2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

Simultaneous Object Detection And Association In Connected Vehicle Platform*

Rui Guo, Member, IEEE, Shalini Keshavamurthy and Kentaro Oguchi

Abstract— **The connectivity in vehicular network extends the sensing capability in both the range and quality of the sensing data. One of the most significant benefits is the availability of sensing data from connected vehicles. Leveraging this data in useful ways is an attractive research topic in the community. In this paper, a novel one-pass deep neural network is proposed to implement object detection and the association simultaneously. Considering the bandwidth limitation in the typical vehicular network communication, the proposed algorithm not only highly compresses the feature representation but also maintains the high quality in the detection and association performance. The learning architecture is delicately designed to enhance the task by incorporating multi-modality features. Each modular unit in the system can be appropriately deployed in the vehicle onboard electrical control unit (ECU) and on remote servers to realize a pratical implementation in many applications.**

I. INTRODUCTION

The connected vehicle technology (CVT) propels the ever-increasing race toward intelligent transportation by advancing connectivity among and within the roadway infrastructure to significantly improve the safety and mobility [1], [2]. Meanwhile, automotive sensing and data exchange become utmost important, as it enables the vehicular platform to timely and precisely acquire the traffic and the environmental information at an extended range. Although the intuition has been discussed theoretically for years, the fundamental and systemic solution is rarely implemented due to the realistic constraints in the connected vehicle environment, in which the communication bandwidth is very limited [1].

The objective of the paper is to present the technologies that enable simultaneous dynamic objects detection and association in the connected vehicular platform. The primary idea is to link the individual observations from connected vehicles that may potentially detect the same objects by using the minimum communication bandwidth for the shared information. Real-time sensing capability in vehicles is also a critical requirement to make the detection of the dynamic objects, like moving vehicles and pedestrians possible. With this in mind, a novel one-pass deep twin network (ODTN) is proposed to tackle the challenges. The proposed architecture dedicatedly incorporates the detection neural network with the association network in a single pass pipeline. Considering the computational overhead in deeper feature generation, ODTN reuses the deeply learned features from the detection subnet into the association process. To alleviate the communication bottleneck in the vehicular network, the learned features are greatly compressed into a very compact format that will not cause any communication burden. The discriminative features are mostly preserved in the twin structured association network to satisfy the objective. Due to the twin structure, each branch of the network is identical in terms of the network configuration. Thus, the deployment of the detection and feature generation network is feasible in each ECU. Only the compact representation of the detected object is shared over the vehicular network. The association task is processed at the server side to aggregate detection results from multiple connected vehicles.

Fulfilling the simultaneous object detection and association in a single shot is obviously challenging, since the prevailed detectors are mostly appearance oriented only, whether the features are hand engineered or deeply learned [3], [4]. However, the association task requires more semantic correlation rather than the appearance. In the proposed ODTN, a comprehensive representation considering the object texture, color, context and viewpoint are adopted, aiming to exploit the contextual information so that the semantic attributes of the observations are consistent. As a result, the appearance coherence, geometry proximity and orientation consistence are fully utilized.

Driven by the emergence of the performance-proven deep learning technology in the visual sensing tasks, the proposed ODTN is also a convolutional neural network (ConvNet) based deep structure including a cascaded deep detection subnet (Dnet) and a twin structured association subnet (Anet) [5], [6]. Different from existing detectors, the emphasis of the work is on linking the detection and association in a single pipeline. The Dnet is capable of handling detection from single image. At association stage, a pair of the detections is fed into Anet. The output of ODTN should be a discriminative value that could be used to distinguish the detected objects from different views. Dnet and Anet are dedicatedly connected by sharing deeply learned features at the intermediate layer. The system is trained and tested on our benchmark dataset. Both the quantitative evaluation and systematic performance convincingly demonstrate the functional advantages and deployment potentials in connected vehicular platform for the emerging object detection and association requirements.

The contributions of this paper are list as the following,

- The representation learning units in ODTN consider a comprehensive feature set that can be adopted to improve the association. The association of objects benefit largely from factors comprising of the appearance,

*Research supported by Toyota Motor Corporation.

Rui Guo, Shalini Keshavamurthy and Kentaro Oguchi are with the Intelligent Computing Division at Toyota InfoTechnology Center USA, Mountain View, CA 94043 USA { rguo,skmurthy,oguchi }@us.toyota-itc.com

978-1-5386-4451-5/18/$31.00 ©2018 IEEE

geometry and viewpoint;

- This is the first to provide compact representation learning oriented multiple view observation association. The system could greatly compress the learned representation, which resolves the communication bandwidth limitation in connected vehicular network;
- The advantages of ODTN include its modular attribute. The easy decomposability of the proposed learning network makes it suitable to be deployed in the vehicle-server architecture.

The paper is organized as follows. Section II presents the related prior works. In Section III, we describe the methodology of the proposed system. Next, comprehensive evaluation is discussed in Section IV. Finally, Section V summarizes and concludes the paper.

II. PRIOR WORKS

The proposed work is related to methods for object detection and association in connected vehicular platform. Prior arts conventionally focus on either detection or association of objects but not the combination of the two. In this section, the literatures discussed would focus on the most related domains, the ConvNet based object detectors and the multiple view observation association in terms of object re-identification in surveillance.

A. *Deep Detectors*

Recently, deep convolutional neural networks have led to radical improvements in the detection of more general object categories. This shift came about when the successful application of deep ConvNets to image classification was transferred to object detection in the RCNN system of Girshick et al [7]. The serial of RCNN including its succeeding Fast and Faster RCNN works directly to solve the problem of “object proposal generation” [7], [8], [9]. A general pattern for current successful object detection systems is to generate a large pool of candidate boxes and classify those using ConvNet features. The algorithms derived from RCNNs are mainly preceded the new trend: aiming to accelerate the training and testing process. Although RCNN has excellent object detection accuracy, it is computationally intensive because it first warps and then processes each object proposal independently. Consequently, some well-known algorithms, which aim to improve its efficiency, such as SPP-net [10], RPN [11], YOLO [12] etc. have been developed. These algorithms detect objects faster, while achieving comparable accuracy with state-of-the-art benchmarks. Among them, Single shot Multibox Detector (SSD) is notable due to its excellent detection performance and speedy processing in real time applications [13]. Improved from RCNNs, the SSD completely eliminates proposal generation and subsequent pixel or feature resampling stages and encapsulates all computation in a single network [13]. Following the SSD pipeline, the detection task for vehicle emerges in the recent works [14], [15], [16].

B. *Observation Association*

Observation association is originally proposed in visual tracking tasks [17]. The concept relates to associate the detections in adjacent frames into a trajectory. Since the detections are usually captured in a very short time interval, the appearance change is marginal. In such a context, the appearance-based feature is good enough for the association. Considering the intention of the association work here, the most related study is about person re-identification (re-ID). The primary goal of person re-ID is to distinguish whether the person captured in the surveillance video has been observed in another place (time) by another camera. The initial work of the person re-ID is from Wojciech Zajdel, etc in 2005 [18]. In their method, a unique, latent label is assumed for every person, and a dynamic Bayesian network is defined to encode the probabilistic relationship between the labels and features (color and spatial-temporal cues) from the tracklets. The ID of an incoming person is determined by the posterior label distributions computed by an approximate Bayesian inference algorithm. Later on, re-ID research divides into two directions, one is following discriminative feature study, and another one is focusing on similarity metric learning. Impressive works have been explored either in hand-crafted systems or in deeply learned systems [19], [20]. However, the objective of the re-ID works is to identify the single person, but not pairing two observations. This separates the research into another track.

Motivated by the deep detectors and the deep re-ID, a novel one-pass detection and association framework is proposed in this paper. Different from the tracking and re-ID, the task here has the following attributes. Firstly, the observations from multiple-view are simultaneous. The geometric constraint plays crucial role in association. Secondly, despite the viewpoint difference, the background should convey strong evidence for the association. The one-pass processing requirement is also advocated in the work, which boosts the novel architecture creation.

III. METHODOLOGY

The diagram of the proposed system is illustrated in Fig. 1. We present a novel one-pass deep twin network (ODTN) that implements object detection and association simultaneously. In particular, the ODTN is comprised of a ConvNets based detection subnet (Dnet) that is in charge of locating the object of interest, and the twin structured association subnet (Anet) adopts paired inputs to distinguish them. As an important step, the detection output is augmented with color, context and viewpoint geometric information in order to enhance the semantics. Since the communication bandwidth limitation is present in the connected vehicular platform, ODTN emphasizes on the compact representation learning more than other methods. As a result, the ODTN network is distributable in vehicle-server structure, which is dominant in current connected vehicular architecture.

A. *Deep Feature From Object Detection*

In this section, we describe the object detection and deep feature extraction via the Dnet. The detailed description of Dnet is present in [13], and for the sake of completeness we summarize it here. The SSD network is built using the 16-layer VGG network [21]. It computes multiple-scaled

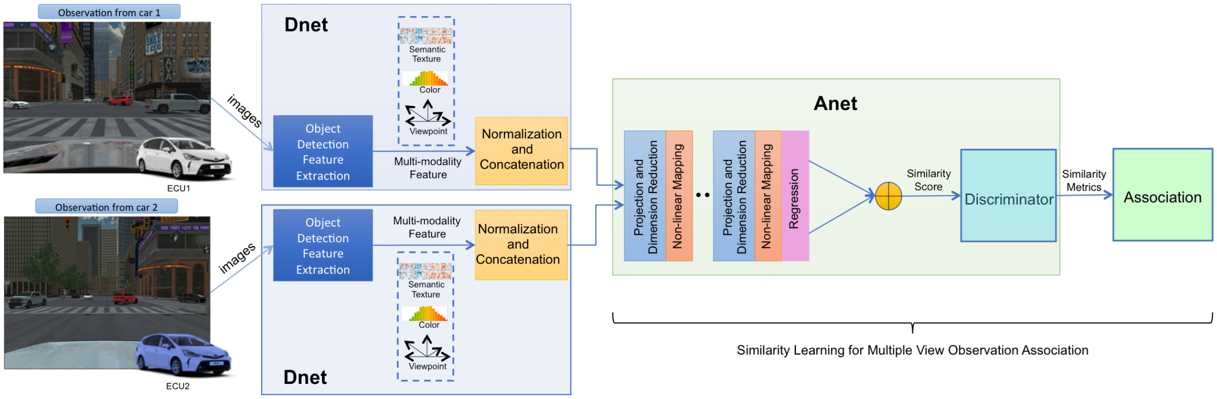

Fig. 1. System diagram for the proposed ODTN in simultaneous object detection and association.

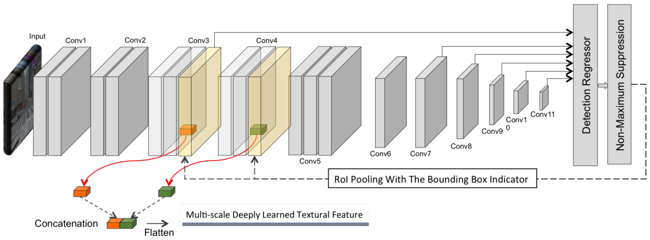

Fig. 2. Deeply learning and RoI based feature extraction network.

convolutional features from the image. Instead of generating the object proposals and resampling on the feature map, SSD creatively uses the default boxes over the multiple scaled feature maps, thus converts the detection problem into a classification task. The improvement also includes a regressor that helps to predict the bounding box offset. The obtained bounding box covers the object more tightly, and the detection process is done in a single pass.

The detected object comes with a tight bounding box identified its location in the image. After the detection, the bounding box coordinates are not abandoned, but utilized as RoI indicator. We conduct the RoI pooling over the deeply learned feature maps across multiple convolutional layers. The investigation on the feature maps reveals that different layers encode different types of features. The receptive field zooms as the neural network goes to deeper. As a result, higher layers capture semantic concepts on object categories, whereas lower layers encode more discriminative features to capture intra class variations. In order to obtain the optimal representation of the detection, the deeply features within the RoI in conv3-3 and conv4-3 of the VGG-16 network are extracted as the most relevant deep feature. It is noticed that, the RoI is normalized respectively according to the resolution of the feature map. The pooling kernel is of size 7 × 7. The feature maps are flattened and concatenated into a d-dimensional vector.

B. Multi-modality Feature Aggregation

The proposed approach exploits multi-modality features for associating the detections. Beyond the discriminative deep features, other attributes, like color information, context of the object existence and observation viewpoint essentially contribute to the observation association. We next illustrate the comprehensive feature in Fig.3 and discuss each of these features in the detail.

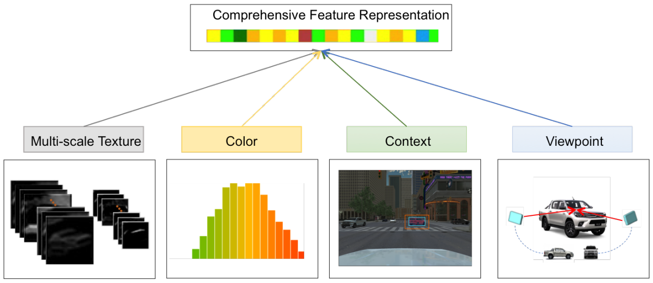

Fig. 3. Comprehensive feature composition.

- ***Color Histogram.*** As an important component, color information is the most discriminative feature to distinguish or associate different observations. We calculate the color histogram of the RoI in HSV color space and use the extracted feature as the auxiliary to compensate the textural deep features. Retinex color correction is applied beforehand to adjust the color distortion due to the shadow and rendering issues [22].
- ***Context Information.*** In the real traffic situation, similar vehicles with same color may be present at nearby locations. To distinguish them is one of the challenges. The context feature encodes the presence of contextual information, e.g. the surrounding texture of the road, background buildings or unique lane marker where the object is standing on [14], [23]. We use an extended rectangle around the detection bounding box as the contextual region. More concretely, we set the rectangle as 1/5 larger than the bounding box size, then apply the RoI pooling within the region to extract the deep features, as the computation as in III.A. The visual demonstration is illustrated in Fig. 4. By combining the context information, we are aiming to distinguish the objects that are difficult to be justified from their own appearances.
- ***Viewpoint.*** The geometric consistence in multiple views shed the light on providing another dimension of feature for observation association. Different from the re-ID problem, the association task considers matching the

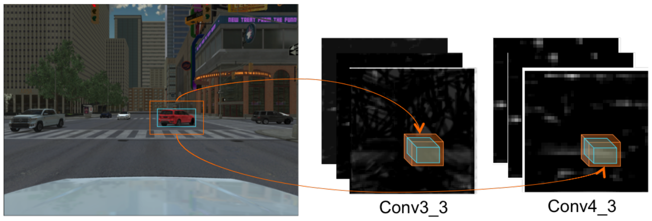

Fig. 4. Demonstration of the context information. The orange cubic demonstrates the actual RoI pooling in Conv3 and Conv4 layers with context information.

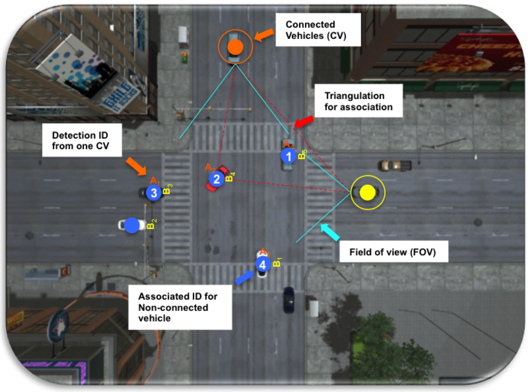

Fig. 5. Demonstration of the viewpoint constraint.

detections from multiple views at the same time. As illustrated in Fig. 5, the positions of the detected object and observers are formulating a *fixed* triangle. The viewpoints from observers' perspective are constrained by such geometry. The appearance of the detected object is implicitly related to the viewing angle of the observer. Inspired from this point, we use viewpoint as an important component for associating observations from multiple views. Quantitatively, as the orientation of the observer and the field of the view (FOV) of the camera are known parameters, the viewpoint is defined as the angle between the bearing and the center of the object bounding box.

The comprehensive feature to represent the detected object is the concatenated deeply learned feature, color feature, context information and viewpoint. By considering the multi-modality, the association process combines the appearance coherence, geometry proximity and orientation consistence. Later on, the experimental results also manifest the concept in terms of receiving better performance in the association.

C. Deep Twin Structured Network For Association

Till now, we have depicted the Dnet in the role of providing deeply learned features for the detected objects, and explained the detailed composition of the comprehensive representation. As the particularly focused task, the deep twin structured association network (Anet) addresses the multiple-view association with its unique characteristics, e.g. identical configuration in twin branches, dimension reduction capability and similarity metric learning [24], [25], [26].

Anet consists of twin branches that accept high dimensional paired input vectors but are joined by a contrastive loss function at the top. The loss function computes a similarity

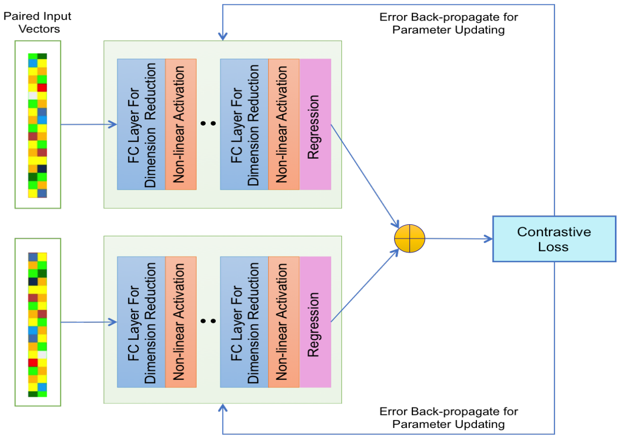
Paired Input Vectors
Error Back-propagate for Parameter Updating
FC Layer For Dimension Reduction
Non-linear Activation
..
FC Layer For Dimension Reduction
Non-linear Activation
Regression
Contrastive Loss
FC Layer For Dimension Reduction
Non-linear Activation
..
FC Layer For Dimension Reduction
Non-linear Activation
Regression
Error Back-propagate for Parameter Updating

Fig. 6. Demonstration of the twin structured Anet.

metric between the highly compressed representations on each branch. The compressed low dimensional representations are acquired by layer-by-layer non-linear mapping. The parameters and weights in the twin branches are tied. The identical weights encoded in the manifold preserved transformation guarantee that two extremely similar images could not possibly be mapped by their respective branches to very different locations in feature space because each branch maps the input along the same projection. In this way, we associate input pairs by evaluating the contrastive loss with a trained discriminator.

For a single branch in Anet, the input vector is fed into a fully connected (FC) layer. The number of neurons in the fully connected layer is much less than the length of the input. Following the FC layer, the rectified linear unit (ReLU) is exclusively performed to exploit the non-linear transformation. The FC layer and ReLU are combined as the core unit in Anet and sequentially applied several times with varying numbers of neurons. The feature dimension could be further reduced. The output vector after the transformation with core units is defined as *observation signature*.

To enforce the learning of a meaningful mapping, the contrastive loss function is introduced, which runs over the paired samples [24]. Let X_1, X_2 be a pair of observation signatures, and Y be a binary label assigned to this pair. Y is set to numerical “1” if X_1 and X_2 are deemed from the same object, otherwise Y is set to “1”. We adopt Euclidean distance D_W as the metric to measure the similarity between the signatures. That is,

$$D_W(X_1, X_2) = \|X_1 - X_2\|^2 \tag{1}$$

The general format of the contrastive loss is defined as

$$L(W, Y, X_1, X_2) = (1-Y)\frac{1}{2}(D_W)^2 + Y\frac{1}{2}\max(0, m-D_W)^2 \tag{2}$$

where W is the layer-wise transformation parameters and $m > 0$ is a margin [24]. The margin defines the radius of the observation signature. The different-object pair only contributes to the loss when their distance is within the margin range. The best setting in this work is $m = 3$. The learning force comes from the minimization over the contrastive loss. By applying the stochastic gradient descent algorithm, the updating gradient of L is

$$\frac{\partial L}{\partial W} = \begin{cases} D_W \frac{\partial D_W}{\partial W} & Y = 1 \\ -(m - D_W) \frac{\partial D_W}{\partial W} & Y = 0 \end{cases} \tag{3}$$

The contrastive loss drives the network to converge at the global equilibrium from both the same-object and different-object pairs.

At the testing stage, the learned Anet configuration and parameters are fixed. A simple regressor format discriminator (comparison function) that thresholds on the D_W is adopted to distinguish the testing pair. The parameters associated with the discriminator are also learned from the training dataset.

D. Deployment In The Vehicle-Server Architecture

Due to the identical structure of the twin network, the parameters in one branch are the same configuration as the other one. After the training, the single sub-network is deployed in each vehicle on-board Electronic Control Unit (ECU). Once the new observation comes in, the Dnet and feature compression layers in Anet process the input and generates a compact representation of the observation, which is suitable for the following associating process. The new observation representation greatly alleviated the data communication burden in the limited bandwidth environment.

The processed compact representation vector in each ECU, which represented the observed object, is transferred to the server through V2V/V2X network in real-time. The server is a physical/virtual computation node that aggregates and processes all the information from the connected vehicles in a certain covered range. It is facilitated with larger memory and more powerful computation capability. The trained comparison function is deployed in the server. Based on the comparison function, the computing unit in the server evaluates each two vectors in the observation dataset to determine if they are from the same object (with the output label "1") or not (with the output label "0"). If the one-time observations contain multiple objects, the association assignment is optimized in terms of minimizing the overall contrastive differences between all paired combinations of inputs.

IV. EXPERIMENTAL EVALUATION

We collected a relatively large dataset UnityCar in Unity3D simulator as the evaluation benchmark [27]. The images generated simulate the natural traffic scene. In each image, the vehicle is randomly placed in an intersection with a street view background. The observing vehicle is installed with a pre-calibrated camera that facing the intersection with an arbitrary orientation. The images are captured from the camera in a distance between 5 to 30 meters range. The whole dataset is using daytime neutral sunlight illumination. The dataset contains 11 different vehicle models. For each model, it includes 5 colors, with 500 images for each color. Benefited from the simulation environment, the vehicle position and bounding box information are known. We split the dataset as train/validation/test in 80%/10%/10% partitions to evaluate the performance of the proposed vehicle detection and association pipeline. The partition ensures that the randomly generate training and testing pairs are completed separated. In association learning, the images with the same vehicle model and the color are labeled as "1", otherwise, the label is "0". The detailed dataset information is summarized in Table I.

TABLE I
UNITYCAR DATASET DESCRIPTION

UnityCar Dataset (Res. 640×480**)**		
Vehicle Models (11)	**Colors (5)**	**FOV (°)**
Altima, Altis, Camry, cla_amg, Forester, gl_amg, Leaf, model S, Prius, Sorento, Tacoma	black, blue, gray, red, white	H: 56.8 V: 45

TABLE II
DETECTION PRECISION

Methods	AP (%)
Fast[8]	79.5
Faster[9]	82.8
SSD300[13]	84.7
ODTN	**87.3**

A. Fine-tuned Detection

The Dnet is pre-trained using the PASCAL VOC 2007 benchmark [13]. For the specific purpose in our automotive application, the fine-tuning is necessary to be applied with the UnityCar. The process includes optimal feature space sampling from Conv3-3, batch process and learning rate annealing. As reported in original SSD [13], Conv4-3 feature is too abstract to differentiate the intra-class variations. In such context, we adopted Conv3-3 to emphasize on the low-level features that are potentially beneficial. Batch process and learning rate annealing contribute to faster convergence and better generalization. We report the Dnet performance over the UnityCar with the same comparison metric as in [8]. The detection performance is obviously improved in terms of Average Precision (AP) without large-scale data augmentation.

B. Balancing The Compact Representation Learning And Association

As the most important claim, the proposed Anet could compress the comprehensive feature into a compact signature that enables the data sharing among the connected vehicle network. The functionality is implemented by using a three-layered FC and non-linear activation structure. Specifically, the first two FCs are of 1024 neurons in size, which reduce the input dimension into $2K$ bytes. The vector length is further reduced. We vary the number of neurons in the experiments, aiming to search for the balancing between the compact representation learning and association performance. The association performance is reported as accuracy metric in Table III. Clearly, the best performance achieved at the optimal dimension 32 (floating accuracy). The encouraging results show the compact representation reduces data amount of one observation from about 30 KB level to less than a hundred bytes level. The new representation reduces more

TABLE III

ASSOCIATION ACCURACY WITH VARIOUS FEATURE LENGTH

Margin	Length of Feature Vector	Accuracy
3	16 bytes	95%
3	**32 bytes**	**97%**
5	16 bytes	94%
5	32 bytes	87%

TABLE IV

ASSOCIATION ACCURACY WITH VARIOUS FEATURE COMBINATION

Feature Combination	**Accuracy**
Texture (Deeply Learned)	90%
Texture + Color	93%
Texture + Color + Context	97%
Texture + Color + Context + Viewpoint	**98%**

than **99%** data needed to transfer in the vehicular V2V/V2X network.

C. Empower The Association Learning

Aggregating features from different domains is not adhoc but considers their internal consistency. We study the effects of different features on the multiple-view association performance. The accuracies directly reflect the importance of adding these features. As the baseline, the accuracy of the association using deeply learned feature alone achieves around 90%. However, when we retrieve failure cases, mostly the discriminator cannot simply distinguish the vehicles with the same color. Other results also explain the value brought by the contextual information and viewpoint. Empowering the association learning from the comprehensive semantics can be addressed as the conclusion that enabled the twin structured Anet to perform well.

V. CONCLUSIONS

This paper presented a novel one-pass solution ODTN for simultaneous object detection and association in the connected vehicular platform. To tackle the challenges in multiple-view detection association, a twin structured association network with compact representation learning capability was formulated. Considering the bandwidth limitation bottleneck of the communication in the connected vehicle network, the ODTN dramatically reduces more than 99% of the communicated data, meanwhile maintains a very high performance in associating objects. The designed ODTN architecture is completely distributable, which could be realistically deployed in the connected vehicles to boost the potential applications in ADAS and autonomous driving.

ACKNOWLEDGMENT

The authors would like to thank Toyota Motor Corporation for sponsoring the research project.

REFERENCES

[1] G.-M. Hoang, B. Denis, J. Härri, and D. T. Slock, "On communication aspects of particle-based cooperative positioning in gps-aided vanets," in *Intelligent Vehicles Symposium (IV)*. IEEE, 2016, pp. 20–25.

[2] T. Tiedemann, C. Backe, T. Vögele, and P. Conradi, "An automotive distributed mobile sensor data collection with machine learning based data fusion and analysis on a central backend system," *Procedia Technology*, vol. 26, pp. 570–579, 2016.

[3] A. Chakraborty, A. Das, and A. K. Roy-Chowdhury, "Network consistent data association," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1859–1871, 2016.

[4] R. Guo, L. Liu, W. Wang, A. Taalimi, C. Zhang, and H. Qi, "Deep tree-structured face: A unified representation for multi-task facial biometrics," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–8.

[5] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[6] R. Guo, W. Wang, and H. Qi, "Hyperspectral image unmixing using autoencoder cascade," in *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2015 7th Workshop on*. IEEE, pp. 1–4.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[8] R. Girshick, "Fast r-cnn," *arXiv preprint arXiv:1504.08083*, 2015.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *european conference on computer vision*. Springer, 2014, pp. 346–361.

[11] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[14] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2147–2156.

[15] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals using stereo imagery for accurate object class detection," *IEEE Trans. on pattern analysis and machine intelligence*, 2017.

[16] C. Szegedy, S. Reed, D. Erhan, D. Anguelov, and S. Ioffe, "Scalable, high-quality object detection," *arXiv:1412.1441*, 2014.

[17] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3119–3127.

[18] W. Zajdel, Z. Zivkovic, and B. Krose, "Keeping track of humans: Have i seen this person before?" in *Robotics and Automation, Proceedings of the 2005 IEEE International Conference on*, pp. 2081–2086.

[19] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv:1610.02984*, 2016.

[20] W. Wang, A. Taalimi, K. Duan, R. Guo, and H. Qi, "Learning patch-dependent kernel forest for person re-identification," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, 2016.

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

[22] E. H. Land, "The retinex theory of color vision," *Scientific American*, vol. 237, no. 6, pp. 108–129, 1977.

[23] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2874–2883.

[24] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2. IEEE, 2006, pp. 1735–1742.

[25] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, vol. 2, 2015.

[26] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996.

[27] "Unity3D," https://unity3d.com/, accessed: 2018-01-05.