Front-End Smart Visual Sensing and Back-End Intelligent Analysis: A Unified Infrastructure for Economizing the Visual System of City Brain

Yihang Lou, *Student Member, IEEE*, Ling-Yu Duan, *Member, IEEE*, Shiqi Wang, *Member, IEEE*, Ziqian Chen, *Student Member, IEEE*, Yan Bai, *Student Member, IEEE*, Changwen Chen, *Fellow, IEEE*, and Wen Gao, *Fellow, IEEE*

***Abstract*— The visual data, which are acquired from the ubiquitous visual sensors deployed in metropolitans, are of great value and paramount significance to enhance the effectiveness and pursue the future development of smart cities. In this paper, the essential building blocks of the unified visual data management and analysis infrastructure that serve as the foundation for the economical visual system in the city brain, are introduced to facilitate the utilization of the visual signal in the artificial intelligence era. In particular, we start by the discussion of the front-end smart visual sensing in the context of economical communication and service with the heterogeneous network, and the functionalities and necessities of compact visual feature and deep learning model representations are detailed. Subsequently, the utilities of the infrastructure are demonstrated through two intelligent applications at the back-end, including vehicle re-identification and person re-identification. The standardizations regarding compact feature and deep neural network representations, which are regarded as the key ingredients in this infrastructure and greatly facilitate the construction of the visual system in the city brain, are also discussed. Finally, we envision how the potential issues regarding the economical visual communications for future smart cities might be pragmatically approached within this unified infrastructure.**

Manuscript received October 13, 2018; revised March 15, 2019; accepted April 21, 2019. Date of publication May 13, 2019; date of current version June 17, 2019. This work was supported in part by the National Basic Research Program of China under Grant 2015CB351806, in part by the National Natural Science Foundation of China under Grant U1611461 and Grant 61661146005, in part by the Shenzhen Municipal Science and Technology Program under Grant JCYJ20170818141146428, and in part by the Hong Kong RGC Early Career Scheme under Grant 9048122 (CityU 21211018). *(Corresponding author: Ling-Yu Duan.)*

Y. Lou, Z. Chen, and Y. Bai are with the Institute of Digital Media, Peking University, Beijing 100871, China (e-mail: yihanglou@pku.edu.cn; wzziqian@pku.edu.cn).

L.-Y. Duan and W. Gao are with the Institute of Digital Media, Peking University, Beijing 100871, China, also with the School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, Shenzhen 518055, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: lingyu@pku.edu.cn; wgao@pku.edu.cn).

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

C. Chen is with the Computer Science and Engineering Department, University at Buffalo, The State University of New York, Buffalo, NY 14228 USA, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: chencw@buffalo.edu).

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

Digital Object Identifier 10.1109/JSAC.2019.2916488

***Index Terms*— Smart city, visual analysis, compact feature representation, vehicle re-identification, person re-identification, standardization.**

I. Introduction

RECENT years have witnessed the dramatically increased demand for the utilization of the visual data acquired in the ambient environments to facilitate the construction of the smart city, including city security, management and planning. The acquisition, managing and analysis of the video big data essentially constitute the digital retina of the smart city and revolutionize the city brain [1], [2]. The recently developed artificial intelligence technologies based on deep learning have greatly advanced the image/video analytics services from the fundamental scientific and technical perspectives. To fully capitalize on these advances in such an environment of extreme growth data and constrained resources such as bandwidths, there is a considerable concern regarding how the economical visual data management and efficient analysis can be reached. In particular, surveillance videos in smart city have already become the biggest big data [3], [4], such that aggregating and processing large-scale surveillance video stream data in real-time are very challenging tasks. Millions of surveillance cameras are deployed in urban areas form large scale camera networks, and without the efficient management and utilization of the visual data, it is impossible to monitor the current dynamics and optimize the resource allocation in the long run.

The traditional cloud computing paradigm requires the video data captured at the front-end entirely compressed and transmitted to the back-end for further processing and analyses. In practice, however, assembling and transmitting thousands-of-thousands video streams simultaneously for real-time analyses are unaffordable, such that the gap between the large scale visual data and constrained resources becomes the bottleneck. Therefore, prevailing attitudes are shifted towards advanced economical techniques for the better utilization of the big visual data in the artificial intelligence (AI) age. In particular, the developments of AI hardware architecture have made the intelligent processing equipped at the front-end more promising and realistic. The essential difference between the surveillance video and traditional broadcasting

video is that ultimate receivers of most surveillance videos are machines instead of humans, and this has made the compact representation of visual data with smart sensing at the front-end feasible and practical.

The purpose of this article is to investigate the economical visual data management in smart city applications based on the infrastructure of front-end smart visual sensing and back-end intelligence. We start by brief discussions of the necessities of such infrastructure, followed by the descriptions on compact features and deep model representation, which constitute the essential building blocks and achieve efficient visual information extraction with feasible model adaptation at the intelligent front-end. The applications of this infrastructure are demonstrated in two main scenarios, *i.e.*, vehicle and person re-identification. To enable the interoperability, the standardization of compact feature representation and deep feature representation are further discussed, where we perceive both great promises and major challenges. Finally, we will provide a vision on the unresolved problems left for future studies.

The contributions of this paper are summarized as follows,

- We investigate the infrastructure with front-end smart sensing and back-end analyses, introduce the essential building blocks and analyze the functionalities to facilitate the economical utilization of the visual data for smart city applications.
- We provide an overview of the roles of the standards in smart city applications, including compact feature representation and deep neural network compression. The new challenges arising from the standardization process in real-application scenarios are also perceived.
- We clarify the future issues confronted and how they might be pragmatically addressed, in an attempt to stimulate fruitful thoughts regarding economical data management and communication in smart cities.

The remainder of this paper is organized as following. The unified structure and application system are presented in Section II. Section III details the front-end visual sensing for economical visual information communication, and the applications of back-end intelligent analyses are introduced in Section IV. Section V discusses the standardizations for economical communication. Section VI envisions the future trend and Section VII concludes the paper.

II. The Unified Visual Data Management and Analysis Infrastructure

In this section, we will describe the unified visual data management and analysis infrastructure, including the motivations, advantages, as well as methodologies. Moreover, the application of this infrastructure in real world scenarios is demonstrated, further providing evidence for economical visual information management in smart cities.

A. The Unified Infrastructure With Front-End Smart Sensing and Back-End Analysis

To begin with, we analyze the design philosophy behind the unified visual data management and analysis infrastructure from multiple perspectives. First, the recent developments of deep learning have greatly facilitated the automatic visual analysis, such that for many smart city applications the ultimate receiver of the visual signal is the machine instead of the human visual system. Second, equipping the smart sensing capabilities such as feature extraction at the front-end camera is feasible, and recently numerous methods emerge aiming to reduce the computational cost for deep learning based feature extraction at the front-end. Third, the feature data are generally much more compact than the texture data, such that transmitting feature is undoubtedly economical. Fourth, the texture decoding given aggregated bit streams and the subsequent feature extraction at the back-end impose high computational cost, which also necessitates the shift of the computations to the front-ends. In Table I, we present an example of transmission cost comparison between compact feature and visual signals in a real surveillance system. These motivations have significantly raised expectations regarding the efficient delivery of the visual data.

TABLE I

The Cost Comparison Between Transmitting Compact Features and Visual Signals in a Real Surveillance System (e.g., 100,000 Cameras)

	Compact Feature (CDVS)	Visual Signals
Size of an Image	512B $\sim$ 16KB	$\sim$ 2MB
Bandwidth	100,000*32Kbps =3.2Gbps	100,000*4Mbps = 400Gbps
Storage in a day	3.2Gbps*3600*24/8=34TB	400Gbps*3600*24/8 = 4218TB

The unified video data management infrastructure is shown in Fig. 1, from which it is obvious that the data transmitted in the network linking the front-end and back-end could be compact representations of features and deep learning models. This is a distinguished feature of this infrastructure as the traditional information transmitted could be the texture bitstream. More specifically, at the front-end, the features are extracted and compactly represented, and further transmitted upstream to the back-end for visual analysis. The scalability which is naturally supported in handcrafted feature representation greatly facilitates the economical utilization of the visual data. For example, there are six operating points from 512B to 16KB in CDVS [5], which provide flexible rate control to deal with the variant scenarios with different bandwidth limitations or fluctuations. Moreover, this infrastructure also supports the downstream updating of the deep learning model at the front-end, based on compact deep learning model representation in the scenario of model communication. At the back-end, many novel visual analysis applications can be supported given the received features, including detection, recognition, tracking and search. In [6], Choi and Bajic proposed a collaborative detection method, which uses the smart front-end devices to extract shallow features, then transmits these features to the back-end where a deeper network receives these features to obtain final detection results. This unified infrastructure is expected to facilitate efficient visual data management by largely reducing the redundancies within the visual signal, which has also been proved to be efficient and useful in many applications [7].

Generally speaking, both the handcrafted features and deep learning features can be represented in a compact way, as they

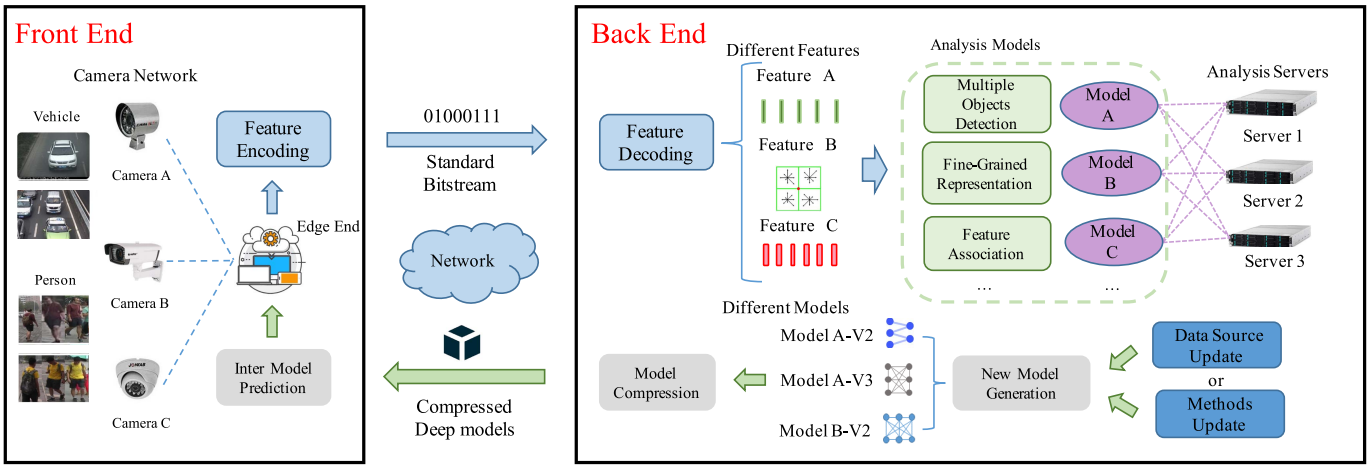


Fig. 1. Illustration of the unified front-end visual sensing and back-end intelligent analysis. Both feature and deep neural network are required to be frequently transmitted in this infrastructure. The front-end edge devices extract features for preliminary analysis and the encoded standardized feature bitstreams are transmitted to the back-end for further analysis. The standard bitstreams are the bistreams encoded in a unified bistream syntax. On the other hand, the deep learning models are trained in the back-end and transmitted to the front-end, and frequent updating may occur at the front-end.

summarize the characteristics of the visual content at the semantic level instead of the visual signal. From the compression point of view, this can be regarded as a different type of redundancy reduction in terms of modality transfer from the raw visual signal space to compact feature space. Moreover, the discrimination capability can be well maintained by compactly representing the visual feature, and the scalability of the feature representation also enables the adaptation of network bandwidth fluctuations towards friendly network transmission. In [5], the CDVS standard supports 6 different rate control for feature compression, which can better adapt to the network transmission, especially in wireless or mobile networks. With these amazing characteristics, the economical visual data management in the smart city scenario can be fullfilled. For handcrafted descriptor, the (CDVS) was standardized by MPEG for mobile visual search. Recently, towards the big video data, MPEG is working on the compact descriptor for video analysis (CDVA) standard, in which both the handcrafted and deep learning features are adopted.

To fully support the smart sensing capabilities at the front-end, recent methods seek to achieve the light-weight feature extraction with deep neural networks based on parameters pruning [8]–[10], matrix factorization [11]–[14], filter selection [15]–[20], quantization [21]–[24], distilling [25], [26], etc. In addition to these techniques for compact deep neural network representation, the deep learning model communication should be particularly supported with enhanced technologies in this scenario [27]. To this end, the redundancy among multiple deep learning models sequentially transmitted is exploited, such that the dynamic and economical updating of the deep neural network is enabled to incrementally improve the visual analysis performance at the back-end.

At the back-end, the visual features of existing images/videos are indexed to support the visual analysis and retrieval applications. Here, we demonstrate the framework based on vehicle and person re-identifications, which are two typical applications in smart cities. In general, the received features from the front-end are compared against the features at the back-end. Moreover, the dataset at the back-end can also be dynamically enlarged and updated, which also facilitate the model updating by distributing the re-trained models to front-end devices to improve the quality of the features.

B. Applications in Smart City

While the field of large-scale video management [28], [29] with front-end smart visual sensing and back-end intelligent analysis is still quickly evolving, it is interesting to discuss how these technologies could be made use of in real-world applications. A straightforward application is city security, which is of paramount importance to the smart city. Here we take the smart city blueprint in China as an example. It is envisioned that by 2020, the video surveillance infrastructure targeting at global coverage, network-wide sharing, full-time availability, and full controllable become mature and well-established. From 2016, the governments have increased their investments and striven to build safe and smart cities, which led to a large number of cameras equipped in various cities across the country and gave birth to a large amount of video data.

The efficiency and effectiveness requirements of security applications make smart security an absolute demand. The traditional strategy based on human viewing brings high cost of manpower and resources, and is certainly inefficient. In addition, the video data stored in each HD camera every month has reached the PB level, and it is difficult to perform simple playing back and retrieval. Thus, new challenges arise in both academic and industry realms, and intelligent security has become the sole way to address above problems.

One major purpose of intelligent security is to transform the unstructured visual data into structured feature representation that can be understood by computer vision algorithms. As such, the “mass video data” can be converted into “intelligent knowledge”, and the traditional requirement of “seeing clearly” is upgraded to “well understood”.

The unified infrastructure with front-end visual sensing and back-end intelligent analysis well supports such requirements, and with the breakthroughs of artificial intelligence technology, the desired visual target (pedestrian, vehicle, etc.) can be quickly and accurately identified in an economical way in the large quantity video data.

There is also an abundant menu regarding the real-world applications based on the front-end visual sensing and back-end analysis infrastructure, including the following,

- Together with the industrial partners, in the big data platform of Huangdao District, Qingdao, we have realized the application of pedestrian and vehicle re-identification with hundreds of video streams, to provide the warning services as well as the post investigations, which have greatly improved the efficiency of police services.
- The large-scale video management system in Guangdong Province China is established with the functionalities of cross-region video analysis, management and mining, facilitating real-time search and accurate identification based on the dataset of billion-level target objects.
- We established the real-time analysis system in a certain community, and the visual data generated by the public cameras are converged to the cloud for real-time analysis. Within less than one month after launching, the system successfully identifies a theft by a quick target search.

III. Front-End Visual Sensing Based on Compact Feature Representation

As described in Section II, the compact feature representation plays an important role in the front-end visual sensing. In principle, there is a vast and increasing proliferation of surveillance videos acquired, and correspondingly the extracted features are continuously transmitted as "feature stream" over networks. The feature compression aims to eliminate the redundancy for compact visual information representation and meanwhile maintain the discriminative capability. From the perspective of economizing the visual system of the city brain, the intelligence analysis, which is elegantly shifted from the back-end to the front-end, is able to significantly alleviate the computational cost on the server side. The smart front-end can perform preliminary feature extraction for the video content, then these features are transmitted to the back-end for further analysis. In the traditional visual system, the front-end only transmits the raw visual signals to the back-end, such that the workload of feature extraction and analysis are both completed at the back-end.

The upstream feature streams, which convey the visual information required for analysis, are more transmission friendly, enabling the massive video data handled simultaneously in a city scale. In principle, such paradigm can better support an economical visual data management system in which the compact features and deep learning models are able to communicate in an ecological closed loop. Towards visual search and video analysis in various application scenarios, we have intensively explored compact feature representation from both the handcrafted and deep learning feature. Furthermore, the techniques in MPEG regarding compact feature representation such as CDVS and CDVA standards are also discussed.

A. Handcrafted and Deep Features

The early visual analysis mainly relies on handcrafted features such as SIFT [30], SURF [31], ORB [32], which all own invariance properties in terms of scale and rotation to some extent. In essense, these features can be combined with vocabulary trees to achieve image retrieval task, which is meaningful to the surveillance data analysis. In the application of large scale retrieval, based on these traditional local features, the aggregated features such as the VLAD [33] and Fisher Vector [34] are proposed to improve the discriminative capability and achieve compact feature representation.

Recently, the CNN based features have also been widely used due to the powerful semantic information in representation. As a preliminary study, in [35], Babenko *et al.* proposed to use the output of fully connected layers of CNNs to generate feature representation, which outperformed canonical handcrafted feature such as SIFT [36]. In [37], Azizpour *et al.* proposed to use the max pooling of the intermediate output of CNNs (*e.g.*, the last pooling layer, named *pool5*) which can generate more effective representations for image retrieval. More recently, Tolias *et al.* [38] proposed Regional Maximum Activation of Convolutions (RMAC), which performs region-wise max pooling over a set of multi-scale regions in feature maps to generate final feature representation. In the context of front-end smart visual sensing, the deep models can be upgraded in the cloud and deployed downstream to the front-end, which is to support the progressive update of state-of-the-art CNN based feature extractors.

B. Compact Feature Representation

The compact feature representation is also a major concern in standardization working groups, such as the completed CDVS standard in MPEG which targets for visual search. In CDVS, both handcrafted local and global descriptors are standardized with scalable descriptor lengths (from 512B to 16KB). In particular, the local descriptors convey the invariant characteristics of local patches, and the global descriptors represent the aggregated statistics of the whole image. The handcrafted feature achieved great success in CDVS standard [5], and the exploration of deep learning features has posed open issues in the ongoing CDVA standard. The emerging video analysis standard CDVA which achieves state-of-the-art video analysis performances with compact descriptors, adopts both handcrafted features and deep learning features in an integrated framework. For deep learning features, the adopted NIP descriptors are generated via three stage nested pooling operations from the intermediate outputs in a VGG16 network.

Compact feature representation is of great significance to the economical feature communication, which substantially reduces the size of raw features, such that thousands upon thousands of bitstreams can be assembled and transmitted simultaneously for efficient analysis, especially in real-time smart city application scenarios. For example, for the MPEG

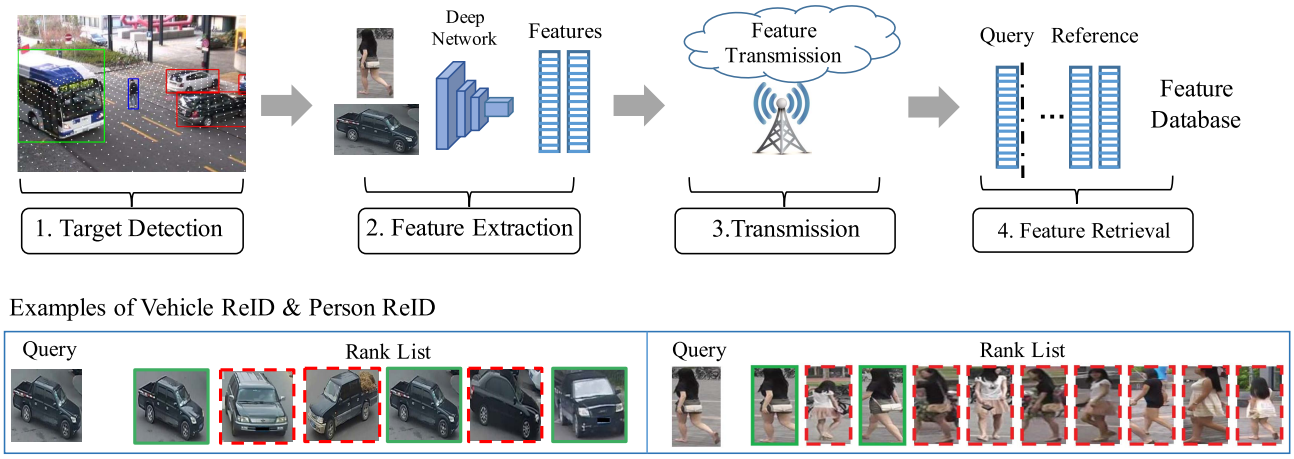


Fig. 2. Illustration of the person ReID system, which contains 4 distinct stages. The person targets are first detected from input video streams, then the corresponding feature representation is extracted from deep networks. After receiving the features, the matching operation is performed with the reference features in the database to generate the final rank list.

CDVS standard, the compression ratio between the raw features and compact feature representations can achieve up to 25. The MPEG CDVA standard, which inherits the hand-crafted features of CDVS [5], provides the promising basis for intelligent video applications. The proliferation of such economical services has greatly pushed forward the horizon of the compression paradigm. Moreover, scalability is naturally supported in feature representation, as a subset of features can also capture the image characteristics. For example, in MPEG CDVS standard, six operating points (from 512B to 16KB) are provided. In CDVA, there are also three operating points (16KB, 64KB and 256KB) defined [7]. As such, efficient and economical feature communication between the front-end and back-end can be naturally supported even when there are fluctuations in network dynamics. Furthermore, the recent research [39] also explored to compress the video deep features from the aspect of inter-frame deep feature prediction.

In addition, from the perspective of minimizing the overall energy and/or latency of the system, the widely use of deep learning features motivates some methods [6], [40] to explore the collaborative paradigm for efficient deployment of deep neural networks across the mobile cloud infrastructure, which divides the network between the mobile and the cloud to distribute the computational workload such that a more scalable solution for front-end smart sensing can be expected. In [40], an effective near-lossless deep feature compressor was proposed. In [6], the impacts of both near-lossless and lossy compression of feature data are also investigated. With the collaborative strategy, the transmission cost can be reduced by 70% without performance drop.

IV. Back-End Intelligent Analysis

The back-end server is responsible for training deep models for delivering and performing analyses on the received features. In this section, we first briefly introduce the role of back-end in the paradigm for model training, deploying and updating. Based on such paradigm, we then introduce the two representative visual analysis tasks, *i.e.*, vehicle and person ReID, which is to search the target vehicle or person images from a large-scale dataset with a given query vehicle or person image (please refer to Fig. 2). In particular, our works as well as the state-of-the-art methods on ReID will be presented and analyzed. Subsequently, we discuss the significance of these two tasks and challenges in the application of smart city. Finally, based on the promising performance improvements from the progressively developed deep models, the economical towards model deployment and updating are discussed and evaluated along with the deep learning model compression approaches by exploiting inter model redundancy.

A. Back-End Model Training, Deploying and Updating

The deep learning feature is data-driven, implying that the discriminative capability of features can be incrementally improved with more available training data and better model structure. In particular, the visual analysis tasks of ReID have drawn numerous attentions towards efficiently utilizing different versions or different types of deep neural networks. The widely used deep models derives a new problem, *i.e.*, how to efficiently deploy and update the deep learning models from the back-end server to the front-end devices.

Considering the limited computation and storage capability in the front-end, the back-end is expected to deliver the light-weight deep models with high performance. In such scenarios, the deep network compression is also a crucial technique for producing compact deep models. Moreover, in such paradigm, the model updating may dominate the form of model transmission, since the new models will be deployed to replace the previous models in front-end from time to time. The average updating interval varies from several days to several months, and also depends on the specific tasks on the front-end devices. When the better models are achieved by the improved optimization methods or more available data on the back-end side, the previous models will be updated on the front-end devices. In the updating procedure, both the inter-model prediction as well as the model reusing should

be paid attention to, in an effort to reduce the redundancy in the model transmission. Correspondingly, towards the model training, deploying and updating paradigm on vehicle and person ReIDs, we will specifically discuss the two application scenarios, performance as well as the insights on economical for model deployment and updating.

B. Vehicle Re-Identification

Recently, vehicle ReID [3], [41] are highly concerned in surveillance videos for the public security. Several methods [42], [43] attempt to search vehicles with the vehicle attributes such as color, vehicle models and spatial-temporal characteristics. Not only the visual appearance characteristics, but also the spatial-temporal information of vehicle tracks on the camera networks were used by Liu *et al.* [42] to design a coarse-to-fine search strategy for vehicle ReID. Shen *et al.* [44] also used a two-stage framework that incorporates complex spatial-temporal information for effectively regularizing the ReID results. To derive robust feature representation, promising research efforts have been devoted to deep metric learning. Significant progress has been made as reviewed below, which shows the necessity of updating deep models in the proposed infrastructure of front-end sensing and back-end intelligent analysis.

Recent methods [45]–[47] tend to utilize deep metric networks (*i.e.*, siamese or triplet network) to learn an embedding space where "the samples of the same vehicle ID are closer than those of the different". The vehicle images are mapped into such feature embedding space by the deep networks, and the feature distance between the vehicle samples can well represent their visual similarities. Liu *et al.* [45] introduced a coupled cluster loss to simultaneously use vehicle model and ID information to optimize a deep embedding model. Bai *et al.* [3] proposed a group sensitive triplet embedding model, which focuses on alleviating negative impacts of intra-class variances by incorporating group-wise variances into a structured triplet embedding model.

In particular, there are some more efforts aiming to further improve the ReID performance, such as mining hard examples for more efficient training [48] and strengthening the cross-view feature representation [49], [50]. Undoubtedly, the deep learning models have been improved in a variety of ways, which is expected to generate robust feature representation. Yuan *et al.* [48] proposed a hard-aware cascaded embedding method that ensembles a series of models via a cascaded way to select hard examples from the training set for efficient training. Zhou and Shao [49] focused on multi-view feature representation, and proposed a viewpoint-aware attention model to select core regions at different viewpoints for more discriminative feature representation. Analogously, Zhou *et al.* [50] proposed a long short-term memory (LSTM) network to model appearance transformations across different views of vehicles for generating cross-view feature representation.

C. Person Re-Identification

Compared with vehicle ReID, the challenge in person ReID is that the person targets are of small scale and have more pose variations. The current works regarding person ReID can be divided into two branches: learning a discriminative feature representation and learning a discriminative embedding space. Cheng *et al.* [51] proposed a multi-channel parts fusion strategy as well as an improved triplet loss to learn the discriminative feature representation. Su *et al.* [52] proposed a semi-supervised attribute learning scheme to learn binary attribute features. In [53], Zheng *et al.* used both identification and verification model to learn the feature representation. In [54], Su *et al.* used a pose-driven feature representation scheme where a sub-network first estimates a pose map which is then used to crop the localized body parts. The local and global person representations are then fused. Although great research efforts have been made, the person ReID performance gap in real surveillance scenarios is still challenging the deep feature representation learning.

Likewise, the feature embedding has been widely adopted to learn the discriminative distance metrics for person images. Yi *et al.* [55] proposed a siamese network for optimizing the feature distances between person image pairs, and several part regions are divided and trained separately. Ahmed *et al.* [56] designed a novel layer to capture local relationships between the image pair. In [57], Zhang *et al.* proposed a local deep descriptor alignment strategy to promote better similarity matching between global descriptors. In [58], a comparative attention network was designed to adaptively compute image similarity. To further advance the embedding learning, chen *et al.* [59] extended the pairwise similarity criterion to the quadruplet.

Moreover, in a broader sense, the domain adaption is also the recent research focus to further facilitate person ReID. Since the domain gap will reduce the generalization capability of ReID methods, the aim of domain adaption is to minimize the distribution gap between the source and target domains. Wei *et al.* [4] proposed a PT-GAN to bridge the domain gap by transferring person images from source domain to target domain and meanwhile it maintains the foreground of person images unchanged during transferring. Deng *et al.* [60] used Cycle-GAN to reduce domain bias via an extra similarity constraint to change styles during person image transferring.

D. Our Methods and Discussions

In this subsection, we present our effective model training for ReID tasks in the proposed infrastructure of front-end sensing and back-end intelligent analysis, including adversarial learning based EmbeddingGAN and group sensitive triplet embedding (GSTE). We then present the performance comparison with the state-of-the-art methods on benchmarks and discuss the open issues in the current ReID fields.

1) Embedding Adversarial Learning: Embedding learning has been widely used in ReID, within which the feature distances between the vehicles are optimized for discrimination. In the learning stage, selecting hard negatives for efficient training can promote the model discriminative capability. Here, we proposed an embedding adversarial learning scheme to actively generate more hard negatives to further facilitate embedding learning. Accordingly, the resulting strengthened model can be updated to the front-end.

The triplet network [61] is used for embedding learning. Let $< x, x^p, x^n >$ denote a triplet unit, where x is an anchor sample, x^p is a positive sample with the same ID as x, while x^n is a negative sample with different IDs. The triplet constraints can be formulated as $d(x, x^p) + \alpha \leq d(x, x^n)$, where α is the minimum margin between the positive and negative. The triplet unit that breaks the margin constraint is regarded as a hard triplet unit, and the x^n is a hard negative for x. We design a novel adversarial scheme on feature distance for generating hard negatives. For the generator, the generated samples are constrained to be located at a specific close region to input x, which can be formulated as follows:

$$\beta_1 \leq d(D(x), D(G(x))) \leq d(D(x^n), \quad D(G(x))) - \beta_2, \tag{1}$$

where the β_1 and β_2 specify the desired regions. D and G is the discriminator and generator in GAN. Therefore, the loss for the generator can be formulated as:

$$\begin{aligned} L_{G\_emb} = \mathbb{E}_x[\max\{\beta_1 - d(D(x), D(G(x))), 0\}] \\ + \mathbb{E}_x[\max\{d(D(x), D(G(x))) \\ - d(D(x^n), D(G(x))) + \beta_2, 0\}]. \end{aligned} \tag{2}$$

Correspondingly, the discriminator tries to push the $G(x)$ away from x to enlarge their feature distances. Hence, the loss for the discriminator can be represented as:

$$\begin{aligned} L_{D\_emb} = \mathbb{E}_x[\max\{d(D(x), D(x^p)) + \alpha \\ - d(D(x), D(G(x))), 0\}]. \end{aligned} \tag{3}$$

The generated hard negatives and real negatives are alternatively selected for optimization. The entire model can be trained end-to-end via adversarial learning. As the training continues, the generated hard negative would be more difficult to differentiate in terms of visual appearance and feature distance, and accordingly the discriminator becomes more discriminative. In other words, the adversarial learning has provided an effective approach to augment the discriminative power of feature representation, which brings about the availability of updated models. In our implementation, we set $\alpha = 0.3$ in optimizing embedding discriminator. According to the statistics of feature distribution in the embedding space, the β_1 and β_2 is set to 0.6 and 0.3, respectively.

2) Group Sensitive Triplet Embedding: The captured vehicle/person images from different cameras present large variances in visual appearance, also known as the intra-class variance such as viewpoints, poses, illuminations. To alleviate such issue for feature learning, we further employ a novel group sensitive triplet embedding (GSTE) method. With an introduced intermediate structure "group" between samples IDs, GSTE tends to build a sort of "similar attribute, closer distance" distribution in the embedding space.

Denote S^p and S^n as the set of samples of a given ID p and a set of different IDs, respectively. The samples of each ID are divided into G groups, and distinct groups represent intra-class variance. Denote $S^{p,g}$ ($g \in \{1, 2, \ldots, G\}$) as a set of instances in group g for the sample p. The objective is to minimize the distances of samples of the same group and meanwhile push the samples away from the different IDs, which can be denoted as:

$$\begin{aligned} \min_M \sum_{g=1}^{G} \sum_{x_i, x_j \in S^{p,g}} \|x_i - x_j\|_M^2 \\ s.t. \sum_{x_i \in S^p, x_n \in S^n} \|x_i - x_n\|_M^2 \geq \alpha, \quad M \succeq 0, \end{aligned} \tag{4}$$

where M is a metric matrix computed by the deep network, and α is the minimum margin gap between the samples of different IDs.

We design an intra-class variance structure in embedding optimization. Given ID p, denote a^p as an anchor in the set S^p and $a^{p,g}$ as a group anchor in $S^{p,g}$. The intra-class within group structure can be imposed as follows:

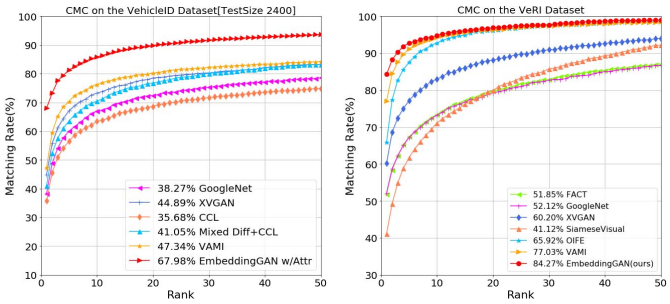
$$\|f(a^{p,g}) - f(x_g^p)\|^2 + \beta \leq \|f(a^{p,g}) - f(x_i^p)\|^2, \quad x_g^p \in S^{p,g} \tag{5}$$

where β is the minimum margin between different groups of the same IDs, $x_g^p \in S^{p,g}$ and $x_i^p \notin S^{p,g}$. This can be explained as a stronger distance constraint for samples with a similar attribute of the same ID.

The inter-class relationship is also modeled by the conventional triplet loss. The joint supervision from inter-class and intra-class will build a group sensitive structure in feature embedding. Thus both the inter-class and intra-class constraints are incorporated in the embedding learning, and the relationship among multiple groups is also characterized. We set $\alpha = 0.3$, which is a widely used setting in the deep embedding learning. Compared to the α constraint between the inter-class relationships, the extra constraint β is a loose constraint to make the samples of one ID with the same attribute to be closer in intra-class feature distribution. Therefore, we set $\beta = 0.1$ in our implementation.

3) Results Comparisons and Discussions: In this subsection, we first briefly introduce the experiments setup, then analyze the comparison results with the state-of-the-art methods. The significance and practical issues of the ReID techniques in the smart visual system are also discussed.

Vehicle ReID experiments were performed on VechileID [45] and VeRI [42] datasets. The VehicleID dataset consists of 22,1763 images of 26,267 vehicles (13,134 for training and 13,133 for testing). The VeRI [42]



(a) (b)

Fig. 3. The CMC curve of different methods on VehicleID and VeRI dataset.

TABLE II

PERFORMANCE COMPARISONS OF VEHICLE RE-IDENTIFICATION METHODS ON VEHICLEID (VEID) [45] AND VERI-776 (VERI) [42] BENCHMARKS. THE PERFORMANCE IS EVALUATED BY THE MAP

Methods	Dims	Size	Year	VEID	VERI	GFLOPS
Triplet	1024	4KB	2015	37.3	-	2.3
Softmax	1024	4KB	2015	58.0	34.3	2.3
Triplet+Softmax	1024	4KB	2014	65.0	55.8	2.3
CCL [45]	2048	8KB	2016	38.6	-	2.0
PROID [42]	1024	4KB	2016	-	27.77	2.5
STR [44]	1024	4KB	2017	-	40.26	3.8
Mixed+CCL [45]	1024	4KB	2016	45.5	-	3.5
OIFE [62]	2048	4KB	2017	-	48.0	3.8
HDC [48]	1024	4KB	2017	57.5	-	12.2
VAMI [49]	2048	8KB	2018	63.12	50.13	3.8
C2F-Rank [63]	2048	8KB	2018	63.5	-	3.8
EmbeddingGAN	1024	4KB	2018	75.1	55.36	2.3

TABLE III

PERFORMANCE COMPARISONS OF PERSON RE-IDENTIFICATION METHODS ON MARKET1501 AND DUKEMTMC DATASETS. THE PERFORMANCE ARE EVALUATED WITH TOP1 PRECISION AND MAP

Dataset	Market1501 [66]					
Method	Year	Dim	Size	R1	mAP	GFLOPS
BoW [64]	2015	16828	66KB	44.42	20.76	-
CRAFT [65]	2017	2048	8KB	68.7	42.3	7.3
P2S [66]	2017	800	3.2KB	70.72	44.27	4.7
CADL [67]	2017	256	1KB	73.84	47.11	3.8
USG-GAN [68]	2017	2048	8KB	78.06	56.23	3.8
LDCAF [69]	2017	256	1KB	80.31	57.53	13.9
SVDNet [70]	2017	1024	4KB	82.3	62.1	3.8
PAR [71]	2017	512	2KB	81.00	63.4	12.2
GSTE [3]	2017	2048	8KB	85.8	69.3	2.1
TriNet [72]	2018	8KB	2048	84.92	69.14	3.8
AACN [73]	2018	1024	4KB	85.90	66.87	12.2
DuATM [74]	2018	2048	8KB	91.42	76.62	3.8
Dataset	DukeMTMC [77]					
Method	Year	Dim	Size	R1	mAP	GFLOPS
BOW [64]	2016	16828	66KB	25.13	12.17	-
LOMO [76]	2015	26960	106KB	30.75	17.04	-
USG-GAN [68]	2017	2048	8KB	67.68	47.13	3.8
OIM [77]	2017	2048	8KB	68.10	-	3.8
SVDNet [70]	2017	1024	4KB	76.70	56.80	3.8
DPFL [65]	2017	4096	16KB	79.20	60.60	12.2
REDA [78]	2017	512	2KB	79.31	62.44	3.8
GSTE [3]	2018	2048	8KB	77.8	61.5	2.3
AACN [73]	2018	1024	4KB	76.84	59.25	12.2
DuATM [74]	2018	2048	8KB	81.8	64.58	3.8

dataset consists of 49,357 of 776 vehicles (576 for training and 200 for testing). The evaluation metrics are the widely used mean Average Precision (mAP) and Cumulative Matching Curve (CMC). Person ReID experiments were performed on Market1501 [64] and DukeMTMC [75] datasets. The Market1501 dataset contains 32,668 images of 1501 person (751 for train and 750 for test). The DukeMTMC dataset contains 36,411 images of 1404 person (751 for training and 750 for testing). The evaluation metrics are the widely adopted mAP and Top 1 Precision Rate (R1).

We report the comparison results of vehicle and person ReID in Tables II and III, respectively. Moreover, the runtime complexity FLOPs (Floating Point Operations) of listed methods are also presented for comprehensive evaluation. For vehicle ReID, our proposed EmbeddingGAN achieves the best results over VehicleID [45] and VeRI [42] benchmarks. The method HDC [48] has a similar motivation with our approach, which uses hard aware cascade scheme to select hard examples for better training. The performance comparisons with [48] demonstrates the superiority of the active hard negative generation scheme in the embedding space. It is worth noting that the network structure we used in EmbeddingGAN is VGG_CNN_M_1024 which is a smaller and more shallow network compared to the Googlenet used in [48]. In [49], a viewpoint aware multi-view inference (VAMI) scheme also involves generative adversarial network, which attempts to facilitate ReID from the perspective of cross-view vehicle generation to alleviate the negative impacts of viewpoint variation on feature representations. The vehicle ReID performance is incrementally improved on the benchmarks. However, in practical applications, more complex scenarios may impose critical challenges, and one important aspect is that the vehicle images captured from the low illumination, poor quality and small scale are not fully considered in the current benchmarks and methods. The vehicle ReID by means of 3D information and the skeleton of vehicles are also worth to be studied. In one word, the continuously updated models for better ReID performance can be expected, which is aligned with the design principle of this infrastructure of front-end sensing and back-end intelligent analysis.

As for the person ReID, our proposed GSTE method also achieves the competitive performance compared with the state-of-the art methods on Market1501 [64] and DukeMTMC [75] datasets. The recent trends on person feature representation are more on the attention scheme and part based scheme. The state-of-the-art approach DuATM [74] used the dual attention matching network to learn context-aware feature sequences and performed attentive sequence comparisons simultaneously. The AACN [73] also utilized pose-guided part attention and attention-aware feature composition to strengthen the feature representation under the pose variances of body parts and the part occlusion. However, the scale of the current person ReID benchmarks is still limited. For example, the widely used benchmark Market1501 only contains 1501 persons and the DukeMTMC contains 1404 persons.

In summary, the back-end analysis techniques are extremely crucial in surveillance systems, especially with the explosive growth on the data scale. As can be observed from both Table II and Table III, the features tend to be more compact and more discriminative. Moreover, the recent work [79], [80] on hashing aimed to compress the feature into the binary form further promotes the compactness of the discriminative features. The recent attempts [81], [82] used deep network within an end-to-end optimization framework to generate binary feature representation. The compactness is of great significance for economizing the smart visual sensing system in this unified infrastructure. To deploy the ReID techniques to the practical applications, there remains open issues. One is the feature extraction complexity in the front-end, as the multiple targets appearing in the captured images may also require the light weight deep models to generate feature representation with low computational cost. Moreover, the ReID models should be able to achieve scalable feature representations and meanwhile the interoperability is expected in the practical application scenarios. To resolve the interoperability, the standardization will be involved, which will be introduced in Section V.

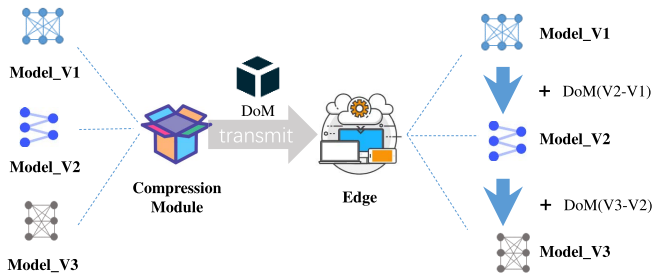


Fig. 4. Illustration of progressive model updating via DoM.

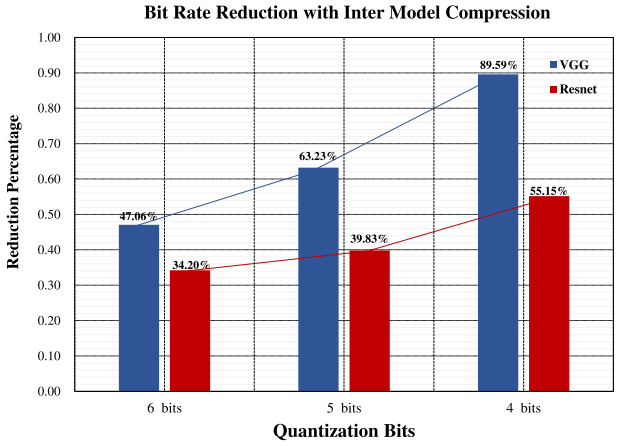


Fig. 5. The bit rate reduction with the inter-model compression strategy under different quantization bits.

E. Economics for Model Deployment and Updating

In this unified infrastructure, the dynamically acquired and sampled data can also be continuously converged to the central server for augmenting the training samples in model learning. The progressively enhanced deep learning models at the back-end further promote the smart visual sensing at the front-end, such as the mentioned ReID system with necessary model updating. For this purpose, the trained models are required to be sequentially distributed to the front-end, enabling stronger capabilities in feature extraction and representation. This demands efficient deep learning model transmission, which has been largely ignored in the existing literature. In view of this, in [27], we studied the sequential deep learning model transmission and attempt to remove the inter redundancy among multiple deep neural networks, as illustrated in Fig. 4.

Formally speaking, given the to-be-compressed model G with L layers, assuming the corresponding weights are represented with W_l (can be empty in the layer without weights such as ReLU), the classical deep learning model compression such as Han *et al.*'s approach [8] can be represented with the following formulation.

$$W_i' = C(W_1, W_2, \ldots, W_L), \quad i = 1 \ldots L \tag{6}$$

where the reconstructed weights W_i' in the i_th layer. Since the compression are determined based on the redundancy inside the given model through the compression algorithm C (including pruning, quantization, entropy encoding,etc.).

Distinctly, regarding economicalal design for model deployment and updating, the inter model redundancies are further taken into account in the scenario that there already exists one or multiple deep learning models G_{V1}, G_{V2}...G_{VJ} having high correlation with the to-be-transmitted one at the receiver side. As such, we reformulate this with the inter-prediction based compression algorithm C' as follows,

$$W_i'' = C'(W_1, W_2 \ldots, W_L; G_{V1}, G_{V2}, \ldots G_{VJ}), \quad i = 1 \ldots L \tag{7}$$

At the receiver side, the model can be reconstructed utilizing the inter model prediction with the selected prediction model. For simplicity, we validate this strategy by assigning only one model (e.g., the latest updated model) as the prediction model.

We firstly conduct experiments on two classical deep learning models, *i.e.*, Resnet50 [83] and VGG-16 [84] pretrained on Imagenet [85]. In particular, they are retrained 20 epoches with the learning rate of 0.001 on IMS dataset [86] using Triplet loss, and the layers after 'pool5' are removed. The inter-model prediction is applied to the retrained models and the latest updated model is used to calculate the differences of the models. Subsequently, we apply Scalar Quantization [21] to the two retrained models with and without inter-model prediction under different quantization levels by allocating bit-per-layer from 6-bit to 4-bit. The results are shown in Fig. 5, providing the bit rate savings when transmitting the model differences compared to the original model. Clearly, the inter model prediction can significantly remove the model redundancy, leading to better compression performance.

Furthermore, experiments are conducted in the scenario of person ReID to demonstrate the real-world model deployment and updating. We train a ResNet-50 on Market1501 dataset using Triplet and softmax loss [72] for 150 epochs with the learning rate of 0.001. The first model with 100 epochs is treated as the model-V1 and subsequently models V2 to V5 are obtained for every 10 epochs after the first 100 epochs. Along with the increment of the training, the performance is further enhanced and all the improved models are aimed to be subsequently transmitted. This simulates the real world model updating, as the updating models are continuously transmitted to the front-end for feature extraction. In particular, we follow the strategy in [27] and adopt the Differences of Models (DoM) representation, quantization and lossless encoding to compress the delivered models based on the inter model redundancy. Thus Eq. (7) can be reformulated as follows,

$$DoM_i = C'(W_i - P_i), \quad i = 1 \ldots L \tag{8}$$

where scalar quantization with encoding algorithm C' is applied to the differences between the to-be-transmitted model W and prediction model P of L layers. In essence, the prediction model can be assigned as the latest model has already been transmitted. For example, DoM V2 can be calculated based on the difference between V2 and V1.

The results of the model updating are shown in Fig. 6, where the whole model V1 is firstly transmitted, and subsequently the DoMs are conveyed to represent models from V2 to V5. From Fig. 6, we can observe that based on the progressively

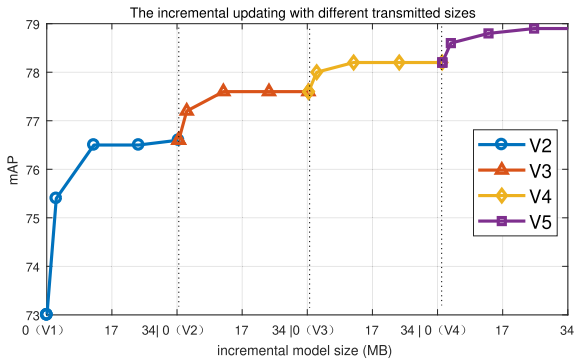


Fig. 6. The performance of person ReID [72] on market1501 dataset with incremental model updating. The incremental transmission occurs where the newly transmission is built upon on the former ones (*i.e.*, V3 is built upon V2 and V4 is based on V3). The x axis represents the transmission cost based on the former model. The x axis of 0MB represents the start-point of the next model version.

transmitting of the DoM, the performance on the front-end has been gradually improved. For example, for Model-V3 in Fig. 6, we incrementally transmit the Model-V3 based on the existing model Model-V2 using inter model prediction. The increasing sizes with the performance variations are also given. Similar trends regarding the performance improvements with the increase of the transmitted incremental model can be found from the initial version of the model (Model-V1) to the model with the best performances (Model-V5) based on the incremental updating built upon the model of the previous version. It is also worth mentioning that the original model size is around 56MB, and it is illustrated in Fig. 6 that when around 34MB is conveyed based on the former models, almost all the information in representing the differences between V3 and V2 have been transmitted. As such, the inter-model prediction strategy can provide great potentials regarding the economic model deployment and updating.

V. Standardizations for Economical Knowledge Communication

With the advance of the Internet of Thing (IoT) [87] along with the generation of big data, the pre-prossessing of raw data before transmission is under great demand due to the underlying unstructured and redundant characteristics. Similar considerations have been witnessed, and Knowledge Centering Networking (KCN) [88] is an innovative concept turning raw data into knowledge. The modality of such knowledge varies as different knowledge creation methods will result in different types of knowledge, such as compact feature as well as deep learning models. In this section, we discuss the standardization activities towards future knowledge communication, including both compact feature and network representation.

A. Compact Descriptors for Visual Search

Considering the significance of the feature transmission in the smart visual sensing system, MPEG has completed the standardization of Compact Descriptors for Visual Search (CDVS) [5], which was published in Sep. 2015. The CDVS uses handcrafted local and global descriptors to represent the visual characteristics of images, and the normative blocks of CDVS include the extraction procedure of both local and global descriptors. More concretely, the local descriptors in CDVS are the highly compressed SIFT descriptors via a low-complexity transform coding. Moreover, a Scalable Compressed Fisher Vector (SCFV) is aggregated from the raw local descriptors, which owns competitive matching accuracy and low memory cost. In view of the fluctuation of the available bandwidth in the mobile environment, the CDVS supports interoperability between different size bitstream. More technical details regarding CDVS can be found in [5], [89].

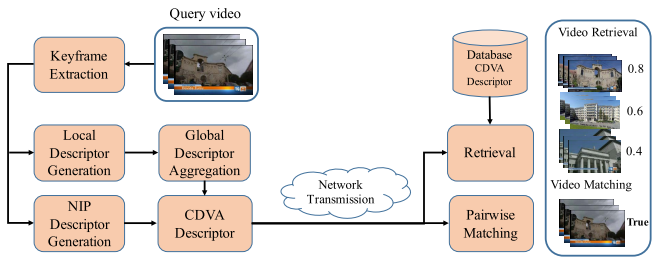


Fig. 7. Illustration of the MPEG-CDVA framwork [7].

TABLE IV
The Performance Comparison of CDVS and CDVA on CDVA Evaluation Benchmarks

	CDVS	CDVA (VGG16)	CDVA (ResNet50)	CDVA (ResNet101)	CDVA (ResNet152)
mAP	72.1	80.1	81.7	82.0	82.3
Precisian@R	71.2	76.7	78.6	78.9	79.2
TPR@FPR	83.6	87.9	87.0	86.6	87.3
Localization	66.2	72.5	70.3	69.8	70.9

B. Compact Descriptors for Video Analysis

The increasing demand of large-scale video analysis pushes the MPEG to move forward to standardize Compact Descriptors for Video Analysis (CDVA) [90]. Due to the sequentially correlated characteristics of video frames, generating the feature from each frame will result in both high feature redundancy and computational costs. The multi-keyframe based strategy was adopted in the ongoing CDVA standard, which converts the problem of video retrieval into image retrieval. In particular, the local and global descriptors of standardized CDVS descriptors are extracted on the sampled keyframes of a given query video, which are further packed together to constitute the CDVA descriptors. Moreover, the deep learning based NIP descriptors [91] are also adopted to further boost the analysis performance. Table IV provides the performance comparisons between CDVS and CDVA with NIP descriptors extracted from different deep neural networks. The evaluation framework of the ongoing CDVA is also illustrated in Fig. 7. In the pipeline, both the handcrafted and deep learning features are extracted and compressed. It is also worth mentioning that the NIP descriptor [92] has been adopted into the Draft of International Standard (DIS).

C. Deep Neural Network Compression

MPEG is also working towards the deep neural network compression in view of the applications of video surveillance,

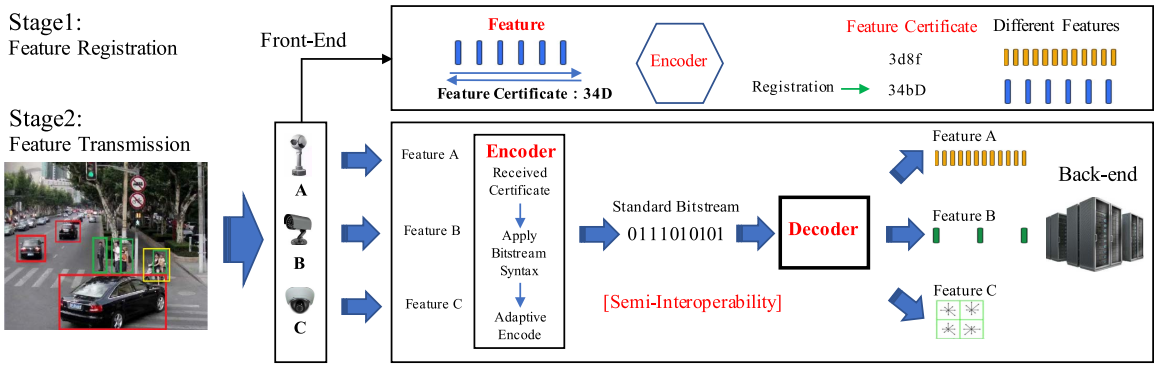


Fig. 8. The semi-interoperability-enabled feature transmission framework. First, new features are required to register the form of the feature in the encoder and obtain the corresponding certificate. The notation of 34D is an example to represent the certificate ID denoting the feature's form, which is used to help the encoder to recognize the input feature and perform adaptive encoding. Second, the features extracted at the front-end will be encoded as the standard feature bitstream by the feature encoder according to its certificate. In the other side, the back-end uses standard algorithm to decode the received features for further analysis.

self-driving vehicle, etc. Recently, the requirements towards the Compressed Representation of Neural Networks (NNR) has been identified, aiming to define an efficiently coded, interpretable and standardized representation for trained neural networks [93]. In [93], NNR seeks representations towards different types of neural networks with scalability as well as the compressed neural network inference under resource limited environments. Moreover, the incremental updates of compressed representation of neural networks is also treated as one particular goal of NNR.

In the perceived requirements, not only the neural network distribution and deployment, but also the retraining as well as image/video processing with neural network are concerned. Use cases have been wildly collected, including object recognition, translation app, public surveillance as well as CDVA, to extend possible usage of efficient neural networks and represent it using NNR. As a result, a number of specific NNR requirements have been considered, such as efficient representation of neural networks and efficient incremental representation of neural network. In our recent proposal, we study the incremental representation [94] from different use cases apart from the classical deep model compression in Call for Evidence(CFE) [95].

D. Toward Standardization of Deep Learning Features

In the context of big video data, it is essential to further ensure the interoperability for deep learning based video analysis. In particular, the feature coding differs from video coding in that feature coding pipeline involves both feature extraction and compression, while video coding treats visual pixel values as input. Moreover, the diversity of the deep models also creates dramatically different features. Therefore, a complete and exhaustive standard that should be able to fully ensure the interoperability typically standardizes the procedure of both feature extraction and compression. However, such standardization scheme requires the deterministic deep model and parameters. In the current stage, it is unlikely to use a generic deep model to cover a wide range of tasks. In view of this, the concept of semi-interoperability for feature coding was proposed in [96], in which only the procedure from raw features to the compressed feature bitstreams is considered, and the final syntax specifying the deep learning features is standardized. In this manner, the feature extraction procedure is still open for research exploration in the future. The differences between the semi-interoperability and full interoperability lies in the interoperable way. In semi-interoperability, the standard bitstream can achieve interoperation between different encoder and decoder sides, however the features from different feature extractors are still unable to interoperate with each other. By contrast, the full-interoperability can achieve the interoperation both at the standard bitstream and feature levels.

In Fig. 8, we show the semi-interoperability feature transmission framework. It can be observed that such semi-interoperability standard is dual to the video coding standard, since only the decoder is standardized. In particular, the decoder can only recover the features, but cannot account for the explanation of the features as the deep network models are not specified. Such semi-interoperability can ensure that any feature bitstreams from the same deep model conforming to such standard can be matched after decoding. In essence, the proposed standard does not possess fully interoperability, and the feature bitstreams under such standard may contain different information. In this way, this standard can be kept with long-lasting vitality, and the new emerging deep models are also able to seamlessly collaborate with this standard.

VI. Envisioning the Future

Despite significant recent progress on this typical infrastructure has been made at both technology and standard levels, there remains significant room for further improvement. In particular, with the fast development of deep learning technologies, there is a gap between the current economical visual information management system and what is desired in the context of video big data. As such, it is expected that a wide spectrum of functionalities could be further supported, to meet the grand challenges faced by smart city initiatives. Here, we envision the future technologies from four perspectives, including scalability, interoperability, utility and feasibility.

- *Scalability:* In principle, scalability requires the encoded stream containing one or more subset bitstreams such that the subsets themselves generated by dropping packets can be successfully decoded. The scalability of compact feature and deep neural network representations is a desirable property that supports the adaptation according to dynamic network conditions and decoding capabilities. Scalable video coding has been extensively studied in both literatures and standardizations. As the feature representation is able to summarize the visual characteristics at different granularities, the scalability is also naturally supported in feature and neural network representation by focusing on appropriate feature dimensionality or network layers. Regarding MPEG-CDVS, where handcrafted features are adopted, six descriptor lengths including 512 bytes, 1K, 2K, 4K, 8K and 16K are supported. The global feature descriptor adopted in CDVS-Scalable Compressed Fisher Vector (SCFV) achieved rate-scalable representations by choosing a subset of Gaussian components. However, to achieve scalability in deep feature representation, it is necessary to identify the importance of each component. This is a challenging task as the deep learning features are usually end-to-end learned. For deep neural networks, the significance of the layer should also be well interpreted to maximize the representation capability given the constraint. This demands to open the "black-box" of deep learning, which is a meaningful task with grand challenges.
- *Interoperability:* The standardizations of compact feature and deep neural network representations enable the interoperability such that any bitstream that conforms the standard can be properly decoded and utilized. This excellent property has greatly facilitated the applications of video compression based on a series of coding standards, including H.264/AVC [97], HEVC [98] and the on-going VVC [99]. It can be anticipated that the deep learning techniques will gradually become mature and generic. By then, the full interoperability may be achieved by a standardized unified model that can be applied to various tasks. However, instead of the well-established image/video signal, the rapid advancements of the deep learning techniques are now dramatically upgrading the network structures and feature representations. This presents technical barriers in standardizing the compact features and deep learning models. Overall, here we are trying to claim that at the current transitional development stage, there are flexible and practical alternative solutions for the standardization that can be considered and deployed. The concept of semi-interoperability is presented in [96], which suggests that only the pipeline from raw features to the compressed bitstream is standardized and specified, while how the feature can be generated is left open. Moreover, conveying the intermediated features instead of the top-layer features to improve the generalization ability of feature representation could also be an alternatively and promising solution [100].
- *Utility:* An excellent economical image/video management strategy is to achieve the good trade-off between the bitrate for representation and the ultimate utility. This is based on the widely applied rate-distortion theory, inspired by the fact that there is a growing interest of using the ultimate utility as the optimization function. However, one important issue that could possibly impede this progress is the lack of mathematical understanding of the properties of the utility functions, such that the rate-utility relationship is very difficult to be theoretically characterized. In particular, when there are multiple tasks simultaneously required to be performed, how to achieve a good balance between these tasks also requires further study, especially in the scenario where different tasks are evaluated by different utility functions. Therefore, a unified function that can well characterize the utility in the analysis tasks can lay the groundwork for the potentials of deploying the infrastructure in more practical applications.
- *Feasibility:* Though the proposed infrastructure has several satisfying characteristics that can be well exploited in economizing the visual system of city brain, a major drawback is that it is extremely difficult to recover the visual signal at pixel level given the compact feature representation, leading to difficulties for further investigations such as human viewing. This may limit the feasibility of the infrastructure as there is a deficit of generative models based on the compact feature in analysis tasks. A promising strategy is combining feature and texture streams coherently, and this could be useful when both analysis and human viewing are required. However, the texture bitstream will largely increase the bitstream volume, such that some excellent properties of feature compression no longer exist. It is also worth to explore extracting images/videos feature at compressed domain. Performing feature extraction in compressed domain has obvious advantages. First, the feature extraction complexity can be reduced by removing the overhead of decompression and less computing workload in compressed domain. Second, compression algorithms already implicitly reveal some features of the images which can provide good insights for image description based on feature representation. Overall, we hope to make the point that each technology has its corresponding application scenarios.

VII. Conclusion

In this paper, we took a certain viewpoint regarding how efficient and economical visual data management and analysis ought to be achieved in smart city applications. Though the recent advances of deep learning techniques have greatly facilitated many visual analysis tasks, there are still unprecedented challenges regarding how these techniques can be feasibly and practically applied in real application scenarios. Therefore, rather than providing an exhaustive survey of all relevant techniques, we have emphasized on the front-end visual sensing and back-end analysis, both of which rely on the economic and efficient visual data representation.

The relevant technologies, standardizations as well as applications are all discussed, and instructive examples are also provided to illustrate the importance of visual information in ensuring the city security. It is expected that the relevant techniques and standardizations will greatly benefit the construction of the visual system in the city brain.

REFERENCES

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

[2] Y. Lou *et al.*, "Towards digital retina in smart cities: A model generation, utilization and communication paradigm," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2019.

[3] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L. Duan, "Group-sensitive triplet embedding for vehicle reidentification," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2385–2399, Sep. 2018.

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

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

[6] H. Choi and I. V. Bajic. (2018). "Deep feature compression for collaborative object detection." [Online]. Available: https://arxiv.org/abs/1802.03931

[7] L.-Y. Duan *et al.* (2017). "Compact descriptors for video analysis: The emerging MPEG standard." [Online]. Available: https://arxiv.org/abs/1704.08141

[8] S. Han, H. Mao, and W. J. Dally. (2015). "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding." [Online]. Available: https://arxiv.org/abs/1510.00149

[9] Y. Guo, A. Yao, and Y. Chen, "Dynamic network surgery for efficient DNNs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1379–1387.

[10] X. Dong, S. Chen, and S. Pan, "Learning to prune deep neural networks via layer-wise optimal brain surgeon," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4860–4874.

[11] H. Zhou, J. M. Alvarez, and F. Porikli, "Less is more: Towards compact CNNs," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 662–677.

[12] X. Yu, T. Liu, X. Wang, and D. Tao, "On compressing deep models by low rank and sparse decomposition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7370–7379.

[13] M. Masana, J. van de Weijer, L. Herranz, A. D. Bagdanov, and J. M. Alvarez, "Domain-adaptive deep network compression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 16, Oct. 2017, pp. 4289–4297.

[14] S. Lin, R. Ji, X. Guo, and X. Li, "Towards convolutional neural networks compression via global error reconstruction," in *Proc. IJCAI*, 2016, pp. 1753–1759.

[15] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, vol. 2, 2017, pp. 1389–1397.

[16] J.-H. Luo, J. Wu, and W. Lin. (2017). "ThiNet: A filter level pruning method for deep neural network compression." [Online]. Available: https://arxiv.org/abs/1707.06342

[17] A. Aghasi, A. Abdi, N. Nguyen, and J. Romberg, "Net-Trim: Convex pruning of deep neural networks with performance guarantee," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3180–3189.

[18] Y. Wang, C. Xu, C. Xu, and D. Tao, "Beyond filters: Compact feature map for portable deep model," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3703–3711.

[19] Y.-D. Kim, E. Park, S. Yoo, T. Choi, L. Yang, and D. Shin. (2015). "Compression of deep convolutional neural networks for fast and low power mobile applications." [Online]. Available: https://arxiv.org/abs/1511.06530

[20] J. Ye, X. Lu, Z. Lin, and J. Z. Wang. (2018). "Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers." [Online]. Available: https://arxiv.org/abs/1802.00124

[21] Y. Gong, L. Liu, M. Yang, and L. Bourdev. (2014). "Compressing deep convolutional networks using vector quantization." [Online]. Available: https://arxiv.org/abs/1412.6115

[22] A. Zhou, A. Yao, Y. Guo, L. Xu, and Y. Chen. (2017). "Incremental network quantization: Towards lossless CNNs with low-precision weights." [Online]. Available: https://arxiv.org/abs/1702.03044

[23] C. Leng, H. Li, S. Zhu, and R. Jin. (2017). "Extremely low bit neural network: Squeeze the last bit out with ADMM." [Online]. Available: https://arxiv.org/abs/1707.09870

[24] C. Louizos, K. Ullrich, and M. Welling, "Bayesian compression for deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3290–3300.

[25] P. Luo *et al.*, "Face model compression by distilling knowledge from neurons," in *Proc. AAAI*, 2016, pp. 3560–3566.

[26] J. Lin, Y. Rao, J. Lu, and J. Zhou, "Runtime neural pruning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2178–2188.

[27] 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.*, 2018, pp. 1625–1633.

[28] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 803–818.

[29] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *Proc. ICCV*, 2017, pp. 2914–2923.

[30] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[31] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Computer Vision—ECCV 2006*. Berlin, Germany: Springer, 2006, pp. 404–417.

[32] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2012, pp. 2564–2571.

[33] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. Comput. Vis. Pattern Recognit.*, 2010, pp. 3304–3311.

[34] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed Fisher vectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3384–3391.

[35] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 584–599.

[36] V. Chandrasekhar, J. Lin, O. Morère, H. Goh, and A. Veillard, "A practical guide to CNNs and Fisher vectors for image instance retrieval," *Signal Process.*, vol. 128, pp. 426–439, Nov. 2016.

[37] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 36–45.

[38] G. Tolias, R. Sicre, and H. Jégou. (2015). "Particular object retrieval with integral max-pooling of CNN activations." [Online]. Available: https://arxiv.org/abs/1511.05879

[39] L. Ding, Y. Tian, H. Fan, Y. Wang, and T. Huang, "Rate-performance-loss optimization for inter-frame deep feature coding from videos," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5743–5757, Dec. 2017.

[40] H. Choi and I. V. Bajic. (2018). "Near-lossless deep feature compression for collaborative intelligence." [Online]. Available: https://arxiv.org/abs/1804.09963

[41] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan, "Veri-wild: A large dataset and a new method for vehicle re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019.

[42] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 869–884.

[43] R. S. Feris *et al.*, "Large-scale vehicle detection, indexing, and search in urban surveillance videos," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 28–42, Feb. 2012.

[44] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1900–1909.

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

[46] X. Zhang, F. Zhou, Y. Lin, and S. Zhang. (2015). "Embedding label structures for fine-grained feature representation." [Online]. Available: https://arxiv.org/abs/1512.02895

[47] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1386–1393.

[48] Y. Yuan, K. Yang, and C. Zhang. (2016). "Hard-aware deeply cascaded embedding." [Online]. Available: https://arxiv.org/abs/1611.05720

[49] Y. Zhou, A. Dhabi, and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6489–6498.

[50] Y. Zhou, L. Liu, and L. Shao, "Vehicle re-identification by deep hidden multi-view inference," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3275–3287, Jul. 2018.

[51] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1335–1344.

[52] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 475–491.

[53] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, 2017, Art. no. 13.

[54] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2017, pp. 3960–3969.

[55] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 34–39.

[56] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3908–3916.

[57] X. Zhang *et al.* (2017). "AlignedReID: Surpassing human-level performance in person re-identification." [Online]. Available: https://arxiv.org/abs/1711.08184

[58] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[59] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, vol. 2, no. 8, pp. 403–412.

[60] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 994–1003.

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

[62] Z. Wang *et al.*, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2017, pp. 379–387.

[63] H. Guo, C. Zhao, Z. Liu, J. Wang, and H. Lu, "Learning coarse-to-fine structured feature embedding for vehicle re-identification," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 6853–6860.

[64] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1116–1124.

[65] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2018, pp. 2590–2600.

[66] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3741–3750.

[67] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 6, Jul. 2017, pp. 5771–5780.

[68] Z. Zheng, L. Zheng, and Y. Yang. (2017). "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*." [Online]. Available: https://arxiv.org/abs/1701.07717

[69] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 384–393.

[70] Y. Sun, L. Zheng, W. Deng, and S. Wang. (2017). "SVDNet for pedestrian retrieval." [Online]. Available: https://arxiv.org/abs/1703.05693

[71] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. ICCV*, Oct. 2017, pp. 3219–3228.

[72] A. Hermans, L. Beyer, and B. Leibe. (2017). "In defense of the triplet loss for person re-identification." [Online]. Available: https://arxiv.org/abs/1703.07737

[73] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang. (2018). "Attention-aware compositional network for person re-identification." [Online]. Available: https://arxiv.org/abs/1805.03344

[74] J. Si *et al.* (2018). "Dual attention matching network for context-aware feature sequence based person re-identification." [Online]. Available: https://arxiv.org/abs/1803.09937

[75] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis. Workshop Benchmarking Multi-Target Tracking*, 2016, pp. 17–35.

[76] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2197–2206.

[77] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3415–3424.

[78] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. (2017). "Random erasing data augmentation." [Online]. Available: https://arxiv.org/abs/1708.04896

[79] Z. Wang, L.-Y. Duan, J. Yuan, T. Huang, and W. Gao, "To project more or to quantize more: Minimize reconstruction bias for learning compact binary codes," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 2181–2188.

[80] Z. Wang, L.-Y. Duan, T. Huang, and W. Gao, "Affinity preserving quantization for hashing: A vector quantization approach to learning compact binary codes," in *Proc. AAAI*, 2016, pp. 1102–1108.

[81] J. Lu, V. Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1979–1993, Aug. 2018.

[82] Y. Duan, J. Lu, Z. Wang, J. Feng, and J. Zhou, "Learning deep binary descriptor with multi-quantization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1183–1192.

[83] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[84] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[85] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[86] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," in *Proc. ECCV*, 2016, pp. 3–20.

[87] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

[88] D. Wu, Z. Li, J. Wang, Y. Zheng, M. Li, and Q. Huang. (2017). "Vision and challenges for knowledge centric networking (KCN)." [Online]. Available: https://arxiv.org/abs/1707.00805

[89] L.-Y. Duan *et al.*, "Fast MPEG-CDVS encoder with GPU-CPU hybrid computing," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2201–2216, May 2018.

[90] *Call for Proposals for Compact Descriptors for Video Analysis (CDVA)—Search and Retrieval*, document I. JTC1/SC29/WG11/N15339, Warsaw, Poland, Jun. 2015.

[91] J. Lin *et al.*, "HNIP: Compact deep invariant representations for video matching, localization, and retrieval," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 1968–1983, Sep. 2017.

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

[93] *Use Cases and Requirements for Compressed Representation of Neural Networks*, document I. JTC1/SC29/WG11/N17740, 2018.

[94] *Study for the Incremental Representation for Neural Networks*, document I. JTC1/SC29/WG11/m44050, 2018.

[95] *Call for Evidence on Neural Network Compression*, document I. JTC1/SC29/WG11/N17757, 2018.

[96] 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 Mag.*, to be published.

[97] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H. 264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, 2003.

[98] G. J. Sullivan, J.-R. Ohm, W.-J. 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.
[99] J. Chen, B. Bross, and S. Liu, *Versatile Video Coding (Draft 2)*, document JVET K1001, 2018.
[100] Z. Chen, W. Lin, S. Wang, L. Duan, and A. C. Kot. (2018). "Intermediate deep feature compression: The next battlefield of intelligent sensing." [Online]. Available: https://arxiv.org/abs/1809.06196

Yihang Lou (S'18) received the B.S. degree in software engineering from the Dalian University of Technology, Liaoning, China, in 2015. He is currently pursuing 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 (M'06) received the Ph.D. degree in information technology from The University of Newcastle, Callaghan, Australia, in 2008. He has been an Associate Director of the Rapid-Rich Object Search Laboratory, a joint lab between Nanyang Technological University, Singapore, and Peking University (PKU), China, since 2012. He is currently a Full Professor with the National Engineering Laboratory of Video Technology, School of Electronics Engineering and Computer Science, PKU. His research interests include multimedia indexing, search, and retrieval, mobile visual search, visual feature coding, and video analytics. He received the *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, the China Patent Award for Excellence in 2017, and the National Information Technology Standardization Technical Committee Standardization Work Outstanding Person Award in 2015. He is serving as a Co-Chair for the MPEG Compact Descriptor for Video Analytics. He was a Co-Editor of MPEG Compact Descriptor for Visual Search Standard (ISO/IEC 15938-13). He is currently an Associate Editor of the *ACM Transactions on Intelligent Systems and Technology* and the *ACM Transactions on Multimedia Computing, Communications, and Applications*.

Shiqi Wang (M'15) received the B.S. degree in computer science from the Harbin Institute of Technology in 2008, and the Ph.D. degree in computer application technology from the Peking University, in 2014. From 2014 to 2016, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada. From 2016 to 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. He has proposed over 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.

Ziqian Chen received the B.S. degree in computer science from the Dalian University of Technology, Liaoning, China, in 2017. He is currently pursuing 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.

Yan Bai (S'18) received the M.S. degree from the School of Electrical Engineering and Computer Science, Peking University, Beijing, China, in 2018. She is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Computer Science, Peking University. Her research interests include large-scale video retrieval and fine-grained visual recognition.

Changwen Chen (F'04) received the B.S. degree from the University of Science and Technology of China in 1983, the M.S.E.E. degree from the University of Southern California in 1986, and the Ph.D. degree from the University of Illinois at Urbana-Champaign in 1992. He was Allen Henry Endow Chair Professor with the Florida Institute of Technology from 2003 to 2007. He was on the faculty of Electrical and Computer Engineering at the University of Rochester from 1992 to 1996 and on the faculty of Electrical and Computer Engineering at the University of Missouri-Columbia from 1996 to 2003. He is currently the Dean of the School of Science and Engineering, Chinese University of Hong Kong, Shenzhen. He also serves as the Deputy Director of the Peng Cheng Laboratory. He continues to serve as an Empire Innovation Professor of computer science and engineering with the University at Buffalo, State University of New York. His research has been supported by NSF, DARPA, Air Force, NASA, Whitaker Foundation, Microsoft, Intel, Kodak, Huawei, and Technicolor. He and his students have received nine Best Paper Awards or Best Student Paper Awards over the past two decades. He has also received several research and professional achievement awards, including the Sigma Xi Excellence in Graduate Research Mentoring Award in 2003, Alexander von Humboldt Research Award in 2009, the University at Buffalo Exceptional Scholar Sustained Achievement Award in 2012, and the State University of New York System Chancellor Award for Excellence in Scholarship and Creative Activities in 2016. He has served as a Conference Chair for several major IEEE, ACM, and SPIE conferences related to multimedia video communications and signal processing. He has been the Editor-in-Chief for the IEEE TRANSACTIONS ON MULTIMEDIA from 2014 to 2016. He has also served as the Editor-in-Chief for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2006 to 2009. He has been an Editor for several other major IEEE TRANSACTIONS and Journals, including the PROCEEDINGS OF IEEE, the IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS, and the IEEE JOURNAL OF EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS. He is an SPIE Fellow since 2007.

Wen Gao (M'92–SM'05–F'09) received the Ph.D. degree in electronics engineering from The University of Tokyo, Tokyo, Japan, in 1991. 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 is currently a Professor in computer science with the Institute of Digital Media, School of Electronic Engineering and Computer Science, Peking University, Beijing, China. He has chaired a number of prestigious international conferences on multimedia and video signal processing, such as IEEE ICME and ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations. He served on the Editorial Boards for several journals, such as the IEEE TRANSACTION ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTION ON AUTONOMOUS MENTAL DEVELOPMENT, the *Eurasip Journal of Image Communications*, and the *Journal of Visual Communication and Image Representation*.