES)

Scale	Market+CUHK	MSMT+CUHK	MSMT+Market
γ=1	39.42	40.37	41.27
γ=2	39.81	40.23	**41.59**
γ=4	**40.57**	**40.82**	40.97
γ=5	39.25	39.42	39.64
γ=8	39.43	38.19	38.96
γ=16	38.26	38.17	38.49
baseline	MSMT only	CUHK only	Market only
31.53	38.52	36.98	38.09

The hyper-parameter analysis: The parameter γ is used to balance the empirical loss and regularization term for reuse in the training objective (Eq. 5). It is crucial to choose proper value of γ to improve the performance of model reuse. We vary the value of γ and show the performances in Fig. 6(b) to investigate the sensitivity of model w.r.t. γ. Across a wide range of γ from 2^{-3} to 2^3, the performance variation of model remains fairly stable. Moreover, in Table II, we show the performances of reusing two models on different scales. We can find that the overall performances under different scales are close.

Importance factors α_m between reused models: The α_m indicates the importance of the m_{th} model in model reuse. To explore the effect of α, we present the performance of two model reusing with different importance factors in Fig. 6(c). The peak point on the performance curves in Fig. 6(c) indicates that choosing proper importance factors between the different reused models can achieve maximum model utilization in performance improvements. However, as the number of reuse models increases, it becomes more difficult to assign proper parameters.

TABLE III
THE mAP PERFORMANCE WITH/WITHOUT ATTENTION BASED MODEL WEIGHTING ON THE DUKE TEST SET

Resume models	w/o attention	w/ attention
Market+CUHK	37.91	39.42
MSMT+CUHK	39.19	40.37
MSMT+Market	39.63	41.27
MSMT+CUHK+Market	40.59	**42.67**

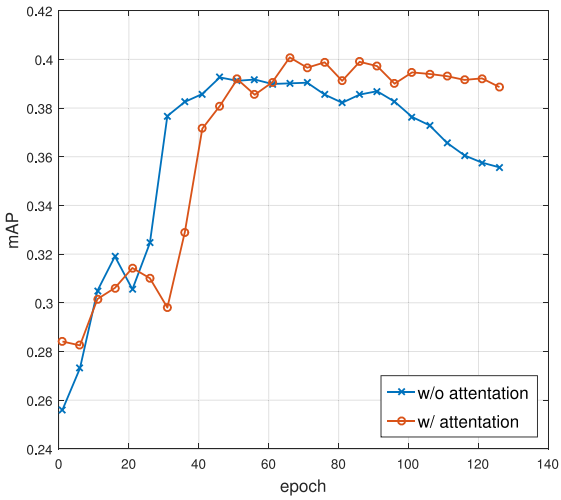

Fig. 7. The performance comparison between with and without attention block for every epoch during model reuse training.

Adaptive model weighting: In order to automatically set hyper-parameter α_m, we use attention block in reuse for adaptive model weighting. We conduct the adaptive and manual parameter setting experiments. The λ is set to 1, and α_m is set to 0.5 and 0.333 in two and three model reuse, respectively. As shown in Table III, the proposed method with attention block can achieve higher performance, which implies that attention mechanism can obtain more optimal hyper-parameters. Fig. 6(c) provides the performance under different parameter combinations. Compared with the optimal results under manual parameter setting, the adaptive weighting method still has significant advantages.

Using the adaptive model weighting mechanism brings some unexpected merits. (1) It can reduce the potential of over-fitting. Fig. 7 shows the performance variations during the training stage, which illustrates that the method with attention block can obtain more stable performance gains, while the performance without attention drops from 39.19% (at 70th epoch) to 35.56% (at 130th epoch). (2) The attention block can improve the efficiency and effectiveness of model reuse mechanism. As shown in Fig. 8, the introduce of attention block makes the learning converge faster with lower loss (with attention 0.02 vs. without attention 0.04).

Multi-model reuse: We provide the results of multi-model reuse in Table IV. Compared to the baseline models, we can find that reusing additional models achieves better performance both in mAP and Rank 1. By increasing the number of reused models, the incremental performance gains can be consistently achieved. With three reused models, we can achieve 42.67% mAP, which significantly outperforms the baseline 31.5% mAP. We also compare with the recent state-of-the-art methods such as BagofTricks [52], which has much superior performance over baseline softmax method. With incremental multi-model reuse, our baseline model can be significantly improved. When reusing three additional models, we can outperform BagofTricks [52] by 4% mAP without any complicated feature representation or loss functions. Such performance improvements can adequately prove the effectiveness of the proposed reuse strategy.

We also visualize the retrieval results comparison in Fig. 9, which lists Top 15 retrieval results of softmax baseline and three model reuse with (MSMT17 + Market1501 + CUHK03). In both methods, the recalled images at the top position show similar attributes with query images, such as color. Moreover, the false positives at high position in model reuse method often present

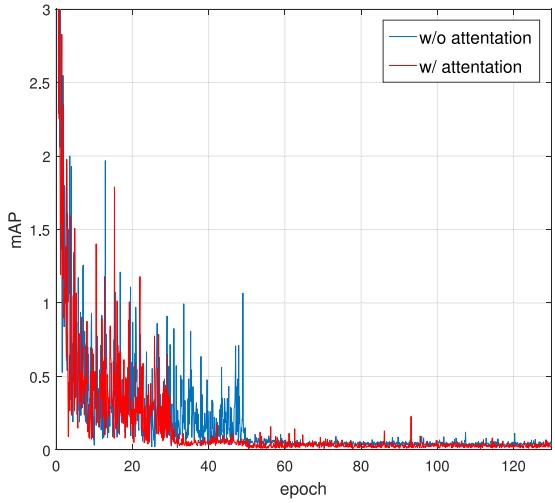

Fig. 8. The variation of loss during the training stage with and without attention block reusing.

TABLE IV
THE PERFORMANCE COMPARISON AND THE INCREMENTAL GAIN BY INCREASING REUSED MODELS ON THE DUKE TEST SET. (70% SAMPLES IN DUKE TRAIN SET ARE USED AS UNLABELED SAMPLES)

Model	mAP	Rank-1
Triplet [47]	34.31	54.30
PCB [9]	36.62	57.05
DefenseTriplet [10]	35.96	55.97
BagofTricks w/o tricks [52]	36.60	57.58
BagofTricks w/ tricks [52]	38.86	59.49
AlignedReID [51]	35.35	55.38
AlignedReID+Mutual Learning [51]	36.60	55.48
Softmax Baseline	31.53	49.55
+MSMT	38.52	58.61
+CUHK	36.98	58.34
+Market	38.09	57.40
+Market+CUHK	40.57	60.96
+MSMT+CUHK	40.82	60.87
+MSMT+Market	41.27	61.66
+MSMT+CUHK+Market	**42.67**	**62.17**

Fig. 9. Top 15 retrieval results on Duke test set. The images with red box are the wrong recall results. For each query, the first and second rows are results from softmax baseline and model reuse (MSMT + CUHK + Market).

TABLE V
THE mAP PERFORMANCE COMPARISON OVER DUKE TEST SET BY TRANSMITTING MODELS WITH DOM (FIRST LINE) AND WITHOUT DOM MODULE (SECOND LINE) IN TERMS OF DIFFERENT QUANTIZATION LEVELS

Model	Market+CUHK	MSMT+CUHK	MSMT+Market
original	40.57	40.82	41.27
compression bits=8	40.61	40.84	41.12
	40.57	40.82	41.06
compression bits=7	40.58	40.81	41.19
	40.54	40.69	40.79
compression bits=6	40.39	40.83	40.98
	38.56	38.49	39.35
compression bits=5	40.12	40.68	40.53
	35.61	34.11	34.05
compression bits=4	39.66	39.09	39.62
	8.73	12.31	8.69
compression bits=3	31.88	25.84	27.14
	0.07	1.23	0.10

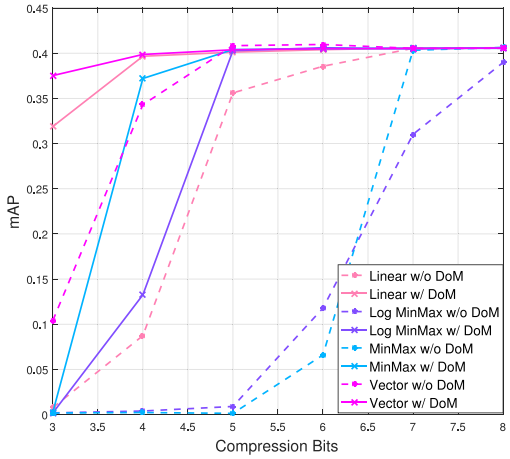

Fig. 10. The performance comparisons among different quantization methods.

similar backgrounds as query. Besides, reusing additional models can achieve more robust capability in analyzing low resolution images, which is important for person re-identification.

C. Results of Model Prediction

It is necessary to deploy the enhanced model by multi-model reuse to the front-end. In Table V, we show the results of DoM with the decrease of compression bits. It is worth noting that the k bits compression means taking k bits memory cost to represent each weight parameter in the model. The smaller compression bits indicates coarser quantization. We use models of different performances as existing models and to-be-deployed models to simulate model prediction. For model reusing with Market1501 and CUHK03, the performance of existing model is 38.76% mAP and the to-be-deployed model is 40.57% mAP. The model prediction is able to obtain better compression performance by using model sharing information. Given the same to-be-compressed models, better performance can be achieved with the DoM, say compression bits = 6 to 3. When compression bits = 3, the DoM strategy can well maintain the performance (40.57→31.88), while the single model compression strategy even collapses (40.57→0.07).

For better understanding of DoM, we plot the mAP changes in terms of different quantization methods with/without DoM in Fig. 10, *i.e.*, linear quantization, MinMax quantization, log MinMax quantization and vector quantization. The linear quantization and vector quantization can achieve higher performance compared with other methods. Besides, the DoM strategy significantly outperforms the simple single model compression scheme under all quantization configurations. The reason is that DoM compresses the differences of the models while the single model compression is applied to the whole model. Thus, the DoM is more suitable for delivering incremental information under constrained transmission environments.

VI. CONCLUSION

In this work, we explore the model-centric paradigm with model generation, utilization and communication to better shape

the digital retina system and support front-end smart visual sensing in smart cities. Within such paradigm, we propose a multi-model reuse scheme for better model utilization in digital retina. To economically transmit the updated model to the front-end, a model prediction method based on Difference of Model (DoM) is proposed. The challenging task of person ReID is utilized to illustrate the effectiveness of the paradigm in the context of model domain bias and model transmission. In the future, we will systematically integrate model generation, utilization, communication and standardization to build an intelligent, economic and efficient digital retina in smart cities.

REFERENCES

[1] H. Wässle, "Parallel processing in the mammalian retina," *Nature Rev. Neuroscience* vol. 5, Oct. 2004, Art. no. 747.

[2] L. Bao *et al.*, "Artificial shape perception retina network based on tunable memristive neurons," *Scientif. Reports*, vol. 8, Sep. 2018, Art. no. 13727.

[3] W. Gao and Y. Tian, "Digital retina: Revolutionizing camera systems for the smart city," *Sci. China Inf. Sci.*, vol. 48, no. 8, pp. 1076–1082, 2018.

[4] G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[5] L. Yu, S. Chen, and J. Wang, "Overview of AVS-video coding standards," *Signal Process.: Image Commun.*, vol. 24, no. 4, pp. 247–262, 2009.

[6] L.-Y. Duan *et al.*, "Overview of the MPEG-CDVS standard," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 179–194, Jan. 2016.

[7] L.-Y. Duan *et al.*, "Compact descriptors for video analysis: The emerging MPEG standard," *IEEE MultiMedia*, vol. 26, no. 2, pp. 44–54, Apr.-Jun. 2018.

[8] Y. Lou *et al.*, "Compact deep invariant descriptors for video retrieval," in *Proc. IEEE Data Compression Conf.*, Apr. 2017, pp. 420–429.

[9] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 480–496.

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

[11] Y. Lou, Y. Bai, J. Liu, S. Wang, and L. Duan, "VERI-Wild: A large dataset and a new method for vehicle re-identification in the wild," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 3235–3243.

[12] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2167–2175.

[13] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan, "Embedding adversarial learning for vehicle re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 3794–3807, Aug. 2019.

[14] Y. Bai *et al.*, "Group-sensitive triplet embedding for vehicle re-identification," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2385–2399, Sep. 2018.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 815–823.

[16] B. Bhattarai, G. Sharma, and F. Jurie, "CP-mtML: Coupled projection multi-task metric learning for large scale face retrieval," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4226–4235.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[18] L.-Y. Duan *et al.*, "Compact descriptors for video analysis: The emerging MPEG standard," *IEEE MultiMedia*, vol. 26, no. 2, pp. 44–54, Apr.-Jun. 2019.

[19] L. Duan, Y. Lou, S. Wang, W. Gao, and Y. Rui, "AI-Oriented large-scale video management for smart city: Technologies, standards and beyond," *IEEE MultiMedia*, vol. 26, no. 2, pp. 8–20, Apr.–Jun. 2019.

[20] R. Woodworth and E. Thorndike, "The influence of improvement in one mental function upon the efficiency of other functions," *Psychol. Rev.*, vol. 8, 1901, Art. no. 247.

[21] N. R. Council, *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. Washington, DC, USA: National Academies Press, 2000.

[22] P. Sterling and S. Laughlin, *Principles of Neural Design*. Cambridge, MA, USA: MIT Press, 2015.

[23] Y. Lou *et al.*, "Towards digital retina in smart cities: A model generation, utilization and communication paradigm," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 19–24.

[24] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand, "Domain-adversarial neural networks," 2014, *arXiv:1412.4446*.

[25] M. Ghifary, W. B. Kleijn, and M. Zhang, "Domain adaptive neural networks for object recognition," in *Proc. Pacific Rim Int. Conf. Artif. Intell.* Springer, 2014, pp. 898–904.

[26] Y. Yang, D.-C. Zhan, Y. Fan, Y. Jiang, and Z.-H. Zhou, "Deep learning for fixed model reuse," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 2831–2837.

[27] X.-Z. Wu, S. Liu, and Z.-H. Zhou, "Heterogeneous model reuse via optimizing multiparty multiclass margin," in *Proc. 36th Int. Conf. Mach. Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun. 2019, pp. 6840–6849.

[28] Y. Y. D.-C. Z. Xiang and Y. G. Y. Jiang, "Modal consistency based pre-trained multi-model reuse," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 3287–3293.

[29] R. Jha, A. Marin, S. Shivaprasad, and I. Zitouni, "Bag of experts architectures for model reuse in conversational language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., Vol. 3 (Ind. Papers)*, 2018, pp. 153–161.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[31] M. Denil, B. Shakibi, L. Dinh, and N. De Freitas, "Predicting parameters in deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2148–2156.

[32] J. Wu *et al.*, "Deep k-means: Re-training and parameter sharing with harder cluster assignments for compressing deep convolutions," *Proc. 35th Int. Conf. Mach. Learn.*, 2018, vol. 80, pp. 5363–5372.

[33] G. Huang, S. Liu, L. Van der Maaten, and K. Q. Weinberger, "CondenseNet: An efficient densenet using learned group convolutions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 2752–2761.

[34] X. Zhang, J. Zou, X. Ming, K. He, and J. Sun, "Efficient and accurate approximations of nonlinear convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 1984–1992.

[35] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Speeding up convolutional neural networks with low rank expansions," in *Proc. Brit. Mach. Vision Conf.*, 2014, pp. 1–13.

[36] J.-H. Luo, J. Wu, and W. Lin, "ThiNet: A filter level pruning method for deep neural network compression," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5058–5066.

[37] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 1389–1397.

[38] W. Chen, J. Wilson, S. Tyree, K. Weinberger, and Y. Chen, "Compressing neural networks with the hashing trick," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2285–2294.

[39] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," in *Proc. 4th Int. Conf. Learn. Representations*, San Juan, Puerto Rico, May 2–4, 2015, pp. 1–14.

[40] D. Wu *et al.*, "Vision and challenges for knowledge centric networking (KCN)," *IEEE Wireless Commun.*, vol. 26, no. 4, pp. 117–123, Aug. 2019.

[41] C. Westphal, "Method for network coding packets in content-centric networking based networks," U.S. Patent 9002921, Apr. 7, 2015.

[42] Z. Chen, S. Wang, D. O. Wu, T. Huang, and L.-Y. Duan, "From data to knowledge: Deep learning model compression, transmission and communication," in *Proc. ACM Multimedia Conf. Multimedia Conf*. ACM, 2018, pp. 1625–1633.

[43] B. Long, P. S. Yu, and Z. Zhang, "A general model for multiple view unsupervised learning," in *Proc. SIAM Int. Conf. Data Mining*. SIAM, 2008, pp. 822–833.

[44] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[45] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 3754–3762.

[46] L. Zheng *et al.*, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 1116–1124.

[47] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 79–88.

[48] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 152–159.
[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
[50] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, vol. 37, pp. 97–105.
[51] X. Zhang *et al.*, "AlignedReID: Surpassing human-level performance in person re-identification," 2017, *arXiv:1711.08184*.
[52] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, Jun. 2019, pp. 4321–4329.

Yihang Lou received the B.S. degree in software engineering from the Dalian University of Technology, Liaoning, China, in 2015. He is currently working toward the Ph.D. degree with the School of Electrical Engineering and Computer Science, Peking University, Beijing, China. His current interests include large-scale image retrieval and video content analysis.

Ling-Yu Duan (Member, IEEE) received the Ph.D. degree in information technology from The University of Newcastle, Callaghan, NSW, Australia, in 2008. He is a Full Professor with the National Engineering Laboratory of Video Technology (NELVT), the School of Electronics Engineering and Computer Science, Peking University (PKU), Beijing, China, and has served as the Associate Director of the Rapid-Rich Object Search Laboratory (ROSE), a joint lab between Nanyang Technological University (NTU), Singapore, and Peking University (PKU), China, since 2012. He has also been with the Peng Cheng Laboratory, Shenzhen, China, since 2019. His research interests include multimedia indexing, search and retrieval, mobile visual search, visual feature coding, and video analytics, etc. He has authored and coauthored about 200 research papers. He is currently an Associate Editor of the *ACM Transactions on Intelligent Systems and Technology* and *ACM Transactions on Multimedia Computing, Communications, and Applications*, and serves as the area chairs of ACM MM and IEEE ICME. He received the IEEE ICME Best Paper Award in 2019, the IEEE VCIP Best Paper Award in 2019, and EURASIP Journal on Image and Video Processing Best Paper Award in 2015, the Ministry of Education Technology Invention Award (First Prize) in 2016, the National Technology Invention Award (Second Prize) in 2017, China Patent Award for Excellence (2017), and the National Information Technology Standardization Technical Committee "Standardization Work Outstanding Person" Award in 2015. He was a Co-Editor of MPEG Compact Descriptor for Visual Search (CDVS) standard (ISO/IEC 15938-13) and MPEG Compact Descriptor for Video Analytics (CDVA) standard (ISO/IEC 15938-15). He is a member of the MSA Technical Committee in IEEE-CAS Society.

Yong Luo received the B.E. degree in computer science from the Northwestern Polytechnical University, Xi'an, China, in 2009, and the D.Sc. degree from the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, in 2014. He is currently a Research Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests are primarily on machine learning and data mining with applications to visual information understanding and analysis. He has authored or co-authored more than 30 papers in top journals and prestigious conferences including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MULTIMEDIA, IJCAI, AAAI, CIKM, ICDM, and ICME. He received the IEEE Globecom 2016, IEEE ICME 2019, and IEEE VCIP 2019 Best Paper Awards, and was nominated for the IJCAI 2017 Distinguished Best Paper Award.

Ziqian Chen received the B.S. degree in computer science from the Dalian University of Technology, Liaoning, China, in 2017. He is currently working toward the master's degree with the School of Electrical Engineering and Computer Science, Peking University, Beijing, China. His current interests include image retrieval and deep learning model compression.

Tongliang Liu is a Lecturer with the School of Computer Science, The University of Sydney, Camperdown, NSW, Australia. His research interests include machine learning and computer vision. He has authored and co-authored 60+ research papers including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, ICML, NeurIPS, CVPR, ECCV, AAAI, IJCAI, KDD, and ICME, with Best Paper Awards, e.g., the 2019 ICME Best Paper Award. He is a recipient of the Discovery Early Career Researcher Award (DECRA) from the Australian Research Council (ARC) and was shortlisted for the J G Russell Award by the Australian Academy of Science (AAS) in 2019.

Shiqi Wang (Member, IEEE) received the B.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2008, and the Ph.D. degree in computer application technology from the Peking University, Beijing, China, in 2014. From March 2014 to March 2016, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. From April 2016 to April 2017, he was with the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore, as a Research Fellow. He is currently an Assistant Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong. He has proposed more than 40 technical proposals to ISO/MPEG, ITU-T, and AVS standards. His research interests include video compression, image/video quality assessment, and image/video search and analysis.

Wen Gao (Fellow, IEEE) received the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991. He is currently a Professor of computer science with the School of Electronic Engineering and Computer Science, Institute of Digital Media, Peking University, Beijing, China. He has also been the Director of the Peng Cheng Laboratory, Shenzhen, China, since 2018. Before joining Peking University, he was a Professor of computer science with the Harbin Institute of Technology, Harbin, China, from 1991 to 1995, and a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He has authored extensively including five books and more than 600 technical articles in refereed journals and conference proceedings in the areas of image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interfaces, and bioinformatics. He is a member of the China Engineering Academy. He has been the Chair of a number of prestigious international conferences on multimedia and video signal processing, such as the IEEE International Conference on Multimedia and Expo and ACM Multimedia, and served on the Advisory and Technical Committees of numerous professional organizations. He served or serves on the Editorial Board of several journals, such as the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON AUTONOMOUSMENTAL DEVELOPMENT, the *EURASIP Journal of Image Communications*, the *Journal of Visual Communication*, and *Image Representation*.