Learning Deep Neural Networks for Vehicle Re-ID withVisual-spatio-temporal Path Proposals

Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, Xiaogang Wang
Department of Electronic Engineering, The Chinese University of Hong Kong
SenseTime Group Limited1{ytshen, xiaotong, hsli, xgwang}@ee.cuhk.edu.hk2yishuai@sensetime.com

Reference: http://openaccess.thecvf.com/content_ICCV_2017/papers/Shen_Learning_Deep_Neural_ICCV_2017_paper.pdf

Intelligent Information Fusion Research Group

Motivation

- 行人重识别很热，榜单刷新的很快。

- 车辆重识别与行人重识别类似，2016年开始火起来。
行人重识别的方法或许可以迁移过来这边

- 车辆重识别中很多车辆款式一样，但是ID是不同的，
要识别他们不容易。同样在行人重识别中，有衣服一样，
但是ID不同的人。或许车辆重识别的方法也可以给行人重识别一些启发。

Vehicle Dataset

- VeRi-776 [Project] [paper] ✔
- PKU-VehicleID [Project] [pdf]
- PKU-VD [Project] [pdf]
- VehicleReId [Project] [pdf] ✔
- PKU-Vehicle[Project] [pdf]
- CompCars[Project] [pdf]

唯一有时空信息

Reference: https://github.com/knwng/awesome-vehicle-re-identification

榜单

VeRi				
Settings	Query = 1678, Test = 11579			
Methods	mAP	r = 1	r = 5	r = 20
LOMO [11]	9.78	23.87	39.14	57.47
DGD [28]	17.92	50.70	67.52	79.93
GoogLeNet [29]	17.81	52.12	66.79	78.77
FACT [15]	18.73	51.85	67.16	79.56
XVGAN [41]	24.65	60.20	77.03	88.14
SiameseVisual [23]	29.48	41.12	60.31	79.87
OIFE [26]	48.00	65.92	87.66	96.63
VAMI (Ours)	**50.13**	**77.03**	**90.82**	**97.16**
SiameseCNN+PathLSTM [23]	58.27	83.49	90.04	96.03
SiameseVisual([23])+STR([15])	40.26	54.23	74.97	91.68
VAMI (Ours) + STR([15])	**61.32**	**85.92**	**91.84**	**97.70**

Method	mAP (%)
FACT [27]	18.49
FACT+Plate-SNN+STR [28]	27.77
Siamese-Visual	29.48
Siamese-Visual+STR	40.26
Siamese-CNN	54.21
Chain MRF model	44.31
Path-LSTM	54.49
Siamese-CNN-VGG16	44.32
Path-LSTM-VGG16	45.56
Siamese-VGG16+ PathLSTM-VGG16	46.85
Siamese-CNN+Path-LSTM	**58.27**

Table 1: mAP by compared methods on the VeRi-776 dataset [28].

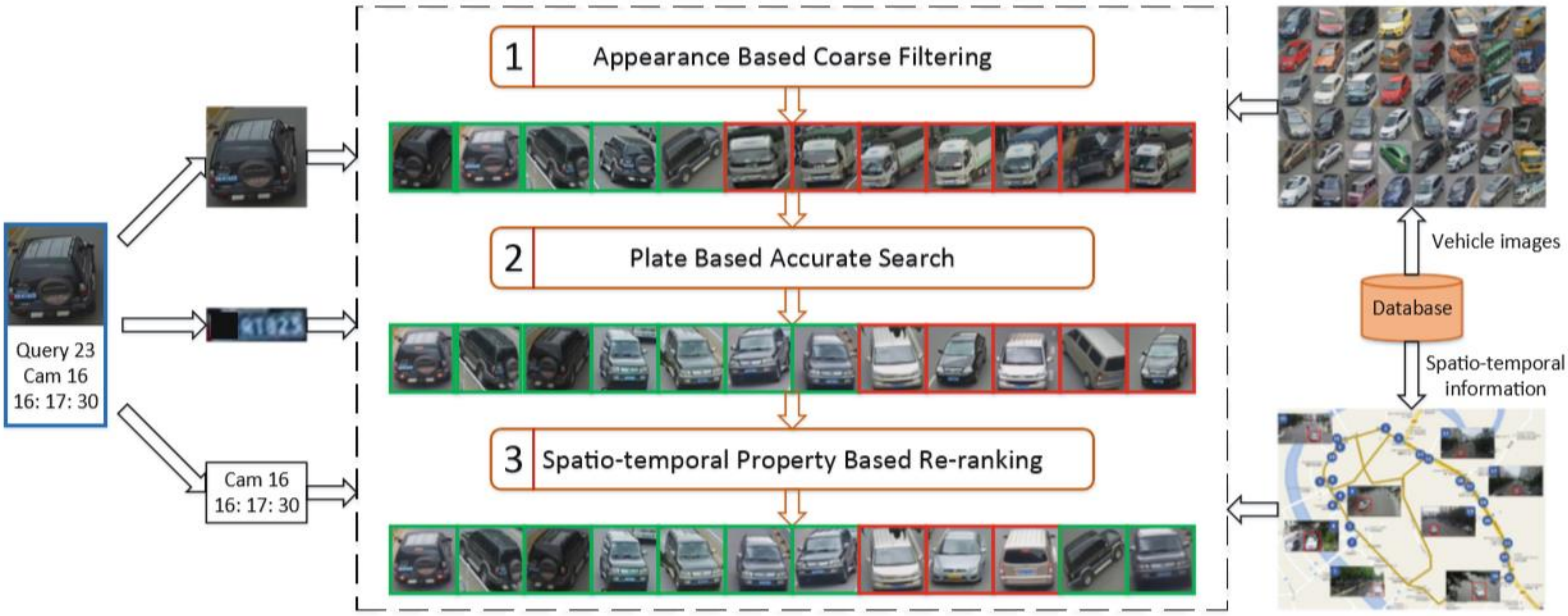
Method	top-1 (%)	top-5 (%)
FACT [27]	50.95	73.48
FACT+Plate-SNN+STR [28]	61.44	78.78
Siamese-Visual	41.12	60.31
Siamese-Visual+STR	54.23	74.97
Siamese-CNN	79.32	88.92
Chain MRF model	54.41	61.50
Path-LSTM	82.89	89.81
Siamese-CNN-VGG16	54.41	61.50
Path-LSTM-VGG16	47.79	62.63
Siamese-VGG16+ PathLSTM-VGG16	50.95	61.62
Siamese-CNN+Path-LSTM	**83.49**	**90.04**

Table 2: Top-1 and top-5 accuracies by compared methods on the VeRi-776 dataset [28].

榜单

车辆重识别(Vehicle Re-identification)

与行人重识别类似，都是一个图像检索问题，给定一组图片集(probe)，对于probe中的每张图片，从候选图片集（gallery）中找到最可能属于同一辆车的图片.



Reference: Liu X., Liu W., Mei T., Ma H. A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance. In: European Conference on Computer Vision. Springer International Publishing, 2016: 869-884.

Visual-spatio-temporal

简化的时空模型

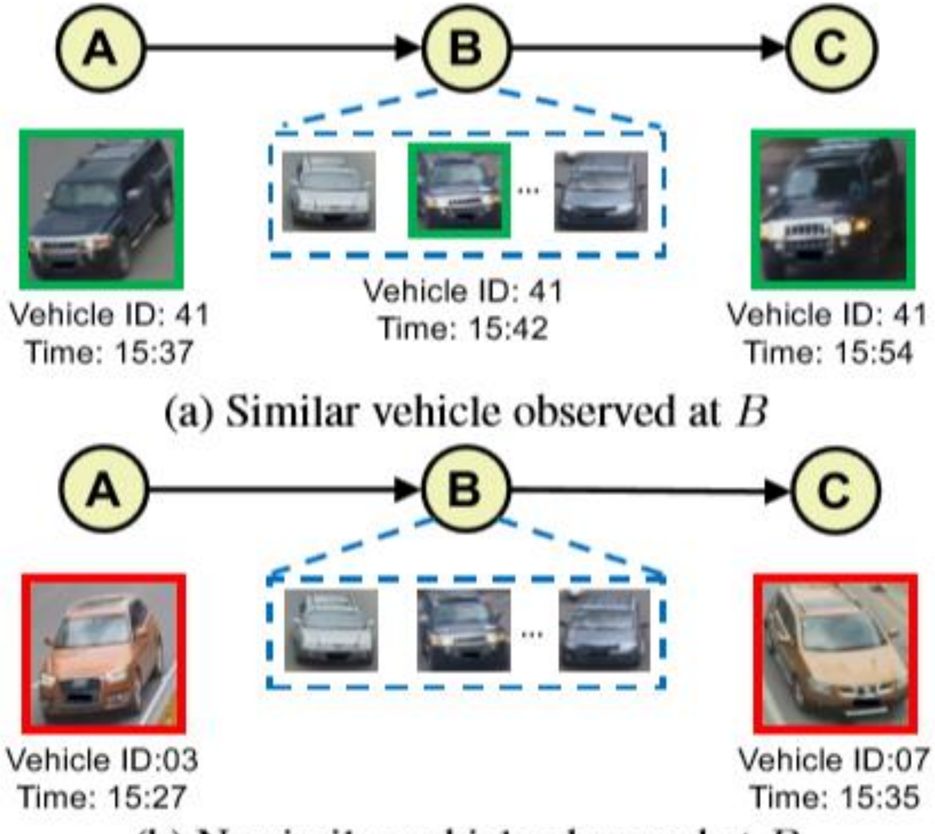


Figure 1: Illustration of spatio-temporal path information as important prior information for vehicle re-identification. (a) For vehicles with the same ID at A and C, it has to be observed at B. (b) If a vehicle with similar appearance and proper time is not observed at B, vehicles at A and C are unlikely to be the same vehicle.

Overall framework

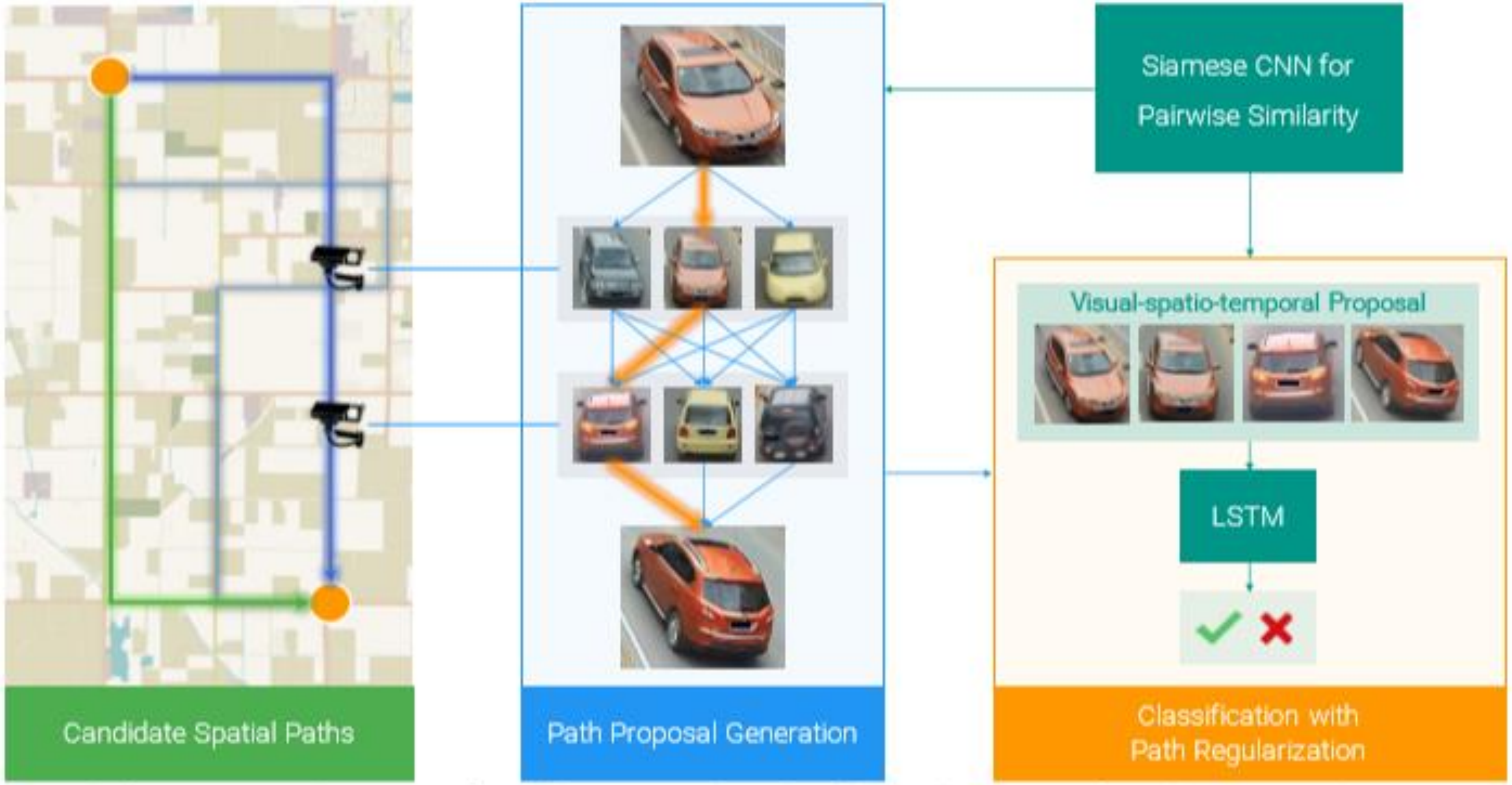


Figure 2: Illustration of the overall framework. Given a pair of vehicle images, the visual-spatio-temporal path proposal is generated by optimizing a chain MRF model with a deeply learned potential function. The path proposal is further validated by the Path-LSTM and regularizes the similarity score by Siamese-CNN to achieve robust re-identification performance.

Chain MRF model for visual-spatio-temporal Path Proposals

- $$p(\mathbf{x}|x_1 = p, x_N = q) = \frac{1}{Z}\psi(p, x_2)\psi(x_{N-1}, q)\prod_{i=2}^{N-2}\psi(x_i, x_{i+1}), \quad (1)$$

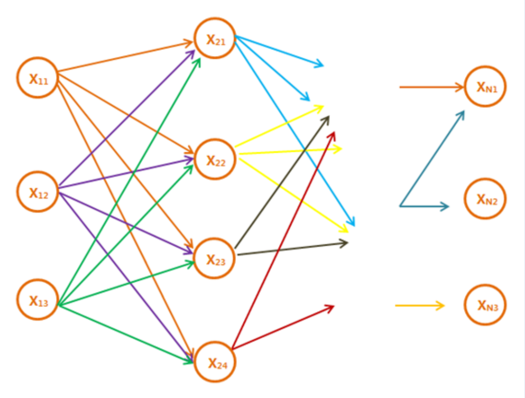
$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \; p(\mathbf{x}|x_1 = p, x_N = q), \quad (2)$$

$$\text{subject to} \quad t_{i,k_i^*} \le t_{i+1,k_{i+1}^*} \quad \forall i \in \{1, \cdots, N-1\}, \quad (3)$$

- $$\max_{\mathbf{x}} \; p(\mathbf{x}|x_1 = p, x_N = q) \quad (4)$$

$$= \frac{1}{Z}\psi(p, x_2)\psi(x_{N-1}, q)\max_{x_2}\cdots\max_{x_{N-1}}\prod_{i=2}^{N-1}\psi(x_i, x_{i+1}) \quad (5)$$

$$= \frac{1}{Z}\max_{x_2}\left[\psi(p, x_2)\psi(x_2, x_3)\left[\cdots\max_{x_{N-1}}\psi(x_{N-1}, x_q)\right]\cdots\right] \quad (6)$$



- $$S(\mathbf{x}^*) = \frac{1}{N-1}\left(\psi(p, 2) + \sum_{i=2}^{N-2}\psi(x_i^*, x_{i+1}^*) + \psi(x_{N-1}^*, q)\right) \quad (7)$$

Examples

(a) Invalid path. Empirical averaged potential: 0.946

(b) Valid path. Empirical averaged potential: 0.916

Figure 5: Examples of empirical averaged potential favoring longer paths. The invalid longer path in (a) has a higher averaged potential than the valid path in (b).

Siamese-CNN for chain MRF model

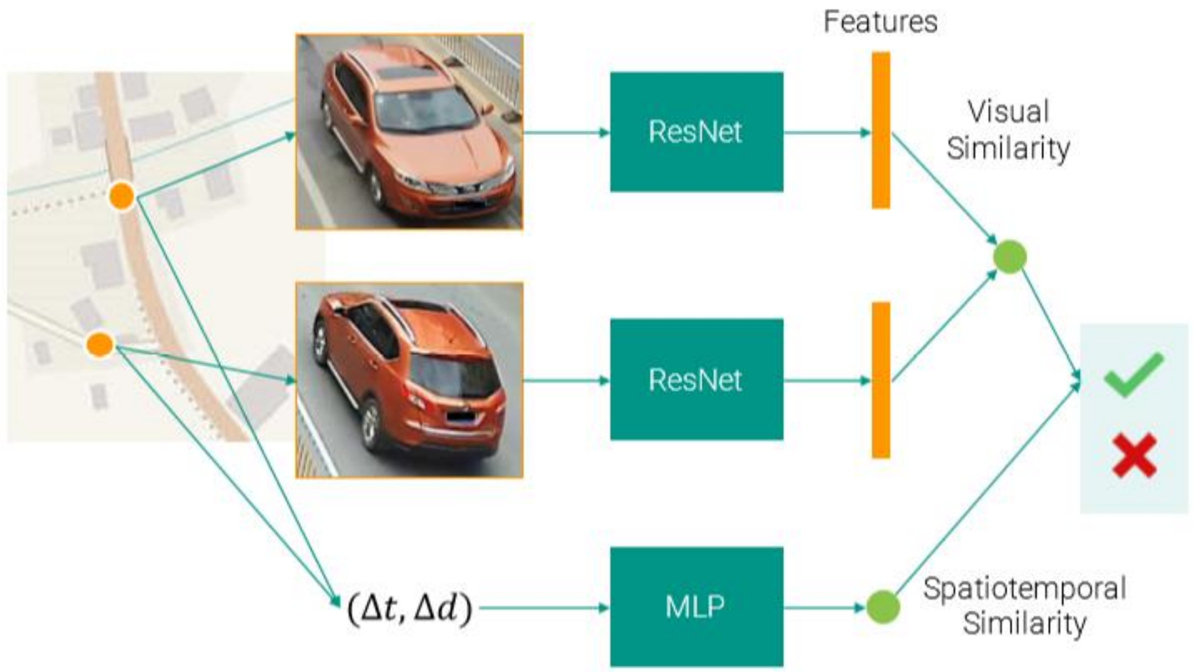


Figure 4: A Siamese-CNN is learned as the pairwise potential function for the chain MRF model, which takes a pair of visual-spatio-temporal states as inputs and estimates their pairwise similarity.

Figure 3: An example visual-spatio-temporal path proposal on the VeRi dataset [28] by our chain MRF model.

Path-LSTM

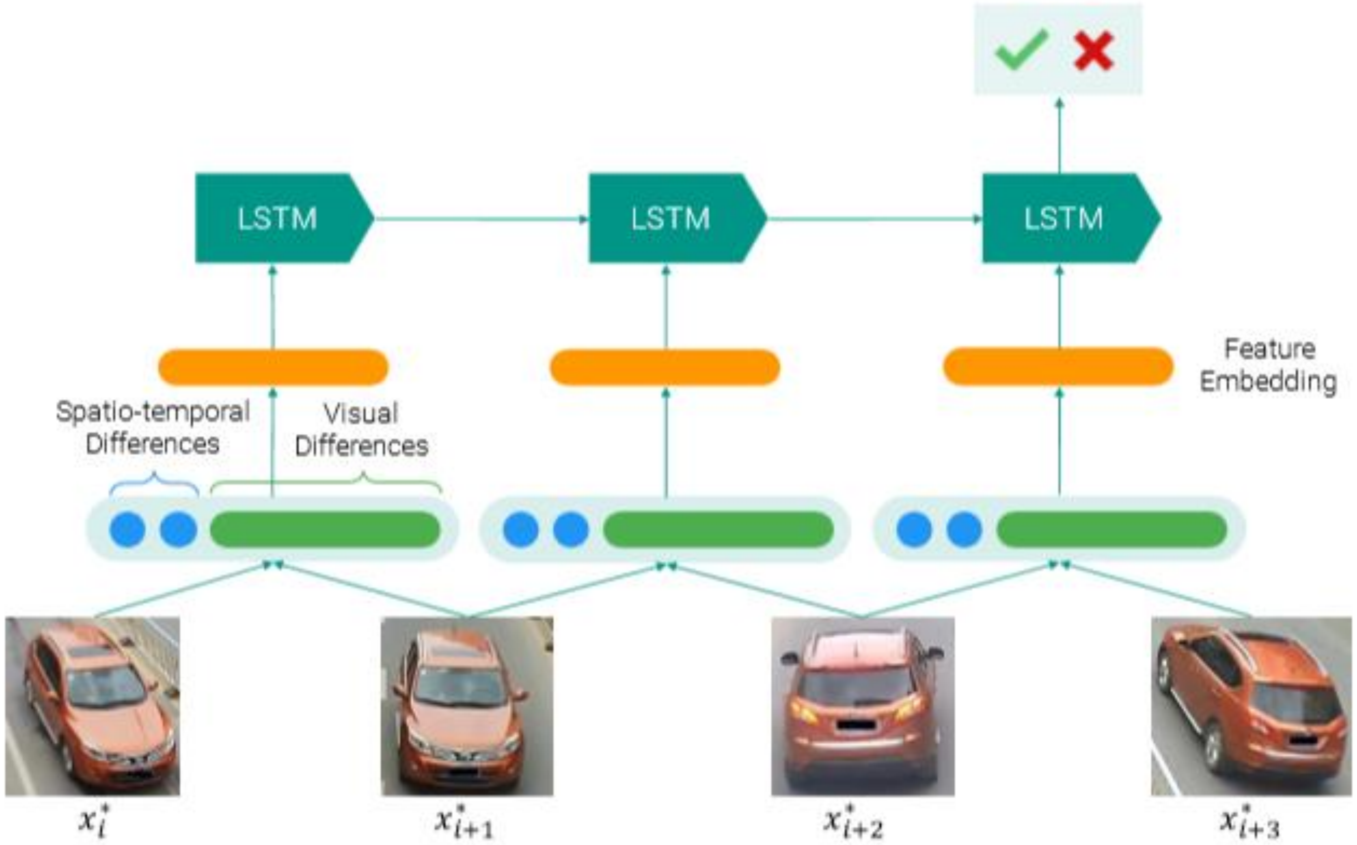


Figure 6: The network structure of the Path-LSTM. It takes visual and spatio-temporal differences of neighboring states along the path proposal as inputs, and estimates the path validness score.

Experiments

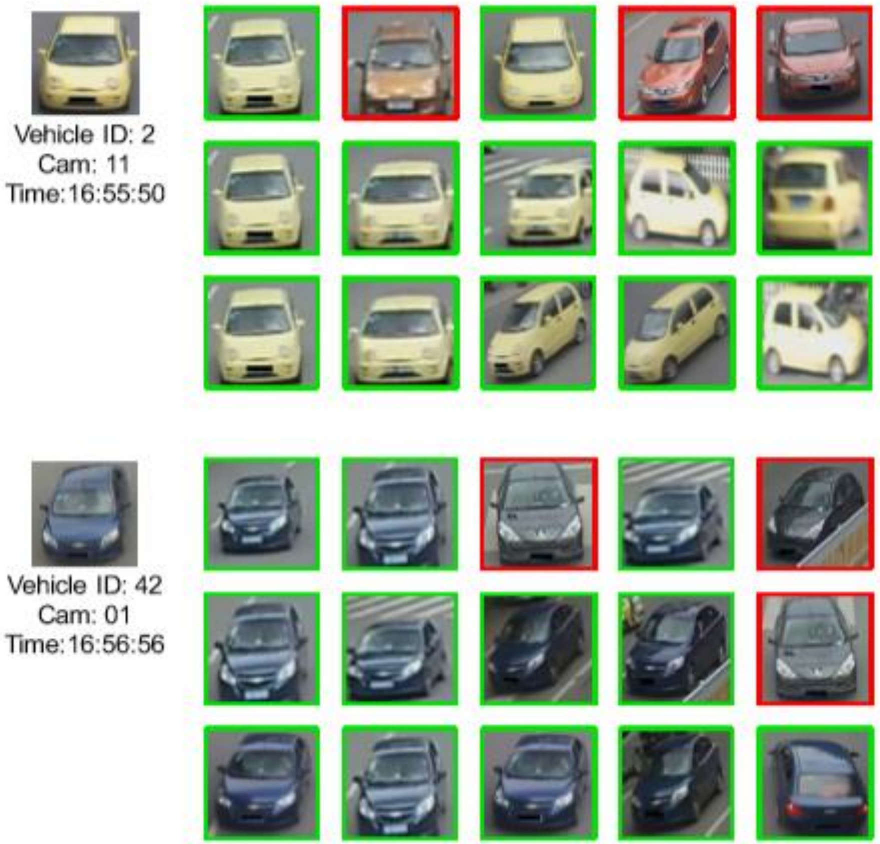


Figure 8: Example vehicle re-identification results (top5) by our proposed approach. The true positive is in green box otherwise red. The three rows are results of Siamese-Visual, Siamese-CNN and Siamese-CNN+Path-LSTM.

Method	mAP (%)
FACT [27]	18.49
FACT+Plate-SNN+STR [28]	27.77
Siamese-Visual	29.48
Siamese-Visual+STR	40.26
Siamese-CNN	54.21
Chain MRF model	44.31
Path-LSTM	54.49
Siamese-CNN-VGG16	44.32
Path-LSTM-VGG16	45.56
Siamese-VGG16+ PathLSTM-VGG16	46.85
Siamese-CNN+Path-LSTM	**58.27**

Table 1: mAP by compared methods on the VeRi-776 dataset [28].

Method	top-1 (%)	top-5 (%)
FACT [27]	50.95	73.48
FACT+Plate-SNN+STR [28]	61.44	78.78
Siamese-Visual	41.12	60.31
Siamese-Visual+STR	54.23	74.97
Siamese-CNN	79.32	88.92
Chain MRF model	54.41	61.50
Path-LSTM	82.89	89.81
Siamese-CNN-VGG16	54.41	61.50
Path-LSTM-VGG16	47.79	62.63
Siamese-VGG16+ PathLSTM-VGG16	50.95	61.62
Siamese-CNN+Path-LSTM	**83.49**	**90.04**

Table 2: Top-1 and top-5 accuracies by compared methods on the VeRi-776 dataset [28].

Overall framework

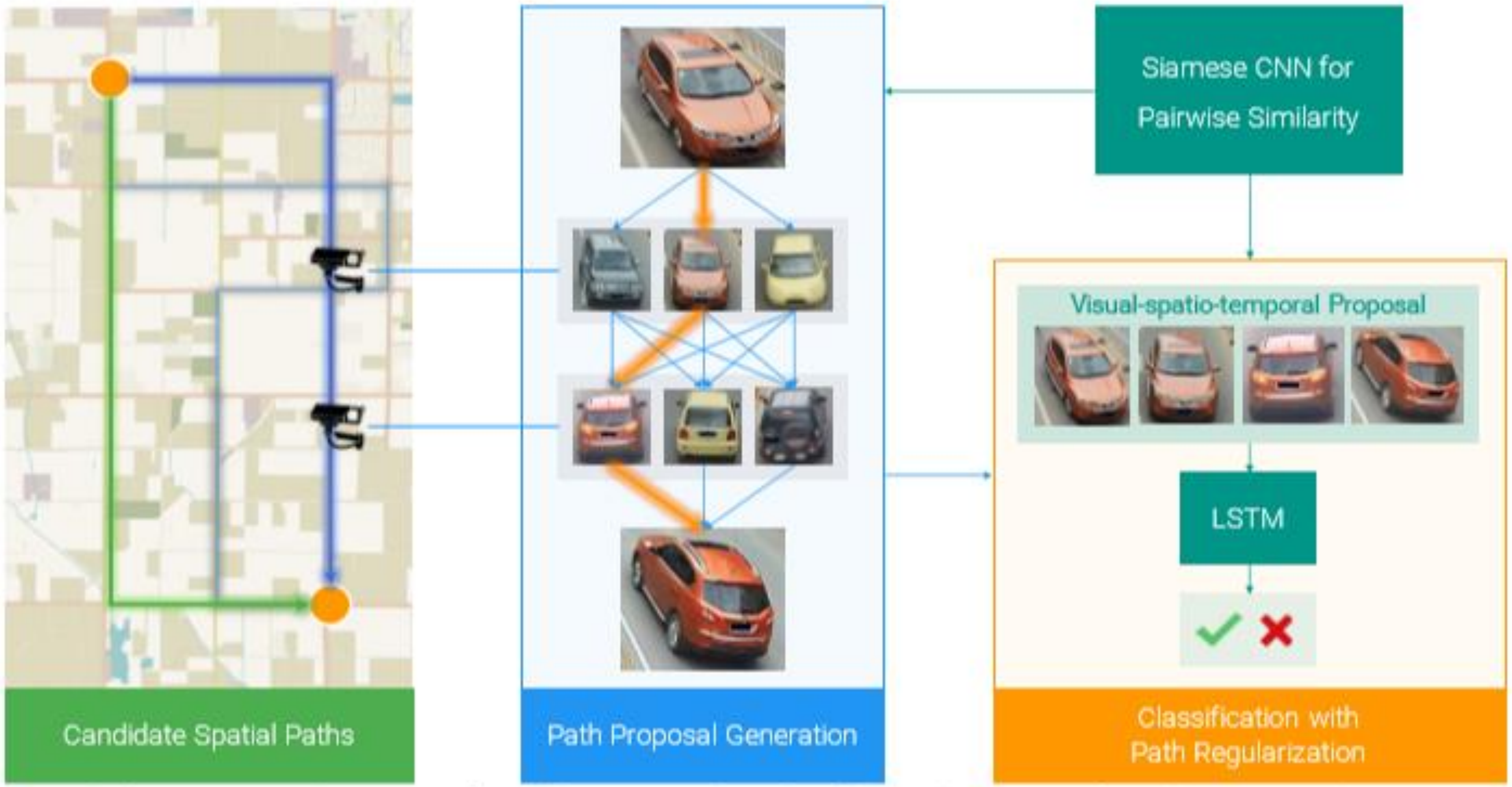


Figure 2: Illustration of the overall framework. Given a pair of vehicle images, the visual-spatio-temporal path proposal is generated by optimizing a chain MRF model with a deeply learned potential function. The path proposal is further validated by the Path-LSTM and regularizes the similarity score by Siamese-CNN to achieve robust re-identification performance.

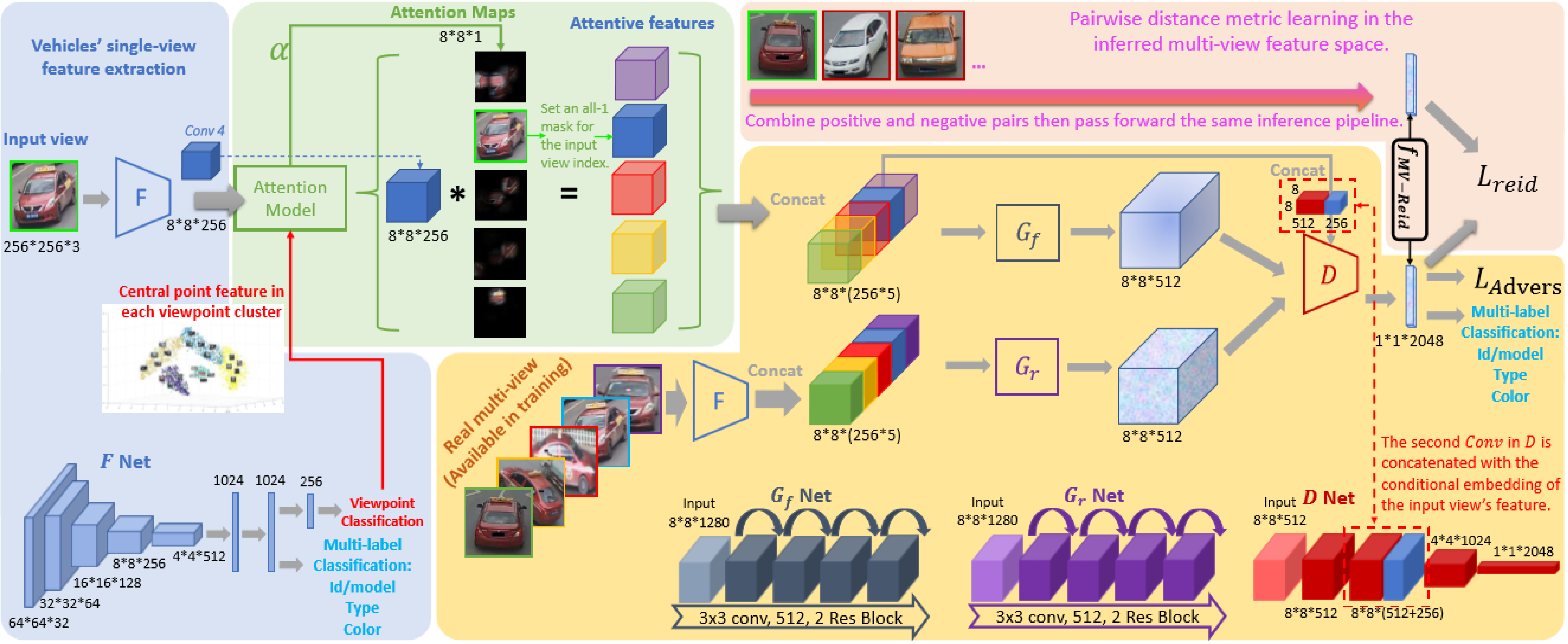
总体框架

Viewpoint-aware Attentive Multi-view Inference for Vehicle Re-identification

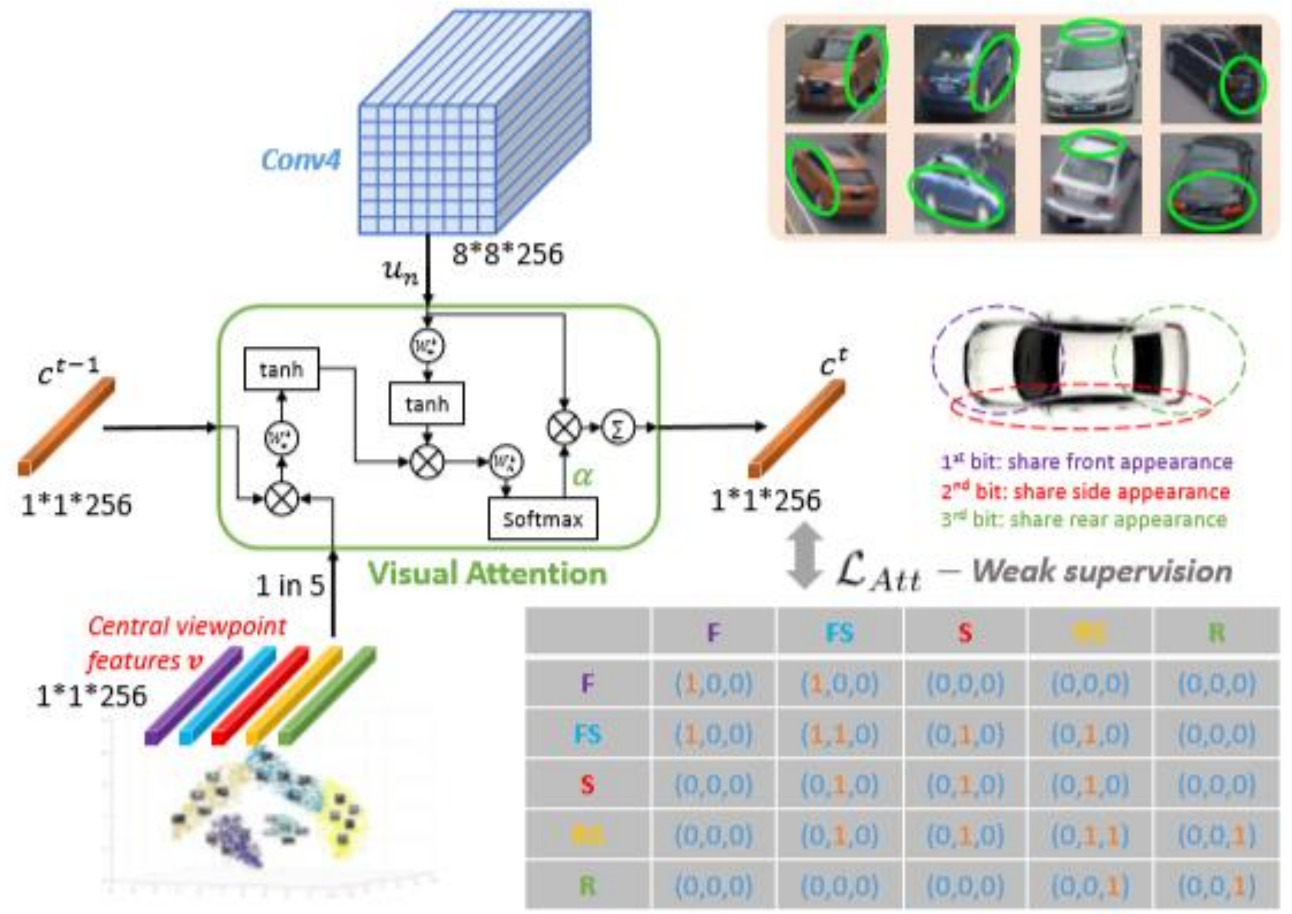
Yi Zhou Ling Shao
Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE School of Computing Sciences, University of East Anglia
y.zhou1@uea.ac.uk ling.shao@ieee.org

Reference: http://openaccess.thecvf.com/content_cvpr_2018/papers/Zhou_Viewpoint-Aware_Attentive_Multi-View_CVPR_2018_paper.pdf

Single-view --> Multi-view



Attention model



	F	FS	S	[illegible]	R
F	(1,0,0)	(1,0,0)	(0,0,0)	(0,0,0)	(0,0,0)
FS	(1,0,0)	(1,1,0)	(0,1,0)	(0,1,0)	(0,0,0)
S	(0,0,0)	(0,1,0)	(0,1,0)	(0,1,0)	(0,0,0)
[illegible]	(0,0,0)	(0,1,0)	(0,1,0)	(0,1,1)	(0,0,1)
R	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,1)	(0,0,1)

Figure 3. The details of the viewpoint-aware attention model. The top-right part gives examples of overlapped regions of certain arbitrary viewpoint pairs.

Attention results

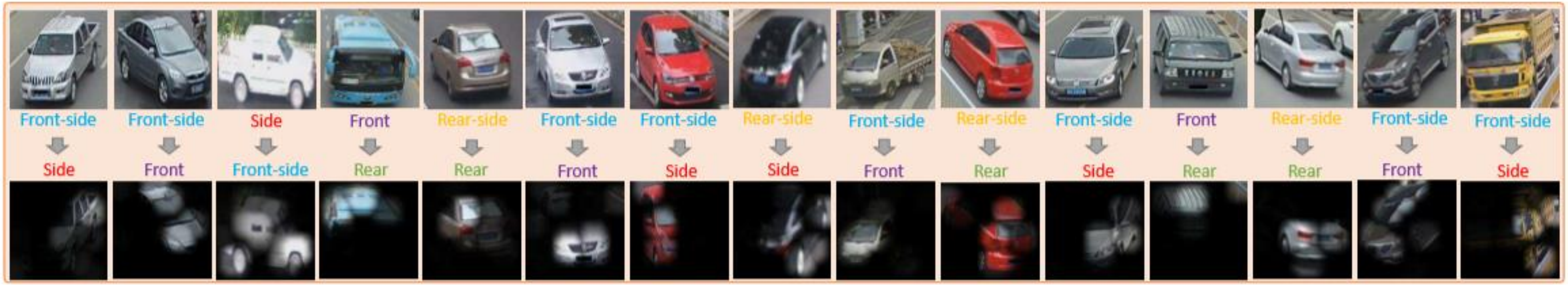


Figure 4. Viewpoint-aware attention maps. The upper row shows the input images and the bottom row shows the output attention maps. The highly-responded region is obtained by the input view attended with the central viewpoint feature of the target viewpoint.

Experiments

Table 3. Comparisons (%) with state-of-the-art re-ID methods. Methods in the last three rows include spatial-temporal (ST) information.

VeRi					VehicleID									
Settings	Query = 1678, Test = 11579				Settings	Test Size = 800			Test Size = 1600			Test Size = 2400		
Methods	mAP	r = 1	r = 5	r = 20	Methods	r = 1	r = 5	r = 20	r = 1	r = 5	r = 20	r = 1	r = 5	r = 20
LOMO [11]	9.78	23.87	39.14	57.47	LOMO [11]	19.76	32.01	45.04	18.85	29.18	39.87	15.32	25.29	35.99
DGD [28]	17.92	50.70	67.52	79.93	DGD [28]	44.80	66.28	81.52	40.25	65.31	76.76	37.33	57.82	70.25
GoogLeNet [29]	17.81	52.12	66.79	78.77	GoogLeNet [29]	47.88	67.18	78.46	43.40	63.86	74.99	38.27	59.39	72.08
FACT [15]	18.73	51.85	67.16	79.56	FACT [15]	49.53	68.07	78.54	44.59	64.57	75.30	39.92	60.32	72.92
XVGAN [41]	24.65	60.20	77.03	88.14	XVGAN [41]	52.87	80.83	91.86	49.55	71.39	81.73	44.89	66.65	78.04
SiameseVisual [23]	29.48	41.12	60.31	79.87	VGG+CCL [13]	43.62	64.84	80.12	39.94	62.98	76.07	35.68	56.24	68.41
OIFE [26]	48.00	65.92	87.66	96.63	MixedDiff+CCL [13]	48.93	75.65	88.47	45.05	68.85	79.88	41.05	63.38	76.62
VAMI (Ours)	**50.13**	**77.03**	**90.82**	**97.16**	VAMI (Ours)	**63.12**	**83.25**	**92.40**	**52.87**	**75.12**	**83.49**	**47.34**	**70.29**	**79.95**
SiameseCNN+PathLSTM [23]	58.27	83.49	90.04	96.03	No ST information	-	-	-	-	-	-	-	-	-
SiameseVisual([23])+STR([15])	40.26	54.23	74.97	91.68		-	-	-	-	-	-	-	-	-
VAMI (Ours) + STR([15])	**61.32**	**85.92**	**91.84**	**97.70**		-	-	-	-	-	-	-	-	-

Summary

- 我们使用时空模型，一般是把其看作辅助的信息，并且推测这些信息怎么转换。
我们现在的想法有**贝叶斯**和看成**图**去考虑

- 第一篇论文把这些特征看成时序数据，并用时序的解决方式来对结果进一步蒸馏，
或许在行人识别中，也可以尝试把这些信息作为时序转换来正则化视觉抽取的效果。

- 第二篇论文用了attention，GAN，MultiTask等多种方法，实际目的就是把输入图像的
不同**空间映射**到一个共同的而且维度挺高的**空间**。但这个映射过程几乎靠着堆叠硬生生地完成的。
启示就是：把维度有限的图像空间映射到高维的图像空间，这个过程不要企图依赖一个网络就能完成。需要多个网络共同协助。而且这个方法类似行人重识别的**pose生成**

- 我们会发现用了时空信息的方法，比单纯的使用视觉特征**优雅**了很多，
而且效果也不会差太多了。